World Scientific News

An International Scientific Journal

Twitter Opinion Mining Using Sentiment Analysis

Saumil Maheshwari[1], Shubham Shukla[2], Dazy Kumari[3]

[1,3]Atal Bihari Vajpayee Indian Institute of Information Technology, Gwalior, India

[2]KIET Group of Institutions, Ghaziabad, India

[1-3]E-mail address: saumilmaheshwari@yahoo.com , shukla.Shubham1989@gmail.com ,

dazikumari08@gmail.com

ABSTRACT

This is very important to extract verdict or opinion of others about any product, topic or about some person. The rich sources of this opinion rich data are blogs, online review sites and social networking sites. Among the social networking sites, twitter one of such largest source of microblog has gained popularity with more than 500 million tweets per day. Because of this Twitter has become a primary source for opinion mining. Twitter messages called tweets, are much focused because of the restricted characters size of 140 characters. Social network data is one of the most effective and accurate indicators of public sentiment. In this paper, twitter data is analyzed to determine the opinion of public. Twitter data about iPhone 6 is collected for analysis using the Twitter public API which allows developers to extract tweets from twitter programmatically. In this paper, Naïve Bayes classifier is used to calculate the sentiments of tweets and compared with baseline algorithms.

Keywords: Naive Bayes, machine learning, Opinion Mining

1. INTRODUCTION

Twitter is one of the rich wellsprings of social information. Twitter provides everybody the ability to create and share their thoughts and data right away with no hindrances. Tweet is a statement of a thought or minute which contains content, recordings and photographs. Tweet

is constrained with one hundred forty characters [1]. Tweets occurrence to be the speed of thought ‖are accessible for utilization continuously. More than five hundred million tweets being sent on twitter every day. When you take after individuals, their tweets are appeared on our timetable right away. Also, your tweets on your follower's course of events in the real time. Intriguing tweets are advised on your course of events when you take after fascinating individuals which incorporates companions, famous people, news sources and so on.

With the advancement in the technology like hand held smart devices, the use of social media is prominently incresing. This has opened up a path for companies for advertisement of product or services or also they can boost up the brand name. With these vast available data on the social platforms opens up the application of opinion mining.

A. Twitter Application Program Interface

Twitter gives a stage to an application or site to interface with overall occurrence on Twitter. Information generation from Twitter is prompting because of the very much archived API, rich engineer apparatuses, inalienable openness for open utilization of tweets and expansive interest to clients from each stroll of life. Twitter utilizes OAuth to send approved and secure demands through its Programming interface. Secure in the sense clients are not required to impart their passwords to any sort of outsider applications there by expanding account security [2]. Sentiment Analysis is an equivalent word for opinion mining. Grouping the extremity of the given record, content or highlight in light of the feeling in the archive, content or highlight as positive, negative or perhaps impartial [3]. Author in [4] has connected distinctive strategies for deciding the extremity of film surveys. If there should be an occurrence of the confinement of the length of the tweet, it can said as deciding the assessment of a sentence. The motivation behind deciding the assessment of a tweet is to know the state of mind of the client regarding a subject or a general relevant extremity of the tweet. User's judgment or assessment, passionate condition of the client (full of feeling state), or the expected enthusiastic correspondence which the client needs to convey i.e., user's audit on an item. Go, [5] gives a thought regarding how assessment examination is done.

Sentiment Analysis will be examination of sentiments about a man/protest and the way the assessments after some time. It is utilized industrially to calculate the consumer loyalty about the items they make particularly the new ones [6]. It additionally gives a thought regarding the gauges and open acknowledgment of contending organizations items. It at last prompts finding of company's notoriety administration and brand observation taking the investigation of the great and awful items sold by the organization. A man dependably wishes to take after the pattern, patterns can be anticipated by supposition mining.

The organization of the paper is divided into sections with Section 2 describing related work, section 3 describes background and motivation; Section 4 describes the concepts that were used for calculating the sentiment analysis; Section 5 describes the implementations and other set-up for the experiment; Section 6 with the results obtained and detailed discussion; section 7 with the conclusion of our experiment.

2. RELATED WORK

Sentiment Analysis [7] is a developing territory of research with noteworthy applications in both industry and the scholarly community. A large portion of proposed arrangements

revolved around supervised, and machine learning approaches. Twitter's special qualities give rise to new issues for current sentiment analysis method, which initially centered around expansive stubborn corpora, for example, item audits. This paper exhibits another substance level slant examination technique for Twitter.

The strategy initially embraces a vocabulary based way to with perform substance level slant investigation. This technique can give high precision, however low recall [8]. To enhance review, extra tweets that are probably going to be obstinate are distinguished consequently by exploiting the data in the consequence of the vocabulary based technique. A classifier is then prepared to relegate polarities to the substances [9].

Author [10] identified that little normal change in opinion related with mainstream occasions (commonly 1% and 6% for Tiger Woods' admissions) is predictable with occasions bearing blurbs chances to fulfill previous individual objectives more regularly than inspiring natural responses. With rise of social networking epoch, there has been a surge of client produced content.

Microblogging destinations have a large number of individuals sharing their considerations every day due to its trademark short and straightforward way of expression. In [11] authors proposed and explored a worldview to mine the opinion from a well known constant microblogging administration, Twitter, where clients present continuous responses on and suppositions about ‖everything‖. In [12], authors elucidate a crossover approach utilizing both corpus based and word reference based strategies to decide the semantic introduction of the supposition words in tweets. A contextual analysis is displayed to show the utilization and viability of the proposed framework.

The expansion of sites and informal organizations introduces another arrangement of difficulties and openings in the way data is looked and recovered. Despite the fact that realities still assume a vital part when data is looked for on a subject, conclusions have moved toward becoming progressively critical too [13].

Suppositions communicated in online journals and interpersonal organizations are assuming a critical part affecting everything from the items individuals purchase to the presidential applicant they bolster. Therefore, there is a requirement for another kind of internet searcher which won't just recover realities, however will likewise empower the recovery of sentiments. Such a web index can be utilized as a part of various different applications like item audits to conglomerating conclusions on a political competitor or issue. Endeavors can likewise utilize such a motor to decide how clients see their items and how they remain with deference to rivalry [14].

This paper introduces a calculation which not just examines the generally speaking assumption of an archive/audit, additionally distinguishes the semantic introduction of particular segments of the survey that prompt a specific estimation. The calculation is coordinated in a conclusion web crawler which presents results to a question alongside their general tone also, a rundown of notions of the most vital components.

3. BACKGROUND AND MOTIVATION

A large portion of the analysts are working on Sentiment analysis by distinguishing effective words from the announcements that are dependable on to one's opinion on a specific subject. In recent years scientists have done a decent measure of work in the field of sentiment

analysis by distinguishing different possibilities of adverb-adjectives and adjective-adverb-verbs. Likewise there has been work on sentiment analysis of social issues in view of verb as most critical term in distinguishing opinions behind surveys. In others researches, researchers have done a condition level assumption investigation by separating assessments frame autonomous provisos of proclamations.

The surveys of electronic contraptions, motion pictures, political issues, books prior were known to individuals just through the exhortation of other individuals and promoting abilities of that organization. Yet, in the present circumstance we have numerous sites which gives us data about the nature of items, unwavering quality of merchandise, criticism for motion pictures, books, political issues and the client is additionally offered flexibility to survey them.

These sites incorporate informal communication destinations, web based business sites and furthermore numerous others. Web-based social networking like twitter is one of the best hotspot for a man to express his perspectives. By ascertaining the supposition of a tweet we can know the way of the tweet as for a specific subject. There by gathering every one of the tweets identified with a specific subject and examining the notions of these tweets will give us the data about the effect of a point on the general population.

4. METHODOLOGY

In this research Methodology is divided into 3 steps:

- Predition using movie review
- Predition using Twitter
- Prediction using Twitter data and movie reviews A. Data Extraction

Twitter has provided the public API for accessing the data [15]. Authentication requests using valid login ID and password is required for accessing the dataset. Twitter provides authentication keys for extractions of the tweets. Authentication keys are obtained by following some steps as shown in Fig 1. Firstly an application on twitter by signing in to https://apps.twitter.com/app/new is created.

Secondly, we have to manage application. Followed by changing the application permissions to read and write. After all the steps, we have the consumer key, consumer secret key, access token, access token secret key that are required to fetch tweets from twitter.

Tweet Extracted from twitter having complete information. The tweets are collected into the csv database. Here python library Tweepy is used to download tweets. We have to enter the credentials access token, access token secret, consumer key and consumer key secret. To extract the tweets related to a particular keyword query parameter is passed to extract, which consist of the keywords we wants to extract tweets. We use these 3 main steps in our program:

- Authorize twitter API client.
- Make a GET request to Twitter API to fetch tweets for a particular query.
- Parse the tweets, categorize every tweet as positive tweet, negative tweet or neutral tweet.
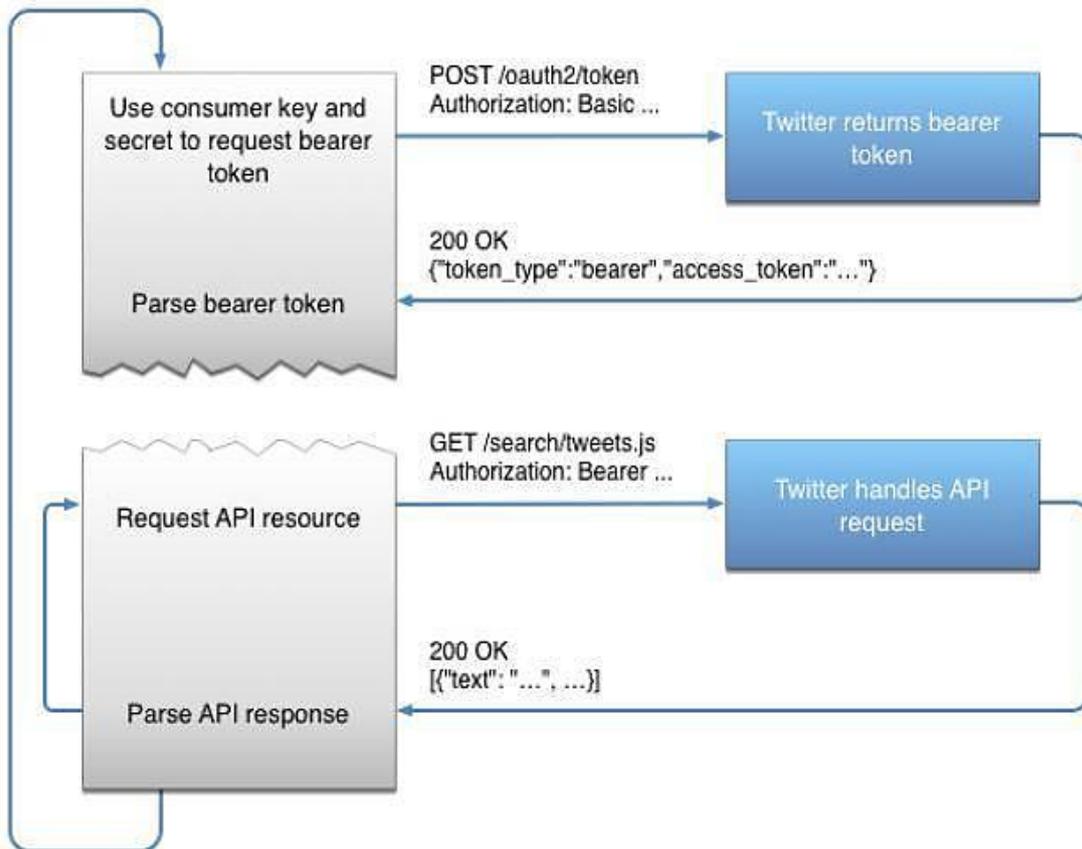
Fig. 1. Application Only Authentication

B. Calculating Sentiments

Baseline Method

In Baseline strategy, analyzing the thought of given tweets is done by summarizing the positive and once-over of negative words. We have assembled word records. We figure the amount of positive and negative words in tweet substance and we incorporate +1 for positive word and - 1 for negative word.

Over all summation is the appraisal of the tweet. In this we have used 2500+ positive words and emoticons of positive feeling and 4500+ negative words and emoticons of negative sentiments. Emoticons[18] gives us the wellspring of emoticons from which each one of the emoticons are being rejected and secured in a social event. [16] gives us the dictionary of web slang words which are dismisses similarly like emoticons.

Naive Bayes Classifier

As said in the acquaintance Bayesian grouping has a place with the Supervised Learning and Statistical technique for report characterization. This gives viable learning calculations and

earlier information on watched information can be joined. Bayesian Classification gives a valuable point of view for assessing and understanding many learning calculations.It ascertains the probabilities expressly to analyze speculation in information data.

- First extract the distinct word from Training set.
- Calculate the probability for each class.

$$P(C) = Nc/N \tag{1}$$

- Take testing data set and then calculate the conditional probability of each word with respective classes.

Choosing Class:

Calculate total probability of each word from testing data set.

P(pos/ti) = p(pos).Пp(ti/pos)

P(neg/ti) = p(neg).Пp(ti/neg) If P (pos/ti)>P (neg/ti)

Then text belong to positive class. Else text belong to negative class.

C. Variations in Testing and Training Sets of Naive Bayes Classifier

Method 1

In method 1, we have utilized 5332 positive motion picture audits and 5332 negative motion picture. In this arrangement of information we have utilized same set for both testing and preparing purposes. This motion picture audit information is for the most part utilized for opinion investigation tests. In this accumulations are accessible with separate marking in view of their general feeling polarity. Testing set: One-fourth of the set.

Training set: Three-fourth of the entire set (positive + negative)

Method 2

In method II, we have changed the data sets for training reason which suits the Twitter condition. We tested and trained on the social networking site like Twitter data.

Testing set: One-fourth of Tweets data set.

Training set: Three-fourth of Tweets.

Method 3

Method 3 upgraded informational index utilized above by including motion picture audits likewise (Twitter dataset + motion picture reviews)

Testing set: One-Tenth of the information

Training set: Nine-Tenth of the information.

5. EXPERIMENTAL SETUP AND DATASET

Experiment is done on iPhone 6. All the tweets and retweets are found in the streamed data. The data is stored in the MongoDB database Preprocessing is done using regular expression. After preprocessed the processed data is stored in another database. For Naive Bayes classifier movies review and tweets were used as train data set. Total 12704 tweets had been collected.

6. RESULTS AND DISCUSSION

A. Naive Bayes Classifier

7915 tweets have positive sentiments.

4790 tweets have negative sentiments.

The classifier classified tweets into 2 different categories:

–One gives 62.29% positive sentiment.

–And other gave 37.701% negative sentiment. Graph is shown in Fig.2:

B. Baseline Method

12704 is assembled tweets on 1st October.

7210 tweets were classified in positive sentiments class.

5494 tweets were classified in negative sentiments class. Tweets from twitter are classified in two categories:

–One is 56.7% with positive.

–Another is 43.2% with negative. Graph is shown in Fig.3:

Accuracy achieved by different method is shown in Fig. 4 and Table 1.

Table 1. Accuracy by varying dataset

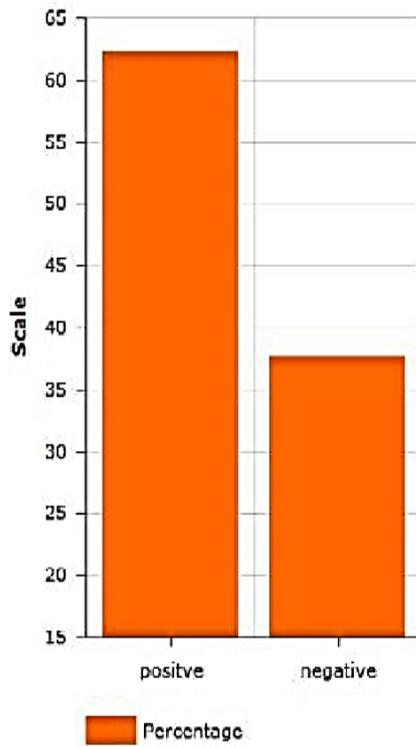| Algorithms	Accuracy (%)
Baeline	88.32
Method I	77.27
Method II	73.58
Method III	82.77

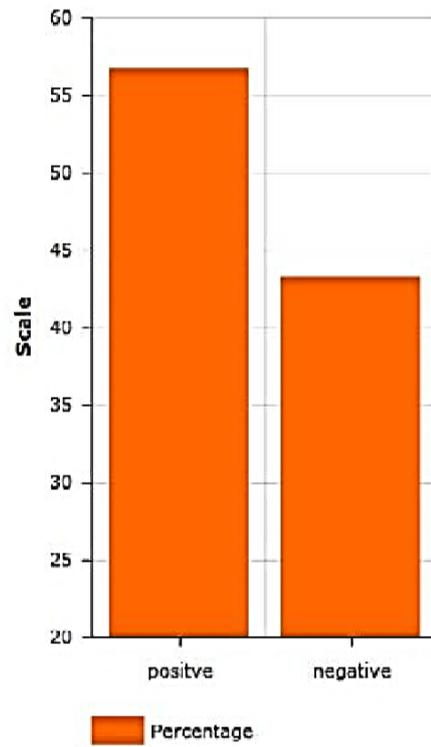**Fig. 2.** Naive Bayes classifier
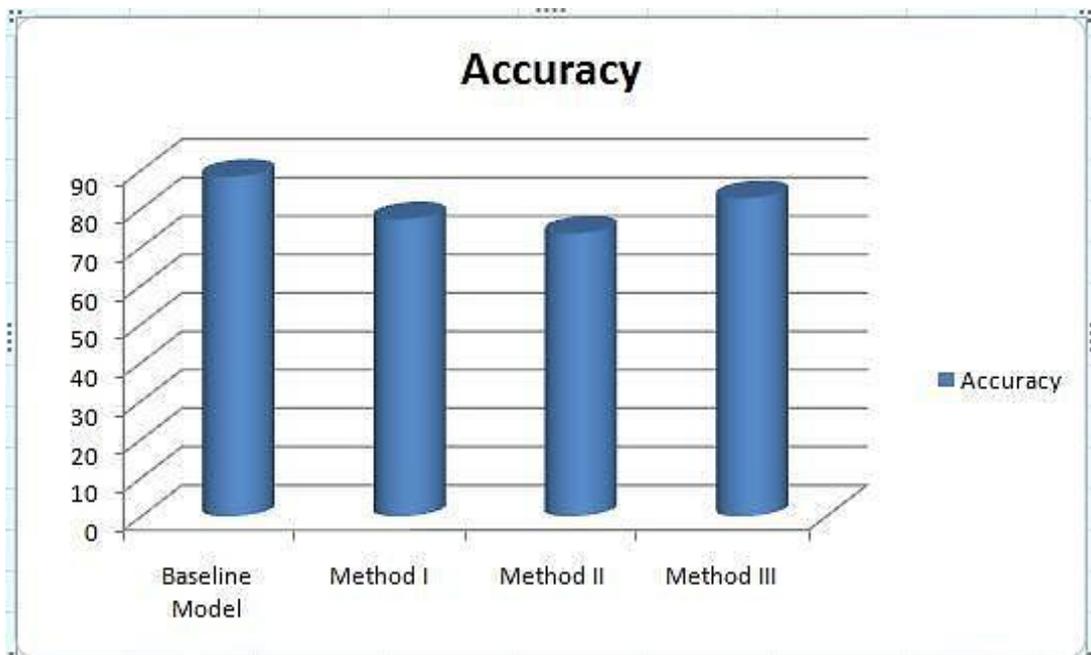

Fig. 3. Baseline Method

Fig. 4. Baseline Method

7. CONCLUSION

Naive Bayes Classifier gives 84.6% accuracy when it is tested and trained on movies review. But when the same is done on tweets 62.4% accuracy is achieved. So the combination of movies review and tweets from twitter is used as a data and 82.7% accuracy is achieved. Likewise, utilized one more procedure called Baseline approach. For this situation, it groups the tweets in negative and positive classes with 88.32% of accuracy.

References

[1] K. Ravi and V. Ravi, A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems,* vol. 89, pp. 14–46, 2015.

[2] S. Asur and B. A. Huberman, Predicting the future with social media, in Web Intelligence and Intelligent Agent Technology (WI-IAT), *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* vol. 1. IEEE, 2010, pp. 492–499.

[3] K. Denecke, Using SentiWordNet for multilingual sentiment analysis, in Data Engineering Workshop, 2008. ICDEW 2008. *IEEE 24th International Conference on Data Engineering Workshop*, IEEE, 2008, pp. 507–512.

[4] F. H. Khan, S. Bashir, and U. Qamar, Tom: Twitter opinion mining framework using hybrid classification scheme, *Decision Support Systems,* vol. 57, pp. 245–257, 2014.

[5] H. Chen and D. Zimbra, Ai and opinion mining, *IEEE Intelligent Systems,* vol. 25, no. 3, pp. 74–80, 2010.

[6] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger, Pulse: Mining customer opinions from free text, in International Symposium on Intelligent Data Analysis. Springer, 2005, pp. 121–132.

[7] Z. Khan, M. Atique, and V. Thakare, Combining lexicon-based and learning-based methods for twitter sentiment analysis. *International Journal of Electronics, Communication and Soft Computing Science & Engineering* 89, 2015.

[8] Balahur, R. Steinberger, E. Van Der Goot, B. Pouliquen, and M. Kabadjov, Opinion mining on newspaper quotations,‖ in Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on *Web Intelligence* and *Intelligent Agent Technology*, vol. 3. IEEE, 2009, pp. 523–526.

[9] M. Thelwall, K. Buckley, and G. Paltoglou, Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, vol. 62, no. 2, pp. 406–418, 2011.

[10] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, ―New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems,* vol. 28, no. 2, pp. 15–21, 2013.

[11] Kumar and T. M. Sebastian, ―Sentiment analysis on twitter. *International Journal of Computer Science Issues,* vol. 9, no. 3, pp. 372–378, 2012.

[12] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger, Pulse: Mining customer opinions from free text,‖ in international symposium on intelligent data analysis. Springer, 2005, pp. 121–132.

[13] M. Eirinaki, S. Pisal, and J. Singh, ―Feature-based opinion mining and ranking,‖ Journal of Computer and System Sciences, vol. 78, no. 4, pp. 1175–1184, 2012.

[14] M. M. Mostafa, More than words: Social networks text mining for consumer brand sentiments, *Expert Systems with Applications,* vol. 40, no. 10, pp. 4241–4251, 2013.

[15] E. Cambria, R. Speer, C. Havasi, and A. Hussain. Senticnet: A publicly available semantic resource for opinion mining.in *AAAI fall symposium: common sense knowledge,* vol. 10, 2010.