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, and the OOS category contains 1,200 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. The original source of this dataset and the academic paper showing the respective business purpose are available at https://www.aclweb.org/anthology/D19-1131/.

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, ensuring that words are reduced to their base form. Finally, stemming is applied using the Snowball Stemmer to further standardize the text. The prepared text is then transformed into a vectorized form using Bag-of-Words (BoW), which captures the presence of particular words within each query. BoW is appropriate in this case, since the classification between “out-of-scope” and “in-scope” may depend on the presence of specific keywords that indicate alignment with the topics provided. For example, terms that frequently appear in in-scope queries are unlikely to be present in out-of-scope queries.

However, BoW can be vulnerable to adversarial attacks. If an out-of-scope query includes too many words related to the scope, the model can be confused. For example:

If a query is: “help with marketing topic,” and “marketing” is an in-scope topic, the model may incorrectly classify it as “in-scope” even though the query is not asking for something relevant.

Logistic Regression 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 the balanced accuracy metric. Using balanced accuracy is advantageous because it adjusts for class imbalance by calculating the average recall obtained on each class. This ensures that all classes, including the less frequent ’out-of-scope’ category, are equally represented in the performance evaluation. The balanced accuracy of the out-of-scope classification indicates how well the model captures all relevant queries, providing a fair assessment even when the dataset is imbalanced. This metric allows us to better understand the model’s effectiveness across all categories, ensuring that minority classes are not overshadowed by the majority classes.

The word ”make” appears as relevant to the classification, but this is a fairly generic action word. Its presence as relevant may be a sign that the classifier is picking up frequent patterns in the text, but it may not be a good discriminator between “in-scope” and “out-of-scope” classes. This word can appear in many types of queries (e.g., “How do I make a reservation?”), which makes it a common word but not necessarily significant for the real purpose of the query. The word `reservation` appears to be the most relevant to the classifier. This makes sense, as there are several scopes that include the reservation of services, such as `book_hotel`, `cancel_reservation`, `accept_reservations`, `confirm_reservation`, and `restaurant_reservation`. And yet, the word `possible` appears as one of the ten most relevant words, and since it is a very broad and generic word, appearing in different types of queries without necessarily indicating a specific topic, this word is possibly confusing the classifier.

	accept_reservations	account_blocked	alarm	application_status	apr	are_you_a
accept_reservations	23	0	0	0	0	
account_blocked	0	25	0	0	0	
alarm	0	0	28	0	0	
application_status	0	0	0	38	0	
apr	0	0	0	0	32	
are_you_a_bot	0	0	0	0	0	
balance	0	1	0	0	0	
bill_balance	0	0	0	0	0	
bill_due	0	0	0	0	0	
book_flight	0	0	0	0	0	

Fig. 1. Confusion matrix displaying the classification performance, indicating how well the model classified queries into ’out-of-scope’ and ’in-scope’ categories.

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.

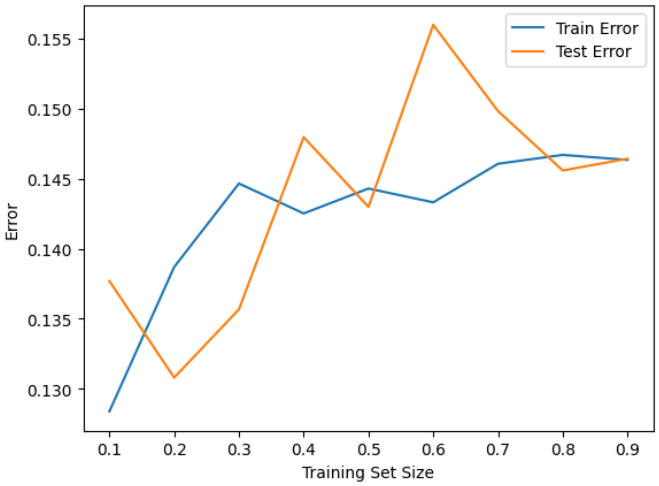

Fig. 2. The graph illustrates the difference between training and testing errors as the dataset size increases. As shown, the difference between train and test errors tend to decrease as the dataset size increase.

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 classify wrong certain queries.