www.ijecs.in

International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 11 November 2017, Page No. 22854-22857
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i11.05

Review Paper on Text mining and Sentimental analysis

Jyoti Sharma[1,] Gurvinder Singh[2,] Ankur Sharma[3]

[1]Arni University Dept. of Computer Science Engineering
Kangra Himachal Pradesh, India
[2]Arni University Dept. of Computer Science Engineering
Kangra Himachal Pradesh, India
[3]Arni University Dept. of Computer Science Engineering
Kangra Himachal Pradesh, India

Abstract:
In this paper we discussed about most emerging field of computer science engineering name as Big Data. This paper provides you detailed information about big data from basic to advance level. In this paper we presented detailed review of the sentimental analysis and text analysis which will provide you deep insight of the work done in this and provides outline about the scope in this area and motivate research work in this and apply it on the real life application. It is solitary of the majority vigorous research areas in natural language processing and text mining in latest years. Its attractiveness is primarily due to two reasons. Due to extensive assortment of applications because opinions are innermost to approximately all human activities and are key influencers of our behaviors. Whenever we necessitate making a pronouncement, we want to hear others' opinions it presents many challenging research problems, which had never been attempted before the year 2000. Part of the reason for the lack of study before was that there was little opinionated text in digital forms

Keywords: Big Data, Data Mining**,** Sentimental analysis, Text Mining

1. Introduction

Data mining can be referred as extracting the knowledge or 'mining' the knowledge from massive amount of data. It comprises of two things knowledge discovery and knowledge deployment. In knowledge discovery process the main conclusion is discovered behind the given information and in knowledge deployment the discovered knowledge can be used in various decision making process of various industries. Data mining has attracted many information industries and social society in the recent year due to the availability of large amount of data which is growing day by day and need to be analyzed to discover knowledge from it. The application of this knowledge can be ranged from market analysis, production monitoring, fraud detection and customer relation management.

Data mining is actually a part of KDD (Knowledge Database Discovery). It is a sub discipline process which leads to uncover the hidden pattern among the large set of data. Some people consider knowledge discovery process as a data mining process and some consider data mining is an essential step in KDD process. The steps of KDD process are given below:

- **Data Cleaning:** To remove noise and inconsistent data.
- **Data Integration:** In this step multiple data sources are combined.
- **Data Selection:** Data relevant to the analysis task are retrieved from the database.
- **Data transformation:** Data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
- **Data mining:** An essential process where intelligent methods are applied in order to extract data patterns.
- **Pattern evaluation:** In this patterns of interest are identified to represent the knowledge.
- **Knowledge presentation:** In this step visualization and knowledge representation techniques are used to present the knowledge that is mined to the user. So that user can grab the knowledge as soon as they see the visualization dashboard. day which require very good tools for the handling this huge amount of data for various purpose such as predictive analysis and decision support system.

Mining of data is the process of finding concealed useful information from stored data. It used different types of approaches to perform mining on stored data such as clustering, classification, natural language processing, statistical analysis etc. In the process of text analytics, mostly classification procedure to castoff. Classification process is administered wisdom technique that supports in allocating a category to uncategorized tulle rendering to a previously classified instance set.

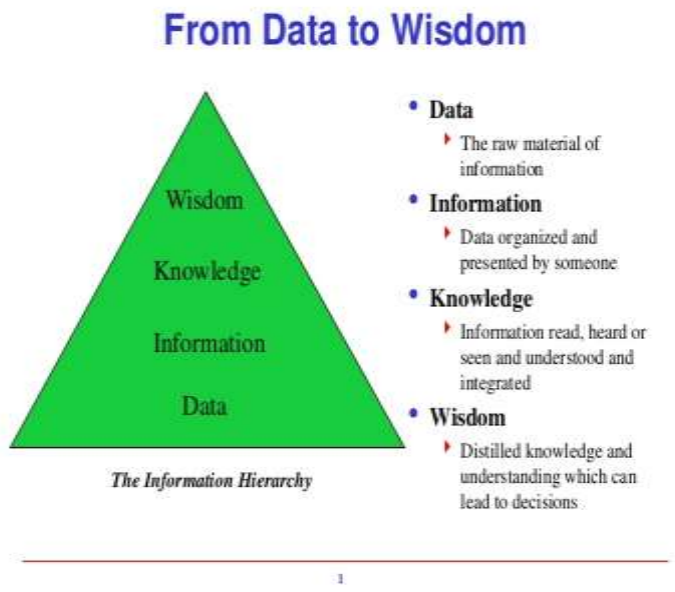
From Data to Wisdom
Wisdom
Knowledge
Information
Data
The Information Hierarchy
Data
The raw material of information
Information
Data organized and presented by someone
Knowledge
Information read, heard or seen and understood and integrated
Wisdom
Distilled knowledge and understanding which can lead to decisions
1

Figure 1: Basic representation of Data Mining

Text classification is a part of data mining which classify the text based on the content that it creates. When we talk about text classification, we usually talk about the supervised classification, which has two stages: the training stage and the testing stage. Usually in the training the classifier is trained. The testing stage includes preprocessing of testing text and classification of the testing text. Text classification is a supervised classification which has two stages: the training stage and the testing stage. Usually the training stage includes creating the labeled corpora dataset, pre-processing the training text, vectorization of the text, and training of the classifier. The testing stage includes preprocessing of testing text, vectorization and classification of the testing text.

- **Creating Corpus:** Collection of text based on categories. Every text belongs to one category and has been corrected labeled. Sometimes we divided this corpus into two sets: the training set and the testing set.
- **Preprocessing:**It is a data mining technique. It is castoff to eliminate noisy, unreliable, inadequate data and for achievement of classification. It transforms the data into the format that can be easily interpreted by users. Text preprocessing and feature extraction is a preliminary phase. Preprocessing can be done in three steps: (i) Tokenization or segmentation (ii)Removal of stop word (iii) Stemming and extracting features. The Major tasks in the Preprocessing are:
    - **Data Cleaning:** In this various errors and missing values and unnecessary information is detected and removed.
    - **Data Integration:** In this task data that is collected from various heterogeneous sources is collected and converted into consistent format.
    - **Data Reduction:** Reduced representation of data in volume is obtained but this reduced data contain the same analytical result as that of original data.
    - **Data Transformation:** In this aggregation and Normalization techniques are used to transform the data into the consistent format.

Vectorization of Text: Transform the text into vector that can be recognized for computer. All text will be represented as feature vector based on the features we selected.

Training of the Classifier: Choose one of the text classification algorithms and feed the training corpus to the classifier to get a training model.

Classification: After we get the training model, we can feed the testing data into it and get the prediction of classification.

2. Literature Review

Sentiment analysis refers to the use of natural language processing to identify and extract one-sided information in source materials or simply it refers to the process of detecting the polarity of the text. It also referred as opinion mining, as it derives the opinion, or the attitude of a user. A common approach of using this is described how people think about a particular topic. Sentiment analysis helps in determining the thoughts of a speaker or a writer with respect to some subject matter or the overall contextual polarity of a document. The attitude may be his or her decision or estimate, the emotional state of the user while writing.

Sentiment Analysis can be used to determine sentiment on a variety of level. It will score the entire document as positive or negative, and it will also score the reaction of individual words or phrases in the document.

Sentiment Analysis can track a particular topic, many companies use it to track or observe their products, services or status in general. For example, if someone is attacking your brand on social media, sentiment analysis will score the post as enormously negative, and you can create alerts for posts with hyper-negative sentiment scores.

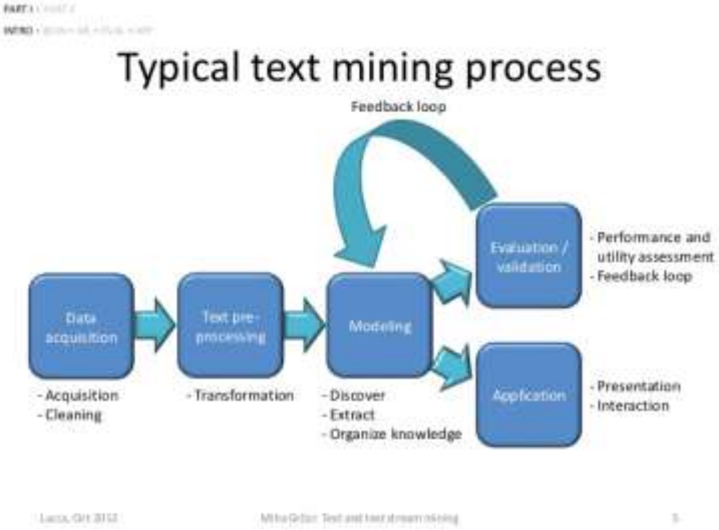

Fig: Text Mining Process

infrastructures Applying different mining techniques to derive usefulness about stored information. Different mining approaches are classification, clustering, statistical analysis, natural language processing etc. In text analytics, mainly classification technique is used. Classification is a supervised learning method that helps in assigning a class label to an unclassified tuple according to an already classifiedinstance set. Data classifying and identifying is all about to tag the data so it can be create quickly and efficiently.

But various organizations can gain from re-transforming their information, which helps in order to cut storage and backup costs, with increasing the speed of data searches. Classification can help an organization to meet authorized and regulatory requirements to retrieve specific information within a specific time period, and this is most important factor behind implementing various data classification technology.

Jiao Wu et al. [1]: The user review is very useful to generate information through which we can estimate current trends in the business, politics and in various field . This paper tells how the online sentiment analysis can be performed on campus network. This system of analysis is based on clustering service. Online sentiment examination system is poised of front-monitor node, data warehouse node, spider node, control node and back Analysis node. By using the system, the internet topic sentiment on campus network can be discovered immediately.

LingyanJi et al. 2010 [2]:This Paper describes the design and implementation of opinion mining applied in product reviews. Generally customers use these online review to take more informed decision regarding the particular product. These reviews can be further classified as positive, negative or neutral. In this semantic Role Labeling was used which is based on natural language processing technology. The analysis system has three modules: (1) web crawler (2) opinion processing (3) Interface display module. Web crawler is mainly used to extract web pages and after collecting it classification is applied on them. The opinion processing establish polarity dictionary and then feature tendency analysis extract the features from sentence and match it with dictionary and then give the orientation of the sentence that either it is positive or negative. In Interface display module graphics is used to display characteristics of goods.

Chaudhari Deptii et al.[3]: Sentiment mining is very important to manage customer relationship, to study current trends in market and popularity of the products. This paper proposed the main framework to classify the reviews using sentiment analysis.The framework uses appraisal words lexicon and product feature extraction for review categorization. Sentiment analysis focuses on two types of sentiment i.e. positive or negative. Sentiment analysis can be done by creating a lexicon of appraisal words. These adjectival appraisal words are classified as attitude, graduation, orientation and polarity. This work aims to use simple and efficient techniques such as the process of structured data generation and natural processing for sentiment analysis. The proposed work uses sentiment aware attitude for classification of reviews instead of the traditional bag of words approach.

Wu Jiao et al. 2011:Classical detection and analysis system of internet topics has low analysis efficiency and large process delay. The functions of cluster-based analysis system are internet data collection, real-time analysis and off-line data analysis. This paper compares the cluster approach and physical server approach for sentiment analysis. Cluster approach mainly has lesser Average Job Time and average waiting time than physical servers.Service Cluster scheme can detect a job failure and react much faster than physical servers.

Mihanovic Ana et al. 2014[6]: This paper describes that how task of dictionary making and sentiment analysis are done by the means of KNIME, which is a user friendly graphical workbench capable of entire analysis process. This compares the analysis work of online reviews and of tweets. In review analysis dictionary is difficult to make but for social sites like twitter this task is easy. In online review analysis is done at category and at phrase level but for tweet analysis is done at word level.

Raut B.Vijay et al. 2014 [4]: This paper describes various methods that can be used for opinion extraction, opinion classification and for opinion summarization. Mainly opinion can be extracted online from various sites like yelp.com etc. opinion classification can be done by either using machine learning method or by using lexicon method. In machine learning method various classification algorithms like SVM and Naïve Bayes can be used. And Opinion summarization can be done on the basis of feature opinion of frequent features. This paper is very helpful for studying the various techniques those can be used for opinion mining.

Bingwei Liu et al. 2013 [8]: This paper is aim to evaluate the scalability of Naïve Bayes classifier in large-scale datasets. NBC is implemented for analysis procedure by using Hadoop. For the analysis purpose raw data is collected form large dataset of movie review and after preprocessing the data the task is divided into three steps: 1.) Training Job: In this all training reviews are fed into job to extract all frequent unique words with their frequency in positive and negative review documents. 2.) Combining Job: In this all test review result and model are combined to obtain necessary information that is needed for final classification. 3.) Classify Job: Reviews are classified to their respective class of positive and negative and result is written to HDFS. And user can see the result from user terminal.

Bhatia Surbhi et al. 2015 [9]: This Paper tells the various existing method for performing Sentiment Analysis and also describes the challenges which are present in current work. Present methods for opinion detection are: Lexicon based method, Text classification method and Machine Learning. After the opinion detection and classification the concise view of large number of opinions is provide. For summarization present methods are: Feature based method, Term frequency based method Aspect based methods etc.

Maharani Warih et al. 2013[10]: This paper describes Lexical and Machine Learning approach for opinion mining. Basically Lexical based approach is dictionary based approach. In this WordNet and Sentiwordnet dictionary is used for the classification of opinion which are written in English language. But it can be used for small text opinion only and it does not take any training process. So, Machine leaning possibly alternative approach is used which take training process. The most often used method for Machine Learning Approach are: SVM, Naïve Bayes, Maximum Entropy and k- Nearest Neighbor (k-NN). The comparison Between Lexical and ML Approaches shows that scoring result is able to classify the opinion by using lexical approach. But Model Based approach is more accurate than Lexical based approach.

Chien-Liang Liu et al. 2012 [11].This paper describes the design to develop Movie-Rating system on mobile environment. Information is rated based upon the Sentiment Analysis result. For producing product feature it describe the use of Latent Semantic Analysis (LSA). This approach is used to filter the frequent feature of products in which people are interested. This Method is used to provide the concise view of review summarization. LSA is very useful for Feature-Based summarization.

Lin Lv et al. [12]:This paperdescribes the limitation of previous approaches like bag-of words for text classification. Two approaches are used for text classification to achieve more accuracy than previous ones are: WordNet and Semantic Indexing . These approaches are used to realize Naïve Bayes Classifier and support Vector for text classification. These two text classification method are gradually become advanced with more and more in depth of semantic analysis. Semantic meaning is very necessary for text classification..

3. Performance Metrics

Now for evaluating the result, different parameter semantic analysis following metrics are to be calculated. True positive, True negative, False positive and False negative are used for comparing the class label that have been assigned to a document by the classifier with the classes the item actually belongs.

a) **Accuracy:** It is measured as the proportion of correctly classified instances to the total number of instances being evaluated. Classification performance being evaluated by using this parameter [4].

(True positive +True negative)/(True positive +True negative +False positive +False negative)

True positive – that are truly classified as positive

False positive- not labeled by the classifier as positive but should be

True negative- that are truly classified as negative

False negative- not labeled by the classifier as negative but should be

b) Precision: It is widely used in evaluating the performance in different field such as text mining, information retrieval. Precision is also referred to measure the exactness. It is defined as ratio of the number of correctly labeled as positive to the total number that has been classified as positive [4].

precision= (true positive)/(true positive+false positive)

c) Kappa Accuracy: This metrics is used to measure prediction performance of a classifier. It compares the observed accuracy with the expected accuracy of the classifier. Observed accuracy means instances which are correctly classified. Expected accuracy is that which every classifier randomly expected to achieve. It is shown through confusion matrix.

d) Reliability: Reliability metrics shows the predicted probability of the classifier is precisely measured by comparing with standard one classifier. By this matrices consistency and validity of test result is determined by repeated trials.

4. Conclusion

The biggest Data mining can be referred as extracting the knowledge or 'mining' the knowledge from massive amount of data. It comprises of two things knowledge discovery and knowledge deployment. Today, Sentiment analysis plays an important role where various machine learning technique is used in determining the sentiment of very huge amounts of text or speech. Various application tasks include such as determining how someone is excited for an upcoming movie, correlates different views for a political party. In this paper we presented detailed review of the sentimental analysis and text analysis which will provide you deep insight of the work done in this and provides outline about the scope in this area and motivate research work in this and apply it on the real life application

References

1. Jiao Wu, WeihuaGao, Bin zhang, Yi Hu, Jinsong Liu, " Online Web Sentiment Analysis on CampusNetwork" Fourth International Symposium on Computational Intelligence and Design "
2. LingyanJi, Hanxiao Shi, Mengli Li, MengxiaCai,PeiqiFeng, "Opinion Mining of Product Reviews Based on Semantic Role Labeling" The 5th International Conference on Computer Science & Education Hefei, China. August 24–27, 2010.
3. Deptii D. Chaudhari, Prof. R.A. Deshmukh, Dr. A.B. Bagwan, Dr. P.K.Deshmukh, "Feature Based Approach for Review Mining Using Appraisal Words"2014.
4. Vijay B. Raut, Prof. D.D. Londhe "Survey on Opinion Mining and Summarization of User Reviews on Web" International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1026-1030.
5. Chien-Liang, Wen-Hoar Hsaio, Chia-Hoang Lee, Gen-Chui Lu, Emery Jou "Movie Rating and Review summarization in Mobile Environment" , IEEE Transactions on Systems, Man and Cybernetics-Part C: Applications and Reviews, Vol. 42, No. 3, May 2012, pp. 397-406.
6. Ana Mihanović, HrvojeGabelica, ŽivkoKrstić, "Big Data and Sentiment Analysis using KNIME:Online Reviews vs. Social Media" , 2014.
7. Ana¨ıs Collomb ,Crina Costea, Damien Joyeux ,Omar Hasan , Lionel Brunie "A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation".
8. Bingwei Liu, Erik Blasch, Yu chen, Dan Shen and Genshe Chen " Scalable Sentiment Classification for Big Data Analytics Using Naïve Bayes classifier" ,IEEE International Conference on Big Data, 2013,pp. 99-104.
9. Bhatia Surbhi, Sharma Manish, Bhatia Komal, "Strategies for Mining Opinions: A survey",2nd International Conference on Computing for Sustainable Global Development (INDIACom),2015, pp. 262-266.
10. Maharani Warih, "Microblogging Sentiment Analysis with lexical Based and Machine LearningApproaches",International Conference of Information and Communication Technology, 2013, pp. 439-443.
11. Chien-Liang Liu, Wen-Hoar Hsaio, Chian-Hoang lee, Gen-Chi Lu and Emery Jou, "Movie Rating and Review summarization in Mobile Environment" ,IEEE Transaction on System, Man, And Cybernetics-Part C Application and Reviews, Vol. 42, No. 3, May 2012, pp. 397-407.
12. Lin Lv, Yu-Shu Liu, "Research of English Text Classification Methods Based on Semantic Meaning" , School of Computer Science and Technology, Beijing Institute f Technology, pp. 688-70