# Intent Detection for Customer Queries: Dataset Annotation and Validation

**Meghana Maringanti**

## Abstract

Intent detection is essential for improving customer support efficiency. Detecting Intent can help us build systems to automate responses, prioritize tasks, and enhance the overall user experience. This project focuses on creating an annotated dataset of customer support queries. The queries are categorized into five distinct intent classes namely, Inquiry, Request for Action, Technical Issue, Complaint, Praise, and Unclear. The dataset consists of 179 queries collected from public data, online reviews, and self-authored examples. Annotation is performed by 2 annotators using the Label Studio tool, complying with the category definitions to ensure consistency. To validate the dataset, quality check is conducted using Cohen's Kappa to measure inter-annotator agreement.

## 1 Introduction

In customer service, accurately identifying user intent can greatly improve support operations. It helps automate responses, prioritize urgent tasks, and enhance the overall customer experience. Intent detection involves classifying a customer's query based on their needs. This makes it possible to provide targeted and efficient resolutions.

This project focuses on creating a high-quality annotated dataset for intent detection in customer queries. The dataset is divided into six distinct categories that represent common user intents: Inquiry, Request for Action, Technical Issue, Complaint, Praise and unclear.

To ensure the dataset is reliable, a quality check is performed using Cohen's Kappa. This statistical measure evaluates how consistent the annotations are among different annotators.

Additionally, the dataset's validity is tested by training a model for intent detection. The model's performance is then assessed on the annotated queries.

The dataset and methodology from this project aim to support research and applications in customer intent detection. This has the potential to improve automation, prioritize tasks, and enhance the customer experience.

## 2 Dataset Preparation

### 2.1 Sources

The dataset consists of 179 queries compiled from three sources:

1. Publicly available data. - from sources such as blog/Twitter posts
2. Customer reviews observed online. - from sources such as Amazon/Flipkart reviews
3. Self-authored examples

By collecting the data from different sources I aim to create a dataset that reflects real-world customer queries across various contexts.All queries are initially stored in a text file and subsequently converted into a CSV file for easier formatting and structured organization.

### 2.2 Categories

1. **Inquiry**

   Queries where the user is explicitly seeking information, clarification, or details.

   Examples :

   1. How can I update my email address on my account?

   2. How do I delete my account permanently?

2. **Request for Action**

   Queries that involve asking for a specific action or service.

Examples :

1. Can you expedite the delivery for my order?

2. Please add a note to my order to deliver after 5 PM.

3. **Praise** Positive feedback or compliments about products, services, or experiences.

   Examples :

   1. I love the variety of items available on your platform.

   2. Your attention to detail in product design is amazing.

4. **Complaint** Expressions of dissatisfaction, frustration, or negative experiences.

   Examples :

   1. It is not worth the price at all. The quality does not match the cost.

   2. The product broke within a few days of use. Clearly, it is not durable.

5. **Technical Issues** Queries describing technical malfunctions, errors, or troubleshooting needs.

   Examples :

   1. I cannot log into my account despite using the correct password.

   2. The app keeps freezing when I try to access my previous orders.

6. **Unclear** If intent cannot be discerned, mark it as Unclear and provide a comment explaining why.

   Example: "I want to upgrade my plan!"

   Here it is not clear whether they are asking for details on how to upgrade or asking customer support to upgrade the plan.

### 2.3 Annotation Guidelines

To ensure consistency and reliability in annotation, detailed guidelines were provided to annotators. These guidelines included:

1. **Category Definitions**: Each category was clearly defined, with specific criteria outlining when it should be selected.

2. **Examples:** Annotators were provided with representative examples for each category to aid in accurate labeling.

3. **Handling Ambiguity:** Clear instructions were given on how to deal with ambiguous queries, including strategies to resolve uncertainties.

4. **Multiple Categories:** For queries that could fit into multiple categories, annotators were directed to select the most dominant intent.

**Key principles:**

- Each query is labeled with only one primary intent category.
- When a query fits multiple categories, the most dominant intent is chosen.
- Definitions and examples are provided to guide annotators.

### 2.4 Annotation Process

Annotation is performed using **Label Studio**, a widely used open-source annotation tool. It provides a user-friendly interface for labeling data and supports efficient categorization of queries.

### 2.5 Statistics

The dataset consists of 179 customer support queries categorized into six categories: Inquiry, Request for Action, Technical Issue, Complaint, Praise, and Unclear.

Table 1: Query Distribution by Category

| Category | Count | Percentage |
|---|---|---|
| Inquiry | 63 | 35% |
| Request for Action | 31 | 17% |
| Complaint | 31 | 17% |
| Praise | 30 | 17% |
| Technical Issue | 22 | 12% |
| Unclear | 2 | 1% |

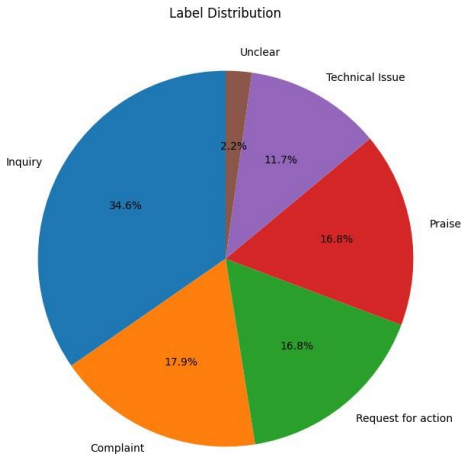


Figure 1: Label-distribution

## 3 Quality Check

To check the reliability of the annotated dataset, we measured inter-annotator agreement using Cohen's Kappa.

This metric was chosen because it accounts for the agreement occurring by chance, providing a robust evaluation of consistency between annotators. The agreement was computed by having two annotators independently label a subset of queries, ensuring their decisions aligned with the predefined category definitions.

Cohen's Kappa (k) evaluates the level of agreement between two annotators, where:

- k = 1 indicates perfect agreement.
- K = 0 indicates No agreement beyond chance.

### 3.1 Threshold

The following thresholds guided our interpretation of

- 0.61–0.80: Substantial agreement.
- 0.81–1.00: Almost perfect agreement.

For this project, an acceptable threshold was set at k > 0.70 to ensure substantial agreement.

### 3.2 Results

The computed Cohen's Kappa for the annotated subset was k = 0.91, which indicates Almost perfect agreement. This result validates the consistency of the annotations, confirming that the dataset is reliable for use in intent detection tasks.

## 4 Validation

To evaluate the dataset and measure its utility for intent detection, we validated it using the "roberta-large-mnli" model

This model was chosen due to its strong performance on a wide range of natural language inference tasks, making it well-suited for understanding and identifying intent in diverse contexts. Its pre-trained architecture enables effective zero-shot applications.

### 4.1 Model Performance

Using the model we predicted labels for the dataset. These predictions were compared with the annotated labels to assess model performance. The evaluation metrics included overall accuracy, and precision, recall, F1-scores per label to highlight category-specific performance.

The model achieved an overall accuracy of 0.53, reflecting its varying performance across different intent categories. The detailed per-label scores are as follows:

Table 2: Detailed Per-Label Scores

| Category | Precision | Recall | F1-Score |
|---|---|---|---|
| Inquiry | 0.79 | 0.42 | 0.55 |
| Request for Action | 0.44 | 0.93 | 0.60 |
| Complaint | 0.83 | 0.16 | 0.26 |
| Technical Issue | 0.63 | 0.48 | 0.54 |
| Praise | 1.00 | 0.87 | 0.93 |
| Unclear | 0.00 | 0.00 | 0.00 |

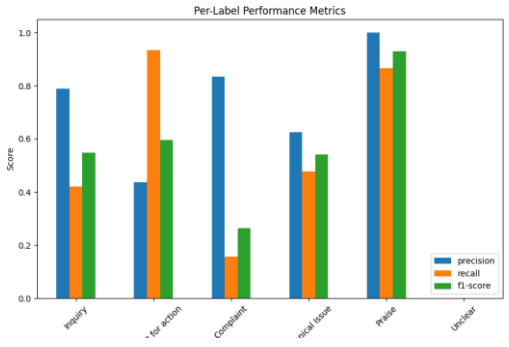


Figure 2: Scores for each Label.

### 4.2 Analysis

The validation results highlight key observations:

- **High Performance - "Praise":**

  The "Praise" category achieved the highest scores across all metrics, with an F1-score of 0.93, indicating the ease of classification.

- **Challenges in "Unclear":**

  The model failed to classify the "Unclear" category, scoring zero in all metrics. This suggests insufficient distinguishing features in the queries.

**Mixed Results for Other Categories:**

1. Inquiry:

   High precision (0.79) but low recall (0.42), indicates missed instances.

2. Request for Action:

   High recall (0.93) but moderate precision (0.44), indicates over-classification into this category.

3. Complaint:

   Precision (0.83) far exceeded recall (0.16), showing confusion with other categories.

4. Technical Issue: Balanced precision and recall, but overall F1-score (0.54) suggests room for improvement.

## Limitations

This project has a few limitations that can impact the applicability of this dataset to the real-world customer interactions.

- **Dataset**:

  The dataset we prepared contains only 179 queries distributed across 6 categories. This may provide a foundation for analysis but it may fail to fully represent the diverse range of customer intents across various industries.

- **Data quality**:

  Though the Cohen's kappa value is 0.91, which indicating good data quality, we can observe that most queries do not pose any ambiguous conditions. So, even though the dataset performs well, there is a need to add more queries that focus on edge cases and ambiguous scenarios.

## Ethics Statement

- No sensitive or personally identifiable information is included; the dataset uses public and self-authored data only.

- This dataset is for research and educational purposes and should not be used in production without additional validation and expansion.

- Efforts will be made in future iterations to represent more diverse user intents.

## Conclusion

In this study, we developed and validated a dataset for intent detection in customer queries, consisting of 179 queries distributed across six categories. Annotation reliability was confirmed with a Cohen's Kappa value of 0.91, indicating almost perfect agreement among annotators.

Validation using the "roberta-large-mnli" model yielded an overall accuracy of 0.53. While the model performed well in certain categories, such as "Praise," it faced challenges in others, particularly the "Unclear" category.

These findings emphasize the need for further refinement of category definitions and expansion of the dataset to include a broader range of ambiguous and diverse queries.

## Future Scope

1. Dataset Expansion:

   Add more queries, particularly focusing on edge cases and ambiguous scenarios, to improve the dataset's diversity.

2. Category Refinement:

   Revisit and refine category definitions to reduce overlap and ambiguity between intents.

3. Integration with Real-World Data:

   Incorporate real-world customer queries from live support systems to make the dataset more representative of practical scenarios.