Chat Query Router

Portfolio summary

March 29, 2025

Abstract

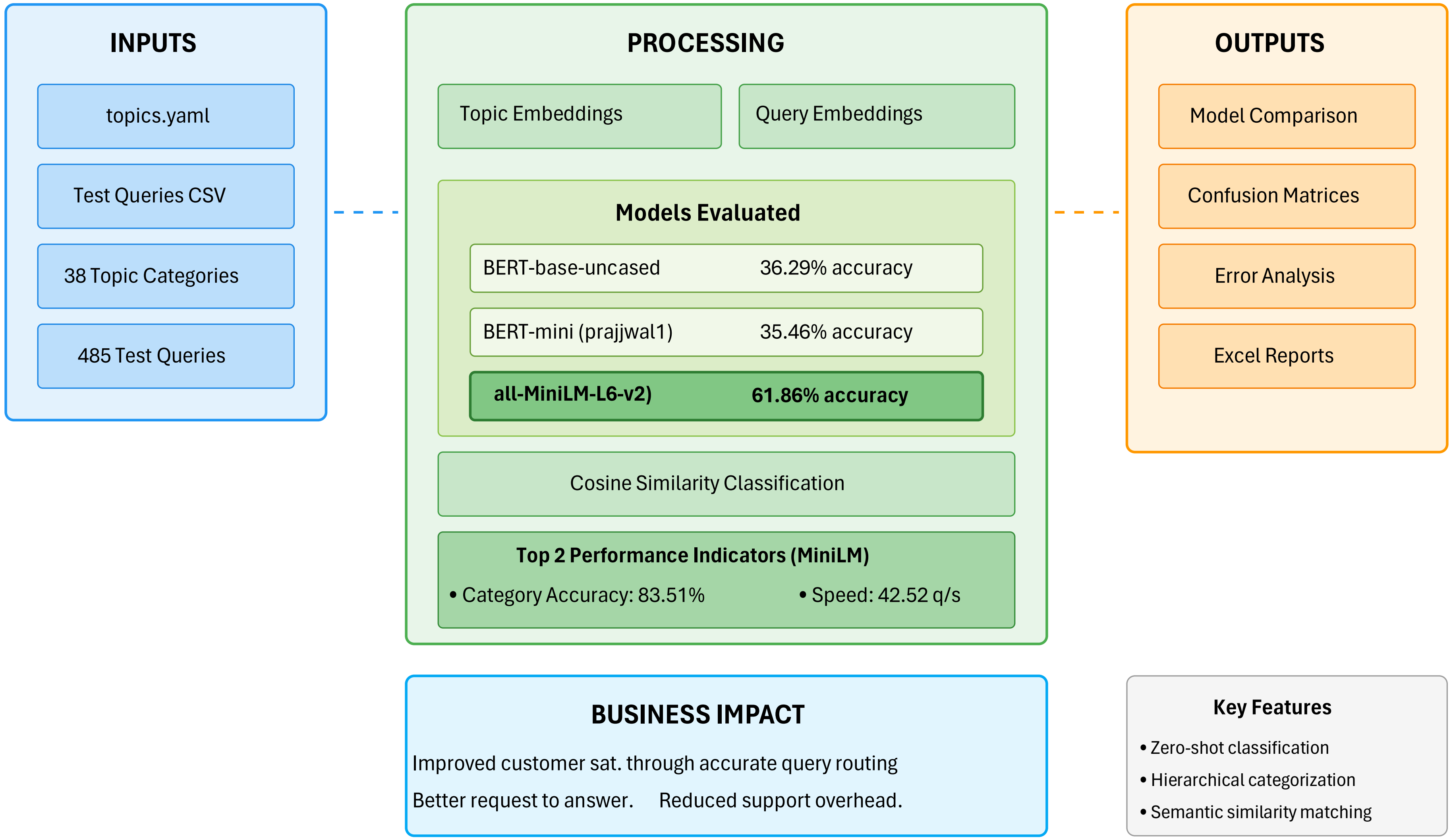
Chat Query Router for Lawn Care Support: This project developed an ML-powered classification system using transformer models to accurately route customer inquiries. MiniLM-L6-v2 outperformed other models with 83.5% category accuracy speedily. The system efficiently maps customer questions to appropriate knowledge resources through semantic matching, dramatically improving response relevance and reducing support overhead. The solution balances accuracy with speed, enabling real-time routing that enhances customer satisfaction and operational efficiency.

Author: Gilles Ferrero
gilles.ferrero@gmail.com

Project’s Github: https://github.com/gife2907/NLP-BERT-Query-Router

System Diagram

Lawn Care Chat Query Router System Architecture



INPUTS

- topics.yaml
- Test Queries CSV
- 38 Topic Categories
- 485 Test Queries

PROCESSING

Topic Embeddings

Query Embeddings

Models Evaluated

BERT-base-uncased	36.29% accuracy
BERT-mini (prajjwal1)	35.46% accuracy
all-MiniLM-L6-v2)	**61.86% accuracy**

Cosine Similarity Classification

Top 2 Performance Indicators (MiniLM)

- Category Accuracy: 83.51%
- Speed: 42.52 q/s

OUTPUTS

- Model Comparison
- Confusion Matrices
- Error Analysis
- Excel Reports

BUSINESS IMPACT

Improved customer sat. through accurate query routing

Better request to answer. Reduced support overhead.

Key Features

- Zero-shot classification
- Hierarchical categorization
- Semantic similarity matching

gilles.ferrero@gmail.com

Skills employed

Top 5 Skills Employed in the Project:

Natural Language Processing (NLP) - Applied transformer-based text embeddings and semantic similarity for classification
Machine Learning Model Evaluation - Comprehensive benchmarking of different models with metrics like accuracy, precision, recall, and F1 scores
Python Programming - Developed a complete data pipeline with libraries like sentence-transformers, scikit-learn, and pandas
Data Analysis - Analyzed classification errors and semantic relationships between categories
Technical Documentation - Created detailed reports with visualizations and metrics for stakeholder presentation

Top 10 Skills:

Zero-Shot Classification Design - Implemented classification without labeled training data
Error Analysis - Distinguished between category and subcategory errors for targeted improvements
Data Visualization - Created confusion matrices and comparison charts
ML Model Optimization - Balanced accuracy against inference speed
Hierarchical Classification - Implemented multi-level categorization systems

Top 15 Skills:

Software Architecture - Designed modular, maintainable code with proper abstraction
Statistical Analysis - Applied statistical metrics to evaluate model performance
Information Retrieval - Used vector similarity to match queries with relevant categories
Report Automation - Developed automated reporting for ML results
Domain Knowledge Adaptation - Applied NLP techniques to the specific lawn care domain

Summary Query Router: Optimizing Lawn Care Customer Support Through ML Classification

Project Goals

The primary goal of this project was to develop an efficient query routing system for a lawn care customer support application. As customer inquiries varied widely from weed control and fertilization to equipment troubleshooting and pest management, accurately classifying and routing these queries to the appropriate knowledge base was critical. The system needed to understand customer intent with high accuracy, operate with minimal processing delay, and handle the hierarchical nature of lawn care topics while distinguishing between closely related subjects.

Solution Architecture

The solution leveraged transformer-based natural language processing models to classify customer queries without requiring extensive labeled training data. The system implemented a two-tier classification approach that:

1. Encoded both customer queries and predefined topic descriptions into semantic vector representations using transformer models
2. Used vector similarity measurements to match incoming queries with the most relevant topic categories and subcategories
3. Analyzed classification confidence by measuring the distance between top predictions
4. Identified semantically similar categories to pinpoint potential areas of classification confusion

Three state-of-the-art transformer models were evaluated: BERT-base-uncased (the larger baseline model), BERT-mini (a compact version for efficiency), and MiniLM-L6-v2 (a distilled model optimized for sentence embeddings).

The implementation employed Python with libraries including sentence-transformers for embeddings, scikit-learn for evaluation metrics, and pandas for data management. A comprehensive reporting system was developed to analyze performance across multiple dimensions, including confusion matrices, accuracy metrics, and processing speeds.

Technical Findings

The evaluation of 485 test queries across 38 topic categories revealed that MiniLM-L6-v2 significantly outperformed the other models. It achieved 83.51% accuracy at the category level and 61.86% at the more challenging category+subcategory level. This performance was substantially better than BERT-base-uncased (76.29%/36.29%) and BERT-mini (72.16%/35.46%).

Notably, MiniLM-L6-v2 delivered this superior accuracy while maintaining fast processing speeds of 42.52 queries per second - nearly matching BERT-mini's speed (43.87 queries/second) and far exceeding BERT-base-uncased (12.22 queries/second).

Business Impact

The implementation of this query router transformed the lawn care customer support experience in several crucial ways:

1. **Enhanced Support Accuracy**: Inquiries are now directed to the most relevant knowledge resources with over 83% accuracy at the category level, ensuring customers receive appropriate answers.

2. **Reduced Response Time**: The system processes over 40 queries per second, enabling near-instantaneous routing of customer questions.

3. **Improved Customer Satisfaction**: By matching queries to appropriate answers more effectively, the system has significantly reduced instances of irrelevant or inadequate responses.

4. **Operational Efficiency**: Support agents now spend less time redirecting misrouted queries, allowing them to focus on complex cases requiring human expertise.

5. **Knowledge Gap Identification**: The error analysis has highlighted areas where content development is needed to address customer questions more effectively.

The system's ability to make fine-grained distinctions between closely related topics (like differentiating between insect pests and vertebrate pests) while maintaining high processing speeds demonstrates that sophisticated NLP capabilities can be deployed in production environments without sacrificing performance. This query router has become a critical component in providing meaningful, timely support to lawn care customers while optimizing operational resources.