THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

DATS 6312: NLP FOR DATA SCIENCE

NLP for Financial Consumer Protection

Automating Complaint Classification: A Comparative Analysis of Classical Machine Learning vs. Fine-Tuned Transformers

Presented by:Vaijayanti Deshmukh, Lasya Raghvendra, Amogh Ramagiri

The Problem Statement

The Context

The Consumer Financial Protection Bureau (CFPB) receives thousands of consumer complaints daily regarding products like Mortgages and Credit Cards.

The Challenge

Manual reading and routing of these narratives is slow, expensive, and inconsistent, leading to delayed resolutions.

The Objective

Develop an NLP model to automatically classify unstructured complaint text into the correct product category with high accuracy.

Exploratory Data Analysis: Initial Discovery

Data Specs

- **Source:** CFPB Consumer complaint Dataset
- **Size:** 12,218,152 complaints, 18 columns
- **Time & Domain:** Financial product and service complaints received by the CFPB
- **Goal of EDA:**
  1. Understand structure and content of the complaints data
  2. Check data quality (missing values, duplicates)
  3. Explore distribution of products, channels, and complaint text.

Data Quality

- 71% missing in consumer complaint narrative text column
- 94% missing in tags column
- 15-16% missing in consumer consent provided
- Duplicates: 0 duplicate rows

```
df.shape
```

```
(12218152, 18)
```

```
df.dtypes
```

	0
Date received	object
Product	object
Sub-product	object
Issue	object
Sub-issue	object
Consumer complaint narrative	object
Company public response	object
Company	object
State	object
ZIP code	object
Tags	object
Consumer consent provided?	object
Submitted via	object
Date sent to company	object
Company response to consumer	object
Timely response?	object
Consumer disputed?	object
Complaint ID	int64

dtype: object

	0
Date received	0.000000
Product	0.000000
Sub-product	1.925782
Issue	0.000049
Sub-issue	7.119293
Consumer complaint narrative	71.049296
Company public response	48.148288
Company	0.000000
State	0.471643
ZIP code	0.247411
Tags	94.391730
Consumer consent provided?	15.650141
Submitted via	0.000000
Date sent to company	0.000000
Company response to consumer	0.000164
Timely response?	0.000000
Consumer disputed?	93.711946
Complaint ID	0.000000

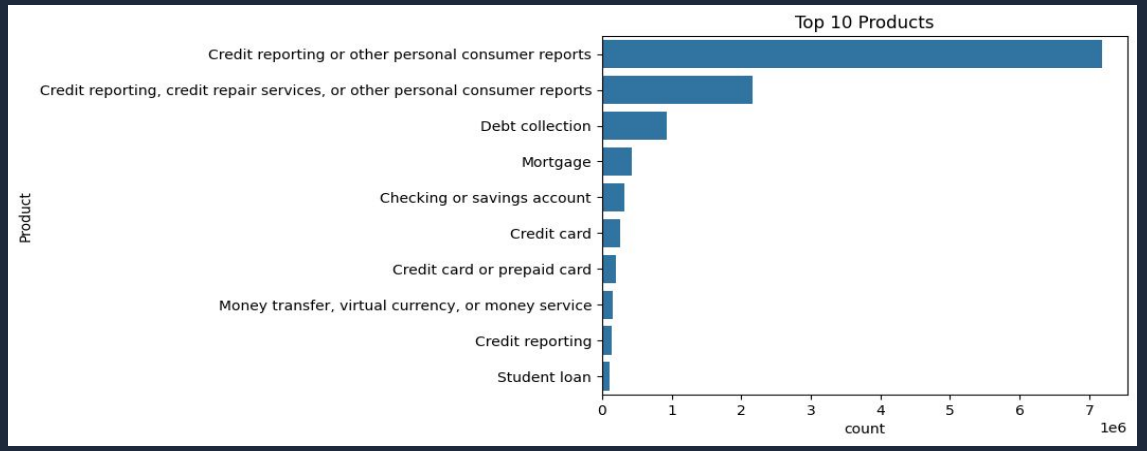
dtype: float64

Insight: High missingness in text fields means preprocessing will require filtering or imputation before modeling.

Univariate Analysis: Product, Geography and Text Patterns

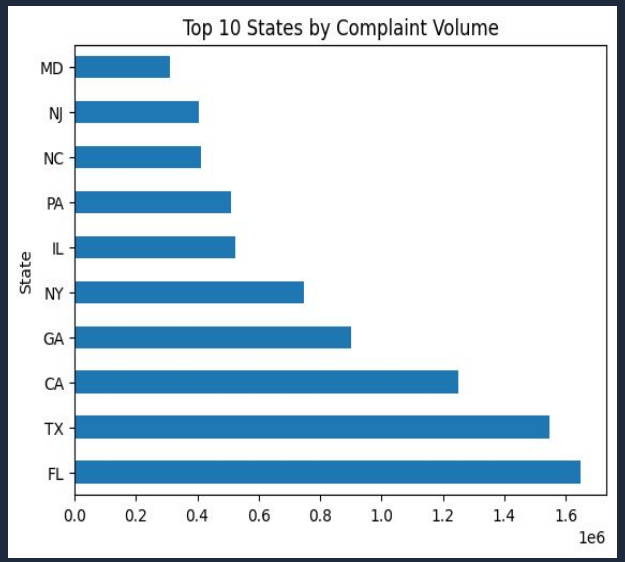
Product Distribution

- Credit Reporting - related complaints (~7M)
- The top 3 categories account for ~80% of total complaints
- Severe class imbalance observed across product categories.



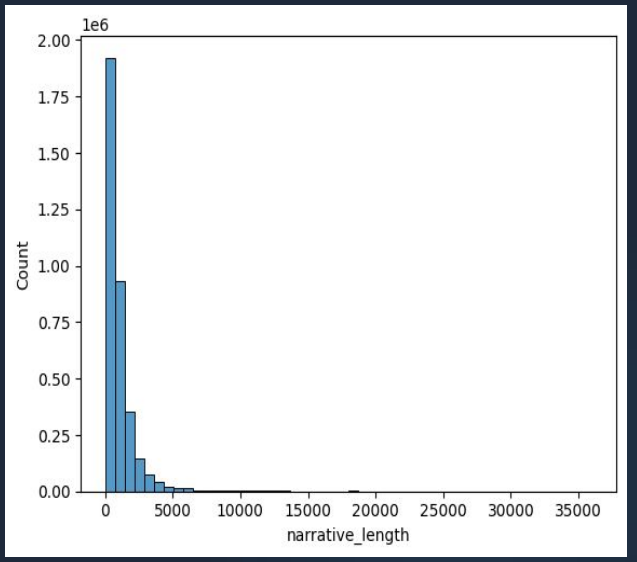
State Distribution

- Complaints are concentrated in CA, FL, TX, NY, GA
- Follows population trends - highly populated states submit the most complaints.



Narrative Text Availability

- 71% of complaints have no narrative text.
- Among available narratives (~3.5M):
  - Median text length: 115 words.
  - Mean text length: 176 words.
  - Long Tail with max: 6469 words.



Insight: Narratives are long-tailed and sparse, meaning text preprocessing will be essential for modelling.

Univariate Analysis: Distribution and Temporal

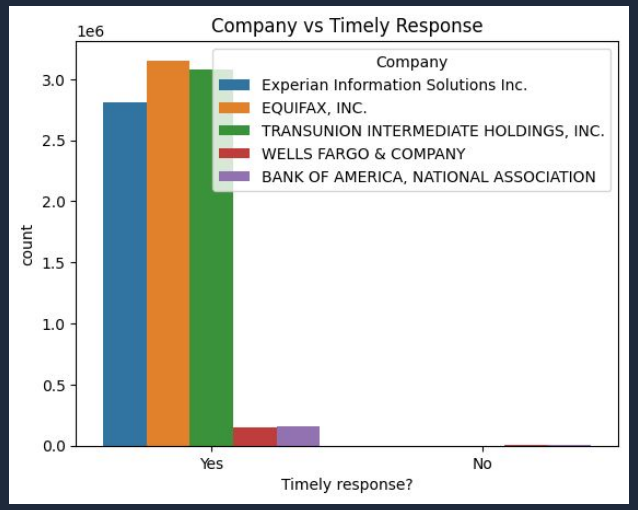
- Complaints steadily increased from 2015 to 2024, indicating growing consumer activity and CFPB usage.
- Sharp spike in 2025, driven by reporting changes, awareness, or major financial events.
- Sudden drop at the end of 2025, often due to incomplete reporting for recent months, not an actual decline.

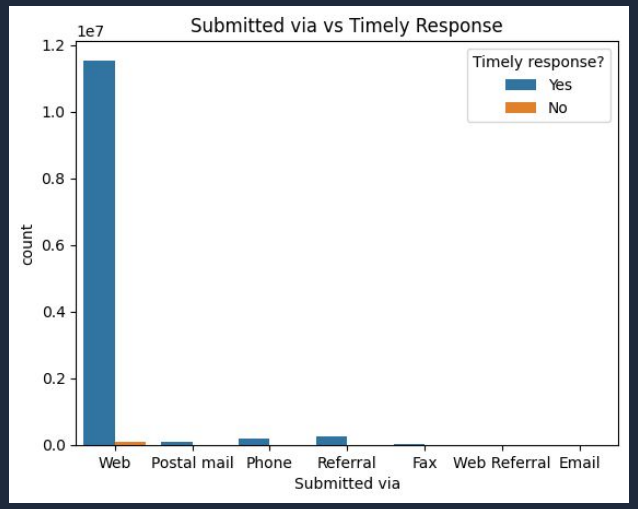
Insight: *Monthly complaint volume shows long-term growth, a major spike in 2025, and an end-of-year drop due to incomplete data.*

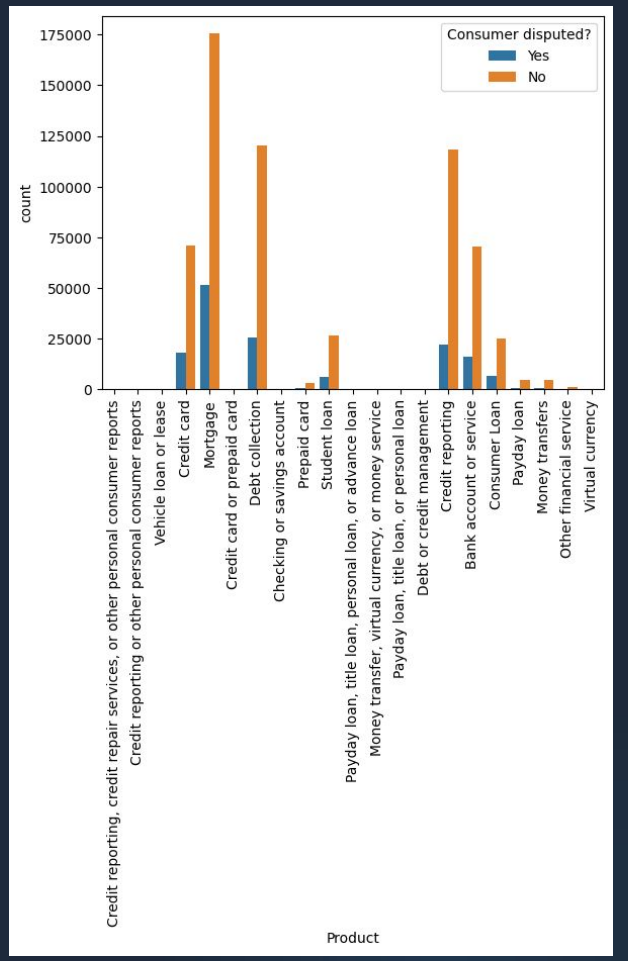
Bivariate Analysis: Company, Channel and Consumer Behavior

Key Observations

- Almost all major companies respond on time, with small differences across firms.
- Complaints submitted via web have the highest volume and strong on-time response rates.
- Phone and postal mail submissions show slightly higher late responses.
- For many products, most consumers do not dispute the company's resolution.
- The proportion of disputes varies by product category.







Insight: Timely responses remain high across companies and channels, while dispute behavior varies strongly by product type

Text Analysis & Key Insights

Narrative Availability

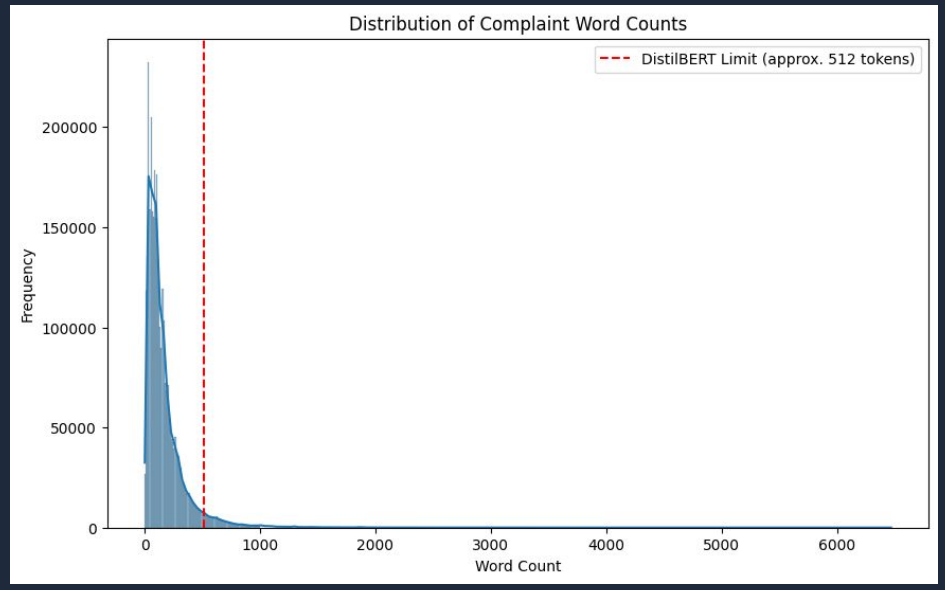
- 71% of complaints have no narrative text

Remaining Text

- After dropping missing narratives → 3,537,241 rows remain

Narrative Length Characteristics

- Median length: 115 words
- Mean Length: 176 words
- Long Tail:
  - 90th Percentile : 364 words
  - 95th percentile : 515 words
  - 99th percentile : 1021 words
  - Max : 6469 words

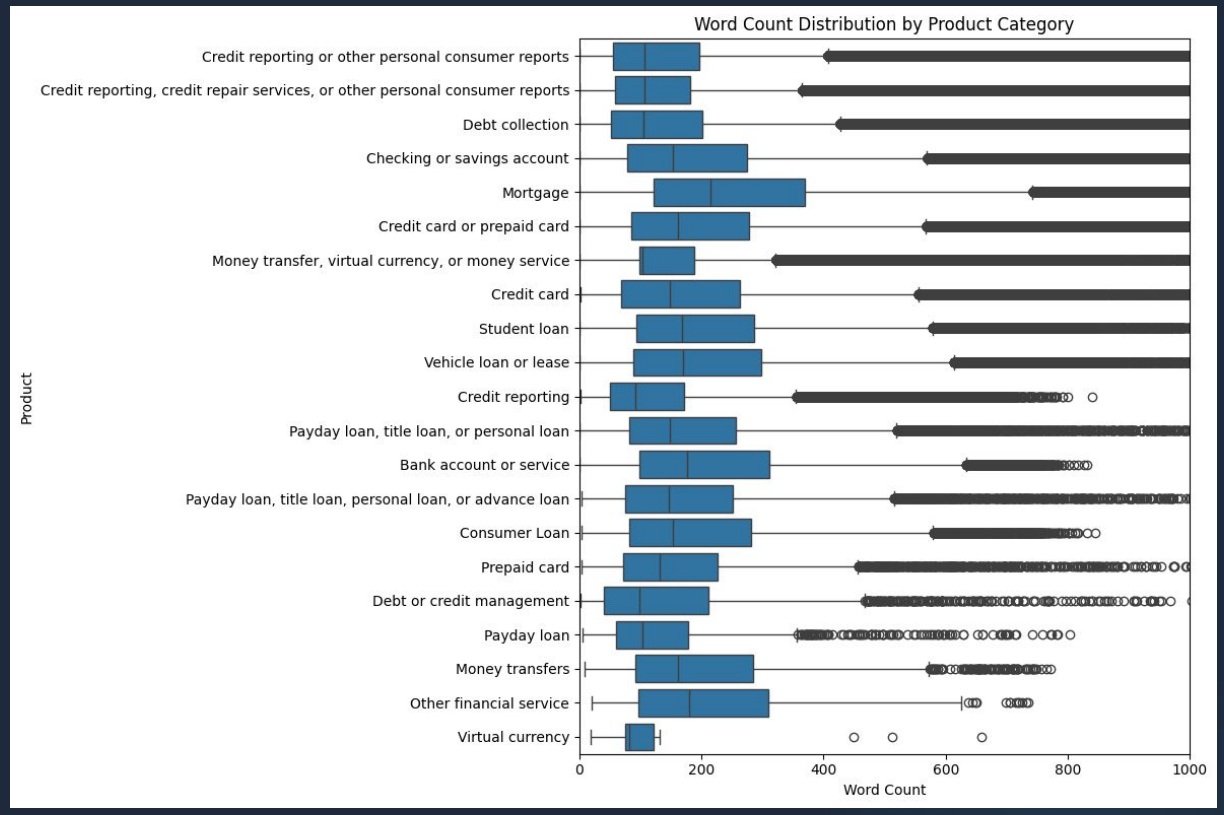


Word count shows strong right skew; most complaints fall under 400–500 words.

Insight: Narrative texts are long, messy, heavily redacted, and require cleaning + tokenization limits (512 tokens for DistilBERT)

N-gram Insights

- The three categories starting with “Credit reporting...” all show identical top N-grams, these categories are semantic duplicates.
- Frequent tokens (xxxx, xx) represent CFPB redacted PII, not meaningful words.
- “Debt collection” is distinct but still shares overlapping vocabulary with credit-reporting categories, high lexical overlap explains classifier confusion.



Some product categories contain longer narratives, but all show heavy-tailed length distributions.

Insight: Category labels in the dataset are inconsistent, and text is extremely noisy — transformer-based models (like DistilBERT) are better suited to handle this than simple bag-of-words models.

Preprocessing for Baseline

Multinomial Naive Bayes

BASELINE

The CFPB dataset has multiple slightly different names for “Credit Reporting,” such as:

1. Credit reporting
2. Credit reporting or other personal consumer reports
3. Credit reporting, credit repair services, or other personal consumer reports

We combined all these variations into one clean label: “Credit Reporting.”

- Cleaned PII redaction tokens (“XXXX”)
- Stratified train-test split (80/20)

Baseline Results (Naive Bayes)

83%

ACCURACY

The "Accuracy Paradox"

The model achieved decent accuracy only because it over-predicted the majority class ("Credit Reporting").

36%

MACRO F1-SCORE

It failed miserably on rare classes. It learned frequency, not meaning.

Methodology

Multinomial Naive Bayes

Technique: Bag-of-Words with TF-IDF.

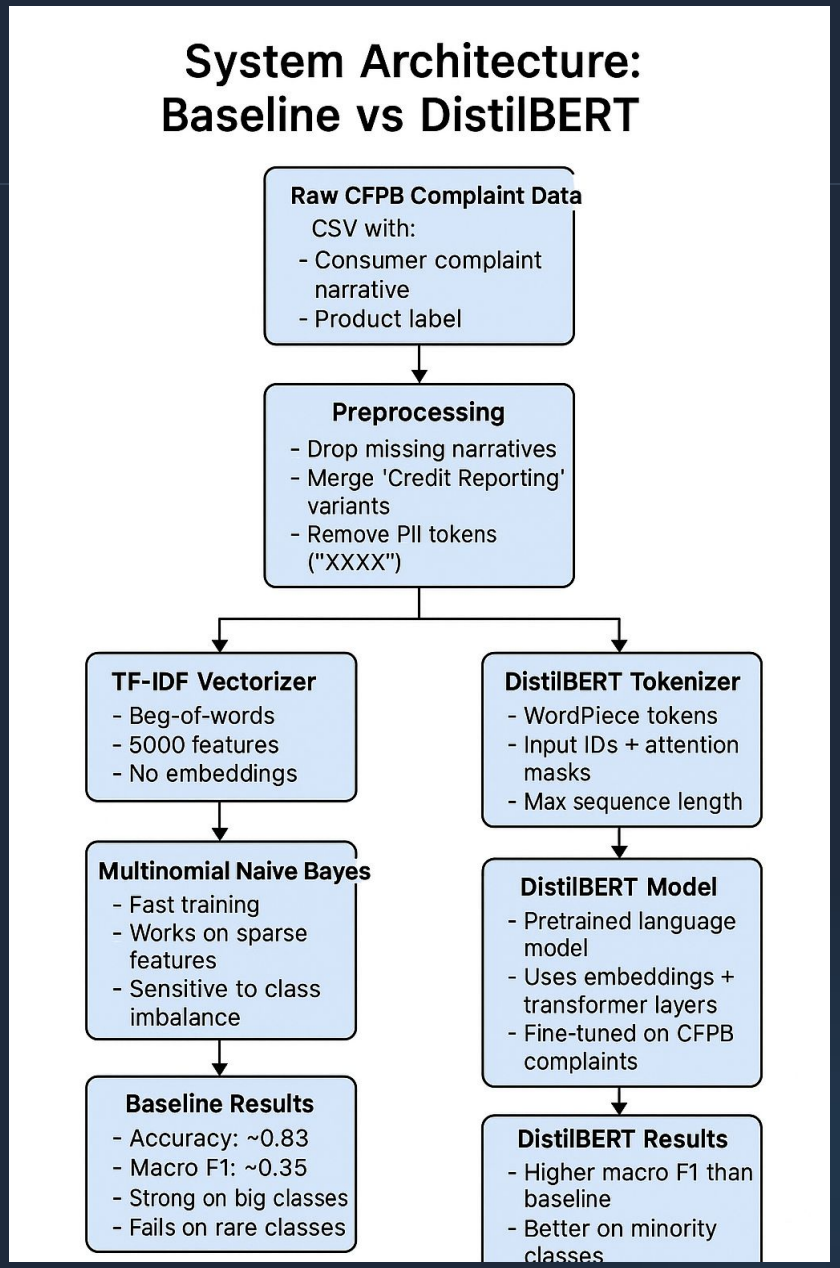
Hypothesis: Fast and interpretable, but will struggle with context (e.g., "credit card" vs "credit report").

Fine-Tuned DistilBERT

Technique: Pre-trained Transformer (Bidirectional).

Why: Understands semantic context and reading direction.

Fine-tuned for 2 epochs



The Cleaning Pipeline: Label Consolidation

The Consolidation Logic

```
# Merge Duplicate Categories credit_categories = [
'Credit reporting, credit repair...', 'Credit
reporting or other...' ] clean_name = 'Credit
Reporting' df[label].replace(credit_categories,
clean_name)
```

This explicit mapping prevented the model from treating identical products as separate classes, a crucial step for model stability.

Dataset Statistics

ORIGINAL LOAD

12,218,152 rows

AFTER FILTER & CLEANING

3,537,241 rows

FINAL DISTINCT CLASSES

19 Labels

DistilBERT Results

DistilBERT achieves 91% accuracy and a macro F1 of 60%, which is a substantial improvement over the baseline. The model understands context better and performs well even on minority complaint categories.

91%

ACCURACY

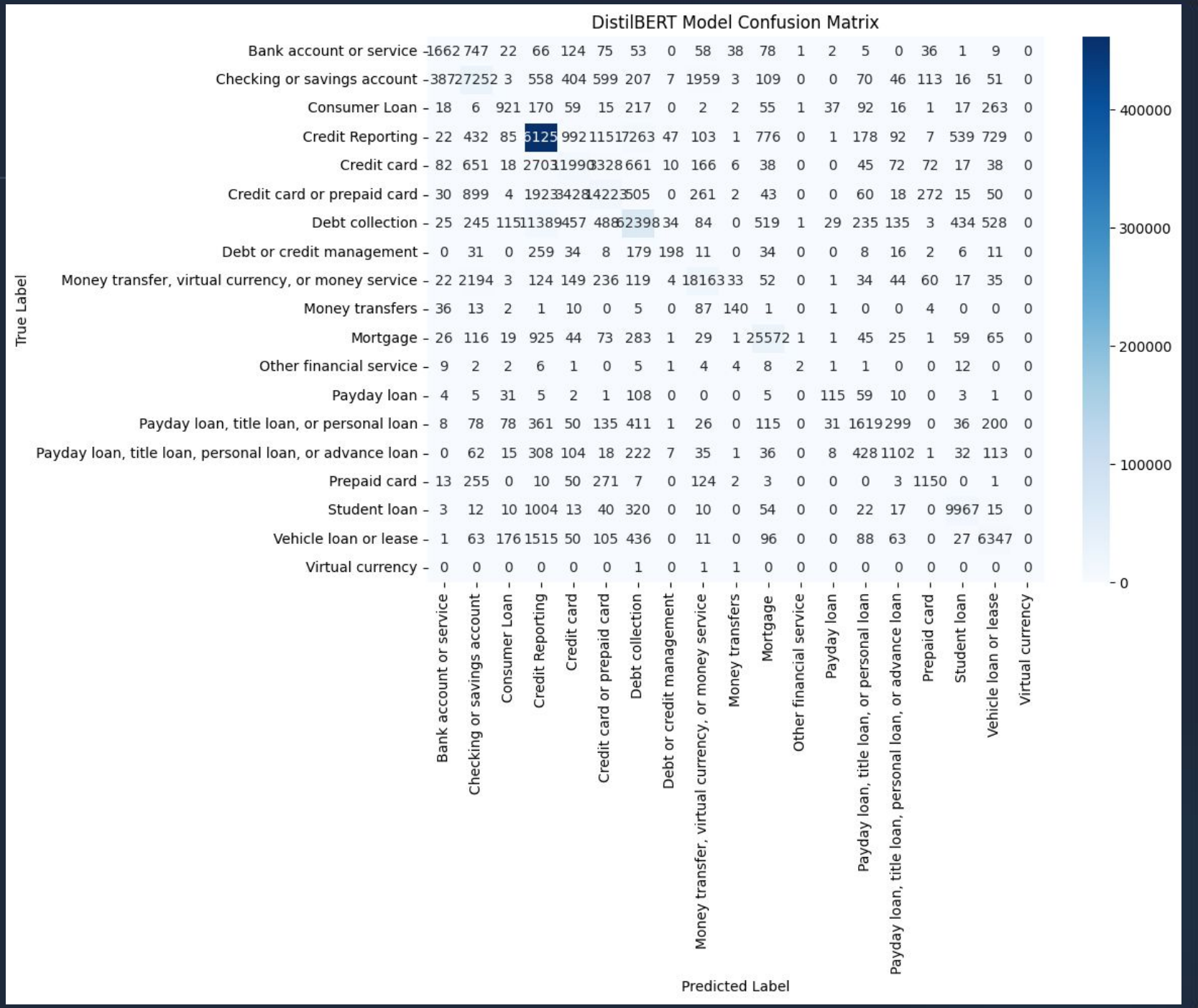
Strong overall classification performance

60%

MACRO F1-SCORE

Better balance across minority classes

DistilBERT Results



Why the Transformer Won

Contextual Awareness

Naive Bayes sees isolated words. DistilBERT reads "My card was declined at the pump" and understands the transactional context, distinguishing it from reporting errors.

Handling Noise

Attention mechanisms allowed the model to focus on relevant signals (verbs, nouns) and ignore the 'xxxx' redaction noise better than simple TF-IDF.

Robustness

The model generalized well to the validation set with minimal overfitting, proving its viability for real-world application.

CONCLUSION