Keyword-Based Correspondence Classification

A Lightweight, Auditable Framework for Categorizing Consumer Communications in Restricted Environments

Rishi Bharaj - 2025

Abstract

This project introduces a lightweight, Excel-based framework for categorizing consumer communications using a keyword-driven approach. Tailored for environments with strict data confidentiality and limited access to external tools, the system enables structured interpretation of unstructured inputs such as scanned documents, emails, and SMS messages. By leveraging rule-based logic and statistical validation, the framework offers operational and quality teams a scalable, auditable, and interpretable method for correspondence classification and outcome prediction.

While the framework does not employ machine learning, it embodies core principles of artificial intelligence:

- Rule-Based NLP: Simulates intent detection through structured keyword logic
- Knowledge Representation: Uses categories and keyword sets to form a semantic structure
- Decision Support: Facilitates structured decision-making from unstructured inputs
- Statistical Validation: Applies confidence scoring and variation analysis akin to model evaluation techniques

Problem Statement

Consumer-facing teams frequently handle large volumes of unstructured communication, often relying on manual interpretation that varies by agent experience and training. In settings where data is proprietary and tooling is restricted to Microsoft Excel; there is a pressing need for a standardized and auditable system that ensures consistent classification of consumer concerns while maintaining interpretability and compliance.

Scope & Input Types

The framework processes three primary types of consumer communication:

- **Scanned Correspondences**: Manually interpreted and summarized by agents based on physical or image-based documents
- **Email Correspondences**: Direct consumer language, often detailed and multi-paragraph, allowing for comprehensive expression
- **SMS Correspondences**: Direct consumer language in a concise format, typically limited by character count and structure

Stakeholders & Intended Usage

- Quality Team: Validates agent interpretations and ensures consistency
- Operations Team: Guides agents and reduces repetitive queries through structured insights

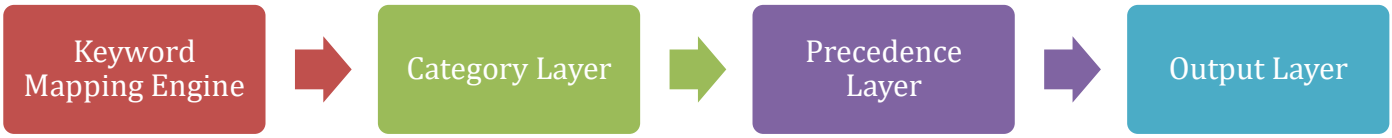
Methodology

- Defined 25 correspondence categories
- Mapped 400 keywords across categories
- Applied keyword logic in Excel to extract and classify them
- Generated structured outputs including:
  - Source type
  - Consumer concern and detailed concern
  - Communication preferences
  - Document-based consumer requests

Subject matter experts from the quality control team were engaged to identify relevant keywords.

Outcome Derivation Logic

Before final outcomes are assigned, the framework follows a structured multi-step logic to ensure accuracy and precedence in category selection:



1. Keyword Mapping Engine:
→ Keywords are matched against the input text to highlight relevant categories

2. Category Layer:
→ Categories are flagged based on the presence of mapped keywords

3. Precedence Layer:
→ In cases where multiple categories are highlighted, precedence rules are applied
Example: If the consumer states, "do not call me" and "do not contact me", this triggers both the "No Phone Outreach" and "Full Contact Restriction" flags respectively.
→ In such cases, the broader **Non-Engagement Directive** supersedes the **Call Suppression Request**, as it reflects a comprehensive withdrawal from all communication channels.

4. Outcome Layer:
→ The highest-precedence category is selected as the final outcome, ensuring clarity and consistency

Visual Example of Keyword Mapping Engine Functionality:

	12	2	3	4
Sr No	**Category 1**	**Category 2**	**Category 3**	**Category 4**
1	Keyword 1	Keyword 1		Keyword 1
2	Keyword 2	Keyword 2	Keyword 1	Keyword 2
3	Keyword 3		Keyword 2	Keyword 3
4	Keyword 4		Keyword 3	Keyword 4

Category 1 is configured with a set of 12 distinct keywords, enabling broad detection coverage.
Category 2, by contrast, is defined by 2 specific keywords, indicating a narrower scope.
This pattern continues across categories, with each category mapped to a unique set of keywords based on relevance and specificity.

Visual Example of the Category Layer:

Sr No	Consumer Corespondence	Category 1	Category 2	Category 3	Category 4
1	XXXXXXX		Category 2		
2	XXXXXXX	Category 1			Category 4
3	XXXXXXX				
4	XXXXXXX		Category 2	Category 3	

Detected keywords trigger associated categories based on predefined mappings.
Multiple categories may be flagged simultaneously, forming the basis for further prioritization and outcome selection.

Excel Tips for Efficient Keyword Mapping & Auditing:

1. **Use Separate Tabs for Clarity & Performance:**
   Mapping all keywords in a single sheet can slow down Excel. Instead, store keywords in a dedicated tab and use the main data sheet to highlight matched categories. This improves clarity and helps ensure accurate flagging.
2. **Streamline Keyword Auditing:**
   During audits, you may delete certain keywords, leaving empty cells. To avoid manually shifting keywords and using memory-heavy formulas, consider this efficient alternative:

```
=IF(SUM(--ISNUMBER(SEARCH(FILTER(References!$A$2:$A$40,References!$A$2:$A$40<>""),H2)))>0, "Category 1", "")
```

H2 = input text, References = keyword sheet

This formula filters out empty cells and checks for keyword matches, returning the appropriate category when found.

Keyword Optimization Process

This process involves refining and strategically selecting search terms to enhance visibility, relevance, and ranking in digital content.

1. Keyword Identification & Sorting - Initial keyword list sorted by frequency of occurrence
2. Heatmap Visualization - Used to visualize keyword frequency and distribution
3. Weight Distribution Analysis - Top keywords initially carried 70–80% of identification weight
4. Contextual Expansion - Added synonyms and contextual variations
5. Keyword Replacement & Balancing - Adjusted top keywords to carry ~30% weight, supported by secondary terms
6. Final Audit - Re-audited categories for balance and confidence

Visual Example of Keyword Heat Map:

Category 1	Count	%	Category 2	Count	%	Category 3	Count	%
Keyword 1	41	24%	Keyword 1	530	38%	Keyword 1	1446	21%
Keyword 2	30	18%	Keyword 2	479	34%	Keyword 2	928	13%
Keyword 3	24	14%	Keyword 3	197	14%	Keyword 3	688	10%
Keyword 4	19	11%	Keyword 4	78	6%	Keyword 4	674	10%
Keyword 4	13	8%	Keyword 4	22	2%	Keyword 4	592	8%
Keyword 6	12	7%	Keyword 6	22	2%	Keyword 6	410	6%
Keyword 7	8	5%	Keyword 7	21	2%	Keyword 7	326	5%

Keyword Heat Map *visually represents the frequency and distribution of keywords across a dataset, highlighting areas of high relevance or concentration to aid in pattern recognition and content analysis.*

Visual Example of Keyword Identification Sheet:

Count of Detects	70	43	5
Category 1	Keyword 1	Keyword 2	Keyword 3
Correspondence 1	314		
Correspondence 2		2077	
Correspondence 3	22	840	
Correspondence 4	314		
Correspondence 5	47		2136
Correspondence 6	1741		

For long emails, it was hard to tell which keyword triggered the category. So, a simple sheet was created to show where each keyword appears in the message.

Excel Tip:

1. Embedding sample size calculations and corresponding audit scores within the same worksheet can significantly impact Excel's performance. To maintain efficiency and clarity, it is recommended to separate audit results into distinct sheets.
2. Additionally, enabling partial or manual calculation modes allows Excel to compute formulas only after task completion, reducing processing overhead during data entry and analysis

Sampling Methodology

The process uses a statistically grounded sample size formula to ensure reliable data validation, applying 99% confidence for high-priority categories and 95% for others. It incorporates variance analysis through mean, standard deviation, and coefficient of variation (CV) to check for strong consistency across samples.

Confidence Scoring & Statistical Validation

Sample Size Formula:

n = ROUND((Z^2 * p * (1 - p) / e^2) / (1 + ((Z^2 * p * (1 - p) / e^2 - 1) / N)), 0)

Where:

- Z: Z-score (2.576 for 99%, 1.96 for 95%)
- p: Estimated proportion (0.5)
- e: Margin of error (0.05)
- N: Population size per category

Confidence Levels:

- 99% confidence: 9 high-priority categories
- 95% confidence: 36 remaining categories

Variance Analysis:

- Calculated mean, standard deviation, and coefficient of variation (CV)

Visual Example of Audit Sampling Mechanism

Total Found	17	200	458	2	3434	1944
Confidence level	95%	95%	95%	95%	99%	99%
z score	1.96	1.96	1.96	1.96	2.56	2.56
Samples needed	15	131	208	1	549	489
Samples correct	15	124	189	0	524	474
Sample proportion	1.00	0.95	0.91	0.00	0.95	0.97
Finite pupulation correction	35%	59%	74%	100%	92%	87%
Standard error	0.00%	1.96%	2.00%	0.00%	0.89%	0.78%
Margin of error	0.00%	3.85%	3.92%	0.00%	2.28%	2.00%
Upper Value	100.00%	90.81%	86.95%	0.00%	93.17%	94.94%
Lower Value	100.00%	98.51%	94.78%	0.00%	97.72%	98.93%
Keywords >	20	31	23	2	59	85
Total Categories >	Category 1	Category 2	Category 3	Category 4	Category 5	Category 6

*A **Z-score** (or standard score) tells you how many standard deviations a data point is from the mean of a dataset. It is commonly used in confidence intervals and hypothesis testing.*

In the context of confidence intervals, the upper and lower limits define the range within which the true population parameter is expected to fall, based on your sample data.

Example: You can be 95% confident that the true value lies between these two limits.

Visual Example of Variance across input types

		Standard Deviation		Mean		Variation across Input Types	
Category	Confidence Level	Lower Value	Upper Value	Lower Value	Upper Value	Lower Value	Upper Value
Cagegory 1	95.0%	0.19	0.08	97%	100%	2%	0%
Cagegory 2	95.0%	0.58	0.58	67%	87%	87%	87%
Cagegory 3	99.0%	0.04	0.01	97%	99%	4%	1%
Cagegory 4	99.0%	0.03	0.01	99%	100%	2%	0%
Cagegory 5	99.0%	0.04	0.01	95%	99%	4%	1%
Cagegory 6	99.0%	0.04	0.01	98%	99%	4%	1%
Cagegory 7	99.0%	0.00	0.00	100%	100%	0%	0%
Cagegory 8	99.0%	0.02	0.01	96%	99%	2%	1%
Cagegory 9	95.0%	0.09	0.06	93%	97%	10%	6%
Cagegory 10	95.0%	0.20	0.14	81%	90%	25%	15%

Interpretation

Category 1

- ***Confidence:*** *95%*
- ***Mean Accuracy:*** *97–100%*
- ***Variation:*** *Low (2–0%)*
- ***Standard Deviation:*** *Moderate (0.19–0.08) → Reliable with slight variability*

Category 4

- ***Confidence:*** *99%*
- ***Mean Accuracy:*** *99–100%*
- ***Variation:*** *Minimal (2–0%)*
- ***Standard Deviation:*** *Very Low (0.03–0.01) → Highly consistent and precise*

Excel Tip:

Allocate separate worksheets for each input method to minimize processing load and enhance performance

Learnings and Outcome

- **Study Scope:** The analysis encompassed 89,000 correspondences, with approximately 22% audited based on the established sampling methodology.
- **Keyword Evolution:** The initial set of ~400 keywords expanded to 692 through iterative optimization—this included additions, removals, and strategic reallocation across categories.
- **Category Refinement:** Due to input source complexity, the number of categories increased from the originally planned 25 to a final count of 45. Of these, 36 were consistently applied across all input types, while 9 were exclusive to certain sources. Notably, some of these exclusive categories also supported data cleansing and preparation for classification.
- **Reliability & Performance:** The framework exhibited robust baseline reliability across most categories. Confidence scores and consistency metrics improved steadily with each reassessment and audit cycle.
- **Stability Metrics:** Coefficient of variation remained within acceptable thresholds, underscoring stable classification performance across varied input types. These metrics continue to strengthen with ongoing refinements to keyword sets and category mapping.
- **High-Confidence Audits:** Several priority categories underwent auditing at a 99% confidence level, with lower bounds exceeding 90%—demonstrating high accuracy and classification consistency.
- **Validation Outcome:** The successful completion of initial testing and parameter tuning affirmed the framework's readiness for broader deployment and scalability.

Future Integration

- The framework has been integrated into Quality Control and Operations workflows, with initial deployment serving as a beta testing phase within the quality team to assess usability and effectiveness.
- Addition of "Incorrect Categorization" field within the operational flow, enabling streamlined review of keyword and category summaries. This enhancement supports the detection of potential inconsistencies and simplifies issue identification during routine analysis.
- Ongoing validation and keyword refinement are planned, alongside expansion of scope to include additional categories and use cases.