Review on Sentiment Analysis in Natural Language Processing

Neha Gaur[1], Neetu Sharma[2]

[1]M.Tech. Student ,Computer Science & Engineering

Ganga Institute of Technology and Management Kablana, Jhajjar, Haryana, India

[2] *HOD*, Computer Science & Engineering

Ganga Institute of Technology and Management Kablana, Jhajjar, Haryana, India

[1]neha.gaur20@yahoo.com; [2]pg.gitam@gmail.com

Abstract— Sentiment analysis is text analysis techniques that automatically detect polarity of text. *Sentiment analysis also called as opinion mining which is one of the major tasks of NLP (Natural Language Processing). Sentiment analysis has gain much attention in recent years.* ***People are intended to develop a system that can identify and classify opinion or sentiment as represented in an electronic text. Consumers regularly face the trade-off in purchase decisions so nowadays if one wants to buy a consumer product one prefer user reviews and discussion in public forums on web about the product. Many consumers use reviews posted by other consumers before making their purchase decisions. People have a tendency to express their opinion on various entities. As a result opinion mining has gained importance. Sentiment Analysis deals with evaluating whether this expressed opinion about the entity has a positive or a negative orientation. Consumers need to decide what subset of available information to use. The process of identifying and extracting subjective information from raw data is known as sentiment analysis. An accurate method for predicting sentiments could enable us, to extract opinions from the internet and predict online customer's preferences, which could prove valuable for economic or marketing research. Till now, there are few different problems predominating in this research community, namely, sentiment classification, feature based classification and handling negations. This paper presents a survey covering the techniques and methods in sentiment analysis and challenges appear in the field.***

I.INTRODUCTION

Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. In other words we can say Sentiment analysis is a type of natural language processing for tracking the mood of the public about a particular product or topic. Its major task is Identify and extract sentiment in given string . It takes an input string and assigns a sentiment rating in the range [-1 t0 1] (very negative to very positive).It involves in building a system to collect and examine opinions about the product made in blog posts, comments, reviews or tweets. Sentiment analysis can be useful in several ways. For example, in marketing it helps in judging the success of an ad campaign or new product launch, determine which versions of a product or service are popular and even identify which demographics like or dislike particular features.

Sentiment analysis concentrates on attitudes, whereas traditional text mining focuses on the analysis of facts. There are few main fields of research predominate in Sentiment analysis: sentiment classification, feature based Sentiment classification and opinion summarization. Sentiment classification deals with classifying entire documents according to the opinions towards certain objects. Feature-based Sentiment classification on the other hand considers the opinions on features of certain objects. Opinion summarization task is different from traditional text summarization because only the features of the product are mined on which the customers have expressed their opinions. Opinion summarization does not summarize the reviews by selecting a subset or rewrite some of the original sentences from the reviews to capture the main points as in the classic text summarization. Sentiment Analysis uses various classification techniques to identify the tone of a given piece of text. It indicates whether the text is positive, negative or neutral. This analysis can be aggregated over large sets of data and the resulting information can be helpful in different contexts.

Nowadays, social media has become a popular platform for people to convey their voice/views to the public. The Internet has rapidly advanced from a static to an interactive medium. Today's users cannot only obtain information but also actively generate content. News reports, forums, blogs, etc. These are the main sources of public opinion information. This online word-of-mouth represents new and measurable source of information with many applications, this process of identifying and extracting subjective information from raw data is known as sentiment analysis. The text contains both cases and opinion which could be extracted using natural language processing to get some opinionated views. Sentiment analysis is also called as opinion mining. Sentiment analysis not only helps in allowing the user to get more and relevant information about different products and services on a mouse click, but also helps in arriving at a more informed decision. The analysis of sentiments may be document based where the sentiment in the entire document is summarized as positive, negative or objective. It can be sentence based where individual sentences, bearing sentiments, in the text are classified. SA can be phrase based where the phrases in a sentence are classified according to polarity. In fact, to identify the emotion analysis task views expressed in a text is positive or negative weather. Natural language processing (NLP) computer science, Artificial intelligence, and computers and human (natural) concerned with interactions between languages is an area of Linguistics. For instance, in a product review, it identifies features of the product that have been commented on by the reviewer and determines whether the comments are positive, negative or neutral. For example, in the sentence, "The life of the battery of this mobile is too compressed", the opinion is on "life of the battery" of the mobile object (target) and the opinion is negative. Many day to day life applications require this level of detailed analysis because in order to make product upgrade one needs to know what components and/or features of the product are liked and disliked by consumers. Such information has not come across by sentiment and subjectivity classification.

II. **LITERATURE REVIEW**

One of the major problem in sentiment analysis is categorization of sentiment polarity . Given a piece of written text, the problem is to categorize the text into one specific sentiment polarity, positive or negative (or neutral). There are three levels of sentiment polarity categorization, namely the document level, the sentence level, and the entity and aspect level. The document level concerns whether a document, as a whole, expresses negative or positive sentiment, while the sentence level deals with each sentence's sentiment categorization; The entity and aspect level then targets on what exactly people like or dislike from their opinions.

Nasukawa and Yi [1] specify that instead of state the complete document into positive or negative, they express sentiments connect with positive or negative for a particular topic from a document. Also, they clarify the fundamental issue in sentiment analysis which is knowing the sentiment expressed in texts whether the sentiment shows positive or negative opinion.

Ding et al [3] proposed an effective method for identifying semantic orientations of opinions expressed by reviewers on product features. It is able to deal with two major problems with the existing methods, (1) opinion words whose semantic orientations are context dependent, and (2) aggregating multiple opinion words in the same sentence. For (1), a holistic approach is proposed that can accurately infer the semantic orientation of an opinion word based on the review context. For (2), a new function to combine multiple opinion words in the same sentence is proposed.

Taylor et al [4] presented a generic design of a tourism opinion mining system that aims to be useful in many industries. They also used their proposals to successfully implement the system and solve a specific problem in the Lake District tourism industry.

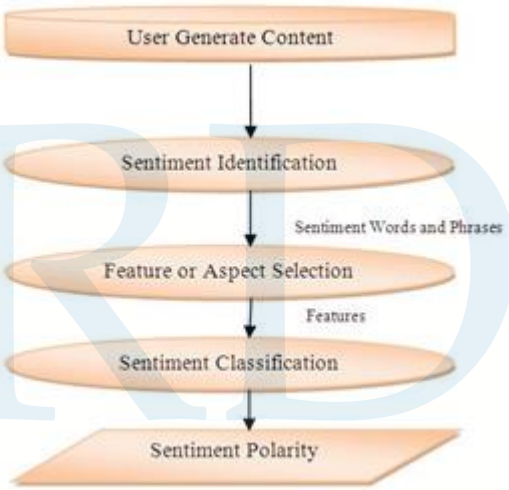

Fig. *1* Sentiment analysis process on user generated content

In Zhu et al [5] an aspect-based opinion polling system takes as input a set of textual reviews and some predefined aspects, and identifies the polarity of each aspect from each review to produce an opinion poll.

In [6] Haddi, Lui and Shi investigated the sentiment of online movie reviews. They used a combination of different pre-processing methods to reduce the noise in the text in addition to using chi-squared method to remove irrelevant features that do not affect its orientation. Authors have reported extensive experimental results, showing that, appropriate text pre-processing accuracy achieved on the two data sets is comparable to the sort of accuracy that can be achieved in topic categorization, a much easier problem.

In Moraes, Valiati and Neto [9] concentrated on comparing between SVM and ANN under the condition of the requirement to achieved good classification accuracies. Also, experiments evaluated all methods as a function in bag-of-words (uni grams) approach in particular terms. Related sentiment learning literature the necessary contributions/findings are in two points. The first point is that in term of classification accuracy on a benchmark dataset of movies reviews. The second point is as a complete comparison in the context of balanced data.

III. SENTIMENT CLASSIFICATION

Sentiment classifications are based on polarity, which may become positive, negative, or neutral. That's mean opinions may be classified into positive, negative, or neutral. Moreover, there is a forth type which is a constructive opinion which obtains suggestion to make the product better . In relation to sentiment analysis, the literature survey done indicates two types of techniques including machine learning and semantic orientation. Sentiment classification assumes that the opinion document express opinions on a single entity or object and opinions are from a single opinion holder. Opinionated documents contain information which can be broadly categorized in two categories: facts, which are typically objective statements about some entity (object) or event and sentiments, which are subjective in nature expressing sentiments and feelings of the opinion holder about the entity. Both facts and opinions are useful in decision making.

In addition to that, the nature language processing techniques (NLP) is used in this area, especially in the document sentiment detection. Current-day sentiment detection is thus a discipline at the crossroads of NLP and Information retrieval, and as such it shares a number of characteristics with other tasks such as information extraction and text-mining, computational linguistics, psychology and predicative analysis.

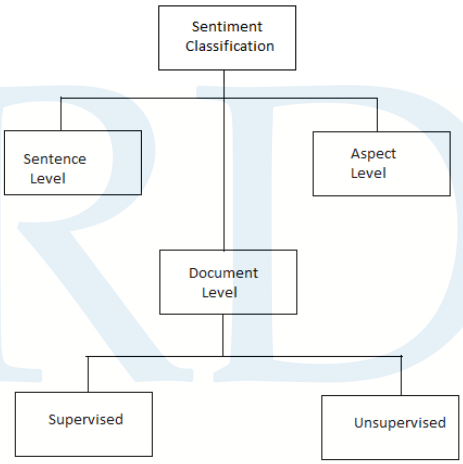

Fig. 2 Sentiment classification

Opinions can also be classified into three types: direct opinions, comparative opinions and indirect opinions. In direct opinion, opinion holder directly attack to target. Indirect opinions are either implied as in idioms or expressed in a reverse way as in sarcasm. In comparative, opinion holder generally compare among entity. In these studies, sentiment analysis is often conducted at one of the three levels: the document level, sentence level, or attribute level.

3.1 Document Level:

Document level sentiment classification aims to classify the entire document as positive or negative. There is much actual work use one of the two types of classification techniques which are a Supervised method and Unsupervised method to build level document sentiment.

3.1.1 Supervised method:

Sentiment classification is performed at document level. Sentiment classification can be used as a supervised classification problem with four classes positive, negative, neutral, and constructive. Also, supervised request machine-learning algorithms like SVM Support Vector Machines to conclude the relationships between the

opinions that expressed and text segment. A lot of researchers found that supervised learning techniques can perform well in SVM and Naïve Bayes .

One of the most fundamental tasks in sentiment classification is selecting an appropriate set of features. Some of the important features are:

Terms and their Frequency: These features are individual words (unigram) and their n-grams with associated frequency counts. They are also the most common features used in traditional topic-based text classification.

Part of speech: The part-of-speech (POS) of each word can be important too. Words of different parts of speech (POS) may be treated differently for example adjectives carry a great deal of information regarding a document's sentiment.

Sentiment words and phrase: Sentiment words or opinion words are words in a language that are used to express positive or negative sentiments. For example, good, awesome, and nice are positive sentiment words, and defective, poor, and risky are negative sentiment words.

Rules of opinions: Apart from sentiment words and phrases, there are also many other expressions or language compositions that can be used to express or imply sentiments and opinions.

Sentiment shifters: These are expressions that are used to change the sentiment orientations, e.g., from positive to negative and vice versa or from negative to constructive.

Negation words are the most important class of sentiment shifters. For example, the sentence "I don't like this smart phone" is negative.

Syntactic dependency: Words dependency-based features generated from parsing or dependency trees. These are also tried by researchers.

3.1.2 Unsupervised method:

Unsupervised classification is performed at the sentence level. There are two types of unsupervised classification, which are lexicon-based, and syntactic-pattern based. Sentence and aspect level sentiment classification for the lexicon-based can be used. Opinion or sentiment words and phrases are the dominating indicators for sentiment classification. Thus, using unsupervised learning based on such words and phrases would be quite natural.

Turney displayed a straightforward unsupervised learning calculation for characterizing a review as suggested or not suggested. He figured out whether words are positive or negative and how solid the assessment is by figuring the words' point wise mutual information (PMI) for their co-occurrence with a positive seed word ("excellent") and a negative seed word ("poor"). He called this value the word's semantic orientation. This technique checked through an audit searching for expressions that match certain grammatical feature designs (descriptive words and intensifiers), computed the semantic orientation of those phrases, and added up the semantic orientation of all of those phrases to compute the orientation of a review. He accomplished 74% accuracy classifying a corpus of item reviews.

3.2 Sentence Level:

It is one level of sentiment classification its work is to determine each sentence in the document as positive or negative opinions. Sentence level sentiment analysis has classified the polarity. This level is close to document

level but here it accomplished by every sentence. However, there may be complex sentences in the text which make the sentence level is not helpful. There are two phases in level sentence sentiment done in every single sentence: first, each sentence classified, as subjective or objective and the second one is the polarity of subjective sentence are concluded.

The task of classifying a sentence as subjective or objective is often called subjectivity classification. The resulting subjective sentences are also classified as expressing positive or negative opinions, which is called sentence-level sentiment classification. In the sentence level sentiment analysis, the polarity of each sentence is calculated. This is similar to a document level sentiment analysis but done at a sentence level [24]. It assumes each sentence contains an opinion for one entity and aspect, and some of the sentences may not be opinionated (objective). The subjective sentences contain opinion words which help in determining the sentiment about the entity. A two stage inference is done for each sentence: first, each sentence is classified as subjective or objective and then the polarity of each of the subjective sentences are inferred. There may be complex sentences also in the opinionated text. In such cases, sentence level sentiment classification is not useful.

3.3 Aspect Level:

It supposes that a document has a hold opinion on many entities and their aspects. Aspect level classification needs discovery of these entities, aspects, and sentiments for each of them.

IV. SENTIMENT ANALYSIS APPLICATION

Some of the applications of sentiment analysis includes online advertising, hotspot detection in forums etc. Online advertising has become one of the major revenue sources of today‟s Internet ecosystem. Sentiment analysis find its recent application in Dissatisfaction oriented online advertising Guang Qiu(2010) and Blogger-Centric Contextual Advertising (Teng-Kai Fan, Chia-Hui Chang ,2011), which refers to the assignment of personal ads to any blog page, chosen in according to bloggers‟ interests. When faced with tremendous amounts of online information from various online forums, information seekers usually find it very difficult to yield accurate information that is useful to them. This has motivated the research on identification of online forum hotspots, where useful information is quickly exposed to those seekers. Nan Li (2010) used sentiment analysis approach to provide a comprehensive and timely description of the interacting structural natural groupings of various forums, which will dynamically enable efficient detection of hotspot forums. In order to identify potential risks, it is important for companies to collect and analyze information about their competitors' products and plans. Sentiment analysis find a major role in competitive intelligence (Kaiquan Xu , 2011) to extract and visualize comparative relations between products from customer reviews, with the interdependencies among relations taken into consideration, to help enterprises discover potential risks and further design new products and marketing strategies. Opinion summarization summarizes opinions of articles by telling sentiment polarities, degree and the correlated events. With opinion summarization, a customer can easily see how the existing customers feel about a product, and the product manufacturer can get the reason why different stands people like it or what they complain about. Ku, Liang, and Chen (2006) investigated both news and web blog articles. Algorithms for opinion extraction at word, sentence and document level are proposed. The issue of relevant sentence selection is discussed, and then topical and opinionated information are summarized. Opinion summarizations are visualized by representative sentences. Finally, an opinionated curve

showing supportive and non-supportive degree along the timeline is illustrated by an opinion tracking system. Other applications includes online message sentiment filtering-mail sentiment classification, web blog author‟s attitude analysis etc. Review Seer is a tool that automates the work done by aggregation sites. Naive Bayes classifier is used with positive and negative review sets for assigning a score to the extracted feature terms. The classifier did not perform well for web pages crawled from the result of a search engine. It displays attributes and score of the attribute along with review sentences. Web Fountain uses beginning definite Base Noun Phrase (bBNP) heuristic for extracting product features. To assign sentiments to the features, reviews are parsed and traversed with two linguistic resources namely the sentiment lexicon and the sentiment pattern database. The sentiment lexicon defines the polarity of terms and sentiment pattern database defines sentiment extraction patterns for a sentence predicates (Yi and Niblack, 2005).

Red Opal is a tool that enables users to find products based on features.

V. **EVALUATION & DISCUSSION**

Opinion mining has a different methods and their performance is ebaluated by calculating various metrices like Precisions ,recall and F- measure. Precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved. 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. An overview of the work done in the task of Sentiment Analysis is shown in Table 1. This table represents a sample of work done and some works published on the topic of Sentiment Analysis. It is evident from the Table 1,as far as the data source is concerned, a lot of work has been done on movie and product reviews. Internet Movie Database (IMDb) is used for movie reviews and product reviews are downloaded from Amazon.com. Movie review mining is a more challenging application than many other types of review mining. The challenges of movie review mining lie in that factual information is always mixed with real-life review data and ironic words are used in writing movie reviews. Product review domain considerably differs from movie review domain because of two reasons. Firstly, there are feature specific comments in product reviews. People may like some features and dislike some others. Thus reviews consist of both positive and negative opinions, which make the task of classifying the review as positive or negative tougher. Such feature specific comments occur less frequently in movie reviews. Secondly, there are a lot of comparative sentences in product reviews and people talk about other products in reviews. This makes the task of opinion target detection an important aspect of the problem.

VI. **TOOLS USED FOR SENTIMENT CLASSIFICATION**

There are number of open-source text-analytics tools used for natural language processing such as information extraction and classification can also be applied for sentiment analysis. Tools are listed below:-

NTLK: The natural language toolkit is a tool for text processing, classification, tokenization, stemming, tagging, parsing etc. It provides easy-to-use interfaces to over 50 corpora and lexical resources.

GATE: Useful if you want to develop a pipeline. Language analysis modules for various languages are contributed by developers are available to be used plugged in your pipeline.

Open NLP: perform the most common NLP tasks, such as POS tagging, named entity extraction, chunking and co-reference resolution.

Stanford Core NLP: If you need part of speech categories, syntactic analysis (phrase structure or dependency analysis), co-reference or named entities in text.

Opinion Finder: It aims to identify subjective sentences and to mark various aspects of subjectivity in these sentences, including the opinion holder of the subjectivity and words that are included in phrases expressing positive or negative sentiments.

Ling Pipe: Ling Pipe is used for linguistic processing of text including, clustering classification and entity extraction etc.

VII. **METHODOLOGY**

Data collection :The data which is used in this paper is a set of product reviews collected from amazon.com. From February to April 2014, we collected, in total, over 5.1 millions of product reviews in which the products belong to 4 major categories: beauty, book, electronic, and home Those online reviews were posted by over 3.2 millions of reviewers (customers) towards 20,062 products. Each review includes the following information: 1) reviewer ID; 2) product ID; 3) rating; 4) time of the review; 5) helpfulness; 6) review text. Every rating is based on a 5-star scale.

Sentiment sentences extraction and POS : It is suggested by Pang and Lee that all objective content should be removed for sentiment analysis. Instead of removing objective content, in our study, all subjective content was extracted for future analysis. The subjective content consists of all sentiment sentences. A sentiment sentence is the one that contains, at least, one positive or negative word. All of the sentences were firstly tokenized into separated English words.

Negation phrases Identification Words: such as adjectives and verbs are able to convey opposite sentiment with the help of negative prefixes. For instance, consider the following sentence that was found in an electronic device's review: "The built in speaker also has its uses but so far nothing revolutionary." The word, "revolutionary" is a positive word according to the list . However, the phrase "nothing revolutionary" gives more or less negative feelings. Therefore, it is crucial to identify such phrases. In this work, there are two types of phrases have been identified, namely negation-of-adjective (NOA) and negation-of-verb (NOV).

Feature vector formation :Sentiment tokens and sentiment scores are information extracted from the original dataset. They are also known as features, which will be used for sentiment categorization. In order to train the classifiers, each entry of training data needs to be transformed to a vector that contains those features, namely a feature vector. For the sentence-level (review-level) categorization, a feature vector is formed based on a sentence (review). One challenge is to control each vector's dimensionality. The challenge is actually twofold: Firstly, a vector should not contain an abundant amount (thousands or hundreds) of features or values of a feature, because of the curse of dimensionality ; secondly, every vector should have the same number of dimensions, in order to fit the classifiers. This challenge particularly applies to sentiment tokens: On one hand, there are 11,478 word tokens as well as 3,023 phrase tokens; On the other hand, vectors cannot be formed by simply including the tokens appeared in a sentence (or a review), because different sentences (or reviews) tend to have different amount of tokens, leading to the consequence that the generated vectors are in different dimensions.

VIII. **CONCLUSION**

Sentiment analysis or opinion mining is a field of study that analyzes people's sentiments, attitudes, or emotions towards certain entities. This paper tackles a fundamental problem of sentiment analysis, sentiment polarity categorization. Sentiment Analysis still need to improve and progress. Moreover, there are many challenges like the polarity in a complex sentence. In addition, the vocabulary of natural languages is a lot which causes difficulty. This survey highlights the basic ideas about Sentiment Analysis and then explains in details the Sentiment Classification, Technique Classification, tools that available for Sentiment Analysis, and a new feature which is Product Aspect Ranking.

REFERENCES

[1] *T. Nasukawa, "Sentiment Analysis : Capturing Favorability Using Natural Language Processing Definition of Sentiment Expressions," pp. 70–77, 2003.*

[2] *K. Dave, I. Way, S. Lawrence, and D. M. Pennock, "Mining the Peanut Gallery : Opinion Extraction and Semantic Classification of Product Reviews," 2003.*

[3] *X. Ding, S. M. Street, B. Liu, S. M. Street, P. S. Yu, and S. M. Street, "A Holistic Lexicon-Based Approach to Opinion Mining," pp. 231–239, 2008.*

[4] *E. Marrese-Taylor, J. D. Velasquez, and F. Bravo-Marquez, "Opinion Zoom: A Modular Tool to Explore Tourism Opinions on the Web," 2013 IEEE/WIC/ACM Int. Jt. Conf. Web Intell. Intell. Agent Technol., pp. 261–264, Nov. 2013.*

[5] *J. Zhu, H. Wang, M. Zhu, B. Tsou and Matthew M, "Aspect-Based Opinion Polling from Customer Reviews", " Ieee Transaction On Affective Computing", vol. 2, NO. 1, January-March 2011.*

[6] *E. Haddi, X. Liu, and Y. Shi, "The Role of Text Pre-processing in Sentiment Analysis," Procedia Comput. Sci., vol. 17, pp. 26–32, Jan. 2013.*

[7] *D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," vol. 3, pp. 993–1022, 2003.*

[8] *T. Hofmann. P. latent, " semantic indexing," In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in inFmnation retrieval, SIGIR '99, pages 50-57, New York, NY, USA, 1999. ACM*

[9] *R. Moraes, J. F. Valiati, and W. P. Gavião Neto, "Document-level sentiment classification: An empirical comparison between SVM and ANN," Expert Syst. Appl., vol. 40, no. 2, pp. 621–633, Feb. 2013.*

[10] *R. Arora and S. Srinivasa, "A Faceted Characterization of the Opinion Mining Landscape," pp. 1–6, 2014.*