International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Naive Bayes Classification Technique for Opinion Mining in Data Mining

Sarpreet Kaur[1], Rasleen Deol[2]
[1]Research Scholar, of M. Tech in CSE, Global Institute of Management & Emerging Technologies, Amritsar, Punjab, India
[2]Assistant Professor, of M. Tech in CSE, Global Institute of Management & Emergent Technologies, Amritsar, Punjab, India

***Abstract*—** *The data mining is the technique which is applied to extract the useful information from the rough information. The opinion mining is the technique of data mining which is applied to extract the opinion of the users. In this work, the technique is been applied which will analyze the opinions of the students towards the collage. The technique is been applied which is based on three phase, the first phase has pre-processing phase in which input data is tokenized. In the second phase, the features from the input data are extracted. In the last phase the technique of classification is been applied which will classify the data into the classes. The proposed algorithm is implemented in python and it is been analyzed that proposed algorithm performs well in terms of various parameters.*

***Keywords*—** *Data Mining, Clustering, Classification, Naïve Bayes Classifier, SVM, Neural network, GA, Opinion Mining, MLP.*

I. INTRODUCTION

The process of extraction of useful information and patterns from large amount of stored data is known as data mining. There are other names for this process as well, such as knowledge discovery process in databases (KDD), information processing, knowledge extraction or data/pattern analysis. This technique is also known as data dredging, data fishing, and data snooping. Various types of data are analyzed with the help of certain data mining tools. The large amount of data which needs certain powerful data analysis tools are thus put for the here which is also known as the data rich but information poor condition [1]. There is an increase in the growth of data, its gathering as well as storing it in huge databases. It is no more in the hands of humans to do it easily or without the help of analysis tools. There are certain data archives created here which can be visited when the data is required [2]. The insightful, interesting and novel patterns of data are discovered from large-scale data sets using the data mining. The knowledge discovery in databases process is a very important step in data mining. Opinion mining can be characterized as a sub-discipline of computational linguistics that concentrates on extracting people's opinion from the web [3]. The current expansion of the web encourages clients to contribute and communicate by means of websites, recordings, and interpersonal interaction sites, and so on. Every one of these stages gives a tremendous amount of valuable information that researchers are interested to break down. Client's opinion is a noteworthy rule for the improvement of the quality of services rendered and enhancement of the deliverables. Blogs, review sites, information and miniaturized scale blogs give a decent comprehension of the reception level of the products and services [4]. Opinion mining concludes whether user's views are positive, negative, or neutral about a particular product, topic, event and so forth. Opinion mining and summarization process include three fundamental steps, first is Opinion Retrieval, Opinion Classification and Opinion Summarization. Review Text is recovered from networking sites. Opinion text in journals, analyses, remarks and so forth contains subjective report about topic. Remarks classified as they are negative or positive reviews. Opinion summary is produced in view of features opinion sentences by considering frequent features about a topic. Various classifiers are utilized within the opinion mining process.

a. Naïve Bayes Classifier: The most popular approach in the theory of supervised parametric classifiers is the quadratic discriminate function which utilizes the Bayesian approach [5]. The objective here it to propose a rule which allows assigning the future objects to a class when a set of objects is given for each class.

b. Support Vector Machine (SVM) Classifier: It is the main objective of SVM to determine the best function by maximizing the margin between the two classes. This is due to the fact that there are many such linear hyperplanes. The amount of space or distance amongst two classes is known as hyperplane [6]. The shortest between the closes data points to a point on the hyperplane is known as margin.

c. Decision Tree Classifier: Decision Trees (DTs) are a non-parametric supervised learning method utilized for classification and regression. The goal is to make a model that predicts the value of a target variable by learning simple decision rules gathered from the data features. It is a method for approximating discrete-valued target functions, in which the learned capacity is spoken to by a decision tree.

d. K-Nearest neighbor: K-Nearest neighbor classifiers depend on learning by analogy. The training samples are depicted by n dimensional numeric attributes. Every sample represents a point in an n-dimensional space [7]. Along these lines, the greater part of the training samples is stored in an n-dimensional pattern space.

e. Multi-layer Perceptron (MLP): The generally utilized feed forward ANN is the multi-layer perceptron classifier. For the purpose of providing simple computations, initially a single hidden layer is utilized. The numbers of neurons are basically involved for the simplification of the process.

II. LITERATURE REVIEW

LI Bing, et.al, (2014) stated in this paper that [8], majority of these works were unable to accurately extract clear indications of general public opinion from the ambiguous social media data. They additionally lacked the capacity to

Sarpreet Kaur and Rasleen Deol, "Naive Bayes classification technique for opinion mining in data mining," *International Research Journal of Advanced Engineering and Science*, Volume 2, Issue 3, pp. 44-47, 2017.

summarize multi-characteristics from the scattered mass of social data and use it to compile useful models. This paper proposes a novel matrix-based algorithm to determine the defined multilayered Twitter data. They additionally lacked the capacity to summarize multi-characteristics from the scattered mass of social data and use it to compile useful models, likewise lacked any efficient mechanism for managing the Vast data.

Dhanalakshmi V., et.al, (2016) explored within this paper [9] opinion mining utilizing supervised learning algorithms to discover the polarity of the student feedback based on pre-defined features of teaching and learning. The study conducted involves the utilization of a combination of machine learning and natural language processing techniques on student feedback data gathered from module evaluation survey results of Middle East College, Oman. The results are compared to locate the better performance with respect to various evaluation criteria for the different algorithms.

Gaurav Dubey, et.al, (2015) proposed in this paper [10], that the automatic summarization and classification is different for different domains and varies with the verify criteria. With the help of this paper, we are deliberating the efficiency of mining the consumer point of view (i.e. opinion mining) and experimenting its practicality in the mobile domain. This implementation in mobile domain will be based on three main steps which are, Applying Part-of speech Tagging(POST), Rule-Mining and identifying remarks, and Summarizing and showing the end results. The automatic summarization and classification is different for different domains and varies with the testing situations.

Ram Chatterjee, et.al, (2015) analyzed in this paper [11], the various methods available through which the twitter data can be extracted. An overall picture is provided of how each method is different from another for extracting tweets. In this research the method to extract twitter data utilizing various sentiment tools is defined. Each sentiment tool is different from other. The choice of sentimental tool to be used is entirely depend on user and his/her need. This classifier will be able to determine the tweets as positive, negative or neutral. This helps the end user to frame a decisive opinion on the query search.

Pooja Kherwa, et.al, (2014) proposed an approach is this paper [12] that diligently scans every line of data one by one, and produce a cogent summary of every review (categorized by aspects) alongside various graphical judgements. A unique application of this method is helping out product manufacturers or the government in gaging response. The paper aims to improve our system as discussed in the prior section, and further more preparing many pilot tests to further enhance the summarization results of the system to develop in more detail.

Lokmanyathilak Govindan et.al, (2015) proposed in this paper [13], study on the computational infrastructure for fast-feedback opinion mining. This is especially ambitious since, when encountering buggy software, customers would simply switch to free software with comparable functionality without giving any feedback. Our framework makes use of real-time Twitter data stream. These data streams are filtered and analyzed and fast feedback is obtained through opinion mining. The framework is based upon Apache Hadoop to deal with huge volume of data streamed from Twitter. The experiments have indicated 84% accuracy in the sentimental analysis. Our framework is therefore able to allocate rapid, inestimable feedbacks to companies.

III. RESEARCH METHODOLOGY

This work is based on the opinion mining in which the features of the input data are classified using the SVM classifier.

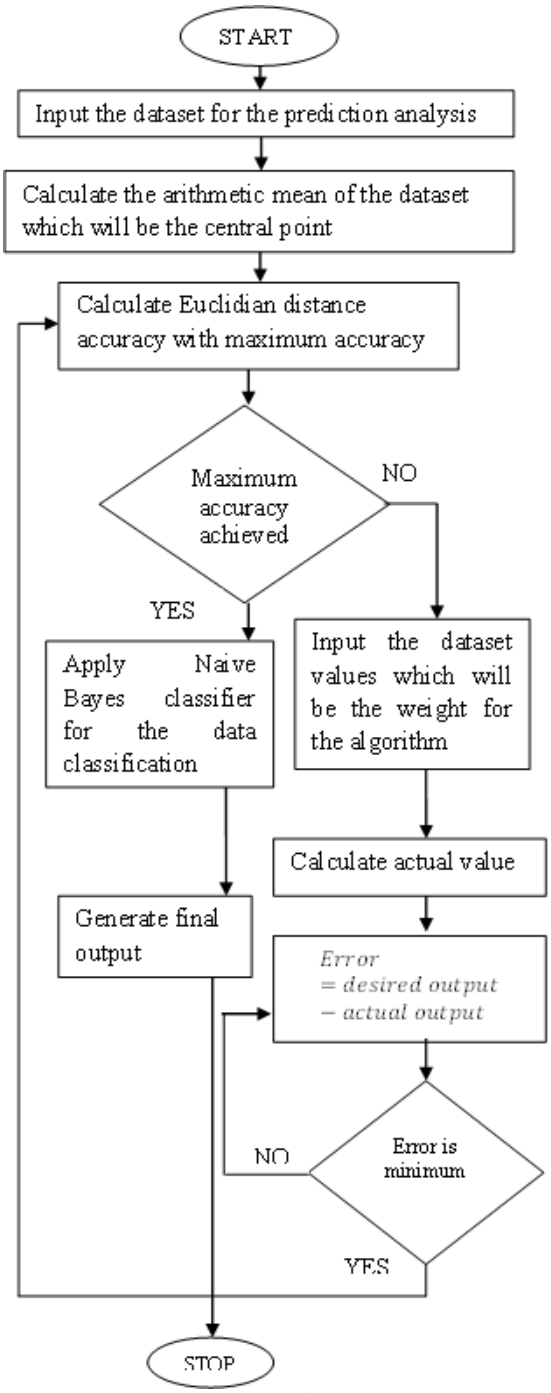

Fig. 1. Proposed flowchart.

Sarpreet Kaur and Rasleen Deol, "Naive Bayes classification technique for opinion mining in data mining," *International Research Journal of Advanced Engineering and Science*, Volume 2, Issue 3, pp. 44-47, 2017.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

The SVM classifier can be replaced with the naïve bayes classifier which is less complex and most efficient as compared to SVM classifier. A simple technique which is used for constructing the classifiers is known as the Naïve bayes classifier technique. For certain problem instances, the models are created which assign class labels to those problems. They are represented as vectors of feature values in which the class labels are drawn from certain defined set. It is a collection of algorithms which have single objective which is that all the Naïve Bayes classifiers predict that the value of a specific feature is not dependent on the value of any other feature provided in the class variable. In consideration of the supervised learning setting, there are various probability models for which the Naïve bayes classifiers are trained in a very efficient manner. The method of maximum likelihood is utilized for parameter estimation in case of Naïve Bayes models. Without accepting Bayesian probability or utilizing any Bayesian methods, the Naïve Bayes model can work. In various complex real-world situations, the Naïve Bayes classifiers have provided efficient results. Even if they have a naïve design and oversimplified assumptions the results are very efficient. Only a small number of training data is required for estimating the parameters which are required for classification. This is an important merit of the Naïve Bayes technique.

IV. EXPERIMENTAL RESULTS

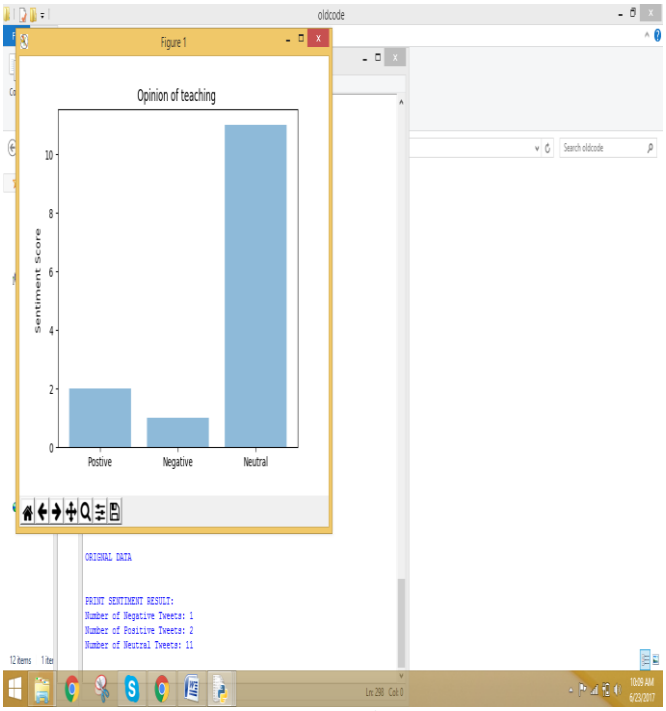

Fig. 2. Graphical analysis of teaching opinion.

As shown in the figure 2, the opinion of the teaching is analyzed and it is been analyzed that number of positive, negative and neural opinions are plotted in the form of bar graph.

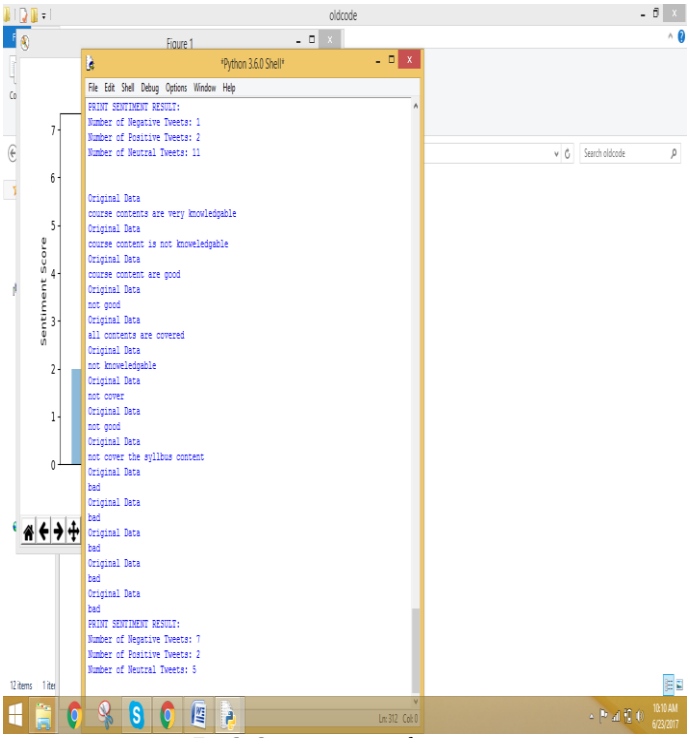

Fig. 3. Opinion mining of courses.

As shown in the figure 3, the dataset is considered in which the various attributes are considered like time stamp, lab assessments etc. In the above figure, the opinion of the courses is shown.

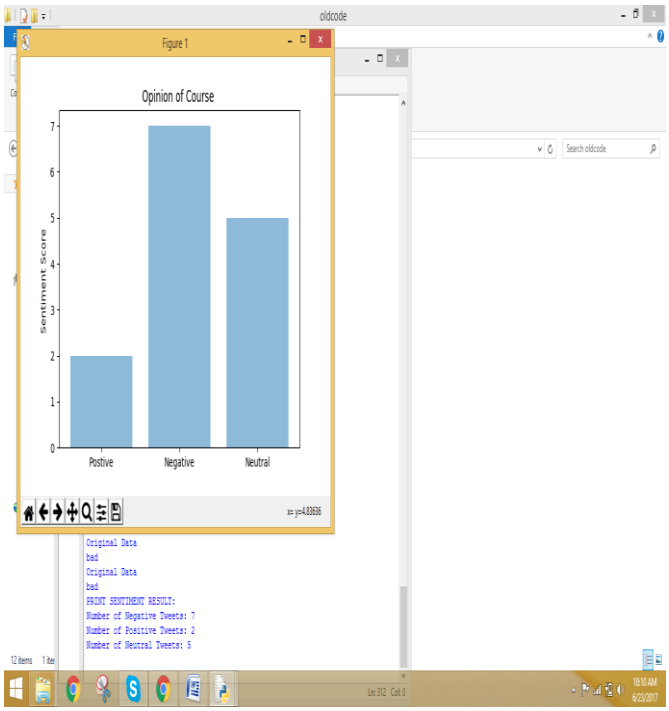

Fig. 4. Graphical Analysis of courses opinion

As shown in the figure 4, the opinion of the courses is analyzed and it is been analyzed that number of positive,

Sarpreet Kaur and Rasleen Deol, "Naive Bayes classification technique for opinion mining in data mining," *International Research Journal of Advanced Engineering and Science*, Volume 2, Issue 3, pp. 44-47, 2017.

negative and neural opinions are plotted in the form of bar graph.

V. CONCLUSION

The opinion mining is the technique which is applied to analyze the opinions of the users. This work is based on the opinion mining to analyze the collage performances. In this work, it is been concluded that for the opinion mining the three steps has been followed. In the first step, the pre-processing is applied which will tokenize the input dataset. In the second step, the feature extraction process is started which will extract the features from the dataset. In the last step, the technique is applied which will classify the dataset into three classes which is neutral, negative and positive. The proposed technique is implemented in python and it has been analyzed that performance of the method increases in terms of various parameters as compared to existing technique.

REFERENCES

[1] Q. Miao, Q. Li, and R. Dai, "AMAZING: A sentiment mining and retrieval system," *Expert Systems with Applications*, vol. 36, issue 3, Part 2, pp. 7192-7198, 2009.

[2] Q. Miao, Q. Li, and D. Zeng, "Fine-grained opinion mining by integrating multiple review sources," *Journal of the American Society for Information Science*, vol. 61, issue 11, pp. 2288-2299, 2010.

[3] S. M. Mudambi and D. Schuff, "What makes a helpful online review? A study of customer reviews on Amazon.Com," MIS Quarterly, vol. 34, issue 1, pp. 185-200, 2010.

[4] R. Prabowo and M. Thelwall, "Sentiment analysis: A combined approach," *Journal of Informetrics*, vol. 3, issue 2, pp. 143-157, 2009.

[5] S. Senecal and J. Nantel, "The influence of online product recommendations on consumers' online choices," *Journal of Retailing*, vol. 80, issue 2, pp. 159-169, 2004.

[6] A. Sharma and S. Dey, "A comparative study of feature selection and machine learning techniques for sentiment analysis," *Proceedings of the 2012 ACM Research in Applied Computation Symposium*, pp. 1-7, 2012.

[7] P. J. Sher and S. H. Lee, "Consumer skepticism and online reviews: An elaboration likelihood model perspective," *Social Behavior and Personality: An International Journal*, vol. 37, issue 1, pp. 137-143, 2009.

[8] L. Bing and K. C. C. Chan, "A fuzzy logic approach for opinion mining on large scale twitter data," *IEEE/ACM 7th International Conference on Utility and Cloud Computing (UCC)*, 2014.

[9] V. Dhanalakshmi, D. Bino, A. M. Saravanan, "Opinion mining from student feedback data using supervised learning algorithms," *IEEE 3rd MEC International Conference on Big Data and Smart City (ICBDSC)*, 2016.

[10] G. Dubey, A. Rana, and N. K. Shukla, "User reviews data analysis using opinion mining on web," *IEEE International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE)*, 2015.

[11] R. Chatterjee and M. Goyal, "Tactics of twitter data extraction for opinion mining," *IEEE 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2015.

[12] P. Kherwa, A. Sachdeva, D. Mahajan, N. Pande, and P. K. Singh, "An approach towards comprehensive sentimental data analysis and opinion mining," *IEEE International Advance Computing Conference (IACC)*, 2014.

[13] L. G. Sankar Selvan and T.-S. Moh, "A framework for fast-feedback opinion mining on twitter data streams," *IEEE International Conference on Collaboration Technologies and Systems (CTS)*, 2015.

Sarpreet Kaur and Rasleen Deol, "Naive Bayes classification technique for opinion mining in data mining," *International Research Journal of Advanced Engineering and Science*, Volume 2, Issue 3, pp. 44-47, 2017.