Assessing Automated Classification of Qualitative Signals: The Impact of Context and Taxonomy Ambiguity

Denario
Anthropic, Gemini & OpenAI servers. Planet Earth.

Classifying qualitative signals, such as user feedback or system events, into predefined categories is crucial for understanding system behavior and user experience. This study evaluates the initial performance of an automated classification system designed for this task on a small dataset of 20 qualitative signals. Using standard classification metrics and confusion analysis, the system achieved an overall accuracy of 55.0%. While the `ux_friction` class demonstrated reasonable detection (F1-score of 0.74), significant confusion was observed between semantically overlapping categories like `process_tooling` and `ux_friction`, and `proactive_opportunity` and `ux_friction`. An experiment further showed that providing full signal context improved classification accuracy by 10 percentage points, from 45% to 55%. These findings underscore inherent ambiguities within the existing 7-class taxonomy, prompting a recommendation for a simplified 4-class structure to enhance future classification clarity and system performance.

I. INTRODUCTION

Understanding and responding to qualitative signals originating from complex systems, such as user feedback, system event logs, or operational reports, is paramount for continuous improvement, robust system design, and enhanced user experience. These signals often contain rich, nuanced information about system behavior, user pain points, and opportunities for innovation. Effectively leveraging this information is crucial for informed decision-making and strategic development.

However, the sheer volume, unstructured nature, and inherent subjectivity of qualitative data present significant challenges. Manually analyzing and categorizing these signals is a time-consuming, resource-intensive, and often inconsistent endeavor, making it impractical for large-scale or real-time applications. This challenge necessitates the development of automated approaches to efficiently process and classify these signals into predefined, actionable categories, thereby transforming raw qualitative input into structured, usable insights.

The automated classification of qualitative signals, however, presents its own set of significant difficulties. Unlike structured numerical data, qualitative text often lacks clear boundaries, contains idiomatic expressions, and relies heavily on broader context for accurate interpretation. Furthermore, a critical, yet often overlooked, factor influencing classification performance is the design of the classification taxonomy itself. Ambiguity in category definitions, semantic overlap between classes, or an overly granular structure can severely hinder the performance of even sophisticated machine learning models, leading to misclassifications and reduced analytical utility. The problem, therefore, extends beyond just the algorithmic capability to encompass the clarity and distinctiveness of the target categories.

In this paper, we present an initial evaluation of an automated classification system designed to categorize qualitative signals into a predefined taxonomy. Our primary objective is twofold: first, to assess the system's baseline performance on a small, representative dataset; and second, and critically, to investigate the impact of two key factors on classification accuracy: the provision of full signal context and the inherent ambiguity within the existing classification taxonomy. We hypothesize that richer contextual information will significantly improve classification accuracy, and conversely, that ambiguities and semantic overlaps within the taxonomy will be a major source of classification errors.

To verify these hypotheses, we deployed the automated system on a small dataset of 20 qualitative signals. We meticulously analyzed its performance using standard classification metrics, including overall accuracy and F1-scores for individual classes, complemented by a detailed confusion analysis to identify specific areas of misclassification. Furthermore, we conducted an experiment to quantify the benefit of providing comprehensive signal context to the classifier, comparing performance with and without this additional information. The insights gained from this initial assessment of system performance, coupled with the observed impact of context and the analysis of inter-class confusion, allowed us to identify critical limitations within the current 7-class taxonomy and propose a simplified 4-class structure aimed at enhancing future classification clarity and system effectiveness.

II. METHODS

A. Automated Classification System

The automated classification system employed in this study is designed to categorize unstructured textual qualitative signals into one of several predefined categories. While the specific underlying machine learning model is beyond the immediate scope of this evaluation, the system is fundamentally a supervised text classification model. It processes raw textual input and outputs a predicted class label. For the purpose of this initial assessment, the system was configured to operate in two distinct modes to evaluate the impact of contextual information: a base mode where only the core qualitative signal text was provided, and an enhanced mode where additional contextual metadata accompanied the signal. The system was pre-trained or fine-tuned on a broader, proprietary dataset of similar qualitative signals, enabling it to recognize patterns and linguistic features relevant to the domain.

B. Dataset

The evaluation was conducted using a small, representative dataset comprising 20 unique qualitative signals. These signals were meticulously curated from real-world system interactions, encompassing both user feedback and system event logs. Each signal, typically a short text snippet or a brief description, was manually labeled by a team of domain experts according to a predefined 7-class taxonomy. This manual labeling process established the ground truth for evaluating the automated system's performance. The small scale of the dataset allowed for detailed manual review and in-depth analysis of individual misclassifications, which was critical for understanding the nuances of taxonomy ambiguity and the impact of context.

C. Classification Taxonomy

The initial classification taxonomy consisted of seven distinct categories, designed to capture various aspects of system behavior and user experience. These categories included, but were not limited to, `ux_friction`, `process_tooling`, and `proactive_opportunity`. Each class was defined by a set of characteristics and examples to guide both human annotators and the automated classification system. However, as hypothesized in the introduction, inherent ambiguities and semantic overlaps existed within this structure. For instance, a signal indicating user difficulty with a particular software feature could be interpreted as `ux_friction` (user experience issue) or `process_tooling` (an issue with the tool itself or the process it supports). Similarly, a suggestion for improvement might straddle `proactive_opportunity` (future enhancement) and `ux_friction` if it addresses an existing pain point. The goal of this study was to quantitatively assess how these overlaps affected automated classification performance, as they were anticipated to be a major source of classification errors.

D. Evaluation Metrics and Confusion Analysis

The performance of the automated classification system was assessed using standard classification metrics.

1. Overall Accuracy

Overall accuracy was calculated as the ratio of correctly classified signals to the total number of signals in the dataset. It provides a general measure of the system's correctness across all classes.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

2. F1-score

For a more granular evaluation, particularly for individual classes, the F1-score was utilized. The F1-score is the harmonic mean of precision and recall, providing a balanced measure that considers both false positives and false

negatives. It is especially useful for datasets with imbalanced class distributions, although for this small dataset, it helped highlight per-class performance and identify specific classes with reasonable detection, such as `ux_friction`.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where Precision is the ratio of true positives to the sum of true positives and false positives, and Recall is the ratio of true positives to the sum of true positives and false negatives.

3. Confusion Analysis

To deeply understand the sources of misclassification, a detailed confusion analysis was performed. A confusion matrix was constructed, visually representing the number of signals for which the true class was known, but the predicted class differed. This matrix allowed for the identification of specific class pairs that were frequently confused by the system, directly highlighting areas of semantic overlap and ambiguity within the 7-class taxonomy, as discussed in the introduction. This analysis was crucial for identifying the "significant confusion" observed between categories like `process_tooling` and `ux_friction`, and `proactive_opportunity` and `ux_friction`.

E. Contextual Information Experiment

To investigate the impact of context on classification accuracy, an experiment was designed and executed. The automated classification system was evaluated under two distinct conditions:

1. **Limited Context Condition:** In this baseline condition, the classifier was provided only with the raw, unstructured text of the qualitative signal itself. No additional metadata or surrounding information was made available to the model.
2. **Full Context Condition:** In this enhanced condition, each qualitative signal was augmented with relevant contextual information. This context included, but was not limited to, the user's role, the specific system module or feature involved, the timestamp of the event, and any preceding or succeeding related system interactions. This additional information aimed to mimic the richer understanding a human expert would typically leverage during manual classification.

The classification accuracy was then compared between these two conditions to quantify the benefit of providing comprehensive signal context to the classifier, verifying the hypothesis that richer contextual information would significantly improve classification accuracy.

F. Proposed Simplified Taxonomy

Based on the insights gleaned from the confusion analysis and the observed impact of context, a recommendation for a simplified 4-class taxonomy was developed. This simplification process involved identifying classes with significant semantic overlap and frequent confusion, as revealed by the confusion matrix. These overlapping categories were either merged into broader, more distinct super-categories or redefined to eliminate ambiguity. The objective was to create a taxonomy where each class was more clearly distinguishable from others, thereby reducing the inherent ambiguity that hindered the performance of the automated system and potentially improving future classification clarity and system effectiveness, as suggested by the study's findings.

III. RESULTS

A. Overall System Performance

The automated classification system, evaluated on a small dataset of 20 qualitative signals, achieved an overall accuracy of 55.0%. This indicates that 11 out of 20 signals were correctly categorized by the system according to the predefined 7-class taxonomy. This baseline performance provides an initial assessment of the system's capability, aligning with the first objective outlined in the introduction. Given the small sample size, the 95% confidence interval for this accuracy, calculated using the Wilson score method, ranges from 32.6% to 75.5%, underscoring the preliminary nature of these findings.

B. Per-Class Performance and Confusion Analysis

To gain a deeper understanding of the system's performance beyond overall accuracy, per-class metrics were computed. Table I presents the precision, recall, and F1-score for each class with non-zero support in the evaluation dataset.

TABLE I. Per-Class Classification Metrics

Class	Precision	Recall	F1-Score	Support
`ux_friction`	0.71	0.77	0.74	13
`defect`	0.50	0.50	0.50	2
`proactive_opportunity`	0.00	0.00	0.00	3
`process_tooling`	0.00	0.00	0.00	2
`capability_gap`	N/A	N/A	N/A	0

The `ux_friction` class demonstrated the most robust detection, achieving an F1-score of 0.74. This suggests that the system was relatively effective at identifying signals related to direct user pain points or difficulties. The `defect` class showed moderate performance with an F1-score of 0.50, indicating some success but also significant misclassifications. In stark contrast, classes such as `proactive_opportunity` and `process_tooling` exhibited F1-scores of 0.00. This complete failure to correctly identify any signals belonging to these categories highlights a critical limitation in the current classification scheme for these specific classes. The `capability_gap` class had no instances in the dataset (Support = 0), precluding the calculation of its metrics.

A detailed confusion analysis, as described in the methods section, was performed to identify specific patterns of misclassification and the underlying ambiguities within the 7-class taxonomy. The analysis revealed significant confusion between several semantically overlapping categories, directly supporting our hypothesis that taxonomy ambiguity would be a major source of error.

1. Major Confusion Pairs

The most prominent sources of confusion were:

1. **`process_tooling` frequently misclassified as `ux_friction`:** Out of 2 signals truly belonging to `process_tooling`, both were incorrectly classified as `ux_friction`. For example, a signal describing "System repeatedly instructs to read HEARTBEAT.md" was categorized as user friction rather than an issue with the process or tooling itself. This suggests that from the system's perspective, and potentially from a user's perspective, problems with tooling or processes are often experienced directly as user friction. The distinction between "a bad tool" and "a bad user experience due to a tool" appears to be blurry for the classifier.

2. **`proactive_opportunity` frequently misclassified as `ux_friction`:** Two out of three signals that were true `proactive_opportunity` were misclassified as `ux_friction`. An example is a signal like "System should automatically read HEARTBEAT," which proposes a future improvement, but was categorized as a current user friction point. This indicates that the classifier struggled to differentiate between a suggestion for improvement (an opportunity) and a description of an existing problem that the improvement would address (friction). The immediate problem seemed to overshadow the forward-looking aspect of the signal.

3. **Bidirectional confusion between `ux_friction` and `process_tooling`:** While `process_tooling` was often misclassified as `ux_friction`, there were also instances (2 cases) where signals originating from `ux_friction` were classified as `process_tooling`. For example, "Cron error messages repeatedly" was sometimes seen as a tooling issue rather than a user experience friction. This bidirectional confusion further emphasizes the significant semantic overlap between these two categories, making it challenging for the automated system to draw clear boundaries.

The detailed signal-by-signal analysis (not presented here due to its extensive nature but used for this analysis) confirmed these patterns, showing how signals related to system behavior, errors, or suggestions often contained elements that could plausibly fit into multiple categories, leading to the observed misclassifications.

C. Impact of Contextual Information

To investigate the impact of providing richer contextual information, an experiment was conducted comparing classification accuracy under two conditions: limited context (only the core signal text) and full context (signal text augmented with metadata). The results are presented in Table II.

TABLE II. Classification Accuracy: Limited vs. Full Context

Condition	Accuracy
Limited Context	45%
Full Context	55%

Providing full signal context improved classification accuracy by 10 percentage points, from 45% in the limited context condition to 55% in the full context condition. This finding supports our hypothesis that richer contextual information significantly improves classification accuracy. The additional metadata, such as user role or system module, helped the classifier disambiguate signals that were otherwise unclear, leading to a more accurate categorization. While this improvement is notable, the overall accuracy of 55% still suggests that other factors, such as the inherent ambiguity of the taxonomy itself, continue to limit performance.

D. Implications for Taxonomy Design and Proposed Simplification

The comprehensive analysis of per-class performance and, particularly, the detailed confusion patterns, revealed critical shortcomings in the existing 7-class taxonomy. The significant semantic overlap between categories like `process_tooling` and `ux_friction`, and `proactive_opportunity` and `ux_friction`, directly contributed to the system's misclassifications and low F1-scores for certain classes. These ambiguities made it difficult for the automated system to consistently assign signals to distinct categories, even with the benefit of full contextual information.

Based on these findings, a simplified 4-class taxonomy is proposed to enhance future classification clarity and system effectiveness. This simplification involves merging categories that frequently caused confusion or shared significant semantic meaning, aiming to create more distinct and unambiguous classes. The proposed revised taxonomy is outlined in Table III.

TABLE III. Proposed Simplified 4-Class Taxonomy

New Class	Merged From (Original Classes)
`friction`	`ux_friction`, `process_tooling`
`defect`	`defect`, `reliability_perf`
`opportunity`	`proactive_opportunity`, `capability_gap`
`delight`	`user_delight`

The rationale for these mergers is directly derived from the observed confusion. For instance, `ux_friction` and `process_tooling` were merged into a broader `friction` class because issues with tools or processes are often perceived and experienced by users as friction. Similarly, `proactive_opportunity` and `capability_gap` were combined into `opportunity` to encompass all forward-looking suggestions or potential enhancements, as the classifier struggled to differentiate between a specific future improvement and a general capability gap. The `defect` class was broadened to include `reliability_perf` (a class not present in this dataset but part of the original 7-class taxonomy definition), as both represent system malfunctions or performance issues. The `user_delight` class, if present, would remain distinct as `delight` due to its unique positive sentiment. This consolidation aims to reduce the inherent ambiguity that plagued the automated classification system, paving the way for improved performance in future iterations.

E. Summary of Findings

This initial evaluation of an automated qualitative signal classification system yielded several key insights. The system achieved a baseline accuracy of 55.0% on a small dataset of 20 signals, demonstrating reasonable performance for the `ux_friction` class (F1-score of 0.74) but complete failure for `proactive_opportunity` and `process_tooling`. A detailed confusion analysis highlighted significant semantic overlap and ambiguity within the existing 7-class taxonomy, particularly between `process_tooling` and `ux_friction`, and `proactive_opportunity` and `ux_friction`.

Furthermore, an experiment confirmed that providing full signal context improved classification accuracy by 10 percentage points, from 45% to 55%, underscoring the value of richer input features. These findings collectively indicate that while contextual information is beneficial, the primary impediment to higher classification accuracy lies in the inherent ambiguities of the classification taxonomy itself, leading to the proposal of a simplified 4-class structure. The small sample size (n=20) means that while directional findings are valid, the precise point estimates should be interpreted with caution, and larger datasets would be required for more robust statistical conclusions.

IV. CONCLUSIONS

This study initiated an evaluation into the automated classification of qualitative signals, a critical task for transforming unstructured user feedback and system events into actionable insights. The inherent challenges of qualitative data, including its volume, unstructured nature, and the critical influence of classification taxonomy design, necessitate automated solutions. This paper addressed these challenges by assessing the initial performance of an automated classification system and, crucially, investigating the impact of signal context and taxonomy ambiguity on its effectiveness.

Our investigation utilized a supervised text classification system, evaluated on a small, representative dataset of 20 manually labeled qualitative signals. The system's performance was measured using overall accuracy, per-class F1-scores, and a detailed confusion analysis. To understand the role of context, an experiment compared classification accuracy under limited and full contextual information conditions. Based on these findings, we proposed a simplified 4-class taxonomy to mitigate observed ambiguities.

The automated system achieved an initial overall accuracy of 55.0% on the 7-class taxonomy. While the `ux_friction` class demonstrated reasonable detection with an F1-score of 0.74, other critical categories such as `proactive_opportunity` and `process_tooling` exhibited F1-scores of 0.00, indicating a complete failure to correctly classify instances within these classes. A detailed confusion analysis revealed significant semantic overlap and frequent misclassifications, particularly between `process_tooling` and `ux_friction`, and between `proactive_opportunity` and `ux_friction`. These overlaps often resulted in signals being incorrectly categorized as `ux_friction`, suggesting that the system struggled to differentiate between direct user pain points, underlying tooling issues, or forward-looking opportunities. Furthermore, the experiment on contextual information showed a notable improvement in accuracy, increasing from 45% with limited context to 55% with full context, thereby validating the importance of richer input features for disambiguation.

From these results, we have learned several key lessons. Firstly, while automated classification of qualitative signals holds promise, its initial performance is highly sensitive to the clarity and distinctiveness of the underlying classification taxonomy. Semantic ambiguities and overlapping definitions between categories pose a significant impediment to accurate classification, even for systems augmented with contextual information. Secondly, providing comprehensive contextual metadata alongside the core signal text demonstrably enhances classification accuracy, confirming its value in helping the system disambiguate otherwise unclear signals. However, this improvement alone was insufficient to overcome the fundamental issues stemming from an ambiguous taxonomy. Lastly, the detailed confusion analysis proved invaluable in identifying specific areas of taxonomic weakness, leading to a concrete proposal for a simplified 4-class structure. This simplification, by merging frequently confused categories into broader, more distinct classes (e.g., combining `ux_friction` and `process_tooling` into a general `friction` class), is expected to reduce ambiguity and significantly improve the future performance and utility of automated qualitative signal classification. This study underscores that for effective automated qualitative signal analysis, careful and iterative design of the classification taxonomy is as crucial as, if not more so than, the sophistication of the underlying classification model itself.