A Survey Paper on Social Media E-Customers ‘Behavior Mining

Ms. Rekha Dimke[1], Ms. Reshma Lakhani[2]
[1, 2] Department of Computer Engineering
[1, 2] Parul University Waghodia, India

Abstract- *As nowadays, Social media has become an important source of information, where people can express their opinions and their views.To analyze social media data of e-customer’s there isneed to extract information for making predictions of how they behaves duringshopping in an e-commerce market place. After collecting data from an e-commerce market, performed a data mining application for extractingabout how customers’ behaves online whether to buy product or not. The model which is presented predicts whether customers will not or will buy their items or products added to shopping baskets on a market place. As there is massive growth of online social networks (OSN) like Twitter, Facebook and other social networking portals have created a need to determine people’s opinion and moods. Posting user feedback on products has become increasingly popular for people to express their opinions toward products and services. The companies think that there is a chance for an improvement in market for a product to people to aware and feel about it. In this study, there is use of sentiment dictionary as an Affine for tokenization and preprocessing process and also there is use of machine learning techniques to find about e-commerce site that is more useful and good for e-customers by predicting and analyzing through their reviews.*

Keywords— Twitter Data, Sentiment dictionary,Social media,Naïve Bayes, Support Vector Machine, Artificial Neural Network.

I. INTRODUCTION

E-commerce means to sell and to buy goods and services, or transmission of data or funds, through an electronic network, such as an Internet. Social media is a promising link which helps to build connection on social networks, personal information channels and mass information. Nowadays, Social media plays an important and precious role for information, where people can express their views.

When these opinions are related to company, accurate analysis can provide them with information such as quality of products and hinders that affects other customer decisions, feedback which are given earlier for launching products,trends, news of companies and also knowledge about other company which are in competition. The massive growth of online social networks (OSN) like Twitter, Facebook and other social networking portals have created a need to determine people’s opinion and moods.

The advantage of an electronic market place is to offer many choices, low price, easy way to search to access customers online.Thus Internet market share is important each and everyday[2].Opinion Mining is also known as Sentiment Analysis[1]. The company think that there is a way for improvement of product in market if they are aware of how people feel about specific product.Twitter data is use for studying, analyzing data.Using data mining methods such as Support Vector Machine and ANN it can examine about customer behaviour. Classification is an important in data mining. Data mining is an area which is included in machine learning,

With the rapid expansion of E-Commerce and social networking sites, so there is huge amount of information available in social media. Views of various products which are expressed in OSN’s plays an important role in market business analysis. In this paper, using machine learning technique with Semantic analysis which is used to classify the people’s opinions and views based on twitter data. Support Vector Machine is used with unigram model and Negation model for giving better performance than using it alone. Emoticons and affine dictionary is used so that it will be decided which E-commerce site is more useful and good for customer based on reviews.

II. RELATED WORK

Two types of approachesare taken towards sentiment analysis the affine and techniques of machine learning .Related work which is present in this section that helps us in getting a better understanding of Sentiment analytics.There are manyways for handling data. There are many sentiment analysis algorithms, tools which are still in research.For 100% accuracy there is no algorithm presented.

Apache Hadoop tool or platform is used to handle the problem for opinion mining in stream of data from social networking site. Hadoop was chosen to deal with the large amount of data from Twitter. Hadoop is designed in such a way so that it can convert unstructured data into structured data form because twitter data are unstructured and it is in JSON format [1].

One of the most important topic in sentiment analysisis classification of sentiment which classifies the viewsthat are expressed in a document, a sentence is positive, negative or neutral. Sentiment classification and Text classification are very different from each other. Sentiment classification use different types of unsupervised, supervised classification. Thus, these methods are used to classify the text into negative and positive. An example of an unsupervised classification is Lexicon classifier because it can function without any reference corpus and doesn't require any training. Lexicon based dictionary is based on list of words [7].

For this paper sentiment analysis has been proved the first encouragement for text classification into negative, positive and neutral.To analyze emotion, propermachine learning algorithms such as Naïve Bayes and Support Vector Machine are used which is the basis of classification [9].

In this Paper, classification is used to classify between positive and negative sentences. Information that are extracted from the Web and label the set of word which requires a lot of unrequired effort. SVM is also used for sentiment analysis of document level, which is used to extract the whole document polarity of each word by using sentiment dictionary such as affine. In this work,word polarity is being calculated ofthe sentence, which can either be positive or negative depending on the related sentence structure. Lakshmi and Edward have proposed to clean the data for the improvement of the quality of thetext or sentence [6].

It involves intelligent opinion mining .It also contains a method which allowsclassifyingopinion and sentiment score classes. The hybrid neural network contain of a modified probabilistic, neural network combine with a single layer classifier. The inputs of the network comprise binary images of potential opinion samplers. The consecutive bits represent evaluative words which were scored on a scale of intensity in an appropriate manner. The output provides the detection of potential opinions and the classification of opinion type classes[11].The method uses modified hybrid multilayer neural networks to recognize whether it is opinion type for finding its score. The network is a pattern classifier.

Kushal Dave in 2003, proposed feature extraction based on scoring approach for product reviews. The method finds the opinion of review as positive or negative and training is done using Machine Learning methods. The approach classifies opinion and then words are classified, finally opinions are classified [4].

III. METHODOLOGY

Affine based approaches are quite popular. These approaches involve tokenization means breaking sentence into words of a particular text or sentence into unigrams and also negation model is used to find polarity score of each word of text, based on dictionary.The average value of these scores determines the sentiment of the text or sentence whether text is negative, neutral or positive based on aggregated and averaging score of text.

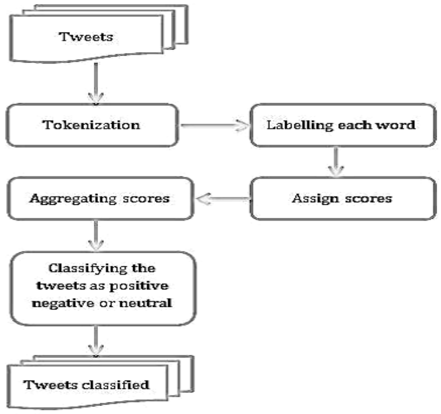

Figure 1:Flow chart of general approach

There are five steps which include:

- Tweets from Twitter API
- Data Cleaning
- Extracting Features
- Classifying based on algorithms
- Prediction of Best E-Commerce site from tweets

A. Tweets from Twitter API

There is Twitter Search API which is used to get tweets online. Firstly there is need to register on Twitter API online then they will provide us with token access key, consumer ID, Owner Id. By using theses Id user can get real time tweets that are available online. The input is given to analyzer is a user entered keyword based on which recent tweets can get from Twitter from Search API. Each request will return up to more than 100 tweets, gathered our dataset from the Twitter API and making use of word based on occurrence of the word are querying the recent tweets.

B. Data Cleaning

As Twitter data is unstructured data so, there is need to process before it come to use. So the tweets which are obtained are needed to clean for removing unwanted discrepancies so that it will give only information that is required for determining the emotional expression. This makes data very easy way for processing in further steps.

The procedure for pre-processing consists of the following steps:

- Removing non-English Tweets.
- Tweets should be converted in to lower case.
- URLs should remove – erased all string that describes links or hyperlinks from tweets.
- Replacing any usernames present in tweets to @username – remove the username because they are not considers in sentiments.
- Hash tags are required to convert into words thare normal because hash tags can provide some helpful information, so it is useful for replacing them literally in same word without usinghash. E.g. #Happy replaced with Happy.
- Removing any extra spaces and unrequired characters etc.
- Remove all the number from tweets and also remove words which are not started with an alphabet, for example 9th, 9:15am.
- At the beginning and at the end of the tweet there is need to remove punctuation like question marks, single quotes, double quotes commas, etc.

E.g. Happy!!!!!! Replaced with Happy.

- Two or more letters should be replacedwhich are repeated in a tweet by two letters of the same in tweets, sometimes users repeat letters to stress the emotion or feelings. E.g. Happppyy, Happyyyyyyyy for 'Happy'. Here in above example it is shown the repeated letters are replaced with 2 which are same.

C. Extracting Feature

Extracting feature playsan important role which is responsible to find performance in accuracy of system.

The input given is tweet which is cleaned that is filtered firstly by using steps given below:

- Polarity Score of the Tweet
- All stop words like is, a, the, an should be remove etc.
- Replace the emoticons with similar mining of word i.e. □ with happy.
- Use of Unigram Model: In the feature extraction method, extracts adjective from the dataset. After that, this adjective is used as negative and positive polarity score in text or sentence which is useful for determining the opinion and views by using unigram model. This model is used for extracting adjective and separates it. It ignores successive and preceding word that occurs with adjective in text or sentence. For eg, “Driving Happy” using unigram model, only word Happy from sentence is extracted.

- Use of Negation model: In negation model, the negative word change orientation of opinion.

For e.g., not happy will become sad.

D. Classification Algorithm:

There are two supervised classifier, they are Support Vector Machines and Naive Bayesthat model the probability of an input in a particular class that predict the emotion such as Sad, Disgust, Anger, Fear, Surprise, and Happy. A classifier is a learning model with associated learning algorithms that is used to recognize and analyze data and which can be used for classification. From these two methods, user can use any one method to analyze emotion. Both of these methods have been trained with a predefined dataset for finding perfect performance. For every tweet that is train, takes as input for feature extraction from tweets and then return the class of tweet as there is one out of six universal emotional expressions.

E. Machine Learning algorithms:

1. Support Vector Machines Classifiers (SVM):

SVM is use to determine linear separator in space which is helpful for separating different classes. In figure 2 there are 2 classes x, o and there are 3 Hyperplanes A, B and C. Hyperplane is best linear line that separates between the classes, because the distance which is normal of any of the data points which is large, so plane A show the margin which is maximum of separation [6].

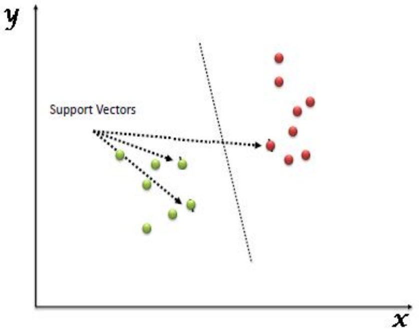

Figure 2: SVM

The main role of SVM is use for determining linear separators in space that is use to separate two different classes.Text data are sparse in nature so it is suited for classification of SVM, but they are correlated with one another and are organized into linearly separable class.

SVM has been used successfully in many real-world problems

a) Text categorization.
b) Image classification.
c) Bio informatics.
d) Recognition of Hand writing.

2. Artificial Neural Network (ANN):

An artificial neural network is collection of many neurons that are artificial whichare linked together based on specific network architecture. The main aim of neural network is to convert the inputs into meaningful and standard outputs.An artificial neural network (ANN), is also known as neural network (NN), is a computational model or mathematical model that is inspired by the structure or function of biological neural networks [11].Neural networks have performed successfully where other methods have not, predicting system behaviour, recognizing and matching complex and incomplete data. Ann is used for pattern recognition, interpretation, prediction, diagnosis.

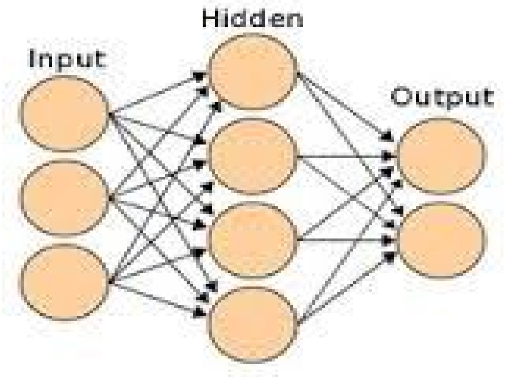

Figure 3: ANN

3. Naïve Bayes(NB):

The Naive Bayes classifier is themost commonly and easy used type of classifier. Naive Bayes classification model computes the posterior probability of a class, based on the distribution of the words in the document. The model works with the BOWs feature extraction which ignores the position of the word in the document. It uses Bayes Theorem to predict the probability that a given feature set belongs to a particular label[9].

Formula:

$$P(A/B) = \frac{P(B/A)P(A)}{P\,(B)}$$

IV. COMPARATIVE ANALYSIS

Method	Application	Advantage	Limitation
Naïve Bayes	Text classification, Spam Filtering, Medical diagnostic	Fast to train Fast to classify Handles streaming data well	Strong feature independence assumption
ANN	Robotics, Spam filtering, Speech recognition	Ann can perform tasks which linear program cannot.	They do not classify and cluster data.you need a lot of chips and a distributed run-time to train on very large datasets
SVM	Text categorization, Hand written recognition, Analyze data used for classification	SVM is less complex. Produce very accurate classifiers. Less over fitting, Robust to noise	SVM is binary classifier to do a multiclass classification

V. EXPECTED SOLUTION

As an E – Commerce industry is considered as an important part in progress point of view. Thus, data set in which tweets that are included has positive effect. So for advertisement of tweets from that data set is very helpful or useful to reflect views of people in understandable and proper way. It is useful to include method such as stemming for cleaning phase to include emotional expressions and considering emoticons that are replaced by words in the analysis phase should be improve so that performance and accuracy of result should be more.

VI. CONCLUSION& FUTURE WORK

In this paper, making use of ML technique with semantic analysis is use to analyze peoples opinion, their views based on twitter data that are real time. Support Vector Machine is used with unigram model and Negation model which will give better performance than using it alone. It will be decided that from customer online reviews that which E-Commerce site is more useful and good for them.The future work will be for improvement of the score dictionary with lexicon and affine combine approach as well as tokenization with large data sets and emoticons features should be included in future.

REFERENCES

[1] LokmanyathilakGovindan Shankar Selvan , Tang-Sheng Moh "A Framework for Fast-Feedback Opinion Mining on Twitter Data Streams", IEEE 2015

[2] GökhanSolahtarolu "Analysis and Prediction of E-Customers' Behavior by Mining Clickstream Data", IEEE 2015

[3] G. Vinodhini, R M Chandra sekaran "Opinion mining using principal component analysis based ensemble model for e-commerce application", Springer 2014

[4] K Unnumalia "Analysis of product using web". Elsevier 2012

[5] Neethu M S ,Rajasree R "Sentiment Analysis in Twitter using Machine Learning Techniques", IEEE 2013.

[6] DivakarYadav, GeetikaGautam"Sentiment Analysis of Twitter Data Using ML Approaches and Semantic Analysis",IEEE 2014

[7] Xujuan Zhou, Xiahui Tao, Jianming, Zhemyu Yang "Sentiment Analysis on Tweets for Social Events", IEEE 2013

[8] Jun Yang, Lan Jiang, ChongJun Wan and , JunyuanXie "Multi-Label Emotion Classification for Tweets in Weibo: Chinese site",IEEE 2014

[9] Uma Nagarsekar, PriyankaKulkarni "Emotion Detection from "The SMS of the Internet" IEEE 2013

[10] HaumaIsah, Paul Trundle, DaneilNeagu "Social media analytics for product safety Using text mining and sentiment analysis" IEEE 2014

[11] Koith Douglas Stuart and Macioj "Intelligent Opinion Mining and Sentiment Analysis Using ANN" Springer 2015