Classifying sentiments in the Amazon Product Reviews dataset

Arthur Gomes Chieppe
Computer Engineering
Insper
São Paulo, Brazil
arthurgc1@al.insper.edu.br

I. Dataset

The chosen dataset for this natural language processing project is the Amazon product review balanced dataset [1], which consists of two essential columns: text and binary sentiment. Each entry is labeled as either positive or negative, providing a clear foundation for sentiment classification. This project addresses the challenge outlined in the paper "A Statistical Approach to Star Rating Classification of Sentiment" [2] recognizing that valuable product opinions often exist solely in text format in comment sections and online forums. In the absence of traditional star ratings or thumbs up/down systems, this research aims to develop advanced sentiment classification methodologies to extract and interpret sentiments from unclassified textual data.

II. Classification pipeline and pre processing

The classification pipeline utilizes a lemmatizer alongside a vectorization method that converts text into a binary format, focusing on the presence or absence of words rather than their frequency. Accents are stripped to ensure uniformity, and stop words are lemmatized for consistent removal of irrelevant terms. A minimum document frequency threshold was applied to exclude rare words that may not contribute meaningfully to the classification, reducing the vocabulary from 10,808 to 9,785. While tests with a stemmer further reduced the vocabulary to 7,745, the lemmatizer was preferred for its superior interpretability.

The bag-of-words approach with binary encoding is appropriate here, as the presence of words with strong positive or negative connotations tends to indicate sentiment more effectively than product descriptions. However, a key limitation of this method is that it does not account for word order, potentially misinterpreting phrases with negations, such as "not perfect." Unlike TF-IDF [3], which could assign undue weight to infrequent or unique words, this approach prioritizes more consistent terms that capture the overall sentiment.

The classification model used in this pipeline is Logistic Regression [4], implemented with scikit-learn's default settings.

III. Model Evaluation

The model was evaluated after being trained 100 times with different train-test splits, consistently using an 80/20 ratio, resulting in a test accuracy of 0.89. The analysis of the most relevant words revealed distinct patterns for both positive and negative reviews. For negative reviews, the top words—"overrated," "disappointing," "yuck," "yikes," and "boo"—are primarily expressions of dissatisfaction and human disgust. In contrast, positive reviews featured terms such as "addictive," "nutritionist," "hooked," "yum," and interestingly, "skeptical."

The predominance of food-related words, particularly in both the positive and negative sets, suggests that a significant portion of the dataset pertains to food products. Additionally, the appearance of "skeptical" as a positive word may indicate that some consumers initially doubted the product but were ultimately satisfied, reflecting a shift in perception.

IV. Dataset Size

The dataset size plays a crucial role in determining the potential for improving model accuracy. By examining the training curves in Figure 1, it is evident that a 5% gap exists between the error (1 - accuracy) of the training and test sets, indicating that the model has not fully generalized and could benefit from a larger dataset.

A higher training-to-test accuracy difference suggests that increasing the dataset size, and consequently the training data, would likely improve test accuracy. Given that product reviews are readily available online, expanding the dataset is feasible and could yield better results. Additionally, the model may also benefit from a more diverse range of product reviews, as the top word analysis suggests a significant portion of the dataset pertains to food products, which could limit the generalization to other product categories.

V. Topic Analysis

Topic analysis of the dataset revealed four main topics: general food products, cat food, coffee, and snacks. Interestingly, despite having fewer than 5,000 related documents, the cat food topic still emerged as one of the most significant, whereas the other topics each had approximately 50,000 samples. By employing a two-layer classification approach—first classifying by topic and then by sentiment with a topic-specific classifier—it was observed that the accuracy remained consistent with the original model. All topic-specific classifiers achieved accuracies around 89% ± 2%, demonstrating robust performance across different categories.

VI. Appendix

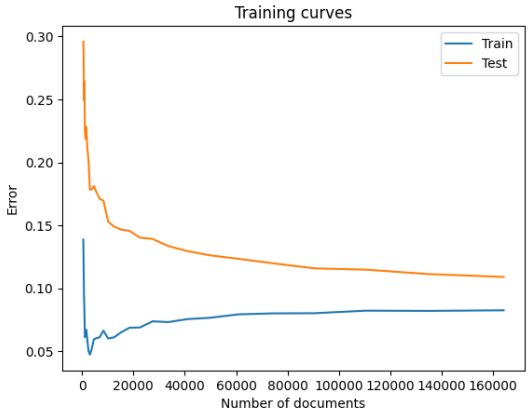

Fig. 1. Relationship between dataset size and error rate

References

[1] S. Iranga, "Amazon reviews balanced dataset", Kaggle. Available at: https://www.kaggle.com/datasets/sashmindairanga/amazon-reviews-balanced-dataset [Accessed: 2024-10-03].

[2] A. Hoogenboom, F. Boon, and F. Frasincar, "A statistical approach to star rating Classification of Sentiment", 2012.

[3] A. Simha, "Understanding TF-IDF for Machine Learning" Available at: https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/ [Accessed: 2024-10-03].

[4] scikit-learn, "LogisticRegression," Available at: https://scikit-learn.org/1.5/modules/generated/sklearn.linear_model.LogisticRegression.html [Accessed: 2024-10-03].