**Technical Assignment Report**

**Author:** Amin Khodamoradi  
**Date:** 14/06/2025

**Objective**

The objective of this assignment was to build a text classification pipeline that:

* Classifies texts into one of the following categories: cult, paranormal, or dramatic (using task.csv).
* Offers a demo web interface for prediction using Gradio.
* Provides a REST API for programmatic access using FastAPI.
* Is containerized via Docker for consistent deployment.
* Optionally, extends to a multi-label classification using bonus\_task.csv.

**Step 1: Data Investigation**

The dataset task.csv contained movie metadata with the following columns:

* Title: Movie title
* Synopsis: Movie description (used for classification)
* Tag: Ground-truth category

Basic statistics and value counts were extracted to understand class distribution.

**Step 2: Preprocessing**

* Converted all text to lowercase.
* Removed special characters and punctuation using regex.
* Created a clean\_synopsis column for the cleaned text.
* Encoded the Tag labels using LabelEncoder into numeric format.

**Step 3: Train-Test Split**

Used an 80/20 split on the cleaned synopsis and encoded labels:

* X\_train, X\_test: Vectorized text features
* y\_train, y\_test: Encoded class labels

Stratification was used to preserve label distribution.

**Step 4: Feature Engineering (TF-IDF)**

Used TfidfVectorizer to convert cleaned text into numerical feature vectors:

* Max features: 5000
* Fitted on training data, transformed both train and test sets

**Step 5: Model Training and Evaluation**

Trained a logistic regression model (LogisticRegression) as a baseline classifier.  
Other models such as RandomForestClassifier and GradientBoostingClassifier were explored.

Evaluation metrics:

* Precision, Recall, F1-score (using classification\_report)
* Label decoding used to interpret predictions

precision recall f1-score support

cult 0.72 0.97 0.83 207

dramatic 1.00 0.03 0.06 34

paranormal 0.82 0.38 0.52 73

accuracy 0.73 314

macro avg 0.85 0.46 0.47 314

weighted avg 0.77 0.73 0.67 314

**Step 6: Gradio Demo App**

Developed a user-friendly demo with Gradio:

* Allows users to input/paste text
* Displays predicted category in real-time
* Bound to 0.0.0.0:7860 for Docker networking

A screenshot of a computer

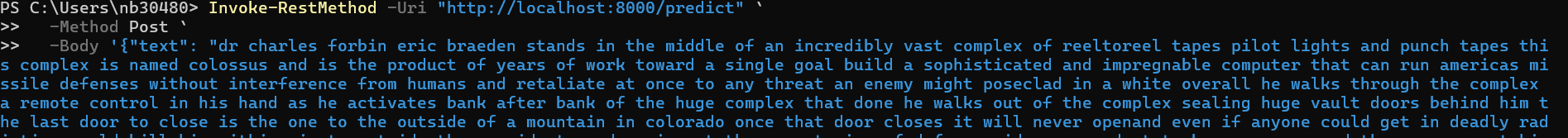
AI-generated content may be incorrect.

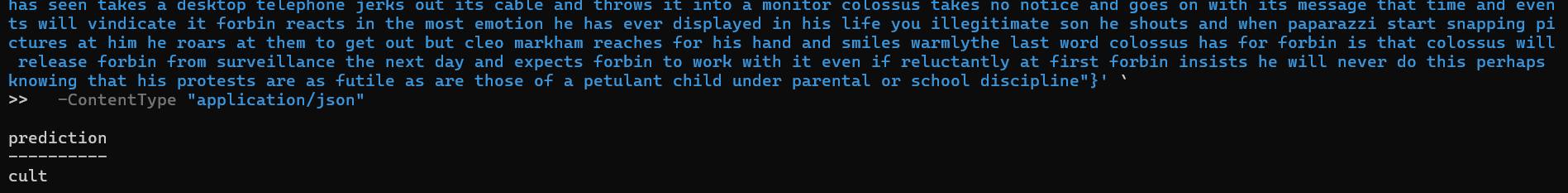
**Step 7: FastAPI REST API**

Built a REST endpoint /predict using FastAPI:

* Accepts JSON payload: { "text": "your input here" }
* Returns: { "prediction": "cult" }

Tested with Swagger UI (/docs) and PowerShell Invoke-RestMethod.





A black and white striped background

AI-generated content may be incorrect.

**Step 8: Dockerization**

Created a Dockerfile to package the app:

* Python 3.11.3 base image
* Installs dependencies via requirements.txt
* Exposes ports 8000 (API) and 7860 (Gradio)
* Entrypoint defaults to FastAPI, but Gradio can be run with CMD override

**Step 9: Docker Compose (Run Both Services Together)**

Used docker-compose.yml to run both FastAPI and Gradio apps concurrently:

1. **Main task:**

* fastapi\_service at port 8000
* gradio\_service at port 7860

1. **Bonus task:**

* RAG\_fastapi\_service at port 8001
* RAG\_gradio\_service at port 7861
* Shared image and codebase

**Bonus Task: Multi-Label Classification using RAG**

For the bonus task, a Retrieval-Augmented Generation (RAG) approach was implemented to perform multi-label classification using bonus\_task.csv. The key steps:

* Extracted unique labels from the dataset and created textual definitions for each.
* Embedded both label descriptions and input synopses using the high-performing all-mpnet-base-v2 SentenceTransformer model.
* Calculated cosine similarities between input text and label embeddings.
* Used a top-k retrieval strategy to select the most semantically similar labels.

**Results:**

* Micro F1 Score: **0.1295**
* Macro F1 Score: **0.0704**

This approach outperformed earlier attempts using smaller models or fixed thresholds and demonstrated the effectiveness of semantic embeddings for zero-shot multi-label prediction.

The RAG-based classifier was kept separate from the original single-label pipeline and was integrated into its own Gradio and FastAPI interfaces for testing and demonstration purposes.

**Tools & Libraries Used**

* Python 3.11
* pandas, scikit-learn
* Gradio
* FastAPI
* Docker & Docker Compose
* sentence-transformers (all-mpnet-base-v2)

**Conclusion**

The project demonstrates a complete NLP deployment pipeline:

* Preprocessing & training
* Evaluation
* Real-time demo (Gradio)
* API integration (FastAPI)
* Scalable deployment (Docker Compose)

The bonus task implementation extends this work into multi-label territory using a RAG-based approach that blends retrieval and inference in a zero-shot setting. The solution is modular and extensible, ready to integrate larger models or full fine-tuning if needed.