Amazon Reviews Sentiment Analysis

1st Mohammad Bilal Gul
Applied Machine Learning
001196113

***Abstract*—The online shopping becomes one of the major trends after covid-19 more specifically. Many companies shift their businesses totally to the online system even the working environments also adopted the work from home scheme. The online reviews after the shopping from specific website implies a high influence on the customers. According to their reviews customers buy or leave a product. So, the companies are also interested into check the behavior of customers like which product they like the most. So, they went to do Analysis of the reviews to see customer behavior. Therefore, the choosing of right Machine Learning Model, to perform the analysis to get proper results so companies can decide their business plan for future. The major parts in the Analysis are the Data Cleaning, Splitting the data, Choosing right ML method and apply it on data set and review the results.**

I. Introduction and related work

In ecommerce, product reviews have a considerable significance as people buy more and more products through online platforms and websites. As computing devices and tools advancing, digital marketing is now more complex than ever. Consumers base their decision of purchasing something online based on the existing reviews and satisfaction level of former customers. These reviews are evidence of the viability of both the product and companies selling that product. The extraction of review data is a very hectic and time taking job due to the unstructured nature of the data. A lot of previous studies show the works done in the field of recommendation systems and sentiment analysis. This study particularly focuses on the predicting the product category and number of stars of an amazon review. This can help guide customer decision on whether to buy a product or not. It also helps sellers improve their product's quality and supply-chain to enhance customers' buying experience. The existing literature does not show much evidence on the work done in this particular area of predicting review score or product category. In 2018, Haque, Saber and Shah used a semi-supervised approach called 'Active Learning' which, in their claim, uses less data for training and gives better performance providing that the data is chosen by the algorithm for learning. They used pool based active learning to label the dataset. They conducted their experiments using different classification algorithms. Their study showed that common features among TF-IDF and bag of words showed the best results. They achieved a peak accuracy of 94.02As after having a literature review, I came to know that there is mostly work which Is already done on majorly some columns in the data set which are the reviews and the score which is given to those reviews. As per literature review, I came to know that the other columns rather than the discussed one doesn't help so much to get any results in analysis. But there is no such work still found which is exactly same or exactly same methods implemented as required in our dataset. Major part of the related work done in the Python as the programming language and platform which used is the Jupiter Notebook. I found the less work which performs on the Google platform. The sentiment analysis which already done is using different techniques of ML (Machine Learning) rather than using Proper ML models such as the Linear regression and Random Bost. As by reviewing their work they mostly used the word extraction techniques and converting of text into the numbers (spacy method, Tf-IDF) and then compare them to the number of stars and get the sentiment analysis. According to some researchers these methods are not the most perfect which gave us the most concise and perfect results on which base we can judge some of the products in the future.

II. Ethical discussion

One of the major use thin online in USA and UK is the Amazon. Amazon got a major users/audience from these two countries. So, there are a million of accounts and products are on it. The trillions of reviews are on the website. Amazon itself providing the cloud computing services to the users of IT industry. And they are so conscious about the data privacy of their users. So, the dataset we have provided is the basic data which only includes the review ids and their reviews and review scores which are not going to harm or arise any legal issues among the users. Moreover, the dataset does not include any further detailed information's of the users such as the Images, Phone Numbers, and addresses which we arise issues. There is just a use of review text and their score for our machine learning models which are not going to hurt any organization or group of people. We are going to see the different data insights after applying the ML methods, which even just get the results of those specific products for only ourselves to go through the procedures and results whether our models are even trained to analysis the real time data or not. There is no such methods or visualizations are used which may affect someone's sentiments. And it's totally ethical to work as this in which there is no such activities perform which are unethical.

III. Dataset preparation

For our data set preparation, I used the most effective and popular Data-Preprocessing method which is the Exploratory Data Analysis (EDA). In our given data set we have five different columns which are labeled as the review id, text

(review text), verified, review score and product category on which the reviews are given. Therefore, the major columns for us on which I am going to work are the text, review score, and the product category. Choosing these columns are on per requirements given in the document to us. The major EDA steps here are the dropping of unnecessary columns, removing empty rows. Dropping and removing these entities helped us to get the precise and correct results in the future. By viewing of dataset, there are some duplicate rows and the review scores which have the values which are negative. So, working on duplicate and negative values are not helpful for us. So, simply we remove these duplicate and negative value with the help of EDA. There is a lot of data cleaning is done in this part. Furthermore, there is another necessary thing which we need is the splitting of the product category into the prime pantry and luxury beauty. After cleaning the data, I balanced the data according to the machine learning models so we can get the results more quickly and correctly. So, First I balanced the product category as a Sampled data so we can train our model properly and secondly, I balanced the review score so we can get the more insights of the data. After all these processes, for our Dataset Preparation, I split the data by default through sklearn which split the data automatically into the proper ration to get insights and visualization. After splitting the data into the sample and original I create the different pipeline for our ML methods. After cleaning of the data, it is hard to choose the ML models for the dataset because even an expert ML engineer cannot tell us about the model which performs well and produce best results. So, choosing the ML model is quite a big decision to take. After having all lectures and labs of my course, and by research I decided to use the Random Forest Classifier, ADA boost classifier, XGB regressor. After cleaning all the data after implementation of EDA, balancing, and splitting the data and creating pipelines, hopefully in future we are not getting any errors during the implementation of the ML models to the data sets.

IV. Methods

There are three different methods which I used on the data to get the results. The methods are the Random Forest Classifier, ADA boost classifier and XGB regressor. We have a limited data for the experiment not having the real-time data or the large amount of data which we cannot be handle or take more time so we can choose the different methods for our dataset. Our data is majorly in the form of 0 and 1 as we converted it into this form to train our models in future. So, the random forest has its own decision trees which operate as the ensemble (using multiple learning algorithms which predict the best results). So, these decision trees perform in a committee and outcome as one single model to get the better results. The results of the Random Forest classifier method are very great. Let's have an example I used a random number generator to get a number and set the constraints that if the selected number is greater than for example the 50 then the predicted number is positive and if it's under the 50 then it is negative. So, that is why this method is very suitable for our dataset if the review number according to text are more than a given constraint number than it would be the positive review and if below then it would be the negative review. The prediction of this method is mostly correlated and provide vast number of results and this methos has a nature of prediction. I split the data into the two different columns as the original dataset and the sample data column. So, second method which I used is the ADA boaster which always take the original data which is always in the less amount and perform the meta-estimate on it and after getting the proper results it copies the same calcification to the sample data to get the results and so we can see different insights of the results. The ADA boaster majorly focuses on the difficult case of the data set, which are hard to solve and always having some problem. This method increases the accuracy of the classifier. In this method, the chosen classifier must be trained interactively on different weighted training examples. For the supervised regression models, we mostly use the XG boost model, for getting the more accurate and accurate results. Therefore, I also used the XGB regressor method as a booster for getting some extra results about our dataset so we can explore the more results rather than the others.

V. Experiments and evaluation

After completing the EDA process of the data, I balanced the original data with the sample data and then split the data into the trained and sample part. Then I create the pipelines for the models. Then First, I created a method for machine learning pipeline which decides whether the experiment is on original dataset or sampled dataset based on arguments. Similarly, this method decides if the XGB Regressor is going to be used or not based on the approach that is intended to be implemented through method arguments. For evaluation, I am using typical metrics of evaluation for machine learning methods like the rate of false positives, true positives, true negatives, and false negatives. This can be visualized through confusion matrix which I have used in my implementation. Moreover, I have used built-in metrics provided by sklearn library that shows us how accurate our model is. In the first approach I have used both ADA Boost and Random Forest Classifiers separately and compared their results for both original and sampled data one by one. In the second method, I have used these two models along with XGB Regressor for ensambling which enhances the working of these models for original and sampled data. All the models have some pre-installed libraries. So, I need to import those libraries which are important for our model and then put the required data into those models and get the visualizations to understand and get some results. In the evaluation, as I mentioned in the upper paragraph, we have the metrics to evaluate our models results. Furthermore, as we know I used the train data and sample data for implementation of the models. So, if the results of sample data are as accurate as the train data, then our models are working properly, and we apply the correct methods. The numbers of results may be change if we change the training and sample data numbers.

VI. Discussion and future work

As by going through the whole report and the work, the almost project is working fine and there is no such errors or warnings which we can discuss here. The results of the models are quite impressive, and they gave us the information of the dataset and we can further implement other models such as the SVM and Linear regression to see more insights of the data. As we used the sk-learn libraries and different high intensity models on our data, so the loading time of the models take a bit long time to load and present the results to us. So, while taking time, that does not mean it is not working or maybe there is any problems. Most of these pipelines and models take time. The following assumption can be made after the work and research done for the future work: If there are some extra columns added into the dataset which are useful then we the ML models can present more results. The additional features can be extract from the dataset on applying more libraries to it. The more characteristics of the speech can be found which we used in this dataset. These ML methods can be used for other type of data sets if they are pre-processed properly.

VII. Conclusions

In the end, the machine learning techniques are very helpful to solve any problem even if the models are trained properly then we can use those models for solving the big problems. Such as the amazon peer review can be used for the eBay and Walmart peer group review. Because the reviews are mostly in the text type and score in the number scale from 1 to 5. We evaluate the behavior of customers on their text reviews in this research paper. The proper built-in functions are available in python to use for the cleaning and pre-processing of the data. The splitting and pipeline creations helped us to train the dataset for the ML methods.

References

[1] Brownlee, J. (2020). How to Develop a Random Forest Ensemble in Python. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/random-forest-ensemble-in-python/.

[2] Brownlee, J. (2021). XGBoost for Regression. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/xgboost-for-regression/.

[3] GeeksforGeeks. (2020). XGBoost for Regression. [online] Available at: https://www.geeksforgeeks.org/xgboost-for-regression/ [Accessed 08 Jul. 2022].

[4] Haque, T.U., Saber, N.N. and Shah, F.M. (2018). Sentiment analysis on large scale Amazon product reviews. 2018 IEEE International Conference on Innovative Research and Development (ICIRD). doi:10.1109/icird.2018.8376299.

[5] IBM Cloud Education (2020). What is Random Forest? [online] www.ibm.com. Available at: https://www.ibm.com/cloud/learn/random-forest.

[6] kaggle.com. (n.d.). An introduction to XGBoost regression. [online] Available at: https://www.kaggle.com/code/carlmcbrideellis/an-introduction-to-xgboost-regression/notebook [Accessed 10 Jul. 2022].

[7] Paperspace Blog. (2020). A Guide To Understanding AdaBoost. [online] Available at: https://blog.paperspace.com/adaboost-optimizer/.

[8] Scikit-learn (2018). 3.2.4.3.1. sklearn.ensemble.RandomForestClassifier — scikit-learn 0.20.3 documentation. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[9] scikit-learn.org. (n.d.). sklearn.ensemble.AdaBoostClassifier — scikit-learn 0.23.2 documentation. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html: :text=An

[10] www.datacamp.com. (n.d.). AdaBoost Classifier Algorithms using Python Sklearn Tutorial. [online] Available at: https://www.datacamp.com/tutorial/adaboost-classifier-pythonadaboost-classifier [Accessed 15 Jul. 2022].

[11] Yiu, T. (2019). Understanding Random Forest. [online] Medium. Available at: https://towardsdatascience.com/understanding-random-forest-58381e0602d2.