Document Level Sentiment Analysis for Product Review using Dictionary Based Approach

Paramita Ray
Computer Sc Dept., Dinabandhu Andrews Institute of Technology and Management, Kolkata, India

Abstract

In recent times, people share their opinions, ideas through social networking site, electronic media etc. Different organizations always want to find public opinions about their products and services. Individual consumers also want to know the opinions from existing users before purchasing product. Sentiment analysis is the computational treatment of user's opinions, sentiments and subjectivity of text. In this paper we propose a framework for sentiment analysis using R software which can analyze sentiment of users on Twitter data using Twitter API. Our methodology involves collection of data from twitter, its pre-processing and followed by a lexicon based approach to analyze user's sentiment.

Keywords

Twitter, Sentiment Analysis, Lexical Analysis.

I.INTRODUCTION

Sentiment analysis (also called opinion mining) means to analyze people's real opinions, sentiments, evaluations, appraisals, attitudes, and emotions regarding specific product, organization, services, movies, individuals, political events, topics, and their attributes[1]. Customer feedback about particular products is very important for commercial organization. They can improve their product quality [2], services on the basis of customer opinion about their product. Twitter is a kind of micro- blogging social networking site and billions of users use it to give their opinion [3] related to a particular topic. On the basis of opinion, sentiments can be estimated through analysis.

In this paper we collected large number of tweets from twitter server by using Twitter API using different keywords [6] within a particular time period and date. We categorized the tweets into positive, negative, or neutral opinion [8] to estimate the overall sentiment of customer or user about particular products or services and that can be utilized as an effective feedback to improve their product or service quality.

II. SENTIMENT ANALYSIS METHODOLOGY: BACKGROUND

Text categorization was started long time ago (Salton and McGill, 1983), however [1] categorization based on sentiment was introduced more recently in (Das and Chen, 2001; Morinaga et al., 2002; Pang et al., 2002; Tong, 2001; Turney, 2002;Wiebe,2000).

In 2012, Federico Neri Carlo et al. [2] had developed an idea of sentimental analysis using 1000 facebook posts about new casts, comparing the[4] sentiments for the Italian public broadcasting service - towards the emerging and more dynamic private companies.

In 2015, Xing fang et al. presented an idea of sentiment analysis using product review data which is collected from [7]. His main aim was to tackle the problem of sentiments polarity categorization of sentiments analysis [6].

A) Two approaches of Sentiment Analysis

1) Supervised approaches or machine learning method:

Machine learning is one of the most prominent techniques gaining interest of researchers [12] due to its adaptability and accuracy. This method comprises of three stages: a) Data collection b) Pre-processing c)Training data Classification and plotting results [9].

2) Unsupervised (or lexicon-based)

Lexical analysis calculates the sentiment from the semantic orientation of words [8] or phrases that occur in a text. In this approach a dictionary containing positive and negative words[4] that are matched with the words containing in tweet. However, these techniques totally depend on lexical resources [6] which are concerned with mapping words [7] to a categorical (positive, negative, neutral) or numerical sentiment score. In this method the unigrams which are found in the lexicon [9] are assigned a polarity score.

III. PROPOSED METHODOLOGY

Here we proceed through the following steps for sentiment analysis.

Step 1:

a) Creating a twitter application- For twitter sentimental analysis we have to create a twitter application. This application allows the connection of the twitter server to crawl the data using the Twitter API [5].

b) Installing R packages- We installed twitteR, ROAuth, plyr, Stringr, ggplot2, tm etc. packages in the R environment.

c) Handshaking- This step is for accessing the Twitter API. This step includes the script code to perform handshaking using the Consumer Key and Consumer Secret of the application.

Step 2: Collecting data from Twitter

We have collected tweets from the twitter server by using Twitter API using different keywords within a particular time period and date.

Step 3: Data pre-processing:

After collecting tweets, the structure of a tweets data is analyzed properly for pre-processing of tweets data to eliminate all unwanted information. Pre-processing step used to handle the following issues:

a) Removing punctuations and digits.
b) Cleaning text: Removal of non alphanumeric characters from the tweet.
c) Removing URLs.
d) Removing un-necessary white spaces and tabs etc.
e) Replacing emoticons: Emoticon is replaced by a word according to a lexicon of emoticons.
The meanings of the emoticons were taken from website [10].

Emoticon synset	Emoticons
Happiness	:-D, =D, xD, (^_^)
Sadness	:-(, =(
Crying	:'(, ='(, (;_;)
Boredom	-_-, -.-, (>_<)
Love	<3, (L)
Embarrassment	:-$, =$, >///<

Table 1: Typical examples of emoticon synsets [10].

f) Stop-words removal- Stop words are words which carry a connecting function in the sentence, such as prepositions, articles etc.

g) Replacing acronyms: Tweet containing acronym is replaced by proper word. The meanings of the acronyms were taken from website.[11].

Acronym	English Expression
gr8,gr8t	great
lol	Laughing out loud
rotf	Rolling on the floor
bff	Best friend forever

Table 2: Typical examples of acronym and their expression [11].

h) Finding target: symbol "@" used by twitter to refer to user.

i) Finding Hash tags:"#" used to mark topics and increase tweet visibility.

j) ***Negations Handling***

Negation word convert positive sentiment to negative or from negative to positive by using special words. if negation word was found with positive word ,we decrement positive score by one and vice versa.

e.g. **I like iphone ----→ Positive**
I don't like iphone -→ Negative

can not	shouldn't	doesn't	didn't
don't	hadn't	hasn't	wasn't
weren't	won't	haven't	wouldn't
neither	nor	without	lacks
lacking	Couldn't	hardly	Daren't

Table 3: Typical examples of negation words

k) Changing tweets text data to lower case and splitting sentences to words.

Step 4:
Classification

Here we use lexical method for classification. We work on dictionary-based approach. The dictionary-based approach depends on finding words from tweets, and then matches the word with the dictionary. If there is a positive match, the positive score is incremented or the word is tagged as positive. If it is negative word then the negative score is incremented or the word is tagged as negative. Otherwise tag neutral word.

Here we used WordNet dictionary for analysis of customer sentiment.

IV. PROPOSED ALGORITHM

- **Pre-processing of Input Tweets**
- **Prerequisites:**

1. File containing list of Positive Sentiment Words
2. File containing list of Negative Sentiment Words
3. File containing list of Negation Words
4. File containing list of acronyms Words
5. File containing list of emoticons with proper words

- **Algorithm Employed**:

Score = 0, Positive_Score=0, Negative_Score=0, Negation_Score=0

/* Negations Handling */

Match words with the dictionary containing negation words.
If Word= Negation word then

Match next words with the dictionary containing positive sentiment words.

If Word== Positive Word
then Negative_Score =
Negative Score +1
Else
Match next words with the dictionary containing negative sentiment words
If Word== Negative Word
then Positive _Score =
Positive_Score +1
Else
If Word== Positive Word then
Positive _Score = Positive_Score +1
If Word==Negative Word then
Negative_Score = Negative_Score +1

/* Replacing acronyms */
If Word==Unknown Word then

Match words with the dictionary containing acronym words.
If Word== acronym word, then

Replace acronym word with proper word.

Match words with the dictionary containing positive sentiment words.
If Word== Positive Word ,then

Positive _Score = Positive_Score +1 Else
Match words with the dictionary containing negative sentiment words.
If Word==Negative Word, then
Negative_Score = Negative_Score +1

/* Detecting emoticon */

If Word==Unknown Word,then
Match words with the dictionary containing emoticon.

If Word== emoticon,then
Replace emoticon word with proper word.

Match words with the dictionary containing positive sentiment words

If Word== Positive Word ,then
Positive _Score = Positive_Score +1 Else
Match words with the dictionary containing negative sentiment words

If Word==Negative Word , then
Negative_Score = Negative_Score +1

/* Score calculation */

Score= Positive_Score - Negative_Score

If Score>0 Then
print “Positive” Else
if Score < 0 then
print “Negative” Else
print “Neutral”

Score Evaluation

Score S_t= Positive Score – Negative Score

let and **T** be the set of tweets t∈T collected for the product. Also define each tweet **t** to carry a score S_t (0>= s_t >0), such that if S_t >0 then tweet shows positive feedback of user. if S_t <0, then tweet show negative feedback for the product.

Now the average score is $\sum_{t \in T} \frac{S_t}{|T|}$

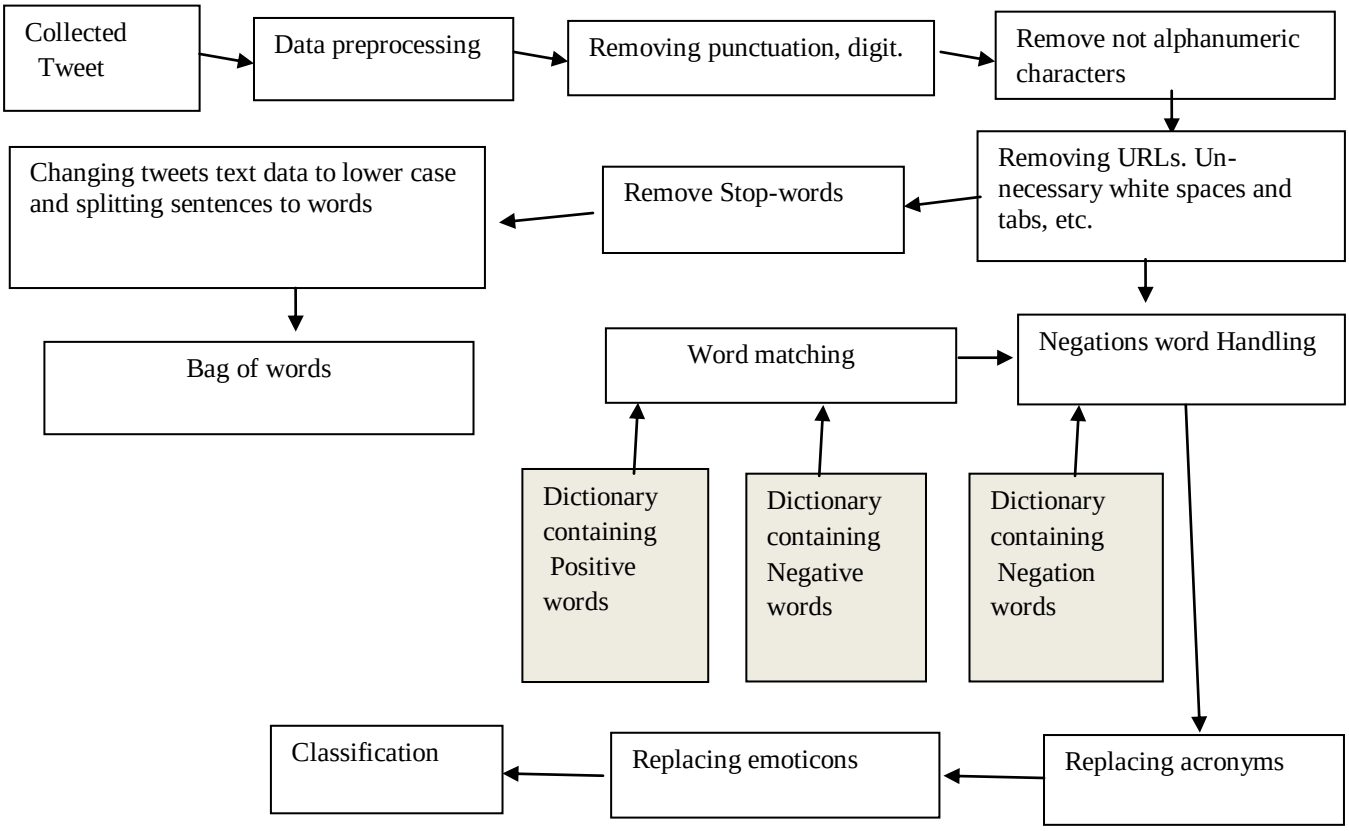

Fig. 1. Working flow of lexical approach

V. RESULT AND ANALYSIS

Sentiment analysis can take place mainly at three levels. a) Document level b) Sentence level c) Aspect level.
Here we tried to analyze the tweets at document level . In document level, the task is to classify whether a whole opinion document expresses a positive or negative sentiment and analyzes customer's feedback on the basis of that.

We have collected 3000 tweets using twitter API from twitter with keywrd "Iphone" from date 12.06.16 to 15.08.16 with different aspects(like Battery life,Services,Price,Size etc) and performed sentiment analysis of twitter users.

On the basis of the analysis we get the following results:

Aspect: GENERAL Positive: 2500 Negetive:300 Neutral :100
Aspect: Battery life Positive: 190 Negative:90 Neutral :10
Aspect: Service Positive: 250 Negative: 20 Neutral :5
Aspect: Size Positive: 250 Negative: 25 Neutral :6
Aspect: Price Positive: 30 Negative:1 40 Neutral :20

Table 4: Iphone:(opinion summary)

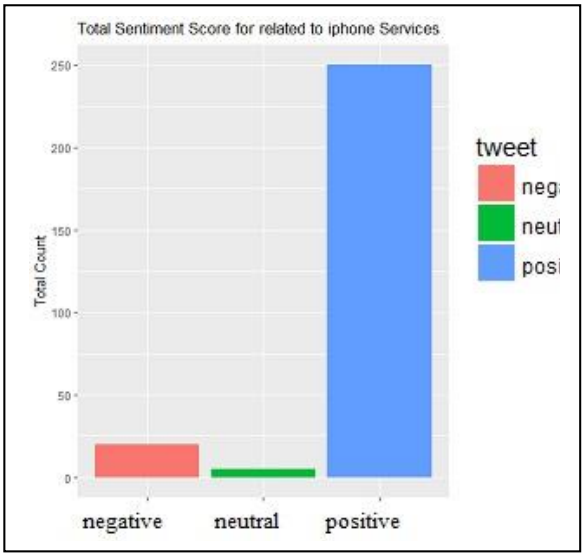

Fig4.Total Sentiment score related to iphone services

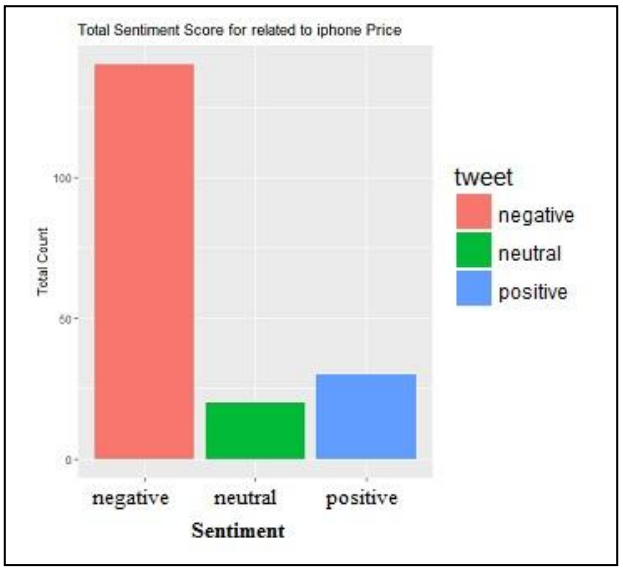

Fig5.Total Sentiment score related to iphone price

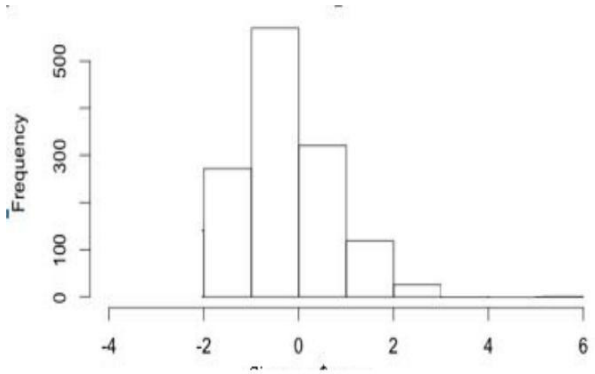

Fig 2: Histogram with entity Size

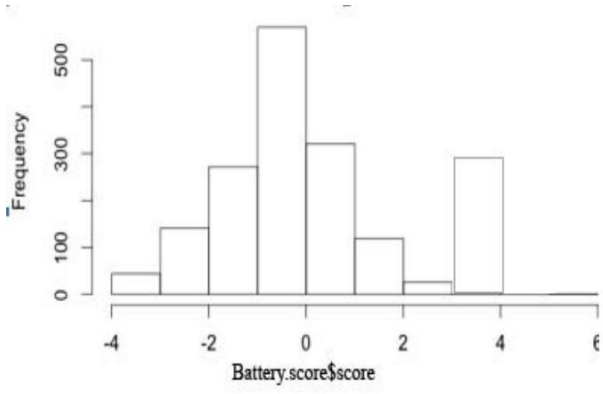

Fig 3: Histogram with entity Battery.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have tried to implement dictionary based methodology of sentiment analysis and developed an algorithm that is employed to large amount of data to estimate the sentiment of public..We replaces acronym word by making acronym dictionary and also detected emoticons in the tweet. We have done both document level and aspect level analysis based on our own approach, which has helped in the decision making. In future, we will work with comparative opinion and try to work with machine learning approaches to develop ₃ hybrid working model for sentiment analysis. Our ethod is applicable in other domain also. Eg.Customer review related to airlines services,Political review etc.

REFERENCES

[1] Olga Kolchyna1, Th´arsis T. P. Souza1, Philip C. Treleaven12 and Tomaso AsteTwitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination , Department of Computer Science, UCL, Gower Street, London,

[2] Bing Liu, Sentiment Analysis and Opinion Mining April 22, 2012

[3] Suvendra Kumar Jayasingh, Jibendu Kumar Mantri, P. Gahan Comparison between J48 Decision Tree, SVM and MLP in Weather Forecasting 10.14445/23488387/IJCSE-V3I11P109

[4] Harsh Thakkar and Dhiren Patel, Approaches for Sentiment Analysis on Twitter: A State-of-Art study , Department of Computer Engineering, NIT, Surat-395007, India.

[5] Apoorv Agarwal Boyi Xie Ilia Vovsha Owen Rambow Rebecca Passonneau,, Sentiment Analysis of Twitter Data , Department of Computer Science ,Columbia University.New York, NY 10027 USA

[6] Cataldo Musto, Giovanni Semeraro, Marco Polignano, A comparison of Lexicon-based approachesfor Sentiment Analysis of microblog Department of Computer Science, University of Bari Aldo Moro, Italy
[7] James Spencer and Gulden Uchyigit,Sentimentor: Sentiment Analysis of Twitter Data.School of Computing, Engineering and Mathematics.University of Brighton
[8] Anna Jurek,Maurice D. Mulvenna and Yaxin Bi,Improved lexicon-based sentiment analysis for social media analyticsScience direct,Published: 9 December 2015
[9] Dr.E.Kesavulu Reddy .14445/23488387/IJCSE-
[10] V3I11P107
[11] http://en.wikipedia.org/wiki/List_of_emoticons.
[12] http://www.acronymfinder.com/
[13] Akshi Kumar and Teeja Mary Sebastian, Sentiment Analysis on Twitter, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012,Dept of Computer Engineering, Delhi Technological University Delhi, India