REPORT CATEGORY CLASSIFICATION MODEL

[Short Report]

1. Dataset Creation

A synthetic dataset of **800 user queries** was created across four categories:

- Career / Job
- Love / Relationship
- Health / Wellness
- Finance / Money

Each category contained **200 queries** generated with realistic, user-like phrases and variations. This ensured a balanced dataset and improved robustness compared to the smaller 400-query dataset.

2. Techniques Used

- **Text Preprocessing** – lowercasing, stopword removal, tokenization, and feature extraction using **TF-IDF Vectorization**.
- **Models Trained**:
  - Logistic Regression
  - Multinomial Naive Bayes
- **Evaluation Metrics** – accuracy, precision, recall, F1-score.
- **Confusion Matrix Visualization** – used to analyse misclassifications.

3. Evaluation Results

Both Logistic Regression and Naive Bayes achieved **very high accuracy** on the balanced dataset.
However, Logistic Regression showed **better generalization** when tested on unseen, tricky queries that overlapped multiple categories.
This makes Logistic Regression more reliable for real-world usage where queries may not be straightforward.

4. Chosen Final Model

The **Logistic Regression model** was selected as the final classifier due to:

- Stronger generalization ability
- Balanced performance across all metrics
- Robustness against overlapping queries

The trained Logistic Regression model and TF-IDF vectorizer were saved as serialized .pkl files using **Joblib**, making them reusable in deployment environments like **Streamlit or Flask** .