IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 96-101

Sentiment Analysis Techniques and Applications a comparative study

Mrs.Deepa Honrao
(Assistant Professor, MCA Department,DES's Navinchandra Mehta Institute of Technology and Development, Mumbai University,India)

Abstract: *E-commerce sites, social media, forums, blogs etc. have become very popular. This has created a new avenue where anyone can exchange views, ideas,suggestions and experiences about products, services or events. This trend accumulated a huge amount of user generated - unstructured data on the web. If this content can be extracted and analyzed properly then it can act as a vital factor in decision making and to gain insights on the entity. But manual extraction and analysis of this huge content is a difficult and time-consuming task. The other challenges being content is unstructured in nature and it is written in natural language. This situation has given rise to new research areas called Opinion Mining and Sentiment Analysis. These terms are sometimes also used interchangeably and are an extension of Data Mining. This paper discusses the key concepts used in Sentiment Analysis and presents comparative analysis of its techniques*

I. Introduction

Human decision making is highly influenced by others experiences, reviews and opinions. With the technological advancements it has become easy and convenient to share experiences or give reviews on product, current issues or even a travelling destination. Buyers likely leave feedback either when they are very happy or not happy. The trend shows 5% of happy buyers and 100% of un happy buyers give their feedback. The reviews are shared in the form of textual comments, in the form of numerical rating or number of stars. Merchants review these ratings to gain insights into its products strength and weaknesses based on the sentiment of the customers.

The ratings and stars give objective feedback on some defined scale whereas the textual comments give brief feedback about the experience. It helps immensely to product developers to gain insight of the experience. But the textual feedbacks are not structured hence are tough to analyze and it becomes challenging when these are in volumes.

In this huge data, evaluating and summarizing the opinions expressed, is a very interesting area for researchers. This new research domain is part of Natural Language Processing and is usually called Sentiment Analysis or Opinion Mining. As per S M Vohra et al[3]
"Sentiment analysis is the automated mining of attitudes, opinions, and emotions from text, speech, and database sources through Natural Language Processing (NLP)."

Sentiment analysis involves classifying opinions in text into categories like "positive" or "negative" or "neutral". It extracts the feeling of the author about some topic. Opinion Extraction &Sentiment classification are the two important tasks involved in OM and SA.
1) Opinion Extraction is about extracting opinionated phrases from free text, in proper context
2) Sentiment classification is about classifying above extracted phrases based on sentiment orientation. It utilizes various ML techniques such as SVM, Naïve Bayes, character Based N-gram model etc. for sentiment classification.

Main research fields in SA are Sentiment Prediction, Subjectivity Detection, Aspect Based Sentiment Summarization, Contrastive Viewpoint Summarization, Text summarization for Opinions, Product Feature Extraction & detecting opinion spam. Determining whether text is opinionated or not is Subjectivity Detection. Predicting the polarity of text is Sentiment Prediction. Providing sentiment summary in the form of star ratings or scores of features of the product is Aspect Based Sentiment Summarization. Generating few sentences that summarize the reviews of a product is called Text Summarization. Emphasizing contradicting opinions is called Contrastive Viewpoint Summarization. Extracting product features from its review is called Product Feature Extraction. Identifying fake or bogus opinion from reviews is called Detecting opinion spam [3]
This paper is an attempt to study the concepts of Sentiment Analysis and compare various techniques used in this field. The paper is organized as follows:

Section 2 presents overview of sentiment analysis , section 3 discusses various techniques used for Sentiment Analysis and in section 4 analysis and comparison of these techniques is discussed. Finally section 5 talks about applications of Sentiment analysis with concluding remarks in section remarks in section 7.

Sentiment Analysis

Sentiment Analysis[SA] is also known as Opinion Mining[OM]. It is the computational study of users' point of views, attitudes & emotions toward an entity. Entities can be individuals, events or topics. These topics are most likely covered by reviews. The 2 terms SA or OM are interchangeably used & express a mutual meaning. Some researchers however stated, both have slightly different notions [1]. OM extracts and then analyzes people's point of view about an entity, whereas what SA does is, it identifies the sentiment expressed by users in a text & then analyzes it. SA aims to to identify users' point of views, emotions, and classify polarity. This is shown in Fig. 1. It can also be considered a process of classification, as shown in Fig. 1. The 3 main levels of classification of SA are:

- Document-level,
- Sentence-level, and
- Aspect-level SA.

Document type of SA aims to identify polarity of point of view or emotion. Whole document is considered a basic information unit (talking about one topic).
Sentence type of SA aims to identify emotion in the sentence. First it is identified if the sentence is subjective or objective. If subjective, polarity of the opinion will be determined.
Aspect / Feature type sentiment classification identifies and extracts product features from the source data.

The flow of activities for sentiment analysis is depicted as below in fig 1.

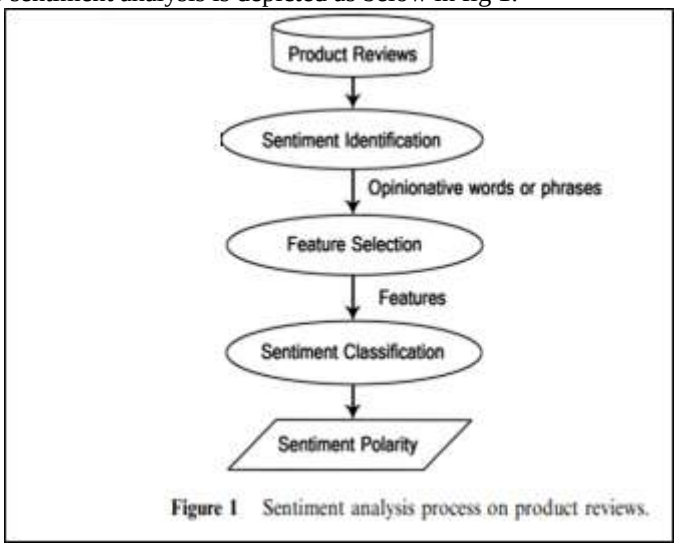

Figure 1 Sentiment analysis process on product reviews.

Source: Source: Walaa Medhat et al,Sentiment analysis algorithms and applications A survey
Feature selection/extraction -is an important step in the SA. From the preprocessed data we need to extract features relevant to sentiment analysis. Some of the features include:
a) Term presence and frequency: It usually consists of n-grams of words and their frequency counts
b) Parts of speech tagging: words in the text are tagged with their respective parts of speech in order to extract adjectives nouns verbs which add meaning to the sentiment.
c) Opinion words and phrases: words or phrases that indicate opinion of the text
d) Negation: presence of words like 'not', 'nor', 'neither' may reverse the sentiment of whole sentence.
E.g. :"not good "
Twitter which is a rich source of user views on various aspects has **Twitter specific features.** Presence of emoticons in tweets, positive or negative hashtags are all twitter specific features which add meaning to the sentiment.

II. Techniques for Sentiment Analysis

Machine learning & Lexicon learning are 2 main approaches of sentiment analysis. Some researchers got better results by combining both calling it hybrid approach. ML classifies text whereas sentiment dictionary

with opinion words is used by lexicon, to match with data to determine polarity. Sentiment scores are assigned to opinion words to describe the intensity of polarity of the words in the dictionary are.

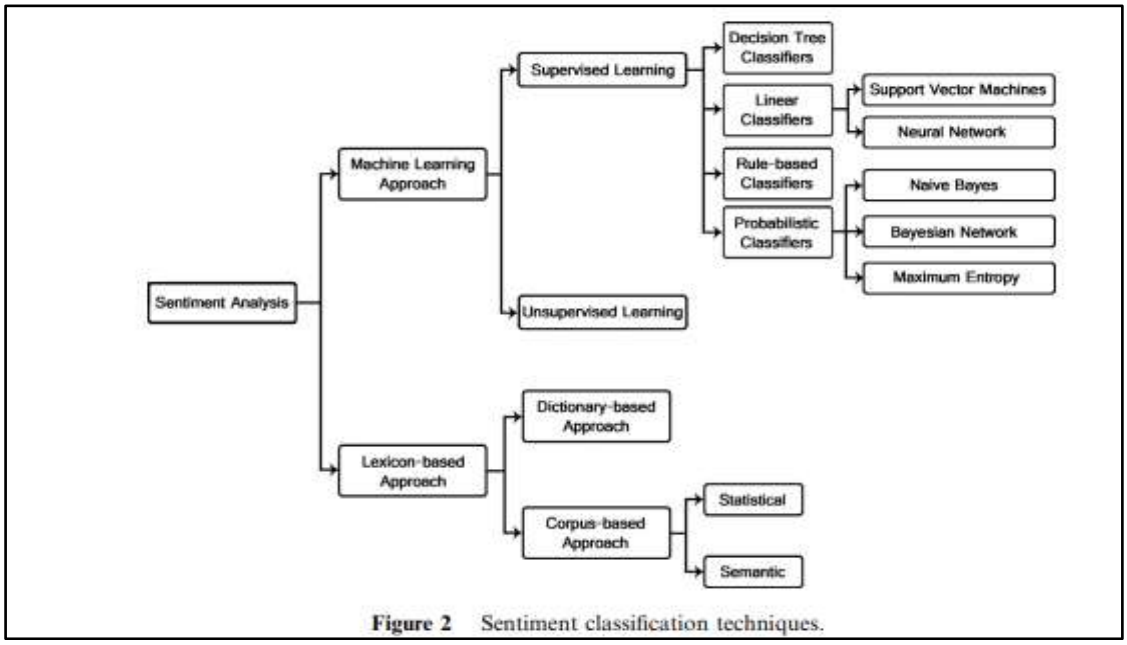

Figure 2 Sentiment classification techniques.

Source: Walaa Medhat et al,Sentiment analysis algorithms and applications A survey

1) Machine learning based techniques:

ML approach applicable to sentiment analysis is based on supervised classification. Two sets viz. training & test are needed in ML based techniques. Document's characteristics are learnt by Training set. Performance of the classifier is learnt using Test set. User reviews classification is usually done using many ML techniques. Most successful ML techniques are Naive Bayes, Support Vector Machines & Maximum Entropy.

a)Naïve Bayes - Naive Bayes (NB) rule says, final probability is the sum of each feature's probability contributed independently to be a class & each one has its distribution. Bayes theorem is the basis of this simple method.

The Naive Bayes classifier is the simplest (as the name suggests) and most commonly used classifier. Naive Bayes classifier works very well for text classification as it computes the posterior probability of a class, based on the distribution of the words (features) in the document. The model uses the Bag of words feature extraction . It assumes that the features are independent of each other.

It uses Bayes Theorem to predict the probability that a given feature set relates to a particular label as denoted in equation(1):

P(label/features) = P(features/label)*P(label)/P(features) P(label) (1)

P(label) is the prior probability of a label or the likelihood that a random feature set the label. P(features|label) is the prior probability that a given feature set is being classified as a label. P(features) is the prior probability that a given feature set is occurred. Given a feature, P(features|label) is the prior probability that feature set is being classified as a label. P(features) is the prior probability that a given feature set is occurred.

Given the Naive assumption which states that all features are independent of eachother , the equation(1) could be rewritten as follows refer equation(2):

P(label|features) = P(label)*P(f1|label)*…P(fn|label) P(features) (2)

b) Support vector machines The main principle of SVMs is to identify linear separators or hyperplane in the search space. These hyperplanes can best separate the different classes. There can be several hyperplanes that separate the classes, but the one that is chosen is the hyperplane in which the normal distance of any of the data points is the largest, so that it depicts the maximum margin of separation. Text classification are perfectly suited for SVMs because of the meager nature of text, in which few features are unrelated, but they tend to be correlated with one another and generally organized into linearly separable categories.

c) Maximum Entropy- This classifier is also called as conditional exponential classifier. The joint features are created by encoding the set of features.The priorities are combined with these joint features to form the parameters .These classifiers are also sometime known as the exponential classifiers, because they work by extracting some set of features from the input set, combining them linearly and then using this sum as exponent. Point wise Mutual Information (PMI) is used in order to find the co-occurrence of a word with positive & negative words when this method is done in an unsupervised manner. This Classifier is one of the models, which do not assume the independent features, but it derives the features with mathematical base..

2) The Lexicon based Technique:

It is a unsupervised learning technique, which classifies the text by comparing the features of testing text under consideration against sentiment lexicons, whose sentiment values are defined prior to their use. Semantic lexicon is lists ofwords with semantic polarity as positive, negative and neutral, defined aprior. It classifies the document by calculating the sentiment score of each opinion word present in the document with reference to lexicons. Then documents with more positive word lexicons is a positive document otherwise it is classified as negative document. The significant steps of lexicon based sentiment analysis are as follows :

a. Preprocessing: This step cleans the document by removing HTML tags, special characters or symbolic characters and noisy characters present in the document, by correcting spelling mistakes, grammatical mistakes and by replacing non- dictionary words withtheiractualterm.

b. Feature Selection: This step Extract the feature present in the document by using techniques like POStagging.

c. Sentiment score calculation: Initialize s with 0. For each extracted sentiment word, checkwhether it is present in the sentiment dictionary, If present with negative polarity, w then s = s-w or If present with positive polarity, w then s = s +w.

d. Sentiment Classification: If s is below a particular threshold value then classifying the document as negative otherwise classify it aspositive.

2.1 Sentiment Lexicon Construction

Sentiment lexicon can be constructed in three ways:

a. **Manual lexicon construction** - In manual lexicon construction, the lexicons are constructed with human effort. It is very difficult and time- consuming task

b. **Dictionary-based lexicon construction**- In dictionary-based lexicon construction, a small set of sentiment words and their polarity are determined manually and then this set is widened by adding more words into it using WordNet dictionary or SentiWordNet dictionary and their synonyms and antonyms.

c. **Corpus-based lexicon construction**- In corpus-based lexicon construction, it considers syntactic patterns of the words in the document. It requires annotated training data to produce lexicons.

III. Comparative Study of Sentiment Analysis Techniques

In most of the cases the supervised machine learning approaches outperformed the unsupervised lexicon based approaches. But the requisite of big labeled training dataset for supervised learning approaches compelled researchers to acquire lexicon based methods, as it is very easy to collect unlabeled dataset. Most domains except movie reviews lack labeled training data in this case unsupervised methods are very useful for developing applications. Most of the researchers observed that SVM has high accuracy than other algorithms. The combined approach of supervised and lexicon based approach, known as hybrid approach gained relatively better performance than the other two.In some of the researches volume and type of dataset also influenced the performance

Table 1: Comparison of Sentiment Analysis Techniques

Paper	Dataset	Approach	Technique	Accuracy
Pang et al [5]	Movie reviews	Supervised	Naïve Bayes	81.0%
			SVM	82.9%
			ME	80.4%
Turney [6]	movie, bank , automobile and travel destinations	Unsupervised	PMI Pointwise Mutual Information	74.39%
M. Hu and B. Liu [7]	Customer Reviews	Unsupervised	Lexicon	84.2%
Alec Go et al [8]	Twitter	Supervised	Naïve Bayes	81.3%
			SVM	82.2%
			ME	80.55
Qiang et al [9]	Travel Blogs	Supervised	Naïve Bayes	80.71%
			SVM	85.14%
			Character based N-gram model	84.05%
Alaa et al [10]	Amazon Product reviews	Unsupervised	Lexicon	67%
Harb et al. [11]	Movie review	Unsupervised	Lexicon	71%
Zhang et al. [12]	Twitter tweets	Hybrid	MLand Lexicon	85.4%
Mudinas et al. [13]	customer reviews	Hybrid	MLand Lexicon	82.3%
Fang et al. [14]	Multi domain	Hybrid	MLand Lexicon	66.8%

IV. Sentiment Analysis Applications

Opinions are vital to almost all human activities because they are key thought drivers of our behaviours. Whenever we need to make a decision, we want to know others beliefs on the entity. Growth of human views on electronic platform has effected in Sentiment Analysis being used in varied fields. Various uses of Sentiment Analysis are as follows:

1.Applications to review related websites

One of the prima focus application of sentiment analysis, is reviewing users social media comments, online reviews, and free-text survey responses. Listen to what they're saying in their own words. Understand exactly how your customers feel, and why. Leverage that information to:

- Grow your market share and presence
- Build customer loyalty and improve retention
- Deliver innovative products in less time
- Perfect your go-to-market strategy

As an individual user analysed product reviews are helpful for product buying decision.

2.Applications as sub component to other technologies

Sentiment analysis in the form of data augmentation tool is used for recommendation system, which will not recommend items with higher negative reviews. It can also be applied to identify phishing emails using subjectivity detection

3.Applications in business and government intelligence

Sentiment-analysis technologies for extracting opinions from unstructured human-authored documents would be excellent tools for handling many business-intelligence tasks. Based on sentiment analysis outcomes prediction models can developed. Government intelligence is another application that has been considered. For example, it has been suggested that one could monitor sources for increases in hostile or negative communications .

4.Applications in different domains

Sentiment analysis is no more related to countable few fields but it has made its presence felt in various domains. As is well known, opinions matter a great deal in Politics, some work has focused in understanding what voters are thinking. Understanding voters thought process has enabled political parties to make strategy and win the elections

On a related note, there has been investigation into opinion mining in weblogs devoted to legal matters, sometimes known as "blawgs" [15].

5.Applications as a service

In recent years, we have seen the democratization of sentiment analysis, in that it's now being offered as-a-service. Companies such as Microsoft, IBM and smaller emerging companies offer REST APIs that integrate easily with your existing software applications.

6.Applications in workforce analytics

Workforce analytics is currently evolving field with sentiment analysis at its core. How employees feel about their companies, managers, and work environments is analysed through the voice of employee.The mail communications in the organization are monitored to gain insights about work environments.

V. Conclusion

Abundant availability of unstructured data in form of reviews and ratings has powered a lot of initial research in sentiment analysis, however, as we look forward, one can be optimistic that the future holds more diverse and more persuasive applications of sentiment analysis. It is coming out as one of the emerging field of Data Mining. Most of the business organizations believe their business success solely depends on customer satisfaction, encouraging academicians and researchers for more accurate results of sentiment analysis. This paper has made a humble effort to touch upon all the aspects of sentiment analysis by comparing the techniques used for it. Though the research in this field is moderately mature, researchers are working towards applying it to various fields of public interest with negligible human effort and improved accuracy.

References

[1]. Walaa Medhat, Ahmed Hassan, Hoda Korashy, "Sentiment analysis algorithms and applications:A survey –",Ain Shams Engineering Journal (2014) 5, 1093–1113

[2]. B. Pang and L. Lee, "Opinion mining andsentiment analysis," Foundations and Trends inInformation Retrieval 2(1-2), 2008, pp. 1–135.

[3]. Mr. S. M. Vohra, PROF. J. B. Teraiya,"A Comparative Study Of Sentiment Analysis Techniques, , Journal Of Information, Knowledge And Research In Computer Engineering",ISSN: 0975 – 6760| NOV 12 TO OCT 13 | VOLUME – 02, ISSUE – 02 Page 313

[4]. Anuja P Jain, Asst. Prof Padma Dandannavar, "Application of Machine Learning Techniques to Sentiment Analysis", 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)pp628-632

[5]. B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learningtechniques," Proceedings of the ACL-02 conference on Empirical methods in natural language rocessing,vol.10, 2002, pp. 79-86.

[6]. P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews", Proceedings of the Association for Computational Linguistics (ACL), 2002, pp. 417– 424.

[7]. M. Hu and B. Liu, "Mining and summarizing customer reviews," Proceedings of the tenth ACM international conference on Knowledge discovery and data mining, Seattle, 2004, pp. 168-177.

[8]. Alec Go, Richa Bhayani, Lei Huang," Twitter Sentiment Classification using Distant Supervision",

[9]. Qiang Ye, Ziqiong Zhang, Rob Law, Sentiment classification of online reviews to travel destinations by supervised machine learning approaches, / Expert Systems with Applications 36 (2009) 6527–6535

[10]. Alaa Hamouda, Mohamed Rohaim," Reviews Classification Using SentiWordNet Lexicon" The Online Journal on Computer Science and Information Technology (OJCSIT) pp120-123

[11]. A. Harb, M. Planti, G. Dray, M. Roche, Fran, o. Trousset and P. Poncelet, "Web opinion mining: how to extract opinions from blogs?", presented at the Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology, Cergy-Pontoise, France,2008.

[12]. L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B.Liu, "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis", Technical report, HP Laboratories,2011.

[13]. A. Mudinas, D. Zhang, M. Levene, "Combining lexicon and learning based approaches for conceptlevel sentiment analysis", Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, ACM, New York,NY, USA, Article 5, pp. 1-8, 2012.

[14]. Ji Fang and Bi Chen, "Incorporating Lexicon Knowledge into SVM Learning to Improve Sentiment Classication", In Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), pages 94–100, 2011.

[15]. Jack G. Conrad and Frank Schilder. Opinion mining in legal blogs. In Proceedings of the International Conference on Artificial Intelligence and Law (ICAIL), pages 231–236, New York, NY, USA, 2007.

[16]. Doaa Mohey El-Din Mohamed Hussein," A survey on sentiment analysis challenges", Journal of King Saud University – Engineering Sciences (2018) 30, 330–338

[17]. Bing Liu,"Sentiment Analysis and Opinion Mining"

[18]. Haseena Rahmath P,, Tanvir Ahmad,"Sentiment Analysis Techniques - A Comparative Study", IJCEM International Journal of Computational Engineering & Management, Vol. 17 Issue 4, July 2014