Survey on Sentiment Analysis

Amreen Shaikh
Department of Computer Engineering
Shree L R Tiwari College of Engineering
Mira Road
amreen1787@gmail.com

Madhuri Rao
Department of Information Technology
Thadomal College of Engineering
Bandra
my_rao@yahoo.com

Abstract **-** With the advent of internet and World Wide Web the field of Opinion mining and Sentiment Analysis is growing rapidly. There are numerous websites available on internet which provides options to users to give reviews about specific product. However the reviews expressed are mostly disorganized. An accurate method for predicting sentiments could help us, to extract opinions from the internet and predict customer's preferences which could prove valuable for economic and marketing research. There are various algorithms available for opinion mining. Opinion mining has three levels of granularities: Document level, Sentence level and Aspect level. In this paper, we study and analyze different issues, data sources, classification methods and evaluation metrics for Sentiment Analysis.

Index Terms ***–*** *Sentiment Analysis, Data Sources, Classification, Clustering, Evaluation Metric*.

*****

I. INTRODUCTION

Opinions are important to all humans as they influence ones behaviour. In today's competitive world, businesses and organizations always want to find consumer or public opinions about their products and services. Consumers also want to know the opinions of existing users of a product before purchasing it. In the past, when an individual needed opinions, he/she asked friends and family. When an organization or a business needed public or consumer opinions, it conducted surveys, opinion polls, and focus groups.[1]

With the explosive growth of social media (e.g., reviews sites, forum discussions, blogs, micro-blogs, Twitter, comments, and postings in social network sites) on the Web, individuals and organizations are increasingly using the content in these media for decision making. How to analyze and summarize the opinions expressed in these huge opinionated text data is a very interesting domain for researchers. This new research domain is usually called Sentiment Analysis or Opinion Mining [7].

This paper is organized as follows: Section 2 presents the Introduction to Sentiment Analysis, Section 3 includes sources used for opinion mining, Section 3 introduces classification for sentiment Analysis and Section 4 presents some evaluation Metrics of sentiment classification. Then we present some comparison of different research on Sentiment Analysis and Last section concludes our study.

II. SENTIMENT ANALYSIS

Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. The term sentiment analysis perhaps first appeared in Nasukawa and Yi, 2003, and the term opinion mining first appeared in Dave, Lawrence and Pennock, 2003[1].

A. Definition

Definition (opinion): An opinion is a quintuple,
(e i , a ij , s ijkl , h k , t l),

where e i is the name of an entity, a ij is an aspect of e i , s ijkl is the sentiment on aspect a ij of entity e i , h k is the opinion holder, and t l is the time when the opinion is expressed by h k .

B. Issues in Sentiment Analysis

Before exploring Sentiment Analysis in depth we need to understand following issues in Sentiment analysis:

1) A sentence containing sentiment words may not express any sentiment., e.g., "Can you tell me which Sony camera is good?" such sentence contain the sentiment word "good", but neither expresses a positive or negative opinion on any specific camera.
2) A positive or negative sentiment word may have opposite orientations in different application domains. For example, "suck" usually indicates negative sentiment, e.g., "This camera sucks," but it can also imply positive sentiment, e.g., "This vacuum cleaner really sucks."
3) A sentence containing sentiment words may not express any sentiment., e.g., "Can you tell me which Sony camera is good?"

4) Sarcastic sentences with or without sentiment words are hard to deal with, e.g., "What a great car! It stopped working in two days."
5) Many sentences without sentiment words can also imply opinions. For example The sentence "This washer uses a lot of water" implies a negative sentiment about the washer since it uses a lot of resource (water).

III. DATA SOURCES

Blogs, review sites, data and micro blogs provide a good understanding of the reception level of the products and services.

1) *Blogs:* With an increasing usage of the internet, blogging and blog pages are growing rapidly. Blog pages have become the most popular means to express one's personal opinions.
2) *Review sites*: The reviews for products or services are usually based on opinions expressed in much unstructured format websites like www.amazon.com (product reviews), www.yelp.com (restaurant reviews), www.CNETdownload.com (product reviews) and [2].
3) *DataSet:* Movie review data are available as dataset http://www.cs.cornell.edu/People/ pabo/movie-review-data). Other dataset which is available online is multi-domain sentiment (MDS) dataset. (http://www.cs.jhu.edu/mdredze/datasets/sentiment). The MDS dataset contains four different types of product reviews extracted from Amazon.com including Books, DVDs, Electronics and Kitchen appliances, with 1000 positive and 1000 negative reviews for each domain. [2].
4) Micro-blogging: Twitter is a popular micro blogging service where users create status messages called "tweets". These tweets sometimes express opinions about different topics. Twitter messages are also used as data source for classifying sentiment [2].

IV. CLASSIFICATION TECHNIQUES

There are different types of algorithms to analyze sentiments. Sentiment Classification techniques can be roughly divide into machine learning approach and lexicon based approach. The Machine Learning Approach (ML) applies the famous ML algorithms and uses linguistic features. The Lexicon-based Approach relies on a sentiment lexicon, a collection of known and precompiled sentiment terms. We discuss them below in brief.

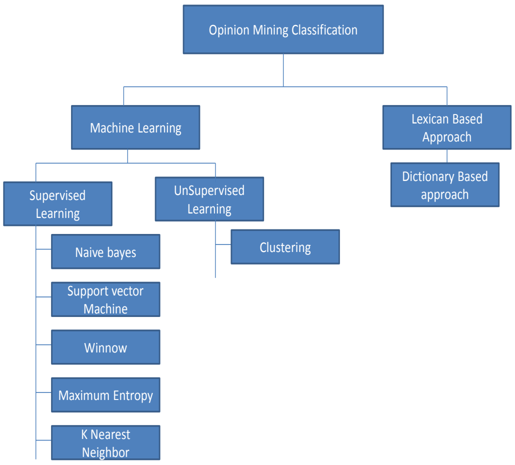

Fig. 1 Classification Techniques

A. *Machine Learning*

A system capable of getting, integrating and analyzing the knowledge automatically is known as Machine Learning.

1) *Naïve Bayes*: It is a simple but effective Learning & Classification algorithm. It is mostly used in Text Classification. The Classification method is based on theory of probability. It plays a vital role in probabilistic classification. It is also used in statistical method for classification and Supervised Learning method.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

2) *K-Nearest Neighbor (KNN):* It is also referred as Lazy Learning, Case-based Reasoning or Memory-based Reasoning.
Given a test document d, the system finds the k nearest neighbors among training documents. The similarity score of each nearest neighbor document to the test document is used as the weight of the classes of the neighbor document.

Support Vector Machine:

3) *SVM: It* is a supervised learning model. This model is associated with a learning algorithm that analyzes the data and identifies the pattern for classification [5]. For example, consider an instance which belongs to either class Circle or Diamond. There is a separating line (Fig 2) which defines a boundary. At the right side of boundary all instances are Circle and at the left side all instances are Diamond.

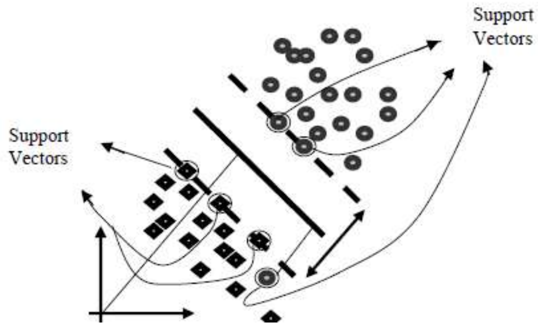

Fig. 2 Support Vector Machine

4) *Centroid classification:* Initially the prototype vector or centroid vector for each training class is calculated, then the similarity between a testing document to all centroid is computed, finally based on these similarities, document is assigned to the class corresponding to the most similar centroid [6].
5) *Winnow:* It is a well-known online mistaken-driven method. It works by updating its weights in a sequence of trials. On each trial, it first makes a prediction for one document and then receives feedback; if a mistake is made, it updates its weight vector using the document.
6) *Maximum Entropy:* Technique used for estimating probability distribution from data. That is when nothing is known the distribution should be uniform as possible. [6]
7) *Clustering Classifier:* Clustering is process of organizing objects and instances in a class or group whose members are similar in some way and members of class or cluster is not similar to those are in the other cluster.[5]

B. *Lexican Based Approach*

Lexican based approaches for sentiment classifications are based on the insight that the polarity of a piece of text can be obtained on the ground of the polarity of the words which compose it.

1) *Dictionary Based Approach*: In this approach first of all a small set of sentiment words which are known as seed words are collected manually with their known positive or negative orientations. Then this set is grown by searching their synonyms and antonyms in WordNet or another online dictionary. The new words are added to the existing seed list. Then next iteration is started. The iteration should be stopped when no new words are found. [4]

V. EVALUATION METRICS

The performance of different methods used for opinion mining is evaluated by calculating various metrics like precision, recall and F-measure.

1) Precision is the fraction of retrieved instances that are relevant.
2) Recall is the fraction of relevant instances that are retrieved.
3) The two measures are sometimes used together in the F1 score (also F-score or F-measure) is a measure of a test's accuracy.

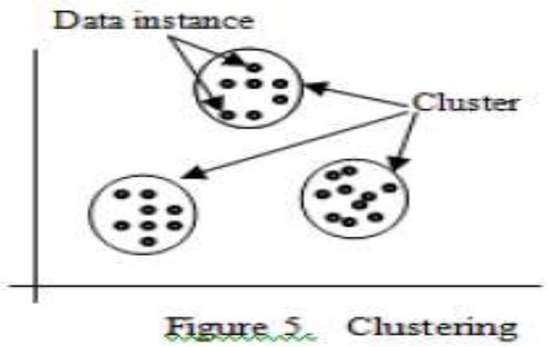

Fig. 3 Clustering

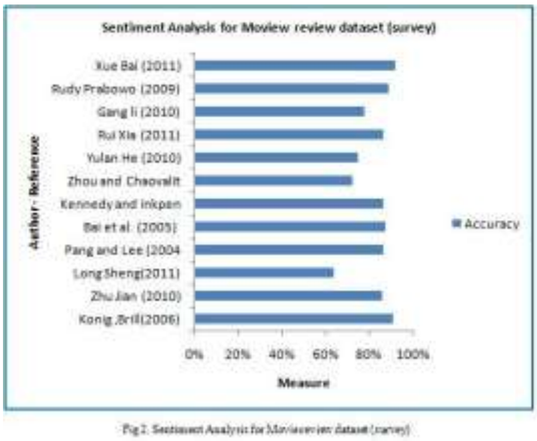

Fig. 4 Movie review Datset

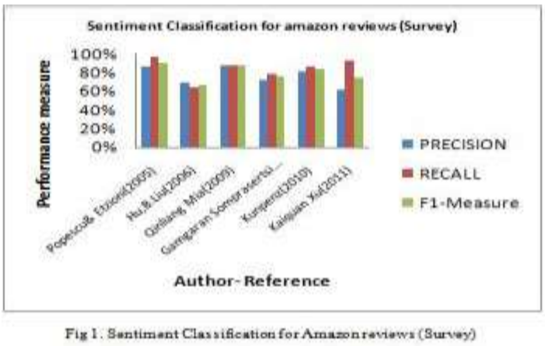

Fig. 5 Amazon Review Dataset

VI. COMPARISON ON DIFFERENT RESEARCH

TABLE I

Research paper/Studies	Technique Used	Data Source	Performance (Accuracy)
Kaiquan Xu(2011)	SVM	Amazon reviews	61 %
Xue Bai (2011)	Naïve bayes	Moview review	92%
Gamgarn somprasti (2010)	Maximum Entropy	Amazon reviews	
Gang li (2010)	K-means Clustering	Moview review	78%
QingliangMiao (2009)	Lexical resource	Amazon reviews	87.6%
Mining Of Product Reviews At Aspect Level	Dictionary based unsupervised learning	Amazon reviews	74%
Polarity Detection at Sentence Level	Lexicon dictionary based approach		67%
Kennedy and Inkpen (2006)	SVM	Movie review	86.2%
Godbole et al. (2007)	Lexical approach	blog posts	82.7–95.7%
Gamon (2005)	Naïve Bayes	Car reviews	86%
Pang and Lee (2004	Nave Bayes	Movie review	86.4%

VII. CONCLUSION

Sentiment analysis has become very popular field of research. A lot has been researched in this field but still there are many issues as sentiment analysis processes text based unstructured data. Dictionary based approach takes less processing time than supervised learning approach but accuracy is not up to the mark. Supervised learning approach provides better accuracy. It is found that sentiment classifiers are severely dependent on domains or topics. From the above work it is evident that neither classification model consistently outperforms the other, different types of features have distinct distributions. It is also found that different types of features and classification algorithms are combined in an efficient way in order to overcome their individual drawbacks and benefit from each other's merits, and finally enhance the sentiment classification performance.

REFERENCES

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

[2] G.Vinodhini, R.M.Chandrasekaran, "Sentiment Analysis And Opinion Mining: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering, June 2012

[3] Lecture from Dr. Sreerama K. Murthy Ph.D., Johns Hopkins Univ.

[4] Sentiment analysis algorithms and applications: A survey, Walaa Medhat, Ahmed Hassan, Hoda Korashy, 2013.

[5] Pravesh Kumar Singh, Mohd Shahid Husain, “Methodological study of opinion mining and sentiment analysis techniques” International Journal on Soft Computing (IJSC) Vol. 5, No. 1, February 2014.

[6] S. Kasthuri, Dr. L. Jayasimman, Dr. A. Nisha Jebaseeli “An Opinion Mining and Sentiment Analysis Techniques: A Survey” International Research Journal of Engineering and Technology (IRJET), 2016.

[7] Mr. Saifee Vohra, Prof. Jay Teraiya “Applications and Challenges for Sentiment Analysis : A Survey” International Journal of Engineering Research & Technology (IJERT), (2013).

[8] Asmita Dhokrat, Sunil Khillare, c.Namrata Mahender “Review on Techniques and tools used for opinion mining” International Journal of Computer Applications Technology and Research, (2015).

[9] Kaiquan Xu , Stephen Shaoyi Liao , Jiexun Li, Yuxia Song, “Mining comparative opinions from customer reviews for Competitive Intelligence”, Decision Support Systems 50 (2011) 743–754.

[10] Gamgarn Somprasertsri, Pattarachai Lalitrojwong , Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization, Journal of Universal Computer Science, vol. 16, no. 6 (2010), 938-955.

[11] Gang Li , Fei Liu , “A Clustering-based Approach on Sentiment Analysis” ,2010, 978-1-4244-6793-8/10 ©2010 IEEE.

[12] Qingliang Miao, Qiudan Li, Ruwei Dai , “AMAZING: A sentiment mining and retrieval system”, Expert Systems with Applications 36 (2009) 7192–7198.

[13] Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," Computational Intelligence, vol. 22, pp. 110–125,2006.

[14] Namrata Godbole, Manjunath Srinivasaiah, Steven Skiena, "LargeScale Sentiment Analysis for News and Blogs", ICWSM‟2007 Boulder, Colorado, USA.

[15] Lecture from Dr. Sreerama K. Murthy Ph.D., Johns Hopkins Univ.

[16] Sentiment analysis algorithms and applications: A survey, Walaa Medhat, Ahmed Hassan, Hoda Korashy, 2013.