Volume 4, Issue 11, November 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A survey: Sentiment Analysis of Online Review

Dharmesh Ramani*
Student, CSE Department,
PIET-GTU, India

Hazari Prasun
Asst. Professor, CSE Department,
PIET-GTU, India

***Abstract*— Sentiment Analysis is a new and emerging field of research which deals with information extraction and knowledge discovery from text using Natural Language Processing (NLP) and Data Mining (DM) technique, which help to track the mood of public about specific products and social or political event. Sentiments of individuals are extremely useful for people and company owner for making several decisions. It is intended to survey on sentiment analysis architecture, sentiment analysis type, level and task. This survey deals with machine learning methods that utilized for mining sentiment analysis and Opinion Mining.**

***Keywords*— Sentiment Analysis, Opinion Mining, Machine Learning Method, Sentiment Classification, Feature Extraction, Subjectivity Classification**

I. INTRODUCTION

Today, huge amounts of informal subjective text statements are accessible online with the growing availability of social networking websites, blogging and micro blogging sites. These statements are represented in several formats such as news articles, comments and review.

Sentiment Analysis (SA) has recently become the focus of many researchers due to its application and various fields. As it analyzes thought and ideas, feelings, attitude, and sentiment of individuals, analysis of this type of online text is helpful and demanded for marketing research, public opinion tracking, product auditing, business research, political surveys, client correspondence surveys, improving of web shopping bases, and so on.

Sentiment Analysis is the procedure, used for automatic extracting the polarity of public's subjective opinions from plain natural language text. Sentiment Analysis is likewise known as Opinion Mining (OM). Based upon opinion of others, one can make a good decision before acquiring any products or items. Sentiment Analysis has an extensive variety of use in e-commerce, which serves to figure out answer of several questions like, What do users think about our product, Which of our clients are unsatisfied, What features of our product are the worst, Who and how impacts our image, What is the public response to some event or some individual.

Opinion can be collected from any individual in the world about anything through review sites, blogs, web forums and discussion groups etc [1]. Organizations and product owners who expect to improve their products/services may strongly benefit from the rich feedback of users or customers. The most generally utilized sources for finding opinion are Blogs, review sites, raw dataset, and Micro-blogging web sites [8].

Online messages that are posted by individual in World Wide Web are mostly informal. Analysis or handling of this kind of text is often more difficult if compared with formal texts. The main difference between formal and informal text is in data preprocessing is formal text often require less preprocessing whereas informal text often contains emoticons, sarcasm, utilization of weak grammar, and non lexicon- standard words [9]. Therefore, extraction of informal content is regularly more troublesome.

People frequently ask their friends, relatives, and field specialists for suggestion during the decision-making procedure, and their opinions and perspectives are based on experiences and observations. One's point of view around a subject can either be positive or negative, which is known as the polarity detection of the sentiment. During sentiment analysis process, it requires very fast and concise information so individual can make quick and accurate decisions.

In sentiment analysis, the information or data collected from the reviews has been investigated mainly at three sentiment analysis level [2]:

A. Document Sentiment Level

The task at this level is to identify whether an entire sentiment document expresses a positive or negative sentiment. For example, given a product review, the system finds out whether the review expresses an overall positive or negative sentiment about any item or product. This task is usually known as document-level sentiment classification.

B. Sentence Sentiment Level

The task at this level goes to the sentences and figures out if each sentence expressed a positive, negative, or neutral sentiment. Neutral usually defines no opinion. This level of analysis is closely related to subjectivity classification, which recognizes sentences as objective sentences, that express factual information about the world and subjective sentences

that express some personal views, beliefs and feelings. This task of classifying whether a sentence is subjective or objective is known as subjectivity classification.

C. Entity and Aspect Sentiment Level

Above described both the document sentiment level and the sentence sentiment level do not analyse what exactly people liked and did not like. Aspect level helps to derive polarity (positive or negative) and a target of sentiment. A sentiment without its target being recognized is of restricted use. Finding out the target of sentiment helps to understand the sentiment analysis problem better.

For example, "although the camera quality is not too much great, I still love this mobile"

This statement is positive about the mobile (entity), but negative about its camera quality (aspect). In this way, the objective of this level of examination is to find sentiments on entities and/or their aspects. Aspect level was earlier known as feature level opinion mining.

In sentiment analysis, the sentiment mainly classified into three types as described below [2]:

A. Regular and Comparative Sentiment

A regular sentiment presents a sentiment only on a specific entity or an aspect of the entity, e.g., "Mango tastes great" which communicates a positive sentiment on the aspect taste of Mango.

It is referred to regularly as a sentiment in the literature and it has two fundamental sub-types:

1) Direct Sentiment: A direct sentiment refers to a sentiment expressed specifically on an entity or an entity aspect, e.g., "The battery life is good."

2) *Indirect Sentiment:* An indirect sentiment refers to a sentiment expressed indirectly on an entity or aspect of an entity based on its effects on some other entities. This sub-type frequently happens in the medicinal area, e.g., "After infusion of the medication, my joints felt more regrettable" describes an undesirable effect of the medication on "my joints", which in a roundabout way gives a negative feeling or opinion to the medication. In the case, the entity is the medication and the aspect is the impact on joints. Much of the current research focuses on direct opinions. They are less difficult to handle. Indirect opinions are regularly harder to manage.

A comparative sentiment communicates a connection of contrasts between two or more entities and/or an inclination of the opinion holder based on some shared aspects of the entities. For example, the sentences, "Mango tastes great than Grapes" and "Mango tastes the best" express two comparative opinions. A comparative opinion is generally preferred to utilizing the comparative or superlative form of an adjective or adverb.

B. Explicit and Implicit Sentiment

An explicit sentiment is a subjective statement that gives a regular or comparative sentiment, e.g., "Coke tastes great," and "Coke tastes better than Pepsi."

An implicit sentiment is an objective statement (usually expresses a desirable or undesirable fact) that implies a regular or comparative opinion, e.g., "India is at mars on his first attempt"

II. SENTIMENT ANALYSIS SYSTEM

Figure 1 shows the architecture of the sentiment analysis for extracting the sentiment scoring and decision making process from the online web document.

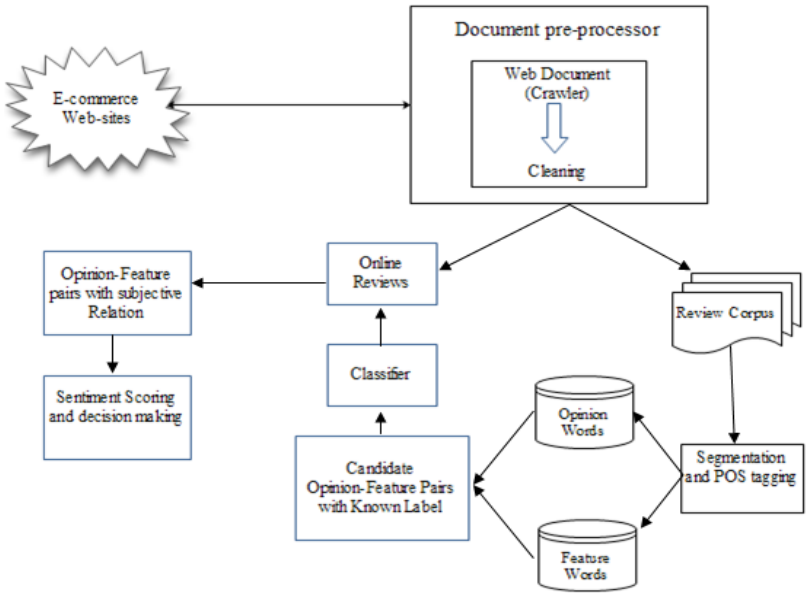

Fig. 1 Sentiment Analysis Architecture

Figure 1 shows the system architecture of sentiment analysis, several steps contains in this architecture are:

A. Document Pre-Processor

This task in this architecture is used to pre-process the review documents by determining relevant portion of a textual document. It includes Markup Language tag filter, that helps to divides an unstructured web document in record-size chunks and after that clean them by removing Machine Learning tags. Web crawler is used to collect web pages on some items or product reviews.

B. Retrieving Collection of Opinion-Feature pair Set

With the help of user generated thesaurus i.e. collection of word represent as dictionary, set of feature terms F and the set of opinion terms O are identified. At last by crossing join F and O, opinion-feature pairs are derived. For the construction of thesaurus of opinion terms, bag of words (BOW) are used to define positive and negative terms, and same as Noun, verb and noun phrases are required for the construction of thesaurus of feature terms. Punctuation utilized as a sign of an individual sentence.

C. Generate Classifier as identifier

Extracting the relative text information as a feature of opinion-feature pair to form training dataset, and develop classifier on training data set [6].

D. Analyse pair set with Subjective relation

Ensemble classifier is then used recognize the presence of subjective relationship in the candidate opinion-feature pairs to test the reviews [6].

E. Sentiment Scoring and decision making

In the last task in this architecture, opinion-feature pairs are used to derive the sentiment scoring for product features, which help user or customer to make a proper decision about some items or product [6].

The principle components of sentiment analysis issue are to identify the sentiment source, sentiment target, and the evaluative expressions or comments made by the opinion holder.

For the most part, an opinion is expressed by an individual person (opinion holder) who expresses a perspective (positive, negative, or neutral) about an entity (target object, e.g., person, item, association, occasion, service, etc.). A broad overview of the sentiment analysis issue [4] is presented in Figure 2.

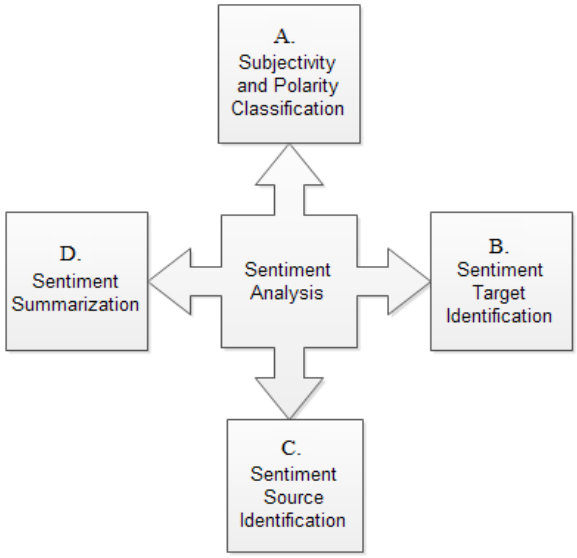

Fig. 1 Task of Sentiment Analysis

A. Subjectivity and polarity classification

The task of analyzing if the sentence is objective sentence or subjective sentence is known as subjectivity classification. In which the sentences known as objective sentences, that express factual information about the world and sentence is known as subjective sentences that express some personal views, beliefs, feelings and opinions.

Sentiment lexicons are created from tremendous corpuses using unsupervised techniques that are then applied for opinion hood determination. Opinion hood determination derives into two subtasks, i.e., subjectivity classification and sentiment polarity classification. Subjectivity classification techniques as described above while sentiment polarity classification techniques are used to classify opinionated terms into positive and negative statements. A few works utilize weighting procedures to recognize the quality of subjectivity, i.e., weak positive and strong positive or weak negative and strong negative.

Sentiment lexical resources play a key part in recognizing and assessing statements of opinion. Opinion lexical resources comprise of a set of two types of words, i.e., positive polar words and negative polar words.

Positive= (great, pleasant, superb, positive, fortunate, right, superior)

Negative= (terrible, dreadful, poor, negative, unfortunate, wrong, inferior)

B. Sentiment target identification

The sentiment target identification refers to the target of the sentence, e.g., individual, item, association, occasion, service, etc about which the sentiment is expressed. Sentiment targets at the sentence sentiment level or document sentiment level, the system should be able to have the capacity to recognize evaluative statements. A background study reveals that the process of opinion target extraction involves several Natural Language Processing (NLP) methods such as document pre-processing, Part-Of-Speech (POS) tagging, sentiment classification, and feature selection.

Regarding the automatic identification of sentiment targets, several approaches have been utilized. These approaches can be extensively separated into two major categories: Supervised approach and unsupervised approach. First, the supervised learning approaches are trained on manually labelled text. In this technique, a machine-learning model is trained on manually labelled data to classify and predict features in the reviews. It also provides a better result for future identification and requires manual work for planning of a training set.

Further for as the classifier to be trained on huge unigrams (feature set), it is very time consuming, memory complexity and sometimes, domain dependent. Additionally supervised techniques are linear classifier (Support Vector Machine, neural network), decision tree, Rule based classifier and probabilistic classifier (Naive Bayesian classifier, Maximum Entropy). Secondly, unsupervised learning approach are mainly based on Sentiment Lexicon (SL) and do not require labelled data. They automatically predict product features based on syntactic patterns and used to cluster the input data in class based on their property.

Some authors have also used the semi-supervised learning approach that combines both the labelled and unlabeled text to derive a classifier.

C. Sentiment source identification

An opinion holder or the source of an opinion is the individual person who expresses the opinion. The opinion holder or opinion source is important when validating the opinion as well as the quality, application and classification of the opinion, as the quality and dependability of a sentiment is significantly dependent on the source of that opinion. For example, a statement has a great strength if it submitted by the expert rather the ordinary person. For instance, a doctor's opinion to health and medical treatment while general public opinion to a political party. This process of identifying the holder of the opinion is problem of a natural language processing.

D. Sentiment summarization

Finally at the last task, the sentiment summarization derive the extraction and categorization of entity, aspect, opinion holder, time and aspect sentiment

Example:

Posted by: corafah Date: 21-NOV- 2013

(1) I purchased a Samsung mobile and my friends brought a soni mobile yesterday.
(2) In the previous week, we both utilized the mobile a lot.
(3) The photos quality from my Samy is not that extraordinary, and the battery life is short as well.
(4) My friend was extremely happy with his mobile and loves its photo quality.
(5) I need a mobile that can take great photos.
(6) I am going to return it tomorrow.

Among the several statement done by one user, entity extraction will retrieve as, "Samsung", "samy", and "soni" and "Samsung" and "samy" together as the same entity.
Aspect will retrieve as "photos quality" and "battery life"
Opinion holder is the person who post that sentence i.e. corafah and one more is corafah's friend.
Time extraction is the time when the post is posted i.e. Nov. 24, 2013
Now the aspect sentiment will derive as, statement 3 gives negative opinion to the photo quality and also negative opinion to the battery life, statement 4 gives the positive opinion about soni mobile and also for photo quality.

III. RELATED WORK

A lot of research has been carried out via researchers in the sentiment analysis area. Some of the methodologies utilized for sentiment classification are discussed here.

A. Naïve Bayes Approach

It is a simple and most usually used classifier model focused around bayes rule that computes post-prior probability of a class focused on distribution of words in documents and utilized for document classification. This methodology work with Bag of Words (BOW) feature extraction which ignore position of words in documents.
The classification approach can be combined with a decision rule, a common rule being, to pick the hypothesis that is most likely which is known as the greatest a posterior model or the MAP decision rule [7].

There are two first order probabilistic models for Naïve Bayes classification are Bernoulli model and the Multinomial model [7]. The Bernoulli model is a Bayesian Network with no word dependencies and binary word features; it likewise produces a Boolean indicator for each one term of the vocabulary relying upon its presence or absence; thus how, the Bernoulli model also takes words that do not appear in the document into account [7]. The

Multinomial model is a unigram language model with integer word counts and when the frequency of a word occurring in a document counts; so, a binarized version Of the Multinomial model is utilized which only takes in to account the presence of a word but not its frequency [7]. It is analyze that the multivariate Bernoulli performs well with small vocabulary sizes, however the multinomial model generally performs even better at larger vocabulary sizes, providing on an average 27% decrease in error over the multivariate Bernoulli model at any vocabulary size [7].

B. Maximum Entropy

Maximum entropy classification (MaxEnt, or ME) is a feature-based [5] probability distribution estimation model and an alternative technique which has proven effective in a number of natural language processing applications.

Principle of maximum entropy is if not much is known about the data, distribution should be as uniform as possible [7]. Importantly, unlike Naive Bayes, MaxEnt makes no assumptions about the relationships between features, and so might potentially performs better when conditional independence assumptions are not met [3]. This implies it should allow adding features like bigrams and phrases to MaxEnt without worrying about its feature overlapping [5]

C. Support Vector Machine (SVM)

Support Vector Machine (SVM) is another popular high margin statistical classification technique proposed for sentiment analysis and highly effective for text categorization [3].

The main idea underlying SVM for sentiment classification is to discover a hyper plane which separates the documents as per the sentiment, and the margin between the classes being as high as possible; it also focused around the Structural Risk Minimization principle [7].

Feature selection is an important task in machine learning methods; there are numerous features that must be considered for text classification, to avoid over fitting and to increase general accuracy [7]. SVM have the potential to handle large feature spaces with high number of measurements.

To deal with a large number of features, traditional text categorization methods assume that some of the features are unimportant, but even the lowest ranked features according to feature selection methods contain considerable information; considering these features as irrelevant often result in a loss of data [7]. Thus how the information loss can be reduced as SVMs does not requires at the time of making an assumption.

Though SVM outperforms all the traditional techniques for sentiment classification, it is a black box technique [7]. It is hard to research the way of classification and to distinguish which words are more important for classification. This is one of the drawbacks of utilizing SVM as a technique for document classification [7].

IV. CONCLUSIONS

It is concluded that above described Machine learning approaches works well for classifying sentiment analysis but among the several methodologies, Support Vector Machine provide high accuracy for sentiment classification. The purpose of using SVM is to extend the accuracy and enhance the performance of the sentiment analysis for better results. In near future enhancing the performance via improving the accuracy and solve the issue of providing significant resources for language detection other than English could be solved.

REFERENCES

[1] Khairullah Khan, Baharum B. Baharudin, Aurangzeb Khan, Fazal-e-Malik, "Mining Opinion from Text Documents: A Survey," 3rd IEEE International Conferences on Digital Ecosystem and Technology, 2009.

[2] Bing Liu, "Sentiment Analysis and Opinion Mining," Morgan & Claypool Publishers, May 2012.

[3] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan, "Thumbs up Sentiment Classification using Machine Learning technique," proceedings of EMNLP, pp. 79-86, 2002.

[4] Khairullah Khan, Baharum Baharudin, Aurnagzeb Khan, Ashraf Ullah, "Mining opinion components from unstructured reviews: A review," Journal of King Saud University – Computer and Information Sciences, 2014.

[5] Alec Go, Richa Bhayani, Lei Huang, "Twitter Sentiment Classification using Distant Supervision," CS224N project report, Stanford, pp. 1-12, 2009.

[6] Lijun Shi, Jing Zhang, Xuegang Hu, "Subjective Relation Identification in Chinese Opinion Mining Based on Sentential Features and Ensemble Classifier," Computer Science and Information Technology, vol. 8, 2010.

[7] Sagar Bhuta, Uehit Doshi, AvitDoshi, Meera Narvekar, "A Review of Techniques for Sentiment Analysis of Twitter Data," Issues and Challenges in Intelligent Computing Technique (ICICT), pp. 583-591, 2014.

[8] Blessy Selvam, S.Abirami, "A survey on Opinion Mining Framework," International Journal of Advanced Research in Computer and Communication Engineering vol. 2, 2013.

[9] Seyed-Ali Bahrainian, Andreas Dengel, "Sentiment Analysis using Sentiment Features," IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT), 2013.

