CUSTOMER SERVICE SUPPORT - UTILIZING MACHINE LEARNING TO CLASSIFY PRIORITIZE AND SUMMARIZE ISSUES

Author: Hoai Nhan Nguyen (sba24098@student.cct.ie) Institutes: CCT College Dublin

ABSTRACT

This project explores the use of machine learning to improve customer support ticket handling. Using a multilingual dataset of customer service tickets, I compared the performance of five classification algorithms for two tasks: ticket queue and priority classification. I also implemented both extractive and abstractive summarization techniques to generate concise overviews of the ticket content.

Among the models tested, Linear Support Vector Classifier demonstrated the best overall performance and extractive summarization was found to be more effective in preserving clarity in technical cases.

INTRODUCTION & BUSINESS CONTEXT

Improving customer experience remains one of the top priorities for organizations. Delays in resolving customer issues is often due to ticket misrouting or incorrect prioritization which can negatively impact customer satisfaction and retention.

To address these issues, I explored the use of machine learning to improve the efficiency of support ticket handling.

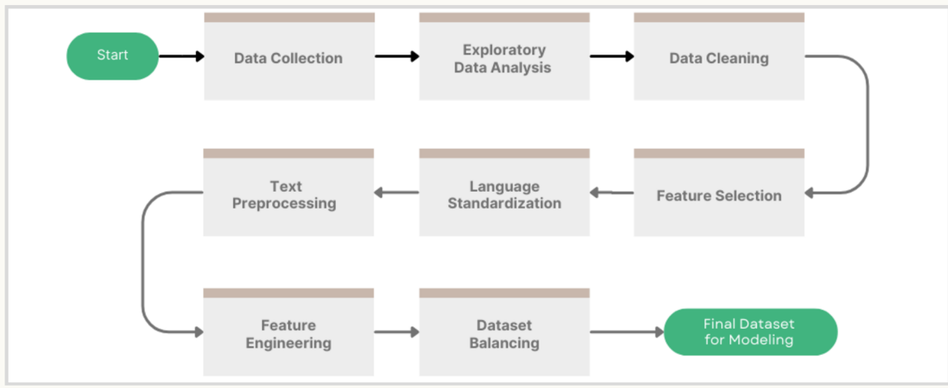
- **Queue Classification** – Automatically categorizing tickets based on its content
- **Priority Classification** – Predicting ticket urgency (Low, Medium, High)
- **Text Summarization** – Generating concise summaries of customer issues

The objective of this project is to evaluate and compare various machine learning models and techniques that can help customer support agents work more efficiently and improve response times.

DATA OVERVIEW & PREPARATION

Dataset Summary

- **Source:** Customer support tickets from Kaggle (Bueck, 2024).
- **Size:** 4,000 rows × 17 columns.
- **Languages:** English, German, Spanish, French, Portuguese.



Data Preparation Pipeline

- **Data Cleaning:** Removed incomplete rows. (No Duplicate Rows)
- **Feature Selection:** subject, body, queue, priority, tags.
- **Language Standardization:** Translate all text to English.
- **Text Preprocessing:** Text Normalization, Tokenization, Stopword removal, Stemming and Lemmatization.
- **Feature Engineering:** Applied TF-IDF to subject and body, Bag of Words to tags, and encoded queue and priority labels.
- **Dataset Balancing:** Used SMOTE to correct class imbalance in target labels.

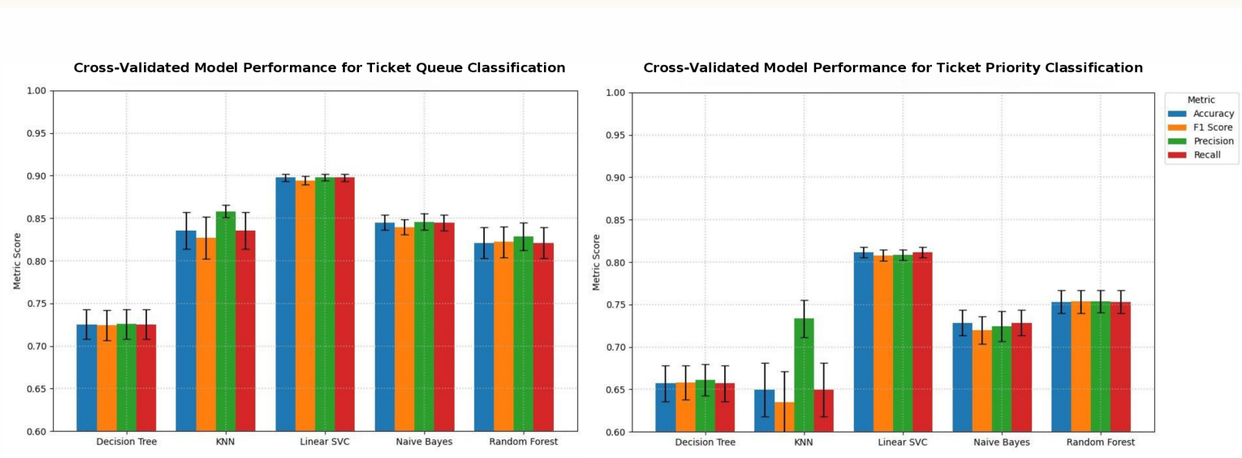
MACHINE LEARNING CLASSIFICATION

To classify the customer support tickets by queue and priority, I implemented and compared five machine learning models:

- Decision Tree
- K-Nearest Neighbors (KNN)
- Linear Support Vector Classifier (LinearSVC)
- Naive Bayes
- Random Forest

I applied **RandomizedSearchCV** for hyperparameter tuning and evaluated each model using cross-validation, training and prediction speed, learning curve analysis, and confusion matrix performance.

Among all models tested, LinearSVC delivered the highest accuracy and most consistent performance across all evaluation metrics, making it the most effective choice for both ticket queue and priority classification.

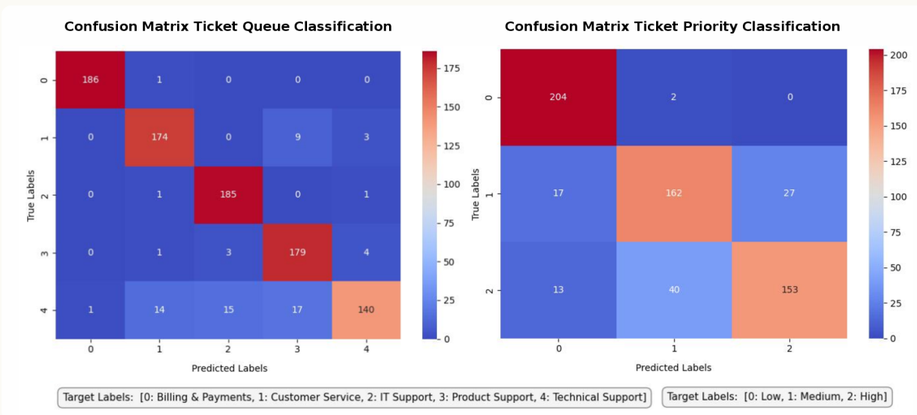


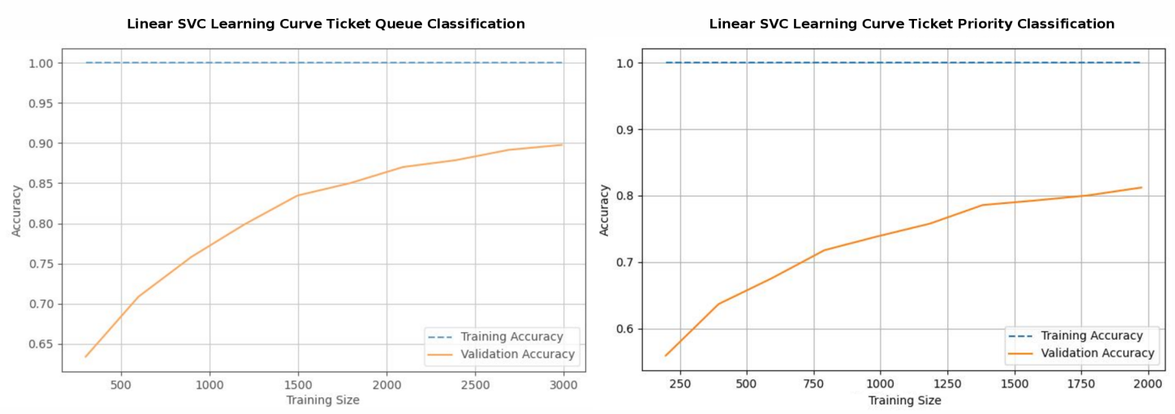
Model Speed and Accuracy for Ticket Queue Classification

	Model	Training Time (s)	Prediction Time (s)	Accuracy
0	Decision Tree	1.231192	0.094584	0.764454
1	KNN	0.671277	22.582993	0.860814
2	Linear SVC	79.838578	0.136117	0.925054
3	Naive Bayes	0.471715	0.253736	0.852248
4	Random Forest	7.441737	0.168920	0.873662

Model Speed and Accuracy for Ticket Priority Classification

	Model	Training Time (s)	Prediction Time (s)	Accuracy
0	Decision Tree	0.942868	0.068399	0.689320
1	KNN	0.314857	6.653142	0.721683
2	Linear SVC	17.349400	0.099388	0.839806
3	Naive Bayes	0.240041	0.101404	0.757282
4	Random Forest	3.765244	0.106764	0.758900





TEXT SUMMARIZATION

To help support agents quickly grasp the core issue in customer tickets, I explored two summarization strategies:

- **Extractive Summarization (BERT) -** Selects key sentences directly from the original text.
- **Abstractive Summarization (BART) -** Generates new text that paraphrases the original content.

After testing both approaches, I found that

- **Extractive Summarization (BERT) -** Was more reliable and accurate.
- **Abstractive Summarization (BART) -** Created fluent, natural text but occasionally omitted critical information.

Final Choice - I selected extractive summarization as the preferred method due to its consistency, accuracy, and ability to preserve essential details.

CHALLENGES ENCOUNTERED

Throughout the project, several challenges arose in data preparation and model development. Strategic solutions were implemented to ensure effective outcomes.

Imbalanced Classes
Ticket queues and priority levels were unevenly distributed.
Solution: Applied SMOTE to oversample minority classes and balance the dataset.

Multilingual Text
Support tickets were written in five different languages.
Solution: Translated all text to English for standardized NLP preprocessing.

Model Selection
Identifying the best model for multi-class classification.
Solution: Evaluated five models using cross-validation, confusion matrix, and learning curve analysis.

CONCLUSION & FUTURE WORK

This project demonstrated the practical application of machine learning that could improve customer support ticket efficiency. Among the five machine learning models evaluated, the Linear Support Vector Classifier consistently achieved the highest accuracy and generalization for both queue and priority classification.

In the summarization task, extractive summarization proved more effective than its abstractive counterpart, especially in technical scenarios where clarity and completeness are critical.

Future Work

Implement deep learning models for ticket queue and priority classification to evaluate their performance against traditional machine learning models used in this project.