README

Project Overview

This project aims to build and evaluate a BERT-based sentiment classification model using both real-world and synthetic datasets. Key steps include data preprocessing, synthetic data generation, model training, evaluation, and performance comparison. The results highlight the strengths and limitations of the model across the two datasets.

Contents of the ZIP File

- `final.py`: The consolidated Python script containing all the steps from data preprocessing to model evaluation.
- `model.ipynb`: Jupyter Notebook detailing the model training and evaluation process.
- `data_cleaning.ipynb`: Jupyter Notebook focused on data cleaning and preprocessing.
- `synthetic.ipynb`: Jupyter Notebook for generating and analyzing synthetic data.
- `data/` Folder containing:
  - `womens_clothing_ecommerce_reviews.csv`: The raw real-world dataset.
  - `cleaned_reviews.csv`: The cleaned real-world dataset.
  - `synthetic_reviews.csv`: The generated synthetic dataset.

Dependencies

The project requires the following Python libraries:

- `pandas`
- `torch`
- `transformers`
- `scikit-learn`
- `matplotlib`

Ensure that Python 3.7+ is installed.

Setup Instructions

- Install the required dependencies:

```
pip install pandas torch transformers scikit-learn matplotlib
```

- Run the final.py script:

```
python final.py
```

Detailed File Descriptions

`final.py`

The consolidated Python script performs the following steps:

1. Loads and cleans the real-world dataset (`data/cleaned_reviews.csv`).
2. Generates synthetic data and saves it as ()`data/synthetic_reviews.csv`).
3. Trains a BERT-based sentiment classification model on the synthetic data.
4. Evaluates the model on both synthetic and real datasets.
5. Displays examples of good and bad predictions for synthetic data in the console.
6. Saves correct and incorrect predictions for the real dataset to CSV files.
7. Visualizes performance metrics in a comparison chart.

`data_cleaning.ipynb`

Focuses on cleaning and preprocessing the real-world dataset, including:

1. Removing rows with missing review text.
2. Mapping ratings to sentiment categories (e.g., Satisfied/Dissatisfied/Neutral).
3. Saving the cleaned dataset as `data/cleaned_reviews.csv`.

`synthetic.ipynb`

Details the generation of synthetic data:

1. Uses predefined templates and sentiment-based word substitutions.
2. Generates synthetic reviews for both positive and negative sentiments.
3. Saves the generated synthetic dataset as `data/synthetic_reviews.csv`.

`model.ipynb`

Provides detailed, step-by-step guidance on:

1. Tokenization and data preparation for the BERT model.
2. Training the BERT-based sentiment classification model.
3. Evaluating the model on validation datasets.

`data/`

Contains the following files:

- `womens_clothing_ecommerce_reviews.csv`: The raw real-world dataset.
- `cleaned_reviews.csv`: The cleaned real-world dataset.
- `synthetic_reviews.csv`: The generated synthetic dataset.