Classifying scope of queries on the ”Intent Classification and Out-of-Scope Prediction dataset”

I. Dataset

The dataset used in this project consists of queries labeled across 151 categories, with a focus on determining whether a query is out-of-scope (OOS). The dataset contains a total of 23,700 samples, divided into training, testing, and validation sets. The OOS category contains 1200 samples, making it one of the largest single categories. This dataset is designed to identify queries that fall outside the supported domains in natural language processing systems, commonly applied in virtual assistants and customer service platforms [1]. The original source of this dataset is [Dataset Reference].

II. Classification Pipeline

The classification pipeline begins with preprocessing steps aimed at improving the quality of the text data. First, stop words are removed to focus on meaningful content, utilizing a custom stop-word list. Following this, the text is lemmatized using WordNet [2], ensuring that words are reduced to their base form. Finally, stemming is applied using the Snowball Stemmer [3] to further standardize the text. The prepared text is then transformed into a vectorized form using Bag-of-Words, which captures the presence of particular words within each query. Logistic Regression [4] was used as the primary classifier, with coefficients reflecting the importance of each word in making the classification decision.

III. Evaluation

To evaluate the performance of the classification model, we used precision, recall, and F1-score metrics. Precision measures the accuracy of the out-of-scope classification, while recall indicates how well the model captures all relevant queries. The F1-score balances these metrics, providing an overall assessment of model performance. Initial results show a precision of X

Dataset Size

The dataset consists of 23,700 queries, with 15,000 samples used for training, 4,500 for testing, and 3,000 for validation. Additionally, there are 1,000 out-of-scope queries reserved for testing and 100 for validation, allowing us to closely monitor the performance of the classifier on OOS data. The balanced distribution of categories ensures a fair evaluation of the model, though the OOS category remains slightly under-represented in comparison to other domain-specific categories [6].

Topic Analysis

An analysis of the most frequent words in out-of-scope queries reveals that many are generic or pertain to topics that are not relevant to the system's domains, such as personal inquiries or highly specialized topics. This suggests that the model relies heavily on the absence of domain-specific keywords to classify a query as OOS. However, this approach could be exploited by deliberately avoiding specific keywords to misclassify certain queries [7]. Future work may focus on enhancing the robustness of the model to such adversarial inputs.