Machine Learning Based Sentiment Analysis Implementation for Multiple Reviews Classification

C. Ramakrishna and Dr. Pramod Pandurang Jadhav

Department of Computer Science and Engineering, Dr. A. P. J. Abdul Kalam University, Indore 452010, India
Corresponding Author Email: cramakrishna537@gmail.com

Abstract: The sentiment is an attitude, judgment or thought comes from feeling. The sentiment analysis is the process of analysis of online writing prices and which are having the emotional tone, that tone can be determined. In the past few years, in the internet social media content is enhanced with high quantity and with group discussions the opinions are expressed about any product or service. Therefore machine learning based sentiment analysis implementation for multiple reviews classification is proposed in this paper. The huge amount of text of sentiment can be identified by the sentiment analysis. The performance of BoW (bag of words) due to some fundamental deficiencies, these are sometimes unlimited in the polarity handling shift problem and other some challenges are hidden state representations, opinions of quality, categorization of polarity and so on. These challenges will be focused on the dual sentiment analysis along with sentiment processes with the all perspectives. The total proposed work will experimented on the Amazon product reviews especially mobile devices reviews. Naïve bayes and support vector machine (SVM) learning techniques are compared in terms of the precision, recall, F1-measure, accuracy for measuring the neutral, positive, negative reviews.

Keywords: Sentiment analysis, reviews, multiple reviews classification, naïve bayes, support vector machine (SVM).

I. INTRODUCTION

The web technology enhancement can be changed based on the perspectives of individual expressions [1]. For analyzing the data or items by the online shopping or else booking the movie tickets to watch the movie in the theaters, the individual perspective is used [2]. Using Facebook, Twitter for tweets, through posts users are together interfacing nowadays.

People opinions or people reviews are the big component of Web in the services of digital communication technology. In online business, which has more advantages than other businesses because of competitive intelligence have sourced, that source is the mining of people's opinion [3]. In research broad applications the sentiment analysis plays an important role which is advertising the organizations, and they are investigating their products and services for improvement of institute brand [4]. The term sentiment analysis is defined as the algorithmic process of grouping or identifying the opinions in the form of text to compute the attitude of users towards the subject of an expressing an opinion is known as sentiment analysis [5]. To analyze generally a product or service can be liked or disliked by the people in similar a movie can be liked or disliked by them.

If a user purchase a product then he give the reviews of the particular product based on his experience. Then these reviews are help to the future customers whether they buy the product or not. The reviews of the particular product can be available in the online and these reviews are significantly utilized for the customer's opinions [6]. Sentiment classification or sentiment analysis concept was rise in the year of 2000 its goal is to estimate the text with polarities of the sentiment opinions or thoughts of the user's example negative or positive. In general, in the form of unstructured data can be represented.

There are two types of sentiment analysis they are, one is classification technique and other one is information based methodology [7]. In first type of methodology, the expensive database which is required predefined feelings. For sentiment reorganization it has proficient information. Again the Machine learning approach is taken for dataset utilization and collection of information can be test to build the classifier. It is most preferable straight forward method over the knowledge based method. In the sentimental analysis, the calculation of improvements is quite difficult. The negative sentiment word can be contingent on the circumstances firstly. Similarly individuals cannot express the conclusions therefore the test is done as secondly. The connection between the outcomes of the reviews and literary audits are occurred in sentiment mining comprehend.

Sentiment analysis can be used to differentiate followers and customers depend on their attitude for specific movie or product or brand by using the reviews.

The review of product is negative or positive and user E-mail is satisfied or not are can be identified. There are four types of feature extractions are semantic feature, syntactic feature, stylistic feature and link based feature [8]. The first two features are most commonly used. The word tags, phrases, punctuations and patterns are utilized by the syntactic feature. The relationship between signs, symbols and words are working by the semantic feature. Through the accurate

language, it is easy to know the human expressions with the help of Linguistic semantics.

II. MACHINE LEARNING AND SENTIMENT ANALYSIS

To dominating the sentiment analysis the two main tools are used namely, natural language processing (NLP) and machine learning [9-10]. The machine learning is having a lot of focus and remaining does not. The decision trees, neural networks, gradient boosting machines and random forests are the machine learning models [11]. These are used in prediction and accept the feature vector. These models are can be learned in a good manner so the required output is provided with the set of feature vectors.

Most of researchers are having more focuses on the feature selection and feature extraction in the applications for datasets represents as 10's or 100's of 1000's of availability of variables [12]. These areas includes the internet documents text processing and combinatorial chemistry, generate the expression analysis of array. The term feature selection can be defined as the process of subset selection of relative features. The Feature selection objective is the three-fold, there are prediction performance can be improved by the classifiers/predictors, it provides more cost effective predictors and faster, good understanding of the underlying process is also provided that data can be generated.

To analyze and discover the attitudes and sentiments of the consumer behavior on the products that are purchased by the user, the algorithms and data mining tools are used under the polarity shift circumstances. The BoW model can have more impact to give the better performance of the machine learning based systems. Moreover it is important to improve performance and machine learning execution models for obtaining the accurate results. The supervised learning is also known as classification. There are some linear classifiers as support vector machines, logistic regression, neural networks, naive bayes classifier, decision trees, random forest, are the machine learning classification algorithms [13].

2.1 Vulnerabilities and threats in the existing systems:-

The process of classification and opinion mining of the product review by taking the keyword, therefore quality of the product is the major considerable factor [14]. When the pre-processing operation is performed it can be reviewed by the dataset, and later vectorization includes classification and (TF-IDF) "Term Frequency — Inverse Document Frequency" can be carried out. The meaningful information will loss if the characters reputation is eliminated. To carry feature extraction process, opinion mining and polarity categorization are two important factors that the authors must understand. The major concerned factor is if increase the data then the complexity and size is also increased.

2.2 Challenges to be addressed:

For sentiment analysis, the literature survey based challenges of the systems are following:

- Having particular data accuracy
- Each text data can be captured at high level semantics
- The opinions quality was not guaranteed because there are fake, meaningless, irrelevant opinions and so on.
- Categorization and polarity shifting [15].
- Eliminating the polarity shifting: It affects the sentiment analysis based machine learning system's classification performance.

III. MACHINE LEARNING BASED SENTIMENT ANALYSIS IMPLEMENTATION

The model framework on the sentiment analysis implementation based on machine learning for multiple reviews classification can be represented in below Fig. 1. The detailed explanation of workflow is described below.
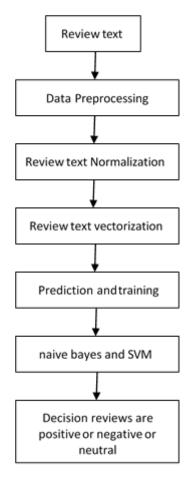

Fig.1: SENTIMENT ANALYSIS BASED ON MACHINE LEARNING WORKFLOW

3.1 Review text: The system can have the review set from the shopping websites like flip kart, amazon and so on. The results are negative and positive. These results are classified by using SVM and Naïve Bayes and so on.

3.2 Data preprocessing: There are some preprocessing techniques thus techniques are stemming, sentence tokenization and stop- word removal and so on. It can help in reviews filtration in required formation and that format is

leads the reduction in uncertainty and difficulty computation for next processing.

3.2.1 Sentence Tokenization: Here text can be divided into tokens which are includes the numbers, words, elimination of white spaces, symbols, individual words and punctuation marks from the sentences etc. the sentence tokenization can be applied and it trains the different languages includes the English language.

3.2.2 Stop Word Removal: The frequent words can be found in sentences or reviews can be removed. For example very, more, too and so on

3.2.3 Stemming: The word stemming can be defined as process of removing the affixes, prefixes to convert the w into its root or stem form. The stemming porter algorithm rule- based algorithm which is mostly used and it introduced by the Martin Porter. It defines only consonants letters not the vowels. This algorithm is applie perform the normalization of text and it converts the tex the standard format.

3.3 Review normalization and vectorization: The text normalization is transforming the text into single canonic form process and it is suitable for the extraction of featur Before processing, the normalizing text can allow for the separation of input and performing the operation. Late reviews on movie are normalized by creating the data featu using TF-IDF, word count vector vectorization, bi-gram techniques. Apply the "Term Frequency — Inverse Document Frequency" (TF-IDF) to the original data. To extract the more stable cluster this includes the voting mechanism. The results are depends on clustering process.

3.4 Training Model: In the training stage, actual training reviews are reversed in opposite for generating the new reversed review dataset that can be referred as reversed training and actual training set. Reversed and actual reviews are corresponding to each other in the technique of data expansion. All possible probability combinations are collected from the reversed reviews and actual training dataset to prepare training classifier.

3.5 Prediction Model: For predicting the final result the both reversed and actual data sets are used. This can also predict how the negative and positive reversed and actual reviews are.

3.6 Classification:

3.6.1 Naive Bayes classifier: Machine Learning is a classification algorithm. One of the algorithms is namely, Naive Bayes used for the classification of text and consideration factor when the input feature dimensionality is high. The probabilistic Bayes theorem can be introduced namely, Thomas Bayes and it is foundation of Naive Bayes classification algorithms. It is also called as naive because of the Bayes theorem depends on the feature assumption that are independent of each other, sometimes it is not correct is a naive. By using Bayes theorem the probability of posterior P(C|X) can be calculated from P(X|C) , P(X) and P(C).

$$P\left(\frac{C}{X}\right) = \frac{P\left(\frac{X}{C}\right)P(C)}{P(X)}$$

Where,

P(C|X) = posterior .

P(X|C) = Likelihood.

P(C) = Proposition prior probability.

P(X) = Evidence prior probability.

3.6.2 SVM (Support Vector Machine) Classifier: This machine has the revolutionized classification and the global regression. the concept of this algorithm is best fitting of hyper plane selection which efficiently divides different data points. However, the data points present in linear world will be divided by linear kernel, so this algorithm is flexible and help to handle the multi-dimensional world. The poly, kernel, and linear techniques used in the data points restructuring according to dataset adversity and length.

3.7 Decisions on reviews: The threshold is applied to segregate the reviews according to their sentiments. For calculate negative_ precision and positive_precision with the ranges from 0 to 1 in the given data are supported by the threshold value. For positive and negative reviews, we determine the class labels the threshold can be set to 50%.

For negative sentiments: <= -0.5; it comes under the Negative Class.

For positive sentiments: >= 0.5; then it comes under the Positive Class.

For Neutral: > -0.5 < 0.5; if it lies in between then it comes under the neutral class.

IV. EXPERIMENTAL RESULTS

Database of the Amazon product has the different datasets there are clothes, furniture, books and so on. It provides the customer details, and reviews of customers, and ID of customer based on the specific product. The text or reviews are generated by the training set and system overall performance is verified by using testing process. These may helpful in the algorithms of machine learning, to make predictions or decision making of particular review.

4.1 Evaluation of performance by using parameters:-

To estimate the presenting system performance, different standards classifiers evaluation metrics are used such as f-measure, recall, precision. The explanations are follows,

Precision:

It is the positive predictive value or the precision can be defined as the accurate positive score number can be divided by the number of positive score predicted by classification algorithm.

$$Precision = \frac{TP}{(TP + FP)}$$

Recall:

It is a measure called sensitivity, recall or TP rate which is defined as the ratio of positive instances to the combination of true positive and false negatives.

$$Recall = \frac{TP}{(TP + FN)}$$

F1-score:-

It is the measured by using recall and precision. It will be one for classification algorithm of good performance, and it will be zero for the bad performance of classification algorithm, the representation is follows.

$$F1 - Score = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Accuracy:
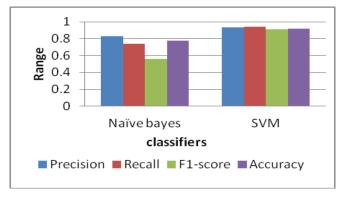
$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

The term accuracy can be defined as the ratio of correctly predicted reviews number as true positive and true negative to the reviews total number which are obtained in the process.

TP is termed as the true positive, and TN is termed as the true negative, FP is termed as the false positive, FN is termed as the false negative. Naïve Bayes and SVM classifiers performance parameters comparisons are described in below Table 1 as:

Table 1: CLASSIFIERS PERFORMANCE PARAMETERS COMPARISONS

Parameters	Naïve bayes	SVM
Precision	0.83	0.93
Recall	0.74	0.94
F1-Score	0.56	0.91
Accuracy	0.78	0.92

The graphical representation of accuracy, recall, F1-score and precision is represented in below Fig. 2.

It depends on threshold, negative_ precision, positive_ precision values computed, the Naïve Bayes and SVM algorithms used to classify those reviews into neutral/negative/positive which are used for predicting and training the reviews which all these are generated with the help of sentiment analysis. The comparison of memory of support vector machine and naive bayes algorithms for the various thresholds values are represents as in Fig. 3. In the graph Y-axis shows the bytes of memory and X- axis shows the respective algorithm. The naive bayes require large amount of memory than the SVM method.

Fig. 3: MEMORY COMPARISION FOR NAÏVE BAYES AND SVM CLASSIFIERS

Time comparison of Naïve Bayes and SVM algorithms with different sizes are represented in the below Fig. 4. In below figure, Y- axis represents the time in milli seconds whereas the X-axis represents the respective algorithm. SVM method takes the less time.
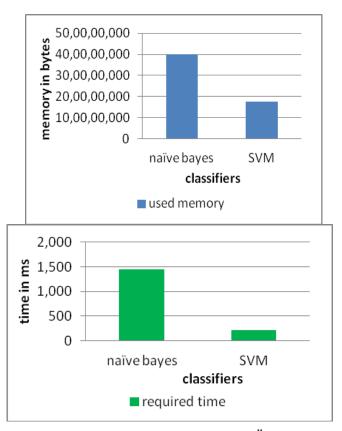

Fig. 4: TIME COMPARISION FOR NAÏVE BAYES AND SVM CLASSIFIERS

On comparing these two algorithms, it was concluded that SVM accuracy is better (92%) compared to the Naïve Bayes (78%).

V. CONCLUSION

Here machine learning based sentiment analysis for multiple reviews classification was analyzed. Therefore this experiment uses the Amazon product data as input data or review data. The data is preprocessed first then classified by using Naïve Bayes and SVM machine learning methods. The comparative performances are done in between the SVM and

Naïve bayes classifiers based on the precision, accuracy, recall, F1-score. Memory and time comparisons are also done. Results show that support vector machine (SVM) is good compared to Naive Bayes algorithm in terms of accuracy, memory and time to produce the good reviews to the users. The presenting method in this paper experimental results analysis are shows that, the work presented in this have good contact on the accurate prediction value used for the final decision making based on the reviews which can be taken from the customers.

VI. REFERENCES

1. Mayura Kulkarni SCET, Vaibhav Chaudhari, Sumit Sakarkar, Aditya Veer, Tanmay Gaurkar, "Web Personalisation based on User Interaction : Web Personalisation", 2021 Third International Conf. on Intelligent Comm. Techs. and Virtual Mobile Networks (ICICV), Year: 2021

2. Hanqing Liu, "A Satisfactory Detection Framework on Online-shopping Comments", 2021 International Conf. on Computer Comm. and AI (CCAI), Year: 2021

3. Tran Thi Thu Ha, Nguyen An Te, Thai Kim Phung, "A ML approach for opinion mining online customer reviews", 2021 21st ACIS International Winter Conf. on Software Engg. , AI, Networking and Parallel/Distributed Computing (SNPD-Winter), Year: 2021

4. B. Bharathi, R Lavanya, "Systematic analysis of Movie Recommendation System through SA", 2021 International Conf. on AI and Smart Systems (ICAIS), Year: 2021

5. K. Sathiya Priya, A. Poornima, "A Comparative Sentiment Analysis Of Sentence Embedding Using ML Techns.", 2020 6th International Conf. on Advanced Computing and Communication Sys. (ICACCS), Year: 2020

6. Sevcan DOĞRAMACI, Mustafa ÇATALTAŞ, Kasım ÖZTOPRAK, Semih YUMUŞAK, "Extraction of Product Defects and Opinions from Customer Reviews by Using Text Clustering and SAs.", 2020 IEEE International Conf. on Big Data (Big Data), Year: 2020

7. Honglei Zhu, Kai Dong, Zhigang Xu, "Text sentiment analysis method based on attention word vector", 2020 International Conf. on Modern Edu. and Information Managt. (ICMEIM), Year: 2020

8. Van Cuong Tran, Huyen Trang Phan, Dosam Hwang, Ngoc Thanh Nguyen, "Improving the Performance of Sentiment Analysis of Tweets Containing Fuzzy Sentiment Using the Feature Ensemble Model", IEEE Access, Volume: 8, Year: 2020

9. Zulfiqar Ali Solangi, Yasir Ali Solangi, Amna Abro, Samreen Aarain, Asadullah Shah, Ghulam Ali Mallah, "Review on NLP and Its Toolkits for Opinion Mining and Sentiment Analysis", 2018 IEEE 5th International Conf. on Engg. Technos. and Applied Scie. (ICETAS), Year: 2018

10. Bernd Markscheffel, Markus Haberzettl, "A Literature Analysis for the Identification of ML and Feature Extraction Methods for SA", 2018 Thirteenth International Conf. on Digital Information Managt. (ICDIM), Year: 2018

11. Sfenrianto Sfenrianto, Achmad Bayhaqy, Emil R. Kaburuan, Kaman Nainggolan, "SA about E-Commerce from Tweets Using DT, KNN, and Naïve Bayes", 2018 International Conf. on Orange Technos. (ICOT), Year: 2018

12. M. Ali Fauzi, Fachrul Rozy Saputra Rangkuti, Eka Dewi Lukmana Sari, Yuita Arum Sari, "SA on Movie Reviews Using Ensemble Features and Pearson Correlation Based Feature Selection", 2018 International Conf. on Sustainable Information Engg. and Tech. (SIET), Year: 2018

13. Henry E. L. Cagnini, Rodrigo C. Barros, Willian Becker, Joonatas Wehrmann, "A character- based convolutional neural network for language-agnostic Twitter SAs", 2017 International Joint Conf. on Neural Networks (IJCNN), Year: 2017

14. Lin Sun, "Research on Product Attribute Extraction and Classification Method for Online Review", 2017 International Conf. on Industrial Informatics - Computing Tech., Intelligent Tech., Industrial Information Integration (ICIICII), Year: 2017

15. Bela Joglekar, Sayali Zirpe, "Polarity shift detection approaches in sentiment analysis: A survey", 2017 International Conf. on Inventive Sys. and Control (ICISC), Year: 2017