DATS 6312: NLP for Data Science - Final Project Proposal

Project Title: A Comparative Analysis of Classical and Transformer-Based Models for Financial Complaint Categorization

1. What problem did you select and why did you select it?

Problem: Financial institutions and regulatory bodies like the Consumer Financial Protection Bureau (CFPB) receive millions of customer complaints, which must be read, categorized, and routed to the correct department (e.g., "Mortgage," "Credit card," "Debt collection"). This manual process is slow, expensive, and error-prone. Our project will build and evaluate a multi-class text classification model to automate this routing process.

Why: This is a high-impact, real-world problem that is a classic and practical application of NLP. This "original investigation" aligns perfectly with the course objectives and the professor's guidance to pursue a non-causal, applied NLP research topic. It allows us to directly compare different generations of NLP models on a large, public dataset.

2. What database/dataset will you use?

We will use the Consumer Complaint Database published by the Consumer Financial Protection Bureau (CFPB).

- Source: https://www.consumerfinance.gov/data-research/consumer-complaints/
- Dataset Details: This is a large-scale U.S. government dataset containing millions of real-world complaints. We will use two key columns:
    1. Feature (X): Consumer complaint narrative (the unstructured text of the complaint).
    2. Label (y): Product (the category, e.g., "Credit reporting, credit repair services...", "Debt collection," "Mortgage," etc.).

3. What NLP methods will you pick from the concept list?

We will implement and conduct a detailed comparison of two models from the "List of key concepts":

1. Rule-Based Model (Classical): We will use Multinomial Naive Bayes. This will serve as our fast, efficient baseline model, trained on TF-IDF features.
2. Pretrained NLP (Transformer): We will use a DistilBERT model. This model will be *customized* by fine-tuning it on the CFPB complaint data to achieve state-of-the-art performance.

Our project's core analysis will be comparing the performance, training time and evaluation metrics of the classical Naive Bayes model versus the fine-tuned Transformer.

4. What packages are you planning to use? Why?

- pandas: To load, explore, and clean the large CSV dataset.
- scikit-learn: To implement our entire baseline pipeline (TF-IDF vectorization, Naive Bayes classifier) and to compute our evaluation metrics (F1-score, precision, recall, confusion matrix).
- Hugging Face transformers: To load the pre-trained DistilBERT model and its tokenizer for fine-tuning.
- PyTorch or TensorFlow: As the backend framework for fine-tuning the Transformer model.
- matplotlib / seaborn: To visualize the data, such as the class imbalance and the final confusion matrices for each model.

5. What NLP tasks will you work on?

The primary NLP task is multi-class text classification. The goal is to build a model that can read a new, unseen Consumer Complaint narrative and accurately assign it to one of the predefined Product categories.

6. How will you judge the performance of the model?

The CFPB dataset is known to be highly imbalanced (e.g., the "Credit reporting" category has far more complaints than others). Therefore, using "Accuracy" as a metric would be misleading.

Our performance evaluation will be based on the following:

1. Primary Metric: Macro-Averaged F1-Score. This is our most important metric as it calculates the F1-score for each class independently and then averages them, giving equal weight to every category, regardless of its size.
2. Supporting Metrics: Per-Class Precision and Recall. These will allow us to see which specific categories our models are good (or bad) at predicting.
3. Visualization: A Confusion Matrix for each model to visually analyze where the misclassifications are occurring (e.g., is the model often confusing "Mortgage" with "Debt collection"?).

7. Group Members : Amogh Ramagiri, Vaijayanti Deshmukh, Lasya Gowda