IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 1, Ver. III (Jan – Feb. 2016), PP 50-54
www.iosrjournals.org

Sentiment Mining and Related Classifiers: A Review

Rehee Mehta[1], Dr. Shaily Jain[2]
[1](Computer Science Department, Chitkara University, India)
[2](Computer Science Department, Chitkara University, India)

Abstract: *Brokers trading goods on the Web often seek for their customer's reviews and feedbacks of their products and the associated services. With the growing popularity of e-commerce, the number of customer opinions that a product receives expands rapidly. This makes it difficult for both the customers as well as the manufacturers to extract an accurate decision or outcome regarding the product's quality.*

This problem can be dealt easily by mining the characteristics of the product on which the customers have expressed their feedbacks and reviews. This paper covers all the fundamental details about opinion analysis. It comprises of current research and forthcoming scope of sentiment mining. Also the information regarding basic workflow of the opinion mining process, recent trends, and applications of sentiment analysis has been explained extraordinarily.

Keywords : *Data Mining, Opinion Mining, surveys, reviews, Sentiment Analysis, Data Pre-processing, Opinions.*

I. Introduction

Data mining is a computer aided process for discovering patterns in large data sets. The overall aim of data mining process is to gather information from a data set and convert it into an understandable form for future use. Web Mining is one of the applications of data mining techniques to discover patterns from the Web. According to the objective of the analysis, web mining can be divided into three broad categories: Web usage mining, Web content mining and Web structure mining. Web Content Mining is the process to fetch useful information from text, image, audio or video data in the web. It can also be referred as web text mining as the text content is the most widely researched area. Opinion mining is a sub discipline of web content mining which is also called as sentiment analysis. It is a process of finding users opinion about particular topic or a product. It aims to make a computer machine capable of understanding human emotions and sentiments in a way a human could understand and respond accordingly [6,8].

Generally before purchasing or launching a new product in the market, companies or individuals excavate different opinions from several people. Depending upon these opinions one can decide to continue in the same direction or to give it a second thought. Such surveys prove to be extremely beneficial economically and practically. Earlier such surveys were conducted manually by distributing pamphlets, collecting information from a sample of individuals, paper-and-pencil interviewing, face-to-face surveys, telephone surveys, mail surveys etc. But from past several years web has dramatically changed the way to express opinions by posting on merchant sites, review portals, blogs, Internet forums and much more. Such type of data is usually referred to as user-generated content or user-generated media. Various review-related websites, businesses, government intelligence applications and several other domains are quite curious about this online 'word-of-mouth', as it provides them information about their customer's choices and demands, as well as the positive and negative comments on their products, thus giving them insight of their product's flaws and an edge over their competitors. It also provides the customers with useful and appropriate knowledge about the products and services to aid in their purchase decision making process.

This paper discusses the existing works on opinion mining and sentimental analysis of customer assessments and reviews online. With the progression of web technology, there is a large amount of data present in the web for the users. These users not only use the existing resources in the web, but also give their effective feedbacks, thus adding up useful information. Due to big amount of user's opinions, feedbacks and suggestions presented by the web resources, it is very much necessary to analyze and systemize their reviews for better decision making. Opinion Mining is an Information Extraction and Natural Language Processing task that classifies the user's comments in the form of positive, negative or neutral categories. There are numerous supervised or data-driven techniques for analyzing the sentiments of the users such as Naive Byes, Maximum Entropy and SVM.

DOI: 10.9790/0661-18135054

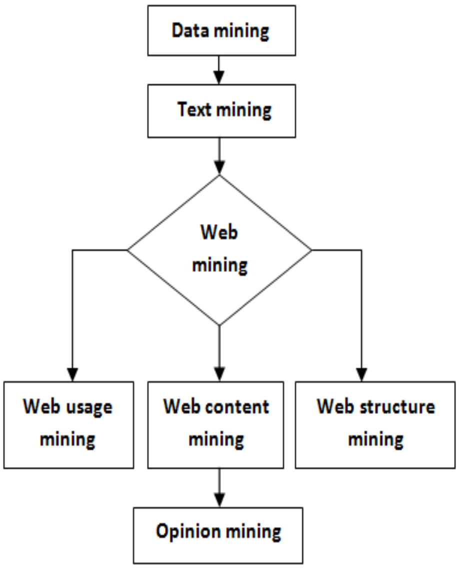

Figure 1 Data Mining and Its Types

II. Related Work

From past few years businesses and research organizations have started focusing on social media and big data. However, the field of supply chain management (SCM) has been comparatively behind in the area of research and practice. Related to this, a research contributing to the SCM community presented a novel, analytical framework (Twitter Analytics) for evaluating supply chain tweets, featuring the current use of Twitter in supply chain context. Thus, observing the possible role of Twitter for supply chain practice and research. The suggested framework associates three methodologies – content analytics (CA) , descriptive analytics (DA) and network analytics (NA) relying upon network visualization and metrics for extracting knowledge from 22,399 #supply chain tweets. The outcome of the paper involved the supply chain tweets used by distinct groups of supply chain professionals and organizations such as news services, IT industries, logistic providers, manufacturers, etc for data sharing, appointing professionals communicating with stakeholders etc. Also several other topics such as logistics and corporate social responsibility to uncertainty, manufacturing, SCMIT and even human rights were examined [1]. The social media collects the data in structured and unstructured, formal and informal form as the users do not care about the spellings and grammatical formation of a sentence while communicating with each other using different social networking websites such as Facebook, orkut, LinkedIn, instagram, etc. The collected data consisted of sentiments and opinions of users which were processed using data mining techniques and were analyzed for capturing the useful information from it [3].

An opinion mining extraction algorithm to jointly explore the essential opinion mining elements was proposed. Particularly, the algorithm automatically creates kernels to join closely related words into new terms from word level to phrase level based on dependency relations and assured the certainty of opinion expressions and polarity based on fuzzy dimensions, opinion rate intensifiers, and opinion patterns. Some interesting observations were acknowledged like the negative polarity of video dimension was greater than the product usability dimension for a product. Still, increasing the dimension of product usability could effectively improve the product [4]. The information on current trends, applications of opinion mining, several areas where it could have been used and also lot of meaningful information on the recent research work that was being carried out in this field of data mining was provided. Also, the primitive work plan of the sentiment analysis process, the challenges and the forthcoming research being planned in the area of sentiment analysis was explained remarkably [5]. A mining approach to mine and gather product characteristics, reviews from various web sources for a particular product in which a rule-based approach system was implemented, was proposed, which practiced linguistic and opinion mining of texts to mine feature-sentiment pairs that have sentence-level co-occurrence in consumer feedback records. The captured feature-sentiment pairs were modeled, classified, distinguished between formal, informal and undefined opinions [6]. A novel approach for contextualizing and enriching massive semantic knowledge bases for sentiment analysis with a focus on Web intelligence platforms and other highly efficient big data applications was presented. The method was not only relevant to traditional sentiment lexicons, but also to broader, complete, multi-dimensional affective resources such as SenticNet [7]. The tool of Opinion mining and Sentiment Analysis processes a set of search results for a given item based on the quality and characteristics. By analyzing customer review one can scale a particular product and provide opinions for it. Research has been carried out in this field to mine opinions in the form of document, sentence and feature level sentiment analysis. It is examined that now the opinion mining trend is shifting to the sentimental analysis of the data obtained from several social media websites such as twitter data, comments used in Facebook on pictures, videos or Facebook status etc. Various techniques and tools of Opinion Mining were discussed in this paper [8].

Opinion Mining/ Sentiment Analysis

▸Preprocessing

In this step, the raw data is collected and pre-processed for feature extraction. The pre-processing phase can further be divided into number of sub phases which are as follows: In Tokenization phase, a text document consisting of number of sentences is broken down into terms or tokens by removing white spaces, commas and other symbols. Stop word removal discard the articles such as a, an, the. Stemming reduces relevant tokens into a single term. Case Normalization is a method that has English texts to be written in both upper and lowercase characters and converts the entire document into lowercase or uppercase.

Feature Extraction

The feature extraction phase deals with feature/characteristic types (which describes the type of features used for sentiment analysis), feature selection (used to select appropriate features for sentiment categorization), feature weighting mechanism (weights each characteristic for better recommendation) reduction mechanisms (features for enhancing the classification process).

Feature Types

Types of features involved in opinion mining process are as follows:

i. Term frequency (number of time the term existed in a given document).
ii. Term co-occurrence (characteristics which exist together like unigram, bigram, trigram, etc).
iii. Part of speech information (POS tagger is used to isolate POS tokens).
iv. Opinion words (Opinion words are the terms which express positive (good) or negative (bad) sentiments).
v. Negations ((such as not, not only, etc) alter opinion orientation in a sentence) and
vi. Syntactic dependency (expressed in terms of a parse tree and consists of word dependency based features).

Feature Selection

i. Information gain (depending upon the presence and absence of a word in a given document, a threshold is set and the words with low information gain are discarded).
ii. Odd Ratio (applicable for binary class domain where it has one positive and one negative class for categorization.
iii. Document Frequency calculates the maximum number of occurrences of a term in the existing document and based on the calculated threshold, the terms are discarded.

Features weighting mechanism

The mechanisms are of two types which are as follows:

i. Term Presence and Term Frequency- word which appears infrequently contains more information than regularly occurring words.
ii. Term frequency and inverse document frequency (TFIDF) - Documents are ranked where highest grading is given for words that occur frequently in a few documents and lowest grading for words that occur frequently in every document.

COMPARASON AMONGST VARIOUS CLASSIFIERS

Table 1 Comparison Between Various Classifiers Based On Their Advantages And Disadvantages.

CLASSIFIER	ADVANTAGES	DISADVANTAGES
Naive Bayes Classifier	➢ Easy to implement. ➢ Excellent computational efficiency & classification rate ➢ Predict accurate results for most of the problems.	➢ Precision decreases if the amount of data is less. ➢ Large number of records required for good results.
Decision Tree	➢ Easy to understand. ➢ Easy to generate rules. ➢ Reduce problem complexity.	➢ Training time is relatively expensive. ➢ One branch ➢ Once a mistake is made at a higher level, any sub tree is wrong. ➢ Does not handle continuous variable well. ➢ May suffer from over fitting.

K-nearest neighbor	➢ Effective ➢ Non-parametric ➢ Classes need not be linearly separable. ➢ Zero cost of the learning process. ➢ Robust to noisy training data. ➢ Suitable for multimodal classes.	➢ Classification time is high ➢ Difficult to find optimal value of k. ➢ High time consumption for large training data set. ➢ Sensitive to noisy and irrelevant attribute.
Support Vector Machine	➢ Gathers the inherent features of the data better. ➢ High accurate	➢ Parameter tuning ➢ kernel selection
Artificial Neural Network Algorithm	➢ Easy to use ➢ Reprogramming is not required ➢ Easy implementation ➢ Can be applied on number of problems	➢ In case of large neural network, high processing time is required ➢ Number of neurons and layers are difficult to compute ➢ Slow learning

Application Areas

Sentiment Analysis And Opinion Mining Covers A Broad Range Of Applications
Some of which are as follows:

i. Commercial markets: For the business investors it is important to analyze the market trends and other investor's opinions about the stocks of a company, to identify price trends.

ii. Goods or commodities: An organization is curious in customers' reviews and feedbacks about its products. Information may be used to enhance the product's quality and recognizing new marketing strategies.

iii. Maps or Location: Tourists are fascinated in gathering information about the best places to visit. Thus opinion mining can be used for this purpose for capturing relevant information before planning a trip.

iv. Analysis of software programs: We can detect users' sentiments from posted reviews on specialized sites.

v. Voting Suggestion Applications: It assists the voters in perceiving which political party has closer positions to their choice. For example, SmartVote.ch asks the voter to establish its degree of agreement with a number of policy statements, and then correlates its position with the political parties.

vi. Computerized content analysis: It assists in processing huge amount of qualitative data. Nowadays there are numerous tools in the market that combine statistical algorithm with semantics and machine learning with human instructions. These results are able to analyze relevant comments and assign positive or negative implications to it (also called as sentiment).

Why It Matters In Governance

Opinion mining applications are the base of large scale collaborative guidelines. It helps to identify initial cautionary system of possible interruption in an appropriate manner, by detecting early reviews from inhabitants. Generally, impromptu surveys are followed to collect feedbacks in a systematic manner. However, this type of data assemblage is not economical, as it requires expenditure in design and data collection; it is quite hard, as people are not responsive in acknowledging surveys and hence it is not profitable, as it detects known issues through pre-existing questions and respondents, but is unsuccessful to detect the most significant problems, the famous 'unknown unknown'. Sentiment analysis aids to detect problems by listening and not by asking, thus assuring a more factual impression of current scenario. Argument mapping application is effective to assure that policy arguments are logical and proof-based, and do not carry out the same debate repeatedly. Such software would ultimately be useful for policy-makers as well as for citizens who could more easily grasp the essential points of a discussion and involve in the policy-making criteria.

III. Conclusion

Sentiment analysis emerges as a challenging field with several barriers. It has diverse applications that could turn out to be extremely beneficial in number of fields such as manufacturing, business analytics, marketing, etc. This study provides a comprehensive understanding of various opinion mining classifiers. From our review we concluded that different algorithms execute differently depending on the data accumulation. In this paper we have examined different classifiers with their pros and cons. Some of these algorithms perform fairly while some perform extraordinarily depending upon the requirements of the user. None of them appears to be exceptionally superior over the others in all contexts.

References

[1] Bongsug (Kevin) Chae, Insights from hashtag #supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research, International Journal of Production Economics, Vol. 165, July 2015, pp. 247–259.

[2] Gerald Petz, Michał Karpowicz, Harald Furschus, Andreas Auinger, Vaclav Stritesky Andreas Holzinger, Reprint of: Computational approaches for mining user's opinions on the Web 2.0, Information Processing and Management, Vol. 51, Issue 4, 2015, pp. 510-519.

[3] Ms. Priyanka Patel, Ms. Khushali Mistry, A Review: Text Classification on Social Media Data, IOSR Journal of Computer Engineering, Vol. 17, Issue 1, Jan – Feb 2015, pp. 80-84.

[4] Haiqing Zhang, aichasekhari, yacineouzrout, abdelazizbouras, Jointly identifying opinion mining elements and fuzzy measurement of opinion intensity to analyze product features, Engineering Applications of Artificial Intelligence, 29 June 2015.

[5] Rushabh Shah, Bhoomit Patel, Procedure of Opinion Mining and Sentiment Analysis: A Study, International Journal of Current Engineering and Technology, Vol.4, No.6, 1 Dec 2014,pp. 4086-4090.

[6] Nidhi R. Sharma, Vidya D. Chitre, Mining, Identifying and Summarizing Features from Web Opinion Sources in Customer Reviews, International Journal of Innovations & Advancement in Computer Science (IJIACS),Vol. 3, 7 Sep 2014, pp. 8-14.

[7] A. Weichselbraun , S. Gindl , A. Scharl, Enriching semantic knowledge bases for opinion mining in big data Applications, Knowledge-Based Systems, Vol. 69, October 2014, pp. 78-85.

[8] G. Angulakshmi, Dr. R. Manickachezian, An Analysis on Opinion Mining: Techniques and Tools, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 7, July 2014, pp. 7483-7487.

[9] Pravesh Kumar Singh, Mohd Shahid Husain, Methodological Study of Opinion Mining and Sentiment Analysis Techniques, International Journal on Soft Computing (IJSC), Vol. 5, No. 1, February 2014, pp. 11-21.

[10] Nidhi R. Sharma, Vidya D. Chitre, Opinion Mining, Analysis and its Challenges, International Journal of Innovations & Advancement in Computer Science (IJIACS), Vol. 3, 1 April 2014, pp. 59-65.

[11] T. K. Das, D. P. Acharjya and M. R. Patra, Opinion Mining about a Product by Analyzing Public Tweets in Twitter, International Conference on Computer Communication and Informatics (ICCCI), 03 – 05 Jan 2014, pp. 1-4.

[12] Nidhi Mishra, C.K. Jha, Classification of Opinion Mining Techniques, International Journal of Computer Applications, Vol. 56, No.13, October 2012.

[13] Ion Smeureanu, Cristian Bucur, Applying Supervised Opinion Mining Techniques on Online User Reviews, Informatica Economica, Vol. 16, No. 2, 2012, pp. 81-91.

[14] Nidhi R. Sharma, Vidya D. Chitre, Opinion Mining, Analysis and its Challenges, International Journal of Innovations & Advancement in Computer Science (IJIACS), Vol. 3, 1 April 2014, pp. 59-65.