Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 3, Issue 2; January-March, 2016, pp. 132-137

http://www.krishisanskriti.org/Publication.html

# Generic Sentiment Analysis

**Kriti Aggarwal[1], Surabhi Anand[2] and Madhu Kumari[3]**

*[1,2]B.Tech Student, CSE, NIT Hamirpur*
*[3]CSE, NIT Hamirpur*
*E-mail: [1]kritiaggarwal.128@gmail.com, [2]san13692@gmail.com, [3]madhu.jaglan@gmail.com*

**Abstract—***Owing to the burgeoning demand by the commercial and marketing sector for monitoring customer's views about the products and services, Sentiment Analysis came into existence. Sentiment Analysis is one of the implementations of text analytics techniques used for the recognition of subjective opinions in text data i.e. classification of text into categories such as "positive", "negative" and "neutral". A lot of work has been done in the area of Sentiment Analysis using social networking sites like Twitter, Facebook etc as a source of data for collecting user's views. This project involves classification of Tweets (collected from Twitter) into three categories: positive, negative and neutral via various pre-processing techniques like stopping, stemming, POS tagging and classification using Naïve Bayes Classifier. Also, the accuracy is calculated for classification of test tweets and effect of increasing training dataset is observed on accuracy of Naïve Bayes Classifier.*

## 1. INTRODUCTION

Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials [1]. That means the main purpose of sentiment analysis is to find the intended emotion or expression of the source (writer or speaker) with respect to the context. The intended emotion can be classified further as positive, negative or neutral towards the topic of interest. Neutral refers to no opinion. For example, the sentence "I love my phone Moto-G. It works pretty well" displays positive emotion of the user towards his/her cell phone, particularly Moto G. On the contrary, the sentence "I hate dogs." displays negative emotion of the writer towards dogs. There are following level of granularities at which sentiment analysis can be classified [2]:

1) Document level: This level of analysis is only applicable to the documents having opinions about single entity only and then classify whether the whole document opinion is positive or negative. (Pang, Lee and Vaithyanathan, 2002; Turney, 2002).

2) Sentence level: This level of analysis is applicable to sentences i.e. it classifies the intended opinion of the sentence as positive, negative or neutral.

3) Entity and Aspect level: This level of analysis is better than above two levels because it helps in determining what the users like or disliked in a better way. The main principle behind this level of analysis is that an opinion has two cardinal components i.e. the sentiment itself (positive, negative or neutral) and a target (of opinion). For example "Although Moto-G has nice camera, but its limited memory restricts user to utilize it to full capacity." The above sentence provides positive emotion in aspect of Moto-G's (target) camera quality but negative in aspect of its storage memory. Therefore, the objective of this level of analysis is to classify the emotions of the speakers in the reference to their aspect.

### Purpose

The project described in this paper intends to provide basic understanding of Sentiment Analysis and develop a functional classifier based on the concepts of Naive Bayes Classifiers to classify unknown Tweet Stream from Twitter. This task of analyzing tweets comes under the domain of Data Mining and uses the techniques of Natural Language Processing to a great extent.

The process of designing a functional classifier for sentiment analysis can be broken down into five basic categories. They are as follows:

I. Data Acquisition

II. Human Labeling/Training Set

III. Preprocessing/Feature Extraction

IV. Classification

V. Analysis of Data

The main purpose behind this document is to describe all the features (mentioned above) and working of the aforementioned application and simultaneously, providing base work for the future enhancements in the same field of work.

### Scope

The scope of this project is not only limited to classifying an unknown tweet stream as positive, negative or neutral. The concepts and techniques explained in this paper can be applied to other social media platforms, providing abundant reliable

amount of user data like Face book, comments, discussion forums, blogs and postings in other social networking sites. Also, it would help to acquire consumer opinions for other applications like marketing, campaigning, public relations, connecting people with same interests etc. thereby decreasing the need to conduct surveys and polls. The scope of this project is boundless and myriad facades of this project are yet to be discovered.

### 1.3. Motivation

Twitter is an online social networking service that enables users to send and read short 140- character messages called "tweets". Twitter was one of the most-visited websites and has been described as "the SMS of the Internet" [3]. Twitter has billions of users and millions of tweets are generated on daily basis. The principle motivation behind this project was to utilize the large amount of data produced on daily basis by the users on Twitter to implement the concept of Sentiment Analysis and to achieve a reflection of public sentiments regarding the entity of interest by developing a functional classifier based on the concepts of Naive Bayes Classifiers.

## 2. RELATED WORK

A vast research has been done in the area of Sentiment Analysis as it has wide prospects in solving many day to day problems.

The research in the area of Sentiment Analysis mainly started from early 2000. However, some books like Hatzivassiloglou and McKeown, 1997; Hearst, 1992; Wiebe, 1990; Wiebe, 1994; Wiebe, Bruce and O'Hara, 1999 provide detailed preliminary text related to the subject.

Various notable earliest researches on opinions and sentiments appeared in (Das and Chen, 2001; Morinaga et al., 2002; Pang, Lee and Vaithyanathan, 2002; Tong, 2001; Turney, 2002; Wiebe, 2000).However, Dave, Lawrence and Pennock, 2003 first mentioned about the term “Opinion Mining” and Nasukawa and Yi, 2003 first mentioned about the term “Sentiment Analysis”.

A survey report from Bo Pang and Lillian Lee, Opinion mining and Sentiment analysis [4], includes a thorough and detailed piece of work in the area of sentiment analysis of blogs, reviews, discussion forums etc. Various algorithms are discussed like Maximum Entropy, SVM and Naive Bayes. It not only includes techniques and approaches which can help to directly enable opinion-oriented information seeking systems but also focuses on methods which aim to address the new challenges raised by sentiment aware applications. [4]

Sentiment Analysis and Opinion Mining, 2012 by Bing Liu, [2] also presents a wide range of information on Sentiment Analysis from Document Sentiment Classification, Sentence Subjectivity and Sentiment Classification to Aspect-based Sentiment Analysis. It also includes topics like Sentiment Lexicon Generation, Opinion Summarization, Opinion Search and Opinion Spam Detection.

Apoorva et al. mentioned some pre-processing techniques along with removal of stop words to deal with slang and short words present in twitter data and also used emoticon dictionary.

## 3. DATA

Tweets are short burst of messages, which people use to express their personal messages, random thoughts, links, or anything, that fits in the character requirements. The limit of 140 characters forces the users to use slang, short words, emoticons, special characters, acronyms, hash tags etc, thus making the sentiment analysis task a little bit more challenging.

The project discussed here classifies the tweets in three categories: “positive, negative and neutral”. The training data set consisted of approx 20 thousand pre classified tweets in three categories positive, negative and neutral. Following table presents detailed description about the data.

**Table 1: Data Statistics**

| Type | Count |
|---|---|
| Positive Tweets | 9665 |
| Negative Tweets | 9665 |
| Neutral Tweets | 2274 |
| Total Tweets | 21604 |

Along with the twitter data, the project also required other datasets like stop words [5]. Though emoticons are a nice way to express emotion without using much text, but at times emoticons can be deceptive too. Consider the case if the writer is writing a positive tweet but using a negative emoticon at the end. For e.g. “I had so much fun today.☹ “.This makes the classification of the tweet ambiguous and hence may not give right results. Therefore, the classification discussed in this paper is purely text based and hence more generic.

## 4. METHODOLOGY

The project discussed here describes the process of designing a functional classifier for sentiment analysis. This process can be broken down into five categories which are Data Acquisition, Human Labeling/Training Set, Preprocessing/Feature Extraction, Classification and Analysis of Data. The machine learning algorithm used in this process is Naive Bayes. The steps listed above are further described in the following sections:

### 4.1. Data Acquisition

In the project discussed here, raw tweets constitute our data set, which are acquired using the python library - tweet stream that is responsible for providing a package for simple twitter streaming API. The two modes of this API are:

a) Sample Stream which is used to deliver a small, random sample of all the tweets streaming at a real time.

b) Filter Stream which is used to deliver tweet, matching a certain criteria ( a) specific keyword(s) to track/search for in the tweet, b)specific Twitter user(s) according to their user-id's , and c) Tweets originating from specific location(s) , only for geo-tagged tweets).

In this project, Sample Stream is used .Since it consists of lots of raw data, several filters can be used to extract the useful information. In this application, iteration is followed through all the tweets in sample tweets and the actual content of the tweets is saved in a separate file. Here a filter is applied on the language of user's account i.e. English so that only the tweets written in English are extracted. Further, another filter is applied on the to-be-labeled tweets so as to have a considerable amount of differential data present. This filter comprises of:

a) Removing short tweets (length is less than 20 characters).
b) Removing similar tweets (discard the tweet where content matching is more than 90%.
c) Removing non- English tweets.
d) Removing re-tweets i.e. the tweet containing the string RT.

### 4.2. Human Labeling/Training Set

In this project, the tweets are classified into following three labels so as to get better results in terms of predicting the sentiments of the writer:

a) Positive Tweet: If more of positive emotions like happy/ enjoying/ merry/ excited/ fun/ delightful/ ecstatic etc are observed in the tweet, then it is classified as a positive tweet. E.g. "I enjoyed very much today."
b) Negative Tweet: If more of negative emotions like sad/ gloomy/ doleful/ depressing/ anxiety/ distressed etc are observed in the tweet, then it is classified as a negative tweet." E.g. "I am feeling sad today."
c) Neutral Tweets: if the tweet doesn't convey any positive or negative emotion or merely is seen as display of some fact or information, then it is considered a neutral tweet. E.g. "The sun is round in shape."

The tweets classified were labeled only on the basis of information provided in the tweet under consideration. No assumptions were made based on past history of the user or some personal information of the user.

### 4.3. Preprocessing Tweets and Feature Extraction

The tweets obtained after applying above mentioned steps are now converted into more appropriate format so that it is easy to train the classifier. The steps required to achieve this agenda are described below:

#### 4.3.1. Initial Processing

Tokenization: The stream of text is broken down into words, symbols and other meaningful elements called tokens, separated by whitespace and/or punctuation characters. To make the comparison to English dictionary easier, tweet is normalized to lowercase and punctuation marks are removed. Since the project is more about analyzing the sentiments of the writer presented in the tweet, URL's and user references (usually marked by the tokens like http/https or @) can also be removed.

#### 4.3.2. Stopping

The most common words in a language are considered as stop words. For e.g. various stop words in English language are "a, the, an, in, are, is" etc. Such words must be filtered out during the pre-processing of the data as they decrease the efficiency of accuracy of the result. Though, mostly stop words don't represent any emotion but some stop words like not, not etc can turn a positive sentiment in to negative one. Therefore, one must refrain from removing such stop words. In this project, the corpus for stop words was obtained from NLTK.

### 4.4. Classification

The process through which the given data can be divided into different classes/label categories based on some common patterns is known as pattern classification. The main purpose of the project discussed here is to design a classifier that can accurately classify the tweets based on their sentiments which are "positive, negative or neutral."

The sentiment classification can be divided into two areas:

a) Contextual Sentiment Analysis: Under this analysis scheme, specific parts of the tweet can be classified according to the context. For e.g. "I like the display of Iphone very much but I don't like the cost of Iphone". In this sentence, positive sentiment is related with the display of Iphone but negative sentiment is related with the cost of Iphone.
b) Generic Sentiment Analysis: Under this analysis, the sentiment associated with the entire tweet is considered.

In the project discussed here, Generic Sentiment Analysis is used.

#### 4.4.1. Naïve Bayes Classifier

Naive Bayes is a very simple technique used for constructing classifiers. It represents a family of algorithms based on Bayes Theorem (conditional probability model) which can be abstractly represented as follows:

If a given problem instance to be classified can be represented by a vector

$$\mathbf{x} = (x_1, \ldots, x_n)$$

, where *n* means features or independent variables, then instance probabilities are assigned as

$$p(C_k|x_1, \ldots, x_n)$$

, for every *K* possible *classes.*

This classifier is based on the assumption that the value of a feature doesn't depend on the value of other feature, given the class variable. One of the advantages of using this classifier is that it needs modicum amount of training data to gauge the parameters required for classification.

Naïve Bayes Classifier from NLTK has been used in this project to test and train the data.

### 4.5 Analysis of Data

A dataset consisting of tweets steamed from Twitter is passed on to the classifier for defining in into a particular label.

## 5. TOOLS USED

### 5.1 NLTK Library

NLTK is a free, open source, community-driven project available for Windows, Mac OS X, and Linux, comprising of library to play with natural language. It is one of the major platforms used worldwide today to build Python programs to deal with human language text. It comprises of a large number of corpora and text processing libraries for various data mining operations such as classification, stemming, tokenization, parsing, pos tagging, named entity recognition etc.

In this project, nltk.classify.naivebayes module is used to perform various operations on the data set**.**
The Naïve Bayes Classifier has two probability distributions as parameters:

a) P (label): If no prior information is given about the input's features, then it gives the probability that an input will get each label.

b) P (fname=f_val|label): Given the label (label), it gives the probability that the given feature (fname) will receive a given value (f_val).

If the classifier comes across an input with a feature that has never been seen with any label, then it ignores that feature. For such features, value of 'None' is reserved.

## 6. EXPERIMENTAL SETUP:

The programming language used is Python. The project is divided into two parts:

a) Fetching the tweets from twitter: For this a twitter application account is created and then secret keys are generated using twitter API, which are further stored in json format. These authentication details are used while trying to connect to https://api.twitter.com/1.1/search/tweets.json? .This twitter API is featured to provide feeds to the authenticated request according to the keywords provided. A time duration along with date and max tweets limit to fetch the content is also provided so as to avoid memory leak in the system. The tweets are then stored in a file.

b) Sentiment Analysis on the fetched tweets: Preprocessing of tweets is done as explained above in section 4.3.Now, the feature vector is used to build a model which the classifier learns from the training data and further can be used to classify previously unseen data. After this, complete processed training dataset with proper labels are passed to the Naive Bayes classifier. The classifier module present inside the NLTK performs learning operation according to probabilistic model. Further it reads the tweets one by one, analyses them & stores the result of each tweet which is printed at the end. This process takes variable time depending upon amount of training set provided. While providing a dataset of approx twenty thousand classified tweets and 5 test tweets it took about 2 hours.

## 7. RESULTS

Multiple code executions were performed & the result set was obtained which can be divided into following parts:

a) Running the fetch code.
b) Sentimental analysis of test tweets.
c) Calculating accuracy.
d) Comparing sentiments of tweets of a given keywords.

### 7.1. Running the fetch code

For the fetch tweet code part, the keyword "new year 2016" was given so as to fetch result in this domain. Now the dataset obtained was limited to 200 tweets and also the tweets were asked to fetch for that particular day. The total time taken to fetch it was less than a minute.

### 7.2. Sentiment Analysis of test tweets

Now sentiment analysis is performed on the sample test tweet. The training data set consisted of approximately twenty thousand pre-classified tweets in three categories positive, negative and neutral. Three sample sentences were tested for their sentiment and analyzing them all correct results were obtained.

**Table 2: Result of Testing Sentiments**

| S.No. | Test Tweet | Result |
|---|---|---|
| 1 | So, you are going to hell for this deed | Negative |
| 2 | It's nice to learn programming language | Positive |

| | | |
|---|---|---|
| 3 | Mind your business or I will complain to neighbors about your bad behavior. | Negative |

### 7.3. Calculating Accuracy

The Naive Bayes classifier was used & accuracy was checked for 9 tweets as shown below:

**Table 3. Checking Accuracy of 9 test tweets**

| | |
|---|---|
| Total Tweets tested | 9 |
| Correct | 9 |
| Wrong | 0 |
| Accuracy | 100% |

The accuracy turned out to be 100% because it was applied on tweets from the training data.

But increasing the number of tweets to twenty thousand non labeled tweets the accuracy reduced to 86.96 as shown in the following figure:

**Table 4. Checking Accuracy of 21604 test tweets**

| | |
|---|---|
| Total Tweets tested | 21604 |
| Correct | 18784 |
| Wrong | 2820 |
| Accuracy | 86.95% |

However, this figure of accuracy can be increased by increasing the training set. To demonstrate this, variable size training sets was taken and corresponding accuracy was calculated as shown in the figure below:

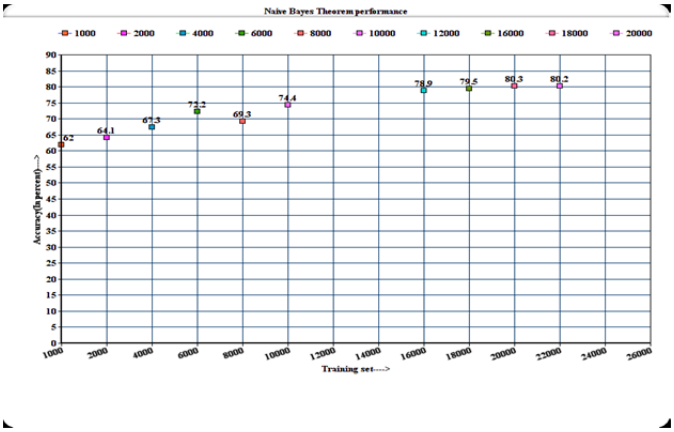


**Figure 1: Performance of Naïve Bayes Classifier**

It can be easily observed from the graph above that Naive Bayes theorem does perform even in less training set available. Although accuracy may not be so great but still its considerable. Also, the accuracy keeps on increasing as data set size is increased. Though some deviation in the trend can be observed which may be due to various reasons like test tweet consists of those words which were earlier having higher positive probability but due to increase in dataset it became less positive or maybe negative.

### 7.4. Comparing sentiments of tweets of a given keywords

The keyword "Iphone" was provided and tweets were fetched. Then, the sentiments related to this keyword were analyzed and following graph was obtained. The results obtained show that 67% tweets show positive review, 28% negative review and 5% neutral. Now it cannot be claimed that these results means review of Iphone instead it simply relates usage of Iphone word at various tweets (given to constraints: number of tweets fetched, date and time at which tweets were fetched as per this project).

**Table 5. Sentiment Analysis of 100 tweets for keyword "Iphone"**

| | |
|---|---|
| Total Tweets | 100 |
| Positive Tweets | 67% |
| Negative Tweets | 28% |
| Neutral Tweets | 5% |

## 8. FUTURE WORK AND CONCLUSION

The project presented in this paper is very generic application of sentiment analysis as it is able to categorize tweets into three categories as positive, negative or neutral. However, this project doesn't give effective results with the sarcastic or ironical sentences. Also, it doesn't take into account the emoticons used in the tweets. More work can be done on inclusion of emoticons. This project only focuses on Naïve Bayes Classifier. Further work can be done by using other classifiers like SVM and Maximum Entropy Classifier and comparison of their performances can be done. We hope to do further work in the areas discussed above.

## 9. ACKNOWLEDGEMENTS

This work was done under the guidance of Prof. Madhu Kumari, Assistant Professor, Computer Science and Engineering Department, NIT Hamirpur. The authors would like to express their deepest sense of gratitude and sincere thanks to the project mentor, Dr. Ms. Madhu Kumari, Assistant Professor, Department of Computer Science & Engineering, NIT Hamirpur for her revered guidance, insightful advice, encouragement, critics and valuable suggestions throughout the course of this project work. The authors would also like to thank Dr. Narottam Chand, Associate Professor and Head of Department of Computer Science and Engineering, NIT Hamirpur for his constant co-operation and support. The authors would also like to thank their team mates Sachin Kumar Vardhan and Amit Kumar for their cooperation.

**REFERENCES**

There are a number of resources and references that have contributed to the successful completion of this project. These resources are listed below:


[1] Wikipedia, Sentiment Analysis https://en.wikipedia.org/wiki/Sentiment_analysis
[2] Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.
[3] Wikipedia,Twitter,https://en.wikipedia.org/wiki/Twitter
[4] Bo Pang and Lillian Lee, Opinion mining and Sentiment analysis
[5] nltk.corpus.stopwords
[6] Shachi H Kumar, University of California, Project Report on Twitter Sentiment Analysis
[7] Christopher D.Manning, Prabhakar Raghavan and Hinrich Schutze, Introduction to Information Retrieval
[8] Tom M. Mitchell, Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression
[9] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments
[10] Hatzivassiloglou, V., & McKeown, K.R. Predicting the semantic orientation of adjectives. In Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL, 2009.
[11] Mukesh Yadav, PIIT NEW PANVEL Varunakshi Bhojane, PIIT NEW PANVEL, Data Analysis & Sentiment Analysis for Unstructured Data In Proceedings of International Journal of Engineering Technology, Management and Applied Sciences December 2014.
[12] Johann Bollen, Alberto Pepe and Huina Mao. Modeling Public Mood and Emotion: Twitter Sentiment and socio-economic phenomena. In Proceedings of AAAI Conference on Weblogs and Social Media (ICWSM), 2011.
[13] Luciano Barbosa and Junlan Feng. Robust Sentiment Detection on Twitter from Biased and Noisy Data. In Proceedings of the international conference on Computational Linguistics (COLING), 2010.