Sentiment Analysis on Social Media Data

Sheeba Farook Patel
BE Computer
SSBT's COET Bambhori
Jalgaon, India

Pranjali Gopal Patil
BE Computer
SSBT's COET Bambhori
Jalgaon, India

Snehal Prabhakar Patil
BE Computer
SSBT's COET Bambhori
Jalgaon, India

Swati Sudhakar Thorat
BE Computer
SSBT's COET Bambhori
Jalgaon, India

***Abstract*—As increasing demand of social networking sites brought a new way of expressing individuals' opinion. Social networking sites have huge amount of information. The information can be seen by other user and helps to take the decision. The sentiment analysis is done by collecting the reviews of customer which are in the tweets form. The tweets opinions are can be positive, negative or somewhat in between the two. The previous approaches used unsupervised approaches. The unsupervised approach do not contain category and there is no accurate result. The proposed approach used supervised approach. The supervised approach, Navie Bayes machine learning algorithms used label datasets for the analysis. It automatically classifies the tweets taken from social networking sites and analyze them. Its main advantage is performance i.e. precision, accuracy will increase.**

Keywords— Machine Learning, Sentiment Analysis, Twitter

I. Introduction

The process of analyzing the dataset having assessment, attitudes or emotions which can be considered as a human things can be considered as sentiment analysis [1]. To find positive, negative or a neutral polarity in the sentence is somewhat difficult to find. There should be very high objective in order to sum up those reviews. As those reviews are written in different approach which seems difficult to judge. By using sentiment analysis of review, user is satisfied or not is come to know.

A. Background

As two types of machine learning techniques which are used for the sentiment analysis [9]. The first technique is unsupervised and another one is supervised [2]. As previously used unsupervised learning does not have category and target is not provided by them correctly and therefore conduct clustering [3]. The proposed system is using the supervised learning. In supervised learning the label data is provided to the model as this learning approach is based on label dataset.

In between the process the training is provided to the given dataset in order to get the desired output.

B. Motivation

As there is fast growth of using online resources, in particular social media. As user use those media for giving their review for the particular product. Those reviews are analyze by the companies for detecting their product level in the market. As they use traditional methods for the reviews which are in the form of interviews, questionnaires and surveys to gain feedback and insight into how customers felt about their products. That methods are very time consuming and expensive. So as by using sentiment analysis there will be valuable feedback on product and services as result in better decision making for the customer along with it help company to check their level in market.

II. Literature Survey

As in recent years many research is taken place on ``Sentiment Analysis”. As sentiment analysis use for the binary classification, comments or tweets to that can have the classes which are bipolar as positive or negative.

Peter David Turney, in paper [4] proposed, predicts the reviews by checking semantic orientation of phrase by using the unsupervised algorithm for calculating the positive or negative polarity for the given thumbs up or thumbs down.

Yan Luo and Wei Huang, in paper[5] proposed, system in which there is comparing the positive and negative sentences, here the web sentences are get extracted and label is provided to them manually which required more efforts.

Rui Liu et. al., in paper [6] proposed, system in which rule based method is used for sentiment analysis. They used the Chinese document and extract the polarity of them by using the sentiment word dictionary and according to the context information it get adjusted.

Lakshmi and Edward, in paper [7] proposed, to improve the quality of the dataset there is preprocessing of the data. The LSA technique is applied along with cosine similarity for the sentiment analysis is used.

Basant Agarwal et. al. in[8] , the phrase pattern is applied for the sentiment classification, uses rules based on part of speech and dependency relation for extracting the syntactic and contextual information from the dataset.

III. Our Approach

In the given proposed system is data from the social media is taken and analysis is done on it. For analyzing those data the unigram extraction technique is used. For this first preprocessed the tweets, after the preprocessing the adjectives are get extracted, then to that selected list machine learning based classification algorithms namely: Naive Bayes is applied with the Semantic Analysis based Synonyms list which gives the synonyms and similarity for the content feature which provides the polarity to the contain.

A. *Preprocessing the datasets*

As it is not easy to analyze the comments, so to make it easy to analyze pre-processing is done on those data. In pre-processing, the repeated words and the punctuation get removed to get quality of data.

B. *Feature Extraction*

After pre-processing the data, feature attraction take place were adjective from the datasets are get extracted. Then those adjective is used to give the polarity i.e. Positive and negative polarity in the given tweets which is useful for examine the sentiment of the individuals. As there is extraction of adjective, as there is disclaim the prefix and suffix word coming with the adjective in the given tweets.

Consider the example, i.e. flower nice by feature extraction, only nice is extracted from the tweets.

C. *Training and classification*

As for the classification of the data some technique is to be used, proposed system is using the supervised approach of machine learning algorithm i.e. Naive Bayes approach.

1) Naïve Bayes Classifier: Proposed by Kang et al., in[3] which is most commonly and simple used classifier. Naive Bayes Classification computes the posterior probability of a class, based on the distribution of the words in the document.

Class *c** is assigned to comments which are denoted by *d*,
Where, $c^* = argmaccP_{NB}(c|d)$

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

In this formula, feature is represented by f and the count of adjectives *fi* find in tweet data *d* represent by ni(d) . There are *m* features. Through maximum estimates, the parameters *P*(*c*) and *P*(*fjc*) are obtained.

For proposed system after preprocessing the dataset naïve bayes classifier the applied to them such that it gives the polarity to that particular dataset. For example, the painting is nice gives the positive polarity

2) Semantic Analysis: There is use of semantic analysis approach. After training and classification this approach is used in the proposed system. Semantic analysis obtains the similar words from the synonyms list where every word is related with one another. This are the list of words which are in English which are associated to each other. If the meaning of two words is similar then they are semantically similar. More emphatically, the algorithm decides the synonyms like similarity. The words contain in user sentence are check with synonyms list and then show the polarity of the statement given by user.

Consider the example in sentence “He is sad” the word “sad” as an adjective gets with the stored list of synonyms. Let’s take the two words ‘sorrow’ and ‘joyless’ are very same to the word ‘sad’. Now in semantic analysis process, ‘joyless’ is supplant ‘sad’ which shows the negative polarity.

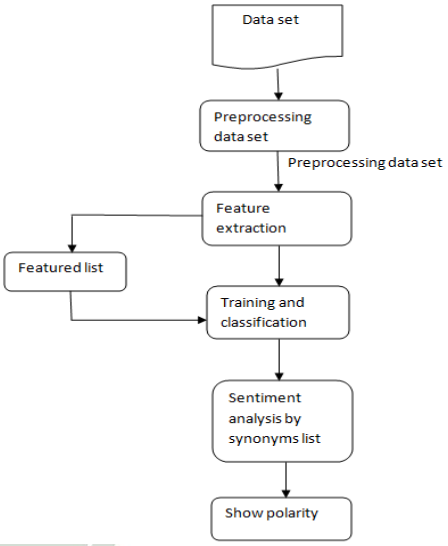


Fig. 1. System Architecture of Proposed System.

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the A4 paper size. If you are using US letter-sized paper, please close this file and download the file “MSW_USltr_format”.

IV. Implementation and Results

The proposed system is get implemented in Java language and the analysis result is shown by using the R language. As when tweets or comments are provided to the system the polarity of those comments i.e positive, negative or the neutral is get calculated and shown the result as the id of the user who gives the comments along with the total number of positive, negative and neutral comments or the tweets given by that particular person.

Algorithm:

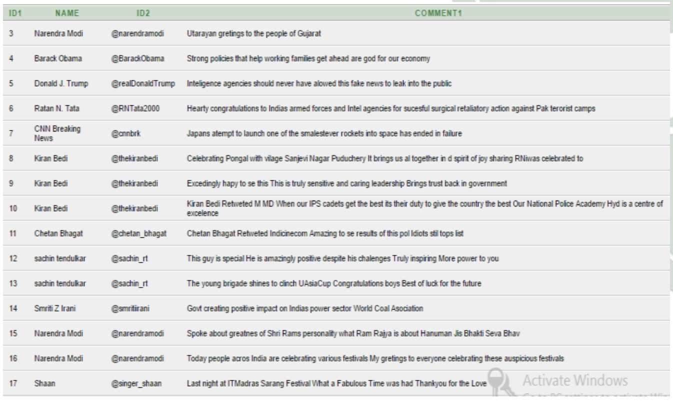
Input: Labeled Tweets Dataset

Output: Polarity of the given tweets.

1. Pre-processing:
Remove duplicate letters from words
Convert lower case
Trimming the dataset
Stemming the dataset

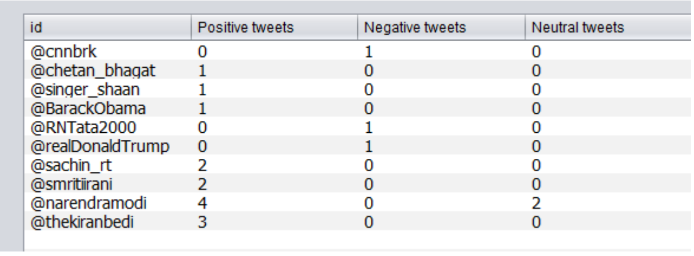
2. Extract the tweet data
Having continue letters in words:
Replace by single one
Remove:

```
    Stop words if there

    3. Extraction of adjectives from extracted tweets:
    For words in extracted tweets
    Find Adjectives=words in extracted tweets
    Return features

    4. Combining step 1 and step 2
    Pre-processed file =file path name
    Stop word file list=path name of file
    Extracted tweets=file path of extracted tweets list

    5. Train the step 4
    And apply classification

    6. Get the Synonym words and Similarity words for extracted tweets
    For each words in adjective list
    Extract adjectives in the tweet data ()
    For every positive words: a
    For every negative words: b
    Find
    If (a>b)
    Classify Positive
    Else if (a<b)
    Classify Negative
    Else
    Classify Neutral

    Print: sentiment polarity of the tweet data.
```



ID1	NAME	ID2	COMMENT1
3	Narendra Modi	@narendramodi	Utarayan gretings to the people of Gujarat
4	Barack Obama	@BarackObama	Strong policies that help working families get ahead are god for our economy
5	Donald J. Trump	@realDonaldTrump	Inteligence agencies should never have alowed this fake news to leak into the public
6	Ratan N. Tata	@RNTata2000	Hearty congratulations to Indias armed forces and Intel agencies for sucesful surgical retaliatory action against Pak terorist camps
7	CNN Breaking News	@cnnbrk	Japans atempt to launch one of the smalestever rockets into space has ended in failure
8	Kiran Bedi	@thekiranbedi	Celebrating Pongal with vilage Sanjevi Nagar Puduchery It brings us al together in d spirit of joy sharing RNiwas celebrated to
9	Kiran Bedi	@thekiranbedi	Excedingly hapy to se this This is truly sensitive and caring leadership Brings trust back in government
10	Kiran Bedi	@thekiranbedi	Kiran Bedi Retweted M MD When our IPS cadets get the best its their duty to give the country the best Our National Police Academy Hyd is a centre of excelence
11	Chetan Bhagat	@chetan_bhagat	Chetan Bhagat Retweted Indicinecom Amazing to se results of this pol Idiots stil tops list
12	sachin tendulkar	@sachin_rt	This guy is special He is amazingly positive despite his chalenges Truly inspiring More power to you
13	sachin tendulkar	@sachin_rt	The young brigade shines to clinch UAsiaCup Congratulations boys Best of luck for the future
14	Smriti Z Irani	@smritiirani	Govt creating positive impact on Indias power sector World Coal Asociation
15	Narendra Modi	@narendramodi	Spoke about greatnes of Shri Rams personality what Ram Rajya is about Hanuman Jis Bhakti Seva Bhav
16	Narendra Modi	@narendramodi	Today people acros India are celebrating various festivals My gretings to everyone celebrating these auspicious festivals
17	Shaan	@singer_shaan	Last night at ITMadras Sarang Festival What a Fabulous Time was had Thankyou for the Love

Fig. 2. Sample Tweets from the Twitter.



id	Positive tweets	Negative tweets	Neutral tweets
@cnnbrk	0	1	0
@chetan_bhagat	1	0	0
@singer_shaan	1	0	0
@BarackObama	1	0	0
@RNTata2000	0	1	0
@realDonaldTrump	0	1	0
@sachin_rt	2	0	0
@smritiirani	2	0	0
@narendramodi	4	0	2
@thekiranbedi	3	0	0

Fig. 3. Polarity of the Tweets.

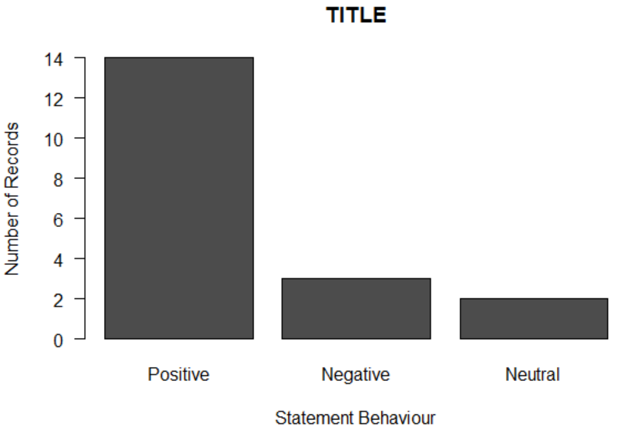


Fig. 4. Graphical Representation of Tweets.

CONCLUSION

As the proposed system is used to detect the sentiment of the given tweets by using supervised machine learning i.e. naïve bayes algorithm along with the semantic analysis. As it seen that naïve bayes gives the better result but when it combine with the semantic analysis its accuracy get increased. Such that it can be used in many application for detecting the reviews from the customer which can be more accurate than which to the unsupervised algorithm.

REFERENCES

[1] R. Feldman, " Techniques and Applications for Sentiment Analysis," Communications of the ACM, Vol. 56 No. 4, pp. 82-89, 2013.

[2] Y. Singh, P. K. Bhatia, and O.P. Sangwan, "A Review of Studies on Machine Learning Techniques," International Journal of Computer Science and Security, Volume (1) : Issue (1), pp. 70-84, 2007.

[3] Vossen Piek Maks Isa. A lexicon model for deep sentiment analysis and opinion mining applications. In Decision Support System, 2012.

[4] P.D. Turney," Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics(ACL), Philadelphia, pp. 417-424, July 2002.

[5] Y.Luo,W.Huang," Product Review Information Extraction Based on Adjective Opinion Words," Fourth International Joint Conference on Computational Sciences and Optimization (CSO), pp.1309 – 1313, 2011.

[6] R.Liu,R.Xiong,and L.Song, "A Sentiment Classification Method for Chinese Document," Processed of the 5th International Conference on Computer Science and Education (ICCSE), pp. 918 – 922, 2010.

[7] L.Ramachandran,E.F.Gehringer, "Automated Assessment of Review Quality Using Latent Semantic Analysis," ICALT, IEEE Computer Society, pp. 136-138, 2011.

[8] B.Agarwal,V.K.Sharma,andN.Mittal,"Sentiment Classification of Review Documents using Phrase Patterns," International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1577-1580, 2013.

[9] Gautam, Geetika, and Divakar Yadav, "Sentiment analysis of twitter data using machine learning approaches and semantic analysis", 2014 Seventh International Conference on Contemporary Computing (IC3), 2014.