A Survey on Opinion Mining

R.Darshini[1], L.Hari Uthra[2], L. K.Megha[3], S. Prince Sahaya Brighty[4]
[1, 2, 3, 4] Department of Computer Science and Engineering
[1, 2, 3, 4] Sri Ramakrishna Engineering College,Coimbatore, India.

Abstract- *Now-a-days due to the rapid growth of internet, people are expressing their views and opinions on the web in large numbers. Opinion mining helps to acquire the public hotspots and trends by collecting and analyzing the massive data on the internet, which is used for public opinion analysis.The user-generated content present on different mediums such as internet forums, discussion groups, and blogs serves a concrete and substantial base for decision making in various fields such as advertising, political polls, scientific surveys, market prediction and business intelligence. They are effective in predicting the polarity of a comment once it is identified as a general opinion.*

Keywords- Opinion mining; Feature extraction; Opinion summarization, Supervised Learning, Unsupervised Learning, Semi-Supervised Learning.

I. INTRODUCTION

The technique of sentiment analysis and opinion mining summarizes the text with the help of machine learning process, statistical data, and the prior knowledge about the languages which extracts the needed data from the database. These types of work help to find the individual opinion from different people and their reviews of the particular information. Opinions from the people or the users play an efficient role in the prediction of results.

There is lots of advantage in this work which helps to collect the various resources from the web. The decisions are taken in a rapid manner for any problem by the government .The opinions and the sentiments of different users can be applied in many areas like polling in the elections, product review analysis, mining of information from the large databases and for social events.
As there are huge number of opinions and reviews published on internet it is difficult to analyze all opinions. All reviews are in the form of plain text which is written in any natural language, therefore to get valuable information from those reviews we need help from other domains like Natural Language Processing (NLP) and Data Mining.

The analysis of the open-answer part is not as easy, but it might contain opinions within and outside of the scope of the questionnaire, objective problem statements, and suggestions. Detection of opinions and their polarities and the domain topic classification are important aspects of the analysis of such open-ended user comments. Sentiment analysis is the process of recognizing and classifying different sentiments conveyed online by the individuals to derive the users view towards a specific product, topic or event is positive, negative or neutral. To analyze and summarize the opinions expressed on web manually is a difficult task. Therefore, we require an automated sentiment analysis system.

In this survey, we aim to review how different methods have been used to build summarization systems and perform reviews analysis.

II.OPINION MINING

Opinions are very important among all human activities as they are prime reflectors of our behaviors. People always tends to know other peoples' opinions, whenever they need to make some decision. Nowadays, businesses and organizations always looks for getting their customers' reviews about their goods and services[5]. Each and every user wants to know the opinion of various other customers on a product before buying it, while in political election, public opinion about a specific political leader is important before making a voting decision. An opinion is the personal view of an individual, it represents the individual's ideas, judgments, beliefs, assessments and evaluations about a specific topic, item or subject. One's opinion about a subject can either be positive or negative, which is referred as the semantic orientation or polarity or sentiment.

An opinion has three main elements, i.e.

- The opinion source: author of the review
- Target of the Opinion: object or its feature.
- Opinion polarity: positive or negative

All of these elements are vital for opinion identification. "Opinion mining is the problem of recognizing the expressed opinion on a particular subject and determining the polarity of opinion" [1]. It provides broad view of the sentiments expressed via text, and to classify and summarize the opinions, it enables further processing of the data.

A. Data sources

The data sources are mainly blogs, review sites and micro blogging.

a) Blogs

As the growth of the internet is increasing day by day, blog pages are also growing at a fast pace. A Blog is a web page where a person or group of users can give feedbacks, opinions, information, etc. on a regular basis [10]. These are written on many different subjects. Blog pages contain s personal opinions. People write about their perceptions or feelings they want to share with others on a blog. Blogging is a good thing because of its feature of simplicity, freely available in form and unedited nature. Many of these blogs contain reviews on many products, services, entities, issues, etc.

b) Review sites

There are millions of websites available where users can write reviews about any products, services, property, etc. Some websites like Amazon, Flipkart, Mantra, Snapdeal, 99acres.com can allow their users to give their reviews on the product page itself. These reviews affect the conclusion of the new user who is going to purchase any product. As new user can know the feedbacks of the previous buyer before purchasing any new product.

c) Micro-blogging

In Twitter information is represented as a short text message called "tweet". The opinions about different topics are expressed in tweets and they are considered for opinion mining [8].

B.Twitter Data Collection Methods

The three possible ways to collect Tweets for research are as follows [11]:

- Data repositories such as UCI, Friendster, Kdnuggets, and SNAP.
- APIs: Twitter provides two types of APIs such as search API and stream API. Search API is used to collect Twitter data on the basis of hashtags and stream API is used to stream real time data from Twitter.
- Automated tools that are further classified into premium tools such as Radian6 [18], Sysmos, Simplify360, Lithium and non-premium tools such as Keyhole, Topsy, Tagboard and Social Mention.

C.Data Preprocessing:

Mining of Twitter data is a challenging task. The collected data is raw data. In order to apply classifier, it is essential to pre-process or clean the raw data. The pre-processing task involves uniform casing, removal of hashtags and other Twitter notations (@, RT), emoticons, URLs, stop words, decompression of slang words and compression of elongated words. The following steps show the pre-processing procedure. Remove the Twitter notations such as hashtags (#), retweets (RT), and account Id (@). Remove the URLs, hyperlinks and emoticon. It is necessary to remove non letter data and symbols as we are dealing with only text data. Remove the stop words such as are, is, am etc. The stop words do not emphasize on any emotions, it is intended to remove them to compress the dataset. Compress the elongated words such as happyyy into happy. Decompress the slag words such as g8, f9. Generally slang words are adjectives or nouns and they contain the extreme level of sentiments. So it is necessary to decompress them.

D.Feature Extraction:

The pre-processed dataset has various discrete properties. In feature extraction methods, we extract different aspects such as adjectives, verbs and nouns and later these aspects are identified as positive or negative to detect the polarity of the whole sentence. Followings are the widely used Feature Extraction methods.

- Terms Frequency and Term Presence: These features denote individual and distinct words and their occurrence counts.
- Negative Phrases: The presence of negative words can change the meaning or orientation of the opinion. So it is evident to take negative word orientation in account.
- Parts Of Speech (POS): Finding nouns, verbs, adjectives etc. as they are significant gauges of opinions.

E.Techniques of opinion mining

There are three approaches for performing Opinion Mining [3].

A) Machine Learning Based Approach:

This includes:

1) Supervised Learning
2) Unsupervised Learning

3) Semi-Supervised Learning

This approach is comparatively more practical as compared to other approaches. This research is supported by the Fundamental Research Grant Scheme because of its automated implementation and ability to handle large collections of data on the Web.Machine Learning-Based methods can be classified into three types: supervised, unsupervised and semi-supervised learning methods.

1) Supervised Learning

Supervised learning is a mature and successful solution in traditional topical classification. It has been implemented to mine opinions and the results obtained are quite satisfactory. Important supervised classification algorithms include: Naïve-Bayes, a generative classifier that estimates prior probabilities of P(X|Y) and P(Y) from the training data and generates posterior probability of P(Y|X) based on these prior probabilities; Support Vector Machine (SVM), a discriminative classifier that does not make any prior assumptions as per the training data and directly estimates P(Y|X); and the lazy learning algorithm KNearest Neighbor (KNN), which does not require prior construction of a classification model. In both topical and opinion classification, Naïve Bayes and SVM are the most common and effective supervised learning algorithms [4]. The biggest disadvantage of supervised learning is that it is sensitive to the quantity and quality of training data and may fail when training data is biased or insufficient. Opinion mining at the sub-document level causes more problems in case of supervised learning based approaches because the amount of information available for the classifier is very less.

2) Unsupervised learning

In the process of text classification, many at times it becomes troublesome to construct labelled training documents. On the contrary, it is easy to construct unlabeled documents. Unsupervised learning methods serve as a solution to this problem. LDA and pLSA are examples unsupervised methods used to elicit latent topics in textual documents. These topics are nothing but features, and each and every feature is basically a distribution over terms. Accurately training a large amount of data is the prerequisite of unsupervised techniques. This is the biggest disadvantage in using unsupervised learning. Fully unsupervised models often produce incoherent topics because the objective functions of topic models do not always correlate well with human judgements. In spite of this disadvantage, unsupervised learning helps in providing a solution to acquire knowledge about the given data.

3) Semi-Supervised learning (SSL)

SSL is a relatively new machine learning approach. SSL models tries to overcome the drawbacks of both supervised and unsupervised methods. SSL learns from both labelled and unlabeled data, hence overcoming the major drawback of supervised learning which is learning only from labelled data. Hence, using SSL with unlabeled data specially when there is fixed amount of labelled data, can help to achieve improvement over supervised learning. Also, the constraints of unsupervised learning approaches are not present with SSL if we include some form of prior knowledge to unsupervised models.

B) Lexicon-based approach

Lexicon Based Approach Techniques based on Natural Language Processing (NLP) and Lexicon Based Approach utilize Parts Of Speech (POS) information and WordNet.This approach searches for the opinion lexicon which is used to analyze the text. It consists of two methods. They are:

1) Dictionary-based method: It finds opinion seed words and then searches the dictionary for their synonyms and antonyms [10].

2) Corpus-based method: It begins with a seed list of opinion words and then searches for other opinion words in a large corpus so as to find words with context specific orientations[3]. This can be done by using statistical or semantic methods.

C) Hybrid Approach

It is basically a combination of Machine Learning Based Approach and Lexicon Based Approach.

F. Background theory

Much research exists on sentiment analysis of user opinion data, which mainly judges the polarities of user reviews. In these studies, sentiment analysis is often conducted at one of the three levels: the document level, sentence level, or attribute level. There are many methods are used for sentiment analysis to find out positive, negative or neutral opinions. Following are useful methods for sentiment analysis. Machine Learning Methods: The machine learning approach applicable to sentiment analysis mostly belongs to supervised classification in general and text classification technique. In a machine learning based classification, two sets of documents are required: training and a test set. A training set is used by an automatic classifier to learn the

differentiating characteristics of documents, and a test set is used to validate the performance of the automatic classifier. A number of machine learning techniques have been adopted to classify the reviews. There are three different machine learning algorithms who achieved great success for text categorization. [2]

1) Naive Bayes: Naive Bayes model is a simplest model. For the categorization of the text Naive Bayes model works well. Naive Bayes classifiers assume that the effect of a variable value on a given class is independent of the values of other variable. This assumption is called class conditional independence. As in [6], it is made to simplify the computation and in this sense considered as "Naive". This classifier is used to find out the probability of the words.

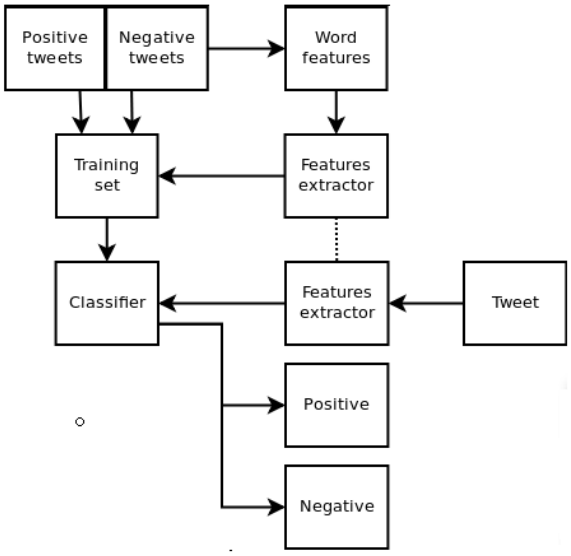

Figure 1: Naïve Bayes classification

2) Maximum Entropy (MaxEnt): This model is Feature based model. MaxEnt do not make any independence assumption for its features, therefore MaxEnt is different than Naive Bayes. MaxEnt can handle features overlapping problems better than Naïve Bayes. Stanford classifier is used for classification in MaxEnt model.

3) Support Vector Machines (SVMs): SVM is used for m statistical learning theory. The class of algorithms called SVMs which are used for pattern recognition. Support vector machines (SVM) are a group of supervised learning methods that can be applied to classification or regression. They can be defined as system which use hypothesis space of linear functions in a high dimensional feature space.

III.PAST RESEARCH WORK

Many algorithms have been proposed in order to understand and implement opinion mining and sentiment analysis. Researchers have developed models for identifying the polarity of words, sentences and whole document.Various tools are also available now for opinion extraction, sentiment analysis and opinion summarization. There have been researches regarding development for better algorithms for such tools.

A) Collaborated opinion value [11]

This paper proposes an algorithm for identifying the polarity of remarks. In the existing work polarity of remarks word by word in a sentence was not considered. The proposed work has been explained with help of a case study. A case has been considered wherein a set of teachers give their remarks about a particular student. The algorithm is applied on every remark to identify the polarity of each remark. The algorithm generates a numeric value for the opinion. If the opinion value are high the opinion are considered positive. Lower opinion value represents negative remarks.

B) Categorization and Summarization [12]

The proposed work consist of three stages to categorize and summarize the given input namely Token Creation, Feature Identification and Categorization and Summarization. It includes Text summarization using Rule reduction [12]. In this paper, for analyzing the text the rule reduction techniques had been used in three stages, Token creation, Feature Identification and categorization and summarization. It produces noteworthy results. Experiment validates the selection of parameter and efficiency of approach.

C) Automatic text summarization [13]

The proposed approach retrieves the important information from the text by performing semantic analysis of the text. In the approach Simplified Lesk Algorithm is used to extract the relevant sentences from a single-document text based on the semantic information of the sentence and WordNet is used as an online semantic dictionary. Another work is text summarization using WordNet [13] which includes unsupervised learning for summarizing the text. WordNet is an online semantic library. Simplified Lesk algorithm is used to evaluate the weights of all the sentences and arranges them in decreasing order of weights. Then according to the percentage of summarization the sentences are selected from the ordered list.

IV.CONCLUSION

In this paper, we have firstly presented the detailed procedure to carryout sentiment analysis process to classify highly unstructured data of Twitter into positive or negative categories. Secondly, we have discussed various techniques to carryout sentiment analysis on Twitter data including knowledge based technique and machine learning techniques.

Hence, the future opportunities in the domain of sentiment analysis include developing a technique to perform sentiment classification that can be applicable to any data regardless of domain. In addition, language diversity in social media data is a key issue which is required to be eliminated in future. Moreover, some of the more crucial challenges of Natural Language Processing (NLP) can also be used as further developments in sentiment analysis, such as hidden or veiled sentiment detection, comparison or association handling and emoticon detection.

REFERENCES

[1] Solanki Yogesh Ganeshbhai,& Bhumika K. Shah (2015, April) Feature Based Opinion Mining: A Survey. International Advance Computing Conference (IACC), 2015 International Conference on (pp. 919-923). IEEE.

[2] Kaijie Guo, Liang Shi , Weilong Ye,& Xiang Le.(2014). A Survey Of Internet Public Opinion Mining. International Conference on Progress in Informatics and Computing (pp.173-179). IEEE.

[3] Ananta Arora, Chinmay Patil, & Stevina Correia. (2015, November) .Opinion Mining An Overview. An International Journal Of Advanced Research in Computer and Communication Engineering Vol.4, Issue 11. IJARCCE.

[4] Mitali Desai, & Mayuri A. Mehta. (2016, ICCCA). Techniques for Sentimental Analysis Of Twitter Data: A Comprehensive Survey. International Conference On Computing, Communication and Automation (PP.149-154).ICCCA.

[5] Pankaj Gupta, Ritu Tiwari , & Nirmal Robert .(2016,April). Sentiment Analysis and Text Summarization Of Online Reviews: A Survey. In proceedings at International Conference On Communication and Signal Processing (ICCSP), April 6-8, 2016, India. IEEE.

[6] Evgeny A. Stepanov, & Giuseppe Riccardi .(2011,IEEE). Detecting General Opinions from Customer Surveys. 11th IEEE Conference On Data Mining Workshops, (PP.115-122) .IEEE.

[7] Apoorv Agarwal, Boyi Xie , Ilia Vosha , & Oven Rambow . Sentiment Analysis Of Twitter Data.

[8] Efthymios Kouloumpis, Theresa Wilson *, & Johanna Moore . (2011).Twitter Sentiment Analysis: The Good the bad and the OMG!. In proceedings of the Fifth International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media ,2011 AAAI.

[9] Zalak M. Patel, & Vishal P. Patel .(2015) .A Survey On Various Techniques Of Sentimental Analysis in Data Mining. In proceedings at International Journal Of Engineering Development and Research (2015 IJEDR) volume 3, Issue 4, ISSN: 2321-9939. IJEDR.

[10] Anu Maheshwari, Anjali Dadhich, & Dr. Pratistha Mathur (2015, January). Opinion Mining: A Survey International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 1, January 2015(pp.187-190) IJARCC.

[11] Deepali Virmani, Vikrant Malhotra, Ridhi Tyagi, Sentiment Analysis Using Collaborated Opinion Mining, Department of Information Technology, conference on knowledge discovery and data mining(pp.168-177).

[12] Devasena, C. L., & Hemalatha, M. (2012, March). Automatic text categorization and summarization using rule reduction. In Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on (pp. 594-598). IEEE.

[13] Pal, A. R., & Saha, D. (2014, February). An approach to automatic text summarization using WordNet. In Advance Computing Conference (IACC), 2014 IEEE International (pp. 1169-1173). IEEE.pp. 593601, 1977.