**Sentiment Analysis on Amazon Product Reviews**

**Introduction**

This project aims to analyze the sentiment of over 34,000 Amazon product reviews. The dataset consists of product names, brands, categories, review dates, review texts, review titles, and sentiment labels. The sentiment labels are classified into three categories: Positive, Neutral, and Negative. This analysis involves handling class imbalance, implementing various classifiers, comparing their performance, exploring topic modeling techniques, and evaluating model performance using different metrics.

**Data Preprocessing**

1. **Loading the Dataset**:
   * The dataset was loaded from a CSV file using Pandas.
   * Basic information and the first few rows of the dataset were displayed to understand its structure.
2. **Dataset Overview**:
   * The dataset contains 8 columns: Name of the product, Product Brand, categories, primaryCategories, reviews.date, reviews.text, reviews.title, and sentiment.
   * There are 4000 entries in the dataset.
   * The sentiment distribution is highly imbalanced: 3749 Positive, 158 Neutral, and 93 Negative.
3. **Visualizing Sentiment Distribution**:
   * A count plot was generated to visualize the distribution of sentiment categories.

**Feature Extraction**

1. **Text Vectorization**:
   * The review texts were converted into numeric features using TF-IDF (Term Frequency-Inverse Document Frequency) vectorization with a maximum of 5000 features.
2. **Handling Class Imbalance**:
   * SMOTE (Synthetic Minority Over-sampling Technique) was applied to balance the sentiment classes.

**Model Training and Evaluation**

1. **Multinomial Naive Bayes**:

* The Multinomial Naive Bayes model was trained on the balanced dataset.
* **Accuracy: 96.27%**
* Classification Report:

precision recall f1-score support

Negative 0.96 1.00 0.98 748

Neutral 0.93 1.00 0.97 733

Positive 1.00 0.89 0.94 769

accuracy 0.96 2250

macro avg 0.96 0.96 0.96 2250

**weighted avg - 0.96 0.96 0.96 2250**

**- AUC-ROC Score: 0.9987**

1. **Support Vector Machine (SVM)**:
   * The SVM model with a linear kernel was trained.
   * **Accuracy: 99.11%**
   * Classification Report:

precision recall f1-score support

Negative 1.00 1.00 1.00 748

Neutral 0.98 1.00 0.99 733

Positive 1.00 0.97 0.99 769

accuracy 0.99 2250

macro avg 0.99 0.99 0.99 2250

**weighted avg - 0.99 0.99 0.99 2250**

1. **Feedforward Neural Network**:
   * A neural network with dense layers and dropout was trained.
   * **Accuracy: 99.69%**
   * Training and validation accuracy showed excellent performance with early stopping.
2. **LSTM Neural Network**:
   * An LSTM (Long Short-Term Memory) network was trained.
   * **Accuracy: 99.95%**
3. **Ensemble Models**:
   * Multinomial Naive Bayes and XGBoost classifiers were trained on the balanced dataset.
   * **Accuracy for Naive Bayes: 96.27%**
   * **Accuracy for XGBoost: 97.64%**
   * An ensemble model using VotingClassifier was created.
   * **Ensemble Model Accuracy: 98.80%**
4. **Additional Feature Engineering**:
   * A sentiment\_score feature was engineered, and its sparse matrix was combined with the TF-IDF features.
   * The combined features improved the SVM model's accuracy.

**Hyperparameter Tuning**

1. **Grid Search for SVM**:
   * Best Parameters: {'C': 1, 'kernel': 'linear'}
   * **Best Accuracy: 99.80%**
2. **Random Search for XGBoost**:
   * Best Parameters: {'n\_estimators': 150, 'learning\_rate': 0.1, 'max\_depth': 5}
   * **Best Accuracy: 98.75%**

**Topic Modeling**

1. **Latent Dirichlet Allocation (LDA)**:
   * Extracted 10 topics from the review texts.
   * Displayed top words for each topic.
2. **Non-Negative Matrix Factorization (NMF)**:
   * Extracted 10 topics from the review texts.
   * Displayed top words for each topic.

**Conclusion**

The sentiment analysis on Amazon product reviews was successfully conducted. Various models were implemented, with the LSTM network achieving the highest accuracy. Feature engineering and balancing techniques significantly improved model performance. Additionally, topic modeling provided insights into common themes within the reviews.