# Automated processing of customer feedback for a retail company

By Jose and Cindy

Project Overview

Given a dataset containing details about products and their reviews, train a model to help classify those reviews into: positive, neutral or negative. Also, using Generative AI, summarize reviews broken down into review score (0-5), and broken down into product categories
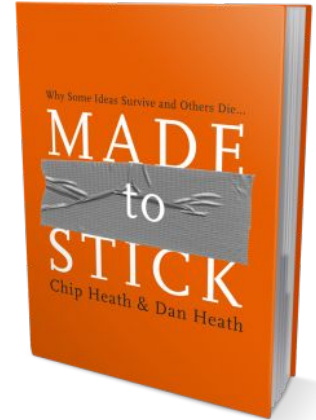
Approach

➜ **Analyze data**
check for particularities and quality of data.

➜ **Preprocessing**
Prepare data to be fed to models.

➜ **Experiment with a fan of models and gather metrics.**
Naive Bayes,Logistic Regression, Linear SVM, Random Forest, LSTM, Transformer Models: Bert

Dataset

Total samples: ~67,992 reviews

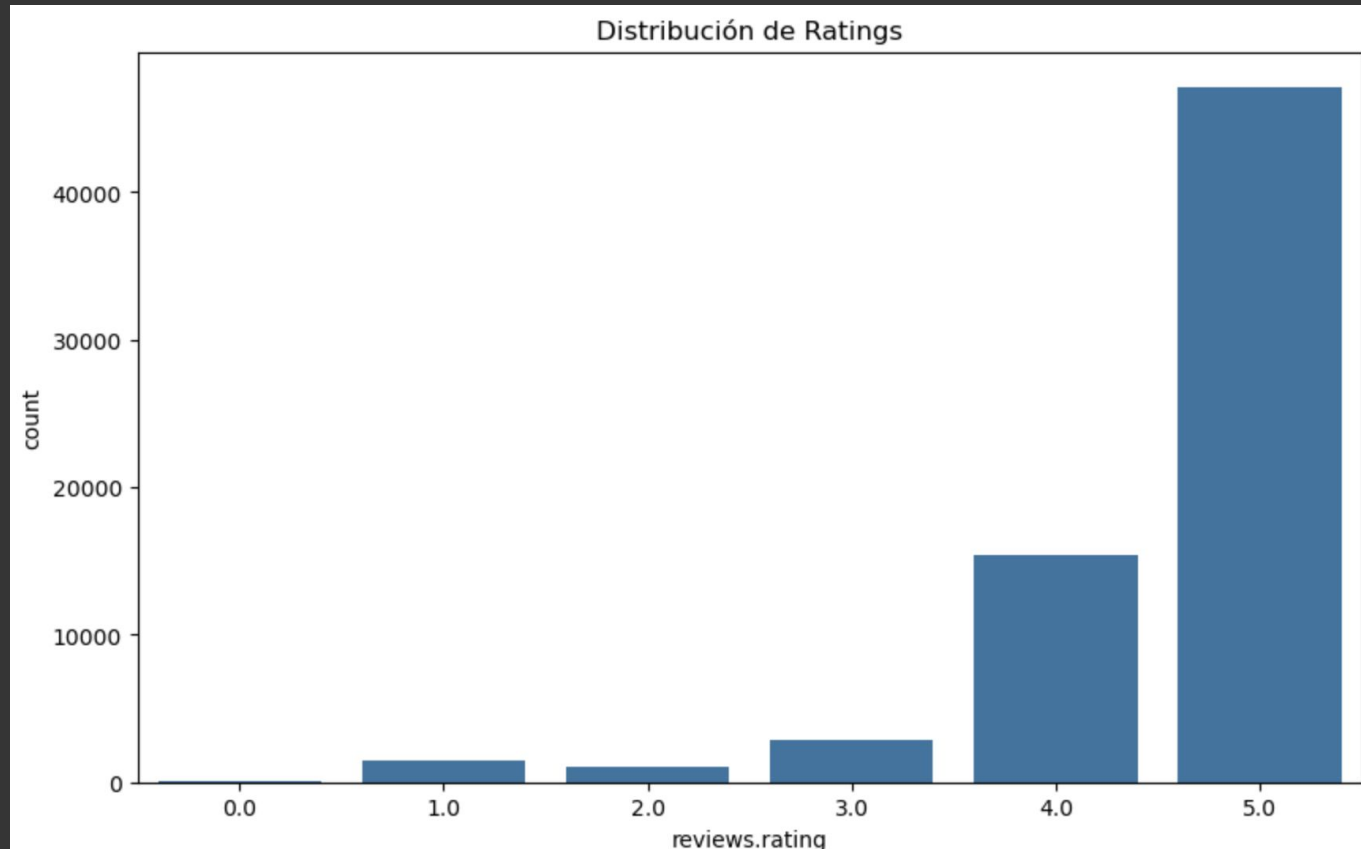Class distribution: Imbalance: too many positives.

Data Insights:

Extracted features: Reviews trend over time, reviews count, average reviews score per category, products or categories with more reviews, sentiment vs review length, most used words per sentiment category
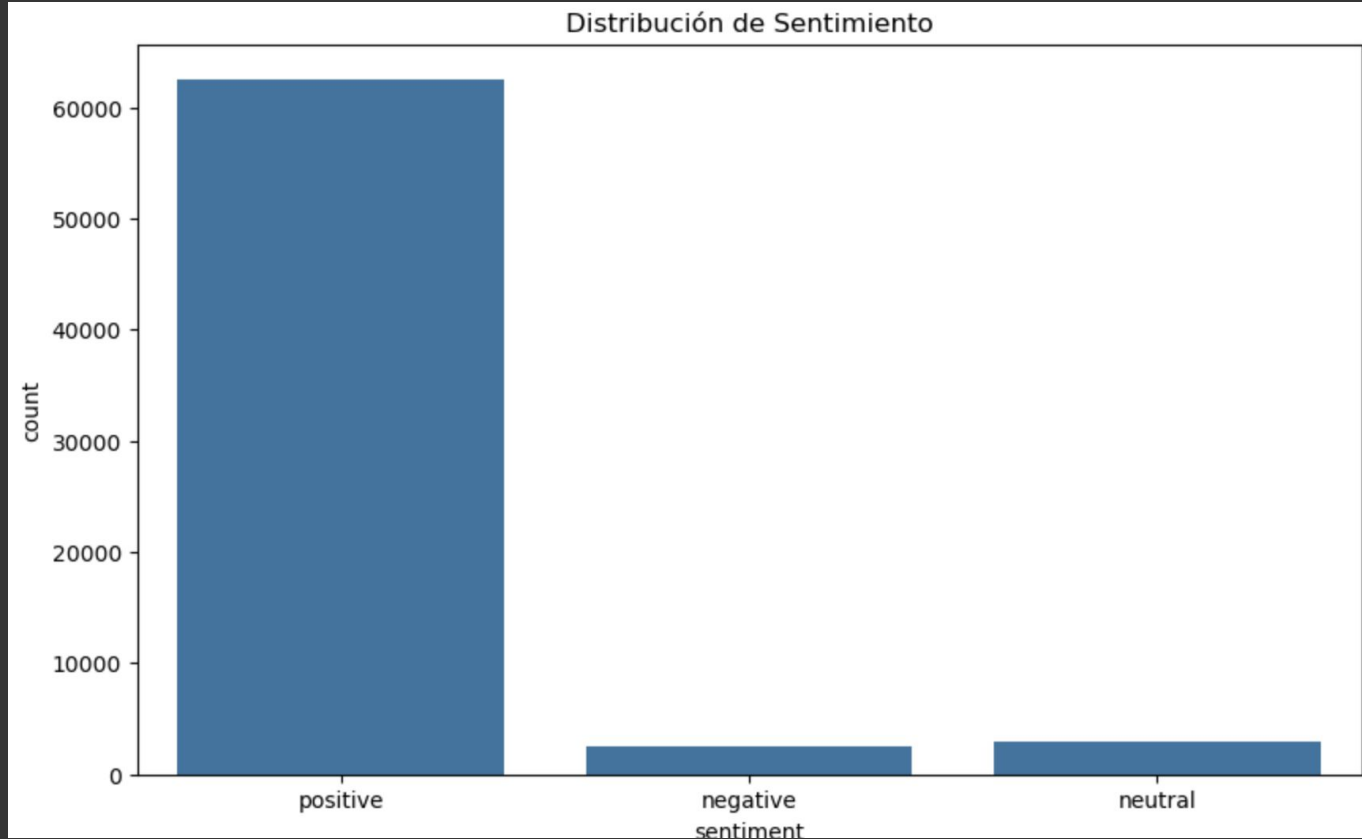
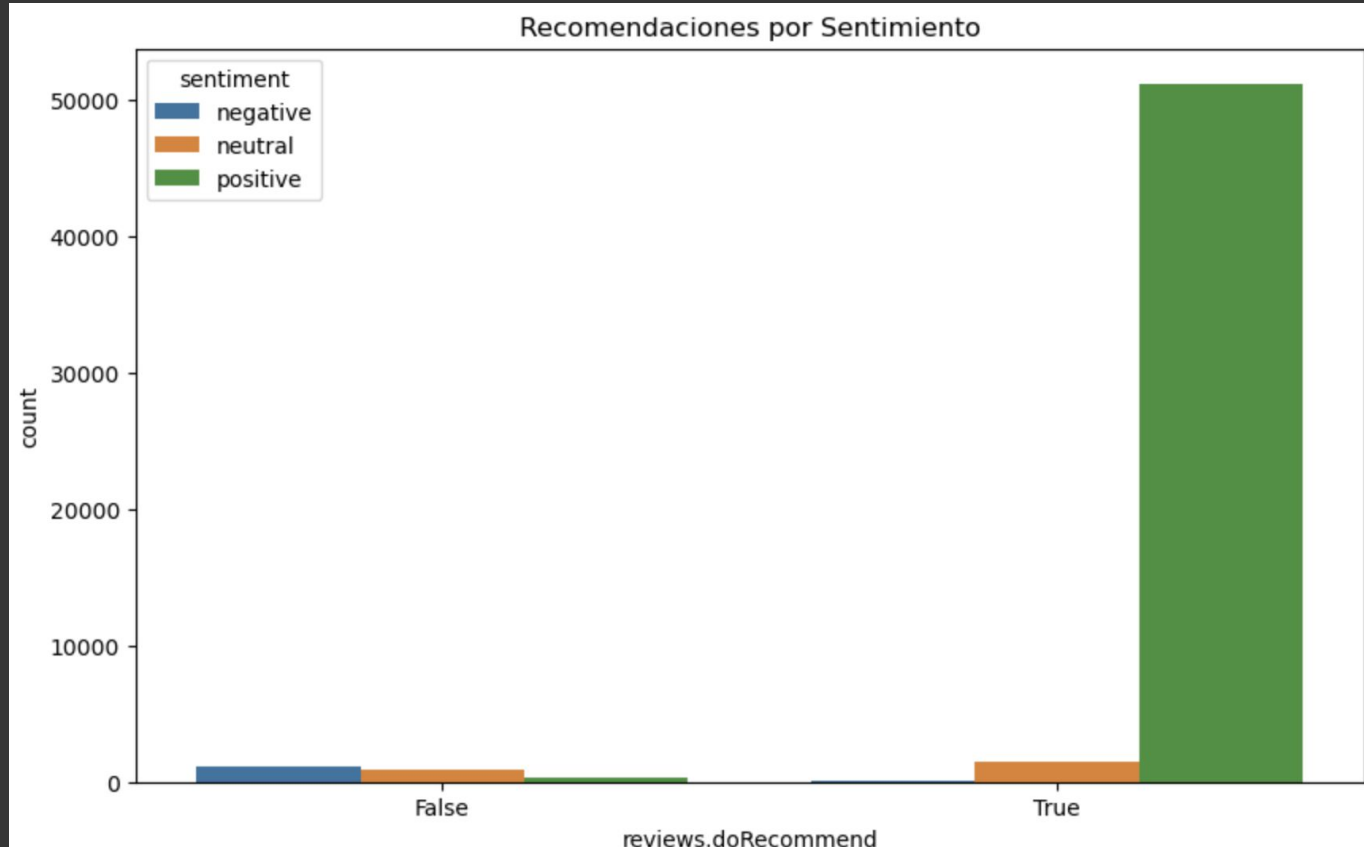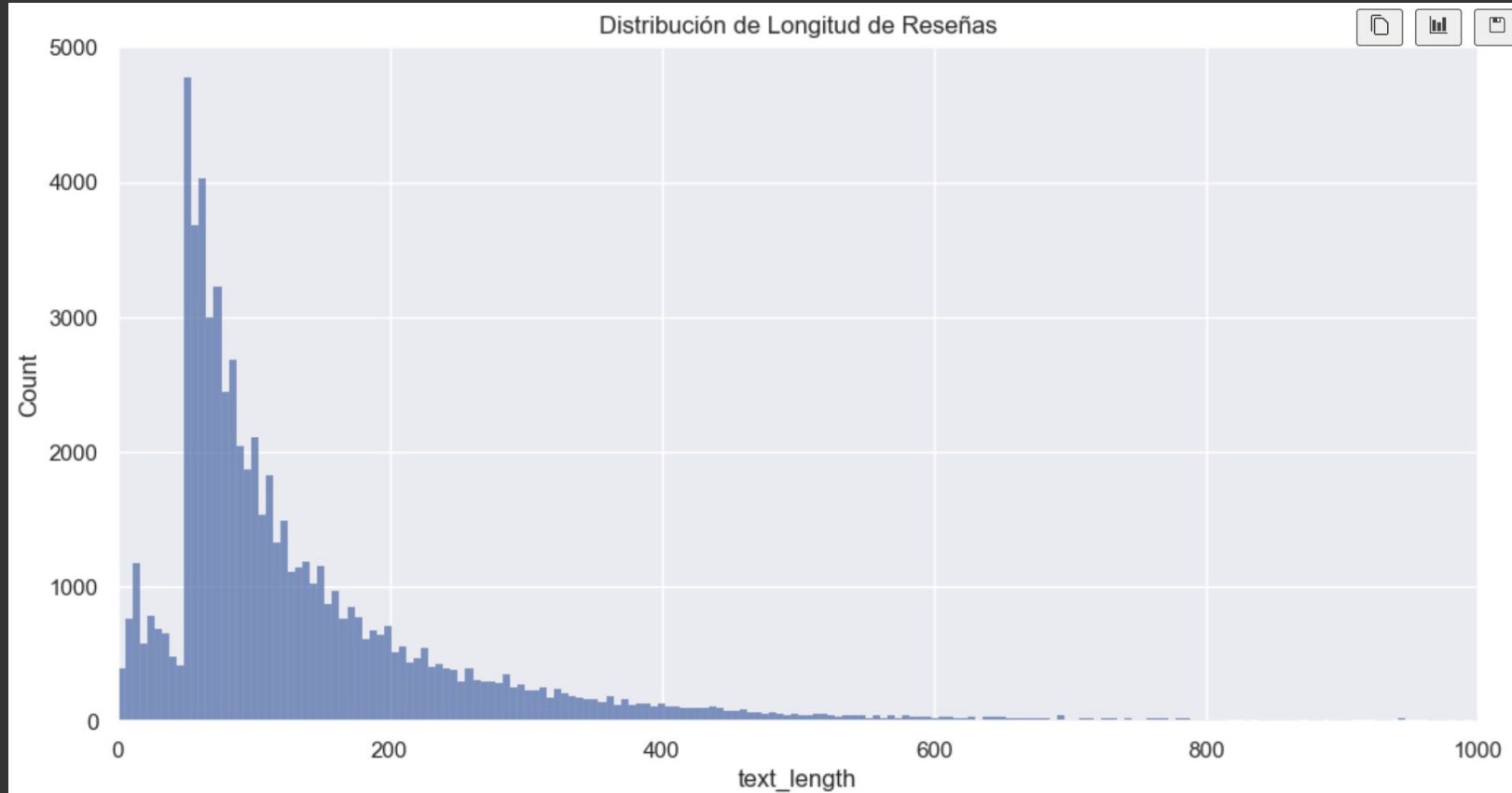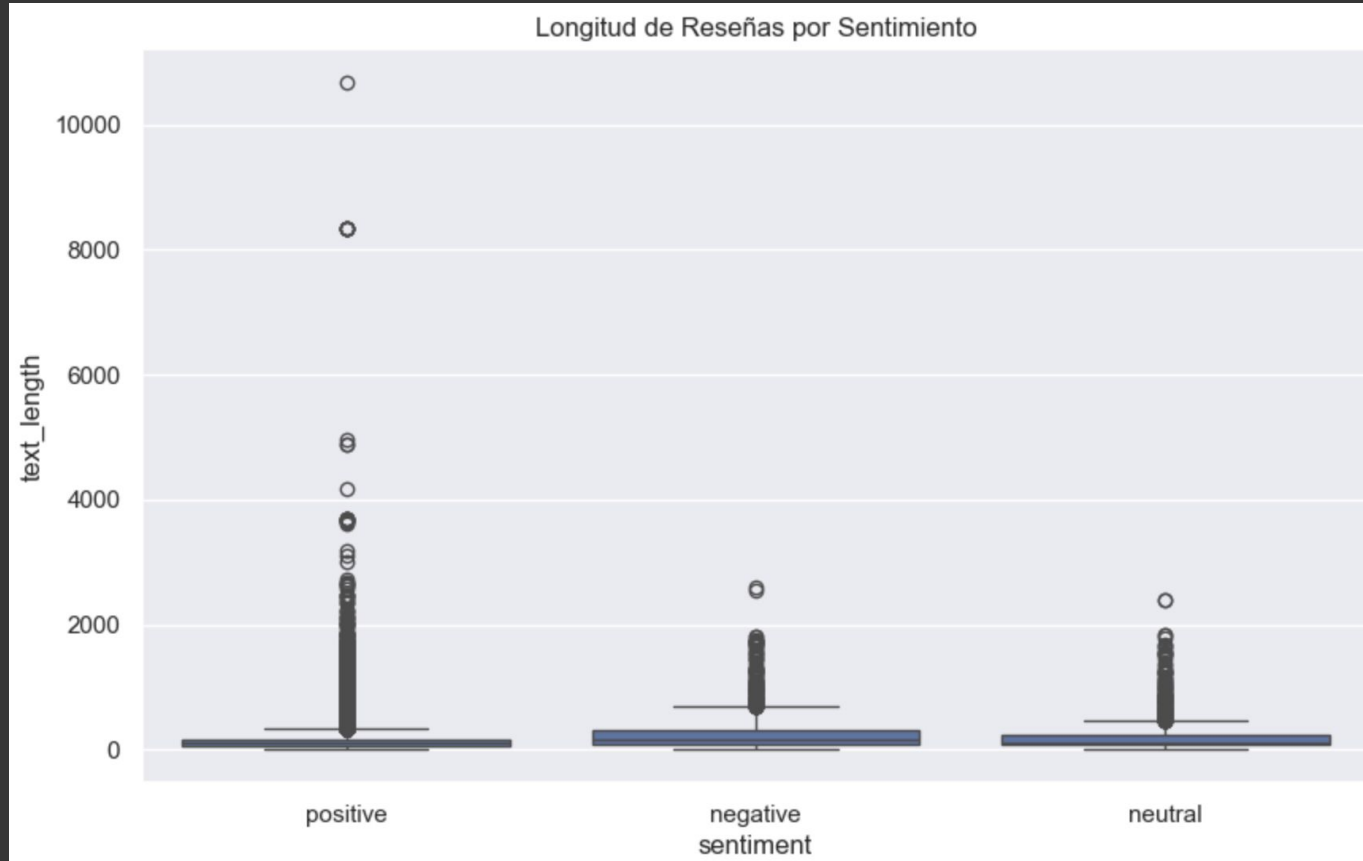Dominant Category of products: Electronics

Data Analisis

Distribución de Ratings

Data Analisis

Distribución de Sentimiento

Data Analisis

Data Analisis

Data Analisis

Longitud de Reseñas por Sentimiento

Data Analisis

Evolución de Ratings Promedio por Mes

Data Analisis

Palabras más frecuentes en reseñas positive

Data Analisis
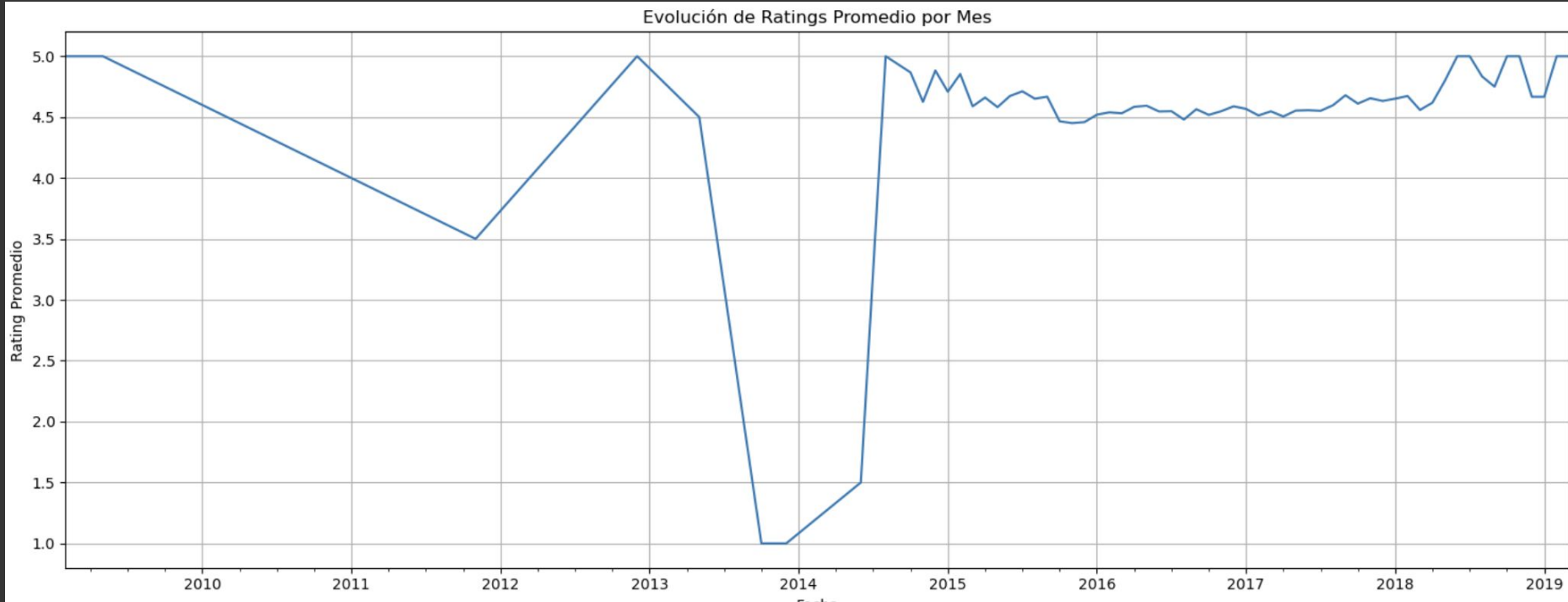
Palabras más frecuentes en reseñas negative

Data Analisis

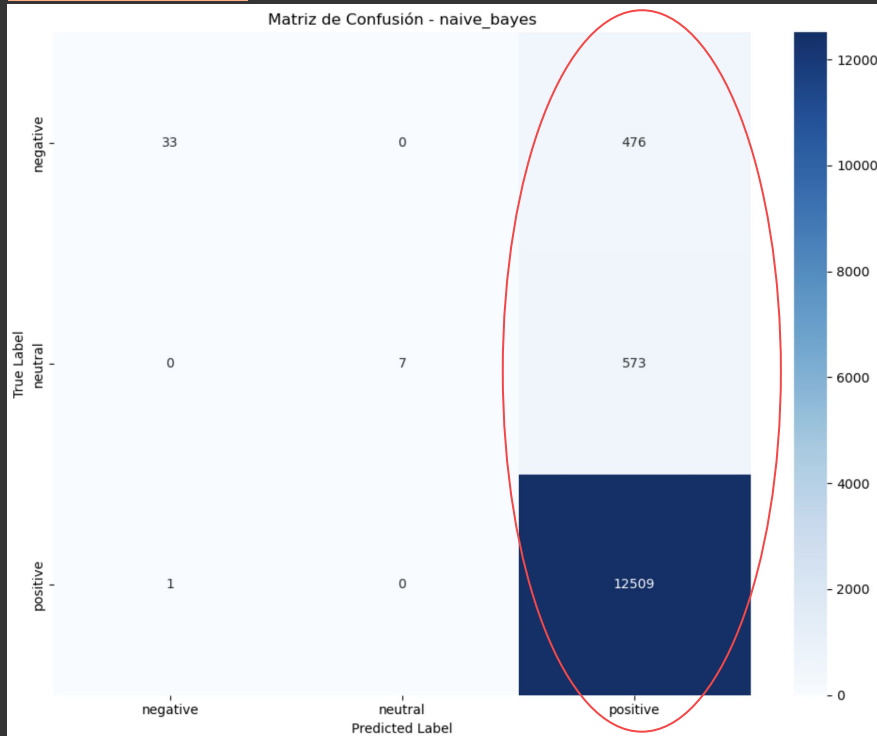Palabras más frecuentes en reseñas neutral

Preprocessing

➜ **Space and punctuation normalization**

➜ **Special characters removal**

➜ **Lowercase conversion**

➜ **Lemmatization using spaCy (es_core_news_md model)**
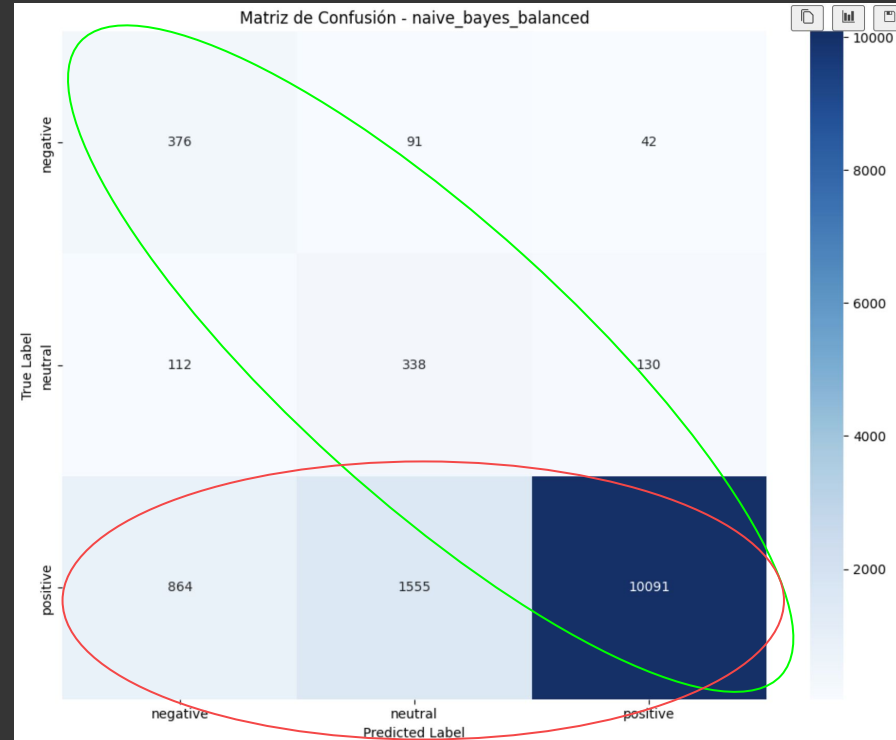
➜ **Stopwords removal using NLTK**

Traditional Models without balancing VS Balanced Traditional Models

Naive Bayes

Logistic Regression
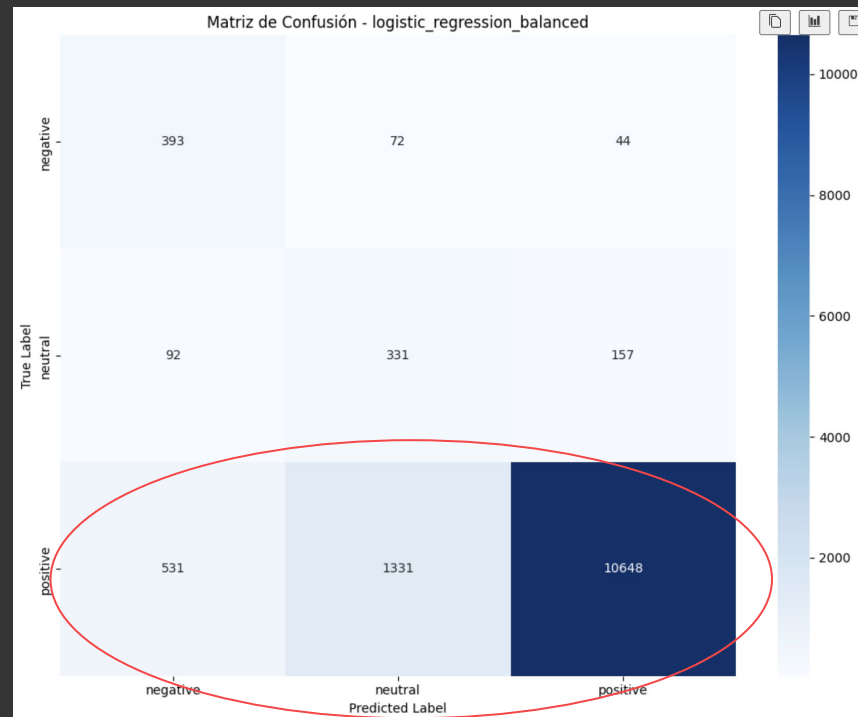
SVM

Random Forest

Accuracy: 0.95

Matriz de Confusión - random_forest

Accuracy: 0.95

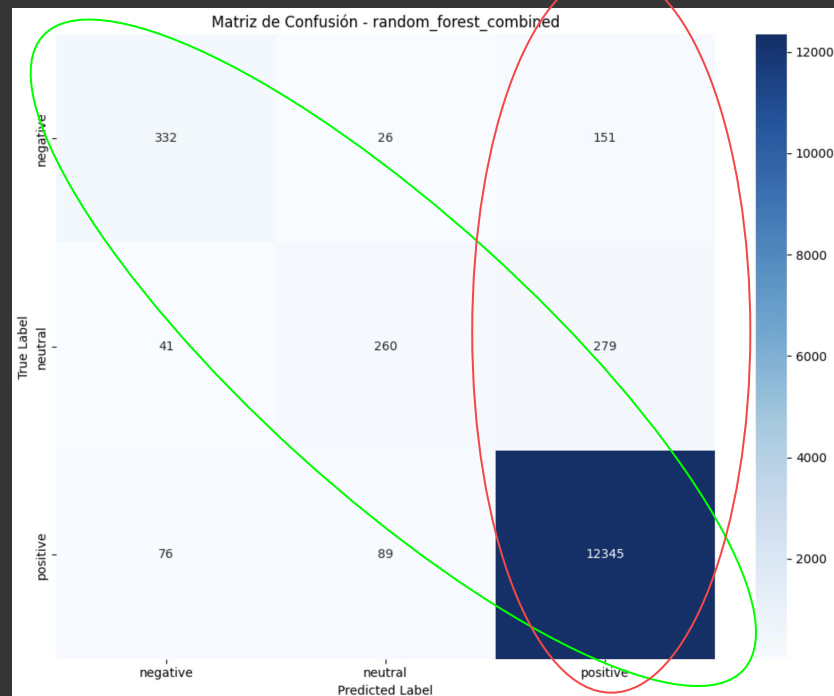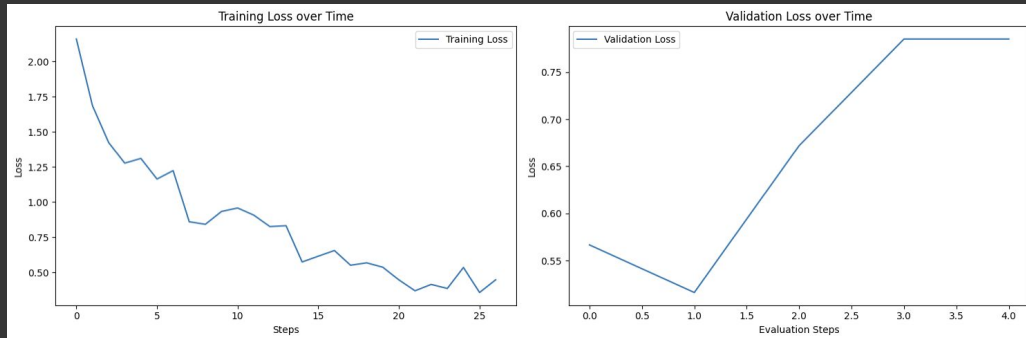Matriz de Confusión - random_forest_combined
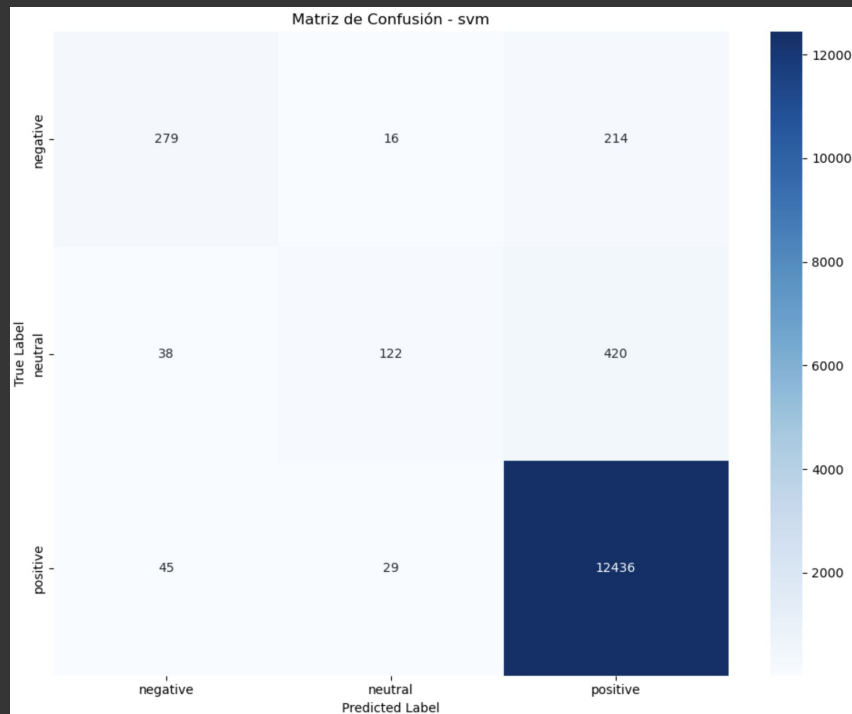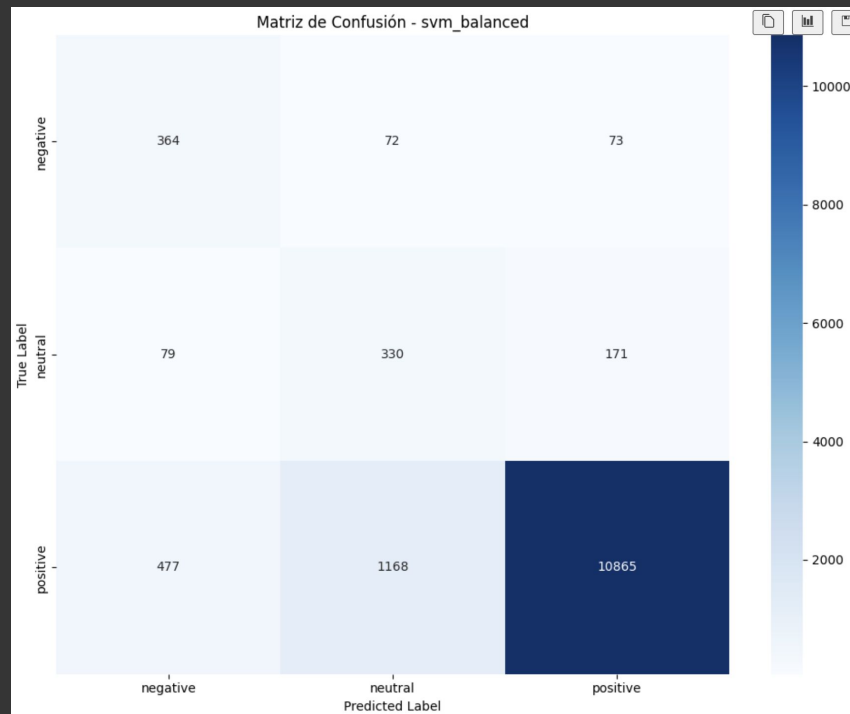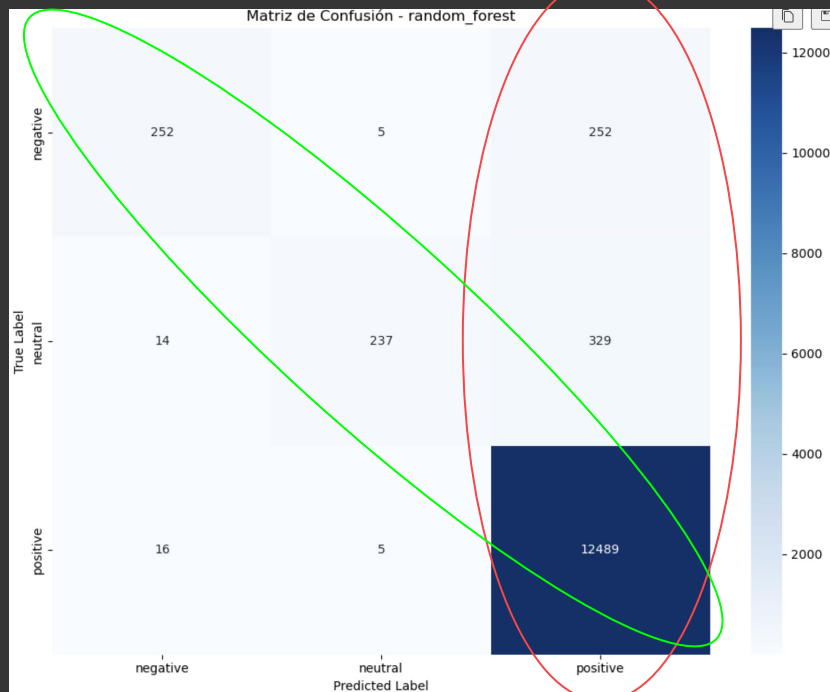
Transformer Model: BERT

configuration

- Base model: BERT (bert-base-uncased)
- Maximum tokenization: 128 tokens
- Batch Size: 32
- Learning Rate: 2e-5
- Weight Decay: 0.01
- Epochs: 4

Results

- Accuracy: 0.960290
- F1-Score: 0.959570
- Precision: 0.958969
- Recall: 0.960290

Conclusions

Transformer (BERT)

- Best overall performance (96% accuracy)
- Greater generalization capacity
- Requires more computational resources

Random Forest

- Best traditional model (95% accuracy)
- Easy to implement and maintain
- Lower computational cost

7. Dashboard