**Results**

All the 75 ambiguous queries were passed through the trained SBERT model, and the cosine similarity were computed between the query and associated clarifying questions. And the resultant clarification with the highest score was selected as the top prediction made by the model. The prediction is only considered as “confident” if the top similarity score ≥ 0.40 threshold**.**

|  |  |
| --- | --- |
| **Metric** | **Results** |
| **Total Queries** | **75** |
| **Confident predictions (≥ 0.40)** | **63(84%)** |
| **Not confident** | **12(16%)** |
| **Average cosine similarity score(TOP)** | **0.563** |
| **Average Cosine(non-confident)** | **0.38** |

These are the metric used and the results. The standard metrics like Precision, re-call and f1 score were not used, as this a ranking task, not a binary classification task with negative labels.

**Discussion and conclusion**

The SBERT model demonstrated a strong semantic understanding capability in confidently ranking the contextually appropriate clarification where the ambiguous queries were diverse and different from the original Qulac data the model was trained on. The **84% confident score** and the **0.563 percent** of average similarity shows that the model capable and it is understanding the semantics of the Query-clarification of a new diverse dataset. Also mentioning that the queries in which the model was not confident the average of similarity is around decision threshold. **(0.38)** which indicate the model’s ability to express it uncertainty when there is less semantic alignment. This reduces the risk of returning irrelevant clarifications. Reduces the chances of overconfident incorrect responses.

Although the model is good at expressing any uncertainties, but it sometimes refrains from answering even when a reasonably close clarification exists. This behaviour can limit the system’s responsiveness in the practical deployment.

The clarification is more subjective and contextual unlike the summary retrieval task. The subjectivity and context sensitive makes it harder to evaluate the model’s performance using the standard classification metrics or generate ground truth labels. This lack of having ground truth can introduce some ambiguity in the interpretation of the results and tuning threshold. This can be improved by training a model to a larger dataset and it can be finetuned for a particular domain0-specific clarification task as well can be trained on a large amount of diverse data to increase the accuracy of the results overall.

**Reference**

1. **Cosine Similarity Utility – SBERT:** [**https://www.sbert.net/docs/package\_reference/util.html**](https://www.sbert.net/docs/package_reference/util.html)
2. **MultipleNegativesRankingLoss – SBERT loss function documentation.**

[**https://www.sbert.net/docs/package\_reference/losses.html#multiplenegativesrankingloss**](https://www.sbert.net/docs/package_reference/losses.html#multiplenegativesrankingloss)

1. **all-MiniLM-L6-v2 pretrained model. Hugging Face Model Card.** [**https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2**](https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2)
2. **Hugging Face – SentenceTransformers Documentation.** [**https://www.sbert.net/**](https://www.sbert.net/)