IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 7, July 2015.
www.ijiset.com
ISSN 2348 – 7968

An Approach towards Real Time Customer Attitude Analysis using ML Technique - Twitter

Akshata B. A[1], Mouneshappa D[2] and Karuna C. G[3]

[1] Department of Computer Science, Krishnamurthy Institute of Technology and Engineering, Hyderabad, India.
akshata_angadi@gmail.com

[2]Assistant Professor, Department of Computer Science, Krishnamurthy Institute of Technology and Engineering, Hyderabad, India
mouneshrd11@gmail.com

[3]Assistant Professor, Department of Computer Science, K.L.E. Institute of Technology, Hubli, India
karuna7674@gmail.com

Abstract

We're living in an amazingly hasty world, where practically every phrase seems to be shortened to save time. Whether solely social or firmly business, shorthand has become must in the language of business. Thus automation has come into existence. To remain competitive in a business milieu and build tangible value for customers there is a necessity of integrating. Massive amount of data is available on web in form of texts on social networking sites. Thus leveraging unstructured data to discover knowledge will be helpful for customers to make right decisions. Considering these thoughts paper helps in knowing the sentiments hidden in the text on company or product or an individual.

Twitter, a micro blogging site has turned into an exclusive means for every updates over the civilization. It is a place where people gather and confer their interests. This high variant data present in sites has increased the prospect of predictions about specific outcomes, without introducing the whole market mechanics. Extracting such data and Analyzing it is an issue in present era. Examining and Mining such information present on sites help in finding people's opinion, sentiments, attitudes and emotions. The paper surveyed on the different techniques of Machine Learning and concentrates on extraction of data like tweets, retweets etc. from popular Social networking site i.e. Twitter, Processing, Applying Feature Selection Method and lastly classifying the test case based on the trained set.

Keywords: *Classification, Data Mining, Feature Selection, Learning Algorithm, Mutual Information, Unstructured Data.*

1. Introduction

Social media analysis is a new epoch in market research exploiting the power of real-time big data and state-of-the-art data analytics. Twitter popularity is growing, and the most interesting aspect from the data analysis point of view, is that a large quantity of data is publically available, as most of the people prefer to publish their posts openly, in contrast to other social networks like Facebook or LinkedIn, where the information is only accessible to people that are friends or connections. [1]Twitter has attracted many users by its exquisite features in a short span. The benefits include Tweeting and Retweeting facility, Adding Favorites, Following people, Sharing, Can find the news, track the Current Trends, Companies, Contacts, Celebrities and many more to pen. Thus a rich source of data is available on twitter due to the increased usage. This rich data help in knowing users passion and excitement towards particular product. Number of people who use twitter actively has increased to about 300M .[2]Twitter users generate more than 500M tweets each day, these users are also overwhelmed by the massive amount of information available and the huge number of people they can interact with.

Marketers aren't the only ones thinking about social media analytics. Twitter active members are from all fields: they may be politicians who tweet as soon as they get a notice on any topic same way youngsters, celebrities, adults, doctors, engineers too. The updates are foreseen on twitter here so quick like kerosene catching a fire. This made to choose Twitter as a case study and implemented a project to know the thoughts of the people by applying Supervised Machine Learning Technique.

"What other people think" has always been an important piece of information for most of us during the decision-making process. Marketers have always needed to monitor media for information related to their brands whether it's for public relations activities, fraud violations, or competitive intelligence. But fragmenting media and changing consumer behavior have crippled traditional monitoring methods [3].These words were taken as motivation and started with this project to improve the marketing strategy by analyzing the opinions and sentiments of the users.

ISSN 2348 – 7968

Sentiment analysis can be cast as a classification problem where the task is to classify texts into three categories depending on whether they convey positive or negative or neutral feelings. To build classifiers for sentiment analysis, we need to collect training data so that we can apply appropriate learning algorithms. In this paper, dataset is collected from social networking site i.e. Twitter. Twitter API is used to get the authentication for accessing data from the site. Collected data set is analyzed based on the Feature Selection Technique to train the system. Using the trained data, the data under test is classified using a suitable Supervised Machine Learning Approach. Finally the system results in consolidated discovered knowledge. Thus in our project we have given a framework which may help improve the tactics of business.

2. Discussion

People usually get mystified with the association of terms Machine Learning and Data Mining. Here it goes:
Data mining and Machine Learning are like two cousins. Though the origin of both the concepts is diverse but both are functioning towards the analogous lines at last. The field of machine learning grew out of the effort of building AI. Its chief anxiety is making a machine learn and adapt to novel information (knowledge discovery through learning). It is often used in finding the relationship in hidden data among large heaps of data available whereas Data Mining is extracting concealed patterns of large data. We can say Machine Learning can be a tool for it, using different programmed algorithms. Data Miners typically have strong groundwork in machine learning, but also have an eager interest in applying it to large-scale problems.
Data miners sort through huge data sets using sophisticated machine learning algorithms to identify undiscovered patterns and establish hidden relationships. Machine learning programs detect patterns in data and adjust program actions accordingly [4].
Machine learning algorithms are so practical because it forces people to think before they use it. The roles of the input data and the model preparation process differs Hence thinking which to use before proceeding matters lot. Selecting the one that is most appropriate for problem in order to get the best result is an important role. Learning paradigms are divided as follows [5]:

- Supervised Learning: Input data is called training data and has a known label or result such as spam/not-spam or a stock price at a time. A model is prepared through a training process where it is required to make predictions and is corrected when those predictions are wrong. The training process continues until the model achieves a desired level of accuracy on the training data. Lastly the trained data is used for testing purpose to find the accuracy of training. Training should be done for huge data for better accuracy. Example problems are classification and regression. Example algorithms are Logistic Regression and the Back Propagation Neural Network. Fig.1. shows stages of supervised method

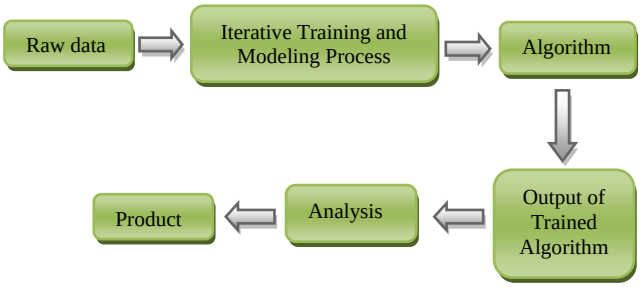

Figure.1. Supervised Learning Model

- Unsupervised Learning: Input data is not labeled and does not have a known result. A model is prepared by deducing structures present in the input data. Clusters/Groups are formed based on certain features of the input data. Lastly manual verification is done to check the outcome as per the requirement. Example problems are association rule learning and clustering. Example algorithms are the Apriori algorithm and k-means. Fig.2. shows stages of unsupervised methods.

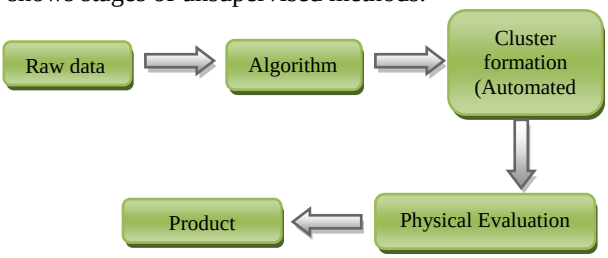

Figure.2. Unsupervised Learning Model

- Reinforcement Learning: Input data is provided as stimulus to a model from an environment to which the model must respond and react.

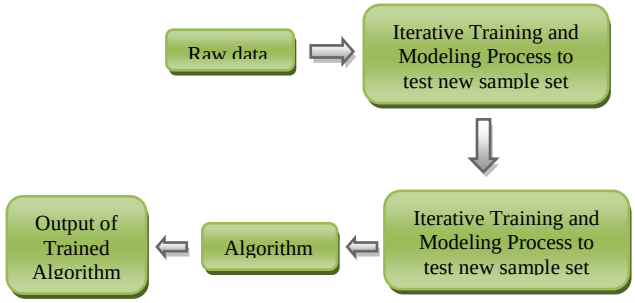

Figure.3. Reinforcement Learning Model

Response/feedback by Iterative training process for new sample set gives as punishments and rewards in the environment; this is unlike from supervised technique.

ISSN 2348 – 7968

Example problems are systems and robot control. Example algorithms are Q-learning and Temporal difference learning. Fig.3. shows stages of unsupervised methods.

- Semi-Supervised Learning: Input data is a mixture of labelled and unlabelled examples. There is a desired prediction problem but the model must learn the structures to organize the data as well as make predictions. Example problems are classification and regression. Example algorithms are extensions to other flexible methods that make assumptions about how to model the unlabelled data. When crunching data to model business decisions, you are most typically using supervised and unsupervised learning methods. A hot topic at the moment is semi-supervised learning methods in areas such as image classification where there are large datasets with very few labeled examples. Reinforcement learning is more likely to turn up in robotic control and other control systems development [5].

3. Machine Learning

There are many groups of machine learning techniques like Association, Neural Networks, Kernel Methods, Sequential Patterns, Decision trees etc. Here we discuss popular types of machine learning methods.

Classification: Classification is supervised learning technique used to assign per-defined label to object on the basis of tokens. Hence data training is must in classification. Data training is required in turn the Model is created from training process. This is further used to classify new test cases.

Suppose you have given a set of vegetables in a basket and said to classify the different elements present in the basket. Now, as you already know the shapes and colors of those, classifying them is not a big task. Here, the process of knowing before is training. It means you are trained before you classify the items. Same way, for Classification the machine needs to be trained before sending the input data for test whereas in Clustering Technique it looks something different. Consider you have given the same basket but you don't know anything about it like its shape or how to categorize. Hence we move for choosing based on physical character as size. So, here we consider size as feature. Example: Cabbage – green and big size, Tomato-red and small size etc. So clustering process doesn't require training as supervised methods need.

Clustering: Clustering is the unsupervised classification of patterns observations data items or feature vectors into group's clusters [6]. Clustering does not require training data. Clustering does not assign per-defined label to each and every group. Clustering is the process of examining a collection of "points," and grouping the points into "clusters" according to some distance measure [7]. The goal is that points in the same cluster have a small distance from one another, while points in different clusters are at a large distance from one another. Clustering is useful to identify different information because it correlates with other examples so you can see where the similarities and ranges agree.

Regression: Classification and regression are learning techniques to create models of prediction from gathered data. Regression is a prediction method that is based on an assumed or known numerical output value. This output value is the result of a series of recursive partitioning, with every step having one numerical value and another group of dependent variables which branch out to another pair such as this. The regression tree starts with one or more precursor variables, and terminates with one final output variable. The dependent variables are either continuous or discrete numerical variables [8].

Association: Association analysis is the task of finding appealing relations in huge data sets. There unseen relationships are then expressed as a collection of association rules and frequent item sets. Frequent item sets are collection of items that frequently occur together. And association rules propose a strong relationship that exists between two items. Consider customer has come to the shop, if he/she buys diaper then there are good chances that they will buy wine too. Thus with the frequent item sets and association rules retailers have a much better understanding of their customers.

Choosing a right method based on the requirement is essential. In this paper we are working on sentiment analysis. Hence we choose classification method. In classification technique there are many algorithms like SVM (Support Vector Machine), ME (Maximum Entropy), (NB) Naive Bayes etc. We go for using Naïve Bayes technique. Before using the classifier it is necessary to make use feature selection method for better vocabulary setup for classification.

4. Proposed Work

Proposed system reveals the process of Sentiment Analysis taking twitter as a case study. The proposed system emphasis on classification phase tweets using supervised learning algorithm, this in turn helps in predicting the opinions of people. Sentiment Analysis of customers opinions assist in gaining the wider community attitudes behind certain subject. This benefits the organization to promote and advance their brand. The layout of the proposed model reveals the ordered steps to meet the objective of the problem definition. This project contributes to the field of sentiment analysis, which aims to extract attitudes, opinions from tweets.

ISSN 2348 – 7968

4.1. Structural Design

The proposed system is shown in Fig 4.

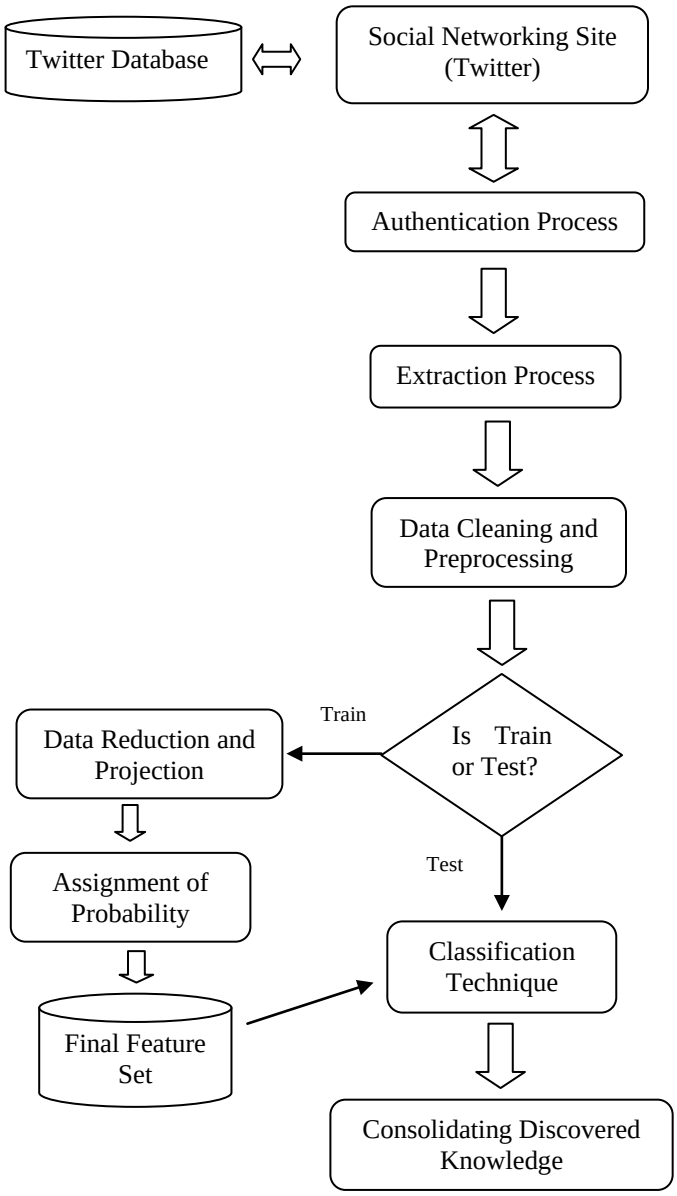

Figure.4.Proposed System

The steps of proposed system are:

1) **Authentication Process[**12]

User has to authenticate his/her application to Social Networking Site (Twitter) via his/her Twitter account. This process issues keys like Consumer key, Token access key etc to the end user.

2) **Extraction Process**

Using the authentication keys user is enabled to extract the status (Tweets, Retweets, Timeline etc) of his/her account. In Twitter, individual message implies the status info of a user.

3) **Data cleaning and Preprocessing**

Removal of unwanted characters (stop words, punctuations, special characters, links, @ etc) from the statuses collected from Social Networking Sites. Take only text part of status information and user names for further processing. The text part is tokenized into set of tokens.

4) **Data Reduction and Projection**

Feature Selection Method selects only those features that are most relevant to assign the probability to the tokens.

5) **Assignment of Probability**

Depending on the Features and their need in finding the polarity (positive, negative or neutral) of tweet under test, assign the probability values to the filtered features.

After these steps, train the system database with final feature set.

6) **Classification Technique**

It includes two levels: Defining the suitable Classifier algorithm and implementing the classifier defined. The former is concerned with the building of a classification model by describing a set of predetermined classes from a training set as a result of learning from that dataset. Each sample in the training set is assumed to belong to a predefined class, as determined by the class attribute label. The later involves the use of a classifier built to predict or classify unknown objects based on the patterns observed in the training set.

7) **Consolidating Discovered Knowledge**

Documenting and reporting results, or using them inside the proposed system.

5. Hypothesis of Feature Selection

A universal problem of intelligent or learning agents is where to focus their attention. The main tasks include:

a) Finding what are the necessary aspects of the problem
b) Recognizing difference between the relevant and irrelevant parts of experience

The main facet is Feature selection defined as “a selection of subset of a learning algorithm’s input variables, upon which it should focus attention, while ignoring the rest i.e. DIMENSIONALITY REDUCTION”.

5.1. Need of Feature Selection

In machine learning, the information about the target class is inherent in the variables. Naive Bayes theoretical view says “More features, More information and More discrimination power”.

Many explored domains have hundreds to tens of thousands of variables/features with many irrelevant and redundant ones. In such domains with many features, the underlying probability distribution can be very complex and very hard to estimate the dependencies between

ISSN 2348 – 7968

variables i.e. Irrelevant and redundant features which can confuse learners, Limit training data, and Limit the computational resources. Hence Curse of dimensionality plays important role**.**

Note: Number of samples to achieve accuracy (as by considering more or maximum number of features), grows exponentially with the number of variables used for training (Fig.5.). And if number of training examples is fixed, chances of lower classifier's performance for a large number of features can be seen.

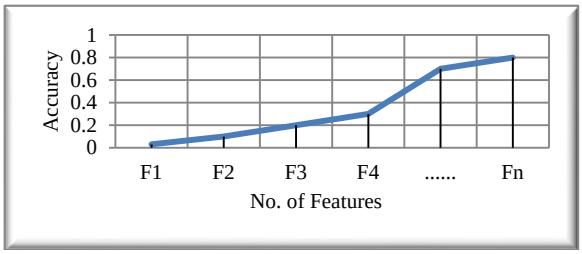

Figure.5. Performance Measurement of Classifier.

When dealing with a large number of variables there is a need for dimensionality reduction. Thus Feature Selection helps in improving learning algorithm's performance. So, the goal is to find an optimal feature-subset i.e. one that maximizes the scoring function.

Information Theoretic Criteria: [9] In case of discrete variables, most approaches use empirical estimates of mutual information between features and the target which is given in the formula (1):

$$I(x_i, y) = \sum_{x_i} \sum_{y} P(X = x_i, Y = y) \log \frac{P(X = x_i, Y = y)}{P(X = x_i) P(Y = y)} \quad \text{----(1)}$$

Here probabilities are estimated from frequency counts. It is a measure for "how much information (in terms of entropy) two random variables share [10].

Thus having a goal to find the optimal feature subset or at least a "good ones", variety of Classification methods can be used like: Filters, Wrappers and Embedded Methods. Choose suitable feature selection measure, which helps in adding the scoring function to features to place them in a feature subset and acts as a strategy to search the space of possible feature subsets.

In this case we used Mutual Information with Wrapper classification method in which Learner is considered as a black-box. Interface of this black-box is used to score the variables of subsets according to the predictive power of the learner when using the subsets. The Results may vary for different learners.

Thus if F is Values of features and B is Predicted Values of behavior, then MI measures the amount of dependencies between the F and B; Given the value of one, how much do we know about the other.

Mutual Information between F and B then defined as in (2)

$$I(F,B) = \sum_{f,b} p(f,b) * \log \frac{p(f,b)}{p(f)*p(b)} \quad \text{----(2)}$$

If there is no relation between F and B then p(F,B)=p(F)*p(B) implies I(F,B)=0.

Now rewriting the above formula (2) as needed by the requirements of project is as below in (3)

$$I(W,B) = \sum_{c_j \in C} p(w, c_j) * log \frac{p(w,c_j)}{p(w)*p(c_j)} + \sum_{c_j \in C} p(\sim w, c_j) * log \frac{p(\sim w,c_j)}{p(\sim w)*p(c_j)} \quad \text{----(3)}$$

Where

$p(w,c_j)$= Probability of word being in class Cj.

$p(\sim w,c_j)$= Probability of word not being in class Cj.

$p(w)$= Probability of word being in all classes.

$p(\sim w)$= Probability of word not being in all classes.

$p(c_j)$= Prior Probability of class Cj.

Usage the formula (3) explained in section 6 in form of pseudo code.

6. Classification using ML Technique

The simple pseudo code to calculate MI values, train the classifier and test the data are given below:

- Extract the tweets from Twitter database.
- Pre-process the tweets.
- Assign the labels such as positive (C1), negative (C2) and neutral (C3) to each preprocessed tweets.
- Create the positive (C1), negative (C2) and neutral (C3) documents containing tweets in tokenized form.
- Create a Vocabulary set containing unique tokens picked from all documents.
- Call the Training Function
- Call MI Feature Selection method
- Call the Testing Function

In Training Phase :

- Find total number of documents.
- Find the Prior probabilities for all documents (C1,C2 and C3) by applying the formula
  - Prior probability of Class (C1/C2/C3) = Number of C1/C2/C3 documents / Total number of documents
- For each token from Vocabulary
  - Find term frequency of picked token for each class (C1/C2/C3)
- For each token from Vocabulary
  - Find the Conditional Probabilities by applying the formula
  - Conditional Probabilities of the token picked for each class (C1/C2/C3) = [Term Frequency of picked token for class (C1/C2/C3) +1] / [Total number tokens in the class (C1/C2/C3) + Vocabulary]
- Return Vocabulary, Prior Probabilities and Conditional Probabilities.

ISSN 2348 – 7968

In MI Feature Selection Phase:

- For each token in the Vocabulary
  - Find the MI value for picked token by applying the formula given below
  - MI(W,C) = P(W, C1/C2/C3)*log[P(W, C1/C2/C3)/P(W)*P(C1/C2/C3)] + P(~W, C1/C2/C3)*log[P(~W,C1/C2/C3)/P(~W)*P(C1/C2/C3)]

  Where
  - MI(W,C) = Mutual Information of Token in class(C1/C2/C3)
  - P(W, C1/C2/C3) = Probability of word being in class (C1/C2/C3) = word count in C1/C2/C3 class / Total number of tweets in C1/C2/C3
  - P(W) = Probability of word being in all classes.= word count in all tweets / Total number of tweets.
  - P(~W,C1/C2/C3)= Probability of word not being in class (C1/C2/C3) = count of word not in C1/C2/C3 class (i.e. Total number of tweets in C1/C2/C3 - word count in C1/C2/C3 class) / Total number of tweets in C1/C2/C3.
  - P(~W) = Probability of word not being in all classes = count of word not being in all tweets (i.e. Total number of tweets - word count in all tweets) / Total number of tweets.
  - P(C1/C2/C3) = Prior Probability of class (C1/C2/C3)
- Add the token into vocabulary
  - if its MI value is greater than or equal to Set Thresh hold value.
  - Otherwise
  - Discard the token
- Return the new Vocabulary

In Testing Phase

- Pick the token from the tweets under test
- For each class (C1/C2/C3)
  - Find the Score for (C1/C2/C3) as logarithm of prior probability of class (C1/C2C3)
  - For each token from tweets under test
    - Find the Score for (C1/C2/C3) as
    - Score for (C1/C2/C3) = Score for (C1/C2/C3) + logarithm of Conditional Probability of the token picked for each class (C1/C2/C3)
- Return Maximum of the Score for (C1/C2/C3)

7. Experimental Results

Figure.6. getting tweets for the query entered. Here the count given is 10 and the query/ the topic is sony.

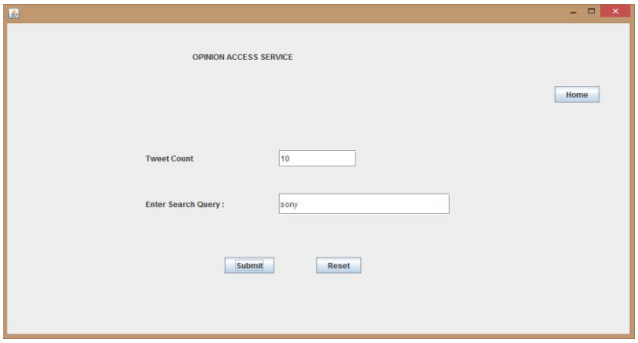

Figure.6-. getting tweets for query entered

Clicking on Submit button process of tweets extraction is carried out. The tweets are sent to training process along with labels or testing process or discarded.

Table 1. Sample tweets collected from Twitter account.

Sample tweets collected from Twitter account for Sony brand Labeling Information	Label-info
1. @Suresh Sony Xperia is Beautiful 5-inch Full HD display	C1
2. @Kiran Sony mobile Sleek, strong, waterproof design	C1
3. @Aks Sony Xperia has Pleasing OS and UI combination	C1
4. @Suresh Sony Xperia Z3 Water Proof Sensational smart phone	C1
5. @Geeta Sony Xperia Z3 Stunning but with minor flaws	C3
6. @Aks Sony's Glass panel at the back scratches easily	C2
7. @Kiran Sony's Flaps covering headphone and usb ports are annoying	C2
8. @Sans The battery of Sony got so hot whilst watching a film, that the screen cracked.	C2
Test Case 1- @Salma Build quality of Sony Xperia Z3 is stunning, its waterproof and the attention to detail is exemplary.	?
Test Case 2- @Seema Sony Xperia is average in lookwise, battery becomes hot soon and found flaws when I used it	?

Table.1. [11] shows the sample tweets collected from twitter account. Here we can see: In 10 tweets 8 are labeled and 2 tweets are used for test. These tweets are further sent for Tokenizing and Feature selection method and lastly to the Classifier to predict the class of test case. Table.2. shows the features selected after evaluating the MI values. MI helps in Data reduction process. The threshold value is set in MI hence the feature values that are greater than the threshold value which has been set are selected as an input to the classifier.

ISSN 2348 – 7968

Table 2. Final sample training set collected after learning and MI evaluation process:

Tweet No.	Keywords in the Tweets					Class
Training set						
1	Beautiful	inch	smile	Display		C1
2	Sleek	Strong	waterproof	Design		C1
3	Pleasing	combination				C1
4	Waterproof	sensational	smart			C1
5	Stunning	Flaws				C3
6	Scratches	problem				C2
7	Headphone	ports	annoying			C2
8	Hot	Whilst	cracked			C2
Test data						
Build	quality	Stunning	waterproof	exemplary	best	?
Average	Hot	Flaws				?

Table.2. features are used as input to the Naïve Bayes Classifier. Figure.7. shows the final Bayes output where in the Test Case 1 is classified as positive and Test case 2 is classified as negative.

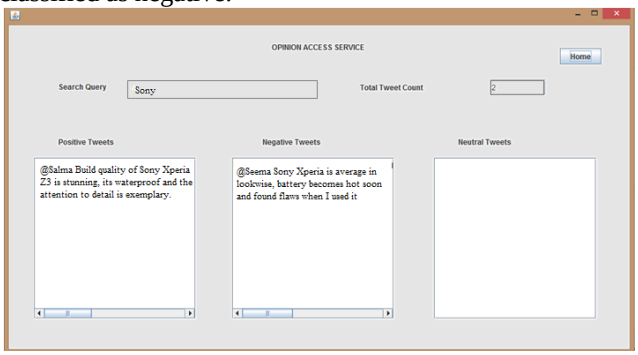

Fig.7. Final Output

8. Conclusion

Social media analysis is a latest epoch in market research binding to the power of real-time big data and futuristic data analytics. Nowadays 96% of the Companies do not obstruct access to Social Networking sites. And even the employees never wish to move to a company that blocks them from using social networking sites. This reveals the significance of the Social Media on people. Every day about 500M tweets are generated. Due to vast amount of user's opinions, views, feedback and suggestions available through the web resources, it's very much essential to explore, analyze and systematize their views for better decision making. This means that beyond simple likes - thumbs up, we can check out qualitatively into people's views, motivation and opinions at quantitative level. Thus in this paper we have used feature Based Technique i.e. Mutual Information for making a better vocabulary input to the Naïve Bayes classifier. The model is trained and the results are stored as a set of patterns, which helps in exploring or to make predictions further.

Naive Bayes Classifier has been sinculcated in paper as it helps in quickly making mining models to discover relations between input features and predictable features. The statement that needs to be stressed while working is "Choose the proper Training data and a Classifier for better results". Huge and proper training data leads to accurate results.

References

[1] Albert Bifet , Geoff Holmes and Bernhard Pfahringer "Detecting Sentiment Change in Twitter Streaming Data", 2nd Workshop on Applications of Pattern Analysis, JMLR: Workshop and Conference Proceedings 17, 5-11, 2011.

[2] http://www.internetlivestats.com/twitter-statistics/.

[3] Dave, Kushal, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of International Conference on World Wide Web (WWW) 2003.

[4] Joomla ,"Predictive Modelling Resources", Publiahed on Saturday ,Jan 1,2011

[5] Jason Brownlee "A Tour of Machine Learning Algorithms" on November 25, 2013.

[6] Furnkranz J., Gamberger D and Lavrac N, "Machine Learning and Data Mining" Springer, ISBN: 978-3-540-75196-0, XVIII, 334 p.,2012.

[7] Martin Brown VP of Technical Publications Couchbase, "Data Mining Techniques", IBM, DeveloperWorks, 2012.

[8] Clarisse "Difference between Classification and regression", difference between.com, 2011.

[9] Isabelle Guyon and Steve Gunn. Nips feature selection challenge. http://www.nipsfsc.ecs.soton.ac.uk/, 2003.

[10] E. Amaldi, V. Kann: The approximability of minimizing nonzero variables and unsatisfied relations in linear systems. (1997).

[11] Karuna Gull, Akshata B. Angadi (2015) " Automated Data Analytics in Measuring Brand Sentiment Using ML Technique from Online Conversations - Twitter", Third International Conference on Emerging Research in Computing, Information, Communication and Applications (ERCICA-15), Springer as a part of the series in advances in Intelligent Systems and Computing (AISC) in Press.

[12] Narashima S. Purohit, Meghana Bhat, Akshata B. Angadi, Karuna C. Gull, "Crawling through Web to Extract the Data from Social Networking Site - Twitter", IEEE National Conference on Parallel Computing Technologies - PARCOMPUTECH India 2015, doi:10.1109/PARCOMPTECH.2015.7084522, ISBN: 978-1-4799-6916-6, pp. 1-6, http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=7084522.

ISSN 2348 – 7968

Akshata B. Angadi has received the B.E. degree in Computer Science & Engineering from Visvesvaraya Technological University, India in the year 2011. She has been working in the area of data mining and social networking since 2012. She has published 5 National and 3 International papers in Conference Proceedings. She has published 6 papers in International Journals. She worked as Lecturer in Computer Science Department in K.L.E.I.T, India. Presently pursuing M.Tech in Computer Science.

Mouneshappa D. has done his B.E. from Visvesvaraya Technological University and M.Tech. in Computer Science and Engineering from JNTU. He has attended 3 conferences. He has 3 years of teaching experience. Presently working as Assistant Professor in Computer Science Department in K.I.T.E., Hyderabad, since 2012.

Karuna C. Gull has received the B.E. degree in Electronics and Communication from Karnataka University, India in the year 1996 and the M.Tech degree in Computer Science and Engineering from the Visvesvaraya Technological University, India in the year 2008. She has been working in the area of data mining and social networking since 2010. She has published 10 International Journal papers on Analysis of Social Network Sites using Mining Techniques, 2 International papers on Image Processing. She has also presented papers in 6 National and 6 International Conferences. She worked as a Lecturer, Senior Lecturer and assistant professor for about 13 years. She is currently working as an Assistant Professor in Computer Science Department in K.L.E.IT, Hubli, India since 2011.