Opinion Mining: A Survey

K.G. Nandha Kumar
Assistant Professor
Dept. of Computer Science
Dr.NGP Arts and Science College
Coimbatore - 641048

T Christopher, Ph.D.
Assistant Professor
PG and Research Dept. of Computer Science
Government Arts College
Coimbatore - 641018

ABSTRACT
Opining mining has gained high level focus in business and commercial fields. The opinions are considered as valuable data. Opinions are customer reviews, comments and feelings and collected from web forums, websites, and user groups which are posted by thousands of end users and also recorded manually. The collected opinions are processed by various techniques, algorithms, methods and software tools in order to extract the meaning from them. Extraction of meaningful information from opinions is given more importance in the field of business analytics. This entire process is popularly known as opinion mining or sentiment analysis and they are used in industries to develop quality of services, products. This article is an outcome of a study which made on various sub tasks, approaches, methods and techniques involved and applied in opinion mining.

General Terms
Data mining, Opinion mining, Sentiment analysis

Keywords
Opinion, Sentiment, Classification, Machine Learning, NLP.

1. INTRODUCTION
Nowadays opinion mining and sentiment analysis are emerging research areas in data mining, natural language processing and computational linguistics. It forms a triangle from three edge points. It may be viewed as a multidisciplinary area within computing sciences. It attracts researchers from all the three above said subject domains. The tremendous development of social websites such as facebook, blogs, tweets, user groups and web forums lead us to a great pool of opinions [1]. Such opinions are about to reviews of products, services, movies like entertainment, games, political movements and also religious activities of people. Some year ago collecting opinions was not an easy task. Opinions were collected from friends, neighbours, and colleagues by individuals who wanted to know about a specific product which he wanted to buy [2]. Industries and market research organizations had been involved in some activities such as interviewing consumers and experts, conducting surveys and field studies etc. They spent much money for collecting various kinds of opinions from different stake holders. But now the scenario is changed. There are thousands of websites which provide uncountable number of opinionated words, sentences and documents. The present challenge is extracting opinions and sentiments from corpus of opinionated data. Extraction, classification, analysis, and finding meanings from different type of input data are collectively called opinion mining [3]. Opinion mining plays a crucial role in business sector. A manufacturer may improve their product quality, design, sales and service by making use of opinions. The objective of this study is to unveil the approaches, methodologies and techniques to opinion mining in the current scenario. Such kind of study helps us to understand the interdisciplinary nature of opinion mining and sentiment analysis. This study clearly shows that the application of opinion and sentiment mining is vast and domain independent.

2. TASKS IN OPINION MINING AND SENTIMENT ANALYSIS
Though it is called as opinion mining, in real it is a collections of activities which are done in a well defined sequence. Under the so called opinion mining there are multiple sub tasks. Detecting subjectivity from corpus or online data, prediction of helpfulness of online comments and reviews, prediction of polarity, product's features extraction, opinion retrieval, opinion based entity ranking, summarization based on aspects, summarization based on sentiments, summarization based on contrastive view points, text summarization for opinions are identified as major tasks under the umbrella of opinion mining [4,8,16].

In polarity prediction a 'sentiment text' will be analysed to find whether it gives negative opinion or positive opinion. Some texts will be classified as neutral opinion based on the ratio of negative or positive measure called threshold. This is done in three levels; document level, sentence level and attribute level. This is also called as sentiment polarity prediction [14]. In detecting subjectivity, an analysis will be done for confirmation whether the taken data have opinions or not. But it is not about detecting polarity of the text.

Product features extraction is another main task. A list of features will be prepared. Opinion regarding the listed features will be extracted from the opinion text. Then they may be classified as positive or negative based on adverbs [5]. Opinion retrieval means identification of opinions from a mixture of opinions which presented in different formats including normal text, short and modernised text forms which are used in Short Message Services and online chatting, emoticons which are special symbols which represent various emotions. In entity ranking, based on customer reviews, comments and preferences all products will be ranked. This kind of ranking will help the manufacturer to improve the particular aspects of their products [11].

Predicting the usefulness of web reviews and comments is another task in opinion mining [12]. In this, a sorting will be made based on reviews by its helpfulness. Because it is not necessary that all the text have contain opinions. There are chances for meaningless words and sentences in online reviews. So this task intends to automatically predict the usefulness of user reviews and comments [6, 7]. Providing a short summary of opinions based on identified aspects is unavoidable. This definitely will help for further improvement. Opinion and sentiment both the words are interchangeably used to mention one another. In some cases sentiments denote feelings especially of reviewer. So a summary may be prepared based only on such sentiments. Highlighting contradictions among opinion is also crucial. At

such situations contradictory ideas may be taken place in the summary of opinions.

Instead of generating structured summaries of sentiments and opinions, there is one more summarization format called text summarization [9]. This point implies that there might be various formats of summarization. In text summarization, all the key opinions will be summarized in few sentences. It will be very useful for quick review.

3. APPROACHES

Opinions can be collected in two ways, from web sources and user generated contents. Cleansing of data is also called pre-processing which is done prior to starting opinion mining. There are many approaches followed for sentiment and opinion analysis including for data cleansing. Natural language processing, artificial intelligence including machine learning, statistical methods and web tools are major approaches to opinion mining [9, 10]. Each of them has its own advantages and limitations.

Natural language processing (NLP) is a linguistic approach with computational techniques. In NLP parts of speech (POS), unigram, bigram, n-gram, tree oriented methods and words of mouth (WOM) are frequently used techniques. Sentiwordnet is a tool used with NLP techniques and it is a lexicon based approach [11]. In machine learning, support vector machines (SVM), Bayesian techniques, and multilayer perceptrons are highly used. From statistics side regression models are repeatedly used. Web tools are highly preferred but they provide less accuracy of extraction in some cases [12]. But plenty of online tools are available for opinion mining and sentiment analysis. Amplified Analytics, Back tweets, IBM Social Sentiment Index, Lithium, OpenAmplify, Opinion Crawl, Reachli, SAS Sentiment Analysis Manager, SAS Sentiment Analysis Studio, Sentiment140, Sentimently, Social Mention, Topsey, Trackur, Tweetbeep, tweetSentiments, Twendz, Twitterfall are some online tools [13,15].

Maximum entropy is a famous method used together with all others methods and it provides reasonably better results. Precision and recalling capacity are scaling measures for all models and techniques.

4. RELATED WORK

4.1. Retrieving Product Features and Opinions from Customer Reviews

Lissete et al. [8] have proposed a language modelling framework which combines a probabilistic model of opinion words and a stochastic mapping model between words. This model extracts product features and opinions from a collection of free-text customer reviews. It does not require any training set of product features. In their frame work they have experimented several data sets with three models and they are W5, drAll, and drSelected.

W5 is a model obtained by using 5-grams as local context of words (mapping model). drAll and drSelected define the local context based on grammatical dependency relations. drAll is built using all the dependency relations obtained with the Stanford dependency parser and drSelected considers only a set of selective relations among opinion words. They have used English and Spanish words. The data sets are collected fromwww2.cs.uic.edu/~liub/FBS/sentiment-nalysis.html. The results show that the retrieval of opinions is better when using drSelected and with respect to speed performance W5 is faster than other two methods. The consolidated report of their result is presented in the following table.

Table 1. Consolidated Statistical Details

Total No. of datasets	3
Total No. of products	15
Total No. of reviews/opinion	17,220
Total No. of sentences	1,72,394
Average precision of W5	0.64
Average precision of drAll	0.68
Average precision of drSelected	0.69

4.2. Arabic Opinion Mining Using Combined Classification Approach

Alaa El-Halees [1] has implemented a combined approach which extracts opinions automatically from Arabic documents. Normally people do opinion mining in English documents through various sources such as blogs, facebook and tweets. But this is very notable. He has experimented different methods and finally found that a combination of different techniques is better than using single techniques. Using a single technique produces better performance when mining opinions from documents in English but in Arabic documents ensemble approaches produce better results. In his ensemble approach he has experimented three methods for opinion classification, lexicon based opinion classifier (Lex), maximum entropy (ME) method and k-nearest neighbour (kNN). All the three methods are applied one by one that is, the second methods is applied to the output of first one and the third method is applied to the output of second one. Data sets have been collected from three domains education, politics and sports. The evaluation metrics are recall, precision and F-measure. F-measure is a combined metric that takes recall and precision for consideration. Totally number of Arabic sentences taken for classification is 8793 statements from 1143 posts. The detailed report of result is given in the table 2.

Table 2. Accuracy of Combined Approach

Lex	Average of Correctly classified documents	48.69
	Accuracy	50.08
Lex+ME	Average of Correctly classified documents	61.81
	Accuracy	60.73

Lex+ME +kNN	Average of Correctly classified documents	80.21
	Accuracy	80.29

5. CONCLUSION

In the name of opinion mining lot of tasks are listed, various approaches and techniques are being followed by researchers based on domains and applications. After an analytical study made on these literature some common challenges have been found. They are; some opinion words which considered positive in a situation may classified as negative in some other situation. People express their opinion in different ways. In some times they are contradictory among their statements. Most number of comments and reviews express both positive and negative opinions. Customers combine different comments in same sentence which are easily understandable only to humans and not to computers. In many cases short pieces of text with less thought are entered as comments. These kinds of comments are difficult to classify. On the other hand, levels of accuracy, precision, recall and error control to be optimized. There are many doors open in the area of opinion mining and sentiment analysis for research.

At present natural language processing (NLP) techniques and tools are highly used for opinion mining in the pre-processing level, because, opinions are mostly represented in natural languages. They are unstructured in many cases. In addition to that emoticons are used with verbal opinions. So three fourth effort of opinion mining process is consumed by NLP techniques and remaining one fourth of effort is spent for actual mining tasks. Mining tasks can be achieved by various regular methods and as well as evolutionary methods including artificial neural networks (ANN).

There are many phases in opinion and sentiment analysis which can be experimented by using some new approaches. Those are left for upcoming researchers. Due to the interdisciplinary nature, a successful opinion mining requires expertness in NLP techniques, traditional data mining techniques and domain knowledge of opinions. In future, ANN techniques have high scope in mining opinions and they could be used either independently or as hybrid techniques with other evolutionary algorithms, statistics and traditional mining algorithms. But there is no fixed list of techniques for opinion and sentiment mining. Novel heuristic methods can be used for the same. The road is waiting for researchers to travel with any kind of algorithms, techniques, or methods to reach the extracted knowledge from different form of opinions.

6. REFERENCES

[1] Alaa El-Halees. 2011, Arabic Opinion Mining Using Combined Classification Approach, International Arab Conference On Information Technology (ACIT), Riyadh, Saudi Arabia.

[2] Aldo Gangemi, Valentina Presutti et al. 2014, Frame-Based Detection of Opinion Holders and Topics: A Model and A Tool, IEEE *Computational Intelligence Magazine*, 20 - 30.

[3] Anais Cadilhac, Farah Benamara *et al*. 2010, Ontolexical Resources for Feature Based Opinion Mining: A Case Study, *Proceedings of the 6th Workshop on Ontologies and Lexical Resources (Ontolex 2010)*, Beijing, 77 - 86.

[4] Chihli Hung and Hao-Kai Lin. 2013, Using Objective Words in Sentiwordnet To Improve Word-Of-Mouth Sentiment Classification, *IEEE Intelligent Systems*, 47 - 54.

[5] Chi-Hwan Choi, Jeong-Eun Lee et al. 2013, Sentiment Analysis For Customer Review Sites, *Proceedings of The 3rdInternational Conference On Circuits, Control, Communication, Electricity, Electronics, Energy, System, Signal And Simulation*, CES-CUBE- ASTL Vol. 25, 157 - 162.

[6] Erik Cambria, Bjorn Schuller et al. 2013, New Avenues in Opinion Mining and Sentiment Analysis, *IEEE Intelligent Systems*, 15 - 21.

[7] Hsinchum Chen and David Zimbra 2010, AI and Opinion Mining, *IEEE Intelligent Systems*, 74 - 80.

[8] Lisette Garcia-Moya, Henry Anaya-Sanchez et al. 2013, Retrieving Product Features and Opinions from Customer Reviews, *IEEE Intelligent Systems*, 19 - 27.

[9] Marie-Catherine de Marneffe, Bill MacCartney et al. 2006, Generating typed dependency parses from phrase structure parses, *Proceedings of international conference on language resources and evaluation*, Vol.6, 449 - 454.

[10] Michel Plantie, Mathieu Roche et al. 2008, Is A Voting Approach Accurate For Opinion Mining? *Proceeding DaWaK '08* - Proceedings of the 10th international conference on Data Warehousing and Knowledge Discovery, 413 – 422.

[11] Poongodi S and Radha N. 2013, Classification Of User Opinions From Tweets Using Machine Learning Techniques, *International Journal Of Advanced Research In Computer Science And Software Engineering*, 1061 - 1065.

[12] Sachin A. Kadam and Shweta T. Joglekar (2013), Sentiment Analysis: An Over View, *International Journal of Research in Engineering & Advanced Technology*, Vol. 1, Issue 4, pp. 1 - 7.

[13] Tamilselvi A and Parveentaj M (2013), Sentiment Analysis Of Micro Blogs Using Opinion Mining Classification Algorithm, *International Journal Of Science And Research*, Vol. 2, Issue 10, pp. 196 - 202.

[14] Xiuzhen Zhang, Zhixin Zhou et al. 2009, Positive, Negative, or Mixed? Mining Blogs for Opinions, *Proceedings of the 14th Australian Document Computing Symposium*, Sydney.

[15] Zhongwu Zhai, Bing Liu et al. 2012, Product Feature Grouping For Opinion Mining, *IEEE Intelligent Systems*, 37- 44.

[16] Zhu Zhang. 2008, Weighing Stars: Aggregating Online Product Reviews For Intelligent E-Commerce Applications", *IEEE Intelligent Systems*, 42- 49.