Sentiment Analysis using Naïve Bayes Classifiers and Logistic Regression

Anuja Bokhare[1], Vijayshri Khedkar[1], Vandana Raj[1], Niraj Bohra[2]

[1]Symbiosis Institute of Computer Studies and Research, Symbiosis International (Deemed University), [2]Software Engineer at ANZ ,Greater Sydney Area, Sydney, New South Wales, Australia
Email: anuja.bokhare@gmail.com

Abstract.

In the current digital world, people are using online shopping for purchasing goods extensively with first preference. Reviews and ratings on the sites will play a vital role in improving global communications among the customers and it has the potential to influence consumer buying patterns as well. Many E-commerce giants like Amazon, Flipkart are providing a platform that empowers users to share their real-time experiences and provide real-time insights about the performance of the product to future buyers. Sentimental Analysis and classification of the reviews as positive and negative will help us in understanding the voice of the consumers. So current study works on the reviews whose ratings are not provided by the customers. Machine learning models like Naïve Bayes, Multinomial NB, Bernoulli NB and Logistic Regression are used during the study. The result predicted the sentiments of those reviews.

Keywords. Machine Learning; Sentiment Analysis; Naïve Bayes classifiers; Feature extraction; Predictions.

1. INTRODUCTION

Current study focuses on Sentimental Analysis of the reviews that customers give for the products they purchase and have classified them as "Positive" or "Negative". Along with the reviews, customers also give ratings for the products and it is one of the filter criteria as well on the e-commerce websites to allow customers to view only those products which have ratings more than the chosen number. Thus, in today's world, it has become important to classify the products accordingly and display correct information to the user. Machine Learning provides different classification models which further helps in predicting the outcome. In this study Naïve Bayes Algorithm and Logistic Regression are used to classify the reviews as positive and negative. Naïve Bayes is a kind of classifier that use Bayes Theorem. It is used to predict membership probabilities for each class. In this work, the membership classes are, if a Positive Review or a Negative review so, the Naïve Bayes classifier is used to predict whether the given review belongs to the Positive or Negative class. Logistic Regression is a supervised machine learning algorithm that is used for classification problems. It is an algorithm used for predictive analysis and is uses the concept of probability. It works best on binary classification problems and since in current study, there are only two classes such as Positive or Negative reviews, so Logistic Regression is also used for comparison purpose.

2. PREVIOUS STUDY

Users today finds it difficult to process all of the data unless an automated summary is available. With fragmented messages as inputs, social media summarization creates a summary of content generated by user. Sentiment analysis and classification, in particular area aim to summarize online customer reviews. These reviews allow consumers to compare consumer feedback or review from different items by highlighting the benefits and drawbacks of their product features. Since web users give their opinions on a product, they have both positive and negative review about various aspects of the product. The ratio of positive and negative review about each product feature across multiple products needs to be compared to offer a quantitative overview and comparison between multiple products. Sentiment classification establishes the alignment of a review opinion irrespective of product features, while sentiment analysis excerpts individual product feature groups from review opinions [1]. Sentimental analysis is the procedure of determining the sentiment of reviews based on the positive and negative implications. Sentiment analysis can be done at three levels: document level, sentence level, and phrase-level [2]. There has been a lot of previous work completed in this area where terms and phrases have been categorized as positive or negative oppositeness [3]. In certain cases, this prior definition is useful, but when contextual oppositeness is involved, the sense obtained from positive or negative oppositeness may be completely different [4]. In addition, a refined method for establishing contextual oppositeness of phrases has been developed, which uses subjective detection to compress reviews while preserving the intended oppositeness [5] [6]. There are different approaches used for sentimental analysis. Also there are many machine learning algorithms that have been used to do sentimental analysis like Decision tree, Support Vector Machine and Rule based Approach. These techniques have its own set of advantages and the limitations as well [14]. Lexicons were regularly used for polarity and intensity classification [7]. Authors have presented SentiSense as an affective lexicon. Its job is to attach emotional categories to WordNet synsets. This lexicon could be useful for opinion mining and analyzing the sentiments. One of its major benefit is the accessibility of diverse set of algorithms and tools. Thus, the lexicon may be extended collaboratively, so that user's extensions may be used to enrich the core of the lexicon. Authors have [8] proposed a technique depend on improvement of bag-of-words method for testing sentiment of positive, negative and neutral and to score repeatedly using the words weight method rather than term frequency. [14] Few of them are Negation handling, language generalization, synonyms gathering, detection of fake and spam reviews and comparative sentences [13]. To do this, use of SVM, KNN models and identify the fake reviews and fake false reviews and do comparative study [15] is an option. A system has been built to take in the sentiment embedding's and does sentimental analysis on feedback made by the customer. To obtain overall score for the product, the polarity scores are added obtained from the feedbacks of a particular product. It is then presented in the form of a graph to the administrator depending upon which administrator can manage the inventory so as to improve the overall quality of the website [9]. Rating, Reviews and Emoticons are the parameters, to measure the quality, quantity or some combination of both. To find out fake reviews MAC based filtering approach was proposed. The method is verified on real time data from online website [10] [11]. CommTrust is proposed to define sellers profile. It shows reputation scores and trust score of seller. [12]. Sentimental analysis has been done using various techniques for example, Information gain, Chi-Square, Mutual information and TF-IDF in order to choose

features from feature set with high dimensionality. The classification of sentiments was performed using Support Vector Machine [16]. Authors have presented methods to normalize the tweets which have noise and classified them based on their polarity [17]. A novel method with a multi-domain active learning framework has been proposed. Term frequency is implemented for weightlifting features and LIBLINEAR SVM is used to generate a better classification model [18]. Authors have proposed a new technique which is based on clustering to oppress the problems of the prevailing methods. The clustering process was implemented multiple times to acquire the final result [19]. Sentimental analysis has been performed using various Machine Learning Techniques in the past such as Decision Tree Classifier, K-Nearest Neighbour Classifier, Support Vector Machine, Bayesian Network, Neural Network etc. Result shows 85% of accuracy was achieved by using supervised learning technique [20-22].

3. PROBLEM FORMULATION

Whenever user intends to buy something from e-commerce, they will tend to buy the products whose ratings are more than 4, and they read reviews of the products whose ratings are good most of the times. Users generally see the reviews as a sorting filter in many e-commerce sites but sadly, sites are being sorted only by the number of reviews irrespective of them being a positive or a negative one. This will impact the sales of the e-commerce site. So, in current study focus given on targeting the reviews for which ratings are not available and try to predict the emotions of those reviews using Naïve Bayes Classifier. This approaches helps in understanding and predicting the sentiments of the users. This approach helps to enhance the products that are having more positive reviews so that the user will feel more secure about the product that is being purchased.

4. PROPOSED METHODOLOGY

Data used for the study is gathered from Kaggle website i.e. Customer reviews of Amazon products. During data cleaning the punctuations and stewards have been removed from the reviews which are having ratings and the reviews with rating Nan. Dataset consists of 21 columns such as Id of the product, Name of the product, brand, categories, keys, manufacturer, and date of the review, review Id, and review rating, review text, review title, username and so on. The data is split into training and test sets. Features have been extracted from the data and then accuracy of the model have been observed. Different classifiers are used during the experiment i.e. NLTK Naïve Bayes, Multinomial NB, Bernoulli NB and Logistic Regression. This helped in predicting and segregating the sentiments into positive and negative. Figure 1 shows the proposed architecture model of the sentiment analysis using different classifiers.

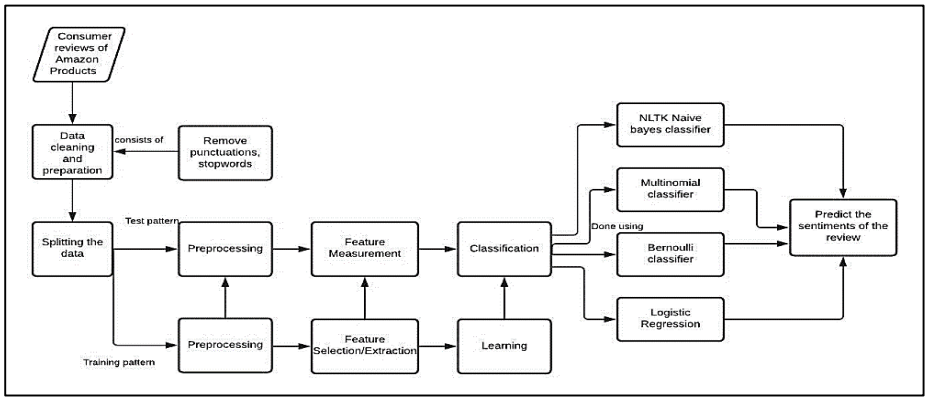

Figure 1. Proposed Sentiment analysis model.

5. DATASET AND EXPERIMENT DISCUSSION

Dataset of Consumer Reviews of Amazon Products is considered during the experiment. The main focus lied only on the columns named "reviews. rating, reviews. Text, reviews. title and reviews. username" from the dataset.

1. Initially Null values in these columns are observed. Then reviews in which there was no rating given are then filtered out. The reviews which possessed some rating also filtered and stored in some variable.
2. Reviews with the rating greater than 4 are classified into positive else classify as negative. It has been observed that there were 32316 reviews as positive and 2311 as negative. Figure 2 shows the bar chart of the same.
3. Next data cleaning was performed in order to feed the data to proposed sentiment analysis model and the sentiment of each review was shown against the reviews given by customers.
4. Naïve Bayes classifier was applied on the training data and the accuracy of the model was observed along with 5 most informative features. However, the accuracy was just 58.9%.
5. Count vector and TF-IDF vector was then constructed and for training, testing and check data.
6. Then predicted the sentiments for check vector which was not having any rating initially using Multinomial NB and found the accuracy to be 93.29%. Bernoulli NB method also checked for accuracy which was approximately 92.04%. The model is also trained using Logistic Regression and predicted the sentiments for check data and accuracy came out to be 93.70%.

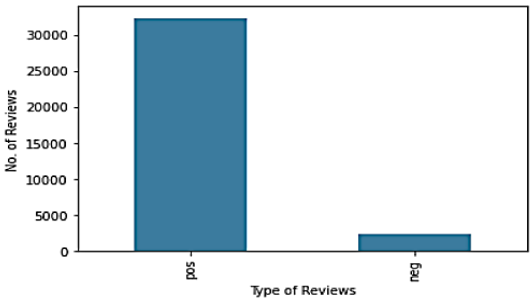

Figure 2. Positive and negative reviews

6. RESULT ANALYSIS

This section discusses about the results obtained from the experiment. It presents the analysis and visualization of the reviews for which ratings are present

6.1. Word Cloud Visualization of the sentiments

Python is used to show the feelings retrieved as features in a WordCloud. Figure 3 describes all of the sentiment words that were retrieved as opinion characteristics.

Fig 3. Word Cloud of all sentiment

Fig4. Word Cloud of positive sentiments

Fig 5. Word Cloud of negative sentiments

Figure 4 and 5 shows the words used for predicting positive and negative sentiments respectively

6.2. Performance of classifier using AUC-ROC curve

Measuring performance is a necessary activity. So authors have used AUC - ROC Curve which helps to validate a classification experiment. The Area Under the Curve (AUC) and Receiver Operating Characteristics (ROC) curve is used to confirm the outcome of a multi-class classification problem The ROC algorithm determines a classifier's threshold. The curve aims to maximize genuine positives while reducing false positives. The area under the curve calculation was utilized (AUC). Fig. 6 shows the performance for different classifier i.e Multinomial, Bernoulli and Logistic Regression. When compared to other methods, Logistic Regression has a high score of 0.86.

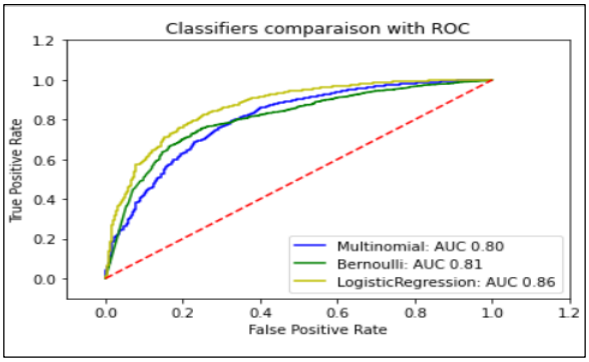

Figure 6. Performance of Classifiers

The ROC algorithm determines a classifier's threshold. The curve aims to maximize genuine positives while reducing false positives. The area under the curve calculation was utilized (AUC). Figure 6 shows the performance for different classifier i.e Multinomial, Bernoulli and Logistic Regression. When compared to other methods, Logistic Regression has a high score of 0.86.

6.3. Performance Metrics of different classifiers

Table 1 shows Performance metrics for the model i.e. precision, recall and accuracy for the Multinomial, Bernoulli and Logistic Regression classifiers.

TABLE I. PERFORMANCE METRICS

Performance Parameter / Types of Machine Learning Models	Precision	Recall	Accuracy
Multinomial	0.87	0.93	0.93
Bernoulli	0.90	0.92	0.92
Logistic Regression	0.93	0.94	**0.94**

Observation for Sentiment model proposed during the study is; Multinomial Precision would be 0.87, or 87 %, Bernoulli Precision would be 0.90, or 90 % and Logistic Regression Precision would be 0.93, or 93%. As a result, when it is predicted that the reviews will be positive, there will be an 87%, 90% or 93% possibility that all of the reviews will be positive. Recall for Multinomial would be 0.93, or 93 %, Recall for Bernoulli would be 0.92, or 92 %, and Recall for Logistic Regression would be 0.94, or 94 % for proposed model. As a result, there would be a 93%, 92%, or 94% chance of getting real positive review from the positive reviews proposed model have identified.

7. CONCLUSION

Sentiment mining is incredibly significant in business since it allows companies to better understand its consumers' opinions and enhance their offerings. Customers also rely on the

opinions of others who have already purchased the item. Reviews or feedback become a deciding factor in whether or not to purchase or sell a product. Current study forecasted the sentiment of reviews with Nan ratings by feeding the reviews with proper ratings into proposed model and that helped to identify the customer's opinion. Current study helps to forecasts the opinion based on unrated reviews. In comparison to multinomial and Bernoulli regression, the logistic regression prediction has received a positive review. The accuracy of the forecast from Logistic regression is higher i.e. 94% than multinomial and Bernoulli.

8. FUTURE SCOPE

Different dataset would be tested on proposed model in future with different feature selection methods. Also the same model can be used for sentiment classification for detecting fake reviews. User can see more positive reviews which will help in building better customer relationships.

9. REFERENCES

[1] Yang CC, Tang X, Wong YC, Wei CP. 'Understanding online consumer review opinions with sentiment analysis using machine learning', Pacific Asia Journal of the Association for Information Systems. 2010;2(3):7, pp.73-89

[2] S. Erevelles, N. Fukawa, and L. Swayne, 'Big data consumer analytics and the transformation of marketing', Journal of Business Research, 2016, 69(2),

[3] V. Hatzivassiloglou and K. R. McKeown, 'Predicting the se- mantic orientation of adjectives', in Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 1997, pp. 174–181.

[4] T. Wilson, J. Wiebe, and P. Hoffmann, 'Recognizing contextual polarity in phrase-level sentiment analysis', in Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics, 2005, pp. 347–354.

[5] B. Pang and L. Lee, 'A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts', in Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004, p. 271

[6] Singla Z, Randhawa S, Jain S. 'Sentiment analysis of customer product reviews using machine learning', In2017 international conference on intelligent computing and control (I2C2) 2017 Jun 23 (pp. 1-5). IEEE.

[7] De Albornoz JC, Plaza L, Gervás P. 'SentiSense: An easily scalable concept-based affective lexicon for sentiment analysis', InLREC 2012 May 23 ,Vol. 12,

[8] Cao Y, Zhang P, Xiong A. 'Sentiment analysis based on expanded aspect and polarity-ambiguous word lexicon', International Journal of Advanced Computer Science and Applications. 2015;6(2).

[9] Murugavalli S, Bagirathan U, Saiprassanth R, Arvindkumar S. 'Feedback analysis using sentiment analysis for e-commerce. Feedback', 2017 Mar 30;2(3),pp. 84-90.

[10] Firake VR, Patil YS. 'Survey on CommTrust: multi-dimensional trust using mining e-commerce feedback comments', the proceedings of the International Journal of Innovative Research in Computer and Communication Engineering IJIRCCE. 2015 Mar;10, pp.1640-1643

[11] Osimo D, Mureddu F. 'Research challenge on opinion mining and sentiment analysis', Universite de Paris-Sud, Laboratoire LIMSI-CNRS, Bâtiment. 2012;508.

[12] Wang J, Jing X, Yan Z, Fu Y, Pedrycz W, Yang LT. A Survey on Trust Evaluation Based on Machine Learning. ACM Computing Surveys (CSUR). 2020 Sep 28;53(5), pp. 1-36.

[13] Tang D, Wei F, Yang N, Zhou M, Liu T, Qin B. 'Learning sentiment-specific word embedding for twitter sentiment classification', InProceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 2014 Jun ,pp. 1555-1565.

[14] Rokade, Prakash & D, Aruna 'Business intelligence analytics using sentiment analysis-a survey', International Journal of Electrical and Computer Engineering (IJECE). 9. 613. 10.11591/ijece.v9i1.2019, pp. 613-620.

[15] Elmurngi E, Gherbi A. 'Detecting fake reviews through sentiment analysis using machine learning techniques', IARIA/data analytics. 2017 Nov: pp. 65-72.

[16] Shahana PH, Omman B. 'Evaluation of features on sentimental analysis', Procedia Computer Science. 2015 Jan 1;46, pp. 1585-92.

[17] Celikyilmaz A, Hakkani-Tür D, Feng J. 'Probabilistic model-based sentiment analysis of twitter messages', In2010 IEEE Spoken Language Technology Workshop 2010 Dec 12 (pp. 79-84). IEEE

[18] Li L, Jin X, Pan SJ, Sun JT. 'Multi-domain active learning for text classification', InProceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining 2012 Aug 12 , pp. 1086-1094.

[19] Li G, Liu F. 'A clustering-based approach on sentiment analysis', In2010 IEEE international conference on intelligent systems and knowledge engineering 2010 Nov 15 ,pp. 331-337. IEEE.

[20] Bhavitha BK, Rodrigues AP, Chiplunkar NN. 'Comparative study of machine learning techniques in sentimental analysis', In2017 International conference on inventive communication and computational technologies (ICICCT) 2017 Mar 10 ,pp. 216-221. IEEE.

[21] Ingale, V. R., & Phursule, R. N. 'Product Feature-based Ratings for OpinionSummarization of E-Commerce Feedback Comments', International Journal of Computer Applications,2016, 135(8),975-8887.

[22] Ingale, V. R., & Phursule, R. N. 'Sentiment Analysis by Visual Inspection of User Data from Social Sites-A Review on Opinion Mining', International Journal of Science and Research, 2014,3(12), 2188-2191.

Biographies

Dr. Anuja Bokhare is working as Assistant Professor in the department of Computer Science at Symbiosis Institute of Computer Studies and Research, Pune, Maharashtra India. She received M.Phil. (Computer Science) at Y.C.M.O.U, Nasik, India and completed PhD from Symbiosis International (Deemed University) in faculty of Computer Studies. She has 20 years of experience in the field of academics. Her research interest includes software engineering, applications of Artificial Intelligence, machine Learning and soft computing. She had published 22 research papers in international journal and conferences along with one book in her account.

Vijayshri Khedkar is an Assistant Professor working at Symbiosis Institute of Technology and skilled in NLP, Applied Machine Learning, Data Analytics, Information Retrieval & Deep Learning. A life-long learner with a strong educational background holding two Master's Degrees (M.B.A. & M.E.) and pursuing Ph.D. in Computer Engineering (NLP) from Symbiosis International University, India. She is a life member of IAENG and Senior IEEE member. She has published around 30 papers at various reputed conferences and journals. She is active reviewer for IJECE journal indexed in Scopus and reviewed papers for various conferences and journals.

Vandana Raj N. currently working as BA in Morgan Stanley. She has completed her masters in IT from Symbiosis International University, with work experience of 7 months as Data Analyst at Colgate Palmolive and 2 years as SDET with an e-commerce website known as blibli.com.

Niraj Bohra is a passionate and forward looking software engineer with over 5+ years of experience in developing and delivering innovative enterprise software solutions to improve the productivity of the business. Experienced in all aspects of the software development life cycle from concept to delivery and always eager to build a network and drive change using technology.