Importance of Neutrality in Sentiment Analysis: A Case Study of Amazon Food Reviews

School of Computing
National College of Ireland
Dublin

***Abstract*—In this era of social media reviews play a major role specifically in B2C and C2C business models. Reviews are given utmost importance as they affect the business performance directly. Most studies in respect to sentiment analysis ignore the neutral reviews thinking that vote of majority has greater impact and that there isnt much to learn from neutral texts. In this work, we propose sentiment analysis considering neutral reviews rather the rational way of analysis using Machine learning techniques Random Forest, Nave Bayes and Radial SVM. We evaluated their performance using k-fold cross validation and found that Random Forest outperforms others. Finally, we will see the importance, effects and advantages of considering neutrality in business perspective.**

***Index Terms*—Sentiment Analysis, Neutrality, Random Forest , SVM, Naive Bayes , Hyper parameter Tuning**

I. Introduction

Whats others thinking This is an important aspect for many people in the world and plays a crucial role in decision making. Each and everyday people discuss many things and shed their views on various social platforms. Many companies find different ways to gather this data as it is rich in content and can help in growing their business. Sentiment analysis is a method of mining and obtaining information with regards to public perspective about a particular topic. In order to reap full benefits of this, these days many E-commerce sites are providing a social platform integrated with their current services and people are opting for this because of convenience,door delivery etc. There is a lot of chances for the customers to get tricked as they dont experience the product they are going to buy. In this virtual world of shopping, reviews play an important role in telling new buyers about their experience with the product which they have already bought and provided a rating for it. These reviews and ratings are important for new buyers in finding how worthy a product is? Hence analysing these reviews and educating the buyers with sentiment of people helps them in understanding about product better. On the other hand, organisations can also learn how people are reacting for their product in different markets. This information provides exact facts and opinions of about the product better than any sales or marketing analyses.

In general sentiment analysis is of three types aspect level, document level and sentence level [1]. In document-level analysis it is considered to be of a single topic and analysis is done. Similarly, sentences are considered and analysed. Naturally when compared, between these two there isnt much difference as sentences are a part of document. In aspect level, it is at the minimum grain level. In these types of analysis, there are a lot of disadvantages and care has to be taken when considering the data as reviews can be posted by anyone so chances of fake reviews are more which directly affect the sentiment analysis. The data that has been considered in this project is from Amazon [2]. The reviews posted are by its customers who have bought and experienced the product.

In general, research on sentiment analysis is performed as binary classification considering the positive and negative reviews. They ignore the neutral reviews leading inappropriate results. In our case study, we are investigating the importance of neutrality and impact of ignoring neutral reviews in sentiment classification on accuracies of model.In this project, we are going to see how sentiment analysis varies when neutrality is considered and how it varies when neutrality is ignored. Our objective is to say how the effect of neutrality on model performance and depending the performance of model, we will deliver meaningful insights to enhance the performance of amazon food industry.

This report is organised as follows: Section II explains related work and model selection. Section III is about methodology chosen. Section IV explains about implementation of data preprocessing and modelling. Section V discusses experimentation and results. Section VI is about drawback of model. Section VII explains conclusion and future work

II. Literature Review

In the paper [3], author considers using neutral samples, unlike other research papers where they have positive and negative samples. Further author proves the effectiveness of using neutral samples by using SVM, KNN and logistic regression and got an accuracy of 86.17% through SVM. In addition, the author has proposed a model to divide neutral samples into positive and negative ones to construct a final opinion polarity classification system.

[4] applied four popular machine learning algorithms (including SVC) for conducting sentiment analysis on two Twitter datasets by experimentally comparing 16 pre-processing methods. They also evaluated these pre-processing techniques on two aspects the number of features that they produce and the resulting classification accuracy. They concluded that basic techniques like removing the numbers, replacing contractions, or lemmatization can potentially enhance the accuracy, while some methods like removing punctuations do not.

[5] discussed about the importance and challenges of conducting a sentiment analysis on social media datasets. They used various pre-processing techniques on their dataset to normalize the impact of bogus factors in social media reviews. The results of their experiment show that the best level of precision, recall, and accuracy for positive and neutral sentiments is produced by Nave Bayes. However, this is not the scenario with negative sentiments.

(Gamal, D., et al, 2017) discussed how clients preferences are affected by social media and how observing and measuring the social media activities is an effective approach to measure loyalty and preferences of customers. They used several algorithms including SVM and Nave Bayes to derive information that the customers opinions. They concluded that sentiment classification accuracy of 92% can be achieved by using SVM with part of speech (POS).

[6] elucidate and applications and demand of predictive classification the difficulty in finding an accurate method for finding the polarity of a dataset. They compared the accuracy of popular methods like SVM, NB, and ME, to conclude that although all classifiers display great performance, NB outperformed the others.

[7] focused on analysing the efficiency of SVM, NB, and ME for classification of online customer reviews (divided as neutral, positive, and negative). Their results show that SVM is the best classifier when it comes to accuracy.

[8] Addresses the issue of fake reviews and talks about how trustworthy are the social media information. A model is developed which can detect fake reviews. Random Forest based supervised classifier outperforms others with an accuracy of 80.6%

[9] discusses the negative and positive polarities with respective to an entity or an item. In here they have performed analysis using Random Forest and SVM and got an accuracy of 88.4 and 91.4 percentage respectively.

[10] he author considers to find sentiment of airline industry customers and chooses twitter as a data source. Six individual classifications namely Naive Bayes, SVM, Bayesian Network, C4.5, Decision Tree and Random Forest are performed. An accuracy of 90.8, 90.4, 90.5 and 87.2 percent was obtained for Random forest, Nave Bayes, Bayesian Network and SVM respectively.

[11] attempted to implement sentiment analysis on the introduction of GST in India to gain residents opinion on this matter. They concluded that most of the people had positive opinion on this issue. However, since it is a fairly untouched area, they couldnt compare the performance of their system with similar experiments.

[12] used sentiment analysis to extract sentiment polarity on customer opinions based on specific characteristics of products using Nave Bayes. As per their results, the system can perform the analysis with F-1 measure of 78.12

[13] compared the dependence of sentiment analysis on emotional dictionary and artificial intelligence with depth models. They combined CNNs model with SVM. Their results showed that the model proposed by them improved the accuracy of sentiment analysis.

[14] talk about the importance of digital reviews is todays era and how it influences consumers shopping patterns. Across all the methods used, SVM produced best level of accuracy (81.75

[15] conducted opinion mining on Twitter data. They also extracted non-text aspects from tweets for training the algorithm which classifies the tweets into relevant categories (Neutral, positive, or negative). Their results indicate that their method outperforms in terms of precision, recall, and F-score.

[16] used a classification approach to enhance the predictive capacity of sentiment analysis. They incorporated five features in their method to train seven classifiers for

performance comparison. Their results depict that the best performance is achieved by random forest classifier.

[17] discussed the exponential growth in and increasing demand for the information shared on social media platforms. They summarized the opinions using ME, SVM, NB, and Random forest methods. Their results show that SVM is better than NB in all areas. Their proposed method, however, generates even better results than any other classifier (including SVM).

[18] presented an approach for analysing user opinions using popular data mining classifiers. Also, they compared the performance of each of the classifier. The results demonstrated that the highest accuracy is generated by k-nearest neighbour classifier. Moreover, ensemble of classifiers under performed as compared to the single classifiers.

[19] did an interesting research by linking peoples opinions about a companys products/services to its stock price. Using sentiment analysis, they analysed peoples emotions about a companys reputation. They used NB and random forest approach for this purpose. Their experiment results show that when hybrid feature and previous stock price are used in prediction models, a prediction in terms of coefficient of determinant 0.999 can be achieved.

After going through the literature review we found out that SVM, Nave Bayes and Random Forest are the best suitable algorithms for us to perform the analysis. The SVM is a supervised learning model which uses the statistical learning theory and Principle of minimal structural risk from VapnikChervonenkis Dimension Theory. Unlike the other models SVM does not use any probabilistic methods, it uses the feature vectors to represent the data in the feature space and generates a hyperplane to separate the data points of different classes. The Naive Bayes is one of the probabilistic classifier model works based on Bayes Theorem. The Naive Bayes gives an excellent classification when the dimensionality is high and outperform the classification methods. Random forest is the simple and ensemble model which can be understable and can be applied to any type of data. It has capability of dealing imbalance data. As per our dataset, we are experimenting on 60000-80000 records. So, we have taken these models instead of deep learning techniques.

III. Methodology

In this project, we used Cross Industry Standard Process for Data Mining (CRISP-DM) methodology. The CRISP-DM is ideal for the project as it can be used for many datasets or applications according to the requirements . The work flow of methodology is shown in Figure 1. More details about each process in methodology are explained in Section IV.

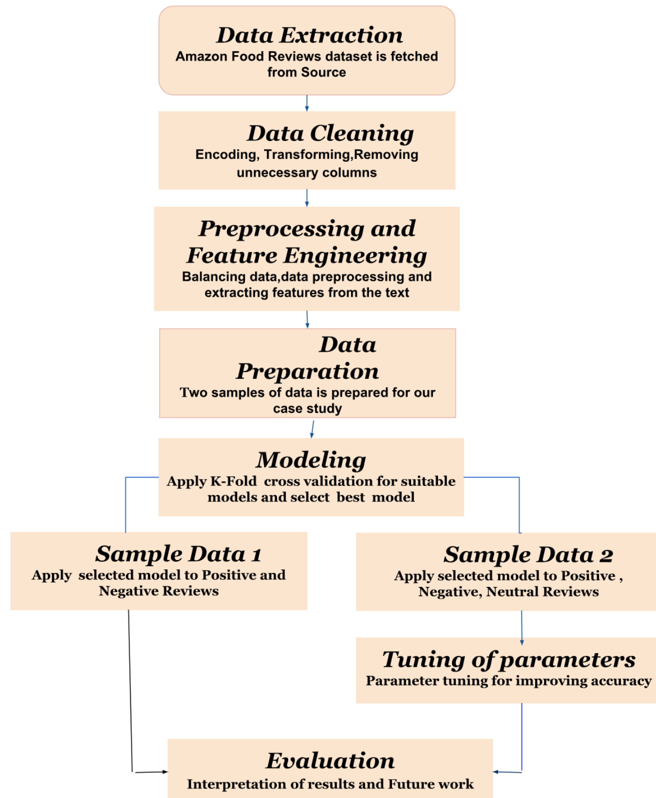

Fig. 1. Methodology

IV. Implementation

A. Data

The dataset is taken from Kaggle [20]. This dataset is about Amazon food reviews and it consists of 0.5 million samples and 10 attributes. This is a labelled dataset. We understood that a good dataset is required for yielding better results. In this study, there are three classes positive, negative and neutral and sentiment analysis is performed on multiclass classification. There are a greater number of positive reviews compared to negative and neutral reviews.

B. Data Cleaning

As mentioned earlier about samples and columns, sentiment analysis requires only text of reviews and score of reviews. Score is dependent variable and text is independent variable. Score has 3 levels and it is encoded into positive (+1), negative (-1) and neutral (0). There are no missing values in the data. As score is dependent variable, it is converted into factor data type. Text is of character data type.

C. Data Preparation

As explained in Section IV-A, dataset from Kaggle consists of positive, negative and neutral reviews. This we call it as **Data-PNN** and used this data for first set of experiments. For the second experiments, we filtered out neutral reviews and created another dataset and calling it as **Data-PN**. From here after **Data-PNN** refers to the data consists of positive, negative and neutral reviews and **Data-PN** refers to the data consists of only positive and negative reviews. Initially we need to explore data before preparing samples for training and testing. In this dataset we found that there is a class imbalance among the reviews. Positive reviews are far more than negative reviews. In the sentiment analysis, balanced positive, negative and neutral reviews are important for improving performance of sentiment classification accuracy. As there are few negative and neutral reviews, classifier tend to classify negative and neutral as positive reviews. This leads to potential misclassification. In general, accuracy will be higher for these class imbalance models. But we have to check other metrics like sensitivity, specificity AUC-ROC curve during evaluation to make sure everything is correct.

Therefore, class balance is necessary, and it can be done by SMOTE, ROSE, down sampling and up sampling methods. In this project, we have multiclass dependent variable (positive, negative and neutral). ROSE package is applicable to binary classification. SMOTE function is used to balance but it is balancing two classes keeping the other class constant. It is not the case in current work. So, we have randomly down sampled positive reviews and negative reviews to balance with neutral class. The plot of before and after balancing is shown in Figure 2.

Hence, we apllied data balance procedure to balance the two datasets **Data-PNN** and **Data-PN**. After balancing **Data-PNN** consists of approximately equal number of positive, negative and neutral reviews. Similarly **Data-PN** consists of equal number of positive and negative reviews.

D. Text Preprocessing

In text analytics, text preprocessing and extracting features is considered as more important than choosing and building a model. The data preprocessing is done using powerful R packages like **tm** and **Snowball**. The following steps were done for preprocessing and

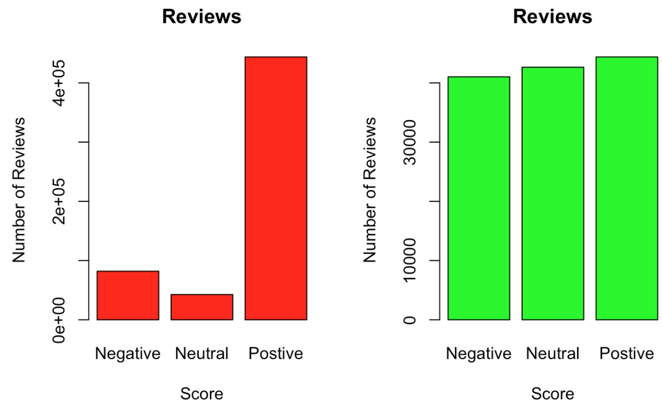

Fig. 2. Class Balance

to transform the data as machine learning models can understand:

- **Lowercase:** All the data is converted to lowercase to avoid case sensitive errors
- **Numbers, Punctuation and Special characters:** Usually the text data consists of unwanted symbols and characters. These wont give any meaning to take part in building a model. All these were removed to make the model robust.
- **Links and Hashtags:** People often use hashtags and post links while sharing reviews. The dataset has plenty of reviews with the combination of website links and hashtags. These do not enumerate any value. So, we got rid of these from reviews to prepare the data which can machine learn flawlessly.
- **Stop words:** Stop words are nothing but words such as *The*, *And* etc. which were filtered out before natural language processing (NLP) because these returns huge amount of unneeded information.
- **Stemming:** Stemming plays a vital role in NLP. Basically, much of NLP models is about sentiment of the text. Stemming means the process of transforming the words to base forms by removing inflection. For example, a stemming algorithm scale down words *Hiking*, *Hiked*, and *Hiker* to the root word, *Hike*
- **White spaces:** These were formed after performing pre-processing steps such as removal of numbers and special characters, stemming document to the corpus. These white spaces might be misclassified by the machine. Hence to avoid that the white spaces were removed from the data.

E. Feature Extraction

Feature Extraction is a type of dimensionality reduction technique which is very often used for deriving the new features from extracted data in order to boost classifier efficiency and greater classification accuracy. The efficient technique which is widely used in NLP is the, term frequency-inverse document frequency (TF-IDF) was implemented to extract the words with weighted average in the corpus. IDF which decreases the weight for frequently used words and increases the weight for less commonly used words in a collection of corpus documents. The term frequency combined with IDF to calculate how important the word is to a document in corpus. By using sparse terms function, words which have least sparse percentage were removed because machine couldnt learn from these terms. Frequent words in two data samples are shown in Figure 3

Fig. 3. frequent words of **Data-PNN** and **Data-PN**

F. Modelling

From the literature review, Random forest (RF), Support Vector Machine (SVM) and Naive Bayes algorithms are widely used for sentiment mining and classification. Radial kernel SVM is used because of the non-linear data and which helps in mapping the data into a higher dimensional space and makes easy to classify. Naive Bayes classifier consider the features of independent variables regardless of their correlation between them to classify the positive, negative and neutral reviews. Random Forest is the other algorithm which can be applied irrespective of input of data. It can be applied to complex and nonlinear data.

V. Experiments and Results

A. Selection of a classifier - **Data-PNN**

First we used the three algorithms (described in Section IV-F), to model the **Data-PNN**, which consists of multiclass dependent variable (positive, negative and neutral). The data is divided into training and testing in the ratio of 80:20. Here we performed k-fold cross-validation for choosing the best model. Generally, k=10 is chosen. As k value increases difference in the size of training dataset is decreases. And therefore bias gets reduced. In the k-fold implementation, during training the is data split into k-folds. Out of k-folds one fold is used to validate the model while training and remaining k–1 folds are used for training.

In Table I, we report the accuracies of the models when k=10. From Table I we can observe that Random Forest model outperforms the other models. We think this is due to the fact that, data is more complex and non-linear and random forest is good at modelling the same. So in the rest of this work we use random forest for sentiment classification.

TABLE I
ACCURACIES OF ALGORITHMS WHEN k=10

Model	Accuracy
Random Forest	61.36
SVM	58.38
Naive Bayes	57.3

B. Parameter tuning for Random Forest

Even though random forest performes well compared to other algorithms (as described in Section V-A), the validation accuracy is not upto the mark. So in order to improve the accuracy we are tuning the parameters using grid search method. Here also we trained the models on **Data-PNN**. In random forest $mtry$ and $ntree$ are the hyper parameters. Generally $mtry$ value is the square root of the number of attributes (=22). We also have choosen a $mtry$=14 because of less **OOB** error. Due to time constraints and computational power we have choosen $ntree$=50. The accuracies are reported in the Table II with these $mtry$ and $ntree$ values. From the Table II, we got better accuracy for the model in which $mtry$=22 and $ntree$=50. As different hyper parameters are given there is no change in the accuracy of the model.

TABLE II
PARAMETER TUNING FOR RANDOM FOREST MODEL

Model	mtry	ntree	Validation accuracy (%)
mtry22_tree50	22	50	62.14
mtry14_tree50	14	50	61.36

1) Evaluation of Random Forest model: Let us evaluate the performance of the models using testing data. The accuracies of the models on the testing data are given in Table III. Here also we got better accuracy for $mtry$=22 and $ntree$=50 (in Table III).

TABLE III
TEST ACCURACY USING THE MODELS TRAINED ON **Data-PNN**

Model	Test Accuray (%)
mtry22_tree50	69.48
mtry14_tree50	67.21

In addition to the accuracies we also report the confusion matrix and the F-score for each class (positive, negative and neutral). The confusion matrix for the $mtry22\_tree50$ on test data is shown in Figure 4. From the confusion matrix we can observe that positive (1) and negative (-1) class accuracies are 75.67% and 76.00%, respectively. Where as the classification accuray of neutral class is 56.45%, which is significantly lower than the other two classes. We can also observe that 28.57% of neutral examples were classified as negative class and 15.11% are classifies as positive class. So in total 44% of neutral examples are misclassified as other classes.

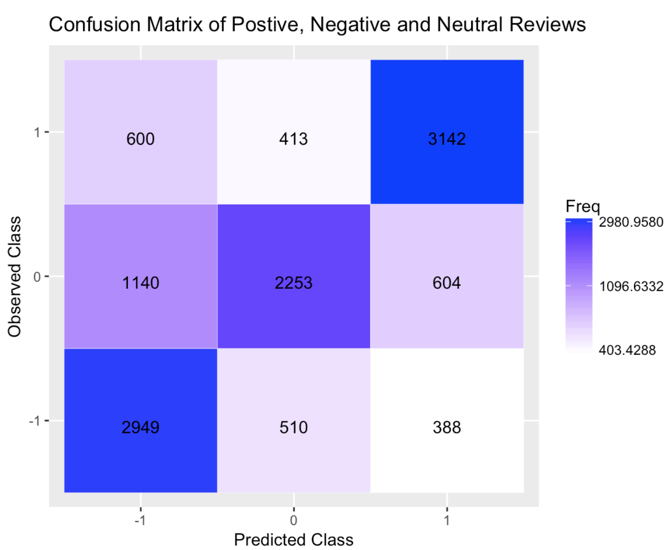

Fig. 4. Confusion matrix of test data

In this data we need to focus on both precision and recall. As F-score is combined metric of precision and recall, we are using F-score for evaluating the performance of model. F-score is more reliable measure for a classifier than accuracy. The F-score for each class is shown in Figure 5. From the Figure 5, we can observe that the F-score for neutral class is 0.688, which is less than the other classes (positive and negative).

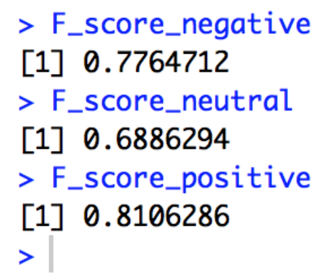

Fig. 5. F-Score for the three classes

C. *Random Forest on* **Data-PN**

Finally, from the confusion matrix (Figure 4) the average accuracy of three classes is 69.53%. Reduction of accuracy and less F-score is mainly due to the misclassification of neutral reviews. So classification of neutral reviews has strong influence on the performance of the model. Hence, the classification of neutral reviews is much harder than the other classes.

After observing the experimental results in Table III and Figure 5, we would like to do another experiment by filtering out the neutral examples (manually) and classify only positive and negative reviews. As described in Section IV-C, we created another dataset called as **Data-PN** by filtering out negative reviews. The **Data-PN** consists of only positive and negative reviews.

Here we took the best hyper parameters from the experiments described in Section V-B, and trained a random forest model on **Data-PN**, consists of positive and negative reviews. For this model we used k=10, $ntree$=50 and $mtry$=22.

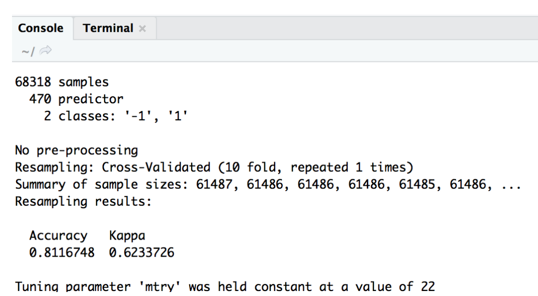

Fig. 6. Validation accuracy when the model is trained on **Data-PN**

The validation accuracy after training is shown on Figure 6. From the Figure 6, the validation accuracy of **Data-PN** is far greater than the validation accuracy of **Data-PNN** (Table II). This is due to the fact that the data consists of only positive and negative reviews and there is no uncertainty introduced by the neutral reviews.

1) Evaluation - **Data-PN***:* We evaluated the model on test data. The accuracy obtained on the test data is 83%, which is far better (20% relative) than the test accuracy reported in Table III. The confusion matrix and F-score are shown in Figure 7 and Figure 8, respectively. From Figure 7 and Figure 8 we can observe that very few number of reviews were misclassified and from F-score we can say that the model can clearly distinguish positive and negative reviews.

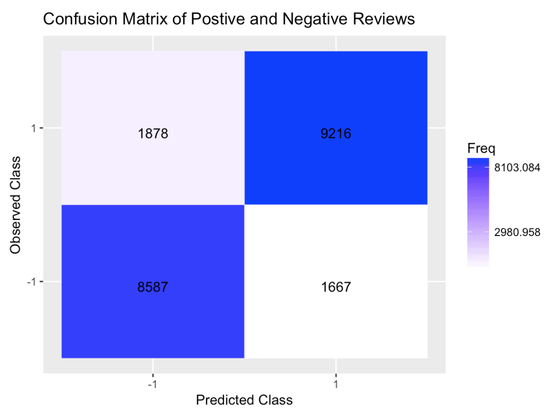

Fig. 7. Confusion matrix of test data **Data-PN**

```
> F_score_positive_negative
[1] 0.8550691
>
```

Fig. 8. F-score for two classes

Apart from confusion matrix on test data and F-score we also plotted the area under receiver operating characteristic (ROC) curve (AUC-ROC). The same is shown in Figure 10. The area under ROC curve indicates that model is good, from Figure 9. Therefore, AUC parameter distinguish positive class from negative class.

```
> auc(tes_neu$Score, pred_tfidf_neu)
[1] 0.8342472
```

Fig. 9. AUC

Finally, we can conclude that if we can some how detect and filtering out neutrality will definitely improve the classification accuracy.

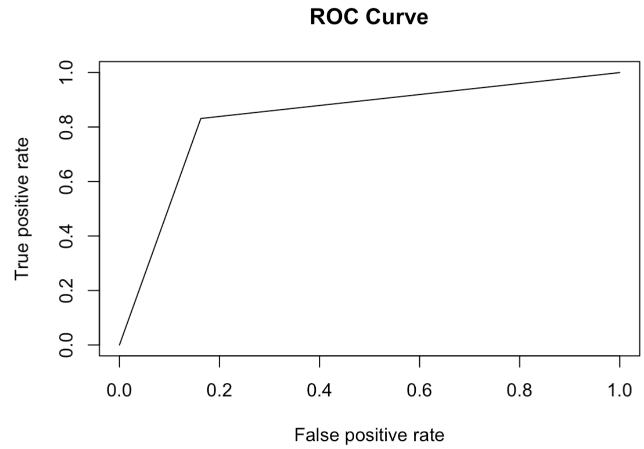

Fig. 10. AUC-ROC curve

VI. Drawback of Model

Though, Random forest is simple to understand it takes lot of time and need high computational power. In this project, it takes almost six hours for each model, even though it is run on 4 core processors in parallel. So, time and computational speed are the two constraints should consider before choosing the model.

VII. Conclusion and Future Work

In this study, it is clearly shown the effect of neutrality on sentiment analysis. The accuracy of **Data-PNN** is not improved even though we applied parameter tuning. This is because of higher misclassification of positive and negative reviews as neutral reviews. The higher difference of accuracies and F- measure of two sample sets indicates impact of neutrality on sentiment analysis. Therefore, Neutrality is the vital factor which cannot ignore. If we detect and filter neutral reviews correctly, then definitely we can see a greater improvement in classification accuracy which leads to better insights for amazon food industry.

Detecting neutral review is not simple and it cant done by single classifier. Some models classify positive reviews as neutral and some classify as neutral. So, there should be agreement among models in detecting neutral reviews. As per time constraints, we have experimented on the importance of neutral reviews and in future we are planning to build agreement model for detecting neutrality which further helps in building best model.

References

[1] P. Sasikala and L. M. I. Sheela, "Sentiment Analysis of Online Food Reviews using Customer Ratings," no.1 June, 2018.

[2] Trinity Hartman,“Bing Liu: The Science of Detecting Fake Reviews — content26 ” [Online]. Available: https://content26.com/blog/bing-liu-the-science-of-detecting-fake-reviews/. [Accessed: 01-Aug-2018].

[3] M. T. Martn-Valdivia, A. Montejo-Rez, A. Urea-Lpez, and M. R. Saleh,“Learning to classify neutral examples from positive and negative opinions,” J. Univers. Comput. Sci., 2012.

[4] S. Symeonidis, D. Effrosynidis, and A. Arampatzis, “A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis,” Expert Syst. Appl., vol. 110, pp. 298310, 2018.

[5] T. N. Fatyanosa and F. A. Bachtiar, “Classification method comparison on Indonesian social media sentiment analysis,” 2017 Int. Conf. Sustain. Inf. Eng. Technol., pp. 310315, 2017.

[6] M. Bordoloi, “Sentiment Analysis of Product using Machine Learning Technique: A Comparison among NB , SVM and MaxEnt Sentiment Analysis of Product using Machine Learning Technique: A Comparison among NB , SVM and MaxEnt,” no.3 July, 2018.

[7] A. Singh, A. Agarwal, and P. Dimri, “ScienceDirect ScienceDirect Comparative Study of Machine Learning Approaches for Amazon Reviews,” Procedia Comput. Sci., vol. 132, pp. 15521561, 2018.

[8] J. Fontanarava, G. Pasi, and M. Viviani, “Feature analysis for fake review detection through supervised classification,” Proc. - 2017 Int. Conf. Data Sci. Adv. Anal. DSAA 2017, vol. 2018Janua, pp. 658666, 2018.

[9] Y. Al Amrani, M. Lazaar, and K. E. El Kadirp, “Random forest and support vector machine based hybrid approach to sentiment analysis,” Procedia Comput. Sci., vol. 127, pp. 511520, 2018.

[10] Y. Wan and Q. Gao, “An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis,” Proc. - 15th IEEE Int. Conf. Data Min. Work. ICDMW 2015, pp. 13181325, 2016.

[11] S. Das and A. K. Kolya, “Sense GST: Text mining & sentiment analysis of GST tweets by Naive Bayes algorithm,” 2017 Third Int. Conf. Res. Comput. Intell. Commun. Networks, pp. 239244, 2017.

[12] M. S. Mubarok, Adiwijaya, and M. D. Aldhi, Aspect-based sentiment analysis to review products using Nave Bayes, vol. 020060, p. 020060, 2017.

[13] M. Li and G. Yulong, Research on Text Sentiment Analysis, Dianzikeji.Org, pp. 27312734, 2018.

[14] Z. Singla, S. Randhawa, and S. Jain, Sentiment Analysis of Customer Product Reviews Using Machine Learning, 2017 Int. Conf. Intell. Comput. Control, 2017.

[15] K. Lavanya and C. Deisy, Twitter Sentiment Analysis Using Multi-Class SVM, 2017 Int. Conf. Intell. Comput. Control, 2017.

[16] J. Ding, Y. Dong, T. Gao, Z. Zhang, and Y. Liu, Sentiment Analysis of Chinese Micro-Blog Based on Classification and Rich Features, Proc. - 13th Web Inf. Syst. Appl. Conf. WISA 2016 - conjunction with 1st Symp. Big Data Process. Anal. BDPA 2016 1st Work. Inf. Syst. Secur. ISS 2016, pp. 6166, 2017.

[17] R. Hegde and S. Seema, Aspect Based Feature Extraction and Sentiment Classification of Review Data sets using Incremental Machine learning Algorithm, pp. 36, 2017.

[18] A. P. Jain and V. D. Katkar, Sentiments analysis of Twitter data using data mining, 2015 Int. Conf. Inf. Process., pp. 807810, 2015.

[19] Y. E. Cakra and B. Distiawan Trisedya, Stock price prediction using linear regression based on sentiment analysis, ICACSIS 2015 - 2015 Int. Conf. Adv. Comput. Sci. Inf. Syst. Proc., pp. 147154, 2015.

[20] [Online]. Availble: https://www.kaggle.com/snap/amazon-fine-food-reviews