Consumer Behavior Prediction through Sentiment Analysis of Web-Data Source

Ms. Savita M. Agewadi[1]*, Prof. R. A. Medar[2]

[1]KLS, GIT/Department of Computer Science & Engg., Belagavi, India

[2]KLS, GIT/Department of Computer Science & Engg., Belagavi, India

Abstract – Sentiment analysis is an emerging subject area, with the availability of wide variety of web data. This web data contains opinions, sentiments and emotions expressed by people and this can be quantified through use of machine learning techniques and data-mining approaches. In this paper, the fundamental steps involved in sentiment analysis and consumer behavior prediction are presented with emphasis on estimating or classifying the level of people's opinions, sentiments towards a subject, product, service or individuals. Sentiment analysis involves data gathering, text pre-processing, feature extraction, sentiment classification and determining the polarity. Two major classification techniques namely, Baye's Classifier and Support Vector Machine based classifier are presented along with their relative merits and demerits. This information extraction is of immense value to companies, marketing teams, sociologists and psychologists who are concerned with opinions, views, and public mood. Sentiment analysis has tremendous potential to change the way the business processes are carried out with an aim of enhanced customer satisfaction and business profitability.

Index Terms- Baye's classifier, Machine learning, Sentiment analysis, Support Vector Machine

---------------------------♦-----------------------------

I. INTRODUCTION

Developments in Internet access have provided an opportunity to vast population on this globe to access Internet easily. Now a day, it has become affordable to access Internet from mobile phones along with regular devices like home PC & laptops. With these improvements in recent years, have witnessed the rapidly expanding e-commerce. A recent study from ComScore reports that online retail spending reached $37.5 billion in Q2 2011. In the U S. Millions of products from various merchants have been offered online. For example, Amazon.com, Shoppers.com offers millions of products online.

Most retail Websites encourages consumers to write reviews to express their opinions on various aspects of the products. Besides the retail Websites, many forum Websites also provide a platform for consumers to post reviews on millions of products. For example, CNet.com involves more than seven million product reviews; whereas Pricegrabber.com contains millions of review on more than 32 million products in 20 distinct categories over 11,000 merchants.

Social media is also providing limitless opportunities for consumers to discuss their experiences with products and companies to receive feedback on their products and services. For example, pharmaceutical companies are prioritizing social network monitoring within their IT departments, creating an opportunity for rapid dissemination and feedback of products and services to optimize and enhance delivery. The twitter platform provides people to express their views and opinions on any topic of their interest like sports, politics, economics, philanthropy etc.

Such numerous consumer reviews contain rich and valuable knowledge and have become important resources for both consumers and business firms. Consumers commonly seek quality information from online reviews prior to purchasing a product, while many firms use online reviews as important feedbacks in their product development, marketing and consumer relationship management.

2. RELATED WORK

There are many applications and enhancements on Sentiment Analysis algorithms that have been proposed in the last few years. Ref. [1] Walaa Medhat, Ahmed Hassan and Hoda Korashy have done a survey on Sentiment analysis algorithms and applications. This survey aims to give a closer look on these enhancements and to summarize and categorize some articles presented in this field according to the various SA techniques. The authors have collected fifty-four articles which presented important enhancements to the SA field lately. This survey can be useful for new comer researchers in this field as it covers the most famous SA techniques and applications in one research paper.

Most work on Sentiment Analysis has been done at the document level, for example, distinguishing positive from negative review. Ref. [2] Early study as the work of Hearst proposes a metaphoric model to determine the directionality of texts. This approach requires a manually constructed lexicon to derive such directionality information. Recently machine learning methods and information retrieval techniques have been employed to address this problem. Ref. [3] Bo Pang investigates several supervised machine learning methods to semantically classify movie reviews.

Mining online opinion is a form of sentiment analysis that is treated as a difficult text classification task. Ref. [4] Yong Shi and Emma Haddi have explored the role of text-preprocessing an sentiment analysis, and report on experiment results that demonstrate that with appropriate feature selection and representation, sentiment analysis accuracies using support vector machines in this area may be significantly improved.

As recent studies reveal, programmer’s emotions have an impact on their performance in the development process. Thus, in Wrobel, the data collected by an online survey of developers shows that emotions like anger have a positive effect on the developers’ productivity and that other negative emotions, like frustration and disgust, must be taken into account from the point of view of the risk. Ref. [5] In the field of Software Engineering, one of the few works that one can find are performed by and Guzman and Bruegge. They propose to improve emotional awareness in software development teams by processing textual collaboration artifacts, such as emails, wikis, commit messages in software repositories, bug reports etc.

Ref. [6] Francisco Jurado and Pilar Rodriguez propose in their article, the introduction of Sentiment Analysis techniques in order to identify and monitor the underlying sentiments in the text written by developers in issues and tickets. They have conducted an exploratory case study analyzing polarity and emotional clues in development issues from nine well-known projects that are freely available. Their results show that although both polarity and emotional analysis are applicable, the emotional analysis looks to be more suitable to this kind of corpus. The developers leave underlying sentiments in the text and that information could be mentioned as any other feature in the development process.

As the Internet is providing the opportunity for investors to post online opinions that they share with fellow investors. There have been a number of studies showing that the sentiment contained in these messages has been correlated with stock prices. Ref. [7] David L Olson and Desheng Dash they have developed a novel decision-support system using sentiment analysis, support vector machine and generalized autoregressive conditional heteroskedasticity modeling. Sina Finance, a widely used Chinese financial website has been selected as an experimental platform where corpuses of financial review data were collected. The intent is to obtain insights on how forecasting accuracy varies across alternative tools under different levels of volatility, providing evidence of how investment decisions might be made under differing conditions.

Intelligently extracting knowledge from social media has recently attracted great interest from the Biomedical and Health Informatics community to simultaneously improve health care outcomes and reduce costs using consumer-generated opinion. Ref. [8] Bjorn –Erik has done two stage study. In the first stage, he employs exploratory analysis using the self-organizing maps to assess correlations between user posts and positive or negative opinion on the drug. In the second stage he models the users and their posts using a network based approach.

3. FUNDAMENTAL CONCEPT

Generally speaking Sentiment Analysis aims to determine the attitude of the speaker with respect to particular topic. A basic task in sentiment analysis is classifying the polarity of a given text at the document level, sentence level or feature or aspect level-whether the given document, sentence or aspect is positive, negative or neutral.

Fig.1 Sentiment analysis involves steps like data gathering, text pre-processing, feature selection, sentiment classification and determining the polarity as shown below.

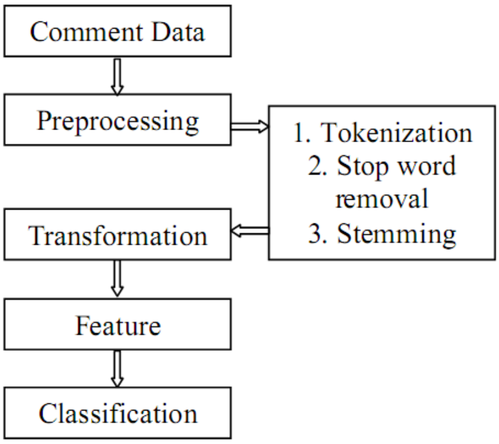

Fig. 1 Steps and techniques involved in Sentiment classification

A. Text Pre-Processing

Text pre-processing is the phase where data is cleaned and prepared for classification. Generally the data will be gathered from online text. This online text will be having lots of noise & HTML tags, scripts & advertisements which are uninformative. In addition, on words level, many words in the text do not have an impact on the general orientation of it.

The whole process involves several steps: online text cleaning, white space removal, expanding abbreviation, stemming, stop words removal, negation handling and finally feature selection. All of the steps but the last are called transformations, while the last step applying some functions to select the required patterns is called filtering.

B. Data Transformation

The text can be cleaned by removing HTML tags. Abbreviations can be expanded using pattern recognition and regular expression techniques. For stop words, a stop list can be constructed from several available standard stop lists, with some changes related to the specific characteristics of the data. For example: the words film, movie, actor, actress, scene are non-informative in movie reviews data. They can be considered as stop words because they are movie domain specific words.

Stemming is the process for reducing derived words to their stem, or root form. Stemming programs are commonly referred to as stemmers or stemming algorithms. A simple stemmer looks up the inflected form in a lookup table; this kind of approach is simple and fast. The disadvantage is that all inflected forms must be explicitly listed in table.eg. “developed”, “development” , ”developing” are reduced to the stem “develop”.

C. Filtering

For Filtering, one of the methods called Chi-square, which is the uni-variable method, can be used. It is statistical analysis method used in text categorization to measure the dependency between the words and the category of the document it is mentioned in. If the word is frequent in many categories, chi-squared value is low, while if the word is frequent in few categories then chi-squared value is high.

D. Feature Selection

Features in the context of Sentiment Analysis are the words, terms or phrases that strongly express the opinion as positive or negative. This means that they have a higher impact on the orientation of the text than other words in the same text. There are several methods that are used in feature selection, like syntactic based on the syntactic position of the word such as adjectives, and some are univariate, based on each feature’s relation to a specific category such as Chi squared (x2) and information gain, and some are multivariate using genetic algorithms and decision trees based on features subsets.

There are several ways to assess the importance of each feature by attaching a certain weight in the text. The most popular ones are: Feature Frequency (FF), Term Frequency Inverse Document Frequency (TF-IDF), and feature presence (FP). FF is the number of occurrences in the document. TF-IDF is given by

$$\text{TF-IDF}=\text{FF} * \log(N/DF) \qquad (1)$$

Where N indicates the number of documents, and DF is the number of documents that contains this feature. FF takes the value 0 or 1 based on the feature absent or presence in the document.

IV. SENTIMENT CLASSIFICATION TECHNIQUES

Fig. 2 Sentiment classification techniques can be roughly divided into Machine learning approach and lexicon approach as shown.

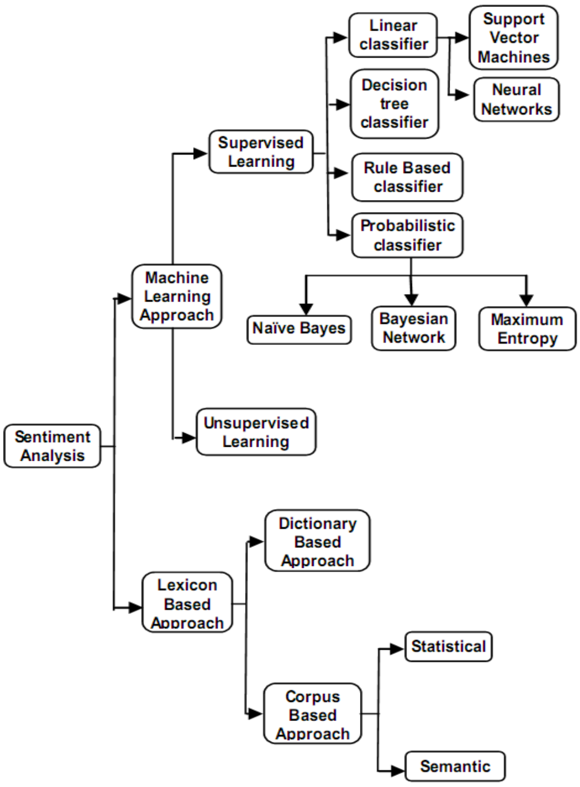

Fig. 2 Sentiment classification techniques

The Machine Learning Approach (ML) applies the famous ML algorithms and uses linguistic features. The Lexicon-based Approach relies on a sentiment lexicon, a collection of known and precompiled sentiment terms. It is divided into dictionary-based approach and corpus-based approach which use statistical or semantic methods to find sentiment polarity.

The text classification methods using ML approach can be roughly divided into supervised and unsupervised learning methods. The supervised methods make use of a large number of labeled training documents. The unsupervised methods are used when it is difficult to find these labeled training documents.

V. MACINE LEARNING APPROACHES

Machine learning approaches rely on the well known machine learning algorithms like Support Vector Machines, Bayesian Classifier etc. to solve the Sentiment Analysis as a regular text classification problem that makes use of syntactic and or linguistic features.

Text Classification Problem Definition: Let us say a set of training records D={X1,X2,…Xn} are chosen, where each record is labeled to a class. The classification model is related to the features in the underlying record to one of the class labels. Then for a given instance of unknown class, the model is used to predict a class label for it. The hard classification problem is when only one label is assigned to an instance. The soft classification problem is when a probabilistic value of labels is assigned to an instance.

VI. NAÏVE BAYES CLASSIFIER (NB)

The Naïve Bayes classifier is the simplest and most commonly used classifier. Naïve 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 Bag Of Words(BOW) 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.

$$\mathrm{P(label|features)} = \frac{\mathrm{P(label)*P(features|label)}}{\mathrm{P(feautres)}} \qquad (2)$$

P(label) is the prior probability of a label or the likelihood that a random feature set the label. P(features|label) is the prior probability that a given feature set is being classified as a label. P(features) is the prior probability that a given feature set is occurred. Given the Naïve assumption which states that all features are independent, the equation could be rewritten as follows:

$$\mathrm{P(label|features)} = \frac{\mathrm{P(label)*P(f1|label)*\ldots*P(fn|label)}}{\mathrm{P(feautres)}} \qquad (3)$$

Advantages and disadvantages of Naïve Bayes technique:

Computations are very simple in Naïve Bayes. It works on simple probability. And this method is very easy to implement. But it requires large number of training sets to achieve greater accuracy.

VII. SUPPORT VECTOR MACHINE CLASSIFIER (SVM)

The main principle of SVMs is to determine linear separators in the search space which can best separate the different classes. In Fig. 3 there are 2 classes x, o and there are 3 hyper-planes A,B and C. Hyper plane A provides the best separation between the classes, because the normal distance of any of the data points is the largest, so it represents the maximum margin of separation.

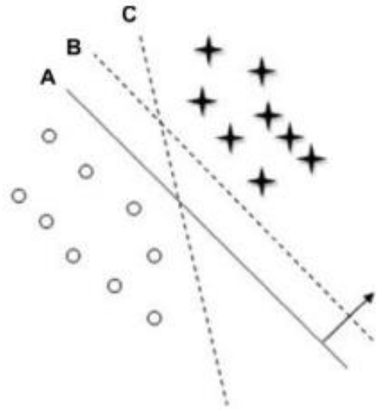

Fig. 3 Using support vector machine on a classification problem

Text data are ideally suited for SVM classification because of the sparse nature of text, in which few features are irrelevant, but they tend to be correlated with one another and SVM can construct a nonlinear decision surface in the original feature space by mapping the data instances non-linearly to an inner product space where the classes can be separated linearly with a hyper-plane.

Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. A schematic example is shown in Fig. 4 for illustration below. In this example, the objects belong either to class GREEN or RED. The separating line defines a boundary on the right side of which all objects are GREEN and to the left of which all objects are RED. Any new object (white circle) falling to the right is labeled, i.e., classified, as GREEN (or classified as RED should it fall to the left of the separating line).

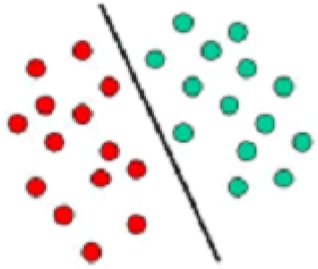

Fig. 4 Linear Classifier for linearly separable problem space

The above is a classic example of a linear classifier, i.e., a classifier that separates a set of objects into their respective groups (GREEN and RED in this case) with a line. Most classification tasks, however, are not that simple, and often more complex structures are needed in order to make an optimal separation, i.e., correctly classify new objects (test cases) on the basis of the examples that are available (train cases). This situation is depicted in fig. 5 for the illustration below. Compared to the previous schematic, it is clear that a full separation of the GREEN and RED objects would require a curve (which is more complex than a line). Classification tasks based on drawing separating lines to distinguish between objects of different class memberships are known as hyper plane classifiers. Support Vector Machines are particularly suited to handle such tasks.

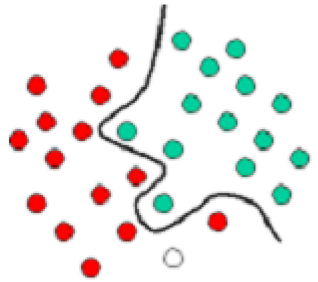

Fig. 5 Non-linear problem space and non-linear Classifier

The illustration in fig. 6 shows the basic idea behind Support Vector Machines. Here one can see the original objects (left side of the schematic) mapped, i.e., rearranged, using a set of mathematical functions, known as kernels. The process of rearranging the objects is known as mapping (transformation). Note that in this new setting, the mapped objects (right side of the schematic) is linearly separable and, thus, instead of constructing the complex curve (left schematic), all that one has to do is to find an optimal line that can separate the GREEN and the RED objects.

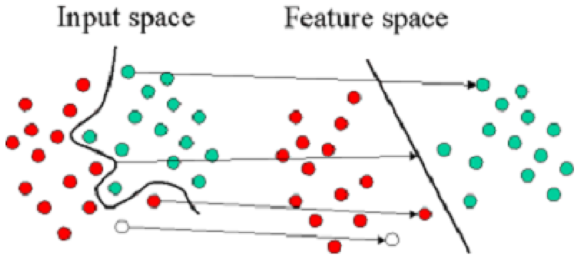

Fig. 6 Application of Kernelling function to transform non-linear to linear space

- **SVM Classifier Models**

Support Vector Machine (SVM) is primarily a classier method that performs classification tasks by constructing hyper planes in a multidimensional space that separates cases of different class labels. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables. For categorical variables a dummy variable is created with case values as either 0 or 1. Thus, a categorical dependent variable consisting of three levels, say (A, B, C), is represented by a set of three dummy variables:

A: {1 0 0}, B: {0 1 0}, C: {0 0 1}

To construct an optimal hyper plane, SVM employs an iterative training algorithm, which is used to minimize an error function. According to the form of the error function, SVM models can be classified into four distinct groups:

- Classification SVM Type 1 (also known as C-SVM classification)
- Classification SVM Type 2 (also known as nu-SVM classification)
- Regression SVM Type 1 (also known as epsilon-SVM regression)
- Regression SVM Type 2 (also known as nu-SVM regression)

VIII. COMPARISION AND EVALUATION

In this paper, two prominent sentiment classification techniques along with the basic steps of sentiment analysis based on feature analysis methods, namely Naïve Bayes and Support Vector Machine have been presented. The Bayes technique is based on very simple probability computations, but its accuracy is not as high as SVM. But for most practical applications where accuracy is not of great importance, Bayes technique is best suited. SVM on the other hand, has considerable computational overheads of first transforming a linearly non-separable domain to linearly separable space and then classifying the inputs. However, if the input is already linearly separable, then it performs the classification much faster. However, most of the practical application scenarios are linearly non-separable and hence, transformation is a must. Despite its high computational overhead, it has much higher accuracy as compared with results of other techniques. With the application of SVM technique the best results can be achieved [4] .

The advantages of SVM techniques are

- High Dimension input space- While text classification has to deal with many features may be more than 1000. Since SVM uses over fitting protection, which does not depend on number of features, so they have ability to handle large number of features.

- Document Vector Space – despite the high dimensionality of the representation, each of the vector contain only a few nonzero elements.

Thus for applications that require very high accuracy, a more computationally intensive SVM is preferred over Baye's classifier.

IX. CONCLUSION

In this paper, Sentiment Analysis concept and the steps involved in the analysis are introduced and its importance is brought out major application domains like product reviews, opinion mining etc. The details of the steps to be performed in Sentiment Analysis are given along with the basic types of feature classification. In Machine learning classification techniques, details of Naïve Bayes and Support Vector Machines types have been presented.

In the present Internet world, Sentiment plays an important role in several fields, like sentiments present in people's reviews about a product, about a company, health care facility, stock market, about a new movie. By carefully observing these reviews one can aim to a high degree of improvement in their particular field along with customer satisfaction. Practical applications of this new area of research include for instance, emotion aware robots, automatic movie classifiers, intelligent computer interfaces, next generation video games design and automatic marketing surveys. Looking at all these points, Sentiment Analysis is a promising field, which has a great scope in future.

REFERENCES

Walaa Medhat Ahmed Hassan Hoda Korashy, "Sentiment Analysis algorithms and applications: A survey," Elsevier, pp. 1093-1113, 2014.

Marti A Hearst, "Direction based Text Interpretation as an Information Access Refinement," 1992.

Bo Pang, "Opinion mining and sentiment analysis," 2008.

Emma Haddi Xiaohui Liu Yong Shi, "The Role of Text Pre-processing in Sentiment Analysis," Elsevier, pp. 26-32, 2013.

Guzman E Bruegge B, "Towards emotional awarness in software developments teams," in Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering, Saint Petersburg, 2013.

Francisco Jurado Pilar Rodriguez, "sentiment Analysis in monitoring software development process:An exploratory case study on GitHub's projrct issues," 2015.

Lijuan Zheng Desheng Dash wu David L Olson, "A decision Support Approach for Online Stock Forum Sentiment Analysis," IEEE Transactions on Systems Man and Cybernetics Systems, pp. 1077-1087, 2014.

Andrei Dragomir Altug Akay Bjorn-Erik, "Network-Based Modeling and Intelligent Data Mining of Social Media for Improving Care," IEEE Journal of Boimedical and Health Informatics, pp. 210-218, Jan 2015.

Ishan Arora Vivek narayanan, "Fast and Accurate Sentiment classification using an enhanced Naive Bayes model".

Indhuja K Reghu raj P C, "Fuzzy Logic Based Sentiment Analysis of Product Review Documents," in 2014 First International Conference on Computational Systems and communications, Trivandrum, Dec 2014, pp. 18-22.

Jose A Olivas Jesus Serrano-Guerrero, "Sentiment analysis: A review and comparitive analysis of web services," Elsevier, pp. 18-38, 2015.

Sentiment Analysis of Movie Review – V.K. Singh, R.Piryani, P.Walia (IEEE)

Issues of Social Data Analytics with a new method for sentiment analysis of social media data. (IEEE).

Feature Selection for Sentiment Analysis by using SVM –Rohini S. Rahate

Issues of Social Data Analytics with a new method for sentiment analysis of social media data. (IEEE)

Corresponding Author

Ms. Savita M. Agewadi*

KLS, GIT/Department of Computer Science & Engg., Belagavi, India

E-Mail – savitamb@gmail.com