A Review Paper on Sentiment Analysis

Miss. Prathana Rajendra Patil
Student, Department of Computer Engineering
K.J.Somaiya Institute of Engineering and Information Technology.
Mumbai, India
prathana.patil@somaiya.edu

Mrs. Nisha Vanjari
Assistant Professor, Department of Computer Engineering
K.J.Somaiya Institute of Engineering and Information Technology.
Mumbai, India.
nvanjari@somaiya.edu

Abstract - Sentiment Analysis can be outlined as "automated extraction of subjective content from digital text and predicting the sound judgment such as positive or negative". Sentiment Analysis is conjointly known as Opinion Mining. The different approaches involved in the sentiment analysis are product review, sentiment identification, feature selection, sentiment classification and sentiment polarity. Though after the analysis of sentiment there are certain open problem, such as the data problem, the language problem and the NLP.

Keywords – sentiment analysis, feature selection, sentiment identification, sentiment classification, sentiment polarity.

I. INTRODUCTION

Sentiment Analysis is the method of determinant whether or not a piece of writing is positive, negative, or neutral. It is additionally notable as opinion mining, account the opinion or angle of a speaker. A typical use of this technology is to discover however folks feel a couple of specific topic.

For example: If you wish to recognize that if folks on Twitter suppose that the Chinese food in point of entry is nice or unhealthy. Twitter Sentiment Analysis can answer this question. You will even learn why folks suppose the food is nice or unhealthy, by extracting the precise words that indicate why folks did or didn't like the food. If "too salty" shows as a common theme, for example, you instantly have a much better plan of why customers aren't sad.

II. LITERATURE SURVEY

Sentiment analysis is one of the quickest growing analysis areas in applied science, making it challenging to keep track of all the activities within the space. We have a tendency to gift a pc - assisted literature review, wherever we have a tendency to utilize each text mining and qualitative writing, and analyse 6,996 papers from Scopus. We discover that the roots of sentiment analysis are in the studies on vox populi analysis at the beginning of twentieth century and in the text sound judgment analysis performed by the computational linguistics Community in1990's. In recent years, sentiment analysis has shifted from analyzing online product reviews to social media texts from Twitter and Facebook. Many topics on the far side product reviews like stock markets, elections, disasters, medicine, software Engineering And cyberbullying extend the Utilization of sentiment analysis.Sentiment Analysis (SA) is associate degree in progress field of analysis in text mining field. Reserves is that the procedure treatment of opinions, sentiments and perspicacity of text. Several recently projected algorithms' enhancements and numerous reserves applications square measure investigated and given shortly during this survey. The connected fields to reserves (transfer learning, feeling detection, and building resources) that attracted researchers recently square measure mentioned.

III. STEPS INVOLVED IN SENTIMENT ANALYSIS

Sentiment Analysis (SA) or Opinion Mining (OM) is that the process study of people's opinions, attitudes associated emotions toward an entity. The entity will represent people, events or topics. These topics square measure possibly to be coated by reviews. The 2 expressions Storm Troops or OM square measure interchangeable. They specific a mutual which means. However, some researchers expressed that Opinion Mining and Storm Troops have slightly totally different notions[1].OM extracts associated analyses people's opinion concerning an entity whereas Sentiment Analysis identifies the sentiment expressed in a very text then analyses it. Therefore, the target of Storm Troops is to search out opinions, determine the feelings they specific, so classify their polarity as shown in figure (1).

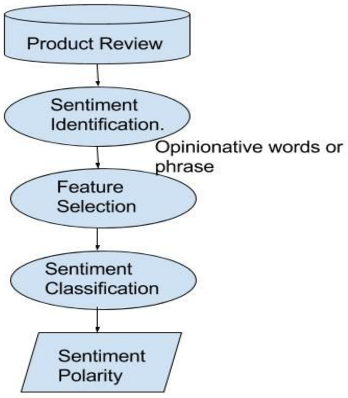

Figure (1): Sentiment analysis process on product reviews.

There are 3 main classification levels in SA: document-level, sentence-level, and aspect level militia.Document-level militia aims to classify Associate in nursing opinion document as expressing a positive or negative opinion or sentiment. It considers the full document a basic data unit (talking regarding one topic). Sentence-level militia aims to classify sentiment expressed in every sentence. The primary step is to spot whether or not the sentence is subjective or objective. If the sentence is subjective, Sentence-level militia can confirm whether or not the sentence expresses positive or negative opinions. Wilson et al.[2] have realized that sentiment expressions don't seem to be essentially subjective in nature. However, there's no elementary distinction between document and sentence level classifications as a result of sentences are simply short documents [3]. Classifying text at the document level or at the sentence level doesn't give the mandatory detail required opinions on all aspects of the entity that is required in several applications, to get these details; we'd like to travel to the side level. Aspect-level militia aims to classify the sentiment with regard to the precise aspects of entities. The primary step is to spot the entities and their aspects. The opinion holders will offer totally different opinions for various aspects of a similar entity like this sentence "The voice quality of this phone isn't sensible, however the battery life is long".

The data sets utilized in militia are a vital issue during this field. The most sources of information are from the merchandise reviews. These reviews are vital to the business holders as they'll take business selections per the analysis results of users' opinions regarding their merchandise. The reviews sources are principally review sites. Militia isn't solely applied on product reviews however can even be applied on stock markets [4, 5], news articles, [6] or political debates [7]. In political debates for instance, we tend to may fathom people's opinions on a particular election candidates or political parties. The election results can even be foreseen from political posts. The social network sites and micro-blogging sites are thought-about a really sensible supply of data as a result of folks share and discuss their opinions a couple of sure topic freely. They're conjointly used as knowledge sources within the militia method.

There are several applications and enhancements on militia algorithms that were projected within the previous few years. This survey aims to provide a more in-depth look on these enhancements and to summarize and reason some articles bestowed during this field per the varied militia techniques. The Sentiment Classification (SC) techniques, as below:

1. *Machine Learning Approach*
    a. Supervised Learning
        i. Decision Tree classifiers.
        ii. Linear classifiers
            1. Support Vector Machines.
            2. Neural Networks.
        iii. Rule-Based classifiers.
        iv. Probalistics-Based classifiers.
            1. Naive Bayes.
            2. Bayesian Network.
            3. Maximum Entropy.
    b. Unsupervised Learning.
2. *Lexicon-based Approach*
    a. Dictionary-based Approach
    b. Corpus-based Approach
        i. Statistical
        ii. Semantic

IV. FEATURE SELECTION MECHANISM FOR SENTIMENT ANALYSIS

Sentiment Analysis task is taken into account a sentiment classification downside. the primary step within the SC downside is to extract and choose text options. a number of this options are[8]

1. *Terms presence and frequency***:** These options area unit individual words or word n-grams and their frequency counts. It either offers the words binary weight (zero if the word seems, or one if otherwise) or uses term frequency weights to point the relative importance of options [9].

2. *Parts of speech (POS):* Finding adjectives, as they're vital indicators of opinions.

3. *Opinion words and phrases:* These area unit words normally wont to specific opinions as well as smart or unhealthy, like or hate. On the opposite hand, some phrases specific opinions while not victimization opinion words. For example: value ME Associate in Nursing arm and a leg.

4. *Negations:* The looks of negative words might amendment the opinion orientation like not smart is akin to unhealthy.

Feature selection method:

Feature choice strategies may be divided into lexicon-based strategies that require human annotation, and applied math strategies that are automatic strategies that ar additional oftentimes used. Lexicon-based approaches sometimes begin with atiny low set of 'seed' words. The feature choice techniques treat the documents either as cluster of words (Bag of Words (BOWs)), or as a string that retains the sequence of words within the document. BOW is employed additional actually because of its simplicity for the classification method. the foremost common feature choice step is that the removal of stop-words and stemming (returning the word to its stem or root i.e. flies → fly).

1. *Pointwise Mutual Information (PMI)*

The mutual system of measurement provides a proper thanks to model the mutual data between the options and therefore the categories. This live was derived from the data theory[65]. The pointwise mutual data (PMI) Mi(w) between the word w and therefore the category i is outlined on the idea of the extent of co-occurrence between the category i and word w. The expected co-occurrence of sophistication i and word w, on the idea of mutual independence, is given by Pi · F(w), and therefore the true co-occurrence is given by F(w) · pi(w).The mutual data is outlined in terms of the magnitude relation between these 2 values and is given by the subsequent equation:

$$M_i(w) = log(\frac{F(w).p_i(w)}{F(w).P_i}) = log(\frac{p_i(w)}{P_i})$$

The word w is completely correlate to the category i, once Mi(w) is bigger than zero. The word w is negatively correlate to the category i once Mi(w) is a smaller amount than zero.Yu and Chinese [4] have extended the essential PMI by developing a discourse entropy model to expand a group of seed words generated from a little corpus of securities market news articles.

2. *Chi-square (χ2)*

Let n be the entire range of documents within the assortment, pi(w) be the contingent probability of sophistication i for documents that contain w, Pi be the worldwide fraction of documents containing the category i, and F(w) be the worldwide fraction of documents that contain the word w. Therefore, the χ2-statistic of the word between word w and sophistication i is outlined as

$$X_i^2 = \frac{n.F(w)^2.(p_i(w)-P_i)^2}{F(w).(1-F(w)).P_i.(1-P_i)}$$

3. *Latent Semantic Indexing (LSI)*

Feature choice ways arrange to scale back the spatial property of the information by choosing from the first set of attributes. Feature transformation ways produce a smaller set of options as a perform of the first set of options. LSI is one among the far-famed feature transformation ways [12]. LSI methodology transforms the text house to a brand newaxis system that could be a linear combination of the first word options. Principal element Analysis techniques (PCA) area unit wont to reach this goal [13]. It determines the axis-system that retains the best level of data regarding the variations within the underlying attribute values. the most disadvantage of LSI is that it's AN unsupervised technique that is blind to the underlying class-distribution. Therefore, the options found by LSI aren't essentially the directions on that the class-distribution of the underlying documents is best separated [8].

V. OPEN PROBLEMS IN SENTIMENT ANALYSIS

Various problems associated with sentiment analysis are as follows:

1. *The Data Problem:*

It's been noticed that there's lack of benchmark information sets during this field. It had been expressed in [1] that few of the foremost far-famed information sets area unit within the field of SA.

2. *The Language problem:*

It had been noticed within the articles given during this survey that the Far East languages particularly the Chinese language has been used additional usually recently. Consequently, several sources of knowledge area unit engineered for these languages. The researcher's area unit currently within the section of building resources of different Latin (European) languages.

3. *NLP (Natural Language Processing):*

The tongue process tools are often accustomed facilitate the SA method. It provides higher tongue understanding and therefore will facilitate manufacture additional correct results of SA. These tools were accustomed facilitate in BR, male erectile dysfunction and additionally SA task within the last 2 years.

VI. CONCLUSION

Thus in this review paper we have studied different approaches for the sentiment analysis. Also we have reviewed the different sentiment classification approaches. After the sentiments have been classified different feature selection techniques are used such pointwise mutual Information, Chi-square, and Latent Semantic Indexing. At last we have also reviewed the different problem areas for the sentiment analysis.

REFERENCES

[1] Mikalai Tsytsarau, Themis PalpanasSurvey on mining subjective data on the web Data Min Knowl Discov, 24 (2012), pp. 478-514

[2] Wilson T, Wiebe J, Hoffman P. Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of HLT/EMNLP; 2005.

[3] B. Liu Sentiment analysis and opinion mining Synth Lect Human Lang Technol (2012)

[4] Liang-Chih Yu, Jheng-Long Wu, Pei-Chann Chang, Hsuan-Shou ChuUsing a contextual entropy model to expand emotion words and Michael Hagenau, Michael Liebmann, Dirk Neumann. Automated news reading: stock price prediction based on financial news using context-capturing features. DecisSupp Syst; 2013.their intensity for the sentiment classification of stock market news
Knowl-Based Syst, 41 (2013), pp. 89-97

[5] Michael Hagenau, Michael Liebmann, Dirk Neumann. Automated news reading: stock price prediction based on financial news using context-capturing features. Decis Supp Syst; 2013.

[6] Xu Tao, Qinke Peng, Yinzhao Cheng Identifying the semantic orientation of terms using S-HAL for sentiment analysis
Knowl-Based Syst, 35 (2012), pp. 279-289

[7] Isa Maks, Piek Vossen A lexicon model for deep sentiment analysis and opinion mining applications
Decis Support Syst, 53 (2012), pp. 680-688

[8] Aggarwal Charu C, Zhai Cheng Xiang. Mining Text Data. Springer New York Dordrecht Heidelberg London:© Springer Science+Business Media, LLC'12; 2012.

[9] Yelena Mejova, Padmini Srinivasan. Exploring feature definition and selection for sentiment classifiers. In: Proceedings of the fifth international AAAI conference on weblogs and social media; 2011.

[10] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, R. Harshman Indexing by latent semantic analysis
JASIS, 41 (1990), pp. 391-407

[11] I.T. Jolliffee Principal component analysis
Springer (2002)

[12] R. Feldman Techniques and applications for sentiment analysis
Commun ACM, 56 (2013), pp. 82-89

[13] Andrés Montoyo, Patricio Martínez-Barco, Alexandra Balahur Subjectivity and sentiment analysis: an overview of the current state of the area and envisaged developments
Decis Support Syst, 53 (2012), pp. 675-679