Predicting Star Rating and Category of products with Amazon Reviews using Natural Language Processing

Neharika Joshi
Applied Machine Learning (COMP-1804)
001200608

***Abstract*—For the coursework, we were given a real-world Amazon Reviews dataset to work with, and we were instructed to investigate how machine learning techniques can be utilised to predict the the star rating and product's category. The algorithms used for this task are Logistic Regression and BERT. The basic purpose of the analysis of review data is to identify the user's subjectivity and categorise their opinions into whether it's a review about video games or a musical instrument and how many stars (out of 5) may the user have given for the product. This is the final report after the completion of carrying out the necessary activities on a dataset chosen. *Key words* : Machine Learning, Classification, Regression, Supervised Learning, Artificial Intelligence, Natural Language Processing, Pattern Recognition, Neural Network.**

I. Introduction and related work

By concept, Machine Learning is a branch of computer science that arose from artificial intelligence research into pattern recognition and computational learning theory. It is the method of learning and developing algorithms that can learn and forecast data sets. Rather than following rigid fixed set of instructions, these processes develop a model from sample entries in effort to allow data-driven projections or judgments. This report is about the multi-classification problem from text data and predicting the star rating and the category of the product. Analyzing Amazon review data can provide useful customer information that can be used to improve products. Organizations must have a thorough awareness of client behaviour and demands when it comes to their products and services. There are numerous solutions related to this problem presented by others. There are numerous research papers on a topic similar to this. In [1], the amazon review texts were fed to five different classifiers to solve the multi-classification problems, the algorithm used were Naïve Bayes , Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression. The reviews were used to predict the star ratings. To sum up, Logisitic Regression attained the highest classification accuracy, on the other hand, Decision Tree attained the lowest average accuracy scores. [4] experimented on classifying the 21 individual product categories by using the amazon review textx. Mutual Information (MI) and Term Frequency – Inverse Document Frequency (TF-IDF), are feature filters that were used to determine feature scores. In order to determine the appropriate feature sizes for different models, multiple thresholds were tested when running the classifiers. They used Random Forest Classifier (RF), Decision Trees Classifier (DT), Linear Discriminant Analysis Classifier (LDA), Linear Supporting Vector Classifier (SVC), and Nearest Centroid Classifier for multi-class classification. To conclude, when the number of features are risen, the Random Forest Classifier and Linear Supporting Vector Classifier's performance was improved and they also prefer TFIDF more than MI because it was more equally distributed.

II. Ethical discussion

Many of the operations that humans used to do individually have been allocated to machines as a result of the emergence of Artificial Intelligence, particularly the field of Machine Learning. This raises questions about the ethical implications of the outcomes that machines send out into the world, which is becoming increasingly prevalent [6] .

Furthermore, billions of individuals use Amazon on a regular basis to meet their wants for purchasing products, and not everyone thinks or like the same things. For example, one person may give a 5 star rating to an item purchased on Amazon, while another person may not like the same item and only give it a 2 star rating. Because of the bias of the algorithms, many people have been affected explicitly or implicitly as a result of the growth of Machine Learning in the world. This particular task to design a machine learning model for Amazon Reviews might have some of it's own social and ethical implications.There are many occasions where reviewers' comments are used for data collecting without their knowledge due to the lack of a superior governing entity that may regulate people's behaviour.

III. Dataset preparation

Explanatory Data Analysis: It is essential to study the data first and what type each column are and what are their relationships with each other. The dataset contains 5 columns altogether containing the following information:

- *review_id* : ID field with unique values
- *text*: Reviews left by the customers regarding a particular product
- *verified*: Whether the given reviews are given after a verified purchase from Amazon
- *review_score*: Star rating given to a particular product that ranges from 1 to 5 stars.

- *product_category*: Actual category of the product the review is about. It's either Video games or it's Musical Instruments.

Findings:

1) **Star Ratings:** Overall, Figure 1 depicts that 64.9% reviews were positive which were rated 5 stars where as only 17.8% reviewers rated 4 stars for the items. Also, 4% were 2 stars, not further along 4/8% of people who gave negative reviews at 1 stars. Lastly, 8.6% reviewers thought that the products were moderate rating them 3 stars.

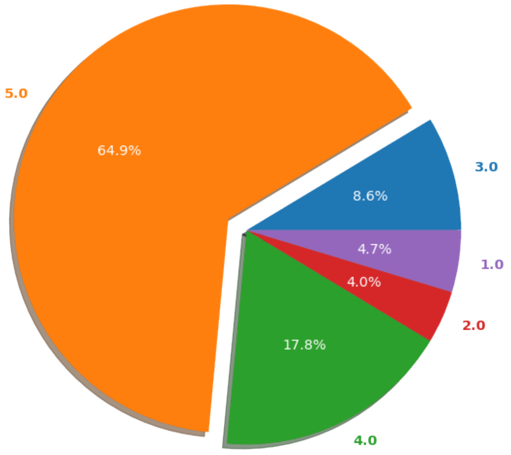

Figure 1: Pie Chart for Star Rating

2) **Product Category:** As depicted in Figure 2, there are 12658 rows of data for the Video games category and 8380 rows of data for Musical Instruments after the data is cleansed.

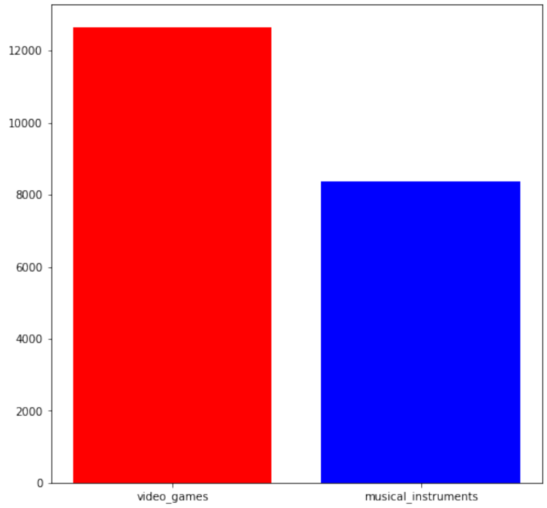

Figure 2: Bar Chart for Product Category

.

3) **Reviews Length:** For the histogram Figure 3, it can be observed that the distribution of words are roughly normal distributions.

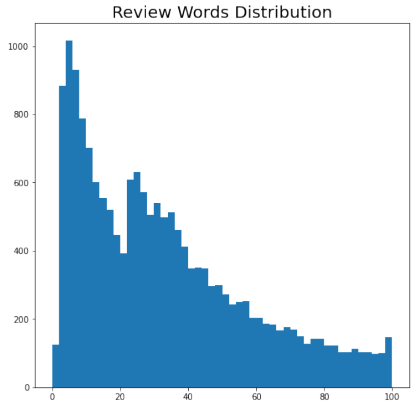

Figure 3: Histogram representing Distribution of length of Amazon Reviews

Data Preprocessing: The data in the given dataset is organised into 32918 rows and five columns. Each row is linked to a customer review as well as the product category. First, 907 rows of negative star ratings (-1) were discovered, which appeared to be an error and were apparently dropped. Also, there were roughly 2758 redundant rows in the 'text' column containing the same reviews, so they were eliminated. Furthermore, only verified reviews were selected for quality data, while unverified reviews(8692 rows) were ignored. Most of the time the unverified reviews meant that the customer actually didn't buy the item or may have received a discount in any other sites; instead, they shared their comments on Amazon, which landed the reviews in the unverified category, which according to Amazon's ranking algorithm isn't very relevant. Finally, 430 rows with null values were dropped from the data set. A total of 11,880 rows from the data set were removed after the data was completely cleaned. Furthermore, the data for the classes is imbalanced so it might be inefficient to directly feed it to the model hence, the data was undersampled for the prediction of rating stars.

Text Preprocessing: Before jumping into creating a model, it is essential for the text data to be processed properly rather than just feeding the data to train. To process the data, a text preprocessing library called 'NEATTEXT' was used. It makes text preprocessing easier as there are numerous built-in functions to remove punctuation, stopwords, HTML tags, special characters, emojis, emails, numbers, phone numbers, URLs and many more along with fixing contractions as well. Finally, the text data is lemmatized using 'nltk' library. ***Splitting Train and Test Set:*** The cleaned data set is chosen to split of 80% training data and 20% testing data because is was found to work the best in this rate.

IV. Methods

Logistic Regression: Logistic Regression unlike it's name is actually used for classification. It is a simple and efficient method of modelling the likelihood of an end result when an input variable is fed. The limit of logistic regression is constrained between 0 and 1 which is the major distinction between linear and logistic regression. Also, logistic regression doesn’t require a linear relationship between the input and the result variable, unlike linear regression.

$$LogisitcFunction(\sigma(z)) = \frac{1}{1+e^{-x}}$$

Instead of fitting a line, Logistic regression uses a ”S” shaped logistic function, the curve goes from 0 to 1. Logistic regression is categorized to further two components. First, if there is only one predictor variable in a model then it is called Simple Logistic Regression. Second, it there are multiple categories and predictor variables in a model, it’s called Multi-variable Logistic Regression (Nick Campbell, 2007).

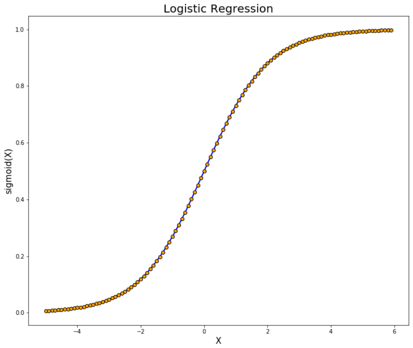

Figure 4: Logistic Regression ’S’ Curve

According to [1], Logistic Regression worked the best for predicting the star ratings using the text reviews. Experimentation on other classifiers were also done which is discussed more in Section V, but Logistic Regression stood out most giving highest accuracy among others.

BERT: Bidirectional Encoder Representation of Transformers(BERT) is a deep learning-based unsupervised language representation model. The use of the bidirectional training of Transformer, a popular attention model, to language modelling is BERT’s fundamental technical breakthrough. In contrast to earlier research, which looked at a text sequence from left to right or a combination of left-to-right and right-to-left training, this study looked at a text sequence from left to right.The utilization of bidirectional training of a Transformer to model languages is the a major breakthrough. In contrast to earlier researches, that only viewed text series either from left to right or right to left or the merging the two techniques by adopting a Cloze-inspired ”masked language model” (MLM) pre-training target [3]. Before the words are given to the model, about 15% of the words in the phrase are written over with mask tokens [MASK]. The model then attempts to anticipate what the masked tokens are by studying the non-tokenized words’ word sequences. Furthermore, BERT training approach also involves feeding the model a couple of sentences and guess what the next sentence from the native file is.During the training process, about 50% of the inputs are set followed by the second sentence from the original text, at the same time the other half of the 50% are just randomly selected sentence from collection. The random sentence will, it is assumed, be detached from the first sentence. It is assumed that the random sentence will be detached from the first sentence. Masked LM and Next Sentence Prediction are coupled for training the BERT model, with the aim of lowering the combined loss function of the two techniques [3]. To help classify the start and finish of each sentence, a [CLS] token is introduced at the beginning of the first sentence and a [SEP] token is introduced at the end.

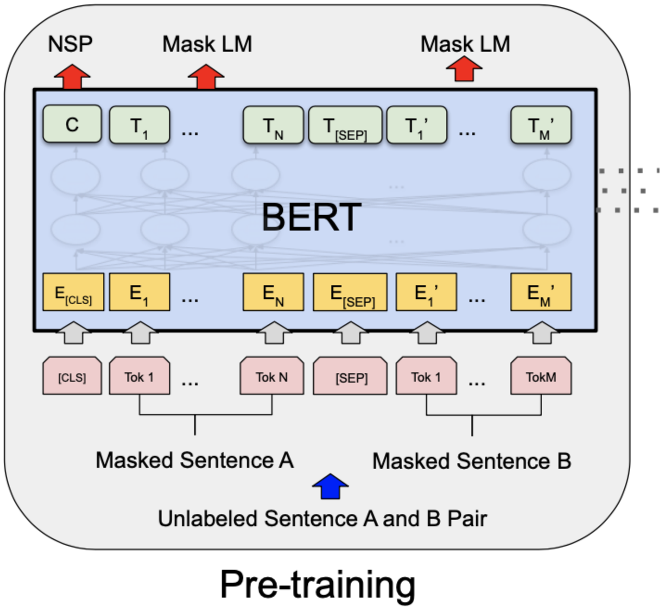

Figure 5: BERT Model pre-training method (Source [3])

Also, Fine Tuning has been employed for BERT to produce cutting-edge outcomes on a range of difficult natural language challenges. Most hyper-parameters in fine-tuning training are the same as in BERT training and comparatively fine tuning is cheaper than pre-training the model. “We just insert the job-specific inputs and outputs into BERT and fine-tune all of the parameters end-to-end for each task” [3].

V. Experiments and Evaluation

Logistic Regression: The metrics used for this method are Confusion matrix and Classification Score.

1) **Confusion Matrix:** A matrix that describes how well a classifier performed on a variety of experimental data for which the true information was available, known as Confusion Matrix.

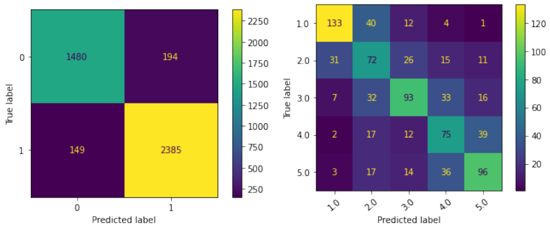

Figure 6: Confusion Matrix of Logistic Regression on Product Categories and Star Rating.

Figure 6 shows that the model worked well when it came to predicting categories, but not so well when it came to predicting star ratings.

2) **Accuracy Score:** One of the metrics for examining the performance of classification model is Accuracy Score which basically is the percentage of correct predictions made by the model. It is a very simple yet efficient method to show how well the model works, that's why it was chosen.

$$Accuracy = \frac{NumberofCorrectPredictions}{TotalNumberofPredictions}$$

In addition, the accuracy of binary classification can be also assessed in terms of positive and negative.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The accuracy for training the model for binary and multi-class classification was exceptional with 97% accuracy score. On the other hand, the accuracy score for testing set was 92% for the product category and only 56% for the star rating.

3) **ROC curve for binary classification:** A graph that shows how well a classification model performs across all classification levels is called ROC (Receiver Operating Characteristic) curve. It plots two parameters against each other, TPR (True Positive) and FPR (False Positive Rate).

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

The ROC AUC Score obtained for training set is 0.993 ROC AUC Score obtained for testing set is 0.976.

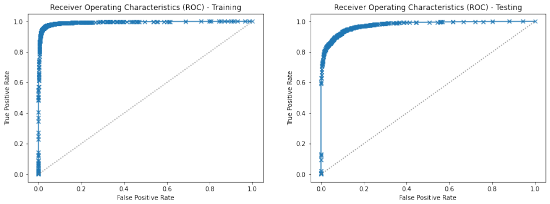

Figure 7: ROC curve for binary classification of Product Categories

For hyper tuning parameter, Grid Search is used where C = 1 was added for parameter along with 500 max_features and saga as the solver.

BERT: Two models were created for binary and multi-class classification. For the binary classification of product categories, BERT Base model is used which is a basic deep learning model but yet efficient. Other models like DistilBERT and Roberta were implemented but didn't quite top the accuracy score like BERT Base did. On the contrary, RoBERTa b5 seems to have done well while classifying the multi-class ratings compared to BERT base. For Hyper-parameter tuning, number of training epochs, batch size, learning rate, warm ups and maximum sequence length were adjusted as suggested by the Hugging Face Transformer.

1) **Accuracy Score:** It is the efficient method to test how the model is performing. For the category classification, accuracy score for the BERT base model was 95% whereas, for the accuracy rating classification was 61% with RoBERTa model.

2) **Classification Report:**A report to assess the accuracy of predictions made by a classification algorithm is knows as Classification Report. It shows the f1-score, precision, accuracy and recall score of all the present classes.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall}$$

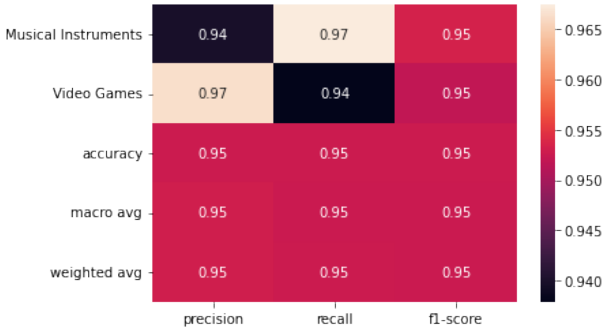

Figure 8: Classification Report for Binary Classification of Product Categories with BERT Base model showing good scores for both classes.

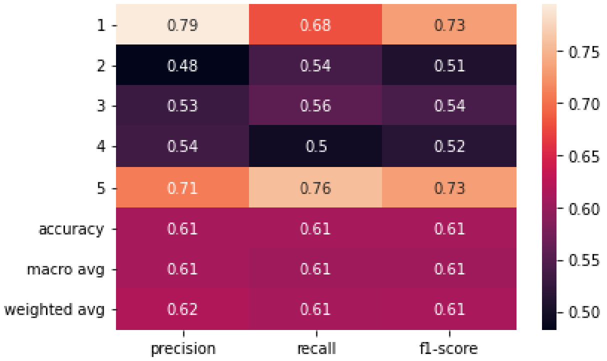

Figure 9: Classification Report for Multi-classification of Star Ratings with RoBERTa model

From the classification report Figure 9, it was found that the model makes the most accurate predictions for 5 stars and 1 stars, followed by 3 stars, 4 stars and at last 2 stars according to the f1-scores. Also, 1 stars has the highest precision score followed by 5 stars but vice-versa in terms of recall score.

Resampling: As seen in Figure 1, the classes for rating labels are unbalanced which can hamper the model performance. So, before training the model, the data was balanced by using Undersampling method where each classes were balanced as 837 (lowest data for a 2 stars class). Let's look at the classification report with imbalanced data is fed to the Logistic Regression Model. Furthermore, for binary classification oversampling was done to balance out the two classes.

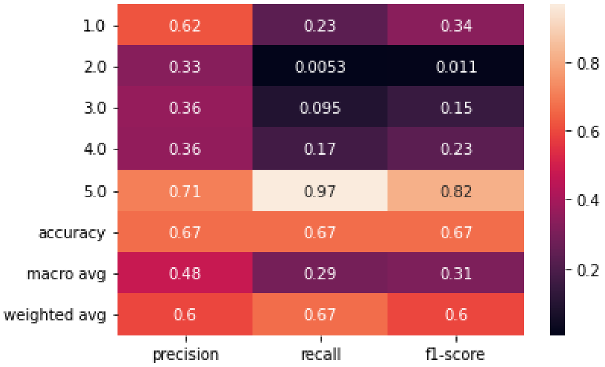

Figure 10: Classification Report for imbalanced data for Star Ratings with Logistic Regression

Figure 10 shows that the f1-score and recall score for a two-star rating are both extremely horrendous. However, the 5 star class has higher scores than the others because it contains the majority of the data. The accuracy of uneven data, on the other hand, is higher than that of balanced data. In terms of metrics, the accuracy score is commonly employed to assess the model's performance, but it is unreliable when dealing with unbalanced data because it is skewed towards classes with the most data.

Comparison of Algorithms:

TABLE I
COMPARISON OF SEVEN DIFFERENT ALGORITHMS FOR BINARY CLASSIFICATION OF PRODUCT CATEGORY

	Accuracy	Recall	Precision	F1-Score
NB	88%	99%	84%	90%
SGD	92%	94%	93%	93%
KNN	55%	62%	62%	62%
RF	60%	100%	60%	75%
SVC	92%	94%	93%	93%
LR	92%	94%	93%	92%
BERT Base	98%	97%	98%	98%

VI. Discussion and Future Work

Overall,after the evaluation by feeding data to numerous models it is essential to choose the best one to work with as data can vary in every dataset. Nevertheless, there are still a lot of issues while classifying the texts as of now. The models needs to be trained on extensive level from which takes a lot of time while working with huge varieties of dataset. Sometimes , there may be some sentences that are relevant to both categories in some reviews that we need to anticipate, such as 'I enjoy playing with this product,' but that doesn't really tell which category the review is about, does it? Furthermore, the words in the comments may contain subjective insight that is unlikely to provide any valuable information, making the prediction task even more difficult. Looking far ahead, it would be better if the model would be optimized further with more research. One of the main goals is to use the reviews written in other languages like French, Italian, Spanish and Japanese etc. by using multilingual deep learning model 'BERT - MULTILINGUAL' to examine if it works well or not like it did for English.

VII. Conclusions

To sum up, binary classification worked extremely well, with a percentage of more than 90% for both methodologies. As previously stated, the main challenge in developing a model was the imbalance of data for the rating classes, however undersampling helped. The hyper-parameter tuning that was done for both techniques which undoubtedly improved the model's performance. Also, it is clear that the classification model's performance solely depends on the data and there isn't a model that magically works for every data.

A. Abbreviations and Acronyms

- MI - Mutual Information
- TF-IDF - Term Frequency – Inverse Document Frequency
- RF - Random Forest Classifier
- DT - Decision Trees Classifier
- SVC - Linear Supporting Vector Classifier
- LDA - Linear Discriminant Analysis Classifier
- LR - Logisitc Regression
- MLM - Masked Language Model
- BERT - Bidirectional Encoder Representation of Transformers
- TP - True Positive
- TN - True Negative
- FP - False Positive
- FN - False Negative
- TPR - True Positive Rate
- FPR - False Positive Rate
- RoBERTa - Robustly Optimized BERT Pretraining Approach
- NB - Naive Bayes
- SGD - Stochastic Gradient Descent
- KNN - k-Nearest Neighbours

References

[1] Pranckevičius, T. & Marcinkevičius, V., 2017. "Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification". Baltic Journal of Modern Computing, 5(2), p. 227.

[2] Nick, T. G. Campbell, K. M., 2007. Topics in Biostatistics. New Jersey: Humana Press.

[3] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," CoRR, vol. abs/1810.04805, 2018.

[4] Y. Liu, T. Hu and H. H, "Multiclass Classifier Building with Amazon Data to Classify Customer Reviews into Product Categories".

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” arXiv, Seattle, 2019.

[6] T. Wilson, “Instagram, Amazon, and Machine Learning: Ethical Implications of Collecting and Analyzing Commercial User Data,” Charlottesville, Virginia, 2020.