A Theoretical Concept on Opinion Mining Process and Sentiment Analysis in Data Mining

Akrati Shrivastava
Dept. of Computer Science & Engg.
SATI, Vidisha
M.P., India
aakratishrivastava93@gmail.com

Satish Pawar
Dept. of Computer Science & Engg.
SATI, Vidisha
M.P., India
Satish_pawar2003@rediffmail.com

Abstract - The review of any product is beneficiary for both the Customer and the Companies. Review always help to know about the market value of any product and companies can actually determine their space of improvement. An essential part of our information-gathering behavior has constantly been to discover “what other people think”. With the explosion of Web 2.0 platforms such as blogs, discussion forums, peer-to-peer networks, and various other types of social media and actively use information technologies to find and appreciate the opinions of others. This paper covers techniques that ability to directly facilitate opinion-oriented information-seeking systems and complete workflow of Opinion mining and sentiment analysis in which those techniques can be implement and also discusses about the application of opinion mining and sentiment analysis and discuss about tools through which are used to track the opinion or polarity from the user generated contents.

Keywords- Opinion Mining, Sentiment Analysis, Mining.

I. INTRODUCTION

An essential part of our information-gathering behavior has constantly been to discover —what other people think? View of other people on a particular data people think? View of other people on a particular data admiration of opinion-rich resources such as online review sites, social networking’s sites, social media and personal blogs, new opportunities and challenges arise as people now do, actively use information technologies to seek out and understand the opinions of others. Opinion mining is the science of using text analysis to understand the drivers behind public sentiment. Sentiment analysis mainly focused on decision of the people’s i.e Positive and negative while opinion mining focused on why they take that decision.

The Opinion mining are collected from public it considered as most valuable data. The Opinion are review from customer, comment are collected from the websites and user groups. The collected Opinion are manipulated by various techniques, method, algorithm and software tools to get opinion from them. This process is called as Opinion Mining and Sentiment analysis [1]. Feature selection is the process of selecting a specific feature from a vast collection of data. All the features are existing on the data so we want to select a particular feature from all the features which belong to the reason are as follow:

- Easier Interpretation with simplification of the data model.
- Reduce the timing of training.
- Reduce over fitting

A topic can be a news, event, product, movie; location hotel etc Opinion mining is a topic in Text mining, Natural language processing (NLP) and Web mining discipline. The goal of Opinion Mining is to make computer able to recognize and express emotions. A thought, view, or attitude based on emotion instead of reason is called sentiment. The hierarchy of Data Mining and the categories of how Opinion Mining is formed under the branch mean the goal of Opinion Mining is to make computer able to recognize and express emotions. To express one’s personal opinions about any product or topic.

The growing importance of sentiment analysis coincides with the growth of social media such as reviews, forum discussions, blogs, micro-blogs, Twitter, and social networks. As a result there is a huge volume of opinionated data recorded in digital form for analysis.

Opinions on all the world wide entities are available on the web. There are many ideas on World Wide Web about various fields like politics, sports, education, marketing, history, agriculture and science.

Opinions are expressed in the form of natural language. All the social media like Twitter, MySpace, LinkedIn, Face book, YouTube and many others have gained so much reputation that they cannot be ignored[1]. Opinions of the person are differing from person to person. In a recent days social networking sites are easy way to collect opinions from the general public.
II. OPINION MINING AND SENTIMENT ANALYSIS IN DATA MINING

An opinion is a subjective statement, view, attitude, emotion, or appraisal about an entity or an aspect of the entity of an opinion holder [10]. Sentiment orientation of an opinion: positive, negative, or neutral (no opinion). It is also known as orientation, semantic used in reference to the automatic analysis of evaluative text and tracking of the predictive judgments [14].

The term "Data mining" was introduced in the 1990s, but data mining is the evolution of a field with a long history. In the early 1960s, data mining was called statistical analysis, and the inventors were statistical software companies such as SAS and SPSS. By the late 1980s, the traditional techniques had been augmented by new methods such as fuzzy logic, heuristics and neural networks [15].

Opinion mining can be defined as a sub-discipline of computational syntax that concentration on extracting people’s opinion from the web. The current growth of the web encourages users to contribute and express themselves via blogs, videos, social networking sites, etc. All these platforms provide a vast quantity of valuable evidence that we are interested to analyze [2]. Data mining is the process extracting significant information from a raw or unorganized form and Miningl extract significant information Data Mining is the nontrivial process of identifying valid, novel potentially useful, and ultimately understandable patterns in data. Data Mining is the process of extracting previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions. [20]

Data Mining is the set of methods used in the knowledge discovery process to distinguish previously unknown relationships and patterns within data [16]. Data Mining is the process of discovering advantageous patterns in data [11]. Data Mining is the decision support process where we look in large data bases for unknown and unexpected patterns of information. Data Mining is the process extracting knowledge hidden from large volume of raw data. The knowledge must be new, not apparent and one must to able to use it. Knowledge discovery database finding useful patterns in data. There is a dynamic requirement for a new generation of computational models and tools to influence humans in extracting useful information (knowledge) from the promptly developing volumes of digital data.

These theories and tools are the subject of the emerging field of knowledge discovery in databases (KDD). Data Mining is the assessment and analysis of huge amounts of data in order to discover meaningful patterns and rules. Data Mining is a collection of powerful

techniques intended for analyzing large amounts of data. Basic need of data mining is that there are too much data and little information there is a need to extract useful information from the data and to interpret the data. There are various techniques in data mining such as Association Rule, Clustering, Decision Trees, and Neural Network.

Opinion Mining is a Big Business if Someone who wants to buy smart phones they definitely look for comments and reviews. Someone who just bought a smart phone on particular brand they comment on this and share their experience and smart phone manufacture they get feedback from customer and improve the quality of their products and adjust marketing strategies. With the explosion of Web 2.0 platforms such as blogs discussion forums peer to peer networks, and various other types of social media consumers have at their disposal a soapbox of unprecedented reach and power by which to share their brand experiences and opinion, positive and negative regarding any product or service.

With the explosion of Web 2.0 platforms such as blogs discussion forums peer to peer networks, and various other types of social media consumers have at their disposal a soapbox of unprecedented reach and power by which to share their brand experiences and opinion, positive and negative regarding any product or service. As a major companies are increasingly coming to realize, these consumer voices can wield enormous influences in shaping the opinions of the other consumers and ultimately their brand advocacy companies can respond to the consumer insights they generate through social media monitoring and analysis by modifying their marketing messages, brand positioning, product development and other activities accordingly [1].

Companies cannot ignore consumer reviews; they have to pay attention to it. In large grade they can influence other people about the product it will affect the company’s reputation. Most of the companies contribute to online discussion on a frequent basis; they continuously interact with their consumer through mailing, chat, talks and improve the quality of the product, services according to the consumer.

III. LITERATURE REVIEW

Predicting the semantic orientation of an adjective is a very important aspect in opinion mining. In this paper [12] use a supervised learning algorithm to infer the semantic orientation of adjectives from constraints on conjunctions. The proposed method used textual conjunctions such as "fair and legitimate" or "simplistic but well received" to separate similarly and oppositely connoted words. Although their method achieves high precision, it relies on large corpus and needs a large amount of manually tagged training data.
In this paper [18] applies a specific unsupervised learning technique based on the mutual information between document phrases and the words “excellent” and “poor”, where the mutual information is computed using statistics gathered by a search engine. The similarity between two words is determined by counting the number of results returned by web searches joining the words with a NEAR operator. The relationship between an unknown word and set of manually-selected seeds was used to place it into a positive or negative subjectivity class.

In [19] examine several supervised machine learning methods for sentiment classification of movie reviews and conclude that machine learning techniques outperform the method that is based on human-tagged features although none of the existing methods could handle the sentiment classification with a reasonable accuracy.

In this paper [9] proposed a novel two-stage method for opinion words and opinion targets co-extraction. In the first stage, a Sentiment Graph Walking algorithm is proposed, which naturally incorporates syntactic patterns in a graph to extract opinion word/target candidates. In the second stage, a self-Learning strategy is used to refine the results from the first stage, specially for filtering out noises with high frequency and capturing long-tail terms.

In this [13] describes an approach to utilizing term weights for sentiment analysis tasks and shows how various term weighting schemes improve the performance of sentiment analysis systems. He proposed to model term weighting into a sentiment analysis system utilizing collection statistics, contextual and topic related characteristics as well as opinion related properties.

In [12] proposes a Proximity based opinion propagation method to calculate the opinion density at each point in a document.

In [17] presents a novel way to extract features from reviews. He proposes a method which crawls the reviews to form a review database. It then identifies the product features from the review database and finally for each feature, the polarity of opinion expressed in reviews is calculated. They restrict to finding explicitly mentioned product features. The frequent features are first generated using association rule miner based on the a priori algorithm. This gives the most frequent itemsets/features which occur above a user-defined threshold. It also considers only those features which have almost two words. The system then prunes some irrelevant and redundant features from the candidate list using Compactness pruning and Redundancy Pruning. In Compactness pruning, the system checks the distance between each word in the feature present in a sentence. If the distance is below a threshold for a certain number of sentences where the feature is present then that feature is selected else rejected. Redundancy pruning involves a term called p-support. For a feature is defined as the number of sentences where a feature (containing single word) is present with none of its superset feature phrase present in it. The features with p-support below a threshold are pruned from the list. The paper also proposes a way to Identify Infrequent Features too whose frequency would be below the threshold. The system finds adjectives near a frequent feature in a given sentence. These adjectives are considered as an opinion word describing that feature. This way, the system builds a list of opinion words. This way, the system builds a list of opinion words. It then proposes to use the words in opinion word list to identify infrequent features by assuming that an opinion word is used to describe various features which can include both frequent and infrequent features. Based on this assumption, the system identifies the noun phrases near an opinion word in a sentence and treats it as an infrequent feature. Their experimental results show that the infrequent features comprise around 15-20% of all the features identified for a product.

IV. OPINION MINING AND SENTIMENT ANALYSIS TECHNIQUES

Opining Mining is a relatively recent discipline that studies the extraction of opinions using Information Retrieval, Artificial intelligence and/or Natural Language Processing techniques. More informally, it’s about extracting the opinions or sentiments given in a piece of text. There are various techniques used to extract information and knowledge are generalization classification, clustering, association rule mining, data visualization, neural networks, fuzzy logic, Bayesian.

1. Supervised Machine Learning

Classification is most frequently used popular data mining technique [3]. Classification used to expect the potential effects from given data set on the basis of well-defined set of attributes and a given prognostic attributes. The given dataset is found to be the training dataset comprise on independent variables (dataset related properties) and a dependent attribute (expected attribute). A training dataset produced model test on test corpus contains the same attributes but no expected attribute. Correctness of tested model that how correct and efficient it is to make expectation. A Naive Bayes Classifier is a simple probabilistic classifier based on Bayes’ theorem and is particularly suited when the dimensionality of the inputs are high. Naive Bayes
classification is an approach to text classification that assigns the class c to a given document d.

c = \arg \max P(c|d) (1)

The Naive Bayes (NB) classifier uses the Bayes rule given in eq (2) \( P(c|d) = P(c)P(d|c) \) (2)

P(d) Where P(c|d) is the probability of instance d being in class c, P(d|c) is the probability of generating instance d given class c, \( P(c) \) is the probability of occurrence of class c and P(d) is the probability of instance d occurring [1].

2. Unsupervised Machine Learning

In compare of supervised learning, unsupervised learning has no logical targeted output associated with input. Class label for any instance is unknown so unsupervised learning is about to learn by observation. Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). Clustering has long been used for feature construction. The idea is to replace a group of similar variables by a cluster cancroids, which becomes a feature.

3. Case Based Reasoning

Case based reasoning is an emerging Artificial Intelligence supervised technique. CBR is a powerful tool of computer reasoning and solve the problems (cases) in such a way which is closest to real time scenario. It is a problem solving technique in which knowledge is personified as past cases in library and it does not depend on classical rules. The solutions are stored in CBR repository called Knowledge base or Case base.

V. WORK FLOW OF OPINION MINING AND SENTIMENT ANALYSIS

Direct opinion gives positive or negative opinion about the object directly [8]. For e.g. the voice quality smart phone music system expresses direct opinion.

2. Comparison

Comparison means to compare the object with some other similar objects [8]. For e.g. —the quality of music system of smart phone-X is better than smart phone music system-Y! expresses the comparison

3. Data Collection and Pre-Process

In this stage it is acquired the raw data from different resource’s such as blogs, social media etc., that will be analyzed for revealing of opinions. It is essential, to remove all substances that not express opinions. In this phase, pre-processing is done to eliminate redundant words or irrelevant opinions. It is compulsory to extract useful keywords from the raw data which can provide accurate information. These useful keywords are generally stored as an array of features A = (A1, A2, ..., An). Each element of array is a word from the original text, called aspect (feature). In short we take raw data and is pre-processed for feature extraction. this phase has been sub-divided into number of sub phases such as Tokenization is that a text document in which it store collection of sentences, It split the sentence into tokens remove white space, comma unwanted symbols etc.

4. Feature Extraction

The feature extraction phase deals with feature types (which identifies the type of features used for opinion mining), feature selection (used to select good features for opinion classification), feature weighting mechanism (weights each feature for good recommendation) reduction mechanisms (features for optimizing the classification process).

5. Feature Types

Types of feature used for opinion mining could be features co-occurrence (features which occurs together like unigram, bigram, or n-gram), features Part of Speech information (POS tagger is used to separate POS token) features Negation (Negation words not, not only) Opinion words (Opinion words are words which express positive or negative emotions).

6. Classification

In this case we will use, for training, a collection of comments, this collection contains various sentences which is already classified as positive and negative opinion. Comments generally contain a number of sentences, but opinion will be determined at sentence level, and then later determine overall comment opinion. Obtained collection consists of two files, one for each set of positive and negative opinions, containing one sentence per line, making it easy to process. For further process, to extract opinions we can use various algorithm but the Naive Bayesian classifier one of the most

© 2019 IJSRET
Frequently used for sentence classification, this type of classifier has the advantage that it is easy to implement, quickly and generate good results.

VI. APPLICATION OF OPINION MINING AND SENTIMENT ANALYSIS
Opinion mining is helpful to detect problems by listening rather than by asking, thereby ensuring a more accurate reflection of reality. Opinion and sentiment analysis on products. A Company is interested in customer’s perceptions about its products. Information may be used to improve products and identifying new marketing strategies [7]. Opinion and sentiment analysis on location. Tourist always curious to know about which place is best for holidays or famous resorts and restaurants. Through sentiment analysis it can be found appropriate information for planning a holiday trip. Opinion and sentiment analysis on election. Through sentiment analysis we can identify opinion of voters about political party. Voting Advise Applications help voters understanding which political party is suitable for that position.

VII. TOOLS USED IN OPINION MINING
The tools which are used to track the opinion or polarity from the user generated contents are:
1. Review Seer tool
This tool is used to automate the work done by aggregation sites. The Naive Bayes classifier approach is used to collect positive and negative opinions for assigning a score to the extracted feature terms. The results are shown as simple opinion sentence [6].
2. Web Fountain
It uses the beginning definite Base Noun Phrase (bBNP) heuristic approach for extracting the product features. It is possible to develop a simple web interfaces.
3. Red Opal
Opinion orientation of products based on their features. It assigns the scores to each product based on features extracted from the customer reviews. The results to be shown with a web base interface [5].
4. Opinion Observer
This is an opinion mining system for analyzing and comparing opinion [7]. On the internet using user generated contents, This system shows the resulting a graph format showing opinion of the product feature by feature, it uses Word Net Exploring method to assign prior polarity.

VIII. OPPORTUNITIES AND CHALLENGES
The detection of spam and fake reviews, mainly through the identification of duplicates, the comparison of qualitative with summary reviews, the detection of outliers and the reputation of the reviewer. [9]
The asymmetry in availability of opinion mining software, which can at present, be afforded only by other words, government has the means today to monitor public opinion in ways that are not available to the average citizens. While content production and publication has democratized, content analysis has not. The incorporation of opinion with behavior and implicit data, in order to authorize and deliver further analysis into the data beyond opinion expressed. The constant need for better quality, usability and user-friendliness of the tools, which are currently usable mainly by data analysts.

IX. CONCLUSION
Opinions are so essential that whenever one wants to make a decision to buy product, one wants to listen to others, opinions about the quality of product and services. This is true for both individuals and organizations. Opinion mining and sentiment analysis is based on information gathering from current available opinion rich It is a tool that enables the users to determine the resources such as blogs, online review sites, social networking’s sites, social media and personal blogs. Opinion mining and sentiment analysis can be very helpful for organization to know what the actual requirement of customer is and how to fulfill their need.

REFERENCES
[9]. Osimo D, Mureddu F; “Research Opinion Mining and Sentiment Universite de Paris-Sud, Laboratoire CNRS, B’atiment, 2012; 508.


[12]. Shima Gerani, Mark J.Carman, Fabio Crestani; b“Proximity-Based Opinion Retrieval”; ACM International Conference on Research and Development in Information Retrieval 2010 - Geneva, Switzerland.


