Text classification – Amazon Reviews

Bhanu Prakash Bandi
Applied Machine Learning
001183470

***Abstract*—For the current digitalized technology, most of the people choose online shopping. People from the online shopping are allowed to post reviews for the product purchased. Also the length of the review and the context depends on the product and how they like it. There are many online shopping sites such as Amazon, Flipkart, eBay, etc. This paper seeks to analyse the reviews posted from the Amazon website. Analyzing reviews to know customer satisfaction, potential areas to improve and identifying the products based on the reviews. This extends the work sentimental analysis in the field of natural language processing. RandomForestClassifier and SupportVectorClassifier are used as models to build the system. The dataset is all about for the video game and musical instrument reviews. The steps involved in experimentation is data preprocessing, data segregation and Machine Learning methods implementation. This paper aims to find the review score(number of stars) and the product category for the amazon reviews.**

I. Introduction and related work

As amazon is a biggest marketing platform where products are buying and selling the goods. There are thousands of products kept on sale. The big question arises for the market leaders whether the delivered product is satisfied by the customer and what are the ways to improve in certain areas. For a single product there are hundreds of reviews available. The paper mainly deals with developing insights for the posted reviews and narrowing down to customer satisfaction analysis. This is a business problem, to survive in the world one must understand in the perception of customers. Understanding the customers is the challenging task, but it's possible in reality by reading the reviews and developing the insights from to understand how well they are satisfied.

There are many literatures on sentimental analysis and recommendation systems which are similar to this paper. Sentimental analysis in amazon reviews using probabilistic machine learning [4]. Comparative study of machine learning approaches for amazon reviews this paper focuses on examining the efficiency of three models for the classifying the review as positive or negative. Sentiment analysis on large scale amazon product reviews was focused on getting accuracy based on the review satisfaction.

The dataset is collected from amazon reviews particularly in the domain of video games and musical instruments. The dataset consists of product Id, text(review), verified, review score and product category. The product Id is unique in data as this column is used to identify the user who wrote the review, text(review) is the main column to be focused, this column contains the text in different sizes, verified column holds the boolean data as whether the user is verified by the amazon certifier, the reviewscore is the target column in which based on the text column review the score is associated ranging from 1 to 5 and product category column contains the category of product either video game or musical instrument.

The preprocessing steps involved is checking the size of the data, dropped unnecessary columns such as product Id and verified column. The aim of this project is to predict the review score and product category. Therefore, the dataset is divided into two sub parts such as dataset1 is having text column and review score column, and dataset2 holding text column and product category column. The analysis is done separately to get the accurate results. The process also involved converting all the text in the text column to lowercase characters with the help of regular expression library which also involved removing any special characters. Next further processed to remove the stopwords with the help of nltk(Natural Language ToolKit) library then the processed text is converted text to numbers from the TfidfVectorizer(Term Frequency and Inverse Document Frequency) function where this function turns the text to weights or a position in the vocabulary by the word tokenizer.

The Machine learning methods used are RandomForestClassifier which has decision tree classifiers that fits number of sub data from the dataset and gives out by averaging the output for better accuracy, and SupportVectorClassifier finds a hyperplane that has a large margin where the maximum distance between data points of different classes.

The systematic representation of this project includes importing data, exploratory data analysis, pre-processing, data segregation, model building and evaluating the results.

II. Ethical discussion

For corporate organisations, this has a greater impact on improving consumer happiness. As part of this effort, Amazon customers' previous purchases and opinions are the primary emphasis. As well as helping consumers better understand a product they are considering purchasing, firms can utilise the data collected through this project to better understand their customers' needs. Using the internet, clients have access to

more information about products, including their quality and quantity, as well as what they should buy.

III. Dataset preparation

Pre-processing is the first step in data analysis. The data is processed in required format. Exploratory data analysis has two approaches, a statistical approach where business assumptions are made and conducting hypothesis test for analysis and another approach is for data modelling. These two approaches are used to explain the results for future modelling. The statistical characteristics are implemented in amazon reviews dataset. Data cleaning is converting raw text data to cleaned text data. This process undergoes removing special characters and making capital letters to lowercase. Also removing stopwords which holds less information for analysis. Data is divided into three parts training data, validation data and test data. In the training data, the model learns the weights to maximum reduce the loss function, altering the hyperparameters in the validation set to further reduce the loss function. Finally testing the results with the test data.

IV. Methods

RandomForestClassifier

Random Forest is widely used and effective method in machine learning involves creative learning models known as ensembles [3]. An ensemble takes multiple individual learning models and combines them to produce an aggregate model that is more powerful any of it's individual model alone. Although each individual models performs well, they will tend to make different kinds of mistakes on the dataset. This happens, each individual might overfit to different parts of the data. By combining different individual models into an ensemble, it can average out their individual mistakes to reduce risk of overfitting while maintaining strong prediction performance. Random Forest prediction, while performing classification, each tree gives probability of each class, probabilities averaged across trees and predict the class with highest probability.

$$P(\frac{H}{X}) = \frac{P(\frac{H}{X}) \cdot P(H)}{P(X)} \tag{1}$$

- The character in parentheses is represented by the letter P, which defines their probability.
- $P(H)$ is the prior probability of the marginal probability of H's prior in the sense that it has not yet taken into account the information contained in X.
- The probability, $P(H/X)$, is also known as the posterior probability since it incorporates the result of event X.
- The conditional probability of X given H is $P(X/H)$.
- A marginal probability of X is known as $P(X)$, and this is often the evidence in the form of data.

The reason for choosing Random Forest, it does not require careful normalization of features or extensive parameter tuning and like decision trees, handles a mixture of feature types. It can also parallelized across multiple CPUs. It is widely used, excellent prediction performance on many problems.

RandomForestClassifiers has key parameters, n estimators defines number of trees to use in ensemble, max features has a strong effect on performance, it influences the diversity if trees in the forest, max depth controls the depth of each tree, n jobs controls number of cores to use in parallel during training and random state reproduces the results.

Support Vector Machine

This algorithm is supervised learning, it is widely used and very powerful in both industry and academia [2]. It is sometimes more powerful way of learning complex non-linear functions. Support Vector Machine works on linear function and predicts the output a binary value by applying a sine function. One way to define a good classifier is to reward classifier for the amount of separation that can provide between two classes. For a given classifier, the margin is the width that the decision boundary can be increased before hitting a datapoint. There is a parameter called c parameter which acts as a regularization for SVM. The strength of regularization is determined by c. Larger values of c meaning lower bias and higher variance, that fits the training data as well as possible each individual data point is important to classify correctly. Smaller values of c means higher bias and lower variance, more tolerate of errors on individual data points.

$$f(\boldsymbol{x1}, \boldsymbol{x2}) = \boldsymbol{e}^{-}|\frac{(x1 - x2)^2}{2\sigma^2}| \tag{2}$$

where,

- x1 and x2 are data points
- sigma is the variance
- (x1- x2) is the Euclidean distance between the x1 and x2

In the case of linear support vector machine, they only use a subset of training points and decision function. Therefore, Kernalized SVM is a powerful extension of linear support vector machines and this can provide more complex models that can go beyond linear decision boundaries. It can also perform both classification and regression. There are different kernels such as Radial Basis Function, polynomial kernel, string kernel, chi-square kernel, histogram intersection kernel and so on.

The reason for choosing support vector is simple and easy to

train, scales well to very large datasets, works well with sparse data, reasons for prediction are relatively easy to interpret, different kernel functions can be specified or custom kernels can be defined for specific data types and works well for both low and high dimensional data.

V. Experiments and evaluation

The dataset is divided into three parts such as training set, validation set and test set. The experimentation and evaluation is divided in to 2 steps because there are two attributes.

The first attribute where review score to be predicted. The models used to analyse the results are SupportVectorClassifier and RandomForestClassifier are trained with 70 percent of data. For the SupportVectorClassifier there are hyperparameters like kernel function used in this model RadialBasisFunction(rbf) and gamma parameter which is a kernel coefficient for RadialBasisFunction used a range of values from 0.001 to 5. As the gamma parameter increases the accuracy is slightly increasing until gamma value reaches to 1, beyond 1 that is gamma value greater than 1 the accuracy starts decreasing. The SupportVectorClassifier outputting a best results when kernel is RadialBasisFunction and gamma value is 1, the accuracy is 64.24 percentage. And for RandomForestClassifier, hyperparameters used are number of estimators which means trees ranging from 100 to 1500 and maximum features used are square root of number of features available in dataset. As the number of estimators increased from 500 the accuracy is decreasing but not drastically reducing.The accuracy is 60.79 percentage for the 100 estimators with the RandomForestClassifier.

TABLE I
ATTRIBUTE-I: THIS TABLE SHOWS THE PERFORMANCE OF SUPPORTVECTORCLASIFIER BY TUNING THE HYPERPARAMETERS. THE COLUMN GAMMA IS THE CO-EFFICIENT OF KERNEL FUNCTION IN THIS CASE RADIALBASISFUNCTION AND THE ACCURACY COLUMN CONTAINS THE ACCURACY VALUES OBTAINED CORRESPONDING TO THE GAMMA VALUES.

gamma	Accuracy
0.001	0.6017
0.01	0.6134
0.1	0.6243
1	**0.6424**
5	0.5995

The second attribute is to predict product category. For this the same models are used such as SupportVectorClassifier and RandomForestClassifier. The SupportVectorClassifier is tuned with the same hyperparameters as used above. As increase in the gamma value the accuracy is increased upto a gamma value 1. The accuracy at gamma value 1 is 94.24 percent. Similarly, for RandomForestClassifier as the number of estimators increased the accuracy is also increased. At 2000 estimators the accuracy is 93.15 percent.

TABLE II
ATTRIBUTE-I: THIS TABLE SHOWS THE PERFORMANCE OF RANDOMFORESTCLASIFIER BY TUNING THE HYPERPARAMETERS FOR THE REVIEW SCORE. THE COLUMN N ESTIMATORS IS THE NUMBER OF TREES TO FORM PURE TREE AND THE ACCURACY COLUMN CONTAINS THE ACCURACY VALUES OBTAINED CORRESPONDING TO THE N ESTIMATORS VALUES.

n estimators	Accuracy
100	**0.6079**
300	0.6064
500	0.6070
800	0.6066
1500	0.6077

TABLE III
ATTRIBUTE-II: THIS TABLE SHOWS THE PERFORMANCE OF SUPPORTVECTORCLASIFIER BY TUNING THE HYPERPARAMETERS FOR THE REVIEW SCORE. THE COLUMN GAMMA IS THE CO-EFFICIENT OF KERNEL FUNCTION IN THIS CASE RADIALBASISFUNCTION AND THE ACCURACY COLUMN CONTAINS THE ACCURACY VALUES OBTAINED CORRESPONDING TO THE N ESTIMATORS VALUES.

gamma	Accuracy
0.001	0.6786
0.01	0.8845
0.1	0.9371
1	**0.9424**
5	0.6877

TABLE IV
ATTRIBUTE-II: THIS TABLE SHOWS THE PERFORMANCE OF RANDOMFORESTCLASIFIER BY TUNING THE HYPERPARAMETERS FOR THE PRODUCT CATEGORY. THE COLUMN N ESTIMATORS IS THE NUMBER OF TREES TO FORM PURE TREE AND THE ACCURACY COLUMN CONTAINS THE ACCURACY VALUES OBTAINED CORRESPONDING TO THE GAMMA VALUES.

n estimators	Accuracy
500	0.9291
1000	0.9302
1500	0.9312
2000	**0.9315**

VI. Discussion and future work

For both the attributes SupportVectorClassifier performed well giving out the accuracy 61 percent for the review score prediction and 93 percent for the product category prediction.

The RandomForestClassifier worked less well for the review score field and product category field outputting accuracy 60 and 93 respectively.

The reason SupportVectorClassifier is working good because it works best for the sparse data than the RandomForestClassifier. Also the hyper parameter used is gamma which determines how fast to converge the data, as gamma value increases the width of the converging curve.

This work can be extended by tuning other hyperparameters such as C as a regularization parameter, defining own kernel

functions for SupportVectorClassifier. Also RandomForestClassifier can be tuned to get better results by tuning maximum depth and maximum features.

VII. Conclusions

For the Amazon review dataset, the models used SupportVectorClassifier and RandomForestClassifier. Among these two the SupportVectorClassifier is showing best results, for the review score the accuracy of the model is 65 percent and for the product category 96 percent. Any review dataset like this, the SupportVectorClassifier model will work best to meet the objectives [1].

References

[1] Sanjay Dey, Sarhan Wasif, Dhiman Sikder Tonmoy, Subrina Sultana, Jayjeet Sarkar, and Monisha Dey. A comparative study of support vector machine and naive bayes classifier for sentiment analysis on amazon product reviews. In *2020 International Conference on Contemporary Computing and Applications (IC3A)*, pages 217–220. IEEE, 2020.

[2] William S Noble. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567, 2006.

[3] Thais Mayumi Oshiro, Pedro Santoro Perez, and José Augusto Baranauskas. How many trees in a random forest? In *International workshop on machine learning and data mining in pattern recognition*, pages 154–168. Springer, 2012.

[4] Callen Rain. Sentiment analysis in amazon reviews using probabilistic machine learning. *Swarthmore College*, 2013.