Master in Data Science AUEB 2024-2025

Practical Data Science - Assignment 2

Stylianos Giagkos f3352410

Food Hazard Detection Challenge

This repository contains the solution for the **Food Hazard Detection Challenge**. The challenge involves classifying food safety-related incidents based on short titles and long descriptions. The solution leverages both **Finetuned PubMedBERT** and **LightGBM (LGBM)** models for classification into hazard-category, product-category, hazard, and product classifications.

Overview

Project Workflow:

1. **Exploratory Data Analysis (EDA)**: Data cleaning, exploration, and visualization.
2. **Modeling**: Two different approaches are used for classification:
    - **Finetuned PubMedBERT**: Fine-tuned on the dataset to classify food hazard-related texts.
    - **LightGBM**: A gradient boosting model for classification based on features derived from the data.
3. **Evaluation**: Both models are evaluated on performance metrics such as accuracy, precision, recall, and F1-score.
4. **Training and Submission**: Generation of final predictions based on the final optimal model and submission in the required format on CodaLab competition.

Subtasks (Performed Separately for Title and Text):

Subtask 1:

- **Classify hazard-category**: Classifies the general hazard type.
- **Classify product-category**: Classifies the general product type.

Subtask 2:

- **Classify hazard**: Classifies the specific hazard type.
- **Classify product**: Classifies the specific product type.

Directory Contents Description:

- **Data**: Contains datasets used for the Food Hazard Detection Challenge.
    - **validation_data**: Folder with data used for validating model performance on CodaLab Food Hazard Detection Challenge.

  - **incidents_train.csv**: Training dataset containing labeled food hazard incidents.
- **Notebooks**: Contains Jupyter notebooks for analysis and model development.
  - **Benchmarks notebooks**: Includes notebooks for baseline models and benchmark results.
  - **Submission notebooks**: Notebooks used to prepare final submission files.
  - **EDA notebook PDS A2 Food Hazard**: Notebook for exploratory data analysis (EDA) related to the Food Hazard Detection Challenge.
- **Submissions**: Contains various submission files.
  - **lgbm_submission_title.zip**: Submission file for the model using LightGBM on title data.
  - **submission_finetuned_PubMedBERT.zip**: Submission file for a fine-tuned PubMedBERT model.
  - **submission_v1.zip**: Version 1 of a submission file using LightGBM.
  - **submission_v3.zip**: Version 3 of a submission file using LightGBM.

Files

The following files are included in the repository:

- **EDA PDS A2 Food Hazard Detection.ipynb**:
  Performs exploratory data analysis on the dataset, including data cleaning and visualization.

- **BENCHMARKS Finetuned PubMedBERT PDS A2 Food Hazard Detection.ipynb**:
  Implements the fine-tuned PubMedBERT model for food hazard detection, classifying hazard-category, product-category, hazard, and product for both titles and descriptions.

- **BENCHMARKS LGBM PDS A2 Food Hazard Detection .ipynb**:
  Implements the LightGBM model for classification, similar to the PubMedBERT model, covering all subtasks for title and text classification.

- **SUBMISSION Finetuned PubMedBERT PDS A2 Food Hazard Detection.ipynb**:
  Retrains the PubMedBERT model based on the results from the benchmark evaluation and generates the final submission file.

- **SUBMISSION LGBM PDS A2 Food Hazard Detection .ipynb**:
  Retrains the LightGBM model based on the results from the benchmark evaluation and generates the final submission file.

Requirements

Make sure to install the required dependencies before running the code. You can use the following pip command to install the necessary packages:

```
pip install torch lightgbm pandas scikit-learn matplotlib tqdm transformers nltk numpy
```

Additional Libraries:

- **pandas**
- **re**
- **nltk**
- **scikit-learn**
- **torch**
- **transformers**
- **lightgbm**
- **matplotlib**
- **numpy**

How to Re-run the Solution

Step 1: Data Preparation

Ensure that the dataset is correctly placed in the expected directory. Adjust file paths if necessary.

Step 2: Exploratory Data Analysis

Run `EDA PDS A2 Food Hazard Detection.ipynb` to clean and visualize the data.

Step 3: Model Training

You can choose between:

- **Finetuned PubMedBERT**: Run `BENCHMARKS Finetuned PubMedBERT PDS A2 Food Hazard Detection.ipynb`.
- **LightGBM**: Run `BENCHMARKS LGBM PDS A2 Food Hazard Detection .ipynb`.

Step 4: Model Evaluation

Both models evaluate accuracy, precision, recall, and F1-score.

Step 5: Retraining for Submission

Retrain using the submission notebooks to create final submissions for competitions based on the best model of the respective benchmarks notebooks:

- **For PubMedBERT**: Run `SUBMISSION Finetuned PubMedBERT PDS A2 Food Hazard Detection.ipynb`.
- **For LightGBM**: Run `SUBMISSION LGBM PDS A2 Food Hazard Detection .ipynb`.

Results

Finetuned PubMedBERT Model (Title-based):

Task	F1-Score
hazard-category (Title)	0.8288
product-category (Title)	0.7494
hazard (Title)	0.5899
product (Title)	0.2172

Finetuned PubMedBERT Model (Text-based):

Task	F1-Score
hazard-category (Text)	0.9459
product-category (Text)	0.7583
hazard (Text)	0.8166
product (Text)	0.2331

LightGBM Model (Title-based):

Task	F1-Score
hazard-category (Title)	0.7614
product-category (Title)	0.5926
hazard (Title)	0.5533
product (Title)	0.0798

LightGBM Model (Text-based):

Task	F1-Score
hazard-category (Text)	0.9129
product-category (Text)	0.6682
hazard (Text)	0.7671
product (Text)	0.0479

Competition Results for Subtask 1 (Hazard-category, Product-category)

The following shows the results for hazard-category and product-category tasks for Subtask 1:

Submission File	Score	Status
submission.zip	—	Failed
submission.zip	0.6428057851	Finished
lgbm_submission_title.zip	0.6428057851	Finished
submission_v3.zip	0.6252022003	Finished
submission_finetuned_PubMedBERT.zip	0.6992400644	Finished

Competition Results for Subtask 2 (Hazard, Product)

The following shows the results for hazard and product tasks for Subtask 1:

But the rating shows for Sub Task 2 a score of 0.3313 .

Submission File	Score	Status
submission.zip	—	Failed
submission.zip	0.00000	Finished
lgbm_submission_title.zip	0.00000	Finished
submission_v3.zip	0.00000	Finished
submission_finetuned_PubMedBERT.zip	0.00000	Finished

Overfitting in the Competition Results

The competition results suggest potential overfitting in the models. The scores for the `submission_finetuned_PubMedBERT.zip` file were significantly higher for the training set compared to the test set, which may indicate that the model is overfitting to the training data.

Conclusion

This solution combines advanced NLP (Finetuned PubMedBERT) and traditional machine learning (LightGBM) techniques to classify food hazard data. The solution addresses both general and specific hazard and product classification tasks. Both models were trained and evaluated, with their results summarized above.The submission models are retrained based on the benchmark results to generate the final predictions. You can re-run the solution by following the provided instructions and reproduce the results with the corresponding datasets and models.

References:

- ChatGPT: For quick assistance, code help, debugging, and guidance on various data science topics.
- Stack Overflow: For troubleshooting and solutions to coding challenges.
- Towards Data Science and Medium: For articles and tutorials on data science techniques and best practices.