ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 1, January 2015

Opinion Mining: A Survey

Anu Maheshwari[1], Anjali Dadhich[2], Dr. Pratistha Mathur[3]

M.Tech Scholar, Banasthali University Jaipur, India[1]

Assistant Professor, Apex Institute of Management & Science, Jaipur ,India2

Associate Professor, Banasthali University Jaipur, India3

Abstract: In the last few years as the growth & use of Internet increases and share of user's opinions increases, the inspiration towards opinion mining also increases. When person wants to conduct some research or making day to day decisions, he has to depend on other people's opinions. We consult "political discussion forums when going to a political vote, read consumer reports when purchasing some products, ask friends to recommend a hotel for the stay". And now Internet has now made it possible to find out the opinions of millions of people on everything from latest gadgets to latest software. Opinion mining is a type of natural language processing for tracking the perceptions or sentiments of the public about a particular product. It can be useful in several ways in marketing. It helps to judge the success of a launch of new product. This paper gives a brief survey on the opinion mining**.**

Keywords: Opinion**,** Opinion Mining, Sentiment Classification

I. INTRODUCTION

Opinions are subjective statements reflecting people's sentiments or perceptions on entities or events. Opinion Mining or Sentiment Analysis is "the process of finding the views of the public about a product, policy or a topic. It involves building a system to collect and examine opinions about the product made in blog posts, comments, reviews or tweets[1]". It is one of the most active research areas in NLP and widely studied in Data Mining, Text Mining, Web Mining. It is a subfield of data mining. It is used to analyze the sentiments expressed by people on the web through reviews. In recent years, large attention has been given to opinion mining because of its wide range of possible applications [2,3]. As an example consumers look for the opinions of the products they want to buy before actually buying them [3]. Opinion mining is very useful to know to get the answers of these questions like **Which phone should I buy? Which movie should I watch? Which camera should I buy? Which novel can I purchase? In which College should I take admission?** So he or/she has no longer depends on the friends or relatives. In general, opinion mining helps to collect information about the positive and negative aspects of a particular topic. Finally, the positive and highly scored opinions obtained about a particular product are recommended to the user. In order to promote marketing, large companies and business people are making use of opinion mining [4].

II. BASIC TERMINOLOGY

The basic terminology of an opinion mining is given by Hu and Liu [8]. He introduces the basic components of an opinion which are:

- **Opinion:** It is a perception, feel, attitude or view of any product given by customer.
- **Opinion holder:** It is the person that gives a particular opinion on an entity like any product.
- **Object:** It is an entity or a thing on which an opinion is being given by customer.

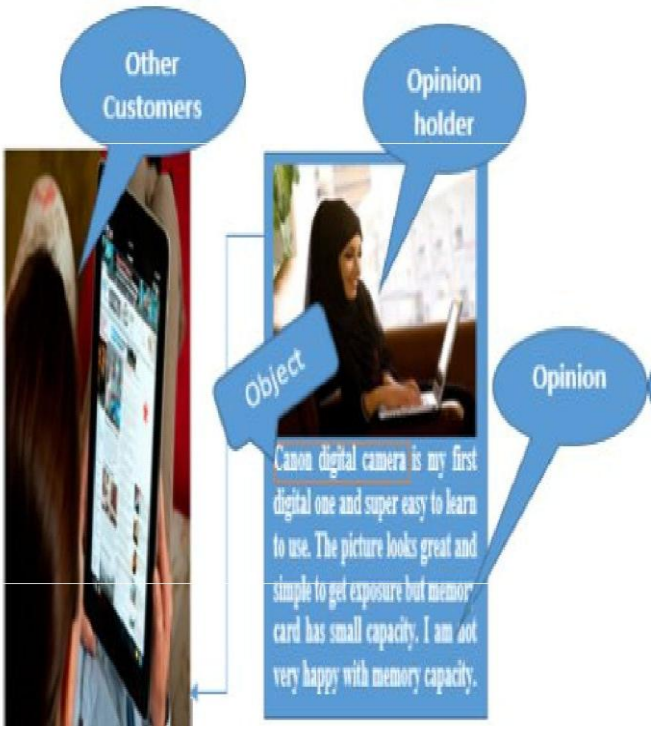

Fig1: Shows basic components of opinion mining

To mine opinions, the reviews collected can be analyzed at three levels.

- The first one is *Document-level*, where the opinion in the whole document is given a summary as affirmative, pessimistic or neutral.
- The second one is the *Phrase level* where the phrases in a sentence are classified according to polarity. In figure 1, the entire framework of opinion mining is represented.Table1 represents insight into opinion mining at different levels.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 1, January 2015

- The third one is the *Sentence level*, which targets the sentences in the document and categorizes it as objective sentences(no opinion) and subjective sentences (with opinion).

III.DATA SOURCES

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

3.1 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". These are written on many different subjects. Blog pages contain the expression of one'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.

3.2. 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 etc 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.

3.3. 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.

IV. SENTIMENT CLASSIFICATION

Sentiment Classification consists basically two approaches machine learning and lexicon based approaches.

4.1. Machine Learning based approaches

It is a supervised learning algorithm in which a system is capable of acquiring and integrating the knowledge automatically. It is all about learning structures from data.It involves text classification techniques. This approach treats the sentiment classification problem as a topic-based text classification problem (Liu, 2007). In this classification, documents require two sets: training set and the test set. For training a set "automatic classifiers are used that learns various characteristics of documents", and a "test set is used to validate the automatic classifier's performance". Some machine learning techniques like Naive Bayes (NB) [13], Support vector machines (SVM) [14] and Maximum entropy (ME) [12] have gained a great success in text category field.

4.1.1 *NAIVE BAYES(NB)*

The Naive Bayes algorithm is widely used algorithm for document classification[12].The basic idea is to estimate the probabilities of categories given a test document by using the joint probabilities of words and categories. The naive part of such a model is the assumption of word independence.

4.1.2 *MAXIMUM ENTROPY(ME)*

The Maximum Entropy classifier is a probabilistic classifier which belongs to the class of exponential models. The Maximum Entropy is based on the Principle of Maximum Entropy and from all the models that fit our training data, selects the one which has the largest entropy. The Maximum Entropy classifier can be used to solve a large variety of text classification problems such as language detection, topic classification, sentiment analysis, etc.

4.1.3 *SUPPORT VECTOR MACHINES(SVM)*

Support Vector Machines (SVMs) are supervised learning methods used for classification . These models are associated learning algorithms that "analyze data", "recognize patterns", used for classification and regression analysis. It is considered to be the best text classification method. "This model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible". New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. Some examples where this model is used are medical science, hand-written characters, pattern recognition, bioinformatics, signature/hand writing recognition, image and text classification, and e-mail spam categorization.

4.2.Lexicon Based Approaches

Lexicon is an important indicator of sentiments called opinion words. Here a list of phrases/words is used and called "sentiment lexicon". Words in a sentence express different opinion i.e affirmative or pessimistic. There are three following techniques used.

1. **The corpus-based techniques**:

Corpus techniques are basically designed to find "domain-specific semantic lexicons" from a collection of domain specific texts. These techniques try to find "co-occurrence patterns of words" to determine their sentiments. "Riloff and Wiebe (2003)" used a bootstrapping process to retain linguistically rich patterns of "subjective expressions" in order to "classify subjective expressions" from "objective expressions[16]". In this research "lexicon strength is computed using point wise mutual information for their co- occurrence with small set of positive seed words and a small set of negative seed words". Finally, the words we get are classified as either affirmative or pessimistic.

2. **The dictionary-based techniques**:

This approaches use lexical resources such as WordNet automatically [17] to retrieve similar words from WordNet. This techniques use "synonyms, antonyms and hierarchies" in WordNet to find word sentiments. Here, five of the six basic emotional categories has been described [18]. For direct affective words, weights from

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 1, January 2015

WordNet-Affect have been used. The affective weights are automatically acquired from a very large text corpus in an unsupervised fashion. The approach of using sentiment orientation of constituting words to find the opinion of the document does not provide good opinion, whereas sentiment often holds the composite meaning of the text, without the use of affect words.

3. **Bootstrapping** is another approach. The idea is to "use the output of an available initial classifier to create labeled data, to which a supervised learning algorithm may be applied". This method was used in synchronization with an "initial high-precision classifier" to learn extraction patterns for subjective expressions [16].

V. RESEARCH CHALLENGES

There are several challenges in Sentiment analysis.

- The *first* challenge is a word i.e "opinion word" is considered to be affirmative in one way may be considered pessimistic in another way.

- The *second* challenge is that sometimes person may convey their opinions in a different way.

- The *third* challenge is "it is difficulty in parsing the sentence to find the subject and object to which verb and/or adjective refer to".

- The *fourth* challenge is the opinion given on twitter is difficult to understand as it has poor abbreviations, lack of capitals, no proper spelling, no proper punctuation, grammaritical error etc.

- The *fifth* challenge is the language i.e most of the work done in opinion mining is focused on basically two language: English and Chinese and other languages will have to be explored.

- The *sixth* challenge is "detection of spam and fake reviews, mainly through the identification of duplicates, the comparison of qualitative with summary reviews, the detection of outliers, and the reputation of the reviewer".

VI. TOOLS

Some tools which are used in opinion mining are:

- **Review Seer tool [1]** – This tool is used to "automates the work done by aggregation sites. The Naive Bayes classifier approach is used to collect positive and negative opinions for assigning a score to the extracted feature terms".
- http://alias-i.com/lingpipe/
- **Opinion observer[1]** –"This is an opinion mining system for analyzing and comparing opinions on the Internet using user generated data. This system shows the results in a graph format showing opinion of the product feature by feature".
- **OpenNLP** – It perform the most common "NLP tasks, such as POS tagging, named entity extraction, chunking and coreference-resolution[9]" "http://opennlp.apache.org/.

- **Red Opal [11]** –It is a tool that "allows the users to find the polarity of products based on their features. Then it can give the scores to each product based on features extracted from the customer review".
- **Web Fountain [10]** - It uses the "beginning definite Base Noun Phrase heuristic approach for extracting the product features".

Some another online tools like Twitrratr, Twendz, Social mention, and Sentimetrics are also available to find the feedbacks in the web.

VII. APPLICATIONS

1. It is used in E-commerce activities. When any user buys any product or service from e-commerce websites, then it allows them to submit their feedback about shopping and product qualities. They provide summary for the product and different features of the product by assigning ratings or scores.

2. It is used in Ad Placement that displays ads as sidebars in online systems. It is helpful to detect WebPages that contain sensitive content which is unsuitable for ads placement.

3. It helps the Government in knowing their potency and failure by analyzing feedback from public.

4. It is used in Entertainment and helps the people to decide which movie or serial to watch.

5. It is used in Stock market to determine whether the stock price will be increasing or decreasing and help the investor to take decision whether to purchase or sell the stock.

6. It is used in Marketing. Nowadays each company provides the facility to its customers to write opinions about its product and services which they delivered. So it is helpful for companies to save a lot of money as well as time because there is no need to conduct surveys as the reviews related to all the products are available on their websites.

7. It is used in Education also, to help students to decide which academic institution is good for study.

8. It is used in Voice of Customer to know what individual customer is saying about products or services.

9. It is used in Voice of Market to determining what customers are feeling about products or services of competitors. It also helps the corporate to get customer opinion in real-time.

DOI 10.17148/IJARCCE.2015.4140

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 1, January 2015

VIII. CONCLUSION

This paper focuses on survey of opinion mining. There are some challenges exists like focus on other languages, dealing with negative statements, produce a summary of opinions based on product features/attributes, etc. More future research could be dedicated to all these challenges and more work has to be done for further enhancement of these challenges.

REFERENCES

[1] G.Vinodhini, RM. Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey", IJARCSSE(2)
[2] Blair Goldensohn. S, Hannan. K, McDonald. R, Neylon .T, Reis.G, andReynar.J," Building a sentiment summarizer for local service reviews". Inthe proceeding www Workshop on NLPChallenges in the Information Explosion Era, 2008.
[3] Zhang.W, Yu.C, and Meng.W, "Opinion retrieval from blogs", In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management
[4] Ghose. A and Ipeirotis. P, "Estimating the Socio Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics", Information Systems Research
[5] Nidhi Misra, C.K.Jha "Classification of Opinion Mining Techniques", International Journal of Computer Applications
[6] B. Liu 2007. Web Data Mining, Exploring Hyperlinks, Contents and Usage data.
[7] http://opennlp.apache.org/
[8] Yi and Niblack 2005. Sentiment Mining in Web Fountain",Proceedings of 21st international Conference on Data Engineering"
[9] "Christopher Scaffidi, Kevin Bierhoff, Eric Chang, Mikhael Felker, Herman Ng and Chun Jin 2007. Red Opal: product-feature scoring from reviews", Proceedings of 8th ACM Conference on Electronic Commerce
[10] Rui Xia , Chengqing Zong, Shoushan Li, "Ensemble of feature sets and classification algorithms for sentiment classification", Information Sciences.
[11] "Sentiment classification of Internet restaurant reviews written in Cantonese", Expert System with applications,2011.
[12] Kaiquan Xu , Stephen Shaoyi Liao , Jiexun Li, Yuxia Song, "Mining comparative opinions from customer reviews for Competitive Intelligence",Decision Support Systems.
[13] Sindhu C, Dr. S. ChandraKala,"A survey on Opinion Mining and Sentiment Polarity classification(2)"
[14] V. S. Jagtap, Karishma Pawar," Analysis of different approaches to Sentence-Level Sentiment Classification"
[15] Alm. C.O, Roth .D and Sproat .R, "Emotions from text: machine learning for text based emotion prediction". In Proceedings of the Joint Conference on Human Language Technology / Empirical Methods in Natural Language Processing ,Vancouver, Canada
[16] Ekman. P,"An Argument for Basic Emotions", Cognition and Emotion