Peoples Opinion on Indian Budget Using Sentiment Analysis Techniques

Bharat Naiknaware, Research Student, Dept. Of CS& IT, Dr. BAMU Aurangabad, India,

bbharat.naiknaware@gmail.com

Seema S. Kawathekar, Assistant Professor, Dept. Of CS& IT, Dr. BAMU, Aurangabad, India

Abstract: **Social media today has becomes a very popular tool in society. Social media not only acts as a proxy for the real world society, it also offers a treasure trove of data for different types of analyses like Trend Analysis, Event Detection and Sentiment Analysis specifically deals with the task of Sentiment Analysis in Twitter. Sentiment Analysis in social media in general and Twitter in particular has a wide range of applications Companies/services can gauge the public sentiment towards the new product or service they launched, political parties can estimate their chances of winning the upcoming elections by monitoring what people are saying on Twitter about them, and so on. Social media is a rich source of data for opinion mining and sentiment analysis. In this paper we use the twitter data of Indian Budget and predict the sentiment analysis. Here we are using last three year budget data for find the peoples opinion and calculate in the technical sense like compute the polarity of particular budget data. we build a sentiment classifier, which is able to determine positive, negative, and neutral sentiment of each tweet.**

Keywords —Sentiment Analysis, SVM, Naïve Bayes, Polarity, LDA, TwitteR, API.

I. INTRODUCTION

Sentiment Analyses is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral. There are there types of Sentiment Analysis [4]:

1) Document level Sentiment Analysis: This is the simplest form of classification. The whole document of opinionated text is considered as basic unit of information. This approach is not suitable if document contains opinions about different objects as in forums and blogs. Classification for full document is done as positive or negative. Irrelevant sentences need to be eliminated before processing.

2) Sentence level sentiment Analysis: This is the most finegrained analysis of the document. In this, polarity is calculated for each sentence as each sentence is considered as separate unit and each sentence can have different opinion. Sentence level sentiment analysis has two tasks: Subjectivity Classification and Objectivity of Classification

3) Feature Level Sentiment Analysis: The basic step in Feature Level sentiment analysis is to identify the piece of text as a feature of some product. For example Battery life is very long lasting. In this review Battery is product feature (noun) and 'very long lasting' is opinion word (adjective).Social media today has becomes a very popular tool in society. Millions of messages are flowing daily in popular social media websites like Twitter, Facebook. On these social media web sites, a huge number of users share their view on a number of topics and discuss current issues of the society. Due to easy accessibility of these social media and free format of message a large number of users shifted from traditional communication tools like email system to these social media sites. These social media sites becomes valuable sources of people's opinions and sentiment as very large number of users post their opinions about products and services, express their views on political and religious issues, and share their thought about current problems of the society. Such data may be used as an important resource for marketing or social studies. Twitter contains a very large number of short messages. The maximum length of a twitter message is 280 characters. The twitter message also called tweet. We use a dataset of tweets downloaded from twitter using its APIs. The contents of the tweet vary from personal opinion to public statements. Due to freely availability of a large number of tweets, twitter data can be used in opinion mining and sentiment analysis. Many manufacturing company may be interested to know public opinions about their product so that accordingly they can improve its quality. Political parties want to know, whether people like their manifesto or not. All these information may be obtained from twitter data, as a large number of tweets on different topics are posted by users daily. In this paper, we study how sentiment analysis may be performs on twitter data. We will show how to use twitter data as a corpus for sentiment analysis. We use twitter data for the sentiment analysis due to following reasons:

• Twitter social media is used by a large number of people to express their opinion about different topics, thus it is a valuable source of people's opinions.

• Twitter contains a huge number of tweets and it grows every day. The collected corpus may be very large.

• Twitter's users vary from regular users to renown, company representatives to politicians, and even country ministers.

Therefore, it is possible to collect tweets from Different interests group of users.

We collected tweets of three consecutive year Union Budget of year 2016, 2017 and 2018
Distributed among three sets of tweets:

1. Tweets contains positive sentiment
2. Tweets that contains negative sentiment
3. Tweets that only state a fact do not express any sentiment

We use statistical methods to build a sentiment classifier that use the collected csv file of budget 2016, 2017 and 2018 as training and test data. This classifier models may be used in determining the sentiment of a new tweet as positive, negative, or neutral.

I. Survey of Literature

Author Jahiruddin works on paper Sentiment Analysis of Twitter Data using Statistical Methods in this paper they use statistical method for computation of sentiment is positive negative or neutral of collected a corpus of 3100 tweets on three different events “Gaza under attack”, “Delhi Assembly Election 2015”, and “Union Budget 2015” and distributed among three sets of tweet. First we use the Latent Dirichlet Allocation (LDA) to identify the key terms. These key terms are used to represent each tweet in n dimensional vector. Using this tweet vectors, they build a sentiment classifier. The binary feature vectors of the tweets are used as input for sentiment analysis. We generate input file for sentiment classifier using these feature vectors. The Weka’s Naive Bayes classifiers are used to classify the tweets as positive, negative, or neutral tweet depending upon their text. Naive Bayes classifier is a probabilistic classifiers based on Bayes' theorem [1]. Authors Satarupa Guha _, Aditya Joshi _, Vasudeva Varma works on paper Sentibase: Sentiment Analysis in Twitter on a Budget approach of this paper is Preprocessing, Vocabulary Generation Feature Engineering The task required us to classify a tweet into positive, negative and neutral polarity categories. This can essentially be treated as a 2-step process Classify each tweet into subjective (positive/negative) and objective (neutral) classes._ Classify subjective tweets into positive and negative ones. Author used the official training and test sets provided for the SemEval 2015 task to train and evaluate system. Tweets in the training data that were not available any more through the Twitter API were removed from the training set. For the evaluation, they compute precision, recall and F1 measures as computed by the scorer package provided for the task [2].Jatinder Kaur1 works on paper A Review Paper on Twitter Sentiment Analysis Techniques author review the techniques of Sentiment Analysis In this paper, we provide a survey and comparative study of existing techniques for opinion mining including machine learning and lexicon-based approaches, Research results show that machine learning methods, such as SVM and naive Bayes have the highest accuracy and can be regarded as the baseline learning methods. Author using supervised and discuss the measure challenges in Sentiment Analysis like Identifying subjective parts of text, Domain dependence, Sarcasm Detection, Thwarted expressions, Explicit Negation of sentiment, Order dependence, Entity Recognition, Building a classifier for subjective vs. objective tweets, Handling comparisons, Applying sentiment analysis to Facebook messages and Internationalization[3].

Data Analysis has a variety of angles and methods that combines many techniques in order to provide better accuracy. One of the most popular methods of data analysis technique is data mining that mainly concentrates on modeling and discovery of knowledge for prediction process rather than descriptive purposes. Predictive analytics is mainly used for predicting forecasting/classification whereas text analytics make use of statistical, linguistic and structural techniques in order to retrieve information from text sources. This text sources are mostly in the form of unstructured data. Sentiment analysis (or) opinion mining plays a significant role [11]. The machine learning based text classifiers learn the set of rules (the decision criterion of classification) automatically from the training data. This clearly indicates that machine learning based text classification is a kind of supervised machine learning paradigm, where the classifier needs to be trained on some labeled training data before it can be applied to actual classification task. Usually the training data is an extracted portion of the original data hand labeled manually. Once the algorithm is trained to correctly classify the documents in the training set, it can be applied to the unseen data. If the learning method is statistical, the classifier is called a statistical text classifier [12].

II. Methodology

Implementation of Sentence level sentiment analysis we are used R Programming Open Source tool for better implementation of Sentiment Analysis approach. A dataset is created using twitter posts of budget related Tweets are short messages with full of slang words and misspellings. So we perform a sentence level sentiment analysis. This is done in three phases. In first phase preprocessing is done. Then a feature vector is created using relevant features. Finally using different classifiers, tweets are classified into positive and negative classes. Based on the number of tweets in each class, the final sentiment is derived.

A. Data Acquisition:

Standard twitter dataset is not available for budget related data domain, we created a new dataset by collecting tweets over a period of time ranging from year 2016 to 2018 year budget day time data. Tweets are collected automatically using Twitter API with Twitter Application Development and they are manually annotated as positive or negative. We are creating Twitter Development Application account after that twitter provide access key, Secret Key, access Token and Application Authentication ID this credential used for fetching data from Twitter Account.We are using following code for extracting the dataset from twitter API

```
library(twitteR)
library(ROAuth)
library(plyr)
CUSTOMER_KEY <- ".............."
CUSTOMER_SECRET <-"............"
ACCESS_TOKEN <- ".............."
ACCESS_secret <- "............."
setup_twitter_oauth(CUSTOMER_KEY, CUSTOMER_SECRET,
ACCESS_TOKEN, ACCESS_secret)
```

```
1
Budget2018 = searchTwitter("#Budget2018 ", n= 2000,
lang="en",since='2018-01-20', until='2018-02-10')
Budget2018
tweets_df = twListToDF(Budget2018)
write.csv(tweets_df,
file='C:/Users/BHARAT/Desktop/DataSet/Budget2018.csv',
row.names=F)
```

B. Preprocessing of Tweets

Keyword extraction is difficult in twitter due to misspellings and slang words. So to avoid this, a preprocessing step is performed before feature extraction. Preprocessing steps include removing URL, avoiding misspellings and slang words. Misspellings are avoided by replacing repeated characters with 2 occurrences. Slang words contribute much to the emotion of a tweet, Remove RT, Remove Hashtags, Remove Controls and special characters, Remove Controls and special characters, Remove Punctuations, Remove leading whitespaces, Remove trailing whitespaces, Remove extra whitespaces . So they can't be simply removed. Therefore a slang word dictionary is maintained to replace slang words occurring in tweets with their associated meanings. Domain information contributes much to the formation of slang word dictionary.

```
Library (stringr)
Library(tm)
Dataset <-
read.csv('C:/Users/BHARAT/Desktop/DataSet/Paper/Budget2018.
csv')
Dataset$text <- as.factor(Dataset$text)
pos.words <-
scan('C:/Users/BHARAT/Desktop/DataSet/Paper/pos_words.txt',
what='character', comment.char=';') #folder with positive
dictionary
neg.words <-
scan('C:/Users/BHARAT/Desktop/DataSet/Paper/neg_words.txt',
what='character', comment.char=';') #folder with negative
dictionary
#pos.words <- c(pos, 'upgrade')
#neg.words <- c(neg, 'wtf', 'wait', 'waiting', 'epicfail')
score.sentiment <- function(sentences, pos.words, neg.words,
.progress='none')
{
 require(plyr)
 require(stringr)
 scores <- laply(sentences, function(sentence, pos.words,
neg.words){
   sentence = gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", sentence)
   sentence <- gsub('[[:punct:]]', "", sentence)
   sentence <- gsub('[[:cntrl:]]', "", sentence)
   sentence <- gsub('\\d+', "", sentence)
sentence = gsub('(RT|via)((?:\\b\\W*@\\w+)+)', ", sentence)
sentence <- tolower(sentence)
word.list <- str_split(sentence, '\\s+')
words <- unlist(word.list)
pos.matches <- match(words, pos.words)
neg.matches <- match(words, neg.words)
pos.matches <- !is.na(pos.matches)
neg.matches <- !is.na(neg.matches)
score <- sum(pos.matches) - sum(neg.matches)
return(score)
}, pos.words, neg.words, .progress=.progress)
 scores.df <- data.frame(score=scores, text=sentences)
 return(scores.df)
}
scores<-score.sentiment(Dataset$text, pos.words, neg.words,
.progress='text')
write.csv(scores,file=('C:/Users/BHARAT/Desktop/DataSet/Paper
/ScoreBudget2018.csv'), row.names=TRUE
```

C. Creation of Feature Vector

Feature extraction is done in two steps. In the first step, twitter specific features are extracted. Hash tags and emoticons are the relevant twitter specific features. Emoticons can be positive or negative and Neutral. So they are given different weights. Positive text is given a weight of '1' and negative text is given a weight of '-1' and neutral text give weight is 0. There may be positive and negative hash tags. Therefore the count of positive hash tags and negative hash tags are added as two separate features in the feature vector. The process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral.

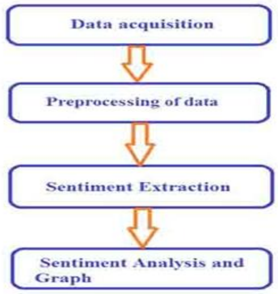

Fig 1Methodology for Sentence Level Sentiment Analysis

III. Results

The approach followed here is to count the positive and negative words in each tweet and assign a sentiment score. This way, we can ascertain how positive or negative a tweet is. Nevertheless, there are multiple ways to calculate such scores; here is one formula to perform such calculations.

Score = Number of positive words - Number of negative words
If Score > 0, means that the tweet has 'positive sentiment'
If Score < 0, means that the tweet has 'negative sentiment'
If Score = 0, means that the tweet has 'neutral sentiment'
Results are computed in R Open Source tool here we are using twitteR, plyr, ROAuth, stringr ggplot2 packages for calculating polarity Score. The scores are weighted averages of the individual scores of news / blogs / twitter during the day. Upon request, we can provide in-depth details on our computation. To put it simple,

we handle textual financial data using Data Mining and Text Mining methods. Our engine does the following steps in order to generate sentiment score:

Filtering and selection of relevant information per topic
Computation of relevance score per tweet
Computation of sentiment score per tweet
Computation of weighted average score of the individual topic.
Three main methods can be provided:

A Statistical Approach:

The statistical approach itself is not binary. It is model based and can be easily calibrated. Below is a non-exhaustive list of the outputs:
Count of positive words, Count of negative words
Weight of the sentence positions, Relevance score
Volume of news available, Buzz score. It still allows for word disambiguation. This method is fast, scalable (especially across languages), but lacks indeed the precision that the grammar based approach can have.

B Grammar Based Approach.

By far the largest selection of technologies for exploiting grammar in sentiment analysis come from the use of HMM- or CRF-type sequence modeling, and consequently, this will be a major component of the course. This type of machine learning uses syntactic and other features as binary-valued functions in learning to label windows of text.

Machine /Deep Learning Based Approach.

The deep learning model actually builds up a representation of whole sentences based on the sentence structure. It computes the sentiment based on how words compose the meaning of longer phrases. This way, the model is not as easily fooled as previous models.

Table 1 Polarity Prediction

	Budget Analysis			
Dataset	**Positive**	**Negative**	**Neutral**	**Total Tweets**
Budget2016	7791	4662	12547	25000
Budget2017	1830	304	2866	5000
Budget2018	495	298	879	1672

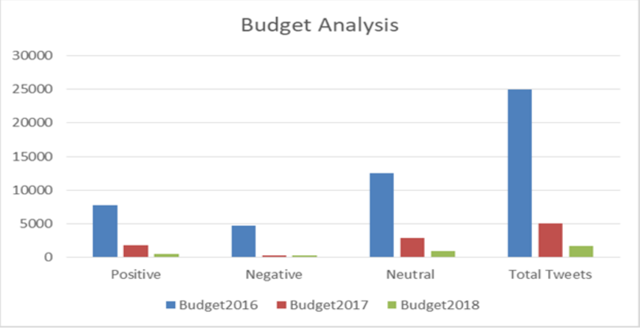

Fig 1 Budget Analysis

IV. CONCLUSION

In this work the Union Budget of India for the year 2016 to 2018 are analyzed and sentiment analysis is carried out on the text data. In this paper we are used Budget data from 2016 to budget 2018 and find peoples opinion on particular budget on the basis of Dataset polarity that is Positive, Negative and Neutral so we got results like in Budget 2016 more peoples are positive polarity 7791 than Negative polarity is 4662 as well as Neutral polarity is 12547 peoples are also high, in Budget 2017 study we got results of positive peoples are again high 1830, Negative peoples is 304 and 2866 is Neutral opinions and in Budget 2018 we studied the opinion of peoples the positive polarity is high 495 and negative polarity 298 and Neutral polarity 879.Overall analysis of peoples of every year peoples are positive opinion against declared Budget.

For the purpose of this we are used Sentence level Sentiment Analysis Techniques and experiments done in R tools.

ACKNOWLEDGMENT

We are author's sincerely thanks to Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University for providing research laboratory facility for completion of this work. Also we would like to thanks Ministry of Electronics & Information Technology (MeitY), Government of India for provided Fellowship under VISVESVARAYA PHD SCHEME FOR ELECTRONICS & IT.

REFERENCES

[1] Jahiruddin, "Sentiment Analysis of Twitter Data using Statistical
Methods," *International Journal of Innovative Research in Engineering & Management (IJIREM) ISSN: 2350-0557, Volume-2, Issue-4, Page no-30-34 July 2015*

[2] Satarupa Guha[1], Aditya Joshi[1], Vasudeva Varma, "Sentibase: Sentiment Analysis in Twitter on a Budget" *SEM 4th Joint Conference on Lexical and Computational Semantics Denver, Colorado, USA Report No: IIIT/TR/2015/-1.*

[3] Jatinder Kaur[1], "A Review Paper on Twitter Sentiment Analysis
Techniques," *International Journal for Research in Applied Science & Engineering Technology (IJRASET) Volume 4 Issue X, October 2016 IC Value: 13.98 ISSN: 2321-9653 page no-61-70.*

[4] Seema Kolkur[1], Gayatri Dantal[1] and Reena Mahe‡*, "Study of Different Levels for Sentiment Analysis" International Journal of Current Engineering and Technology E-ISSN 2277 – 4106, P-ISSN 2347 – 5161Vol.5, No.2 April 2015 page no-768-770.

[5] Neha Upadhyay1, Prof. Angad Singh2, "A Survey on Twitter for Sentiments Analysis Using Machine Learning Methods" International Journal of Engineering Science and Computing, May 2016 Volume 6 Issue No. 5 DOI 10.4010/2016.1113 ISSN 2321 3361 page no 4890-4893

[6] Anuj Verma[1], Kunwar Abhay[1] Pratap Singh[2], Kakali Kanjilal[3], "Knowledge Discovery And Twitter Sentiment

Analysis: Mining Public Opinion And Studying Its Correlation With Popularity Of Indian Movies,” International Journal Of Management (Ijm) ISSN 0976-6502 (Print) ISSN 0976-6510 (Online) Volume 6, Issue 1, January (2015), pp. 697-705.

[7] Moonis Shakeel, Vikram Karwal “Lexicon-based Sentiment Analysis of Indian Union Budget 2016-17),” 978-1-5090-2684-5/16/$31.00 ©2016 IEEE.

[8] Prof. Durgesh M. Sharma, Prof. Moiz M. Baig “Sentiment Analysis on Social Networking: A Literature Review,” International Journal on Recent and Innovation Trends in Computing and Communication Volume: 3 Issue: 2 ISSN: 2321-8169 Page no-22-27.

[9] Yi Ding[1], Bing Li2, 3*, Yuqi Zhao[2], Can Cheng2, “Scoring Tourist Attractions Based on Sentiment Lexicon,” 978-1-4673-8979-2/17/$31.00 ©2017 IEEE page 1990 to 1993.

[10] Salas, A., Georgakis, P., Nwagboso, C., Ammari, A. and Petalas, I.
”Traffic Event Detection Framework Using Social Media” 2017 IEEE International Conference on Smart Grid and Smart Cities 23-26 July 2017 978-1-5386-0504-2/17/$31.00 ©2017 IEEE page no-303-307.

[11] Ms.A.M.Abirami, Ms.V.Gayathri “A SURVEY ON SENTIMENT ANALYSIS METHODS AND APPROACH” 2016 Eighth International Conference on Advanced Computing (ICoAC) 19-21 Jan. 2017 Electronic ISBN: 978-1-5090-5888-4 Print on Demand(PoD) ISBN: 978-1-5090-5889-1 DOI: 10.1109/ICoAC.2017.7951748.

[12] P. Waila*, Marisha**, V.K. Singh***and M.K. Singh* “Evaluating Machine Learning and Unsupervised Semantic Orientation Approaches for Sentiment Analysis of Textual Reviews” 2012 IEEE International Conference on Computational Intelligence and Computing Research 18-20 Dec. 2012 Electronic ISBN: 978-1-4673-1344-5 Print ISBN: 978-1-4673-1342-1 CD-ROM ISBN: 978-1-4673-1343-8 Print on Demand(PoD) ISBN: 978-1-4673-1342-1 DOI: 10.1109/ICCIC.2012.6510235.