RESEARCH ARTICLE

SAMRIDDHI Volume 13, Special Issue 2, 2021 Print ISSN : 2229-7111 Online ISSN : 2454-5767

An Efficient Sentiment Analysis Based on Product Reviews

Vijay Mane[1*], Sanmit Patil[2], Rohan Awale[3], Vipul Pisal[4]

[1,*] Deptt. of Electronics Engineering, Vishwakarma Institute of Technology, Pune, India; e-mail : vijay.mane@vit.edu

[2-4.] Deptt. of Electronics Engineering, Vishwakarma Institute of Technology, Pune, India;

ABSTRACT

Amazon is the most popular online shopping market for most people in the world today. Anything from daily necessities to luxurious items can be bought from here. And especially in recent times where people have to avoid going out to crowded places, platforms like Amazon have emerged as the go-to solution. So, when people want to buy products from these platforms it is important for them to have a look at the reviews before being assured about it. But every product has thousands of reviews for it and it's not easy to analyze them quickly. This paper presents an implementation of a Amazon review sentiment analysis with the web application. Various algorithms are implemented for the experimentation purpose. The combination of logistic regression with CountVectorizer performed well in the term of accuracy. Using the proposed methodology, the user can search for a product on this web-based App and analysis of product reviews, price ranges, ratings and much more will be displayed to the user. The accuracy of the different algorithms is reported in this paper.

Keywords: Amazon reviews, analysis, machine learning, sentiment.

SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology, (2021); DOI : 10.18090/samriddhi.v13spli02.17

INTRODUCTION

In recent times E-commerce has grown massively and almost everyone is familiar and comfortable with it. Whether a person is young and looking for study material regarding his or her education or someone old is looking for a comfortable chair, E-commerce platforms help everyone to get their products without much difficulty. These platforms deliver every product right at our door-step without the customer having to move anywhere. Although this is a massive benefit to the customer, it also results in the customer not being guaranteed about the exact qualities of the product as they are not actually able to examine it before buying. Because of this, reviews and ratings have taken up a very significant role in the entire E-commerce platform.

When a customer buys a product, the E-commerce platform allows them to give a review and rating to give an idea about the product, its packaging, the delivery process and so on. These reviews can be helpful for the sellers as they can improve on their service based on it and it can also be helpful for other users who might be thinking about buying the same product. But, as E-commerce has risen tremendously in recent years and will continue to do so in the coming years, it has been observed that vast amounts of reviews are found for every product available on E-commerce platforms like Amazon. Getting the right information from so many data is not easily possible for every person.

It is necessary to design a system, where the user can search for a product, and they will be provided with short yet accurate analysis regarding those products and then the user can make a decision based on them. So, when a user searches for a product, different graphical visualizations will be shown like the different brands available, the

Corresponding Author : Vijay Mane, Deptt. of Electronics Engineering, Vishwakarma Institute of Technology, Pune, India; e-mail : vijay.mane@vit.edu

How to cite this article : Mane, V., Patil, S., Awale, R., Pisal, V. (2021). An Efficient Sentiment Analysis Based on Product Reviews.

SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology, Volume 13, Special Issue (2), 203-208.

Source of support : Nil

Conflict of interest : None

ratings for those brands, the price ranges and so on. Along with this, all the reviews of products searched will be trained using machine learning and natural language processing. Based on it, every review will be categorized as either positive or negative and a sentiment analysis will be done on it. Finally with those trained models, some of the most common words appearing in all negative and positive reviews will also be shown to give an idea as to what exactly is wrong or good about the products.

LITERATURE REVIEW

The sentiment analysis based on product reviews collected from Amazon are performed in [1]. The experiments for both classifications of reviews and extraction of narratives from the text are performed with promising outcomes. Using Logistic Regression an accuracy of 0.88 on bag of words and 0.89 on TFIDF.

The modern world is becoming more & more digitalized. In this digitalized world the importance of E-commerce is on the rise by making products easily accessible to customers so that the customer doesn't have to go out of their house. A customer first goes through several products reviews before making the decision of buying that product. In today's where machine learning is assuming great importance, the models which would polarize reviews into positive or negative were developed [2]. So supervised learning methods were used on large scale amazon datasets to polarize it and get its outcomes. The best accuracy of 93.2% was obtained using Linear Support Vector Machine algorithm in [2].

The Sentiment Analysis has been implemented on opinion reviews to detect unfair negative, neutral, and positive reviews [3]. They experimented using four supervised machine learning algorithms Naïve Bayes, Decision Tree, Logistic Regression and Support Vector Machine for sentiment classification based on three datasets of reviews regarding Clothing, Shoes and Jewelry, Baby products and Pet Supplies. For evaluation of the implementation parameters like accuracy, precision and recall have been considered. The Logistic Regression (LR) algorithm gave the best accuracy compared to the others.

The aspect level sentiment detection, which focuses on the features of the item has been implemented on Amazon customer reviews [4]. For extracting significant information from the reviews, this approach includes performing pre-processing operations like stemming, tokenization, casing, stop-word removal on the dataset which finally gives a rank for its classification in negativity or positivity. Support Vector Machine classifier has been used for the classification which gave an accuracy of about 97%.

Product reviews found on Amazon not only give an idea about the product but also the service related to it. If users can get clear bifurcation about reviews for product and after sales, it will be easier for them to make decisions regarding buying the product or not. A rule-based extraction of product feature sentiment implemented to perform the classification of customer reviews by finding sentiment of the reviews [5].

Opinion Mining & Sentiment Analysis are some of the most popular fields to analyze data and find insights from it using various social media handles. The classification of review into positive and negative were implemented using machine learning [6]. The Naive Bayes gave the best accuracy of around 98% and Support Vector Machine also performed well by giving an accuracy around 93%.

The sentiment analysis for Amazon's product using Naive Bayes and decision list performed by [7]. The rating given by the user for a product is utilized for training data to execute supervised machine learning. The dataset used for this study consists of 50,000 product reviews from 15 products. The Naïve Bayes classifier gave an accuracy of 84.2 %.

The classification of reviews performed using removal of punctuations, whitespaces and special characters and also performing stemming [8]. Then term frequency-inverse document frequency (TF–IDF) was used to represent the preprocessed data. Finally, algorithms like K-Nearest Neighbor, Decision Tree, Support Vector Machine, Random Forest and Naive Bayes to classify the reviews. The comparison between the accuracies obtained by the different classifiers in which the best was seen in Random Forest classifier with an accuracy of 94.72%, time required for each classifier and the sentiment scores of the various books.

The sentiment analysis of Amazon reviews was done using their corresponding ratings. The vectors were created from product reviews using paragraph vectors, which was then used to train a recurrent neural network with gated recurrent unit. This model incorporated both semantic relationship of review text and product information. Then a web app was also developed which would predict the rating for the review which would be submitted and

using that we can provide feedback to the user if there is a mismatch between the review and the rating. The accuracies of around 81% were obtained in this approach [9].

The sentiment analysis for the investors based on stock messages were proposed using machine learning and voting methodology [10]. This system elaborated the use of their methodology to monitor the influence of various broadcasts and governing alterations on shareholders specifically.

METHODOLOGY

The proposed system is implemented as shown in Figure 1. The sentiment analysis implementation requires a good dataset to create a model to predict the sentiment of the review as positive or negative. The dataset must contain reviews, ratings, brands and product names regarding mobile phones and office tools. Every review which had a rating of 3 or more was given positive sentiment and the rest a negative sentiment. The preprocessing of such data is needed to make the data ready to be applied to different machine learning techniques [11,12]. All the reviews were converted to lower case and all other characters other than the 26 alphabets were removed. Then the stop words which were not going to be useful in determining the sentiment of the review were also removed.

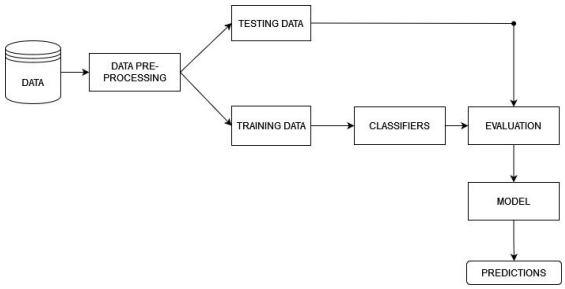

Figure 1: Different steps to be performed

Once the data pre-processing is done, the features are extracted. The Countvectorizer and TF-IDF Vectorizer is utilized in this implementation. The Countvectorizer is used to convert a collection of text to a vector which contains the frequency of every word in the text. And, TF-IDF Vectorizer on the other hand creates a vector which contains calculated values. These values are calculated by multiplying two metrics: how many times a word appears in a review, and the inverse frequency of the word across all reviews. So, using these two Vectorizers we were able to convert our reviews into vectors which could be given to machine learning algorithms as shown in figure 2. Along with this word2vec is also a technique which was used to convert out text reviews into vectors. But the vectors we get from this are then provided to a neural network instead of machine learning algorithms.

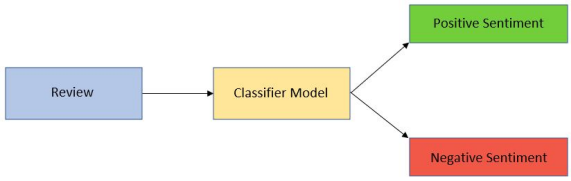

Figure 2: Expected Outcome

The experimentation for the analysis has been implemented using different algorithms like Logistic Regression, KNN, Support Vector Machine and Random Forest. These algorithms are explained in short below.

Logistic Regression: Logistic regression is a classification algorithm which is used when our output is binary in nature like positive sentiment or negative sentiment. This algorithm predicts the probability of our output being positive or negative and the one with higher probability is selected as the output.

K-nearest neighbors (KNN): It implements based on distance, proximity or nearness. The distance for every review will be calculated from the positive sentiment group and the negative sentiment group and the group with the least distance from the review will be assigned to the review.

Support Vector Machine (SVM): Here, based on number of features the data is projected as appoint in a multidimensional space with the value of each feature being the value of a particular coordinate. Then, the hyper-plane is found which helps to classify into one of the two classes very well. So, the positive and negative sentiments will be distinguished by a hyperplane created by the algorithm [13].

Random Forest: In this algorithm a number of decision trees are constructed at the time of training and outputs of the models predicted by the individual trees is found. Every tree has its final nodes as positive sentiment and negative sentiment. Based on outputs of all individual trees the random forest algorithm will give the most common output as the final result [14].

So, the vectors we got earlier were given to these different algorithms to create a model. Based on these models, predictions were made for every and then their accuracy was checked. All these steps were done for the 2 different datasets and then finally the Countvectorizer with Logistic Regression was chosen as the best combination.

The web app is also created, where the user could search for a product and the analysis would be done for those products. For this we would need to scrape the Amazon website for its reviews. This was done using Beautiful Soup. After scraping we were able to create a data frame which would consist of information like brand and product name, average rating for that product, reviews for that product and finally the sentiment predicted for that product. Using this data frame, we were then able to display the analysis in the form of visual graphs like bar charts, histograms and word clouds.

Results and Discussion

As mentioned earlier, we used different machine learning algorithms with different feature extraction algorithms. These algorithms were tested on a mobile phone dataset and an office tools dataset. The mobile phones dataset consists of information like the brand name, model name, price, features and the users review about it. Similarly, the office tools dataset consists of information like the products id number, users review for it and the ratings given by the user. The accuracy of these algorithms is shown below Figure 3.

Model	Accuracy
Count Vectorizer-Logistic Regression	0.952167
LSTM	0.951853
TFIDF Logisitc Regression	0.948399
TFIDF SVM	0.947247
W2V SVM	0.941909
W2V Random Forest	0.941490
W2V KNN	0.939397
W2V Logistic Regression	0.938978
Count Vectorizer-SVM	0.938560
TFIDF Random Forest	0.938141
Count Vectorizer-Random Forest	0.938036
Count Vectorizer-KNN	0.936885
TFIDF KNN	0.926418

Figure 3: Accuracy of different algorithms on office tools dataset

For the office tools dataset, we can see in the above figure that six algorithms have given an accuracy of more than 94%. We can say that Logistic regression works really well with both CountVectorizer and TFID Vectorizer. Then, Support Vector Machine (SVM) also gave decent results. Although the Long-Short Term Memory (LSTM) method gave a decent accuracy as well, the time it required for generating the output was really high which made it inconvenient.

For the mobile phones dataset, the results were obtained as shown in figure 4. Here also, Logistic Regression gave a good accuracy with both the feature extraction algorithms.

After considering the accuracy obtained from all algorithms and the time required for the processing, we found that Logistic Regression with CountVectorizer gave us the most optimum results. So, we decided to use it further while building the web app.

Model	Accuracy
TFIDF SVM	0.946908
LSTM	0.935092
Count Vectorizer-Logistic Regression	0.933959
TFIDF Logisitc Regression	0.931531
TFIDF Random Forest	0.925219
Count Vectorizer-Random Forest	0.923438
W2V Random Forest	0.920524
Count Vectorizer-SVM	0.914212
W2V SVM	0.904014
W2V Logistic Regression	0.890741
W2V KNN	0.888152
Count Vectorizer-KNN	0.839430
TFIDF KNN	0.767077

Figure 4: Accuracy of different algorithms on mobile phones dataset.

Below in figure 5 the home page is displayed. Through this page you can go to two sections. One which will search for products generally without specifying the model's name. The other section searches for a specific model using its Amazon ID which can be found from the Amazon website.

Figure 5: Home page of the web application

Then in figure 6 we have the search page where the user will search for the product.

Figure 6: Search page of the web application

Upon submitting the search query various graphs depicting the analysis of the products will be displayed. Below in Figure 7, the sentiment analysis of different hard disks is shown.

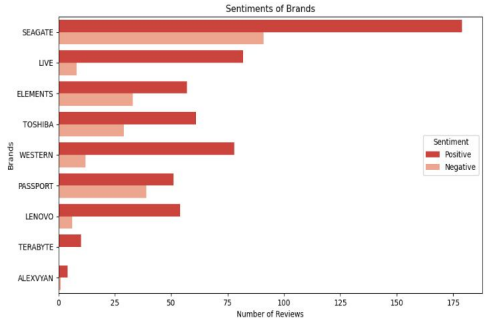

Figure 7: Sentiment Analysis of Reviews

Then the web app also displays some word clouds related to the reviews. Just having a glance at these word clouds can give you an idea about what can be good or bad about your product. Below are two examples where most common words appearing in positive and negative reviews of a JBL earphone are shown in figure 8 and 9.

Figure 8: Common words appearing in positive reviews

Figure 9: Common words appearing in negative reviews

Ideally a lot more data than what we have used can be used to train the models. More the data we have, the more accurate the predictions we will be able to make. But the limitation with huge amounts of data is its processing. Processing this huge amount of data requires a lot of time. Also, the RAM of the device should be able to handle all the processing being done which was not always possible in our case.

The online or cloud-based services to process the data can be utilize in future. This would allow us to process more data and, in less time, as well. Another aspect where improvements could be done is regarding the number of products whose reviews are being collected. At present the web would only consider the products which appear on the first page after searching for it. Including products which are present on the further pages would be beneficial for the user.

CONCLUSION

The presented approach successfully implemented a Amazon review analysis web app. With the help of pre-collected data, we were able to train models to predict the sentiment of the reviews. Several algorithms were used to implement this. The combination of logistic regression with CountVectorizer was seen as the best one. Then we used this combination of algorithms in the web app. So, the user can search for a product on this web app and analysis of product reviews, price ranges, ratings and much more will be displayed to the user.

REFERENCES

[1] Jain, Vineet, and Mayur Kambli. "Amazon Product Reviews: Sentiment Analysis.

[2] Haque, Tanjim Ul, Nudrat Nawal Saber, and Faisal Muhammad Shah. "Sentiment analysis on large scale Amazon product reviews." 2018 IEEE international conference on innovative research and development (ICIRD). IEEE, 2018.

[3] Elmurngi, Elshrif Ibrahim, and Abdelouahed Gherbi. "Unfair reviews detection on amazon reviews using sentiment analysis with supervised learning techniques." J. Comput. Sci. 14.5 (2018): 714-726.

[4] Nandal, Neha, Rohit Tanwar, and Jyoti Pruthi. "Machine learning based aspect level sentiment analysis for Amazon products." Spatial Information Research 28.5 (2020): 601-607.

[5] Bhatt, A., Patel, A., Chheda, H., & Gawande, K. (2015). Amazon review classification and sentiment analysis. International Journal of Computer Science and Information Technologies, 6(6), 5107-5110.

[6] Jagdale, Rajkumar S., Vishal S. Shirsat, and Sachin N. Deshmukh. "Sentiment analysis on product reviews using machine learning techniques." Cognitive Informatics and Soft Computing. Springer, Singapore, 2019. 639-647.

[7] Rain, Callen. "Sentiment analysis in amazon reviews using probabilistic machine learning." Swarthmore College (2013).

[8] Srujan, K. S., Nikhil, S. S., Rao, H. R., Karthik, K., Harish, B. S., & Kumar, H. K. (2018). Classification of amazon book reviews based on sentiment analysis. In Information Systems Design and Intelligent Applications (pp. 401-411). Springer, Singapore.

[9] Shrestha, Nishit, and Fatma Nasoz. "Deep learning sentiment analysis of amazon. com reviews and ratings." arXiv preprint arXiv:1904.04096 (2019)., International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.8, No.1, February 2019.

[10] Das, Sanjiv R., and Mike Y. Chen. "Yahoo! for Amazon: Sentiment extraction from small talk on the web." Management science 53.9 (2007): 1375-1388.

[11] Almjawel, A., Bayoumi, S., Alshehri, D., Alzahrani, S., & Alotaibi, M. (2019, May). Sentiment analysis and visualization of amazon books' reviews. In 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS) (pp. 1-6). IEEE.

[12] Lee, Dong-yub, Jae-Choon Jo, and Heui-Seok Lim. "User sentiment analysis on Amazon fashion product review using word embedding." Journal of the Korea Convergence Society 8.4 (2017): 1-8.

[13] Dey, Sanjay, et al. "A comparative study of support vector machine and Naive Bayes classifier for sentiment analysis on Amazon product reviews." 2020 International Conference on Contemporary Computing and Applications (IC3A). IEEE, 2020.

[14] Rathor, Abhilasha Singh, Amit Agarwal, and Preeti Dimri. "Comparative study of machine learning approaches for Amazon reviews." Procedia computer science 132 (2018): 1552-1561.