International Journal of Research in Information Technology (IJRIT)

www.ijrit.com **ISSN 2001-5569**

Phrase Level Sentiment Analysis for Twitter

Shrikant Wath[1], Somy Kamble[2], Vinod Choudhari[3] and Dinesh Gawande[4]

[1]Student, Department of Computer Science and Engineering, Nagpur University
Nagpur, Maharashtra, India
shrikant.wath08@gmail.com

[2]Project coordinator, Business Development, Alkrun Technologies
Nagpur, Maharashtra, India
somykamble1001@gmail.com

[3]Software Engineer, Maximess
Nagpur, Maharashtra, India
vinodchoudhari90@gmail.com

[4]Assistant Professor, Computer Science and Engineering, Nagpur University
Nagpur, Maharashtra, India
gawande.dinesh@gmail.com

Abstract

In today's era where the competition is on the rise, the success of a business largely depends upon the way it is able to tackle the competition and rise above it. Most of the companies are highly investing in social-media platforms to advertise their products and at the same time to get feedback from the customers about their products. This system will enable these companies to know the reviews of the customers by filtering their reviews in three categories viz., positive, negative and neutral.

Keywords: *Sentiment, twitter, data-set.*

1. Introduction

There are times when an organization wants to know the views of public or its customers. This system classifies tweets in tweeter into three categories viz., positive, negative and neutral. The user will have to just enter the keyword in the application and then this application will download all the tweets relevant to the keyword added by the user and after processing it, it will display the final result in the form of a graph. For example, if your company has 30% negative sentiment, is that bad? It depends. If your competitors have a roughly 50% positive and 10% negative sentiment, while yours is 30% negative, that merits more discovery to understand the drivers of these opinions. Knowing the sentiments associated with competitors helps companies evaluate their own performance and search for ways to improve. This application will not only classify those tweets on the basis of positive or negative words but also by learning the meaning of the phrase. Natural language processing techniques will be used to extract the meaning of the phrase and to classify those tweets into positive, negative or neutral. The main objective of this project will be to increase the accuracy of the application so that it can learn or understand the meaning of the phrase and can classify it accordingly and to generate the final result in a dynamic graph.

2. Literature Review

Sentiment analysis on social networks introduces additional challenges related with noisy content, shorter messages and user-specific issues [1]. Performance on the task is typically worse than in generic domains [2], hence novel approaches are being sought, addressing social network communication specifics by analyzing emoticons or utilizing specific vocabularies [3] [4].

Our method for sentiment analysis of social network posts in Slovak language is based on machine learning with document level classification. We consider one document (a social network post) is assigned one sentiment value, so the mixed sentiment is not detected here (no aspect based analysis is employed). With the appropriate pre-processing and feature selection we aim to achieve accuracy as high as existing solutions in world languages [5]

In this paper a system which tries to answer a specific sentiment analysis problem described in the SemEval-2015 series will be taken as the base challenge to be improved. This falls under SemEval-2015 task 10: Sentiment Analysis in Twitter (Rosenthal et al., 2015). Under the main task, five subtasks (A-E) are provided. Any team taking part can compete for the entire five subtasks or any preferred number of tasks [6]

Multi-class sentiment analysis has proven to be a very challenging task. This is mainly for the simple reason that a tweet usually does not contain a single sentiment, but many ones. In this paper, we propose a pattern-based approach for sentiment quantification in Twitter. By quantification, we refer to the detection of the existing sentiments within a tweet and the detection of the weight of these sentiments. In a first step, we classify tweets into positive, negative, or neutral.

Our approach reaches an accuracy of 81%. We then perform the sentiment quantification on the sentimental tweets (i.e., positive and negative ones) to extract the sentiments within them: we define 5 positive sentiment sub-classes 5 negative ones and detect which exist in each tweet. We define 2 metrics to measure the correctness of sentiment detection, and prove that sentiment quantification can be a more meaningful task than the regular multi-class classification. [7]

3.1 Methodology

In this paper we propose Naïve-Bayes algorithm for analysis the sentiment of the tweets. A training set will be used to identify the tweets or phrase in three categories viz., ‘Positive’ , ‘Negative’ or ‘Neutral’. A training set is a dataset in which a phrase is classified into two groups i.e., ‘positive’ and ‘negative’. A positive phrase is denoted with ‘1’ whereas a negative phrase is denoted with ‘0’. By using Naïve Bayes Algorithm the probability of the phrase/tweet is counted and finally it is classified as ‘Positive’, ‘Negative’ or ‘Neutral’.

In fig. (a) Flow-chart of the diagram is explained. The first step is to train the classifier using the train data-set. In second step the phrase is tested and finally, in the last step, we get sentiment as a result. In fig (b) flow-chart of ‘Train Classifier’ is explained. A data-set is used in train classifier which is classified into ‘negative’ and ‘positive’ on the basis of its labels. In fig (c) flow-chart of ‘Test Classifier and Get Sentiment’ is shown. Here, all the stop words of the test data is removed and then each word is reviewed. If the word is present in train data-set then its probability is taken. Otherwise, its probability is measured using ‘Naïve Bayes algorithm’. [8]

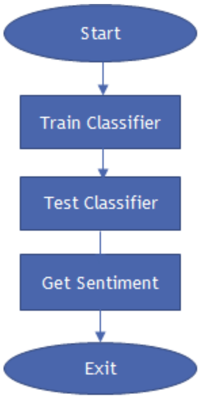


Fig.(a) Flow-chart

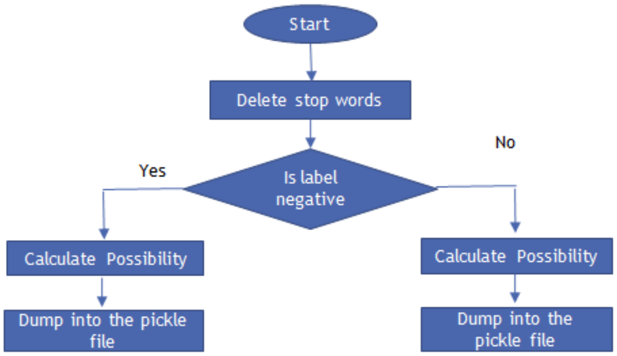


Fig.(b) Flow-chart of Train-classifier

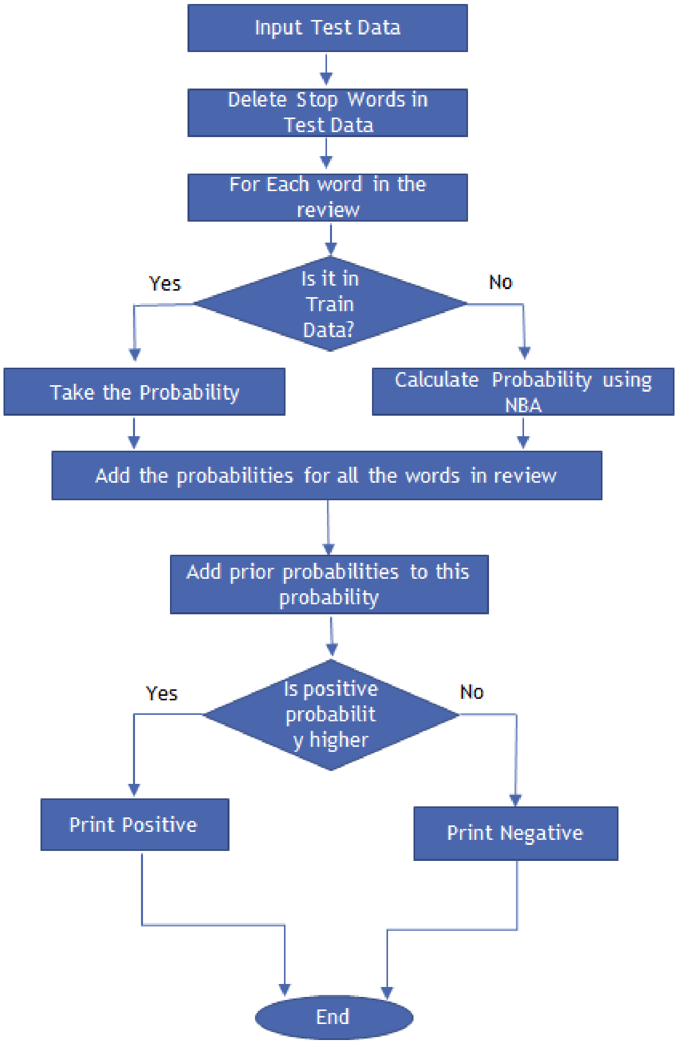

Input Test Data
Delete Stop Words in Test Data
For Each word in the review
Is it in Train Data?
Yes
No
Take the Probability
Calculate Probability using NBA
Add the probabilities for all the words in review
Add prior probabilities to this probability
Is positive probabilit y higher
Yes
No
Print Positive
Print Negative
End

Fig.(c) Flow-chart of Test classifier and Get Sentiment

3.2 Calculations

In order to implement this project I have used Naïve Bayes algorithm. Let us try to understand the Naïve Bayes algorithm with an example. As you can see below in the train set which would be used for identifying the problem phrases into positive, negative or neutral.

Step 1: Determine the train set and test set in your data set -

Document	Text	Class
1	I loved the movie but the songs were terrible ,still a nice movie though	+
2	I hated it	-
3	Great movie, good acting	+
4	A Poor movie	-
5	A good movie, great acting	+

So, there is a total of 12 unique words.

<loved,movie,great,hated,good,acting,poor,songs,terrible,still,nice,but,though>

Step 2: Convert the data set into frequency table –

Here, are the unique words and the frequency of each words in a document –

Doc	loved	movie	great	hated	good	acting	poor	songs	terrible	still	nice	but	though	class
1	1	2						1	1	1	1	1	1	+
2		1		1										-
3		1	1		1	1								+
4						1	1							-
5		1	1		1	1								+

Step 3: Compute the prior –

Prior Probability of the positive docs.-
P (+) = 3/5 = 0.6
P (-) = 2/5 = 0.4

Suppose we have to analyze the sentiment of the problem phrase given below –

“This movie is a waste of time.”

While testing for unknown words we use nk =0 and find its probability being both +ve and –ve

Step 4: Compute the prior conditional likelihood probability -

Computing conditional likelihood probability of each word of the problem phrase –
P (wk |+) = (nk + 1) / (n + |vocabulary|)
Where,

nk= number of times word ‘k’ occurs in the cases of positive and negative
n= number of words in positive and negative cases
vocabulary = total unique words

For Positive Case -

P (This|+) = (0+1)/ (17+12) =(Stop Word)
P (movie|+) = 4/29
P (is|+) = (Stop Word)
P (a|+) = (Stop Word)
P (waste|+) = 1/29
P (of|+) = (Stop Word)
P (time|+) = 1/29

For Negative Case –

P (This|-) = (0+1)/ (7+4) = (Stop Word)
P (movie|-) = 1/11
P (is|-) = (Stop Word)
P (a|-) = (Stop Word)
P (waste|-) = 1/11
P (of|-) = (Stop Word)
P (time|-) = 1/11

Step 5: Compute the Posterior Probability -

Posterior Probability –
P (+) = 4/29* 1/29 * 1/29 *0.6 = 9.840501866*10^{-5}
P (-) = 1/11*1/11*1/11*0.4 = 3.005259204*10^{-4}

3.005259204*10^{-4} > 9.840501866*10^{-5}

Step 6: Determine the class of the test set -

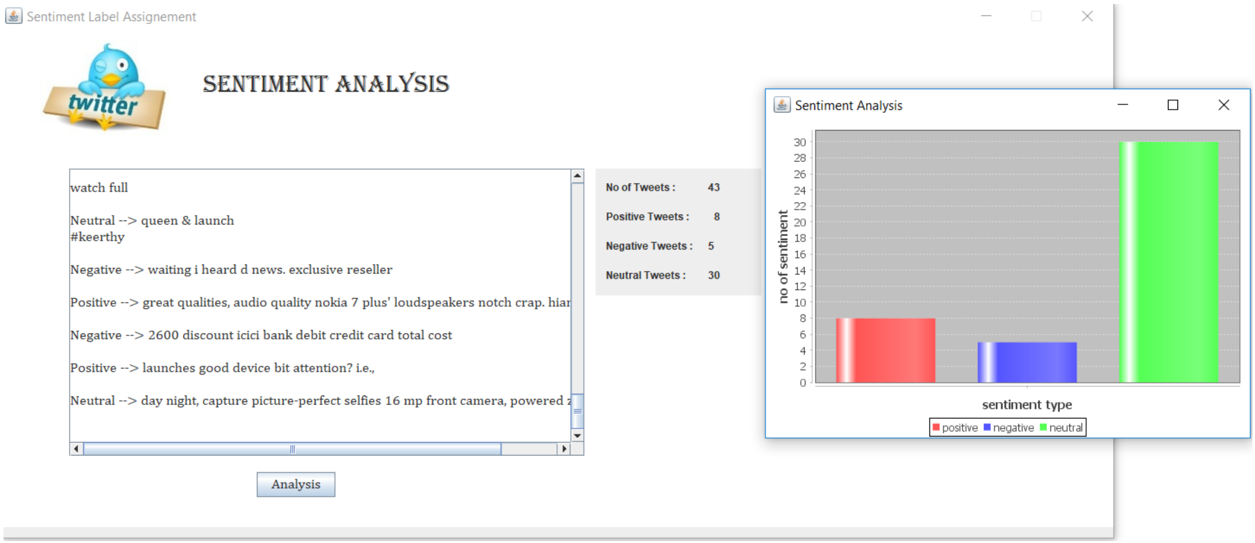
As we can see value of P (-) is greater than P (+), hence, the phrase is **Negative.**

4. Result and Analysis

The result we are getting in this application is in the form of bar-graph. For the sake of analysis I chose 3 different topics and did sentiment analysis on them. Following are the results –

#Nokia7plus

As this product was trending number 1 in India (for May 4, 2018) I did sentiment analysis on this key-word. On manually analyzing total number of tweets which were 43; I found out of 43 tweets 29 (6/8-positive, 3/5 – negative, 20/30 – neutral) tweets were correctly identified giving us an **accuracy of 67.44%.**

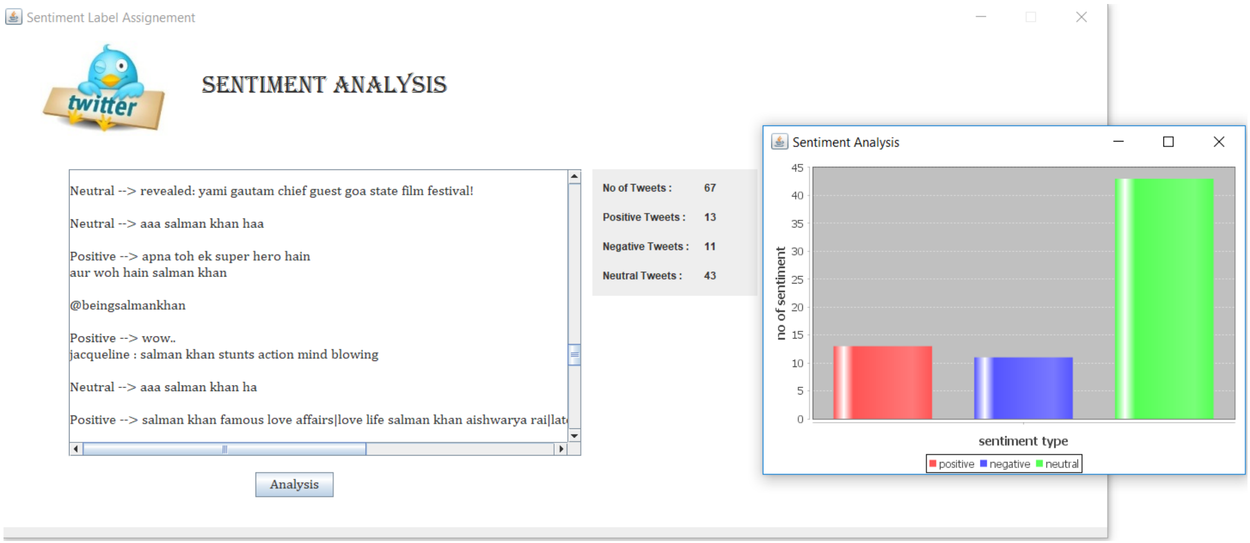


```
watch full

Neutral --> queen & launch
#keerthy

Negative --> waiting i heard d news. exclusive reseller

Positive --> great qualities, audio quality nokia 7 plus' loudspeakers notch crap. hiar

Negative --> 2600 discount icici bank debit credit card total cost

Positive --> launches good device bit attention? i.e.,

Neutral --> day night, capture picture-perfect selfies 16 mp front camera, powered z
```

No of Tweets :	43
Positive Tweets :	8
Negative Tweets :	5
Neutral Tweets :	30

#Salmankhan

Next key-word I analyze for was Salman khan, a famous Bollywood actor. Application fetched 67 total number of tweets. Out of which it identified 46 tweets (7/13 –positive, 6/11 – negative, 33/43 - neutral) correctly giving us an **accuracy of 68.65%.**



Neutral --> revealed: yami gautam chief guest goa state film festival!

Neutral --> aaa salman khan haa

Positive --> apna toh ek super hero hain
aur woh hain salman khan

@beingsalmankhan

Positive --> wow..
jacqueline : salman khan stunts action mind blowing

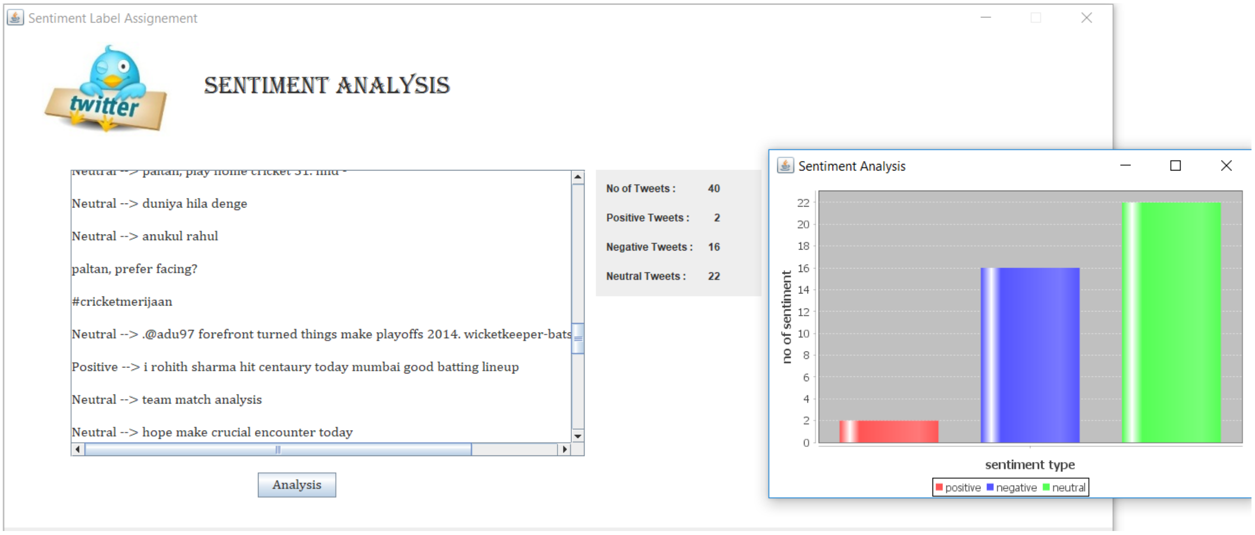
Neutral --> aaa salman khan ha

Positive --> salman khan famous love affairs|love life salman khan aishwarya rai|lat

No of Tweets :	67
Positive Tweets :	13
Negative Tweets :	11
Neutral Tweets :	43

#MumbaiIndians

Being the most successful league in the cricket world, IPL, as its matches were already going on, I tried to do sentiment analysis on one of its team 'Mumbai Indians'. I found out of total no. of tweets which were 40 out of which 26 tweets (1/2 – positive, 11/16 – negative, 14/22 - neutral) were identified correctly giving us an **accuracy of 65%.**



Neutral --> paltan, play home cricket 31. md -

Neutral --> duniya hila denge

Neutral --> anukul rahul

paltan, prefer facing?

#cricketmerijaan

Neutral --> .@adu97 forefront turned things make playoffs 2014. wicketkeeper-bats

Positive --> i rohith sharma hit centaury today mumbai good batting lineup

Neutral --> team match analysis

Neutral --> hope make crucial encounter today

No of Tweets :	40
Positive Tweets :	2
Negative Tweets :	16
Neutral Tweets :	22

5. Conclusions

In-order to determine the nature of the tweet Naïve Bayes algorithm is used. Using datasets it is possible to classify the nature of tweets. This project will be done in English; however, the proposed technique can be used with any other language, provided that language lexicon dictionary.

References

[1] Pozzi, F. A., Fersini, E., Messina, E., & Liu, B.: Challenges of Sentiment Analysis in Social Networks: An Overview. In Sentiment Analysis in Social Networks. Morgan Kaufmann, 2016

[2] Musto, C., Semerano, G., Polignano, M.: A comparison of Lexicon-based approaches for Sentiment Analysis of microblog posts. In Proc. of the 8th Int. Workshop on Inf. Filtering and Retrieval, 2014, pp. 59-68

[3] Korenek, P., & Šimko, M.: Sentiment analysis on microblog utilizing appraisal theory. World Wide Web Journal, 17(4), 2014, pp. 847-867.

[4] Hogenboom, A., Bal, D., Frasincar, F., Bal, M., de Jong, F. Kaymak, U.: Exploiting Emoticons in Sentiment Analysis. Proc. of the 28th Annual ACM Symposium on Applied Computing, 2013, pp. 703–710

[5] Rastislav Krchnavy, Marian Simko. "Semantic and Social Media Adaptation and Personalization (SMAP)", 2017 12th International Workshop

[6] Supun R.Muthutantrige, A.R.Weerasinghe "Sentiment Analysis in Twitter Messages Using Constrained and Unconstrained Data Categories"

[7] Mondher Bouazizi, Tomoaki Ohtsuki "Sentiment Analysis in Twitter: From Classification to Quantification of Sentiments within Tweets"

[8] UW-AI Class, (2016, December) "Sentiment Analysis - Sirisha". Retrieved from https://www.youtube.com/watch?v=doznOnG81xY&t=107s