**Predicting Rating Levels Associated with User Reviews from Amazon Reviews**

By Hogan Lee

Kaggle User ID: 9707656

1. **Overview of the Process**

Text

Description automatically generatedThe flow chart below presents the processes to predict rating levels associated with user reviews from Amazon reviews. For the prediction through machine learning, a total of 518,430 labeled review were loaded from the Kaggle site and cleaned using the Natural Language Toolkit and regular expression modules. The cleaned data were vectorized using TF-IDF method and vectorized data were used to train and validate a total of 8 machine learning models: Multinomial Naïve Bayes, Extreme Gradient Boosting, Support Vector machine, Decision Tree, Random Forest, Multi-layer Perceptron, K-nearest neighbors, and Long short-term memory. The performance of the trained models was evaluated by four metrics (accuracy, precision, recall and F1), and the best performing model was selected to predict the rating level of the test data which included 5,000 reviews with no rating. Additional data analyses were conducted to describe training data and summarize the results. ***RapidMiner and Python*** were used for the analyses.

1. **Brief Description of Data**

|  |  |  |
| --- | --- | --- |
| Reviews Scored Low | Reviews Scored Medium | Reviews Scored High |
| A picture containing text  Description automatically generated | A picture containing text, newspaper  Description automatically generated | Text  Description automatically generated |

The training data included 518,430 reviews on foods which were purchased in Amazon. In addition to review text, the data included reviewID, productid, userId, profileName, helpfulness, score\_level, time and summary. The analyses used only information in review text and score level. Among 518,430 reviews, *78% were scored for ‘high’, 14% for ‘low’ and 8% for ‘medium’*. *The mean length of strings* in cleaned review text was 254.8 (±266) for ‘high’, 292 (±309) for ‘low’, and 314.8 (±320) for ‘medium’. Word could per score level was shown above. Words used in reviews scored as medium and low seemed similar each other while high scored reviews included more positive words like highly recommended, great, enjoy, love, much better, so on.

1. **Data Cleaning**

Before training, we used the Python module of Natural Language Toolkit (NLTK) and regular expression (RE) for data cleaning to remove emojis, links, and punctuations in the text of the reviews. The “stop” words, which do not convey any meaning (e.g., 'is', 'are', 'and,' etc.) were excluded and lemmatization processes were lemmatized to identify the base form of a word, called Lemma (e.g., 'go' from 'went') based on the dictionary. For this processing, all the reviews were tokenized before applying all the methods for data cleaning. These were then detokenized. The data were randomly split into two sets for training (90%) and validation (10%).

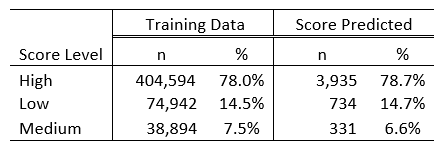
1. **Models and Model Evaluation**

Table below represents the evaluation metrics by the models trained. Overall, the performance of machine learning models was fairly good in predicting score level. While the performance of multi-Naïve Bayes and support vector machine were relatively lower, the better performance was shown in Extreme Gradient Boosting, Decision Tree, Random Forest, Multi-layer Perceptron, K-nearest neighbors, and Long short-term memory. *Multi-layer perceptron* was chosen for predicting score level in test data

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | MNB | XGBoost | SVM | Dtree | RF | MLP | KNN | LSTM |
| Accuracy | 0.79 | 0.86 | 0.78 | 0.86 | 0.88 | 0.90 | 0.86 | 0.86 |
| Precision | 0.82 | 0.85 | 0.61 | 0.85 | 0.9 | 0.90 | 0.86 | 0.87 |
| Recall | 0.79 | 0.85 | 0.78 | 0.86 | 0.88 | 0.90 | 0.86 | 0.86 |
| F1 | 0.71 | 0.83 | 0.68 | 0.85 | 0.86 | 0.90 | 0.84 | 0.86 |

MNB: Multinomial Naïve Bayes; XGBoost = Extreme Gradient Boosting; SVM: Support Vector machine; Dtree: Decision Tree; RF: Random Forest; MLP= Multi-layer Perceptron classifier; KNN: K-nearest neighbors; LSTM= Long short-term memory

1. **Brief Description of Predicted Score Level**

Table shows the frequency of score level between training data and predicted test data. There was no significant difference in score level between two data sets.

Table

Description automatically generated

Vader and TextBlot sentiment scores were the highest when MLP model predicted score level to be high while reviews predicted to be ‘low’ had the lowest sentiment scores, implying that reviews predicted to be high included more positive sentiment than the others, as we expected. In sum, these findings suggested that prediction using MLP may be performed reasonably and well.