An Evaluation of Opinion Mining and Sentiment Analysis

Dr.Md.Safdar

Dept. School of Information & Computer Sciences, Academic Counselor IGNOU, New Delhi, INDIA

***Abstract*- community being express themselves by giving opinions, feedback, suggestions or thoughts about any entity. Opinions can be expressed in many ways such as it can be expressed on twitter, face book, reviews, blogs etc. currently, if someone has to buy a online product then he/she can first view other buyer's reviews on the product site and take the right judgment accordingly. Opinion mining is a procedure of the mining opinion from the review. Analyzing customers review is more important for any user in making right purchasing decision product and organization. Opinion mining is also known as sentiment analysis. In this paper we survey on Opinion mining with respect to their different levels, architecture, techniques applied, tools used, comparative study of techniques and challenges**

Index Terms- Opinion Mining, Sentiment Analysis, Naive Bayes, SVM (Support Vector Machine), Spam Detection

I. INTRODUCTION

Opinions are statement that is a sign of people's perception or sentiment. These statements also provide opinion on objects or events. Opinion Mining or Sentiment analysis is a task under natural language processing for finding the mood of the customers about a purchasing of a particular product or subject. It involves building a system to collect and examine opinions about the product made in many online purchasing sites. Opinion mining is a sub field of web content mining. Web content mining is Since widespread of World Wide Web, internet and extensive growth of social media, organizations feel need to study public opinions for decision making. However to analyze polarity of opinions the exact intelligent information needs to be filtered. Hence automated opinion mining and sentiment analysis systems are needed [1]. Opinion mining techniques are used to extract reviews, opinions, political issues, brand perception automatically from web [2]. And sentiment analysis tracks, examines and evaluates public mood by using natural language processing [3]. Opinion mining and sentiment analysis can be used for business intelligence systems so as to analyze the opinions of public towards their brand and accordingly implement market strategies [4].

II. SENTIMENT CLASSIFICATION

2.1 Document level Sentiment Analysis: There are aims to categorize an opinion or a single review where a single topic is to be studied. The basic source of information is whole documented text [5]. The challenge is that all sentences in document may not contribute to opinion about a specific entity. Hence subjectivity/objectivity classification is very important, so as to discard irrelevant sentences or text. In this context, focus is on supervised learning methods which are used for document level classification. E.g.: Naïve Bayes classification and Support vector machines. The features that can be used for machine learning are individual words and frequency counts, parts of speech such as adjectives, opinion words to indicate polarity of sentiment (e.g.: good, wonderful are positive opinions and bad, poor, cheap, terrible are negative opinions), opinion phrases and idioms, negations and words dependency based features[6]. A combination of the above mentioned features and techniques based on polarity of words are used to further improve classification. Another interesting method used is domain adaptation, as sentiment analysis is highly sensitive to the domain from which data source is used. As language dependencies and their context are different from domain to domain, hence expressing opinions may also vary. Hence domain adaptation is very useful in document level sentiment analysis [6].

2.2 Sentence level Sentiment analysis: In this process, split of every sentence in document is determined. Document level classification methods can be applied to individual sentences. In this technique, we classify a sentence as subjective or objective, and resulting sentences are further classified as positive or negative opinions [5, 6]. This can be classified as: 1) Subjectivity classification 2) Objectivity classification. The above 2 subtasks are very useful as it filters out sentences which contains no opinions and classifies aspects, through which polarity of opinions is determined. Also sentence-level classification may not determine the exact opinion for complex and compound sentences, as it may contain multiple opinions. For e.g.: Sony mobile phone is very good and use easy to handle. Here there are 2 positive opinions, but sentence overall is positive. Also not all subjective sentences forms opinions, objective sentences can also imply opinions. Hence we have to focus on both sentence classifications.

2.3 Entity level Sentiment analysis: Document level and sentence level classification may not be useful in all applications, as it fails to review opinion about a specific entity. Aspect based Sentiment analysis uses a set of problems which follows natural language processing techniques and gives a better opinion set [6]. Hence, the context to which opinion is formed is extracted for every aspect in the sentence. However to solve for complex and compound sentences, lexicon-based approach is used which works as:

1. We mark all words and phrases containing aspects and assign score of +1 and -1 for positive and negative words respectively.
2. We classify or focus on those words that can change the orientation of a sentence. e.g.: Negation words: neither, never.
3. Handling contrary words like „but‟. E.g.: Phone x is great phone but Phone y is better in terms of processor performance.
4. Finally, after classifying opinions, we aggregate all similar opinions to determine final orientation of the sentence.

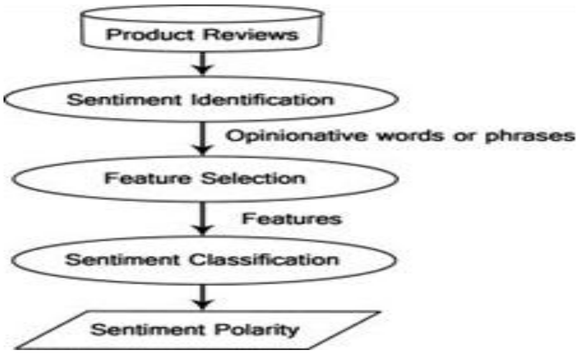

Fig: 1. Sentiment Classification Cycles

III. OPINION MINING

3.2 Opinion/Sentiment Types:
There are two main types:

3.2.1 Regular type: A regular opinion is often referred simply as an opinion in the literature and it has two subtypes.
3.2.1.1. Direct Opinion: A direct opinion denotes to an attitude articulated straight on an object or an object aspect. For example, “The battery life of this mobile phone is good.”
3.2.2.2. Indirect Opinion: It denotes to an opinion that is articulated indirectly on an object or an object Aspect. For example, “After taking this syrup, my body pains relieved”
3.2.2 Comparative type: A comparative opinion states a relation of similarities or differences between two or more entities. For example, the sentences, ”Boost tastes better than Horlicks” and ”Boost tastes the best” express two comparative opinions.

IV. OPINION MINING PROCESS

The data set can be collected from various social media network, website reviews http//www.amazon.com and it is processed to have precise oriented sets. Then we can apply classification techniques at document, sentence or aspect level as per requirements, to classify opinion as positive, negative or neutral. Here main focus is on supervised learning. For supervised learning method, Naïve Bayes, SVM can be used to classify positive, negative opinions [8, 9]. Using document level sentiment analysis, following techniques can be applied to implement the current features of the above analysis [7]:
1) N-gram model for extracting opinion words and phrases: The n-gram model is a contiguous sequence of n items from given data set [7]. Each sentence is spitted into words and its frequency is noted as tf (w, d)=|{w E d}| where tf(w,d) is number of times word occurs in text d. The term tp(w,d) only checks if word w is present, hence its function can be written as tp(w,d)=1 if w E d, else tp(w,d)=0. However if frequent words are uniformly distributed then its power will be low. The E.g.: The perfume smells good. For n=2 (bigrams) “mobile speaker”, “mobile speaker”, “very good”.

2) Frequency-inverse document frequency measure for listing most important words with respect to context of the text: the tf-idf is a statistic that specifies how frequently the word is used and hence states its importance in the given text [7]. The inverse document frequency is used to measure the rareness of the word in the given text. Hence the greater the value of tf-idf, the lesser the words importance. Idf (w,D) of word w in document D is idf(w,D) =|D|/log(df(w,D)). [7] 3) Part of speech Tagger for identifying opinions words in text: Parts of speech as noted earlier are a feature which can be easily used to indicate opinions [7]. Hence POS tagger helps us to identify POS and associate it to the word. E.g.: The design is exquisite. Here “exquisite” is an adjective which is a positive opinion. Hence in a given sentence, there are very few words which indicate sentiment. In English, POS are noun, verb, adjective, and adverb. However adjective are given major focus in identifying sentiments. Here focus is on supervised learning methods applied to sentiment classification - Naive Bayes classification and Support Vector machine. Using unigrams as classification features, the efficiency was to be good in Bayes and SVM.

1) Naive-Bayes Classifier

It is very popular algorithm as it is simple, efficient and shows better performance for real world problems. “Naive” assumes that features are fully independent. In spite of real world not following the above conditions, this algorithm solves the problems suited to normal distribution. This technique is supervised learning and statistical method. It assumes a probabilistic model and allows the capture of uncertain aspects in the text, by calculating probabilities of the outcomes [10]. Bayes theorem specifies mathematically the relation between probability of 2 events A and B. Let P(A) be conditional probability of event A conditioned by B and P(B) be conditional probability of event B conditioned by A. Bayes Formula [10] P(A/B)=P(B/A).P(A) / P(B) This formula helps us to find conditional probability of contrary events and independent probabilities of events. Hence we estimate probability of a document is positive or negative, or likelihood that an event will take place is positive or negative. So we estimate the probability of word with positive or negative meaning by analysing examples of positive and negative series using P(sentiment|sentence)= P(sentiment). P(sentence)/P(sentence) [10] Using above relation, we estimate P(word|sentiment) for all words as: P(word|sentiment)=(no. of words occurrences in a class +1)/(no. of words belonging to a class + total no. of words) [10] Hence using Naive Bayes we can find the polarity of a document by estimating the polarity associated with an opinion word. The major advantage is that we can train this model even by using a relatively small training set.

2) Support Vector Machine

SVM is a supervised classifier [10, 11] which exists in linear and non-linear forms. Using SVM, ideally datasets i.e. classes should be linearly separable. So that a line is found which divide the two classes perfectly in 2 regions? In real world problems classes cannot be perfectly linearly separable. Hence a function of higher order is applied which maps points in non-linear data to linear data. For example, consider an instance which belongs to either class employed or unemployed. There is a separating line which defines a boundary. At the right side of boundary all instances are employed and at the left side all instances are unemployed. For training data set D, a set of n points can be written as: D {(x , c) x R , c { 1,1}} ....(1) Where, xi is a p-dimensional real vector [10]. We find the maximum-margin hyper plane i.e. splits the points having ci = 1 from those having ci = -1. Any hyper plane can be written as the set of points satisfying [10]: w • x - b =1 ....(2) The distance between two hyper planes is w b and therefore w needs to be minimized. The minimized w in w, b subject to ci(w.xi − b) ≥ 1 for any i = 1… n SVM outperforms Naive Bayes by attaining maximum accuracy of approximately 80% using unigram data model.

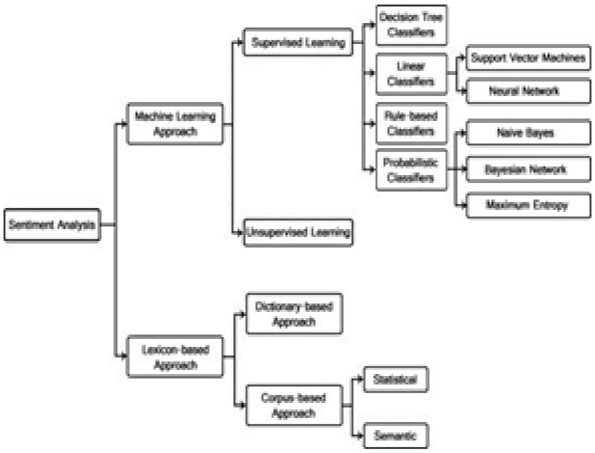

Fig: 2. Opinion Mining Process

V. OPINION SPAM DETECTION

5.1 Spam:

Unwanted, anonymous, commercial and mass email messages, called spam are viewed as a serious problem for Internet, content quality and trust [12]. Spam is flooding the Internet with many copies of the same messages to force the message on people who would not otherwise choose to receive it. Most spam is commercial advertising, often for dubious products, get-rich-quick schemes or quasi-legal services and customer review spam. Spamming is any deliberate action solely in order to boost a web page‘s position in search engine results, incommensurate with page‘s real value [13]. The term spamming (also, spam indexing) is used to refer to any deliberate human action that is meant to trigger an unjustifiably favorable relevance or importance for some web page, considering the page‘s true value. The adjective _spam’s used to mark all those web objects (page content items or links) that are the result of some form of spamming. People who perform spamming are called spammers [14].

5.2 Email Spam:

Email Spam is any email that was not requested by a user but was sent to that user and many others, typically (but not always) with malicious intent. The source and identity of the sender is anonymous and there is no option to cease receiving future emails.

5.3 Web Spam:

Web Spam is the web pages that are the result of spamming. Web spam is the deliberate manipulation of search engine indexes. It is one of the search engine optimization methods. Implementing web spam on a search engine reduces the redundant and non-desirable results [13, 14].

5.4 SMS Spam

SMS spam is any unwanted text message received on a mobile device. Like mail spam, SMS spam can range from unsolicited advertising to social engineering hoaxes to harmful attempts to steal subscriber’s personal and financial details.

5.5 Review Spam:

A review spam is considered to be totally unrelated, untrustworthy & untruthful user opinions on products and services. Such spam reviews are widespread on merchant‘s site and can be very harmful. For e.g a spam review that praises a product that every reviewer likes (gives a high rating) is not very damaging. However a spammer can carefully craft a review that criticizes a product that most people like can be very harmful affecting the customers buying decisions. Detecting untruthful opinion spam by manual reading is very hard, if not impossible, because a spammer can carefully craft a spam review to promote a target product or to damage the reputation of another product that is just like any other innocent review. Also the occurrence of the large number of duplicate and near-duplicate reviews written by the same reviewers on different products or by the different reviewers on the same products or different products are almost certainly considered as untruthful opinion spam reviews [15] .

VI. CONCLUSION

Sentiment analysis and Opinion mining are research area in machine learning problems that are silent moving ahead and improving on concert procedures. In this review, technique and approaches to sentiment analysis are studied. We have tried to underline the main aspect on which opinions are expressed. Since combined intelligence on area related to commerce, tourism, education, health, business and corporate world has its data all over the web, relevant solutions to solve this problem are becoming research interest. Also we have to find sentiment analysis is spam detection which is considered in this paper. On the other hand it still remains a challenge in machine learning field for the (NLP) natural language processing. In our future research more efficient for machine learning techniques.

REFERENCE

[1] Bo Pang and Lillian Lee, “Opinion Mining and Sentiment Analysis”, 2008

[2] Anand Mahendran and Anjali Duraiswany, “Opinion Mining for text classification”, International Journal of Scientific Engineering and Technology (2277-1581),Vol No.2,2013

[3] Deepali Virmani, Vikrant malhotra and Ridhi tyagi,” Sentiment Analysis Using Collaborated Opinion Mining”, 2014

[4] G.Vinodhini and RM.Chnadrasekaran, " Sentiment Analysis and Opinion Mining: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering (2277-128X), Vol No.2, 2012

[5] Raisa Varghese and Jayashree M, "A Survey on Sentiment Analysis and Opinion Mining ", International Journal of Research in Engineering and Technology(2319-1163)

[6] Bing Liu and Lie Zhang, "A Survey on Sentiment Analysis and Opinion Mining ", 2012 International Journal of Computer Applications (0975 – 8887) Volume 131 – No.1, December2015 27

[7] Murthy Canapathibholta and Bing Liu, "Minings in comparative sentences", 2008

[8] Pravesh Kumar Singh and Mohd Shahid Husain, "Methodological Study Of Opinion Mining And Sentiment Analysis Techniques", International Journal On Soft Computing (IJSC) Vol. 5,2014

[9] Ion Smeurean, "Applying Supervised Opinion Mining Techniues on Online User Reviews", Informatica Economics, Vol No. 16, 2012

[10] K. Saraswati and Dr. A. Tamilrasai, "Investigation of SVM for Opinion Mining", Journal of theoretical and Apllied Information Technology, Vol. 59 No. 2, 2014

[11] M. Rushdi Saleh, M.T. Martin-Valdiva, A.Montejo-Raez and L.A. Urena Lopez, "Experiments with SVM to classify opinions in different domains", Elsevier (14799- 14804), 2011

[12] Pedram Hayati, Vidyasagar Potdar Toward Spam 2.0: An Evaluation of Web 2.0 Anti-Spam Methods.

[13] Sumit Sahu, Bharti Dongre, Rajesh Vadhwani, ― Web Spam Detection Using Different Features ― in IJSCE, ISSN: 2231-2307 , Volume-1, Issue -3,July 2011.

[14] Z.Gyongyi & H. Garcia-Molina. ―Web Spam Texonomy‖.Technical Report, Stanford University, 2004.

[15] Arjun Mukherjee , Bing Liu , Junhui Wang, Natalie Glance, Nitin Jindal, ― Detecting Group Review Spam ― in poster March 28 –April 1,2011,Hyderabad, India.