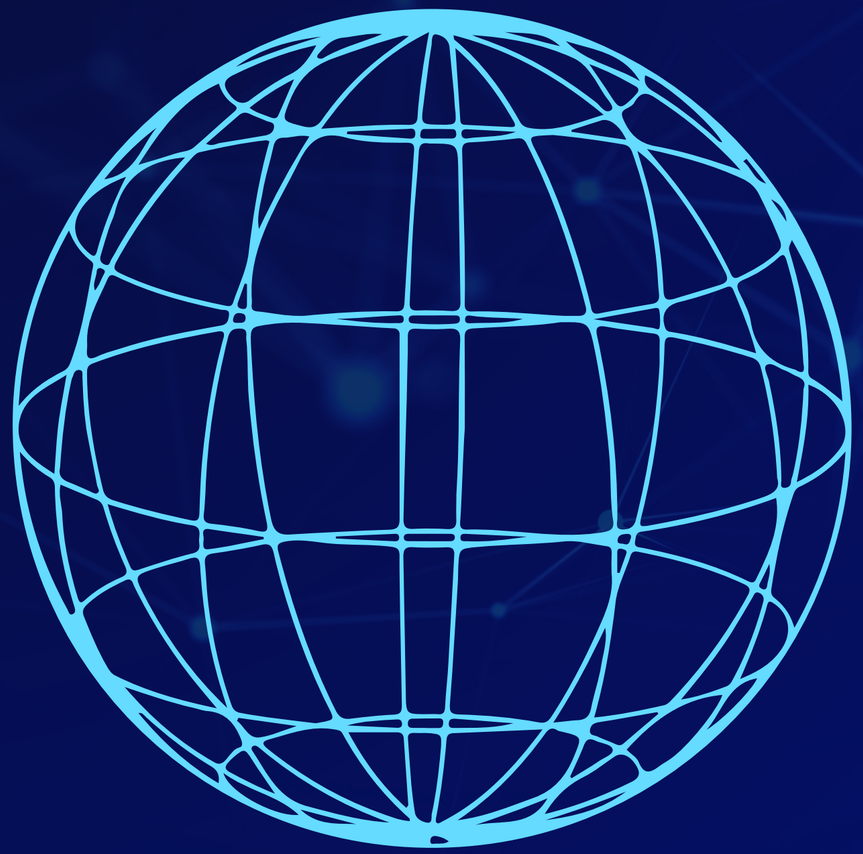
Sentiment Elicitation from Amazon Reviews: A Comparative Analysis of Deep Learning Approaches

Group 4 – Asfandiyar Safi, Emaan Waleed, Shahrez Faisal, Maryam Rizwan

Why Sentiment Analysis? Why Amazon Reviews?

Key Points:

- In the digital economy , reviews are gold:
- Amazon reviews contain unstructured emotional cues
- **Goal:** Build a system to automatically classify review sentiments using modern NLP techniques.
- **Importance:** Helps brands detect issues early, tailor marketing, and improve customer satisfaction



Problem statement

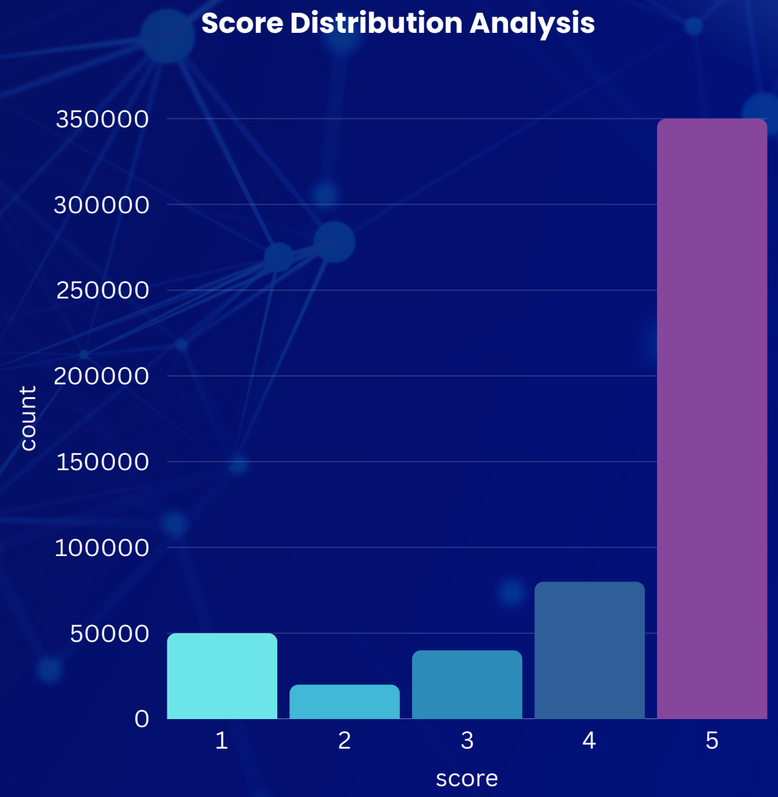
- Classifying into three sentiment classes, not just positive/negative
- Benchmarking three model types: CNN + GloVe, RoBERTa, DistilBERT
- Using TextBlob vs. Star Ratings to validate sentiment alignment
- Extracting frequent phrases using n-grams for aspect-level insights
- Detecting spam and duplicates with text similarity.
- Combining prediction + explainability → actionable outputs

Dataset Overview

- Source: Amazon Fine Food Reviews
- Total Reviews: 568,401
- Key Features:
  - Score (1–5 stars)
  - Text, Summary (review content)
  - Helpfulness votes, Timestamp, User ID
- Time span: 2001–2011
- Skewed distribution: Majority are 5-star reviews

EDA

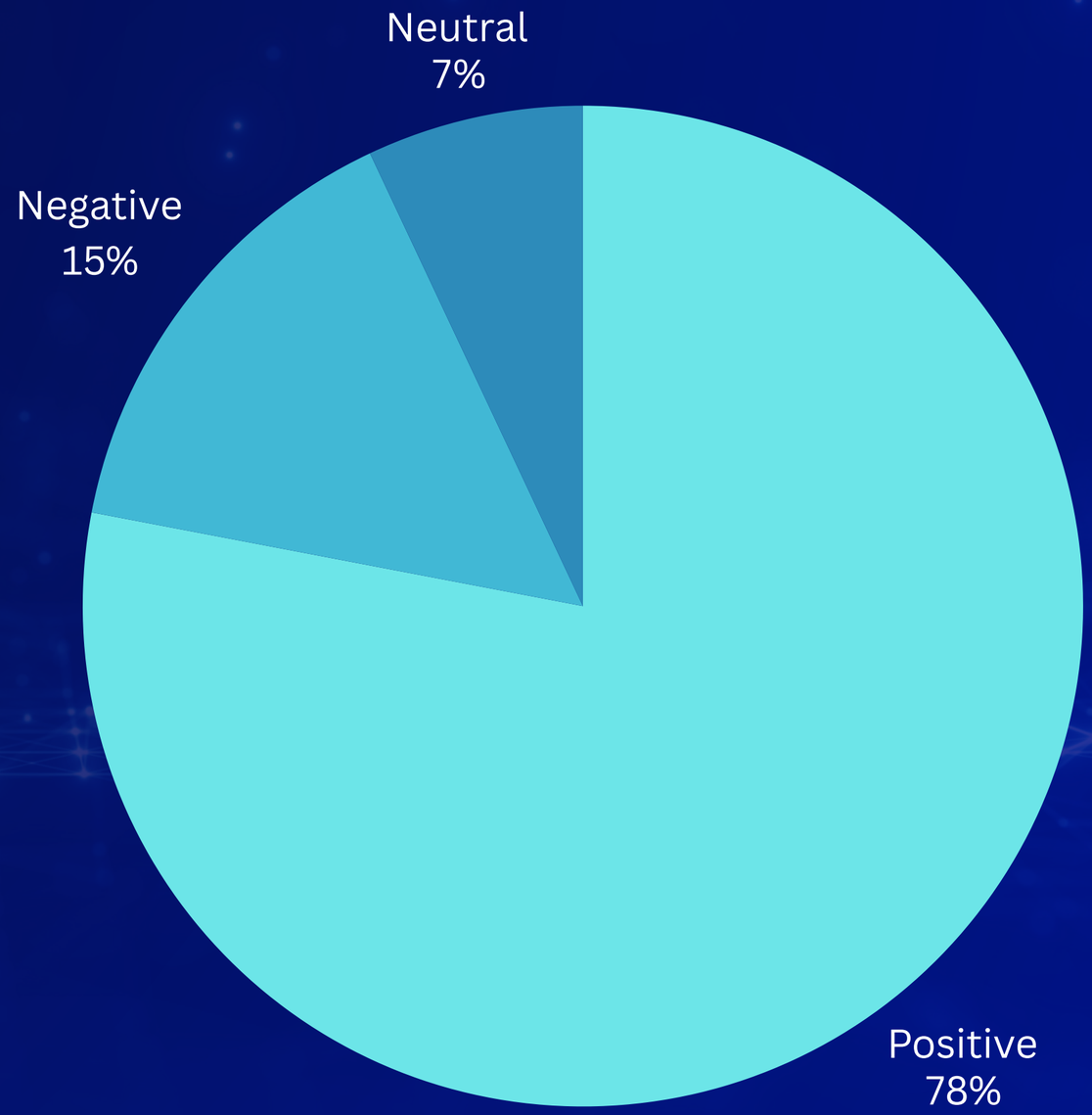
- **Data Cleaning:** Standardized and cleaned the review text by removing noise (HTML, URLs, stopwords, etc.) and tokenizing it—ensuring readiness for meaningful NLP analysis.
- **N-gram Patterns:** Identified common phrases showing customer priorities (taste, quality, recommendations)
- **Sentiment-Score Alignment:** Text sentiment matched numeric scores for positive reviews but less so for neutral/negative
- **Sentiment Validation:** Compared TextBlob polarity with star ratings to catch mismatches and noisy labels.
- **Review Length Patterns:** Found negative reviews (1-2 stars) tend to be longer and more detailed, offering the most actionable insights for product improvement



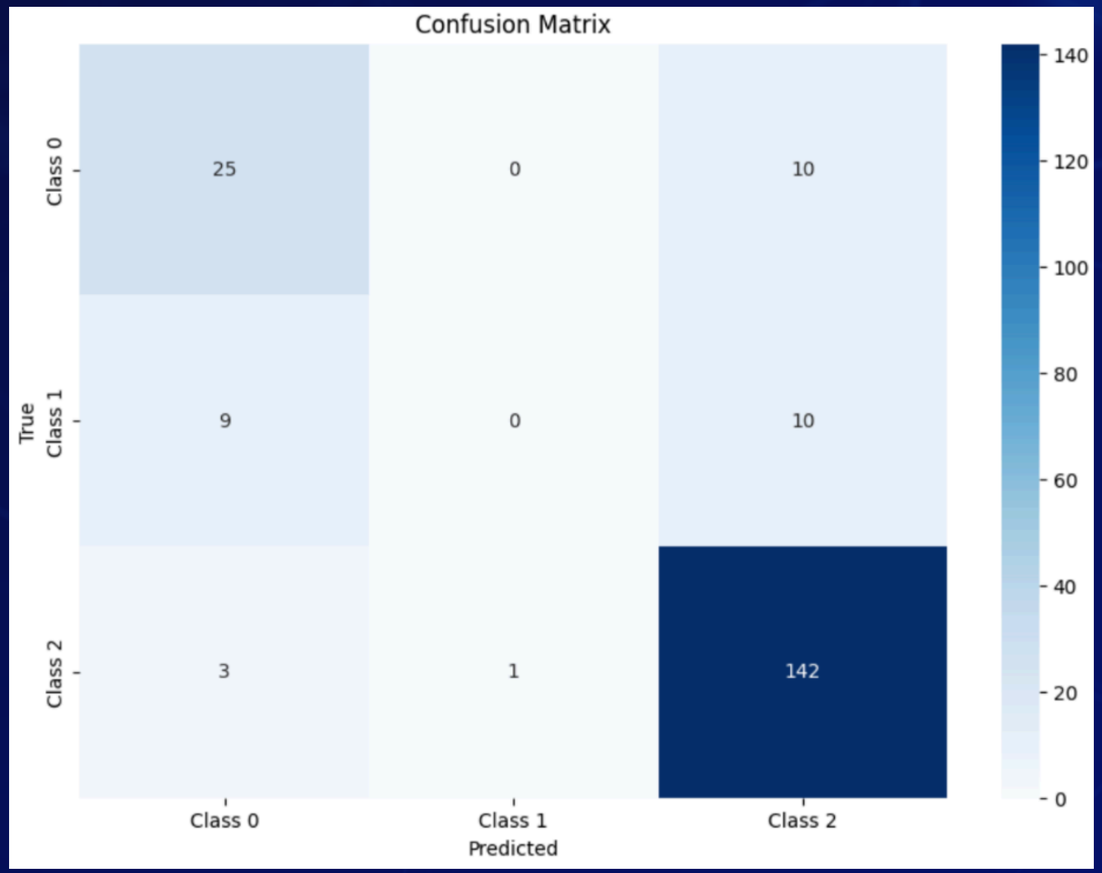
Amazon Review Sentiment

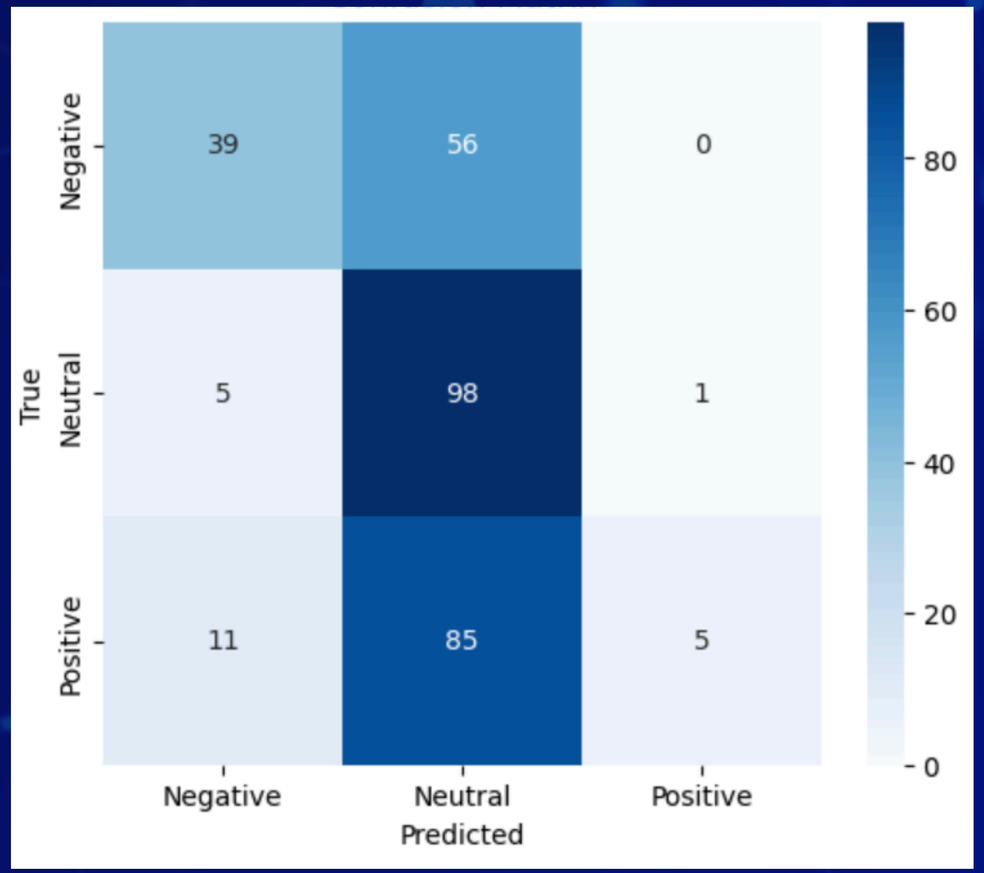
From Failure to Fix

Sentiment Distribution in Reviews

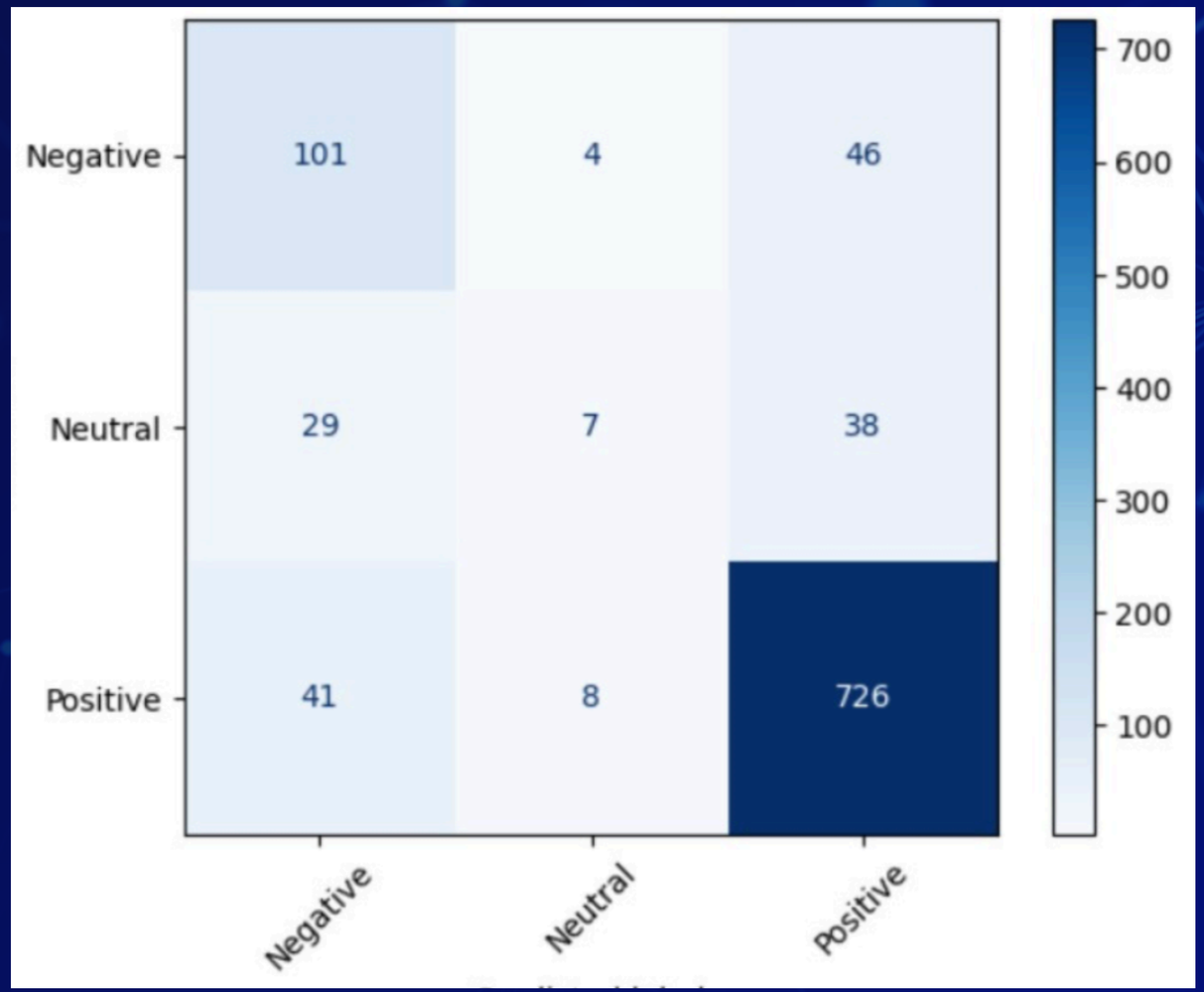


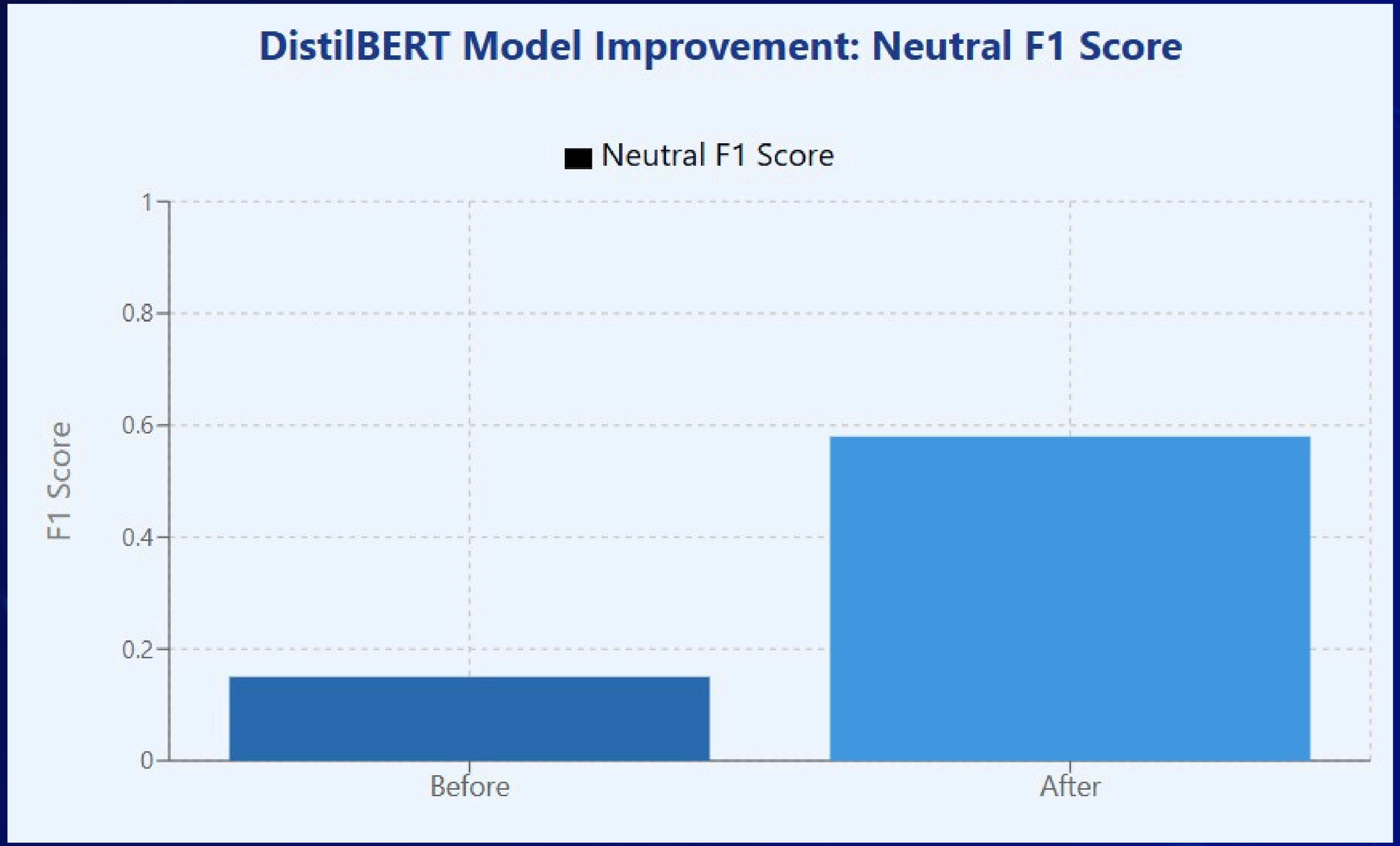
RoBERTa vs CNN

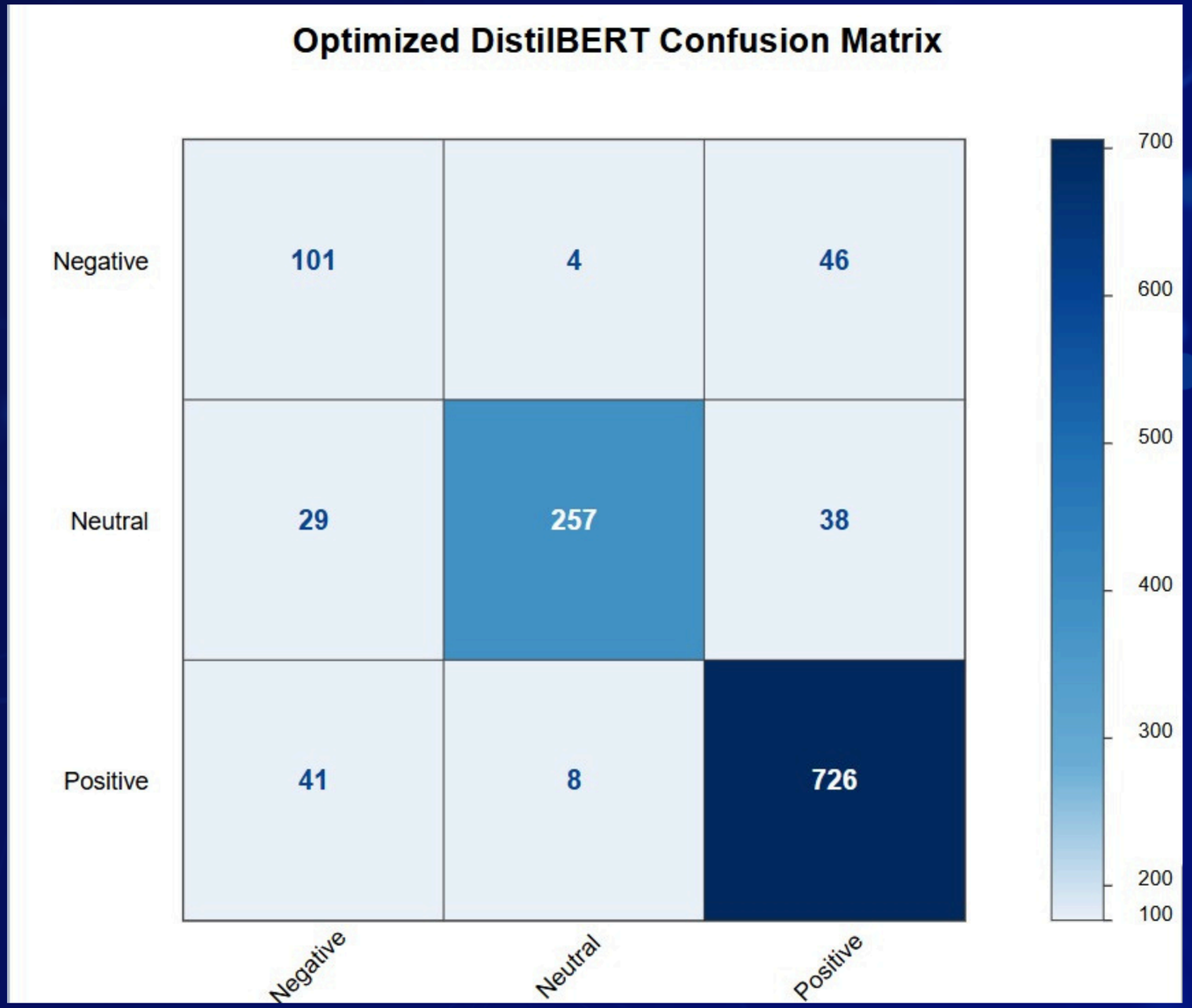




DistilBERT


Negative
Neutral
Positive
101
4
46
29
7
38
41
8
726
Negative
Neutral
Positive
700
600
500
400
300
200
100


DistilBERT Model Improvement: Neutral F1 Score
Neutral F1 Score
F1 Score
1
0.8
0.6
0.4
0.2
0
Before
After


Optimized DistilBERT Confusion Matrix
Negative
Neutral
Positive
101
4
46
29
257
38
41
8
726
Negative
Neutral
Positive
700
600
500
400
300
200
100

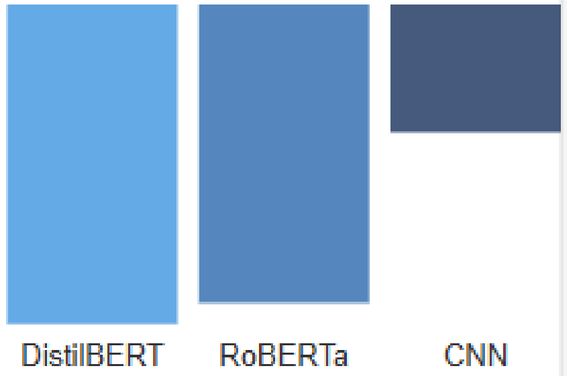
Conclusion

Transformer Excellence

DistilBERT: 86.7% accuracy

Outperformed RoBERTa (85.1%) and CNN (40%)

Better on minority classes



Neutral Sentiment Challenge

F1-score improved 287% (from 0.15 to 0.58)

Remains hardest to identify

Practical Implications

F1 ≥ 0.92 for positive/negative

Strong signals for customer satisfaction detection

Neutral reviews highlight product improvement areas