CASE STUDY

Blanc research

NLP Analysis Exposes Copy-Paste Fraud

How Machine Learning Identified Fraudulent Patterns in Text Data

Blanc Research

November 2025

The Challenge: Copy-Paste Fraud

Open-ended survey questions are uniquely vulnerable. Traditional manual review often misses subtle similarities in massive datasets.

⚠ Fraudsters reuse near-identical answers across multiple responses.

Corrupted qualitative data severely skews insights and themes.

40%
of open-ended responses in some studies show copy-paste patterns.

NLP Technology for Fraud Detection

Semantic Similarity

Detects identical or near-duplicate text using cosine similarity models.

Clustering

Groups similar responses to identify suspicious volume clusters instantly.

Linguistic Fingerprinting

Analyzes writing patterns, vocabulary, and syntax anomalies.

Real-Time Scoring & Automated Flagging

Flags high-risk responses instantly during collection, prioritizing them for review.

The Study: Consumer Preference

5k

Respondents

Across 3 countries

08

Questions

Qualitative feedback

04

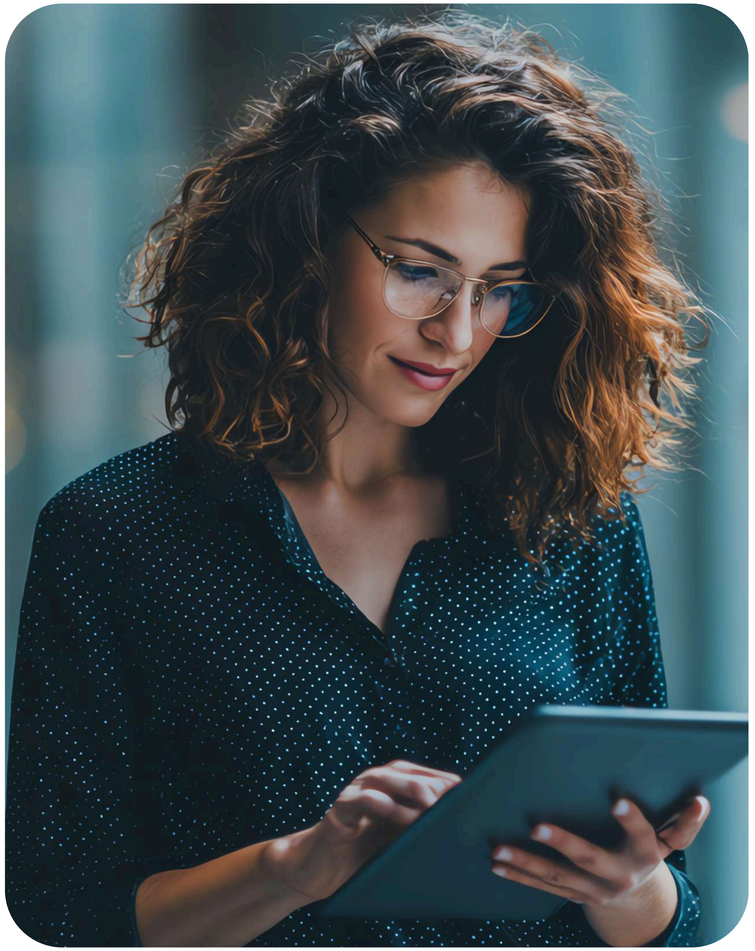
Weeks

Data collection period

⚠

Initial Concern

High clustering detected



Findings: Alarming Fraud Patterns

16.9%

of total responses flagged (847)

89%

Quality Score (improved from 62%)

Duplicate Clusters

12 major clusters identified. Largest cluster had 156 responses with 92% similarity.

Geographic Pattern

78% of fraudulent responses originated from just 3 IP addresses.

Behavioral Speed

Suspicious responses submitted in rapid succession (2-3 seconds).

Data Impact

Removal of these responses drastically shifted the sentiment analysis results.

Real-World Consequences Prevented

✕ Original Insights (Fraudulent)

False Sentiment: Data suggested "frustration with customer service."

Planned Action: Company was planning a $500K customer service overhaul.

✓ Revised Insights (Clean)

True Sentiment: "Satisfaction with product features."

Actual Action: Invested in product feature development instead.

+35% ROI

Lessons Learned & Next Steps

Standardize NLP: Open-ended responses require machine learning analysis, manual review is insufficient.

Set Thresholds: Use semantic similarity thresholds (e.g., >85%) to auto-flag potential fraud.

Real-Time Detection: Prevent costly downstream decisions by catching fraud during collection.

Partner with Blanc Research

Implement next-generation fraud detection in your next study.