SENTIMENT ANALYSIS: A REVIEW

Nilu, Tanuja Shukla
M. tech CSE, Galgotias University

Abstract- Sentiment analysis is also known as Opinion mining. It is the field of analysing an individual's sentiments, feelings, opinions, assessments, perspectives and thoughts towards a particular object. It can be a product, service, individual, any organization, any current issue, any upcoming or recently happened event, any particular topic and their features. The area has become very fascinating area of research. There are many causes for this. First is that it has a no. of operations approximately in every area. Second is, it presents several imposing research problems, which haven't been considered earlier. These above and some other similar reasons bring a great encouragement for research. It takes the text, written in the digital form, as input. Languages that are mostly being studied are English and Chinese. At present, there are only few researchers who are performing research in this area. In this paper we will study about Sentiment analysis, steps to be followed in the sentiment analysis and its' levels.

Index Terms- Sentiments, Opinion, Polarity, Classification, Tools.

I. INTRODUCTION

Sentiment analysis is a technique that is used to fetch out the subjective information in text documents which is written in the digital format. In general, sentiment analysis determines the sentiment of a person about some aspect and also the overall concept of a document. The sentiment may be a person's personal opinion, judgment, mood or evaluation about any particular thing. A main problem in this field is the classification of sentiment, where a document is labelled as either a positive or a negative

polarity regarding a particular thing. The evaluation of opinion can be done in two ways:

• **Direct opinion:** In it, one can give positive or negative opinion about the object directly. For example, "The sound quality of this music system is poor" expresses a direct opinion.

• **Comparison:** It means, correlating the specific object with some other similar objects. For example, "The sound quality of Samsung mobile is better than Sony mobiles." expresses a comparison.

II. POLARITY OF OPINION OR REVIEW

Sentiment analysis checks the reviews that are available in the text form and creates output in as polarities i.e. – Negative, Positive or Neutral. It works for finding out the reasons of important variations in sales of products and later, they can be fixed accordingly. The algorithms that are being used in classifications effects the correctness of the results in polarity therefore, a minor mistake in classification can give a wrong output that result in an extreme outcome of a wrong business auditing approach. Sentiment analysis is a function that comprises extraction of information from customer's feedback and other legitimate sources like some organizations that perform surveys. As the word refers, it's a process of revealing sentiments of any person from the written text. There are large bunch of operations of this approach. This concept became very important since industry got reformed with the change in criteria of "Seller's Market" to "Buyer's Market" in order to take the market share.

III. MAJOR STEPS TO BE FOLLOWED IN SENTIMENT ANALYSIS

Pre-processing:

It is the process, in which raw data is taken, further it is pre-processed to extracting the features or attributes of a particular object or entity. This whole process can be divided into some parts in which following tasks will take place:

- Tokenization: The process, in which a collection of sentences, that are present in the form of text document, is divided into tokens by removing commas, blank spaces and other unnecessary logos etc.
- Stop word Removal: It eliminates articles i.e. a, an, the.

- Stemming: It reduces related tokens into a single unit.
- Case Normalization: An operation through which English text inputs can be presented in two forms i.e., higher and lowercase characters. It converts the whole document in the other form i.e., either from lowercase to uppercase or from uppercase to lowercase.

Feature Extraction: The feature extraction is a process that negotiates:

- Feature types: In it, recognition of the kind of features that are used for opinion mining, takes place.
- Feature selection: It is a process of choosing the good features for sentiment classification.
- Feature weighting mechanism: It measures each n every feature for better approval.
- Reduction mechanisms: It is a process by which one can perform the classification in an optimized way.

Feature Types:

Features that are selected for review analysis can be divided into following few types:

- Term frequency, which means, the existence of the word in a document maintains weight age.
- Term co-occurrence, which means, features which comes together.
- Part of speech (POS) information, in which, POS labels are used to separate POS tokens.
- Opinion words: Those words which express either good/positive or bad/negative emotions or thoughts.
- Negations: These terms, for e.g. no, never, not only, can transform the polarity or orientation of the sentiment in a sentence.
- Syntactic dependency: These are represented as a parse tree that contains features like word dependency based features. It mostly deals with the issues of grammar. Validity of it is dependent upon the parse trees. If the tree is formed in a right way the validity is ensured, else it won't be accepted as it will change the result of analysis.

Feature Selection:

- Information gain: In it, a verge is fixed and the terms with less information gain are eliminated on the basis of the existence of a particular term in a document.
- Odd Ratio: These are suitable for binary class domain where there are two types of classes are present for classification, one positive and one negative class. Features are classified accordingly.
- Document Frequency: It computes the frequency or the number of arrivals of a specific term in the accessible number of documents in the corpus and based on the verge computed, those terms are eliminated.

Features weighting mechanism:

The mechanism of feature weighting is of two types. They are as follows:

- Term Presence and Term Frequency: In it, term that appears hardly, encloses, more information than regularly occurring words.
- Term frequency and inverse document frequency: Records are ranked where top ranking is provided for terms that occur frequently in a few records and low ranking for the terms that occur regularly in each record.

IV. LEVELS OF SETIMENT ANALYSIS

Levels of Sentiment analysis can be divided in the following three levels:

- Document-level: It focuses to divide a document as indicating a bad/negative or good/positive opinion or feeling. It acknowledges the whole document as a single information unit.
- Sentence-level: It intends to categorize the sentiments that are being conveyed in each sentence. At first, one identifies whether the sentence is subjective or objective. If the sentence given is subjective, Sentence-level SA will determine whether the sentence indicates good or bad thought. However, there is not very much difference between the document and Sentence level classifications. The reason behind it is that the sentences can also be defined as small documents. Classifying

information at any of the above two level does not provide all necessary details that are required in many applications.

- Aspect-level: It aims to label the opinion with respect to the particular prospects of an individual. The first task that is to be done is to recognize the individuals and their aspects. The sentiment holders can have different thoughts for different features of the same object. For e.g.: - "The voice quality of this phone is bad, but the battery life is excellent''.

V. LITERATURE SURVEY

Sentiment analysis that is also known as Opinion mining is the process of studying about people's attitude, emotions or feelings toward an object. Objects can represent individuals, organization, events or topics. These topics are most likely to be covered by opinions. Generally, Sentiment analysis and Opinion mining are interchangeable. They express a mutual meaning. However, some researchers stated that Opinion mining and Sentiment analysis have lightly different notions [1]. It includes in building a system to detect and examine reviews about the product made in blog posts, comments, reviews or tweets. Sentiment analysis can be useful in

various ways. For e.g. - In Sales, it helps in predicting the success of any product's launch. There are several challenges in Sentiment analysis. First challenge is, an opinion word that is considered to be positive in one situation, may be considered negative in another situation. For these type of challenges one has to go through the aspect based approach. A second challenge is the way of expressing the opinion of people is not always the same. It can be different. Their response over a particular object can be different in different situations or scenario. Most of the traditional text processing depends on the fact that if there are minor differences between two text sentences, it won't affect the concept behind it, the meaning will remain same. In Sentiment analysis, however, "the picture was great" is very different from "the picture was not great". People can be contradictory in their statements [2]. Sentiment analysis engine performs mining from the textual reviews and generate output in the form of polarities or orientation i.e. Negative, Positive or Neutral [3]. This helps in making a decision about that particular entity. There are few algorithms that are used in classification of sentiments. When we talk about e-business, the algorithm in classification influences the correctness of the result and hence an incorrect classification can result in the wrong observation which will give a flawed business monitoring strategy [4].

VI. CLASSIFICATION IN SENTIMENT ANALYSIS

Classification is a stage in a sentiment analysis that can be described as a process in which an observation is assigned to a category or class. There are many possible classification technique, few of them are as follows:

Naïve Bayes classifier: The basic idea behind this algorithm is Bayes theorem. It assumes previously that the classes for the classification are not dependent, they are independent. In other words, pre-assumption is done. It is a nice approach when it is used with the real life application because it gives a good result even if the probability estimation is low. It computes the posterior probability of a class. The probability is calculated as follows:

$$P(\text{label} \mid \text{features}) = \frac{P(\text{label}) * P(\text{feature} \mid \text{label})}{P(\text{features})}$$

Where, P (label) can be defined as the prior probability of a label. P (label | feature) can be defined as the prior probability that a provided feature set is being categorized as a label [6].

Max Entropy classifier: The basic idea of the MaxEnt classifiers is that the most uniform model should be preferred that satisfies any given constraint. It changes the labelled feature set to vector using the process of encoding. The models that are being used are feature based models. These features are used to find a distribution over the various classes using linguistic regression. The probability is computed as follows:

$$P(c \mid d, \lambda') = \frac{\exp\left[\sum_i \lambda_i f_i(c, d)\right]}{\sum_{c'} \exp\left[\sum_i \lambda_i f_i(c', d)\right]}$$

Where, c is the class, d is the data point and λ is a vector that is treated as weight vector [7].

The **advantages** of using MaxEnt are as follows:

i. **Accuracy**: Correctness of result is more than other algorithms.

ii. **Consistency**: It maintains the firmness of character. It maintains the durability of changes.

iii. **Performance/ Efficiency**: It can handle a large amount of data.

iv. **Flexibility**: It can deal with many kinds of data in a unified platform and classify it [8].

Boosted Trees classifier: It is a classifier that is basically a combination of Boosting and Decision Trees, or, in other words, we can define it as a merger of Decision trees and Boosted trees. In Boosting, weighted trees are created using predictive classifier. Further, these trees are merged into a single prediction model in the final phase. Boosted tree generally merges the fortitude of two algorithms: Regression tree and Boosting. Regression trees are the variants of decision trees which may have single or multiple inputs that results in single output. Boosting is a method in which no. of simple models are added to give an improved anticipated performance [9].

Random forest classifier: Random forests are the classification methods that operate by building a multitude of decision trees at training time and outputting the class that is the mode of the classes output by the individual trees. It creates multi-altitude decision trees at initial phase where input is required and at the end of the process, output is generated in the form of multiple decision trees. [10].

Advantages of Random Forest algorithm are:

- Quick and scalable.
- Easy to depict and grasp
- No need of parameters in input dataset.
- No need of snipping the trees.
- Robust to unneeded text mentioned in document.

VII. TOOLS FOR SENTIMENT ANALYSIS

There are number of tools that can be used to perform the analysis of tracking the opinion or polarity which are provided or expressed by the user. Few of them are as follows:

- **Review Seer tool**– These tools are the tools that are used to automate the work done by aggregation sites. Aggregation means to collect or to summarize. By using this tool one can give result of analysis in the summarized form. The Naive Bayes classifier approach is used to collect positive and negative opinions for assigning a score to the extracted feature terms and final results are presented in the form of simple opinion sentence. This is a very useful tool when there is a need of collective data.
- **Web Fountain**- These are the tools that follow the approach of beginning definite Base Noun Phrase (bBNP). It is a heuristic approach for extracting the product features. By using this tool it is also possible to develop a simple web interface. This is also a very useful tool when one has to deal with selecting a particular product from the bunch of similar products with various attributes or features because it is a tool that focuses mainly on extracting the features of any product.
- **Red Opal–** These tools allow the users to decide the direction or polarities of the sentiments or thoughts regarding any product based on their attributes. It allocates the weights to each product based on attributes taken from the customer reviews or feedbacks. The output is represented with the help of a web based interface.
- **Opinion observer-** It is a system that performs mining for the analysis of the opinions that are later compared on the Internet. The inputs for this process are the contents that are generated by users by providing their reviews or opinion. The output generated by this system gives result in the form of graphs that shows the reviews about any product feature by feature. Word-Net Exploring method is used to allocate prior polarity.

VIII. CONCLUSION

Sentiment analysis has become a wide area of research now-a-days. It has number of applications in information science, like analysis and classification of review, sentiment/opinion summarization, etc. There are various techniques available using which we can classify a review or an opinion. It plays an important role when used in

e-business. E-commerce of any product also depends upon it. In this paper we have studied about Sentiment analysis (SA), levels of SA, few technique of classification of sentiments. According to analysis of these methods, we can conclude that if we need an output in which the accuracy plays an important role then our first choice should be Random forest. Even if it takes high learning time, it gives most accurate analysis result. If one has a strong processor and enough memory but shortage of training time then in that case, Maximum Entropy is a best technique. If the problem is with memory and processing power then the best method will be Naïve Bayes classifier because it consumes low memory and low processing power. If we require a method that is average in all manners then we should select Boosted tree classifier. There are lots of more techniques available that we will be exploring in next survey. In the upcoming time, lots of work is required to be done to further improve the techniques of classification. Various kinds of features and classification algorithms are merged together. It helps them in overcoming their individual disadvantages and to get advantages from each-other's merits. It results in enhancing the performance of the algorithm by producing a better output or result. In near future, the focus should be on further improving the performance of the algorithms. Although the field is progressing at very high speed but there are still so many problems that are not being solved at present. More future research could be focused to these areas. The interest in languages other than English in this field is growing due to lack of resources and researches showing less interest in these languages. This work will surely make a grate change in the field of sentiment analysis. Now-a-days there are only few languages are present in which one can perform its analysis. Data available on social network sites and micro-blogging sites still needs deeper analysis.

REFERENCES

[1]. Tsytsarau Mikalai, Palpanas Themis, "The Survey on mining subjective data on the web". Data Mining Knowledge Discovery 2012; 24:478–514.

[2].G. Vinodhini, RM. Chandrasekaran. Sentiment analysis and Opinion mining: A Survey. June 2012; ISSN: 2277 128X.

[3]. Amit Gupte et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (5), 2014, 6261-6264.

[4]. Bo Pang and Lillian Lee "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts" in ACL '04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, 2004, Article No. 271

[5]. Sasha Blair- Goldensohn, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A. Reis, Jeff Reynar. Building a Sentiment Summarizer for local service Reviews, 2008.

[6]. Kang Hanhoon, Yoo Seong Joon, Han Dongil. Senti - lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. Expert Syst Appl 2012; 39:6000–10.

[7]. Y. M. C. S. Kostas Fragos, "A Weighted Maximum Entropy Language Model for Text Classification," Natural Language Understanding and Cognitive Science, no. NCLUS May 2005.

[8]. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer.

[9]. Jerome Friedman, Trevor Hastie, Robert Tibshirani, "Additive Logistic Regression: A statistical view of Boosting", The Annals of Statistics 2000, Vol 28, No.2, 337-407.

[10]. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, "An Introduction to Statistical Learning: with Applications in R", Springer Texts in Statistics, 2013.