SENTIMENT ANALYSIS OF TWITTER DATA

V.Lakshmi, K.Harika , H.Bavishya, Ch.Sri Harsha

*Under the guidance of **M.Ramesh**, Asst prof.*
Department of Information Technology
VR Siddhartha engineering college, Andhra Pradesh, India

---***---

Abstract – *With the advancement of web technology and its growth, there is a huge volume of data present in the web for internet users and a lot of data is generated too. Internet has become a platform for online learning, exchanging ideas and sharing opinions.*
Social networking sites like Twitter, Facebook, Google+ are rapidly gaining popularity as they allow people to share and express their views about topics, have discussion with different communities, or post messages across the world. There has been lot of work in the field of sentiment analysis of twitter data. This survey focuses mainly on sentiment analysis of twitter data which is helpful to analyze the information in the tweets where opinions are highly structured, heterogeneous and are either positive or negative, or neutral in some cases.

Key Words: Twitter, Sentiment analysis(SA), Opinion mining, Machine learning, Naïve Bayes(NB).

1.INTRODUCTION

Now-a-days, the age of internet has changed the way people express their views, opinions. It is now mainly done through blog posts, online forums, product review websites, social media etc. Nowadays, millions of people are using social network sites like Facebook, Twitter, Google plus, etc to express their emotions, opinion and share views about their lives.

Through the online communities, we get an interactive media where consumers inform and influence others through forums. Social media is generating a large volume of sentiment rich data in the form of tweets, status updates, blog posts, comments, reviews, etc. Moreover, social media provides an opportunity for business by giving a platform to connect with their customers for advertising. People mostly depend upon user generated content over online to a great extent for decision making. For e.g. if someone wants to buy a product or wants to use any service, then they firstly look up its reviews online, discuss about it on social network but the data generated by users is too vast for a normal user to analyse. So there is a need to automate this, various sentiment analysis techniques are widely used. Sentiment analysis (SA) tells user whether the information about the product is satisfactory or not before they buy it. Marketers and firms use this analysis data to understand products or services in such a way that it can be offered as per the users requirements.

1.1 Twitter's simplicity

Twitter data is interesting because tweets happen at the "speed of thought" and are available for consumption in real time, and you can obtain data from anywhere in the world.

We chose because Twitter is predominantly suited for data mining because of the three key features.

- Twitter's API is well designhed and easy to access.
- Twitter data in a convenient format for analysis.
- Twitter's terms of use for the data are relatively liberal as compared to other API's.

1.2 Twitter's API

An Application Programming Interface(API) is a set of programming instructions and standards for accessing a web-based software application. Twitter bases its API of the Reprsentational State Transfer(REST) Architecture. REST architecture refers to a collection of network design principles that define resources and ways to address and access data.

2. LITERATURE SURVEY

In recent years a lot of work has been done in the field of "Sentiment Analysis on Twitter" by number of researchers. In its early stage it was intended for binary classification which assigns opinions or reviews to bipolar classes such as positive or negative only.

Pak and Paroubek(2010) proposed a model to classify the tweets as objective, positive and negative. They created a twitter corpus by collecting tweets using Twitter API and automatically annotating those tweets using emoticons. Using that corpus, they developed a sentiment classifier based on the multinomial Naïve Bayes method that uses features like Ngram and POS-tags. The training set they used was less efficient since it contains only tweets having emoticons.

Parikh and Movassate(2009) implemented two models, a Naïve Bayes bigram model and a Maximum Entrophy model to classify tweets. They found that the Naïve Bayes classifiers worked much better than the Maximum Entrophy model.

Go and I.Hyung(2009) proposed a solution for sentiment analysis for twitter data by using distant supervision, in which their training data consisted of tweets with emoticons which served as noisy labels.They build models using Naïve Bayes, MaxEnt and Support Vector Machines (SVM). Their feature space consisted of unigrams, bigrams and POS. They concluded that SVM outperformed other models and that unigram were more effective as features.

Barbosa et al.(2010) designed a two phase automatic sentiment analysis method for classifying tweets. They classified tweets as objective or subjective and then in second phase, the subjective tweets were classified as positive or negative. The feature space used included retweets, hashtags, link, punctuation and exclamation marks in conjunction with features like prior polarity of words and POS.

Bifet and Frank(2010) used Twitter streaming data provided by Firehouse API , which gave all messages from every user which are publicly available in real-time. They experimented multinomial naive Bayes, stochastic gradient descent, and the Hoeffding tree. They arrived at a conclusion that SGD-based model, when used with an appropriate learning rate was the better than the rest used.

Davidov et al.,(2010) proposed a approach to utilize Twitter user- defined hastags in tweets as a classification of sentiment type using punctuation, single words, n-grams and patterns as different feature types, which are then combined into a single feature vector for sentiment classification. They made use of K-Nearest Neighbor strategy to assign sentiment labels by constructing a feature vector for each example in the training and test set.

Po-Wei Liang et.al.(2014) used Twitter API to collect twitter data. Their training data falls in three different categories (camera, movie , mobile). The data is labeled as positive, negative and non-opinions. Tweets containing opinions were filtered. Unigram Naive Bayes model was implemented and the Naive Bayes simplifying independence assumption was employed. They also eliminated useless features by using the Mutual Information and Chi square feature extraction method. Finally , the orientation of an tweet is predicted. i.e. positive or negative.

3. PROPOSED WORK

3.1 DATA COLLECTION

Twitter equips appliance to amass data through its Application Programming Interface (API). The streaming mechanism grabs the input information Tweets and operates any anatomizing, percolating, or aggregation mandatorily anterior to accumulating the outcome to a data store. The HTTP handling mechanism queries the data store for outcome in reply to user inquiry as shown in Fig 10 and 11. HTTP makes use of requests of GET method & it can gives outputs transformed by the use of Java Structured Object Notation on other hand ATOM. Python scripts were written to interact with streaming API and related data can be collected. These scripts collect data according to keywords and returns up to 17 entities like date, location, gender, language, urls, text etc.

3.2 DATA TRANSFORMATION

Python script returns requested data in a JSON (Java Script Object Notation) file format. JSON is not much compatible file format to process directly on textual data. So the python script can be written in such a way that JSON data is stored into database and a flat file was then generated for the purpose of pre-processing.

3.3 PREPROCESSING

For further processing flat files that are generated can be used. Analyzing data that has not been carefully screened may give incorrect outputs. So affecting representation and quality of data is first and foremost before running analysis.

Feature extractors and different Machine Learning Classifiers are used here. The unigrams and feature extractors are unigrams, with weighted decisive and pesimisive keywords. A framework is designed that separates feature extractors and classifiers as a pair of components.

a. Query Term

Normalizing the key terms should be performed. If user wants to perform sentiment analysis about a product, is classified by the count of positive/negative query terms.

b. Emoticons

Emotions are used by the training process as unwanted labels and it plays significant role in categorization. Emoticons are pruned from the training data. If emoticons are excluded, appropriate result cannot be calculated in the case of Support Vector Machine classifiers and MaxEnt, and less impact on Naive Bayes. Classifier is restricted to analyze all other features in the tweet, if all the emoticons are pruned.

c. Feature Reduction

There are many unique properties for the twitter language model. To reduce the feature space the following properties are taken.

Usernames: Direct messages are given by using Twitter Usernames. Before the username (@SanthiCh20) @ symbol is used.

Usages of links: Links are often included by the users in their tweets. For all the URLs a compatibility class is used.

Stop words: More no. of filler and stop words are present in this. They are like “the”, “is”, and “a”. These are discarded because they do not express the sentiment for a tweet. The complete list of stop words can be found at.

Repeated letters: Tweets contain normal language. A nonempty result set may be generated in twitter, when user take a key word “lucky” with the no. of c’s (For e.g., lucccky, lucccccky, lucccccccccky). If a row contains any letter occurred more than two times then it is swapped with two letters. From the above word, it is modified to “lucky”.

3.4 Classification

Naive Bayes: It is a probabilistic classifier and can learn the pattern of examining a set of documents that has been categorized [9]. It compares the contents with the list of words to classify the documents to their right category or class. Let d be the tweet and c* be a class that is assigned to d, where

$$C^{*} = \arg mac_{c} P_{NB}(c \mid d)$$

$$P_{NB}(c \mid d) = \frac{(P(c)) \sum_{i-1}^{m} p(f \mid c)^{n_{i(d)}}}{P(d)}$$

From the above equation, “ f ” is a “ feature‟, count of feature (fi) is denoted with ni(d) and is present in d which represents a tweet. Here, m denotes no. of features. Parameters P(c) and P(f|c) are computed through maximum likelihood estimates, and smoothing is utilized for unseen features. To train and classify using Naïve Bayes Machine Learning technique ,we can use the Python NLTK library.

4. APPLICATIONS OF SENTIMENT ANALYSIS

Sentiment Analysis has many applications in various Fields.

1.Applications that use Reviews from Websites:

Today Internet has a large collection of reviews and feedbacks on almost everything. This includes product reviews, feedbacks on political issues, comments about services, etc. Thus there is a need for a sentiment analysis system that can extract sentiments about a particular product or services. It will help us to automate in provision of feedback or rating for the given product, item, etc. This would serve the needs of both the users and the vendors.

2. Applications as a Sub-component Technology:

A sentiment predictor system can be helpful in recommender systems as well. The recommender system will not recommend items that receive a lot of negative feedback or fewer ratings. In online communication, we come across abusive language and other negative elements. These can be detected simply by identifying a highly negative sentiment and correspondingly taking action against it.

3. Applications in Business Intelligence :

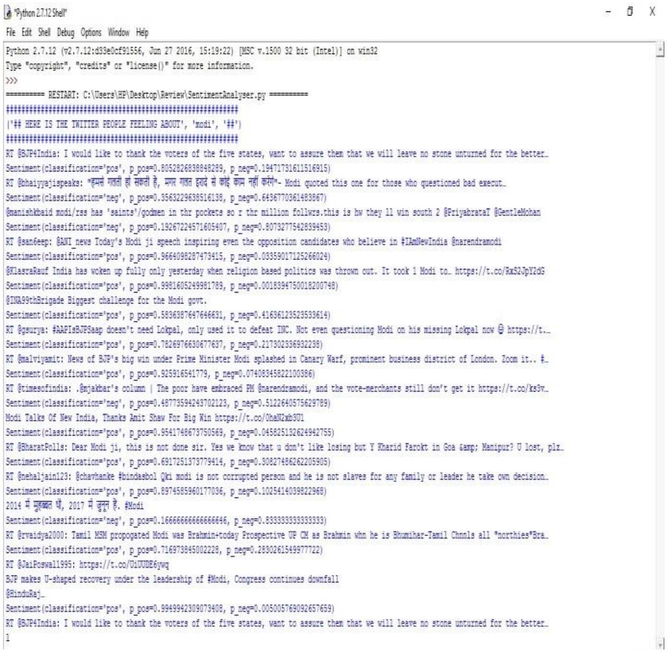
It has been observed that people nowadays tend to look upon reviews of products which are available online before they buy them. And for many businesses, the online opinion decides the success or failure of their product. Thus, Sentiment Analysis plays an important role in businesses. Businesses also wish to extract sentiment from the online reviews in order to improve their products and in turn their reputation and help in customer satisfaction .

4. Applications across Domains:

Recent researches in sociology and other fields like medical, sports have also been benefitted by Sentiment Analysis that show trends in human emotions especially on social media.

5. **Applications In Smart Homes:** Smart homes are supposed to be the technology of the future. In future entire homes would be networked and people would be able to control any part of the home using a tablet device. Recently there has been lot of research going on Internet of Things(IoT). Sentiment Analysis would also find its way in IoT. Like for example, based on the current sentiment or emotion of the user, the home could alter its ambiance to create a soothing and peaceful environment. Sentiment Analysis can also be used in trend prediction. By tracking public views, important data regarding sales trends and customer satisfaction can be extracted.

5. RESULT



```
Python 2.7.12 (v2.7.12:d33e0cf91556, Jun 27 2016, 15:19:22) [MSC v.1500 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
========== RESTART: C:\Users\HP\Desktop\Review\SentimentAnalyser.py ==========
############################################################
('## HERE IS THE TWITTER PEOPLE FEELING ABOUT', 'modi', '##')
############################################################
RT @BJP4India: I would like to thank the voters of the five states, want to assure them that we will leave no stone unturned for the better…
Sentiment(classification='pos', p_pos=0.8052826838848289, p_neg=0.19471731611516915)
RT @bhaiyyajispeaks: "हमसे गलती हो सकती है, मगर गलत इरादे से कोई काम नहीं करेंगे"- Modi quoted this one for those who questioned bad execut…
Sentiment(classification='neg', p_pos=0.3563229638516138, p_neg=0.6436770361483867)
@manishkbaid modi/rss has 'saints'/godmen in thr pockets so r thr million follwrs.this is hw they ll win south 2 @PriyabrataT @GentleMohan
Sentiment(classification='neg', p_pos=0.19267224571605407, p_neg=0.8073277542839453)
RT @san6eep: @ANI_news Today's Modi ji speech inspiring even the opposition candidates who believe in #IAmNewIndia @narendramodi
Sentiment(classification='pos', p_pos=0.9664098287473415, p_neg=0.03359017125266024)
@KlasraRauf India has woken up fully only yesterday when religion based politics was thrown out. It took 1 Modi to… https://t.co/RxS2JpY2dG
Sentiment(classification='pos', p_pos=0.9981605249981789, p_neg=0.0018394750018200748)
@INA99thBrigade Biggest challenge for the Modi govt.
Sentiment(classification='pos', p_pos=0.5836387647646631, p_neg=0.4163612352353614)
RT @gsurya: #AAPIsBJPSaap doesn't need Lokpal, only used it to defeat INC. Not even questioning Modi on his missing Lokpal now 😂 https://t…
Sentiment(classification='pos', p_pos=0.7826976630677637, p_neg=0.217302336932238)
RT @malviyamit: News of BJP's big win under Prime Minister Modi splashed in Canary Warf, prominent business district of London. Zoom it.. #…
Sentiment(classification='pos', p_pos=0.925916541779, p_neg=0.07408345822100386)
RT @timesofindia: .@mjakbar's column | The poor have embraced PM @narendramodi, and the vote-merchants still don't get it https://t.co/ks3v…
Sentiment(classification='neg', p_pos=0.48773594243702123, p_neg=0.5122640575629789)
Modi Talks Of New India, Thanks Amit Shaw For Big Win https://t.co/0haN2xb3U1
Sentiment(classification='pos', p_pos=0.9541748673750569, p_neg=0.045825132624942755)
RT @BharatPolls: Dear Modi ji, this is not done sir. Yes we know that u don't like losing but Y Kharid Farokt in Goa &amp; Manipur? U lost, plz…
Sentiment(classification='pos', p_pos=0.6917251373779414, p_neg=0.30827486262205905)
RT @nehaljain123: @chavhanke #bindasbol Qki modi is not corrupted person and he is not slaves for any family or leader he take own decision…
Sentiment(classification='pos', p_pos=0.8974585960177036, p_neg=0.1025414039822968)
2014 में मुहब्बत थी, 2017 में जुनून है. #Modi
Sentiment(classification='neg', p_pos=0.16666666666666646, p_neg=0.8333333333333333)
RT @rvaidya2000: Tamil MSM propogated Modi was Brahmin+today Prospective UP CM as Brahmin whn he is Bhumihar-Tamil Chnnls all "northies"Bra…
Sentiment(classification='pos', p_pos=0.716973845002228, p_neg=0.2830261549977722)
RT @JaiPoswal1995: https://t.co/UiUUDE6ywq
BJP makes U-shaped recovery under the leadership of #Modi, Congress continues downfall
@HinduRaj…
Sentiment(classification='pos', p_pos=0.9949942309073408, p_neg=0.005005769092657659)
RT @BJP4India: I would like to thank the voters of the five states, want to assure them that we will leave no stone unturned for the better…
1
```

6. CONCLUSION AND FUTURE WORK

There are various types of Machine Learning Algorithms like Naive Bayes, Max entropy that may not produce accurate results for either of unigrams, bigrams or weighted unigrams. Support Vector Machines (SVM) is termed as best data classification technique. Sentiment Analysis on twitter data is nothing different from those techniques on other genres. For further in the future these techniques can be explored to rich linguistic analysis like topic modeling and semantic analysis.

Semantics: The comprehensive sentiment of a tweet is classified by the algorithms. Semantic role labeler can be used which indicates which noun is associated with the verb and accordingly the classification occurs.

Internationalization: Here the focus is only on English tweets but Twitter has a large amount of international audience. This approach should be used to classify sentiment with a language specific positive/negative keyword list in other languages.

REFERENCES

[1] "Twitter passed 500M users in june 2012,140M of them in US; Jakarta 'Biggest Tweeting City'.

[2] Twitter search team(May 31, 2011)."The Engineering Behind Twitter's new search experience".Twitter Engineering Blog. Retrieved June 7, 2014.

[3] Twitter turns six Twitter.com, March 21, 2012. Retrieved December 18,2012.

[4] "Twitter.com Site Info".Alexa Internet. Retrieved 2014-04-01.

[5] D'Monte Leslie(April 29, 2009). "Swine Flu's tweet causes online flutter".Business standard.Retrieved February 4, 2011. Also known as the 'SMS of the internet', Twitter is a free social networking service.

[6] Twitter via SMS FAQ. Retrieved April 13,2012

[7] Williams, Evan(April 13,2012). "It's true…". Twitter. Retrieved April 26,2011.

[8] Sagolla, Dom(January 30,2009). "How Twitter was Born". 140 charecters-A style guide for the short form. Retrieved February 4,2011.

[9] Dorsey, Jack(March 21, 2006). "Just setting up my twitter". Retrieved February 4, 2011.

[10] Dunn, John E(2011-04-06). "Twitter delays hpmepage revamp after service glitch". Retrieved 2011-05-22.