International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 2, ISSN No. 2455-2143, Pages 93-98
Published Online June 2022 in IJEAST (http://www.ijeast.com)

SENTIMENT ANALYSIS USING R PROGRAMMING

Sahil Kumar
Department of Mathematics
Chandigarh University, Gharuan, Mohali, Punjab.

Abstract: **The customers, when they buy the products online from e-commerce websites, often tend to rate these products and give reviews on that product. This rating/review system often helps the other potential customers to decide whether to purchase that product or not. However, reading all of the available reviews on a particular product, often make the customer invest a lot of time in this process as abundance of places such as blogs, review sites etc. contain reviews. The process of sentiment analysis aims at reducing this time of the customer by displaying the data in a compact format in the form of means, analysis score or simply histograms. The sentiment analysis procedure shown in this paper can be extended to the reviews of products in different domains. The experimental results have shown that this method exhibits better performance.**

I. INTRODUCTION:

Nowadays, people prefer online shopping of products from the various e-commerce websites because this helps them to save time and offers them a wider range of selection at their convenience. Focusing our selection on the basis of the consumer reviews of other customers helps us to save time and filter the products based on the reviews. But most of the reviews often contain very less details about that particular product and have more of other sentences which are not useful to the potential buy er. Hence, we need to extract only that information required by the customer and trim out other undesired information, so that theycan be displayed on compact devices such as mobile phones which are often handy among people. So, in this work sentiment analysis is projected for this purpose. Sentiment analysis is one of the stages of opinion mining. In sentiment analysis, we classify the particular word into positive, negative or neutral in order to predict the emotion of the speaker or reviewer towards the product. These reviews are given based on the customers ‟judgment of the product and experiences with it afterusing it. So, basically, the sentiment analysis of the reviews of a set of customers are being used by one customer to analyze the product and make up his mind whether to buy the product or not.

Opinion mining deals with natural language processing in order to identify the important keywords in the given sentences. Opinion mining consists of three stages- Opinion Retrieval, Opinion Classification and Opinion Summarization. Opinion retrieval aims to extract the keywords containing the opinions or comments concerned to a particular subject of the user‟s interest. Opinion retrieval is followed by opinion classification which deals with classifying the extracted keywords into positive, negative or neutral based on an existing dictionary. This can also be referred to as polarization of the words. The next stage, opinion summarization is the process of reproducing summaries from the extracted polarized keywords. This paper aims at gathering important opinion words from the product reviews given by the existing customers taken from the Amazon review dataset, finding their orientation i.e.- positive, negative or neutral and finally outline the views of public to a potential user which enhances his decision making process on whether to choose the product or not. All this has been done using R Programming language. This paper aims at gathering important opinion words from the product reviews given by the existing customers taken from the Amazon review dataset, finding their orientation i.e.- positive, negative or neutral and finally outline the views of public to a potential user which enhances his decision making process on whether to choose the product or not. All this has been done using R Programming language.

II. RELATED WORK:

Min Wang, et al [1] have proposed an approach to realize polarity analysis of new words, and also implement quantitative computation of sentiment words and automatic expansion of polarity lexicon. Their experimental results showed the feasibility and effectiveness of their approach. Their future work includes making fine-grained sentiment analysis possible from the attribute level with an automatically built polarity lexicon.

Basant Agarwal, et al [2] worked on sentiment-rich phrases that were obtained using POS based rules and dependency relations that were capable of extracting contextual and syntactic information from the document. Their experiments prove that by using POS p atterns for the phrases, performance of sentiment analysis can be improved. In future, they would like to explore more patterns using POS Tagging to get better results.

Rui Yao, et al [3] have applied sentiment analysis and

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 2, ISSN No. 2455-2143, Pages 93-98
Published Online June 2022 in IJEAST (http://www.ijeast.com)

machine learning concepts to study the relationship between the online reviews for a movie and the box office collection of the movie. It takes into account only positive and negative reviews leaving behind the neutral ones. Further experiments with larger datasets can be carried out to train and test the model. The comparisons among different movies can also be considered.

Kai Gao, et al [4] conducted tests using SVM based algorithm to do the alternative structural formulation of the SVM optimization problem for classification. Two different datasets which are, micro blogging and e- commerce were used to evaluate the performance. Their experiments proved that the proposed approach, which includes feature extraction & selection and SVM, is effective in microblogging multi-class sentiment classification and e-commerce sentiment classification.

In their paper, Pooja Kherwa, et al [5] gather opinions and review data from e-commerce websites, social networks, popular portals and blogs to find out what exactly people are talking about and the sentiment they are expressing. The Scoring Algorithm, which they have used, scans every line of data and gives a summary and also a graphical representation if required. The efficiency of this algorithm can be improved upon if a self-learning system can be implemented.

Giovanni Acampora, et al [6] introduce an innovative framework consisting of methods to analyze efficiently sentiments of the customer reviews and compute their corresponding numerical data so that companies can plan their future projects. The dimension and imprecision ratings of data are calculated. As a conclusion, they propose a system to reduce the uncertainty between the reviews to validate the reviews to get the useful reviews and produce a more accurate system.

In their work, Siddharth Aravindan, et al [7] put forward a system that obtains the product features automatically from the reviews and divide it into positive and negative. It does feature extraction followed by polarity classification using association rules and supervised machine learning. Their future work is to investigate some more features for opinion mining, and to make use of classifiers that would enhance their work.

Yan Wan, et al [8] conducted a fine grain sentiment analysis to get better results of the customer reviews and use general methods to crawl reviews and find implicit features based on POS rules. They believe that it can help producers make improvements clear and discover niche market and can also help the consumer understand the advantages as well as the advantages of the target product and hence make a wise selection.

III. METHODOLOGY:

The modules of the system designed are illustrated in Figure 1 and explained in the subsequent sections.

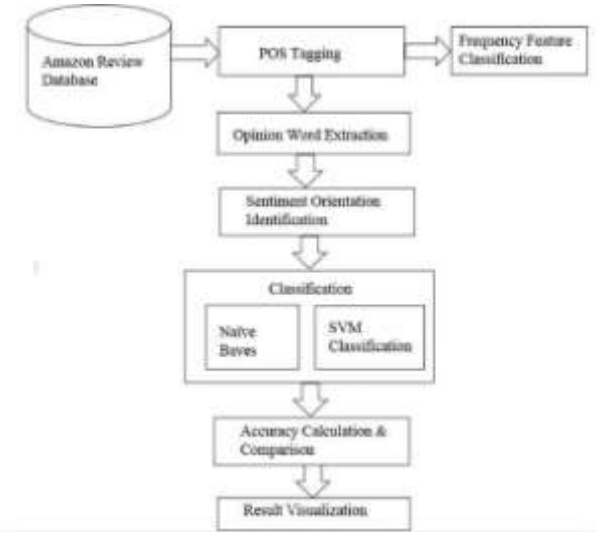

Fig 1-Flowchart of the entire process

POS Tagging:

Part-of-speech Tagging (POS Tagging) [9, 10] is the process of attaching every word of a file (corpus) with its corresponding part of speech, based on its definition and its relation with the adjacent phrases and words. The outcome of this process, is all the words along with their equivalent POS tag from which the words can be identified as nouns, adjectives, pronouns, verbs, etc.

This process is essential because, in the reviews, the product features are often described in the form of nouns or noun phrases and the sentiment regarding those nouns is in the form of adjectives. Therefore, extract ing the noun with its corresponding adjective allows us to identify a feature of the product and the customer‟s emotion towards it.

The process of POS tagging involves converting each word into Unicode Transformation Format (UTF-8) so as to encode all the character vectors into 8 bit code units in order to avoid complications of byte order marks. This is followed by tokenization of all the individual characters to convert them into individual tokens. The next step is to remove the stop words which are the most common words used in a language. The program has been written in such a way that the user can add his desired words into the list of existing stop words. The final step is to apply sentence token

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 2, ISSN No. 2455-2143, Pages 93-98
Published Online June 2022 in IJEAST (http://www.ijeast.com)

annotations and word token annotations to the customer reviews.
From the analysis of the dataset of a particular model of a digital camera, one of the review obtained was- the /DT macro-mode/NN is/VBZ exceptional/JJ and/VBP the/DT pictures/NNS are/VBP clear/JJ
The nouns i.e. - macro-mode and pictures and identified and the POS Tag NN is appended to them which signifies that they are nouns. The adjectives i.e. - exceptional and clear are identified and the tag JJ is appended to them signifying that they are adjectives. The extracted nouns can be used for frequent feature detection whereas the adjectives can be used to identify the polarity of the review.

Frequent Feature Identification:
As a result of the above discussion, the features of a product are described in the form of nouns. Therefore, once, the nouns and adjectives are generated, the nouns along with their respective adjectives can be used to find the most frequently repeating positive (Table 1) and negative feature (Table 2) which the customers have reviewed. From the analysis of the entire dataset of the reviews the following table of the frequent features list can be obtained:

Table 1: Positive Features

Positive Features	Frequency(No of times)
Battery	19
Picture Quality	11
Zoom	11

Table 2: Negative Features

Negative Features	Frequency(No of times)
Lens	5
Auto	4
Price	4

Opinion Word Extraction:
The process of POS tagging is followed by opinion word extraction which is the process of extracting all types of adjectives (i.e.-comparative & superlative) in order to find the customer‟s emotion towards the product i.e.- positive or negative. Whether the word is positive or negative can be determined by comparing each adjective extracted to a dictionary consisting of the list of positive and negative words. For example in the review: awesome camera with great print quality but bad zoom
The words awesome and great can easily be identified as a word having positive sentiment and therefore are counted as positive words whereas bad can be identified as a word having negative sentiment and hence is counted as a negative word.

Sentiment Orientation Identification:
The next step is to calculate the sentiment score of each review. The sentiment score helps us to classify the total score of each review and therefore the positive, negative and neutral reviews can be identified.A score of +1 is assigned to a positive word whereas -1 is assigned to a negative word. The total review score can be calculated by summing up the individual scores of all the adjectives in a review. In thispaper, the reviews that have score greater than 0 are classified as positive, reviews having a score of less than 0 are classified as negative and a score of 0 makes the review a neutral review. For example in the following reviews:
excellent compact digital camera!
In the above sentence, the words excellent and compact are given the score of +1 each giving it an overall score of +2, therefore making it a positive review.
the main drawback of this camera is its lens.
In this sentence, the word drawback gives it a score -1 making the total sentence score -1, therefore making it a negative review.
amazing camera but comes at an exorbitant price
In the above sentence, the word amazing is given a score of +1whereas the word exorbitant is given a score of -1, therefore giving a total score of 0 and makes it a neutral review.

Classification:
Classification should be applied on the data so that the can analyse existing data can be analysed to predict the future trends of the data. There are many classification algorithms which make this job easy. Two classification algorithms, Naïve Bayes Classification [11] and SVM Classification [12] in R Programming language are used in this paper. For applying classification algorit hms, the entire dataset is divided into training set and testing set. The training set is used to predict the results of classification on the dataset while the test ing set is used to validate the results. In our experiments, we have used 33% of the dataset as testing set and the remaining as the testing set.

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 2, ISSN No. 2455-2143, Pages 93-98
Published Online June 2022 in IJEAST (http://www.ijeast.com)

Naïve Bayes:
Naïve Bayes method works on the lines of Bayes Theorem of probability to predict the class of the data. A Naïve Bayes classifier [13] estimates that the presence of one feature in a class is not related to the presence of any other feature. It is a highly scalable problem, requiring a number of parameters linear in thenumber of variables in a learning problem. Naïve Bayes is used so that it can give easy, fast and accurate results compared to other classification algorithms

SVM Classification:
Support Vector Classification [15, 16] is a machine learning algorithm mainly used for classification and regression analysis. The goal of this algorithm is to find a decision boundary between the two classes that is located at the maximum distance from any point in the training data. SVM is mainly used because by introducing a kernel, we can gain flexibility in threshold.

Accuracy Calculation and Comparison:
Once the results of both the algorithms are acquired, the acquired results have to be compared in order to check which algorithm gives a better performance. From our analysis, we come to a conclusion that SVM Classification yields better results compared to Naïve Bayes as SVM has a higher accuracy compared to Naïve Bayes.

Result Visualization:
The final result of classification can be represented in any format like bar graphs, histograms, trees and tables. A histogram is used to show results of sentiment classification. Receiver Operating Characteristic Plot (ROC Plot) to show the results of our analysis. An ROC Plot is a graphical plot that depicts the results of our classification. In this, the true positive rate is plotted against the false p ositive rate.

IV. DATASET, EXPERIMENTAL RESULTS AND ANALYSIS:

Dataset used:
The dataset used for this experiment is the review dataset of a particular model of popular digital camera from Amazon collected over a couple of years. This collected data has been segregated into positive and negative reviews using R Programming.

Implementation:
The basic classes and methods used for POS Tagging of the review dataset have been implemented using„ NLP Package‟[17] in R. Naïve Bayes and SVM classifications have been implemented using „e1071‟ package[18] in R. „ ROCR‟[19] has been found effective to identify and analyse Hit Rate and False Alarm Rate.

Experimental Result and Analysis:
The entire dataset has been given a score based on their adjectives as explained in the previous sections of the paper. Each review is given a score ranging between -2 to 6 based on the positive or negative polarity of the words. In the graph below (Fig 1), frequency of the words is taken on the x-axis and the score of
the review denoted by analysis$score is plotted on y-axis. The length of each bar of the histogram denotes the frequency of reviews of the particular score in the database.
The ROC Curve (Fig 3) plots hit rate or true positive rate on x-axis against false alarm rate or false positive rate. The hit rate shows the part of predictions that have been identified correctly. The false alarm rate refers the expectancy of false ratio. The area under the curve shows the accuracy of the predictions.

The given table (Table 3) shows a comparison of the experimental results of SVM and Naïve Bayes classifications using Precision and Recall functions done using modules of R. Precision the fraction of extracted instances which are correct whereas Recall is the fraction of correct instances that have been extracted. By comparing the accuracy values, it is clear that SVM has a higher accuracy and hence is a better classification method compared to Naïve Bayes.

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 2, ISSN No. 2455-2143, Pages 93-98
Published Online June 2022 in IJEAST (http://www.ijeast.com)

Table 3. Precision and Recall

Classification	Accuracy	Hit Precision	Hit Recall	Miss Precision	Miss Recall
SVM	0.8347	0.90	0.79411	0.7704	0.8867
Naïve Bayes	0.8014	0.8791	0.7698	0.7432	0.8645

V. FUTURE WORK:

In our future work, we further plan to enhance our techniques and refine them in order to extend the process of Sentiment Analysis to wider domains. We then plan to include neutral sentiment along with positive and negative sentiments in order to get better results using multiclass Naïve Bayes and multiclass SVM Classification. We also plan to include several other classification algorithms in order to obtain better results. Implied Sentiments which are not expressed in words are also a means to get user views on a product. In future, we plan to understand and classify these implied sentiments accurately.

VI. REFERENCES:

[1]. Min Wang and Hanxiao Shi " Research on Sentiment Analysis Technology and Polarity Computation of Sentiment words" ,Hangzhou, China ,pages 331-334, 2010S. Owen, et al., Mahout in Action: Manning Publications, 2011 (est.).

[2]. Basant Agarwal, Vijay Kumar Sharma and Namita Mittal " Sentiment Classification of Review Documents using Phrase Patterns", I n 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI) ,pages 1577 -1580

[3]. Rui Yao and Jianhua Chen " Predicting Movie Sales Revenue using Online Reviews" In 396 2013 IEEE International Conference on Granular

[4]. Computing (GrC), pages 396-401

[5]. Kai Gao, Shu Su and Jiu-shuo Wang " A Sentiment Analysis Hybrid Approach for Microblogging and E-Commerce Corpus" In 7th International Conference on Modelling, Identification and Control (ICMIC 2015)

[6]. Pooja Kherwa ,Arjit Sachdeva ,Dhruv Mahajan ,Nishtha Pande and Prashast Kumar "An approach towards comprehensive sentimental data analysis and opinion mining" In 2014 IEEE International Advance Computing Conference (IACC), pages 606 -612

[7]. Giovanni Acampora and Georgina Cosma, " A Hybrid Computational Intelligence Approach for Efficiently Evaluating Customer Sentiments in E- Commerce Reviews" Nottingham, United Kingdom

[8]. Siddharth Aravindan and Asif Ekbal " Feature Extraction and Opinion Mining in Online Product Reviews", Patna, India pages 94 -99

[9]. Yan Wan, Hongzhurui Nie, Tianguang Lan and Zhaohui Wang " Fine-grained Sentiment Analysis Of Online Reviews" In 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) pages 1406 -1411

[10]. Ming Sun and Jerome R. Bellegarda " IMPROVED POS TAGGING FOR TEXT –TO–SPEECH SYNTHESIS", ICASSP 2011 ,pages 5384-5387

[11]. Astha _Gupta, _R. _Rajput, _R. _Gupta _and _M. _Arora _" Hybrid model to improve time complexity of words search in POSTagging", Delhi, India, pages 1 -6

[12]. For Naïve Bayes Classification: http://joshwalters.com/2012/11/27/naive-bayes-classification-in-r.html

[13]. For SVM Classification: https://en.wikipedia.org/wiki/Support_vector_machine

[14]. Liangxiao Jiang, Harry Zhang, and Zhihua Cai, " A Novel Bayes Model: Hidden Naive Bayes",IEEE Transactions on Knowledge and Data Engineering, pages 1371-1371

[15]. GuoQiang " An Effective Algorithm for Improving the Performance of Naive Bayes for Text Classification". Second International Conference on

[16]. Computer Research and Development, 2010.

[17]. Jair Cervantes , Xiaoou Li and Wen Yu " SVM Classification for Large Data Sets by Considering Models of Classes Distribution", In Sixth Mexican International Conference on Artificial Intelligence, Special Session pages 51-60

[18]. Jin Huang, Jingjing, Lu Charles and X. Ling " Comparing Naive Bayes, Decision Trees, and SVM with AUC and Accuracy", in Third IEEE International Conference on Data Mining (ICDM‟03) Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews." Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD- 2004),

Aug 22-25, 2004, Seattle, Washington, USA

[19]. For POS Tagging: https://cran.r-project.org/web/packages/NLP/index.html
[20]. SVM and Naïve Bayes Package: https://cran.r-project.org/web/packages/e1071/index.html
[21]. ROC Plot implementation: https://rocr.bioinf.mpi-sb.mpg.de/