ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

Twitter Sentiment Analysis Using Machine Learning and Ontology

Amol S. Gaikwad[1], Anil S. Mokhade[2]

Lecturer, Department of Computer Engineering, Govt. Polytechnic Gadchiroli, Maharashtra, India[1]

Associate Professor, Department of Computer Science Engineering, VNIT, Nagpur, Maharashtra, India[2]

ABSTRACT: Sentiment analysis is a subfield of natural language processing which involves identifying people's opinion in a given text. Information available from social networks can be efficiently utilised for opinion mining. Twitter is one of the most popular and rapidly growing microblogging service. Twitter sentiment analysis involves exploring people's opinions, appraisals, views, emotions towards events, entities and products. This papers introduces a combine approach to using machine learning and ontology. Sentiment analysis is performed with naive bayes classifier and feature analysis of mobile phone brands is done using domain ontology and result of earlier classification. A shinydasboard application is created which enables real time analysis using proposed approach. The aim of paper is fine grain analysis of tweets and to present the results of analysis in an application with which users can interact . Such approach can help businesses guess customers opinions and preferences and could prove helpful in shaping business strategy.

KEYWORDS: Sentiment Analysis, Twitter, Ontology ,Naive bayes, Document Classification, Shinydashboard, Shiny

I. INTRODUCTION

Sentiment analysis involves extracting, classifying, determining people's opinion in various types of contents. Due to increasing popularity of social networking, blogging and micro blogging websites huge volume of opinion rich information is generated each day which can be effectively utilised to predict online customers opinion and preferences, this can be very useful for business and market research. Twitter is one such source of information. On twitter user creates status messages called as tweets ,these tweets can be of maximum 140 characters in length. Twitter has several unique characteristics such as twitter api, domains, language, emoticons etc. that differentiates it from other sources of text and so is it's analysis.

Twitter sentiment analysis has wide business applications. Marketing businesses can use it to research public opinion about their product and services, or to find customer satisfaction. Organisation can also use this as a critical feedback about problems in their products and services. Companies can compare their performance with that of their rivals and help them reshape their business strategy on time. Twitter sentiment analysis proves to be cost and time efficient compared to traditional methods of collecting feedback and surveys. One of the most promising applications of sentiment analysis of tweets could be in the field of economic and financial modelling. Such analysis not only help businesses but also assists customers who want to search the sentiment of products before purchase. This can give a technological edge to the organisation in a very competitive market.

Large amount of research has been done in the field of sentiment analysis or opinion mining. Various sources of text like blogs, review sites, e-commerce sites, micro-blogging sites, comments have been used for opinion mining. Various methodologies has been used for analysis of twitter. Lexicon based method uses dictionary called as opinion lexicon to identify positive and negative words in the tweets [1]. Machine Learning methods has widely been used in classification problems, machine learning methods makes use of two sets of documents training sets and testing sets [13]. Ontology based approaches make use of ontologies for more fine grained analysis [4], [3]. Label Propogation method uses weighted graphs to solve the problem [14], [5]. This paper proposes a combine approach to twitter sentiment analysing using machine learning and ontology. Most of the existing ontology based techniques have

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

collected tweets on object attribute pair and perform sentiment analysis on such tweets. Our proposed system not only classify tweets but also identifies relevant domain features from it. We used machine learning methods particularly naive bayes classifier to classify tweets and mobile phone domain ontology to identify feature if it is present.

II. SENTIMENT ANALYSIS AND ONTOLOGY

An ontology can be defined as an ‘‘explicit, machine-readable specification of a shared conceptualization’’.An ontology is a formal naming and definition of the types, properties, and interrelationships of the entities that really or fundamentally exist for a particular domain of discourse. It is thus a practical application of philosophical ontology, with a taxonomy. Ontologies are used for modelling the terms in a domain of interest as well as the relations among these terms and are now applied in various fields, like agent and knowledge management systems and e-commerce platforms. Other applications include natural language generation, intelligent information integration, semantic-based access to the Internet and extracting information from texts.

Ontology based techniques can be used for more fine grained analysis of twitter posts. Ontologies can be created for domains under considerations for sentiment analysis. Reference [6] have refined the ontologies for mobile phones into three categories smart phones, wet and dirty mobile phones and simple mobile phones. The Knowledge represented in such domain ontologies can be used to extract relevant tweets and then perform more fine grained analysis of tweets. Reference [4] introduced an ontology based sentiment analysis of twitter posts. In their work they created domain ontology by extracting objects and attributes from retrieved tweets. They have used ‘smartphone’ as particular domain. Various semi-automatic learning techniques like OntoGen or manual methods can be used to create domain ontology. After creating domain ontology they used it to extract tweets that contain objects and attributes. The extracted tweets are then submitted to OpenDover for assigning sentiment score to the tweets. Reference [3] build a custom sentiment analysis tool based on ontology for twitter posts. They used laptops as a domain. They extracted tweets relevant to the feature of a particular domain here laptops and then assigned score for each feature using sentiment package of R.

III. DESCRIPTION OF PROPOSED APPROACH

The proposed approach uses supervised learning method to classify tweets. We particularly used naive bayes classifier using unigram approach.

1. System Architecture

The architecture of the proposed system consist of many phases connected sequentially as shown in figure 1. The proposed system automatically collect and classify tweets. Domain ontology is used at various stages for performing feature analysis.

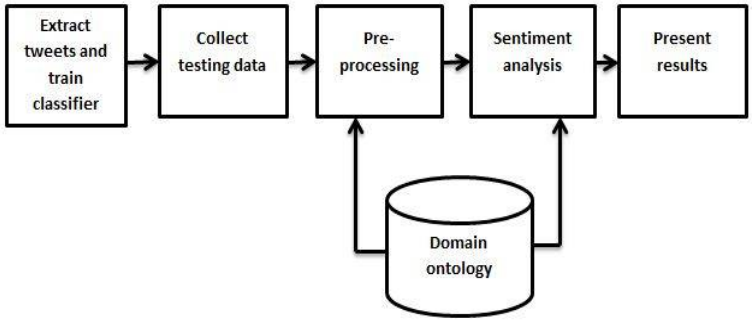

Fig. 1 System Architecture

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

2. Creation of Domain Ontology

Ontologies can be created using various methods, ontologies are created to identify objects, attributes and their relationship for domain under consideration. We have created domain ontology of mobile phone. Such ontology contains general features of simple as well as smart phones. Many tools can be used for crating domain ontology like TODE, OntoGen, Protege etc. Similarly various ontology languages are also available like FOAF, RWF, OWL etc. We have use Protege software and OWL language to create mobile phone ontology. We have also created OWL graph of the ontology for representing the ontology diagrammatically. The following fig shows the mobile phone domain ontology.

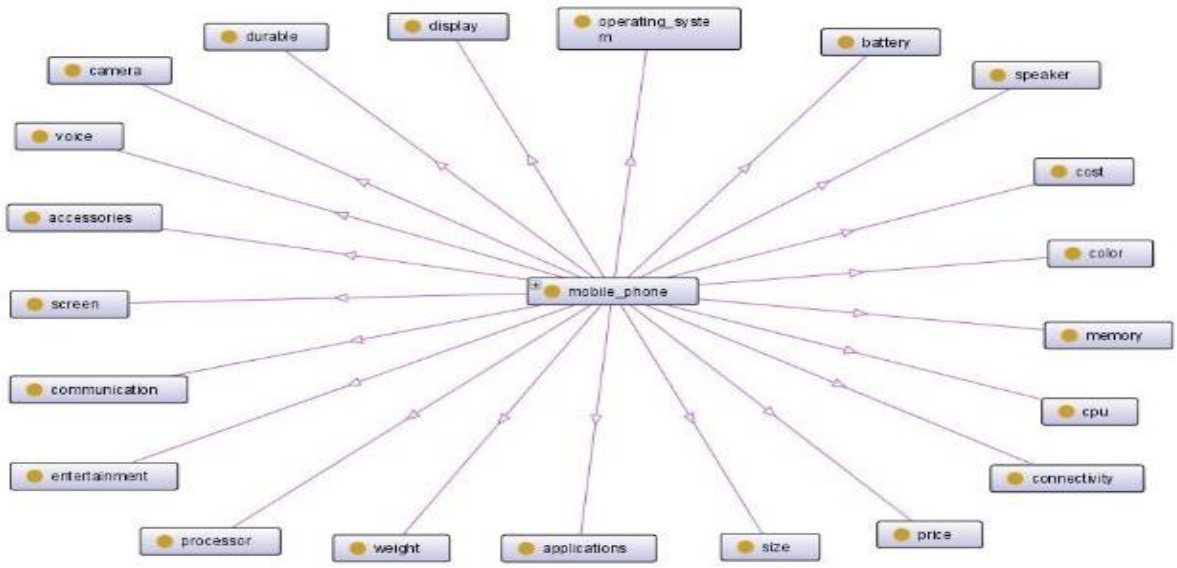

Fig. 2 Mobile phone domain ontology

3. Collecting Training Data

As we are using machine learning method using naive bayes classifier we require large amount of labelled training data to train and build the classifier. With the large range of topics discussed on Twitter, it would be very difficult to manually collect enough data to train a sentiment classifier for tweets. To automatically collect large labelled training data we have used method introduced by Read and also used by [2]. In this method emoticons are used as noisy labels, tweets containing emoticons such as ‘ :) ’ are considered as positive while tweets containing ‘ :(‘ emoticons are considered as negative. The following table gives complete list of emoticons considered as positive and negative.

TABLE I: LIST OF EMOTICONS

Emoticons mapped to :)	Emoticons mapped to :(
:) :-) :) :D =) :P	:(:-(: (

No efficient method exists for collecting neutral tweets so we mainly focused on positive and negative tweets. As we have used R platform/RStudio for our work in this paper , we extracted the tweets using twitteR package of R. Positive tweets are extracted using ‘ :) ‘ as searchterm and negative tweets extracted using : (as a searchterm. Nearly 4 lakhs tweets are collected using above methods, these tweets such forms our large training data set which is used to train the naive bayes classifier.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

Tweets can be directly extracted from twitter api in real time, but to achieve it we first needed to create and register for an application on twitter's official website, after creating account on twitter we can create new applications and use twitter api for it. After application is created it provides you with consumer key, consumer secret, access token, access secret, request url, access url. All these parameters are required for authentication of application. Once you are authenticated you can easily access twitter api to extract tweets. Tweets are searched using searchTerm function of twitteR package. For collecting tweets for testing purpose mobile phone brands are used as keyword for search, whereas for collecting training data emoticons are used as search term. Tweets can be collected in hundreds or thousands.

4. Training Naive Bayes Classifier

For using naive bayes classifier in analysis phase we require to train it first. To train naïve bayes classifier frequencies in data can be used as follow

$$P(cj) = \frac{doccount\ (C=cj)}{N}$$

$$P(wi, cj) = \frac{count\ (wi, cj)}{\sum count\ (w, cj)}$$

Where $P(cj)$ is probabilities of classes, N is total no. of documents used for training, $P(wi, cj)$ is probabilities of features or words. After training naive bayes we get a trained naive bayes model which contains unigram features and their corresponding probabilities.

5. Preprocessing

The language used in twitter is extremely informal, thus large numbers of misspelled words, slang words, emoticons, appear in tweets which renders sentiment analysis difficult. Such tweets contain many words which doesn't play any significant role in determining polarity of the tweets, these words are removed in preprocessing or data cleaning phase. Words such as retweets, @username, punctuations, digits, html links, multiple white spaces/tabs, dots, besides this all tweets are converted to lower case letters for simplicity. Preprocessing or data cleaning plays significant role as it reduces the size of tweets and helps in feature reduction as it reduces number of unigrams in a tweet.

Tweets used as testing data are also processed as explained earlier. In addition to this, testing data is also processed to replace multiple synonyms of feature with only one synonym. Many times users use multiple names or words for same features for example talk time, duration words are used to mention feature battery, games, music, radio are used to mention entertainment etc. We used same names for features that we have used in mobile phone ontology . This helps in the matching process. Following table shows list of brand features replaces with equivalent features.

TABLE II
FEATURE REPLACEMENT OF MOBILE BRANDS

Sr. no.	Multiple brand features	Equivalent single feature
1	internal memory, external memory, card slot, sd/sdhc card, micro sd card	memory
2	size, bright, color, touch screen, responsiveness	screen/display
3	android, symbian	operating system
4	build	durability
5	mega pixel, photo, video, dual camera	camera
6	blue tooth,wifi,gprs,2g,3g,4g,antenna,reception,edge	connectivity
7	battery life, talk time, duration	battery
8	bag, case	accessories
9	games, music, radio	entertainment
10	voice recognition, office, browser, gps, software	applications
11	speaker, volume	voice
12	social network,web,messages,sms,mms,email,im	communication

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

6. Sentiment Analysis of Tweets

*a) Analysis of Tweets:*For performing sentiment analysis of tweets we using supervised approach using naive bayes classifier with unigram features. Naive Bayes classifier is a simple but very power classification algorithm. It is widely used algorithm for document classification. The basic idea is to estimate the probabilities of categories given a test document by using the joint probabilities of words and categories. The naive part of such a model is the assumption of word independence. The simplicity of this assumption makes the computation of Naive Bayes classifier far more efficient. We use a trained naive bayes model. Class c* is assigned to tweet d, where

$$c* = argmac_c P_{NB}(c|d)$$

$$P_{NB}(c|d) := \frac{(P(c)\sum_{i=1}^{m} P(f|c)^{ni(d)}}{P(d)}$$

In this formula, f represents a feature and $n_i(d)$ represents the count of feature f_i found in tweet d. There are a total of m features. Parameters $p(c)$ and $p(f|c)$ are obtained through maximum likelihood estimates, and add-1 smoothing is utilized for unseen features. For sentiment analysis we perform tokenization of both training and testing data, we will create a corpus which set of text or words. This is done with the help of 'tm' package of R. 'tm' is a package which a framework for text mining applications within R. Then a document term matrix is created. Document term matrix is whose rows correspond to documents or tweets and columns corresponds to unigram features or words. The entry $v[i,j]$ in the matrix v determines the frequency of the jth term in ith document. After this a naive bayes model is built with help of a function which gives probability table which contains columns for terms, term or word frequency, add one smoothing, probability of terms and log probability respectively. For testing the tweets, probabilities of terms found is obtained from probability table and probabilities of terms which are not found is calculated as (1/total smoothed words). The total probability of tweet is sum of probabilities of all the words in the tweet. For classification purpose, probabilities for both positive and negative class is calculated If positive probability is greater it is assigned to positive class else to negative class.

b) Analysis of Features of Brands : Determining features in tweets gives more detailed view of users opinion or preferences. Many times user likes or dislikes a particular feature of a product, but mere labelling a document as positive or negative doesn't give this information hence domain ontology can be used to identify various features of a product. Ontologies created in protege can be accessed from R with the help of ontoCAT, rJava package. ontoCAT gives unified, format-independent access to ontology terms and the ontology hierarchy represented in OWL and OBO formats, provides basic methods for ontology traversal, such as searching for terms, listing a specific term's relations, showing paths to the term from the root element of the ontology, showing fattened-tree representations of the ontology hierarchy etc. all the mobile phones features or subclasses are extracted obtained from ontology and are matched with the terms in the tweets. Matching is done with the help of grepl function, this gives a matching matrix with entries of 0s and 1s indicating presence or absence of features. We use 'feature count' to calculate the sentiment values of features, there is high probability that the liked features will appear more in positive tweets while disliked features will appear more in negative tweets. Hence feature count gives a simple and efficient method for sentiment values. The sentiment values are calculated as below.

$$fv_I = tc_p - tc_n$$

if $fv_I > 0$ the feature has positive sentiment else negative sentiment. Where tc_p is total number of times feature appears in positive tweets and tc_n is total number of time feature appears in negative tweets. fv_I is opinion value of feature i.

IV. RESULTS

We collected testing tweets from twitter api of four mobile brands. The tweets are then preprocessed to remove unwanted words and features as explained earlier. The naive bayes classifier model is used to classify tweets into positive and negative classes, this classification gives idea about people's opinion about four mobile brands. The results are plotted in bar graphs using ggplot2 package of R. Figure 3 shows comparision of positive sentiments of four mobile

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

brands.

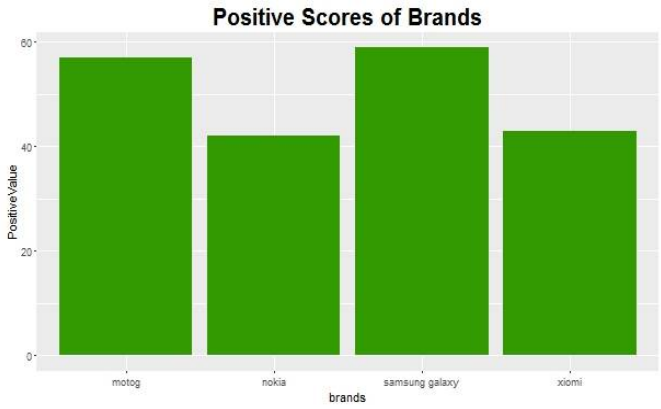

Fig 3. Positive sentiments of brands

After obtaining polarity of tweets feature analysis of the brands is done using ontology and polarity of tweets. Figure 4 and figure 5 shows the results for two such mobile brands Xiomi and Samsung. The feature having opinion value greater than zero is considered positive while opinion value less than zero is taken as negative. Using this we can relatively compare features of different mobile brands.

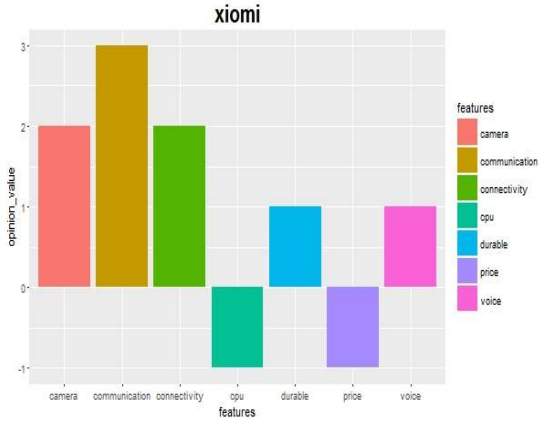

Fig. 4 Opinion value of Xiomi features

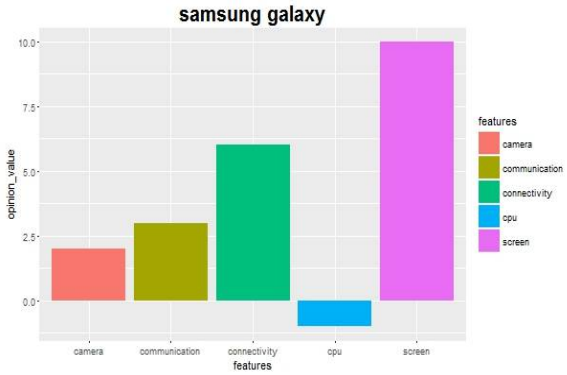

Fig. 5 Opinion values of Samsung Galaxy's features

V. SHINYDASHBOARD APPLICATION

shiny/shinydashboard is a web application framework for R. It makes it incredibly easy to build interactive web applications with R. Automatic ``reactive'' binding between inputs and outputs and extensive pre-built widgets make it possible to build beautiful, responsive, and powerful applications with minimal effort. We can host our Shiny/shinydashboard app somewhere that is publicly available, such as shinyapps.io or on your own Shiny Server. Shinydashboard application has two components user-interface script and server script. The user-interface (ui) script controls the layout and appearance of the app. It is defined in a source script named ui.R. The server.R script contains the instructions that computer needs to build the app. These applications let you specify input parameters using friendly controls like sliders, drop-downs, and text fields; and they can easily incorporate any number of outputs like plots, tables, and summaries. The Shiny and Shinydashboard package is free and open source, and is designed primarily to run applications locally. Shiny/shinydashboard allows customization of the application's user-interface to provide an elegant environment for displaying user-input controls and simulation output–where the latter simultaneously updates

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

with changing input. The flexible nature of the R language makes simulations of population variability possible thus promoting the combination of Shiny with R in model visualization.

An application is created based on the proposed approach. The application is created using shinydashboard package of R .This package provides a theme on top of 'Shiny', making it easy to create attractive dashboards. This application enables real time analysis of twitter using proposed approach. Shinydashboard application consist of user interface and server. The application consists of three parts dashboard header, dashboard body and dashboard sidebar. The application we created consists of four input text fields and one action button. The result of analysis is presented in two tabs. The tabs are further divided four boxes which hosts various plots. These tabs are present in dashboard body. The four text field allows user to enter the name four brands, when user presses action button these input fields are used as inputs for further analysis. The result of analysis are presented in the form of bar graphs in two tabs, user can select any tab for viewing results. This application enables real time sentiment analysis of twitter with which user can interact and perform analysis of any brands. Fig 6 shows the initial empty user interface of the shinydashboard application. Fig 7 shows tab1 one of application after entering mobile
phone brand names
.

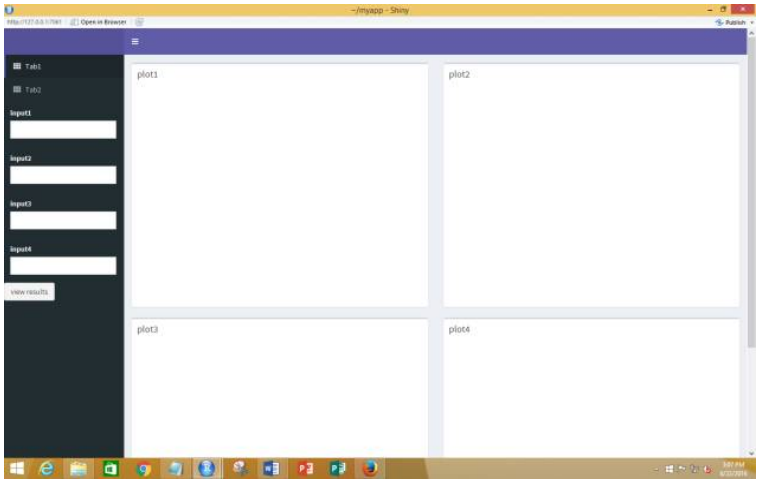

Fig. 6 Initial User-interface of shinydashboard application

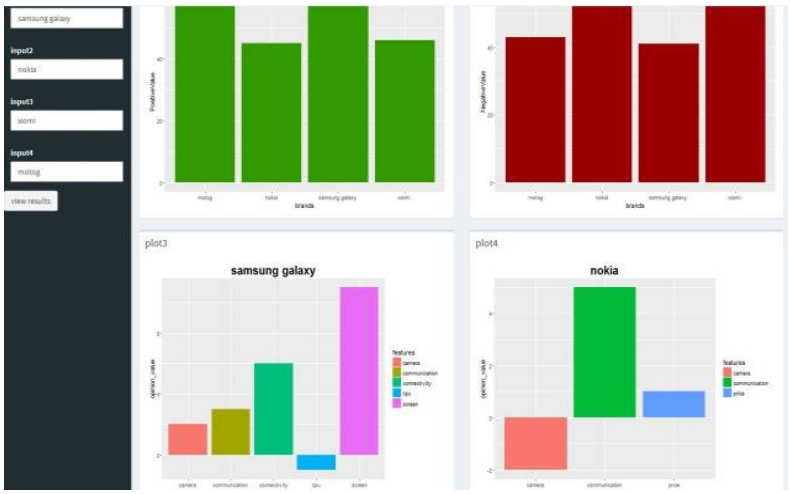

Fig. 7 Shinydashboard application tab1

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization *Volume 6, Special Issue 1, January 2017*

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

VI. CHALLENGES

Sentiment analysis in general do poses a lot of challenges. Sentiment analysis of micro-blogging site twitter does bring its own challenges in addition to traditional ones. Twitter in involves extensive use of jargons in authors tweets. The language used in twitter is extremely informal, thus large numbers of misspelled words, slang words, emoticons, appear in tweets which renders sentiment analysis difficult. Opinion words that appear positive in one context may appear negative in other contexts. People express their opinions in many different ways which is easy for humans to recognise but difficult for machines to parse.

We have used machine learning approach for classifying tweets, training supervised algorithm consumes lot of time. More the amount of training data more time it takes to train algorithm. Naive bayes classifier labels the tweet as positive or negative, but sometime the expressed opinion may not be related to mobile phone or its features, it decreases the accuracy of analysis. While creating shinydashboard real time application, the speed of application depend upon the time required to retrieve tweets from twitter api. If large amount tweets are retrieved the application tends to run slower.

VII. CONCLUSIONS

This paper thus shows an efficient implementation of machine learning and ontology methods for sentiment analysis purpose. This approach not only shows sentiments in tweets but also does in depth analysis to identify further information about product's feature using ontology. Shinydashboard application created shows that such method can be efficiently use to create a dynamic application which enables analysis of any brands in real time.

VIII. FUTURE WORK

We have implemented combination of naive bayes and ontology based approach,.Larger the number of tweets for training more accurate the results will be, hence in future results need to be tested with larger number of tweets. Sometimes opinion expressed in the tweet may not be related with the brand feature, this needed to be checked for more accurate results. Building fully automated ontology is still a distant dream. Automated ontology tools will help to build fully automatic analysis system or application, it will save time and effort hence research should be done to achieve it. More algorithms can be used in combination like neural networks and other machine learning algorithms like support vector machine , k-nearest neighbour etc . for more accuracy and verification of results.

REFERENCES

[1] Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun sd Hsu, Bing Liu. Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis. HPL Laboratories, HPL 2011.
[2] Go, Bhayani, Huang. Twitter Sentiment Classification using Distant Supervision. 2009.
[3] K. Vithiya Ruba and D. Venkatesan. Building a Custom Sentiment Analysis Tool based on an Ontology for Twitter Posts. In IJST 2015. SASTRA University Tamil Nadu.
[4] Efstratios Kontopoulos, Christos Berberidis , Theologos Dergiades , Nick Bassiliades. Ontology-based sentiment analysis of twitter posts. In Elsevier 2013.
[5] Speriosu, Sudan, Upadhay, Baldridge. Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph. University of Texas in Austin.
[6] Syed Zeeshan Haider. AN ONTOLOGY BASED SENTIMENT ANALYSIS a case study. University of Skovde.
[7] James Spencer and Gulden Uchyigit. Sentimentor: Sentiment Analysis of Twitter Data. School of Computing, Engineering and Mathematics University of Brighton, Brighton, BN2 4GJ.
[8] G.Vinodhini,RM.Chandrasekaran. Sentiment Analysis and Opinion Mining: A Survey. In IJARCSSE 2012. Annamalai University, Annamalai Nagar, India.
[9] Akshi Kumar and Teeja Mary Sebastian. Sentiment Analysis on Twitter. In IJCSI 2012. Department of Computer Engineering, Delhi Technological University Delhi, India.
[10] Bing Liu. Sentiment Analysis a Multi-faceted Problem. IEEE Intelligent systems 2010. Department of Computer Science University of Illinois at Chicago.
[11] Aamera Z. H. Khan, Dr. Mohammad Atique, Dr. V. M. Thakare. In IJARCSSE 2015. Amravati University, Amravati, India.
[12] Suhaas Prasad. Micro-blogging Sentiment Analysis Using Bayesian Classification Methods.
[13] Ravi Parikh and Matin Movassate. Sentiment Analysis of User Generated Twitter Updates using Various Classification Techniques. 2009.
[14] Christopher Johnson, Parul Shukla, Shilpa Shukla. On Classifying Political Sentiments of tweets. Department of Computer Science, University of Texas in Austin.
[15] Reynier Ortega, Adrian Fonseca, Yoan Gutierrez, Andres Montoyo. SSA-UO: Unsupervised Twitter Sentiment Analysis. University of Oriente, University of Matanzas, Cuba, University of.Alicante,Spain.