Project Proposal

Title: Data-Driven Insights into iPhone Customer Feedback using NLP

---

1. Background

In large-scale production systems, panic logs such as user-space watchdog timeouts are typically high in volume and complex in nature. A common approach to manage such data is to extract concise summaries, group them by similarity, and label them according to problem domains (e.g., Camera, GPU, Video).

At my workplace, I deal with user-space watchdog timeout problems. However, due to legal restrictions, I cannot share or publish such panic-related datasets externally. To overcome this limitation, I plan to use a publicly available dataset as a proxy. By working on the **Amazon iPhone Customer Reviews dataset from Kaggle**, I aim to gain the **skills and hands-on experience** needed to apply similar methodologies back in my work environment.

Customer reviews are also high-volume, noisy, and unstructured, making them a strong analog for panic log data. By applying Natural Language Processing (NLP) techniques, I will extract, group, and interpret recurring feedback patterns to uncover meaningful insights.

---

2. Objectives

The goals of this project are:

1. Perform **exploratory data analysis (EDA)** to gain a high-level understanding of the dataset.
2. Identify the **Top 10 recurring feedback themes** using NLP-based clustering and topic modeling.
3. Extract and analyze the **Top 3 recurring negative concerns** in greater detail.
4. Build a model/pipeline that can **automatically extract and label customer concerns** from new reviews.
5. Compare model-driven insights with **LLM-based summaries** (OpenAI APIs, Perplexity).

6. Apply this framework to **upcoming iPhone release reviews** to extract emerging concerns in real time.

---

3. Methodology

The project will proceed in the following stages:

a. Data Preparation

- Collect and clean the Kaggle Amazon iPhone Customer Reviews dataset.
- Preprocess reviews (tokenization, stop-word removal, lemmatization).

b. Exploratory Data Analysis (EDA)

- Examine review distributions (ratings, length, sentiment trends).
- Extract preliminary themes using clustering and topic modeling techniques (e.g., TF-IDF, LDA, BERTopic).
- Supplement insights with LLM-based summarization (OpenAI APIs, Perplexity).

c. Deep Dive on Negative Concerns

- Focus on the Top 3 negative themes (e.g., battery life, camera, network issues).
- Perform sentiment and trend analysis for these concerns.

d. Concern Extraction Pipeline

- Develop a pipeline for **automated concern identification and grouping** using embeddings + clustering.
- Train classification models (e.g., logistic regression, transformer-based classifiers) to label feedback by domain (e.g., camera, battery, performance).

e. Validation

- Compare extracted themes with LLM-generated summaries.
- Evaluate overlap, divergence, and precision of concern labeling.

f. Future Application

- Deploy the concern extraction pipeline to upcoming iPhone reviews.
- Surface the **Top 10 concerns** from new feedback in real time.

---

4. Expected Outcomes

- A structured pipeline for analyzing large-scale, unstructured customer review data.
- A labeled set of **Top 10 customer concerns** and supporting evidence.
- Detailed insights into the **Top 3 negative feedback areas** affecting user satisfaction.
- A framework that generalizes to future reviews, enabling real-time extraction of emerging concerns.

---

5. Tools & Technologies

- **Data Source:** Kaggle Amazon iPhone Customer Reviews dataset
- **Languages & Libraries:** Python, Pandas, Scikit-learn, Hugging Face Transformers, NLTK/Spacy
- **Techniques:** TF-IDF, clustering, topic modeling, embeddings, text classification
- **LLM APIs:** OpenAI GPT, Perplexity for external validation
- **Visualization:** Matplotlib, Seaborn, Plotly

---

6. Significance

This project is designed not only to extract insights from iPhone customer feedback but also to serve as a **skill-building exercise** for addressing complex system log challenges in my professional work, particularly **user-space watchdog timeout issues**.

By applying techniques such as **log-like data summarization, clustering, and concern labeling** on a publicly available dataset, I can practice and refine methods that are directly transferable to production environments.

At the same time, the project demonstrates the broader value of combining **NLP-based automation** with **LLM validation** to analyze high-volume, unstructured data. The resulting framework provides actionable insights into customer sentiment and can be adapted both to **future product launches** in consumer domains and to **internal system monitoring in** technical domains.