International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-6, April 2020

An Efficient Algorithm for Text Mining in Business Intelligence using Machine Learning

Devendra Kumar Mishra, Arvind Kumar Upadhyay, Sanjiv Sharma

Abstract: ***Data plays an important role in success of any organization, so organizations required more data to make decision for their planning to improvement. The data that are generating for any organization, in which 80 to 90 percent data belongs to unstructured data type.Text mining is the process that indicate retrieve appealing and unknown information from unstructured text. Social network sites also generate huge amounts of data,with the help of these data people's behavior and thought easily determine but analysis of these data is a difficult task. This paper proposed an efficient approach for text mining using machine learning.***

Keywords : Business Intelligence, Machine Learning, Unstructured Data .

I. INTRODUCTION

The term text mining indicates the process of retrieving unknown knowledge from unstructured text. Text mining is a multidisciplinary field related to information extraction, machine learning approach, analysis tools on statistics, and data mining.Text mining contains some step that involves preprocessing of documents ,classification process, clustering on the basis of the feature, information retrieval and finally, visualization. Machine Learning approach contain algorithms based on statistic methods those are capable of analysis of big volume data in real time. There are so many learning approaches available to solve specific problems, but supervised and unsupervised learning commonly used. K-means clustering is one of the examples of unsupervised learning. In the case of supervised learning labeled training data available that guide to calculating the value of given input, an example includes handwriting recognition, classification of e-mail messages. There are a lot of algorithms available to create learners, for example, Support Vector Machines, Naive Bayes Classifiers, and Neural Networks. In the case of unsupervised learning, no guide is available to make sense of data. This approach is generally used for clustering purpose.

Revised Manuscript Received on April 30, 2020.
***** Correspondence Author
Devendra Kumar Mishra*, Amity School of Engineering and Technology,Amity University Madhya Pradesh, Maharajpura Dang,Gwalior(MP)-474005,India.
Prof.(Dr.) Arvind Kumar Upadhyay, Amity School of Engineering and Technology,Amity University Madhya Pradesh, Maharajpura Dang,Gwalior(MP)-474005,India.
Dr. Sanjiv Sharma, Madhav Institute of Technology and Science,Gwalior(MP) ,India.

Business Intelligence (BI) specifies different techniques that store and analyze the business data so that analyzed data can be used to make a decision for improvement of business process[1].

II. LITERATURE REVIEW

In 2010, Himanshu Vashishtha, Michael Smit and Eleni Stroulia[2] have design an architecture to move the text analysis tools on cloud.they worked on TAPoR model .they describe redesign and reimplement four basic operation on Hadoop,and show the result that indicate performance improvements enabled by the migration.

In 2010, B. Zhou,Y. Jia, Chunyang Liu and X. Zhang[3],proposed a text mining system for real world web mining .this system contain distributed and layered architecture. Layers are divided into crawling and storage,mining and service analysis layer.they used online topic detection application for their experiment and discuss the results.

In 2011, F.S. Gharehchopogh and Z.A. Khalifelu[4] perform analysis between structured and unstructured data,they also performed analysis between natural language processing and text mining.In 2012, S. Jiang,D. Lowd and Dejing Dou[5] ,present an approach for knowledge extraction this approach is based on Markov logic .this system was use ontological constraints ,confidence values and human labeled data for patterns extraction. In 2013, N.Otsuka and M.Matsushita,[6] design the interface in text mining process.this interface is based on trial and error concept. In text mining user interface is used to explore the text.In 2013, Y. Zhao,Kai Niu Zhiqiang He, Jiaru Lin and Xinyu Wang[7] ,propose an approach that can apply into a real time platform. In this approach they optimize data structure, query strategy and parallel processing.they produced results regarding efficiency and accuracy of their approach and other existing methods.

In 2014, Suan Lee,Namsoo Kim and Jinho Kim[8],proposed a text cube model for analysis of unstructured text.in this approach they extend text cube model by including TF-IDF and LM .they provide some experimental results effective then other existing methods. In 2014,W.Sunayama,Y.Takama,Y.Nishihara, T.Kajinami ,M.Kushima and H.Tokunaga[9] offering an approach named TETDM. It provides an environment where many text analysis tools can be combine.it is design with aim to support for mining information from large volume of text data.

In 2014, Dursun Delen and Asil Oztekin[10] specify the use of data mining ,text mining and web mining for decision support . they describe Mini-track system for managerial decision making. In 2015, Renaud Richardet,Jean Cedric Chappelier,

Shreejoy Tripathy and Sean Hill[11] Published a paper based on agile text mining and also introduce a system Sherlok used for develop applications with agile text mining approach .In 2016, Xuan Lv,and Nora El-Gohary[12], proposed opinion mining approach for transportation users. It was used supervised machine learning approach for extraction. In 2017 Liu, Y., J. W. Bi, and Z. P. Fan.[13] specify an approach to rank the products on the basis of online reviews using sentiment analysis, this approach determine the sentiment of customer towards products and also apply fuzzy set theory for determining the result.
In 2018 R. Chen, Y. Zheng, W. Xu, M. Liu, and J. Wang[14] proposed an prediction model that efficiently assessing secondhand sellers. In 2019 Achim Klein , Martin Riekert , Velizar Dinev[15] proposed information retrieval system that automatically collect text and analyze by machine learning based approach.

III. PROPOSED SYSTEM

The proposed system will be used concept of text clustering, this is an unsupervised process and focus to classified objects into groups. In this approach categories are not predefined. Proposed system will automatically create different categories on the basis of different concept appear in the documents.

(A)Working of proposed system:--

Step 1. In this step we get input as reviews and feedback that makes business robust

(a)Scrapping this feedback as plain text
(b)Processing this text with multiple NLP algorithms.
(c)Extracting important information out of this dump of free text by improved pre-processing method

Step 2. After completion of step 1 we get Enhancement in existing knowledge.
Step3. Filter irrelevant concepts to keep precision high.
Step 4.Capture and visualize most frequent words from vast amount of data.
Step 5. Organize dispersed knowledge into categories and key words.
Step6 Display the results.

IV. FRAMEWORK AND METHODOLOGY

The approach of proposed system is not using the concept of knowledge point extraction instead of this, it will focus on extraction of general features available in the documents. This system classifying different features those plays an important role for making decision to business using machine learning. Proposed system will take input from different sources in unstructured form and provide output in terms of different features. On the basis of values of different features organization will take appropriate action to improve the business process.The domain relevance of a concept C in domain □om k is calculated as:-

Domain Relevant (C, □om k) = P(C | □om k) /max P (C | □om k) …….(i)

Where, P(C | □om k) = freq (C ∈ □om k)/ $\sum_{i=1}^{n}$ freq (C ∈ □om k).

The domain consensus of a concept "C" in domain □om k is given as follows:

DC (C, □omK) = □ (P (C, □□)) = ∑P (C, □□) × log2 1/ P (C, □□) ……(ii)

where □□ is documents in □omK, and the probability P (C, □□) will be calculate as follows:

freq (C ∈ □□)/ $\sum_{dj \in DomK}$ freq (C ∈ □□)

The framework of the proposed system is mention in the figure 1.

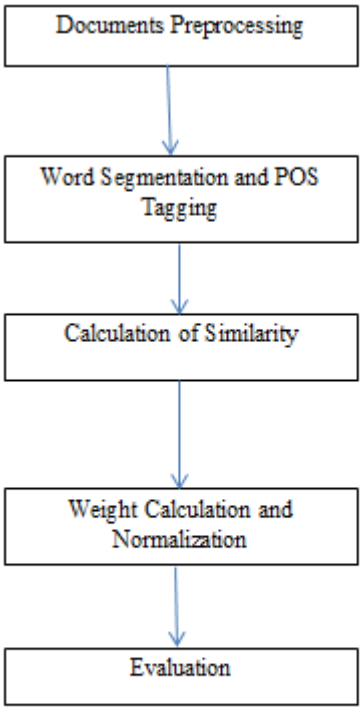
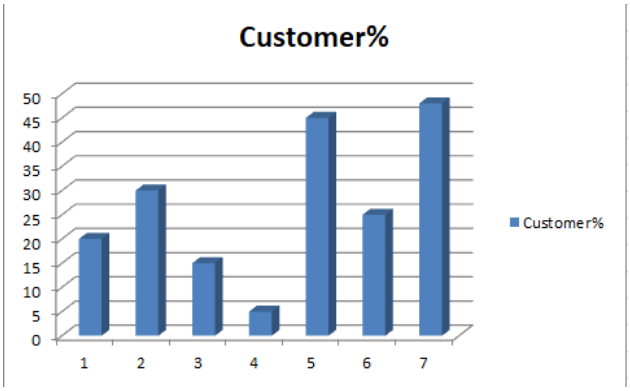


Figure 1 – Framework for proposed system

V. RESULT AND DISCUSSION

When we apply proposed algorithm on the feedback of an product provided by different customers. it will generate different clusters that will specify the customers response toward the different parameters of the product. If we apply algorithm on a data set then expected output is given in the graph



VI. CONCLUSION

proposed system identified the key attributes driving customer satisfaction and dissatisfaction toward products and services . this system is using machine learning approach and classifying all the features.

After analysis it has been found that proposed system will provide information for multiple dimension to make suitable decision for business process because it is considering each and every response provided by customers.

REFERENCES

1. Ishikiriyama Celia Satiko, Miro Diego, and Francisco Carlos, Gomes Simoes "Text Mining Business Intelligence:A small Sample of what words can say",Procedia Computer Science,vol.55,261-267,2015.
2. Vashishtha Himanshu, Michael Smit and Eleni Stroulia. "Moving Text Analysis Tools to the Cloud", 6th World Congress on Services, IEEE, 107-114,2010.
3. Zhou Bin, Jia Yan, Liu Chunyang and Zhang Xu " A Distributed Text Mining System for Online Web Textual Data Analysis",International Confrence on Cyber Enabled Distributed Computing and Knowledge Discovery,IEEE,1-4,2010.
4. Gharehchopogh F.S. and Khalifelu Z.A. "Analysis and Evaluation of Unstructured Data: Text Mining versus Natural Language Processing",978-1-61284-832-7/11,IEEE,2011 .
5. Jiang S., Lowd D. and Dou D. "Learning to Refine an Automatically Extracted Knowledge Base using Markov logic",12th International Conference on Data Mining,IEEE,912-917,2012.
6. Otsuka N. and Matsushita M. Graphical interface that supports user's trial and error process of text mining in proceedings of the 27th annual conference of the Japanese society for Artificial Intelligence (JSAI),2013.
7. Zhao Y., Zhiqiang K. N. , Lin J. and Wang X. "Text Sentiment Analysis Algorithm Optimization & Platform Development in Social Network", sixth International Symposium on Computational Intelligence and Design(ISCID),410-413,2013.
8. Lee S., Kim N. and Kim J..A "Multi-Dimensional Analysis and Data Cube for Unstructured Text and Social Media", Fourth International Conference on Big Data and Cloud Computing ,IEEE,761-764,2014.
9. Sunayama W., Takama Y., Nishihara Y., Kajinami T., Kushima M. and Tokunaga H. (2014). "Practical application in development and use of mining tools with total environment for text data mining", Journal of the Japanese Society for Artificial Intelligence, vol.29, no.1, 100-112,2014.
10. Delen D. and Oztekin A. "Introduction to Data ,Text and Web Mining for Managerial Decision Support" Mini –track 47th Hawaii International Conference on System Science(HICSS), IEEE,768,2014.
11. Richardet R., Chappelier J. C., Tripathy S. and Hill S..Agile Text Mining with Sherlok ,International Conference on Big Data,IEEE,1479-1484,2015.
12. Xuan L., and Nora E.G. "Text analytics for supporting stakeholder opinion mining for large scale highway projects", Procedia Engineering 145,518-524,2016.
13. Liu, Y., J. W. Bi, and Z. P. Fan "Ranking Products Through Online Reviews: A Method Based on Sentiment Analysis Technique and Intuitionistic Fuzzy Set Theory."Information Fusion 36: 149–161,2017.
14. R. Chen, Y. Zheng, W. Xu, M. Liu, and J. Wang, ''Secondhand seller reputation in online markets: A text analytics framework,'' Decis. Support Syst., vol. 108, pp. 96–106, Apr. 2018.
15. Achim Klein , Martin Riekert ,Velizar Dinev "Accurate Retrieval of Corporate Reputation from Online Media Using Machine Learning" vol.18 ,pp 43-46 IEEE ,2019

AUTHORS PROFILE

Devendra Kumar Mishra is a research scholar and presently working as an assistant professor at AMITY University Gwalior (MP) He has obtained his M.Tech degree from BIT Mesra ,Ranchi .He has qualified UGC-NET and GATE Examination various time.His area of research is data mining.

Dr. A. K Upadhyay is a Masters of Engineering from BITS, Pilani, with a PhD in Regression Testing from MNNIT, Allahabad. He has 7 years of Industry Experience with 18 years experience in teaching Engineering students across various Institutes. He has contributed in over 13 conferences across India with Two publications as the main author and various others as co-author. He earned the U.P. State Scholarship from 1979 to 1981 and received Merit Scholarships from U.P. Board and Luckhnow University. He is presently working as a Professor in ASET, Amity University, Gwalior, teaching and guiding students in B.E. and PhD, respectively.

Dr. Sanjiv Sharma works as an assistant professor in Department of Computer science Engineering and Information Technology in Madhav Institute of Technology and Science, Gwalior. He have 12 year of teaching and research experience. He has more than 70 research publications in various reputed international journals and conferences.His area of research is Network security, Data Mining and Social Network Analysis.