SENTIMENT ANALYSIS OF TWITTER DATA

Neha Pandey[1], Ms.Surbhi Madan[2]

[1]*Student of Teerthanker Mahaveer University, CCSIT,Moradabad*

[2]*Assistant Professor Teerthanker Mahaveer University, CCSIT,Moradabad*

nehapandey9685@gmail.com

surbhi.compuers@tmu.ac.in

Abstract— **Sentiment analysis is a process that help to differentiate people point of view in terms of positive, negative,neutral form.Social media contain data in large amount in the form of tweets, blogs, and updates on the status, posts, etc. Through sentiment analysis we can decide the polarity of the data.It is useful in analyzing user view on a particular topic. Twitter is a popular social networking site in which user post called tweets.In this people express their view regarding any topic.Different techniques has been used discussed for sentiment analysis like Naive Bayes,Support Vector Machine.**

***Keywords*—SentimentAnalysis,Twitter, Naïve Bayes,Support Vector Machine.**

I. INTRODUCTION

As the use of social networking websites increases , a large data amount of data generated. People express their view and feelings using post,tweets.Sentiment analysis is concerned with the identification of people opinion.In this paper we review about the micro-blogging site twitter and classify tweets in positive ,negative and in neutral.Various techniques and machine learing process has been for sentiment analysis of twitter data.The several methods are used to extract the feature from the source text. Feature extraction is done in twophases.Tweet consist of positive, negative or neutral.People sometimes uses tweets phrases to express their view is it contain review about any movie,product.It is helpful for companies to track position and know about their product through analyzation of tweets.Here I discussed about different techniques for sentiment analysis like NaiveBayes classifier, SVM algorithmfor the sentiment analysis.The use of internet leads increament to the of texualinformation in the internet from last sometimes in human life. In the internet people uses micro-blogging websites to share their opinion and thoughts about any product or any topic. This creates lost of data in the form of text which is then used analysis to extract the user opinion or view on any topic. Opinions of any user is subjective that describe the user view and feelings towards the product.

II. TWITTER AND TWEETS

Twitter is a social networking website where its users can share their thoughts in the form of tweets.The length of any tweet is maximum of 140 character.People uses abbreviations to increase the words in tweets. Those acronyms lead to a broader dictionary of words, but also make it harder to analysis the tweets, since they create a broader feature space.

Another used term in twitter is retweet which shows the content of tweet posted by other user.People post retweets for sending original message to other followers.

A. Sentiment Classification

We can classify sentiment commonly in three way,which express the user point of view that will be beneficial for extracting emotion of people toward any text.

1) *Positive:*When user show anything good or in favour of topic using happy word is comes under the positive category.
2) *Negative:*When user show anything bad or against the topic using angry words is comes under the positive category.
3) *Neutral:*when user doesn't show any negative and positive reaction toward any topic its comes under neutral category.

III. LITERATURE REVIEW

[1] Batra and Rao collected about 60 millions of tweeets as dataset.Extraction of entity done by the Standard NER,they used augmentation they uses user tags and URL.Corporus of various reviews labeled as negative and positive for traingmodels..By using these corpus they gives the probability like Unigram or bigram used for positive or negative.

[2] An approach is introduce to improve sentiment classifier performance to get away from the noisy content. uppercases,Capitalizedwords,emoticons and@ All these replace by keyword as they are considered as intensifier.But words which doesnot show any opinion were removed.

[3] Po-Wei Liang et.al.(2014) [8] collected twitter data from Twitter API.used Twitter API to collect twitter data.The data for training comes in three category(mobile,camera,movie.Theylabelled their data as positive,negative and neutral.Tweets shows opinion are filtered.They remove the useless features by Mutual information.At the end the Tweet can be oriented as negative and positive.

[4] Barbosa et al.(2010) [4] proposed automatic method of two phase to classify tweets.They classify tweets in subjective and objective and in second phase they again divide subjective in two parts positive and negative.

[5] introduce an approach in which classification is based on punctuation ,words,patterns for sentiment as various features type.then it combines to make a single feature for classification For assigning sentiment they use K-Nearest Neighbhorstraregy.

[6] used data training as emotion for noisy label for performing sentiment analysis of twitter data.They used Naive Bayes, MaxEnt and Support Vector Machines (SVM).for models construction.They suggested that SVM outperformed models and unigram were very effective as features.

[7] They uses feature vector in two different form in which they extract the specific words and remove them.Now these all words converted into normal.They show that Naive Bayes give 75%accuracy and SVM gives 65% accuracy.

[8] Pak presented a way for corpus collection by automatic.ForPOS-tagging he used Tree Tagger.After that they find the positive negative and neutral set.

[9] Ohbyung Kwon, NamyeonLeea.[1]because of informal messages and hey show that Sentiment analysis is very challenging because of informal messages and multi languages.They reveals that big data analytic can be positively influence by quality of data.

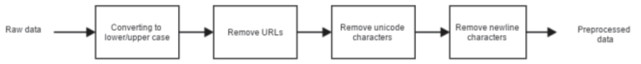
[10] Analyses sentiment on lexicon based approach in which sentiment dictionary is used with opinion words.And matches all of them with the content to find polarity.They allows scores based on sentiment describing positive,negative.

IV. PROPOSED SYSTEM

The system have various stages of development.A set of data called dataset is made using post of twitter.As tweets have misspelling and phrases.It is compulsory to perform a level wise sentiment analysis.This can be done in three stages.In the first stage preprocessing should be done.After that creation of feature vector by using relevant feature.After that we can classify tweets in the form of positive,negativend neutral classes by using different classifiers.

A. *Creation of Dataset:*Creation of data set using tweets about any topic.It can be created by taking review.

B. Pre-processing:


Raw data
Converting to lower/upper case
Remove URLs
Remove unicode characters
Remove newline characters
Preprocessed data

Preprocessing is an important phase in which all tweets converts into the lower case.Now ignore the URl @username changed by AT_USER and removal of hastag is done.Replacement of repeated word or character with the two occurence.And remove all the white spaces.

C. Creation of Feature Vector

Tweeter feature can be extracted in two ways.Firstly tweeter specific feature is extracted.After that hastag is removed from the words like #Product can replaced by Product.Is is possible that all tweets may or may not contain tweeter specific word.

1) *Naive Bayes Classifier:* Naïve Bayes Model classifier relate to the word frequency of any tweet or post.The words in the tweet are matches the sentimental word.In Naive Bayes all the features are different from each other.So it can use all feature in feature vector,The tweets are classified according to their weight of importance so that it can help in correct result.IT has higher precision then other classifier.And its scalable and easy to use.But it contain some disadvantage its recall and accuracy is low.

2) *Support Vector Machine (SVM):*SVM is better than Naive Bayes algorithm it is used for categorization of text.Means SVM divide tweets into single word.The basic idea is to find the hyperplanewhich is represented as the vector w which separatesdocument vector in one class from the vectors in otherclass.

IV. SENTIMENT ANALYIS FIELDS

In Sentiment analysis there are various research fields, computational linguistics,text analysisand natural language.It is used to extract information from the raw data.

It refers to the extraction of subjective information from raw data, often in text form. Although media also contain data in the form of sounds,images and videos.But these are studied less.All kind of media contain different type of sentiments.In accordance, in all media types different kinds of sentiments exist.sentiments indicates emotions and thought of any person. The sentiment can refer to opinions or emotions, even though these two types are related there is an evident difference. Through sentiment analysis a opinion is made something is positive, negative and neutral.Sentiment analysis refers to get person opinion regarding any subject. Other applications try to determine the overall sentiment of a document. Analzation of Sentiment can be difficult.Word used in tweet decide the sentiment of opinion.

Tweets are real time post.The tweets can be fatched using Twitter4j Api.In this Api they only accept a phrase or a line and returntweets lists that contains the searched term.

- Change the URL's (Uniform Resource Identifier)which are in tweets with URL keywords.

- Change the wordthat is in the form of @Person with the person name or user name.

- Avoidall slang wordsin the tweets.

- Remove of articlesfrom the tweets.

- Remove of all adjectives from the tweets.

These operation performed as they do not contain or associated with any sentiment.Large number of positive, negative and neutral tweets used as training dataset.

A Unigram approach was implemented In which each word in a tweets was compare with the training dataset.The probability is calculated for each word in a tweet by using naive bayesalgorithim .Then probability were compared later.the sentiment of the word is computed using three probability. If present positive sentiment probability is larger then it assign as positive sentiment,if present negative sentiment probability is larger then it assign as negative sentiment.

V. CONCLUSION

Analysing tweets is very helpful for determining the user opinion about any topic.Result based on the analyzation of tweets help to give a suggestion for any product or topic.Positive shows that something or some project is accepted by the user or negative result shows opposite.

In this paper we discuss sentiment analysis using machine learing technique in which we studies Naïve Bayes ,SVM.

REFERENCES

[1]S. Batra and D. Rao, "Entity Based Sentiment Analysis on Twitter", Stanford University,2010

[2]S. D. R. N. H. M. E. F. M. Cohen, P. Damiani,

\Sentiment analysis in microblogging: A practical

implementation," University of Buenos Aires, 2011.

[3]Pablo Gamallo, Marcos Garcia, "Citius: A Naive-Bayes Strategyfor Sentiment Analysis on English Tweets", 8th InternationalWorkshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland,Aug 23-24 2014, pp 171-175.

[4]L. Barbosa, J. Feng. “Robust Sentiment Detection on Twitterfrom Biased and Noisy Data". COLING 2010: Poster Volume,pp. 36-44.

[5]Dmitry Davidov, Ari Rappoport." Enhanced Sentiment Learning Using Twitter Hashtags and Smileys". Coling 2010: Poster Volumepages 241{249, Beijing, August 2010

[6]Go, R. Bhayani, L.Huang. “Twitter Sentiment ClassificationUsing Distant Supervision". Stanford University, Technical Paper,2009

[7]Sentiment Analysis of Twitter Data

ApoorvAgarwalBoyiXie Ilia Vovsha Owen Rambow Rebecca Passonneau”Sentiment Analysis of Twitter Data”Department of Computer Science

Columbia UniversityNew York, NY 10027 USA

[8]A.Pak and P. Paroubek.„Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp.1320-1326

[9]A. Cui, M. Zhang, Y. Liu, S. Ma, Emotion Tokens: Bridging the Gap among Multilingual Twitter Sentiment Analysis, Springer-Verlag,

Berlin, Heidelberg, 2011, pp. 238–249.

[10]Taboada, M., Brooke, J., Tofiloski, M., Voll, K., &Stede, M..“Lexiconbasedmethods for sentiment analysis”. Computational linguistics, 2011:37(2), 267-307.