# Project Planning & Management: Sentiment Analysis with Streamlit

Project: Deployed Product Review Sentiment Analysis Application

Author: Abdelrahman

Date: November 1, 2025

Version: 2.0

### 1. Project Proposal

#### 1.1. Overview

This project will develop and deploy a complete web application for analyzing and classifying public or internal complaints. The process involves sourcing or scraping raw feedback data, building an automated pipeline to clean and process this text, and then using a sophisticated machine learning model (e.g., a multi-task BERT) to perform **dual classification**. The model will predict both the **responsible department** (e.g., 'Traffic', 'Health') and the specific **issue type** (e.g., 'Nuisance', 'Vandalism'). The final, user-facing application will be built with Streamlit and deployed publicly via Streamlit Community Cloud.

#### 1.2. Scope

* **In-Scope:**
  + Sourcing or scraping a custom dataset of feedback/complaints.
  + An automated data pipeline script to handle text cleaning and preparation for a multi-task model.
  + Training and fine-tuning a model for high-accuracy **department and issue type** classification.
  + Developing an interactive Streamlit application for users to input text and view **both** classifications.
  + Public deployment of the application using Streamlit Community Cloud.
  + All code and documentation managed through a central GitHub repository.

**Out-of-Scope (for initial delivery):**

* Containerization with Docker.
* Deployment to major cloud platforms like Azure or Google Cloud.
* Real-time, continuous data ingestion (the initial dataset will be a fixed set).

#### 1.3. Objectives

* To demonstrate a full end-to-end MLOps workflow, from data collection to a live web application.
* To apply data sourcing/scraping techniques for a specialized, dual-label dataset.
* To implement a robust data pipeline for complex NLP classification tasks.
* To leverage a state-of-the-art model for **multi-task text classification**.
* To create an intuitive user interface with Streamlit that clearly presents the dual-classification results.
* To successfully manage a team project using version control (Git/GitHub).

### 2. Project Plan

#### 2.1. Timeline & Milestones

* **Phase 1: Planning & Data Collection (Due: Sept 21, 2025)**
  + Finalize all planning, requirements, and system design documents.
  + **Deliverable:** Sourced and cleaned dataset is available on GitHub. Must include labels for department and issue\_type. EDA notebook is complete.
* **Phase 2: Pipeline & Model Development (Due: Oct 30, 2025)**
  + **Deliverable 1:** The automated data pipeline script is functional.
  + **Deliverable 2:** The **classification model(s)** are trained, fine-tuned, and saved. Evaluation report (for **both tasks**) is complete.
* **Phase 3: Application & Deployment (Due: Nov 15, 2025)**
  + **Deliverable:** The Streamlit application is fully functional, displays both outputs, and is deployed to Streamlit Community Cloud.
* **Phase 4: Finalization (Due: Nov 20, 2025)**
  + Complete all testing, final reports, and prepare the project presentation.

#### 2.2. Task Assignment & Roles

The team consists of four members with specialized roles to ensure a smooth workflow.

| **Role** | **Responsibilities** | **Assigned To** |
| --- | --- | --- |
| **Team Lead & ML/DL Specialist** | Manages all project documentation, oversees the GitHub repository, trains and fine-tunes the BERT model for sentiment classification. | Abdelrahman Mohmmed |
| **Data Collection & EDA Specialist** | Scrapes product review data from target websites, performs extensive data cleaning, and creates visualizations for Exploratory Data Analysis (EDA). | Youssef Elgamel |
| **Streamlit Developer** | Builds the user interface and all interactive components of the Streamlit application. Integrates the model and pipeline outputs into the app. | Marwan Emad |
| **Data Pipeline Engineer** | Develops the automated script that connects all stages: takes raw data, processes it, and prepares it for the model. Ensures the pipeline runs efficiently. | Ali Adel |

### 3. Risk Assessment & Mitigation

| **Risk** | **Likelihood** | **Impact** | **Mitigation Strategy** |
| --- | --- | --- | --- |
| **Web Scraping Blocked** | High | High | Implement ethical scraping practices (headers, delays). The **Data Collection & EDA Specialist** will have a backup public dataset from Kaggle ready. |
| **Poor Quality Dataset** | Medium | High | The **Data Collection & EDA Specialist** will perform initial data profiling to assess quality early. The pipeline will include data validation steps. |
| **Low Model Accuracy** | Medium | Medium | The **ML/DL Specialist** will experiment with multiple models (e.g., DistilBERT) and perform hyperparameter tuning. |
| **Pipeline Integration Fails** | Low | High | Adhere strictly to a defined data schema for handoffs. The **Data Pipeline Engineer** and **ML/DL Specialist** will collaborate closely and conduct early integration tests. |
| **Model Too Large for Deployment** | Medium | High | The **ML/DL Specialist** will use a smaller, distilled version of BERT (e.g., DistilBERT) to ensure it fits within Streamlit Community Cloud's memory limits. |

### 4. Key Performance Indicators (KPIs)

* **Model Performance:** F1-Score of > 0.80 for the sentiment model.
* **Pipeline Reliability:** 99% success rate for automated pipeline runs on test data.
* **Data Throughput:** Pipeline can process at least 1,000 reviews per minute.