Sentiment Analysis of The Reviews Of Amazon Products On The Basis of Multinomial Naïve Bayes Approach And Random Forest Classifier

Aarti Majumdar
[a]Department Of Computer Science and Engineering
BRCM-College Of Engineering & Technology, Bahal
[b]Maharishi Dayanand University, Rohtak, Haryana, India
Email Id: aarti8j@gmail.com

ABSTRACT

For the growth of a particular organization and to withstand other competitions, decisions with more pros and fewer cons have to be made. This approach has helped prominent companies soar. Online Transactions have become a huge part of our life in recent days due to their convenience, of the products and fast delivery of the products. Reviews are crucial for understanding a particular product's position in the company and they can predict its sales flow. Opinion Mining, which is also known as "Sentiment Analysis" is a Machine Learning Method that offers a comparison between various brands based on the collected optimistic and negating ratings provided by the consumers. To determine the sentiment analysis of the feedbacks left by customers who had already sampled and bought the Amazon Products, we made this framework by using the Multinomial Naïve Bayes Classifier and the Random Forest Classifier in order to produce the best results with the given dataset. The accuracy results given by Multinomial Bayes and Random Forest Classifier are 90% and 99% respectively.

Keywords- Amazon Product, Reviews, Sentiment Analysis, Stopwords, One Hot Encoding, Multinomial Naïve Bayes ,Random Forest Classifier

1. INTRODUCTION

For the growth of a particular organization and to withstand other competitions, decisions with more pros and fewer cons have to be made. Such decisions are put forth when we have evaluated the opinions of those concerned. This approach has helped prominent companies soar. Amazon is an online-based American company that sells its products and services digitally, providing users a convenient method of shopping for items in this ever-growing electronic industry. Customers give reviews after using their services according to the quality of the product, its condition, and speed of the delivery to express their experiences and preferences which can be further studied to examine the overall customer satisfaction. Reviews are crucial for understanding a particular product's position in the company and they can predict its sales flow. Opinion Mining, which is also known as "Sentiment Analysis" is a Machine Learning Method that offers a comparison between various brands based on the collected optimistic and negating ratings provided by the consumers. In this study, we have presented our findings by using efficient and accurate results-inducing algorithms, namely the Multinomial Naive Bayes approach and the Random Forest Classifier method.

2. RELATED WORKS

The research carried out by Mohan Kamal Hassan, Sana Prasanth, and R Sasikala[1] reported its findings on a sentimental analysis of

Amazon Laptop product reviews using the naïve Bayes algorithm. From the beginning, the review databases are collected from the Amazon Mongodb database, this process is known as data collection. Now the dataset is processed to remove meaningless and unwanted stopwords and then transformed into words that can help to seek out the review's meaning. Finally, extraction of features is done to decide the way of understanding the review. The naïve Bayes classifier approach is to help determine the precise tag word indicating positivity and negativity of the comment where the defined tag word can also tag the required words and these words are needed to be calculated so that score for every review can be obtained. After successively performing the naïve Bayes classification, the duality(both positive and negative) is measured using a decision tree. To form semantic relations between the words, the WordNet dictionary was implemented to make out the foremost sensible comment. They have discussed two classification techniques namely Machine Learning Approach and Bayesian Networks(BN) method. The former has an accuracy of 71.7% approx. for 3 categories and approx. 46.9% for 5 categories and is limited to the English Language. The latter has approx. 73% accuracy. Although BN and Naïve Bayes have some differences, both have almost the same level of performance.

S. PAKNEJAD[2] provided the comparison between two Machine Learning Approaches specifically, the SVM(Support vector machines) approach, which is under supervised learning, and the Naïve Bayes Method. In the SVM algorithm, labeled training data that is situated on a plane is divided into different groups by another hyperplane. While the naïve Bayes method has independent features and feature extraction is carried out before its execution. Feature Extraction, also known as the Bag of Words model assigns each word a unique number. These techniques are offered by Python programming language which specializes in supervised learning methods. Firstly, data is collected from the SNAP dataset, and preparation of the targeted data is done to exclude unnecessary features with the exception of items such as the summary and text of the review itself. Going ahead, Bag of words is applied before training the Naïve Bayes and SVM classifiers, and then finally test data is adapted to the trained classifiers to calculate the level of performance. Two experiments were conducted in this research, for the first experiment, a total of 15000 training data and 48500 testing data were collected. Here, any scale of 3 stars(which denoted mixed reviews) was excluded to avoid complications. Next, in the second experiment, 10 products were selected and 300 data were collected from each product. Here, 3-star reviews were also considered negative. The results from the first experiment came out as naive Bayes having 92.72% accuracy while SVM has 93.20% accuracy on summaries. The second result having more reviews has an average of 89.98% and 80 to 90% accuracy in the case of Naïve Bayes and SVM respectively. This shows that Naive Bayes has more accuracy than SVM in the case of a large sampling review of data.

Nguyen, Heidi; Veluchamy, Aravind; Diop, Mamadou; and Iqbal, Rashed(2018)[3] conducted a study of Sentiment Analysis of Amazon product reviews using term frequency-inverse document frequency (TF-IDF)vectorization along with various Machine Learning and Lexicon based methods. Data is collected through Amazon Standard Identification Number which has links for each review.1000 product reviews are sampled and duplicate reviews are omitted. They then perform various Text Preprocessing operations such as sentence extraction, unescaping HTML escape sequences, removing special characters, lemmatizing text, erasing stopwords, performing tokenization, etc. After completing the process, we move onto TF-IDF Vectorization where we extract useful elements from a summary. Term Frequency(TF) is the frequency of occurring of a bunch of words in a file, also known as Bag of Words model. In this, each file is indicated by 0s and 1s where a nonexistent word is denoted by 0 and vice-versa. IDF(Inverse Document Frequency) is defined as the weighing of the importance of a particular word. They have carried out experiments by making use of three supervised approaches specifically, LR, SVM, and Gradient Boosting classifiers along with three lexicons in NLP which are VADER, Pattern, and SentiWordNet.
The results came out for SVM, Gradient Boosting, and LR methods having an accuracy of 89%,87%, and 90%; precision of 90%,98%, and 97%; FI score of 94%,92%, and 94% respectively. Pattern, VADER and SentiWordNet gained an accuracy of 69%,83% and 80%; precision of 88%,90% and 90%; FI score of 79%, 89% and 88% respectively. This shows that Machine Learning methods surpass Lexicon-based models.

SANA NABIL, JABER ELBOUHDIDI, and MOHAMED YASSIN CHKOURI[4] have set out to discover the best results among the results produced by different Machine Learning algorithms, namely, the naïve Bayes, Support Vector Machine, the Decision Tree, and the

Logistic regression algorithms in this paper. Their goal lies in expressing the sentiments of users derived from Amazon product reviews. Datasets first undergo transformation into fixed-size features by HashingTF, Tokenized and StopWords removed in Text Preprocessing before going through with Feature extraction. Next onwards, Machine Learning Algorithms are applied and implemented through the usage of Spark and Scala languages Thus resulting in results representing the achievement of the accuracy of 100%, 95%, 97%, and 75% for SVM, Naïve Bayes, LR, and Decision Tree classifiers respectively. This highlighted the fact that SVM produced the best results among others.

3. METHODOLOGY

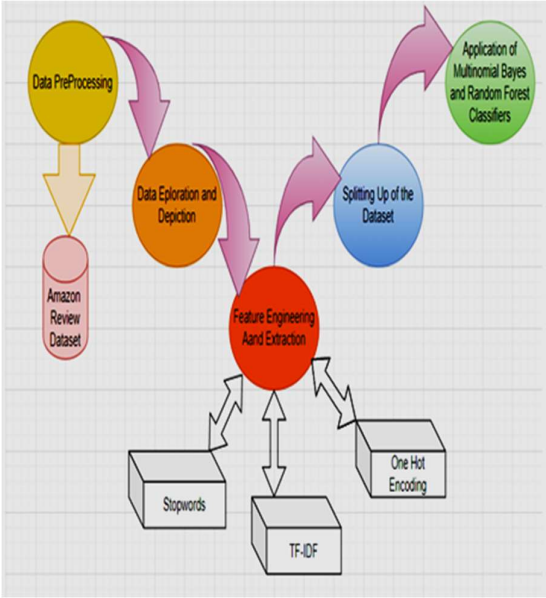

Figure 1.1 Stages of analysis of sentiments in the Amazon Product Reviews

There are 5 stages of methodology in the system:
1.Loading and Preprocessing of data
2. Data Exploration and Depiction
3. Feature Engineering and Extraction
4. Training and Testing the Dataset
5. Application of Two Classifiers:- Multinomial Bayes and Random Forest Classifiers

3.1 STAGE 1: Loading and Preprocessing of Data

Data is loaded from the Amazon Review Dataset which has a collection of fifty thousand plus Reviews of Amazon Products. This Dataset will be studied and operated upon throughout the processes to help train our model and provide results based on it. This Dataset gives an accurate description of the products whose reviews were recorded on their respective websites on the internet and has labeled and ordered columns that provide reliability and consistency to the model. Choosing a proper dataset is crucial to avoid overfitting and interdependency of the variables during the Training and Testing of the model.

Preprocessing of data mainly constitutes data cleansing to enable datasets to become free of errors and duplications. Without cleansing the data, data becomes corrupted which becomes a hindrance to the machine learning process. It also removes any unnecessary characters or punctuations that affect the accuracy results of the model. The dataset contents are also required to be standardized in the same format to make them comparable. Data Preprocessing consists of a wide variety of operations such as splitting of the text, removal of special characters, removal of punctuations(such as question mark , commas, etc .), converting the sentences to lowercase letters, and so on.

3.2 STAGE 2:Data Exploration and Depiction

Plotting and Depiction of data in bar graphs is very much useful in comprehending the relationships between the data. We can gain much more insight from this way of envisioning data and the detection of anomalies in the patterns of data is easily done. Even one single unit of data that is showing misbehavior and taking different paths than the other variables can prove to be the cause of miscalculations in future operations. Every learning model needs this stage to make correct decisions in the successive stages of methodology.

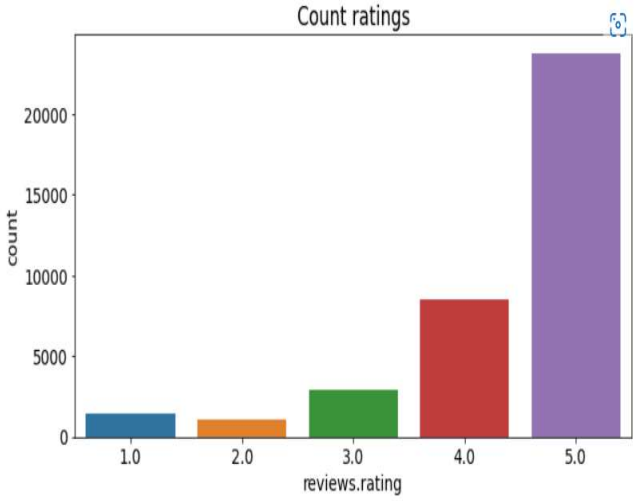

Figure 1.2 Plotting a Bar graph to count the ratings present in the reviews

3.3 STAGE 3: Feature Extraction and Feature Engineering

Feature Extraction is one of the major techniques of Machine Learning that obtain features from the text of the required data that are necessary for producing results for the Training and Testing of data.
Implementing upon raw texts of data ia s a hugely erroneous task that can incriminate predictions of the overall training model.
Feature Extraction has Two Constituents:

- BOW(Bag Of Words Model) and
- TF-IDF Vectorization

3.3.1 Bag Of Words Model

It consists of the creation of features after cleaning of data.

features are the various types of words of the text data that can easily summarize the data and convey an accurate description of data. With the help of Features, we can easily analyze the sentiment of text and we do not need unnecessary information which can slow down the throughput of the model. We can produce features by removing the stopwords from the text and then setting conditions that will determine the groups to which the words of the text will belong to. For example, a text review that contains negative sentiments will have features that are labeled as negative while a group of words containing positive sentiments will have features that are labeled as positive respectively.

3.3.2 Stopwords

Stopwords are lists of words in the English dictionaries that bestow English Grammar to the texts but are an obstacle to machine learning models and slow down the processes with their heavy amount of words. They can also make the training model mistake a bad review asa good review if the review contains the words such as ‘positively disgusting’ and this results in an inaccurate prediction.
Therefore, they are required to be removed so that models can be trained in a much more fast manner.
In this paper, Stopword() function is used to expel all the stopwords.

3.3.3 WordCloud

Stopwords can also be used to visually depict the different types of sentiments contained in the dataset in overall with the assistance of WordCloud. WordCloud is a type of generator of words that are occurring frequently and are highly favored in the depiction of sentiments in a vividly eye-catching manner.
It provides a deeper analysis of the text.
We used the WordCloud() function to derive common positive and negative words.

Feature Engineering is the manipulation of existing lists of data, so that we can create new features by categorizing the variables in the dataset. This action simplifies data and makes the dataset more structured, resulting in the provision of data that can be easily Trained and Tested. It is commonly used Deep Learning to enhance the accuracy of models. It also creates correlation between various attributes and provides Dimensionality reduction.
.

Figure 1.3 Depiction of the WordCloud

3.3.4 One Hot Encoding

One Hot Encoding is one of the most common techniques of Feature Engineering that is applied to datasets to categorize their contents. It creates new features for every content in the columns of the dataset and labels the features as numerical values so that each element can be identified by the values that were created after the implementation of One-Hot Encoding.

3.3.5 TF-IDF(Term Frequency-Inverse Document Frequency)

It represents the words that may or may not be occurring frequently in our compilation of files but are definitely crucial for our present file. It is composed of two parts:

- Term Frequency and
- Inverse Document Frequency

Term Frequency is the number of times a word occurs in a file with respect to the total number of words.
Inverse Document Frequency represents the words that are occurring frequently in a small amount of files.

We will use this vectorizer to covert our reviews into tokens.

3.4 STAGE 4:Splitting Up of Dataset

We have to split our dataset into two parts so that they can be used for the training of the model:

- Training Dataset
- Testing Dataset

3.4.1 Training Dataset

It is the dataset which is submitted to the machine learning model to invent new patterns. These patterns can later be learned to help the model develop further.

3.4.2 Testing Dataset

3.4.1 Training Dataset

It is the dataset which is submitted to the machine learning model to invent new patterns. These patterns can later be learned to help the model develop further.

Figure 1.4 Splitting up of Training DataSet and Testing Dataset

3.5 STAGE 5: Training and Testing The Model

We are going to train our model on the basis of two machine learning algorithms,namely, MultinomialNB() and Random Forest Classifier().we have already tokenized the required values using TF-IDF and have also split the dataset into Training and Testing data.

Training of the model will be done by the Training Dataset to predict the outcome of the model.
Then we test the model with respect to the target variable so that we can find out the model's accuracy by using the test data.

Algorithms:

3.5.1 MultiNomialNB()

It is a machine learning algorithm is a Naïve Bayes Classifier which has features generated from distribution.

These features are related to the occurring frequency of the words in a document and require classifying. Apart from Gaussian Bayes, it is the only other one of the Naïve Bayes classification algorithm.

3.5.2 Random Forest Classifier

It is a type of supervised machine learning algorithm that represents a set of decision trees that were constructed from a randomly selected part of the Training Dataset. This group of decision trees are called a forest. Random forest classifier has two parts, mainly classification and regression. votes are collected from each tree to choose the class with the most votes in classification..
The mean of all the individual trees' results is chosen as the final result in regression.

4. RESULTS

We have the following results after Training our model with the two algorithms: Multinomial Naïve Bayes and Random Forest Classifier.

4.1 MultinomialNB()

```
Train accuracy : 89.90424439216726
Test accuracy : 89.19957593426982

CLASSIFICATION REPORT
              precision    recall  f1-score   support

    Negative       0.95      0.28      0.43      1106
    Positive       0.89      1.00      0.94      6440

    accuracy                           0.89      7546
   macro avg       0.92      0.64      0.69      7546
weighted avg       0.90      0.89      0.87      7546
```

Figure 1.5 Observations of MultinomialNB() Technique

The accuracy of the model for the training and testing data is 89.90% and 89.19% respectively as shown above. Therefore we can deduce that the training accuracy and testing accuracy of the model is nearly similar. It represents that in the

case of real scenario, the model performs perfectly.

4.2 Random Forest Classifier

```
Test accuracy 0.9347115469564449
Train accuracy 0.9993941229930324
```

Classification Report(test)

	precision	recall	f1-score
Negative	0.92	0.60	0.73
Positive	0.94	0.99	0.96
accuracy			0.93
macro avg	0.93	0.80	0.85
weighted avg	0.93	0.93	0.93

Figure 1.6 Observations of Random Forest Classifier Technique

The accuracy of the model for the training and testing data is 93.47% and 99.93% respectively as shown above. Therefore we can deduce that the training accuracy and testing accuracy of the model is nearly similar. It represents that in the case of real scenario, the model performs perfectly.

We can clearly see from the above results that Random Forest Classifier has more percentage of accuracy than Multinomial Naïve Bayes classifier. Hence the former should be preferably used.

5. CONCLUSION

Online Transactions have become a huge part of our life in recent days due to their convenience , of the products and fast delivery of the products. To determine the sentiment analysis of the feedbacks left by customers who had already sampled and bought the Amazon Products, we made this framework by using the Multinomial Naïve Bayes Classifier and the Random Forest Classifier in order to produce the best results with the given dataset. The accuracy results given by Multinomial Bayes and Random Forest Classifier are 90% and 99% respectively and can be developed further by using another advanced NLP classifier.

We hope that in the future our model will not only just limited by Amazon Product Reviews but can be broadened further for other research areas in the field of Data Analytics and produce similar results and prove useful.

6. ACKNOWLEDGEMENTS

The Author has not received any financial funding for the overall research, document preparation, or publishing of this paper.

7. REFERENCES

[1] Mohan Kamal Hassan, Sana Prasanth Shakthi, and R Sasikala “Sentimental analysis of Amazon reviews using naïve Bayes on laptop products with MongoDB and R.” November 2017; IOP Conference Series Materials Science and Engineering 263(4):042090

[2] Sepideh Paknejad “Sentiment classification on Amazon reviews using machine learning approaches”.

[3] Nguyen, Heidi; Veluchamy, Aravind; Diop, Mamadou; and Iqbal, Rashed (2018) "Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches," SMU Data Science Review: Vol. 1: No. 4, Article 7.

[4] Sana Nabil, Jaber Elbouhdidi, Mohamed Yassin Chkouri “Sentiment Analysis Of Amazon’s Reviews Using Machine Learning Algorithms” Journal of Theoretical and Applied Information Technology, E-ISSN: 1817-3195, 30th November 2021. Vol.99. No 22

[5] Anjali Dadhich and Blessy Thankachan “Sentiment Analysis of Amazon Product Reviews Using Hybrid Rule-based Approach” I. J. Engineering and Manufacturing, 2021, 2, 40-52 Published Online April 2021 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijem.2021.02.04

[[6] Sobia Wassana , Xi Chenb , Tian Shenc*, Muhammad Waqard , NZ Jhanjhie “Amazon Product Sentiment Analysis using Machine Learning Techniques” Revista Argentina de Clínica Psicológica 2021, Vol. XXX, N°1, 695-703 DOI: 10.24205/03276716.2020.2065

[7] Aamir Rashid and Ching-yu Huang “Sentiment Analysis on Consumer Reviews of Amazon Products” International Journal of

Computer Theory and Engineering, Vol. 13, No. 2, May 2021

[8] Arwa S. M. AlQahtani "Product Sentiment Analysis For Amazon Reviews",(June 2021), Volume 13, Number 3, International Journal Of Computer Science & Information Technology(IJCSIT), ISSN:0975-3826

[9] Gayatri Khanvilkar, Deepali Vora (January 2019), "Product Recommendation using Sentiment Analysis using Random Forest Approach Gayatri" International Journal of Engineering and Advanced Technology (IJEAT)ISSN: 2249 – 8958, Volume-8, Issue-2S2.

[10] Lemons, K., 2020. "A Comparison Between Naïve Bayes and Random Forest to Predict Breast Cancer". International Journal of Undergraduate Research and Creative Activities, 12(1), pp.1–5. DOI: http://doi.org/10.7710/2168-0620.0287

[11] Alsaeedi, Abdullah & Khan, Mohammad. (2019). "A Study on Sentiment Analysis Techniques of Twitter Data." International Journal of Advanced Computer Science and Applications. 10. 361-374. 10.14569/IJACSA.2019.0100248

[12] Iqbal, Muhammad & Muneeb Abid, Malik & Noman, Muhammad & Manzoor, Engr. Dr. Amir. (2020). "Review of feature selection methods for text classification." International Journal of Advanced Computer Research. 10. 2277-7970. 10.19101/IJACR.2020.1048037.

[13] Karolina Ilieska, (2013). "Customer Satisfaction Index – as a Base for Strategic Marketing Management", TEM Journal, 2(4), 327-331.

[14] Oyamada, Masafumi. (2019). "Extracting Feature Engineering Knowledge from Data Science Notebooks". Conference: 2019 IEEE International Conference on Big Data (Big Data)6172-6173. 10.1109/BigData47090.2019.9006522

[15] Mutlag, Wamidh & Ali, Shaker & Mosad, Zahoor & Ghrabat, Bahaa Hussein. (2020). "Feature Extraction Methods: A Review." Journal of Physics: Conference Series. 1591. 012028. 10.1088/1742-6596/1591/1/012028.

[16] Ul Haq, Ikram & Gondal, Iqbal & Vamplew, Peter & Brown, Simon. (2019). "Categorical Features Transformation with Compact One-Hot Encoder for Fraud Detection in Distributed Environment": 16th Australasian Conference, AusDM 2018, Bahrurst, NSW, Australia, November 28–30, 2018, Revised Selected Papers. 10.1007/978-981-13-6661-1_6.

[17] Abbas, Muhammad & Ali, Kamran & Memon, Saleem & Jamali, Abdul & Memon, Saleemullah & Ahmed, Anees. (2019). "Multinomial Naive Bayes Classification Model for Sentiment Analysis." 10.13140/RG.2.2.30021.40169.