Comparative Analysis of Opinion Mining KDD for Customer Reviews

Gayika Singh[1], Brij Kishore[2]

[1]Student, Dept. of Computer Science and Engg., Apex Institute of Engineering and Technology, Jaipur, India

[2]Assistant Professor, Dept. of Computer Science and Engg., Apex Inst. of Engg. and Technology, Jaipur, India

***Abstract*: Analyzing customer reviews is important for the sellers as well as for the customers who buys product online or even offline. Companies these days put efforts and are quite conscious about the reviews and the sentiments of the consumers about the company products and its policies. Online reviews are being monitored and responded by the companies in accordance with their relevance to the product and their services. Opinion Mining is an option and an algorithm to study and understand the mass opinion about the products and the services of an individual or the company.**

***Keywords*: Opinion Mining, Customer Reviews, Ratings, Knowledge Discovery Databases.**

1. Introduction

This paper deals with the concept of opinion mining and the Knowledge Discovery Database for analyzing current trends and reviews of the consumer for the improvements in the quality of products (QoP) and in quality of service (QoS). Knowledge discovery in databases (KDD) is the process of discovering useful knowledge from a collection of data. This widely used data mining technique is a process that includes data preparation and selection, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed results.

Traditionally, data mining and knowledge discovery was performed manually. As time passed, the amount of data in many systems grew to larger than terabyte size, and could no longer be maintained manually. Moreover, for the successful existence of any business, discovering underlying patterns in data is considered essential. As a result, several software tools were developed to discover hidden data and make assumptions, which formed a part of artificial intelligence.

The KDD process has reached its peak in the last 10 years. It now houses many different approaches to discovery, which includes inductive learning, Bayesian statistics, semantic query optimization, knowledge acquisition for expert systems and information theory. The ultimate goal is to extract high-level knowledge from low-level data.

Sentiment analysis, also referred to as opinion mining, is an approach to natural language processing (NLP) that identifies the emotional tone behind a body of text. This is a popular way for organizations to determine and categorize opinions about a product, service or idea. It involves the use of data mining, machine learning (ML) and artificial intelligence (AI) to mine text for sentiment and subjective information.

Sentiment analysis systems help organizations gather insights from unorganized and unstructured text that comes from online sources such as emails, blog posts, support tickets, web chats, social media channels, forums and comments. Algorithms replace manual data processing by implementing rule-based, automatic or hybrid methods. Rule-based systems perform sentiment analysis based on predefined, lexicon-based rules while automatic systems learn from data with machine learning techniques. A hybrid sentiment analysis combines both approaches.

In addition to identifying sentiment, opinion mining can extract the polarity (or the amount of positivity and negativity), subject and opinion holder within the text. Furthermore, sentiment analysis can be applied to varying scopes such as document, paragraph, sentence and sub-sentence levels.

Vendors that offer sentiment analysis platforms or SaaS products include Brandwatch, Hootsuite, Lexalytics, NetBase, Sprout Social, Sysomos and Zoho. Businesses that use these tools can review customer feedback more regularly and proactively respond to changes of opinion within the market.

Data mining is a process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers to develop more effective marketing strategies, increase sales and decrease costs. Data mining depends on effective data collection, warehousing, and computer processing.

Data mining involves exploring and analyzing large blocks of information to glean meaningful patterns and trends. It can be used in a variety of ways, such as database marketing, credit risk management, fraud detection, spam Email filtering, or even to discern the sentiment or opinion of users.

The data mining process breaks down into five steps. First, organizations collect data and load it into their data warehouses. Next, they store and manage the data, either on in-house servers or the cloud. Business analysts, management teams and information technology professionals access the data and determine how they want to organize it. Then, application software sorts the data based on the user's results, and finally, the end user presents the data in an easy-to-share format, such as a graph or table.

Data mining programs analyze relationships and patterns in data based on what users request. For example, a company can use data mining software to create classes of information. To illustrate, imagine a restaurant wants to use data mining to determine when it should offer certain specials. It looks at the information it has collected and creates classes based on when customers visit and what they order.

2. Literature survey

Marjia Sultana, et.al (2016) presented the various issues that arise due to the input attributes for prediction of heart diseases. Across the whole world, heart diseases are increasing day by day. Since an expert is required along with higher knowledge related to prediction, the prediction of heart diseases is not easy [1]. The hidden information that is very important within decision making is extracted using data mining. In order to identify more accurate technique for heart disease prediction, an experiment is conducted in this paper that utilizes diverse data mining approaches. For each data mining approach, two separate data sets are utilized in this paper. Five different classifiers are investigated here amongst which the performances of Bayes Net and SMa classifiers are shown to be better than others for predicting heart diseases.

Archana Gahlaut, et.al (2017) proposed a load credibility prediction system here in order to provide benefits to various bank organizations [2]. The approval or disapproval of a loan request generated by a client is decided by the banks on the basis of the information related to that client such that proper decision can be made. The information related to the client's family background, his occupation, marital and financial status, are the factors that are important here. The decision related to providing clients with credit cards can be made as per the proposed analysis work. The Random Forest algorithm is chosen to be the best risky credit classification mechanism amongst others once the classification data mining techniques are applied. The performance and accuracy of Random Forest is not degraded even though its performance is slower in the runtime.

Kamaljit Kaur et.al (2015) proposed novel Credit Based Continuous Evaluation and Grading System (CBCEGS) approach continuous evaluation of complete system is utilized in order to assess the performance of students [3]. The performance of students is enhanced by providing them a multistage examination pattern. The two data sets of 1000 students are utilized here such that the performances of students can be analyzed and predicted. The system analysis and design is generated here using this gathered data. In order to predict the grades of students for both subjects, the Classification and Regression Tree (CART) supplemented by AdaBoost is known to be the best classifier as per the evaluations. In order to perform system analysis and design the best performance is shown by J48 that is supplemented by AdaBoost. However, for mathematics, its performance is known to be the worst. In order to predict the marks of students previously in major test, the best results are generated by M5P.

Monali Paul, et.al (2015) proposed a system for predicting the category of analyzed soil datasets which utilized data mining techniques [4]. The yielding of crops is predicted by this category. As a classification rule, the formalization of predicting the crop yield is done which also utilize Naive Bayes and K-Nearest Neighbor methods. The selection of appropriate land such that a better crop can be produced by soil analysts and farmers can be done through this study. Further, by utilizing other data mining classification techniques such that more efficient models can be generated is the future work of this research.

Akshay Raul, et.al (2016) presented the information related to various bodies that are relevant to healthcare domains. An alternative to the medicine that is prescribed by doctor can be identified by the proposed system [5]. Along with an alternative of medicines, an optimum price of drug is provided here. The competitors as well as the price of medicine in which they are selling it can also be identified by the pharmaceutical companies. These companies can advertise their respective medicines in better way. The area that is prone to diseases and that expands their sales out is identified by the Insurance Companies. The human efforts are minimized such that the information can be identified with the help of data mining along with availability of information.

Shan Xu, et.al (2016) presented a study related to the practical issues of Chinese hospital that is dealing with the data of cardiovascular patients. This provides an early detection and risk prediction of the diseases of patients. The basic natural language processing mechanism was used here. For real patents dataset, 50 data mining techniques were tested after performing data preprocessing [6]. In order to generate an ensemble system, the top 6 sub-classifiers were chosen such that a complete advantage of multi-methods was taken and bias was minimized. In order to generate a final result that included risk prediction and confidence, the voting mechanism was adjusted. Thus, for diagnosing the doctors within practical use, the risk prediction confidence and algorithm's accuracy were shown.

Maklan and Klaus [7, p.228] defined customer Experience as "the customer's cognitive and affective assessment of all the direct and indirect encounters with the firm relating to their purchase behaviour". Moreover, this definition is derived from the previous studies [8]-[10] and it is highly consistent with the conceptualization of these researches. The most interesting findings of Maklan and Klaus [7] study was that customers' experience perceptions have a significant impact on customer satisfaction. The results of this study also indicate that customer experience also has significant impact on loyalty intention and customer behavioural intention.

3. Conclusion

Opinion mining is important and is really crucial when it comes to the analysis of the textual data. Opinion mining and KDD (Knowledge Discovery in Database) can really help in creating analysis of the sentiments of the users. Online presence of the business and the competitors have made the market competition healthy and allows the customers to choose from plenty which eventually benefits the customer as the manufacturers and the service providers needs to upgrade and improve themselves.

References

[1] Marjia Sultana, Afrin Haider and Mohammad ShorifUddin, “Analysis of Data Mining Techniques for Heart Disease Prediction”, IEEE, volume 19, issue 55, pp. 155-167, 2016.

[2] Archana Gahlaut, Tushar, Prince Kumar Singh, “Prediction analysis of risky credit using Data mining classification models”, ICCCNT, volume 3, issue 8, pp. 839-846, 2017.

[3] Kamaljit Kaur and Kuljit Kaur, “Analyzing the Effect of Difficulty Level of a Course on Students Performance Prediction using Data Mining”, 1st International Conference on Next Generation Computing Technologies (NGCT-2015), volume 21, no. 10, pp. 422-430, 2015.

[4] Monali Paul, Santosh K. Vishwakarma, Ashok Verma, “Analysis of Soil Behaviour and Prediction of Crop Yield using Data Mining Approach”, International Conference on Computational Intelligence and Communication Networks, volume 8, issue 2, pp. 823- 839, 2015.

[5] Akshay Raul, Atharva Patil, Prem Raheja, Rupali Sawant, “Knowledge Discovery, Analysis and Prediction in Healthcare using Data Mining and Analytics”, 2nd International Conference on Next Generation Computing Technologies (NGCT-2016), vol 19, issue 11, pp. 550- 563, 2016.

[6] Shan Xu, Haoyue Shi, Xiaohui Duan, Tiangang Zhu, Peihua Wu, Dongyue Liu, “Cardiovascular Risk Prediction Method Based on Test Analysis and Data Mining Ensemble System”, IEEE, volume 20, issue 15, pp. 475-489, 2016.

[7] S. Maklan and P. Klaus, "Customer experience: are we measuring the right things?," International Journal of Market Research, vol. 53, pp. 771-792, 2011.

[8] P. C. Verhoef, K. N. Lemon, A. Parasuraman, A. Roggeveen, M. Tsiros, and L. A. Schlesinger, "Customer experience creation: Determinants, dynamics and management strategies," Journal of retailing, vol. 85, pp. 31-41, 2009.

[9] F. Lemke, M. Clark, and H. Wilson, "Customer experience quality: an exploration in business and consumer contexts using repertory grid technique," Journal of the Academy of Marketing Science, vol. 39, pp. 846-869, 2011.

[10] A. L. Roggeveen and L. Schlesinger, "Customer Experience Creation: Determinants, Dynamics and Management Strategies," Journal of Retailing, 2008.