Original Research Paper Management

An Analysis on Opinion Mining: Techniques and Tools

B.Sampath Kumar M.B.A.,M.Com.,(PhD), Research scholar, VIT business school, VIT University, Vellore

Dr.D.Bhanu Sree Reddy PhD, Sr.Professor – General Management, VIT Business School, VIT University, Vellore.

ABSTRACT

The use of social media has created many opportunities for people to publicly voice their opinions, but when they are meant to have a opinion make a serious problem. Opinion mining is a type of natural language processing which could track the mood of the people about any particular product by review. Opinion Mining is a process of automatic extraction of knowledge by means of opinion of others about some particular product, topic or problem. The idea of Opinion mining and Sentiment Analysis tool is to process a set of search results for a given item based on the quality and features. Opinion mining is also called sentiment analysis due to large volume of opinion which is rich in web resources available online. Analyzing customer review is most important, by doing that we tend to rate the product and provide opinions for it which is been a challenging problem today. Thus this paper discusses about Opinion Mining the techniques and tools used.

KEYWORDS Data Mining, Opinion Mining, Opinion Summarization, Sentiment Analysis, Text Mining, Web Mining.

I. INTRODUCTION

The evolution of automated systems and digital information in every field of life is evolving rapidly which tends to generate data. As a result huge volumes of data are produced in field of science, engineering, medical, marketing, finance, demographic etc. Automated systems are meant to automate analysis, summarization and classification of data and number of efficient ways is available to store huge volumes of data. Text mining is an interdisciplinary method used in different fields like machine learning, information retrieval, statistics, and computational linguistics. Web mining is a sub discipline of text mining used to mine the semi structured web data in form of Web Content mining, Web Structure mining and Web Usage mining. Opinion mining also called sentiment analysis is a process of finding users opinion about particular topic or a product or problem. 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.

Hierarchy of Data Mining

II. LITERATURE SURVEY

ArtiBuche et al proposed the work on how text is classified by Navie Bayes algorithm and also Hidden Markov Model to calculate the Entropy and Purity measure. Bakhtawar Seerat et al proposed the work on how opinions are being extracted from online reviews and challenges of opinion mining. Blessy Selvam et al proposed different approaches of sentiment classification and the existing methods with the framework. Dongjoo Lee et al proposed to use the PMI method to use for large corpus to achieve higher accuracy. Tools were also discussed. S.Chandrakala et al proposed a work on recent papers on sentiment analysis and its related tasks with future challenges. S.Padmaja et al proposed a work on commonly used Machine Learning Models for text classification and an overview of the most popular machine learning algorithm used in sentiment analysis. Raisa Varghese et al proposed the different levels of sentiment analysis and the major challenges involved in sentiment analysis. Sindhu.C et al proposed a systematic flow and Machine learning approaches to optimize the performance. Vijay B.Raut et al have compared the methods and produced the synopsis of different approaches used for opinion mining and the results obtained. G.Vinodhini et al presented an overview of different opinion mining techniques with approaches used. Ayesha Rashid et al proposed the drawbacks at different sentiment level and the techniques used in Opinion mining. Nidhi Mishra et al present the insights into opinion mining at different levels. Nilesh M. Shelke et al compared the accuracy using Navie Bayes, Maximum Entropy and Support Vector Machine.Dr.RituSindhu et al proposed different levels of analysis and issues in sentiment analysis. David Osimo et al present an outline for discussion upon a new Research Challenge on Opinion Mining and Sentiment Analysis.

III. DATA SOURCE

User opinion is a major criterion for the improvement of the quality of services. Blogs, review sites, data and micro blogs provide a good understanding for the deliverable level of the products and services provided to customers.

Blogs

The name associated to universe of all the blog sites is called blogosphere. People write about the topics they want to share with others on a blog. Blog pages have become the popular

means to express ones personal opinions about any product or topic.

Review sites
For any user in making a purchasing decision, the opinions of others is being an important factor. A large number of user-generated reviews are available on the Internet. The reviewers data used in most of the sentiment classification studies are collected from the e-commerce websites like www.amazon.com (product reviews).

Data Set
The work in the field uses movie reviews data for classification. The dataset contains different types of product reviews extracted from Amazon.com including Books, DVDs, Electronics and Kitchen appliances.

IV. OPINION MINING OR SENTIMENT ANALYSIS
Opinion mining is a technique which is used to detect and extract subjective information in text documents. In general, sentiment analysis tries to determine the sentiment of a writer about some aspect and also the overall contextual polarity of a document. The sentiment may be his or her judgment, mood or evaluation. A key problem in this area is sentiment classification, where a document is labeled as a positive or negative evaluation of a target object (film, book, product etc.)The evaluation of opinion can be done in two ways:

Direct opinion gives positive or negative opinion about the object directly. For example, "The picture quality of this camera is poor" expresses a direct opinion.

Comparison means to compare the object with some other similar objects. For example, "The picture quality of camera-y is better than that of Camera-x." expresses a comparison.

Workflow of Opinion Mining
The figure workflow of opinion mining have a workflow of Opinion Mining of how the opinions are being extracted from people review over their comment Opinion feature extraction is a sub problem of opinion mining , with the vast majority of existing work done in the product review domain.

Pre-processing In this process, raw data taken and is pre-processed for feature extraction. The pre-processing phase has been further divided into number of sub phases as follows: Tokenization is that a text document has a collection of sentences which is split up into terms or tokens by removing white spaces, commas and other symbols etc. Stop word Removal removes articles (like „a, an, the).Stemming decreases relevant tokens into a single type. Case Normalization is a process that has English texts to be published in both higher and lowercase characters and turns the entire document or sentences into lowercase/uppercase.

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).

Feature Types
Types of features used for opinion mining could be: 1: Term frequency (The presence of the term in a document carries weight age). 2: Term co-occurrence (features which occurs together like unigram, bigram or n-gram), 3: Part of speech information (POS tagger is used to separate POS tokens). 4: Opinion words (Opinion words are words which express positive (good) or negative (bad) emotions). 5: Negations (Negation words (not, not only) shifts sentiment orientation in a sentence) and 6: Syntactic dependency (It is represented as a parse tree and it contains word dependency based features)

Feature Selection
1: Information gain (based on the presence and absence of a term in a document a threshold is set and the terms with less information gain is removed). 2: Odd Ratio (It is suitable for binary class domain where it has one positive and one negative class for classification.3: Document Frequency measures the number of appearances of a term in the available number of documents in the corpus and based on the threshold computed the terms are removed.

Features weighting mechanism
The mechanisms are of two types. They are 1: Term Presence and Term Frequency- word which occurs occasionally contains more information than frequently occurring words. 2: Term frequency and inverse document frequency (TFIDF) - Documents are rated where highest rating is given for words that appear regularly in a few documents and lowest rating for words that appear regularly in every document.

Feature Reduction
Feature reduction reduces the feature vector size to optimize the performance of a classifier. Reduction of the number of features in the feature vector can be done in two different ways in which top n-features can be left in the vector and either low level or unwanted linguistic features could be removed.

Adjectives only
Adjectives have been used most frequently as features amongst all parts of speech. A strong correlation between adjectives and subjectivity has been found. Although all the parts of speech are important people most commonly used adjectives to depict most of the sentiments and a high accuracy have been reported by all the works concentrating on only adjectives for feature generation.

Adjective-Adverb Combination
Most of the adverbs have no prior polarity. But when they occur with sentiment bearing adjectives, they can play a major role in determining the sentiment of a sentence. Adverbs alter the sentiment value of the adjective that they are used with. Adverbs of degree, on the basis of the extent to which they modify this sentiment value, are classified as:

Adverbs of affirmation: certainly, totally
Adverbs of doubt: maybe, probably
- Strongly intensifying adverbs: exceedingly, immensely
- Weakly intensifying adverbs: barely, slightly
- Negation and minimisers: never

Some of the positive Adjectives are as follows dazzling, brilliant, phenomenal, excellent and fantastic. Negative Adjectives: suck, terrible, awful, unwatchable, hideous.

V. ARCHITECTURE OF OPINION MINING
Opinion Mining also called sentiment analysis is a process of finding user's opinion towards a topic or a product. Opinion mining concludes whether user's view is positive, negative, or neutral about product, topic, event etc. Opinion mining and summarization process involve three main steps, first is Opinion Retrieval, Opinion Classification and Opinion Summarization. Review Text is retrieved from review websites. Opinion text in blog, reviews, comments etc. contains subjective infor-

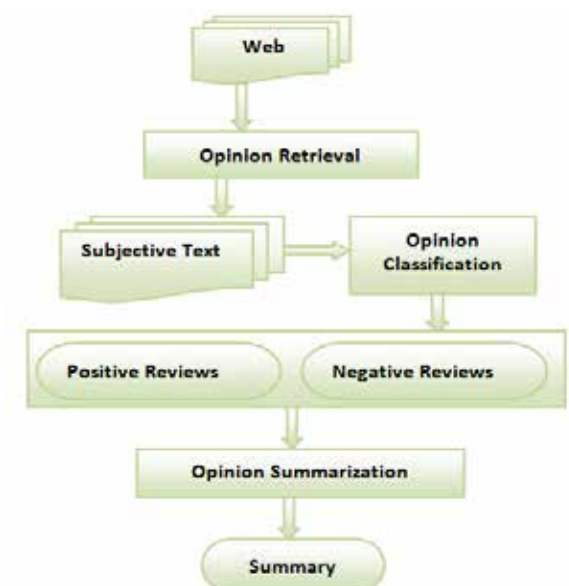
mation about topic. Reviews classified as positive or negative review. Opinion summary is generated based on features opinion sentences by considering frequent features about a topic.

Opinion Retrieval
It is the process of collecting review text from review websites. Different review websites contain reviews for products, movies, hotels and news. Information retrieval techniques such as web crawler can be applied to collect the review text data from many sources and store them in database. This step involves retrieval of reviews, micro blogs, and comments of user.

Opinion Classification
Primary step in sentiment analysis is classification of review text. Given a review document D = {d1…..d1} and a predefined categories set ***C = {positive, negative},*** sentiment classification is to classify each ***di*** in ***D***, with a label expressed in ***C***. The approach involves classifying review text into two forms namely positive and negative. Machine learning and lexicon based approach is more popular.

Opinion Summarization
Summarization of opinion is a major part in opinion mining process. Summary of reviews provided should be based on features or subtopics that are mentioned in reviews. Many works have been done on summarization of product reviews. The opinion summarization process mainly involves the following two approaches. ***Feature based summarization*** a type summarization involves finding of frequent terms (features) that are appearing in many reviews. The summary is presented by selecting sentences that contain particular feature information.

Features present in review text can be identified using Latent Semantic Analysis (LSA) method. ***Term frequency*** is count of term occurrences in a document. If a term has higher frequency it means that term is more import for summary presentation. In many product reviews certain product features appear frequently and associated with user opinions about it. Figure Architecture of Opinion Mining has the architecture of Opinion Mining which says how the input is being classified on various steps to summarize the reviews.



Architecture of Opinion Mining

VI. TECHNIQUES
Major data mining techniques used to extract the knowledge and information are: generalization, classification, clustering, association rule mining, data visualization, neural networks, fuzzy logic, Bayesian networks, and genetic algorithm, decision tree. Techniques of Opinion Mining have the techniques of Opinion Mining.

Techniques of Opinion Mining

Supervised Machine Learning
Classification is most frequently used popular data mining technique. Classification used to predict the possible outcome from given data set on the basis of defined set of attributes and a given predictive attributes. The given dataset is found to be the training dataset consist on independent variables (dataset related properties) and a dependent attribute (predicted attribute). A training dataset created model test on test corpus contains the same attributes but no predicted attribute. Accuracy of model checked that how accurate it is to make prediction. Product features and sentenced words are extracted using Double Propagation Algorithm.

Unsupervised Learning
In contrast of supervised learning, unsupervised learning has no explicit targeted output associated with input. Class label for any instance is unknown so un supervised learning is about to learn by observation. Clustering is a technique used in unsupervised learning. The process of gathering objects of similar characteristics into a group is called clustering. Objects in one cluster are dissimilar to the objects in other clusters.

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.

VII.TOOLS USED IN OPINION MINING
The tools which are used to track the opinion or polarity from the user generated contents are:

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.

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 interface.

Red Opal –It is a tool that enables the users to determine the opinion orientations 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 based interface.

Opinion observer-This is an opinion mining system for analyzing and comparing opinions on the Internet using user generated contents. This system shows the results in a graph format showing opinion of the product feature by feature. It uses Word Net Exploring method to assign prior polarity.

IX. CONCLUSION
Opinion miming is an emerging field of data mining used to extract the knowledge from huge volume of data that may be customer comments, feedback and reviews on any product or topic etc. Research has been conducted to mine opinions in

form of document, sentence and feature level sentiment analysis It is examined that now opinion mining trend is moving to the sentimental reviews of twitter data, comments used in Face book on pictures, videos or Face

book status. Thus this paper discusses about an overview of Opinion Mining in detail with the techniques and tools.

REFERENCES

[1] ArtiBuche, Dr.M.B.Chandak, AkshayZadgoanakar "Opinion Mining and Analysis: A Survey", International Journal on Natural Language Computing (IJNLC) Vol 2 No 3 June 2013Pg No 39-48.
[2] Bakhtawar Seerat, FarouqueAzam, "Opinion Mining: Issues and Challenges (A Survey)", International Journal of Computer Applications, Vol49 No 9 July 2012Pg No 42-51.
[3] BlessySelvam,A.Abirami, "A Survey on Opinion Mining Framework", International Journal of Advanced Research in Computer and Communication Engineering,Vol 2, Issue 9, Sep 2013 Pg No 3544-3549.
[4] Dongjoo Lee et al, "Opinion Mining of Customer Feedback Data on the Web". Seoul National University.
[5] S.Chandrakala, C.Sindhu, "Opinion Mining and Sentiment Classification: A Survey",ICTACT Journal on Soft Computing, Oct 2012 Vol 3 Issue 1,Pg No 420-425.
[6] S.Padmaja et al, "Opinion Mining and Sentiment Analysis – An Assesment of Peoples' Belief: A Survey", International Journal of Ad hoc, Sensor & Ubiquitous Computing IJASUC, Vol 4 No 1, Feb 2013.
[7] Raisa Varghese, Jayasree, "A Survey on Sentiment Analysis and Opinion Mining", International Journal of Research in Engineering and Technology (IJRET), Vol 2 Issue 11 Nov 2013.
[8] Sindhu, Chandrakala, "A Survey on Opinion Mining and Sentiment Polarity Classification", International Journal of Emerging Technology and Advanced Engineering.Vol 3 Issue 1, Jan 2013.
[9] Vijay .B.Raut et al, "Survey on Opinion Mining and Summarization of User Reviews on Web", International Journal of Computer Science and Information Technologies (IJCSIT),Vol 5(2), 2014. 1026-1030.
[10] G.Vinodhini et al, "Sentiment Analysis and Opinion Mining: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol 2, Issue 6, June 2012.
[11] Ayesha Rashid et al, "A Survey Paper: Areas, Techniques and Challenges of Opinion Mining", International Journal of Computer Science (IJCSI), Vol 10 Issue 6 No 2, Nov 2013.
[12] Nidhi Mishra et al, "Classification of Opinion Mining Techniques", Internation al Journal of Computer Applications, Vol 56, No 13, Oct 2012Pg No 1-6.
[13] NileshM.Shelke et al, "Survey of Techniques for Opinion Mining", International Journal of Computer Applications, Vol 57 No 13. Nov 2012Pg No 30-35.
[14] Dr.RituSindhu, RavendraRatan Singh Jandail, RakeshRanjan Kumar, "A Novel Approach for Sentiment Analysis and Opinion Mining", International Journal of Emerging Technology and Advanced Engineering (IJETAE), Vol 4, Issue 4, April 2014.
[15] DavidOsimo and Francesco Mureddu, "Research Challenge on Opinion Mining and Sentiment Analysis".