Classifying Sentiments on the Amazon Product Reviews Dataset

Aluno: **Felipe Lemos** — Email: **felipebpl@al.insper.edu.br**

I. Section 1 - Dataset

We utilize the *Consumer Reviews of Amazon Products* dataset, which contains over 34,000 Amazon product reviews with textual feedback, star ratings, and product information. The dataset is available at Kaggle.

The dataset's business purpose, as detailed on Kaggle, is to analyze Amazon's product launches, extract insights from consumer reviews, and support machine learning models. Businesses can identify the most reviewed products, correlate reviews with less obvious categories for data-driven decisions, and understand customer sentiments to enhance products and services.

II. Section 2 - Classification Pipeline

The classification pipeline predicts review sentiments as positive or negative. Preprocessing involved:

- **Data Cleaning:** Removed duplicates and missing values in texts and ratings.
- **Text Processing:** Converted text to lowercase, removed non-alphabetic characters, stop words, and lemmatized words using NLTK's `WordNetLemmatizer`.
- **Feature Engineering:** Combined cleaned titles and texts into `combined_text`. One-hot encoded categorical variables.

For feature extraction, we used the Bag-of-Words model with binary term frequencies via `CountVectorizer`. This approach captures the presence of words indicative of sentiment, like *excellent* or *terrible*.

However, relying solely on word presence can mislead, as negations (e.g., "not good") and sarcasm (e.g., "great, it broke") invert sentiments and confuse the classifier. Nonetheless, the Bag-of-Words model remains effective for general sentiment classification.

III. Section 3 - Evaluation

The classifiers were evaluated using Repeated Stratified K-Fold cross-validation with 5 splits and 10 repeats. Due to class imbalance, the *balanced_accuracy_score* metric was used.

The mean balanced accuracy scores were: Logistic Regression (0.87), XGBoost (0.87), Linear SVC (0.83), Random Forest (0.70), Gradient Boosting (0.69), and Bernoulli Naive Bayes (0.66).

Logistic Regression and XGBoost achieved the highest scores. The confusion matrices for Logistic Regression (LR) and XGBoost (XGB) are summarized in the table below:

Comparing the confusion matrices, Logistic Regression correctly classified more positive reviews, while XGBoost better identified negative reviews.

Analyzing the most influential words in Logistic Regression revealed that words like *five*, *excellent*, *love*, and *great* contributed to predicting positive sentiment. Conversely, words such as *worst*, *dead*, *terrible*, and *disappointed* were strong indicators of negative sentiment. These findings confirm the model's ability to identify sentiment-laden terms relevant to the problem context.

However, reliance on word presence may misclassify reviews containing negations or sarcasm, highlighting limitations of the Bag-of-Words model in capturing nuanced language.

IV. Section 4 - Dataset Size

To assess the impact of dataset size on model performance, learning curves were generated by evaluating training and validation scores at varying training set sizes. As the dataset size increased, the training accuracy slightly decreased from 99.4% to 97.1%, while the validation accuracy increased from 80.5% to 86.4%.

This trend shows that the model benefits from more data, enhancing its ability to generalize. However, validation accuracy gains plateau around 86%, indicating limited potential for further improvement by expanding the dataset. Given resource constraints and the business context, acquiring additional data for minimal performance gains may not be feasible.

V. Section 5 - Topic Analysis

Topic modeling was conducted using Latent Dirichlet Allocation (LDA) with 5 topics:

- **Topic 1**: battery, light, life, use, performance
- **Topic 2**: tablet, price, screen, use, good
- **Topic 3**: battery, price, brand, buy, good
- **Topic 4**: tablet, kid, love, game, bought
- **Topic 5**: kindle, love, gift, book, use

Classification performance varied across topics, with balanced accuracy scores ranging from 0.61 to 0.90. Specifically, Topic 3 achieved the highest accuracy (0.8968), while Topics 4 and 5 had the lowest (0.6085 and 0.6125, respectively).

A two-layer classifier was implemented, first classifying reviews into topics and then applying topic-specific classifiers. This approach resulted in an overall balanced accuracy of 0.9824, outperforming the single-layer classifier's 0.87. This indicates that topic-based specialization enhances classification performance by tailoring models to specific contexts.