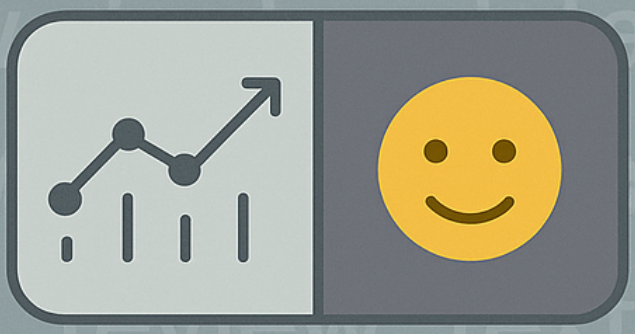
DATA PROJECT

Review Insights & Sentiment Prediction



Project Summary Review Insights: Sentiment Classification

Objective

Classify customer reviews from Amazon into three sentiment classes (positive, neutral, negative) using both traditional ML (TF-IDF + Logistic Regression) and deep learning (DistilBERT).

Dataset

- Source: Kaggle Datafiniti Amazon Reviews
- Size: ~28,000 reviews
- Fields: product name, category, review text, rating
- Labeling logic: ratings 12 = negative, 3 = neutral, 45 = positive

EDA Highlights

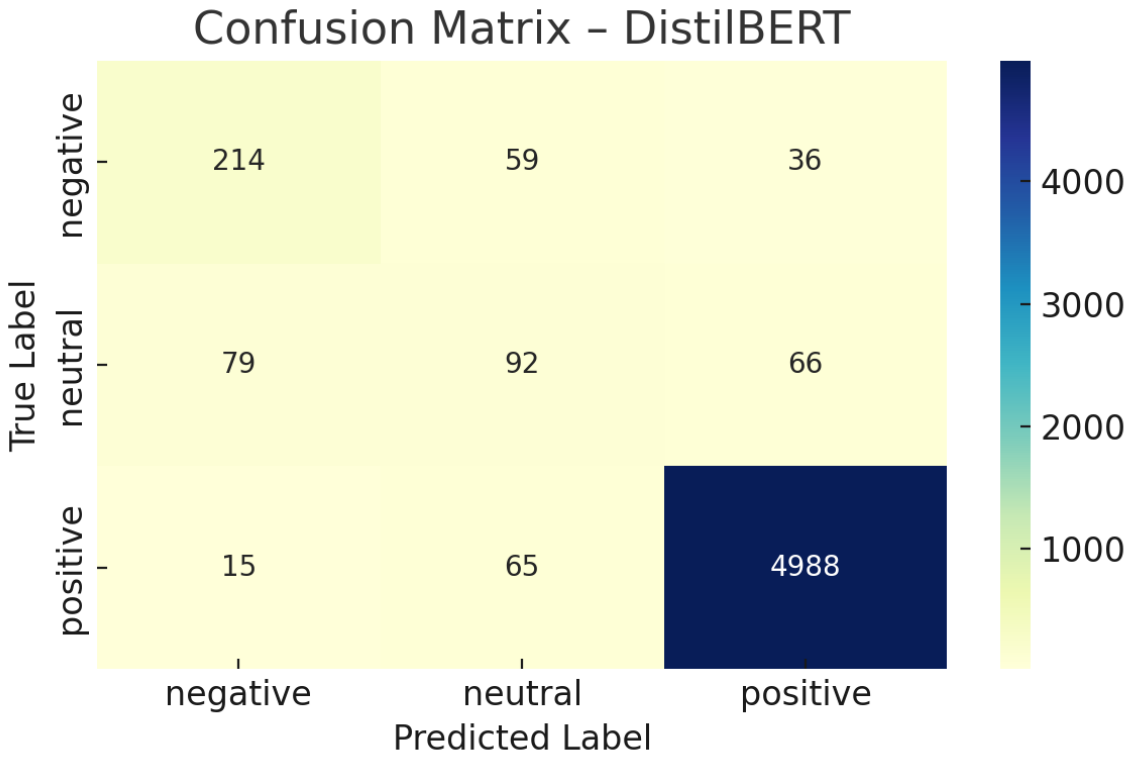
- 89% of reviews are positive (class imbalance)
- Reviews are short (median ~17 words)
- Top categories: Electronics, Health & Beauty
- Key terms: 'great', 'battery', 'return', 'fail', 'love'

Models and Performance

1. TF-IDF + Logistic Regression
   - Accuracy: 86.5%
   - Macro F1: 0.6438
   - Weakness: Poor recall on neutral class

2. DistilBERT (fine-tuned)
   - Accuracy: 94.3%
   - Macro F1: 0.7214
   - Positive F1: 0.9740 | Neutral F1: 0.4612 | Negative F1: 0.7291

Confusion Matrix - DistilBERT



Conclusions

- DistilBERT significantly outperforms TF-IDF baseline
- Shows robustness in positive and negative classes
- Demonstrates ability to implement transformer fine-tuning for NLP
- Business-ready solution for review sentiment monitoring

Author

Diego Alejandro Vlez Martnez

Senior Project Manager | AI Engineer in Training

GitHub: diegoavelez | LinkedIn: diegoavelezm