*International Journal of Engineering Applied Sciences and Technology, 2017*
**Vol. 2, Issue 5, ISSN No. 2455-2143, Pages 234-237**
Published Online March-April 2017 in IJEAST (http://www.ijeast.com)

# DATA MINING TECHNIQUE ON SENTIMENT ANALYSIS AND COMPUTATION OF VIEWS

Jyoti Rusia
Amity institute of Biotechnology
Amity University Uttar Pradesh

**Abstract— In the current scenario of social media, opinion mining shows a remarkable impact on information retrieval and web data analysis. This new research domain becomes important as the use of social media has increased to next fold. Users here generate the content, which is the form of emotions, comments that can be positive or negative, an individual's own view point etc. Using the social networking sites (e.g. Facebook, twitter, multi-media sharing sites (e.g. YouTube, Flickr), blogs and rich web applications as the usage of Web 2.0 increases, user can exchange or share their opinion. In this paper a systematic literature review is conducted that thoroughly discuss the commonly used classification techniques to assist future research in this new emerging area. These techniques are used for Opinion**

**Keywords— Social Networks, Opinion Mining, Sentiment Analysis, Reviews, knowledge discovery**

## I. INTRODUCTION

Sentiment Analysis and Opinion mining is indeed a challenging job. It combines Data Mining and intelligent computing in order to find subjectivity in big data. The word sentiment analysis was first time used by [1], while opinion mining first time appeared in [2]. This area of research has become important as the usage of social media such as Facebook®, Twitter®, Linked-in®, mobile social application, is accelerating at a very high pace. Users here users produce the contents on the internet. This content is in the form of comments, opinions (positive or negative), emotions, review etc. Various applications are developed for this purpose but in opinion mining and sentiment analysis main task is Opinion polarity classification (OPC) [3][4]. We can broadly classify the approaches to polarity classification into three types:

- Unsupervised Approach: It doesn't require prior training but deals with the arrangement of the words. This arrangement may be positive or negative.
- Semi-Supervised Approach: In this, with unlabeled data some supervision data is provided to perform classification
- Supervised Approach (SA): It is a Machine Based Learning Method (MBLM). To train the classifiers, it uses an assembly of data.

In this paper we have presented a critic review of the work done by the researchers of this domain. Initially we discussed the growth of different opinion mining and sentiment analysis systems based on Graph Theory and Artificial Intelligence.

## II. RELATED WORK

There are numerous critical examination endeavors on assessment targets/words extraction (sentence level and corpus level). In sentence level extraction, past techniques principally intended to recognize all conclusion target/word notice in sentences. They viewed it as an arrangement naming errand, where a few established models were utilized, for example, CRFs and SVM. A large portion of past corpus-level systems received a co-extraction structure, where conclusion targets and feeling words strengthen one another as indicated by their assessment relations. Subsequently, how to enhance sentiment relations distinguishing proof execution was their principle center. Abused closest neighbor principles to mine supposition relations among words. What's more, (Qiu et al., 2011) composed syntactic examples to perform this assignment. They embraced some uncommon composed examples to increment review. (Liu et al., 2012; Liu et al., 2013a; Liu et al., 2013b) utilized word arrangement model to catch feeling relations instead of syntactic parsing. The test results demonstrated that these arrangement based systems are more viable than sentence structure based methodologies for online casual writings. Then again, all previously stated techniques just utilized assessment relations for the extraction, however overlook considering semantic relations among homogeneous competitors. In addition, they all disregarded word inclination in the extraction process.

As far as considering semantic relations among words, our technique is connected with a few methodologies in light of point model (Zhao et al., 2010; Moghaddam and Ester, 2011; Moghaddam and Ester, 2012a.The fundamental objectives of these systems weren't to concentrate supposition targets/words, yet to arrange all given perspective terms and assessment words. In spite of the fact that these models could be utilized for our undertaking as indicated by the relationship in the middle of applicants and subjects, exclusively utilizing semantic relations is still uneven and inadequate to acquire expected execution.

Moreover, there is little work which considered these two sorts of relations universally (Su et al., 2008; Hai et al., 2012; Bross and Ehrig, 2013). They normally caught diverse relations utilizing cooccurrence data. That was excessively coarse, making it impossible to acquire expected results (Liu et al., 2012). What's more, (Hai et al., 2012) removed assessment targets/words in a bootstrapping procedure, which had a mistake spread issue. Interestingly, we perform extraction with a worldwide chart co-positioning procedure, where blunder proliferation can be successfully lightened. (Su et al., 2008) utilized heterogeneous relations to discover certain opinion relationship among words.

## III. RESEARCH METHODOLOGY

For performing the opinion mining and sentiment classification literature review we searched various online reading resources like:

- IEEE
- Google Scholar
- KSII Transactions on Internet and Information Systems
- Elsevier (Journal: Expert System with Application)
- ACM
- Springer
- Morgan & Claypool (Synthesis Lectures on Human Language Technologies) [4]

The importance of this research domain can also be determined by the exponential increase of the number of research articles published in different journals and divided in positive, negative, objective categories, here objective means no sentiment.

### *Sentence level*

This task goes to the sentence level. It checks out whether the opinion of a sentence is a *positive, negative, or neutral* one. This level of analysis is closely related to *subjectivity classification* [18].

### *Document Level*

This level classifies whether the whole opinion of the document in general showcases a positive or a negative sentiment [10]. This analysis takes the assumption that a documents showcases opinions or views on a single quantity. Hence, it is not applicable where the document uses multiple quantities.

### *Entity and Aspect level*

This level of sentiment analysis is based on the features [26] of an entity, therefore it is also known as *feature based opinion mining and summarization* [13] [20]. It analyzes the sentiments at a much finer level and helps to identify the objects which people exactly like the most. It also helps to perform comparative study of the multiple entities based upon their attributes. In [60] author try to explain about the contradictions in text based on the features.

Further opinions can also be classified into *regular opinions* and *comparative opinions [3]* Fig. 1.

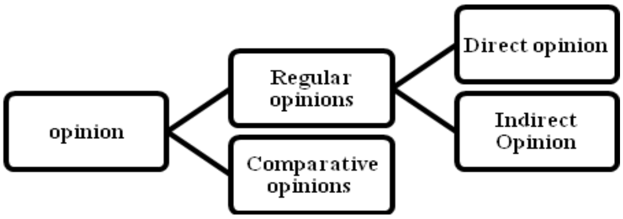

Fig. 1. Basic Opinion Classification

### *Regular opinion*

A *regular opinion* in literature is divided into two main sub-types:

**Direct opinion:** An opinion which is directed towards an entity or a specific aspect of the entity.

**Indirect opinion:** An opinion is showcased indirectly to a quantity or attribute of the quantity.

### *Comparative opinion*

A comparative opinion discovers the similarities and differences between the entities based upon the shared attributes of the entities and identify the inclination of the opinion holder [3]. It includes both negative and positive [83] opinion about the entity as the inclination of the opinion holder shows positive opinion regarding one entity and negative for the other *e.g. the camera quality of phone A* ***is better than*** *that of phone B.*

## IV. APPLICATIONS OF SENTIMENTAL ANALYSIS

When consumers have to make a decision or a choice regarding a product, an important info is the standing of that creation, which is derived from the opinion of others. Sentiment analysis can reveal what other people think about a product. The first application of sentiment analysis is thus giving indication & commendation in the choice of products permitting to the wisdom of the crowd. When you choose a product, you are generally attracted to certain specific aspects of the product. A single global rating could be deceiving. Sentiment examination can regroup the sentiments of the reviewers & estimate ratings on certain aspects of the product. Another utility of sentimentality analysis is for corporations that want to distinguish the opinion of customers on their products. They can then progress the aspects that the clienteles found indecisive. Sentiment analysis can also determine which aspects are more significant for the consumers. Finally, sentiment exploration has been proposed as a component of other technologies. An idea is to develop information removal in text analysis by excluding the most subjective section of a document or to automatically suggest internet ads for products that fit the viewer's opinion (and removing the others). Knowing what persons think gives numerous potentials in the Human/Mechanism interface domain. Sentiment analysis for determining the opinion of a customer on a product (and consequently the reputation of the product) is the main focus of this paper. In the succeeding section, we will deliberate solutions that allow determining the expressed opinion on products.

.

## V. FEATURE EXTRACTION IN SENTIMENTAL ANALYSIS

Text Analysis is a main application field for mechanism learning processes. However the raw information, an order of symbols cannot be fed straight to the algorithms themselves as maximum of them expect arithmetical feature paths with a fixed size somewhat than the raw text forms with variable length. In imperative to address this, sickie-learn offers utilities for the most mutual ways to extract numerical structures from text content, namely:

- Tokenizing strings and giving an integer id for each imaginable token, for example by using white-spaces & punctuation as symbolic separators.
- Counting the existences of tokens in each document.
- Regulating and weighting with diminishing importance tokens that occur in the majority of samples / forms.

In this arrangement, topographies and samples are defined as follows:

- Each individual token incidence frequency (regularized or not) is preserved as a feature.
- The direction of totally the token frequencies for a given article is considered a multivariate sample.

A quantity of documents can thus be signified by a matrix with one row per manuscript and one support per token (e.g. word) occurring in the corpus.

## VI. EXISTING PROBLEMS

Assume a message, categorize whether the communication is of positive, negative, or neutral sentiment. For messages transmission both a positive and negative sentiment, either is the stronger sentimentality should be selected. Today, knowledgeable by the thoughts and exhibitions at the recent Sentiment Analysis Symposium, let's observe the business case for sentiment analysis, as well as some problems related to the detection and analysis processes applied to those forms to mine actionable information for businesses.

In this segment, they discover the problem declarations related to Sentiment analysis. They start with problem of very basic environment and finished with some unsolved problems.

## VII. EXISTING APPROACH

Support vector machines were proposed by Boser et al. in. SVM is supervised machine learning approach specifically designed for pattern matching. SVMs construct a set of hyper-planes that separates the data points into two classes with maximal margin in high dimensional feature space. Mathematically, SVM learns a mapping $\chi \rightarrow \Upsilon$ where x € $\chi$ represents the feature vector & y € $\Upsilon$ represents scene category. The objective is to learn a function defined below, with function parameter α, given a set of training images

$$(x1, y1), (x2, y2)\ldots(xn,yn).$$

$$y=f(x,\alpha)$$

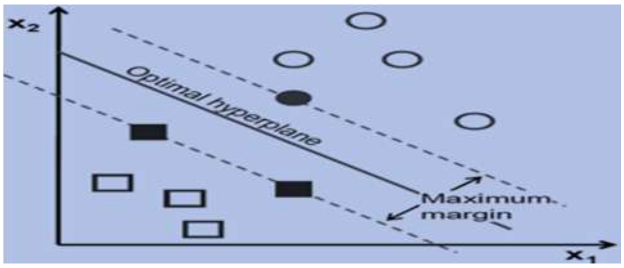

Fig.2. SVM

Once the mapping function is learnt, the classifier can render a category label for unobserved feature vector.[14] According to the linear reparability of data-points, SVM can be linear or non-linear. Non-linear SVM is an extension of linear SVM. It maps the non-linearly separable points to higher dimensional plane in which these can be linearly separable. Linear SVM- In case of non-linear SVM, the hyper plane is defined as given below

$x_i.w+b \geq +1$, if $y_i=+1$

$x_i .w+b \leq +1$, if $y_i= -1$

*where $x_i$denotes observed data points, w is normal vector, b is offset of hyper plane w.r.t. to origin & $y_i$is target value.*

Supports vectors are those data points that lie exactly on hyper plane & satisfy following equations:

$$x_i .w+b= +1$$
$$x_i .w+b= -1$$

This linear SVM is used for binary classification. For multiclass classification multiple one-vs-all linear SVMs can be used.

So this paper, different partition based clustering techniques k-means algorithms have been applied to a sample Twitter dataset and the experimental analysis of the algorithms, how they works with twitter data and the tiny differences found between the two methods has been explained with statistic outputs. For future work, focus will be more ways to understand relation between sentiment and behavior. Some near-future directions are listed below:

- Features, in the current model the simple bag-of-words model is used, but other more meaningful feature models could be used, such as a distance model where sentiment from SentiWordNet is considered for each word
- Size of dataset, as the used Twitter dataset is not very large, building much larger topic-wise and annotated datasets from Twitter and Facebook is a research direction. • Real-time application, as the model matures, testing it on live Twitter or Facebook data is interesting. This direction
- Real-time application, as the model matures, testing it on live Twitter or Facebook data is interesting. This direction requires adaption of our model for real-time usage.

## VIII. CONCLUSION AND FUTURE WORK

In the Future Work we will use semantic-based approach for multi source bioinformatics data integration. In our approach, a metamodel is utilized to represent the master search schema, and an effective interface extraction algorithm based on the hierarchical structure of the web and pattern is developed to capture the rich semantic relationships of the online bioinformatics data sources. Our final goal is to develop a meta-search interface for biologists as a single point of access to multiple online bioinformatics databases. In text mining, some of the challenging issues in mining and searching the biomedical literature are addressed, and I will present a unified architecture Bio-SET-DM (Biomedical Literature]Searching, Extraction and Text Data Mining), discuss some novel algorithms such as semantic-based language model for literature retrieval, semi-supervised pattern learning for information extraction of biological relationships from biomedical literature. In the third part, graph-based data mining, the focus is on graph-based mining in biological networks. In the Future Work We will discuss how to apply graph-based mining techniques and algorithms in the analysis of modular and hierarchical structure of biological networks, how to identify and evaluate the sub networks from complicated biological networks, and present the experimental result.

## IX. REFRENCES

1. Yrd.Doç.Dr. Ayça ÇAKMAK PEHLİVANLI, the comparsion of data mining tools, 16 November 2011
2. Kalpana Rangra and Dr. K. L. Bansal,Comparative Study of Data Mining Tools,Volume 4, Issue 6, June 2014
3. A. Jović*, K. Brkić* and N. Bogunović, An overview of free software tools for general data mining, MIPRO 2014, 26-30 May 2014, Opatija, Croatia
4. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: an update,” SIGKDD Explorations, vol. 11, no. 1, pp. 10–18, 2009.
5. Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining: concepts and techniques: concepts and techniques. Elsevier, 2011.
6. Gonzalez, Hector, Jiawei Han, and Xiaolei Li. "Mining compressed commodity workflows from massive RFID data sets." Proceedings of the 15th ACM international conference on Information and knowledge management. ACM, 2006.
7. Gao, Kun, Qin Wang, and Lifeng Xi. "Controlling Moving Object in the Internet of Things." International Journal of A