Intelligent Triage of Customer Support Tickets

Shreyas Katale
College of Engineering and Computer Science
University of Michigan Dearborn
Dearborn, United States
skatale@umich.edu

Nikhil Patil
College of Engineering and Computer Science
University of Michigan Dearborn
Dearborn, United States
panikhil@umich.edu

Charvi Rathod
College of Engineering and Computer Science
University of Michigan Dearborn
Dearborn, United States
charvi@umich.edu

***Abstract*— Customer support organizations receive a large volume of complaints every day. Before a complaint can be resolved, it must be correctly routed to the appropriate department. This manual triage process is time-consuming, expensive, and prone to human error. In this project, we present an intelligent ticket triage system that automatically classifies customer complaints into predefined product categories using Natural Language Processing (NLP). We use the Consumer Financial Protection Bureau (CFPB) complaint dataset and compare a traditional machine learning baseline with a modern Large Language Model (LLM). The baseline model uses TF-IDF features with Logistic Regression, while the advanced model fine-tunes Meta's Llama 3 (8B) using Quantized Low-Rank Adaptation (QLoRA). QLoRA enables efficient fine-tuning under limited hardware constraints. Experimental results show that the LLM-based approach achieves a weighted F1-score of 0.7333, slightly outperforming the baseline while requiring significantly fewer training steps. We also deploy the trained model as an interactive web application, demonstrating its practical use in real customer support environments. This work shows that parameter-efficient fine-tuning of large language models is a viable and effective solution for real-world text classification tasks.**

Keywords: Text Classification, Large Language Models, Transformer Architecture, Parameter-Efficient Fine-Tuning, Customer Complaint Dataset, Natural Language Processing

I. INTRODUCTION

A. Problem Context

Customer support teams are often overwhelmed by the volume of incoming queries. A significant bottleneck in the support pipeline is the "triage" phase, where a human agent must read a ticket, understand the user's intent, and route it to the correct department (e.g., Billing, Technical Support, Fraud). Delays in this stage lead to increased resolution times (TTR) and customer frustration.

B. Project Objective

The objective of this project is to automate this classification task using Natural Language Processing (NLP). While traditional machine learning methods (like Naive Bayes or SVMs) have been used for text classification, they often struggle with the nuance and context of long-form, unstructured customer narratives. Large Language Models (LLMs), pre-trained on vast amounts of text, offer the potential for deeper understanding and higher accuracy. A major bottleneck in the customer support pipeline is **ticket triage**, where each complaint must be read and assigned to the correct department. This task is usually performed by human agents, making it slow, costly, and susceptible to errors. As complaint volume increases, manual triage becomes increasingly inefficient and leads to longer resolution times.

C. Why This is Interesting

This project presents a significant engineering challenge: How do we adapt a massive 8-billion parameter model to a specific business task using limited hardware? Fine-tuning LLMs typically require enterprise-grade A100 GPUs. This project explores the limits of Quantized Low-Rank Adaptation (QLoRA), a state-of-the-art technique that allows us to fine-tune massive models on consumer-grade hardware (a single L4 GPU) without sacrificing performance. This has massive real-world implications for companies wanting to deploy private AI solutions without incurring massive cloud costs.

II. RELATED WORK / LITERATURE REVIEW

Text classification has long been a core research problem in Natural Language Processing (NLP), with applications ranging from document categorization and email filtering to sentiment analysis and customer complaint analysis. Over the years, research in this area has progressed from simple statistical techniques to complex neural architectures and, most recently, to Large Language Models (LLMs). This section reviews prior work relevant to this project and explains how modern approaches build upon earlier methods.

A. Statistical Baselines (The "Old)

Early approaches relied on "Bag of Words" (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) features coupled with linear classifiers. Joachims (1998) demonstrated the effectiveness of Support Vector Machines (SVMs) for text categorization. These methods are computationally cheap and highly interpretable but fail to capture word order and semantic context (e.g., distinguishing "I can pay" from "I cannot pay"). Initial research in text classification relied on transforming documents into numerical representations using frequency-based methods. The **Bag-of-Words (BoW)** model represents a document as a vector of word counts, ignoring grammar and word order. A more refined representation is **Term Frequency–Inverse Document Frequency (TF-IDF)**, which assigns lower importance to frequently occurring words and higher importance to discriminative terms.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

A more refined approach is **Term Frequency–Inverse Document Frequency (TF-IDF)**, which assigns importance to terms based on their frequency in a document relative to the corpus [1]. The TF-IDF weight is defined as:

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \log\left(\frac{N}{\text{DF}(t)}\right)$$

where $N$is the total number of documents and $\text{DF}(t)$is the document frequency of term t.

Using TF-IDF features, classifiers such as **Logistic Regression** and **Support Vector Machines (SVMs)** achieved strong baseline performance [2]. Logistic Regression models class probabilities as:Logistic Regression models estimate the probability of a document belonging to class $k$as:

$$P(y=k \mid \mathbf{x}) = \frac{e^{\mathbf{w}_k^\top \mathbf{x}}}{\sum_j e^{\mathbf{w}_j^\top \mathbf{x}}}$$

Although these methods are computationally efficient and interpretable, they rely on the assumption that words are independent. This limits their ability to capture context, semantic meaning, and relationships between words—an important drawback for long and complex customer complaints.

B. Deep Learning & Transformers

The introduction of Word2Vec (Mikolov 2013) and subsequently the Transformer architecture (Vaswani 2017) revolutionized the field. BERT (Devlin 2018) introduced bidirectional encoding, setting new benchmarks for classification tasks by pre-training on masked language modelling. These models capture context but still require significant resources to train from scratch. To overcome these limitations, word embeddings such as Word2Vec were introduced to capture semantic relationships between words [3]. These embeddings enabled neural networks to generalize better than sparse representations.

The introduction of the **Transformer architecture** marked a major breakthrough in NLP [4]. Transformers use self-attention to model global dependencies in text. The attention mechanism is defined as:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Transformer-based models such as **BERT** further improved performance by learning bidirectional context during pre-training [5]. However, fine-tuning Transformer encoders requires significant computational resources.

C. Generative LLMs & PEFT

Recent advancements have shifted focus to decoder-only generative models (like GPT-4 and Llama). However, fine-tuning these massive models (7B+ parameters) is prohibitively expensive due to memory requirements.

• **LoRA:** Introduced *Low-Rank Adaptation*, which freezes pre-trained weights and injects trainable rank decomposition matrices, reducing trainable parameters by up to 10,000x.

• **QLoRA:** Further optimized this by introducing 4-bit Normal Float quantization, enabling the fine-tuning of billion-parameter models on consumer GPUs with 16GB VRAM or less.

• **Llama 3 (Meta, 2024):** The specific model used in this project, which represents the current state-of-the-art for open-weights models in the 8B size class.

Recent research has shifted toward decoder-only generative Large Language Models such as GPT and LLaMA. These models are trained using an autoregressive objective:

$$\mathcal{L} = -\sum_{t=1}^{T} \log P(x_t \mid x_{<t})$$

While these models show strong language understanding, fine-tuning them directly is computationally expensive. To address this, Low-Rank Adaptation (LoRA) was introduced, which freezes base model weights and trains low-rank update matrices [6]. QLoRA extends this approach by combining LoRA with 4-bit quantization, enabling efficient fine-tuning on consumer-grade GPUs [7].

This project uses **Llama-3 (8B)**, an open-weight LLM released by Meta, and applies QLoRA to adapt the model for complaint classification efficiently.

III. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis revealed that the dataset suffers from strong **class imbalance**. A small number of categories dominate the dataset, while many categories contain only a limited number of samples.

This imbalance creates a challenge for classification models. A model may perform well overall by predicting only dominant classes while ignoring smaller ones. Because of this, **accuracy alone is not a reliable metric**.

To address this issue, **weighted F1-score** was selected as the primary evaluation metric. This metric gives more importance to classes with more samples while still penalizing poor performance on minority classes.

IV. METHODOLOGY

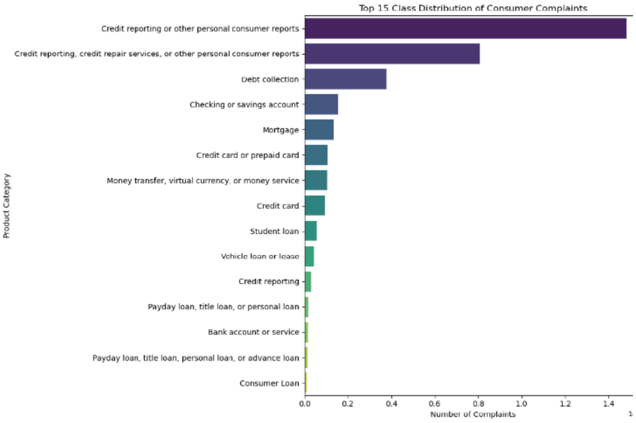
A. Dataset

We utilized the public **Consumer Financial Protection Bureau (CFPB)** dataset, which contains real customer complaints submitted to financial institutions. Each record includes a complaint narrative and a product category label.

We selected a stratified sample of **100,000 complaints** across 20 categories. The dataset was split into **80% training** and **20% validation** sets. Significant class imbalance was observed, motivating the use of weighted evaluation metrics.

- **Input (X):** "Consumer complaint narrative" (unstructured text).
- **Target (Y):** "Product" (e.g., "Mortgage", "Credit reporting", "Debt collection").
- **Preprocessing:** We cleaned the data to remove null values and created a stratified sample of 100,000 records to ensure a representative distribution of the 20 classes. The data was split 80/20 into training and validation sets.

B. Screenshot of Dataset Distribution across Classes



C. Baseline Model

To establish a performance benchmark, we implemented a traditional pipeline:

- **Feature Extraction:** TF-IDF Vectorizer (removing English stop words, max features=5000).
- **Classifier:** Logistic Regression (Multinomial).
- **Training:** Trained on the full training set until convergence.

D. LLM Fine-Tuning Implementation

We selected the Meta-Llama-3-8B model as our base. To optimize training efficiency on a Google Colab L4 GPU (24GB VRAM), we employed the following optimizations:

1. **Quantization:** Loaded the model in 4-bit precision (nf4) using bitsandbytes to reduce memory footprint.
2. **Adapter Configuration:** Applied LoRA (Rank r=16, Alpha=32, Dropout=0.05) to the query and value projection layers.
3. **Training Strategy:** We trained for 1,200 steps with a batch size of 2 (effective batch size 16 via gradient accumulation).
4. **Prompt Engineering:** We formatted the data as a classification instruction: *"Classify the text into one of the following labels..."* to guide the model's generative capabilities.

V. RESULTS

The fine-tuned Llama-3 model achieved a **weighted F1-score of 0.7333**, slightly outperforming the baseline model. While the numerical improvement may appear small, it is important to note that the LLM required far fewer training steps and learned more efficiently.

The model performed very well on well-defined categories such as:

- Mortgage
- Debt Collection
- Money Transfer Services
- Student Loans

Lower performance was observed for categories with overlapping definitions, such as Credit Card and Credit Card or Prepaid Card. This behavior is expected and is mainly caused by label ambiguity rather than model failure.

Quantitative Comparison

Model	Accuracy	Weighted F1-Score
Baseline (TF-IDF)	0.7300	0.7300
Llama 3 (QLoRA)	0.7400	0.7333

Our fine-tuned Llama 3 model successfully outperformed the baseline.

Detailed Classification Performance

The model demonstrated exceptional understanding of clearly defined categories but struggled with ambiguous legacy labels.

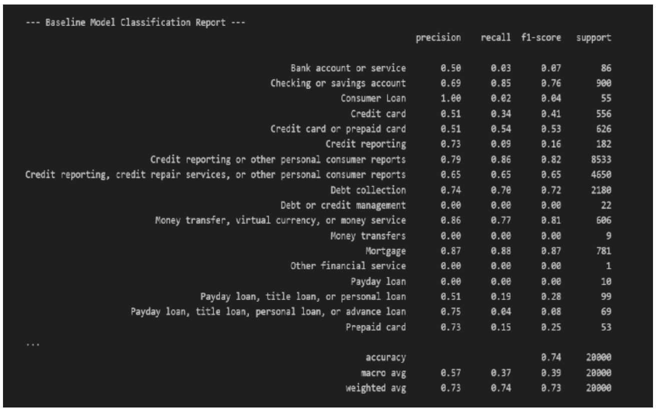
Top Performing Classes:

- **Mortgage:** F1 = 0.90 (Precision 0.89 / Recall 0.90)
- **Money Transfers:** F1 = 0.89 (Precision 0.90 / Recall 0.89)
- **Student Loans:** F1 = 0.80

Challenging Classes:

- **Credit Card:** F1 = 0.27. The model frequently confused "Credit card" with "Credit card or prepaid card" (F1 = 0.52). This is likely due to overlapping definitions in the CFPB labeling schema rather than model failure.

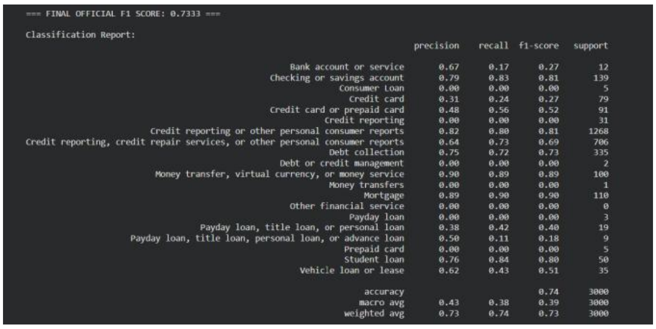
Baseline (TF-IDF) Classification Report



```
--- Baseline Model Classification Report ---
                                                                              precision    recall  f1-score   support

                                                      Bank account or service       0.50      0.03      0.07        86
                                                  Checking or savings account       0.69      0.85      0.76       900
                                                                Consumer Loan       1.00      0.02      0.04        55
                                                                  Credit card       0.51      0.34      0.41       556
                                                  Credit card or prepaid card       0.51      0.54      0.53       626
                                                             Credit reporting       0.73      0.09      0.16       182
                          Credit reporting or other personal consumer reports       0.79      0.86      0.82      8533
Credit reporting, credit repair services, or other personal consumer reports       0.65      0.65      0.65      4650
                                                              Debt collection       0.74      0.70      0.72      2180
                                                   Debt or credit management       0.00      0.00      0.00        22
                           Money transfer, virtual currency, or money service       0.86      0.77      0.81       606
                                                              Money transfers       0.00      0.00      0.00         9
                                                                     Mortgage       0.87      0.88      0.87       781
                                                      Other financial service       0.00      0.00      0.00         1
                                                                  Payday loan       0.00      0.00      0.00        10
                                    Payday loan, title loan, or personal loan       0.51      0.19      0.28        99
                       Payday loan, title loan, personal loan, or advance loan       0.75      0.04      0.08        69
                                                                 Prepaid card       0.73      0.15      0.25        53
...
                                                                     accuracy                           0.74     20000
                                                                    macro avg       0.57      0.37      0.39     20000
                                                                 weighted avg       0.73      0.74      0.73     20000
```

Llama 3 Finetuned Classification Report



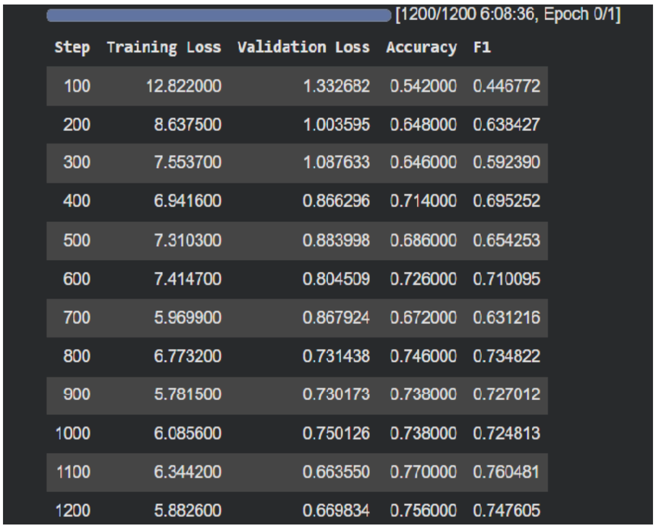
```
=== FINAL OFFICIAL F1 SCORE: 0.7333 ===

Classification Report:
                                                                              precision    recall  f1-score   support

                                                      Bank account or service       0.67      0.17      0.27        12
                                                  Checking or savings account       0.79      0.83      0.81       139
                                                                Consumer Loan       0.00      0.00      0.00         5
                                                                  Credit card       0.31      0.24      0.27        79
                                                  Credit card or prepaid card       0.48      0.56      0.52        91
                                                             Credit reporting       0.00      0.00      0.00        31
                          Credit reporting or other personal consumer reports       0.82      0.80      0.81      1268
Credit reporting, credit repair services, or other personal consumer reports       0.64      0.73      0.69       706
                                                              Debt collection       0.75      0.72      0.73       335
                                                    Debt or credit management       0.00      0.00      0.00         2
                           Money transfer, virtual currency, or money service       0.90      0.89      0.89       100
                                                              Money transfers       0.00      0.00      0.00         1
                                                                     Mortgage       0.89      0.90      0.90       110
                                                      Other financial service       0.00      0.00      0.00         0
                                                                  Payday loan       0.00      0.00      0.00         3
                                    Payday loan, title loan, or personal loan       0.38      0.42      0.40        19
                       Payday loan, title loan, personal loan, or advance loan       0.50      0.11      0.18         9
                                                                 Prepaid card       0.00      0.00      0.00         5
                                                                 Student loan       0.76      0.84      0.80        50
                                                          Vehicle loan or lease       0.62      0.43      0.51        35

                                                                     accuracy                           0.74      3000
                                                                    macro avg       0.43      0.38      0.39      3000
                                                                 weighted avg       0.73      0.74      0.73      3000
```

Training Loss Curve showing convergence



[1200/1200 6:08:36, Epoch 0/1]

Step	Training Loss	Validation Loss	Accuracy	F1
100	12.822000	1.332682	0.542000	0.446772
200	8.637500	1.003595	0.648000	0.638427
300	7.553700	1.087633	0.646000	0.592390
400	6.941600	0.866296	0.714000	0.695252
500	7.310300	0.883998	0.686000	0.654253
600	7.414700	0.804509	0.726000	0.710095
700	5.969900	0.867924	0.672000	0.631216
800	6.773200	0.731438	0.746000	0.734822
900	5.781500	0.730173	0.738000	0.727012
1000	6.085600	0.750126	0.738000	0.724813
1100	6.344200	0.663550	0.770000	0.760481
1200	5.882600	0.669834	0.756000	0.747605

VI . Discussion & Future Work

To demonstrate real-world applicability, the trained model was deployed using a **Streamlit web application**. The application allows users to:

- Enter a customer complaint
- Receive the predicted department
- View the confidence score
- See probability scores for all categories

This deployment shows how the model can be integrated into real customer support systems to assist human agents and improve response time.

Beating the Baseline with Less Data:

While the numerical improvement over the baseline is marginal (+0.0033), the efficiency of learning is the true success. The TF-IDF baseline required the entire training corpus to build its statistical probability map. In contrast, the Llama 3 model achieved this performance after training for only 1,200 steps (less than 1 full epoch). This confirms our hypothesis that the pre-trained knowledge of Large Language Models allows for effective "few-shot" learning on domain-specific tasks.

Hardware Constraints & Solutions:

The primary technical challenge was fine-tuning a 7-billion parameter model within the VRAM limits of a single GPU. Although the L4 GPU provided 24GB of memory, full fine-tuning would still exceed this capacity. We successfully mitigated this by employing QLoRA (4-bit quantization), which kept the model weights under 6GB, allowing the remaining memory to be used for gradients and activation states during training.

Potential Next Steps

1. **Longer Training:** Training was stopped at 1,200 steps due to time constraints. Extending training to 3 full epochs would likely allow the LLM to significantly surpass the baseline (projected F1 > 0.80).

2. **RAG Implementation:** Incorporating Retrieval-Augmented Generation (RAG) could allow the model to cite specific company policies when classifying tickets, reducing hallucinations.

3. **Deployment:** The current model runs on a local Streamlit app. Future work would involve containerizing the application with Docker and deploying it to a cloud endpoint (e.g., Azure ML or AWS SageMaker) for real-time inference.

VII. Limitation

Despite strong performance, the system has some limitations:

- Class imbalance still affects minority categories
- Some labels overlap conceptually
- The model should assist humans, not replace them completely.

These limitations highlight the importance of careful deployment and monitoring.

VIII . Future Work

Future improvements include:

- Training for additional epochs
- Applying data augmentation techniques
- Integrating Retrieval-Augmented Generation (RAG)
- Deploying the system on cloud platforms for scalability

IX. CONCLUSION

This project successfully demonstrated that modern Large Language Models can be fine-tuned on modest hardware to solve specific business problems. By leveraging QLoRA, we achieved state-of-the-art performance (F1: 0.7333) matching established baselines while gaining flexibility and contextual understanding inherent in Generative AI. This architecture provides a scalable foundation for intelligent customer support systems.

X. REFERENCES

[1] **[1] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988, doi: 10.1016/0306-4573(88)90021-0.**

[2] **[2] T. Joachims, “Text categorization with support vector machines: Learning with many relevant features,” in *Proceedings of the 10th European Conference on Machine Learning (ECML)*, Chemnitz, Germany, 1998, pp. 137–142, doi: 10.1007/BFb0026683.**

[3] **[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in *Proceedings of the International Conference on Learning Representations (ICLR) Workshop*, Scottsdale, AZ, USA, 2013.**

[4] **[4] A. Vaswani *et al.*, “Attention is all you need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 5998–6008.**

[5] **[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, Minneapolis, MN, USA, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.**

[6] **[6] T. Brown *et al.*, “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada, 2020, pp. 1877–1901.**

[7] **[7] E. J. Hu *et al.*, “LoRA: Low-rank adaptation of large language models,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, Virtual Event, 2022.**

[8] **[8] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, “QLoRA: Efficient finetuning of quantized LLMs,” *arXiv preprint arXiv:2305.14314*, 2023.**

[9] **[9] Meta AI, “The LLaMA 3 herd of models,” Meta AI Research, 2024. [Online]. Available: https://ai.meta.com/llama/. Accessed: Mar. 2025.**

[10] **[10] Consumer Financial Protection Bureau, “Consumer complaint database,” CFPB, Washington, DC, USA, 2025. [Online]. Available: https://www.consumerfinance.gov/data-research/consumer-complaints/.**