Multilingual Review Rating Prediction using BERT and Regression

George-Cristian Serban

1 Problem Statement

The task is to perform sentiment analysis on product reviews using BERT models and Logistic Regression on multilingual data. Sentiment analysis aims to classify the sentiment expressed in a text as positive, negative, or neutral. For this project, we focus on predicting the star ratings (1-5) based on the review texts. The goal is to find out how training for a specific language can be relevant for another language.

2 Proposed Solution

2.1 Theoretical Aspects

The solution involves using the multilingual DistilBERT model, a smaller and faster version of BERT, for obtaining text embeddings and training classifiers. DistilBERT retains 97% of BERT's performance while being 60% faster.

Formally, let $T = \{t_1, t_2, \dots, t_n\}$ be a set of tokenized review texts, and $Y = \{y_1, y_2, \dots, y_n\}$ be the corresponding star ratings. The DistilBERT model encodes each text t_i into a fixed-size embedding e_i. A classifier f is then trained to predict $\hat{y_i}$ from e_i. We use Logistic Regression classifiers as well as BERT Classifiers.

2.2 Data Set Used in Application

We use two datasets:

- **Amazon Reviews:** Contains English reviews with star ratings.
  https://www.kaggle.com/datasets/tarkkaanko/amazon
- **LaRoSeDa:** Contains Romanian reviews with star ratings.
  https://github.com/ancatache/LaRoSeDa

The data is split 50/50 into training and test sets.

2.3 Application

The application consists of the following steps:

1. Data Preprocessing
2. Embedding Extraction using DistilBERT

3. Training Logistic Regression Classifiers
4. Training Bert Classifiers
5. Evaluation on Test Data

3 Implementation Details

3.1 Libraries/Functions Used

- `pandas, json, numpy`: For data manipulation and analysis.
- `torch`: For building and training neural networks.
- `transformers`: For using the DistilBERT model.
- `sklearn`: For evaluation metrics and logistic regression.

3.2 Original Contribution

- Development of a multilingual sentiment analysis pipeline using DistilBERT and Logistic Regression.
- Manipulation of datasets to facilitate training and evaluation across different languages.
- Implementation of training, saving, and loading mechanisms for multiple models, enabling easy reuse and comparison.
- Comprehensive evaluation of model performance using multiple metrics (accuracy, precision, recall, F1 score) on both English and Romanian reviews.
- Exploration of cross-language model applicability by evaluating models trained on one language with reviews in another language.

4 Experiments and Results

4.1 Regression Model Results and Interpretations

```
Amazon classifier on Amazon test set results:
Accuracy: 0.7928
Precision: 0.6927
Recall: 0.7928
F1 Score: 0.7344

Laroseda classifier on Amazon test set results:
Accuracy: 0.5059
Precision: 0.7187
Recall: 0.5059
F1 Score: 0.5728
```

```
Combined classifier on Amazon test set results:
Accuracy: 0.7908
Precision: 0.6940
Recall: 0.7908
F1 Score: 0.7333

Amazon classifier on Laroseda test set results:
Accuracy: 0.4223
Precision: 0.3888
Recall: 0.4223
F1 Score: 0.2582

Laroseda classifier on Laroseda test set results:
Accuracy: 0.6857
Precision: 0.6272
Recall: 0.6857
F1 Score: 0.6333

Combined classifier on Laroseda test set results:
Accuracy: 0.6857
Precision: 0.6348
Recall: 0.6857
F1 Score: 0.6319
```

- **Amazon classifier on Amazon test set:** High accuracy and F1 score indicate that the model performs well when evaluated on the same dataset it was trained on.
- **Laroseda classifier on Amazon test set:** Moderate accuracy and F1 score suggest that a model trained on Romanian reviews has some applicability to English reviews but is not optimal.
- **Combined classifier on Amazon test set:** Similar performance to the Amazon-specific model, indicating that combining datasets does not degrade nor increase performance.
- **Amazon classifier on Laroseda test set:** Low accuracy and F1 score show that a model trained on English reviews does not perform well on Romanian reviews.
- **Laroseda classifier on Laroseda test set:** Reasonably high accuracy and F1 score, indicating good performance on the same language dataset.
- **Combined classifier on Laroseda test set:** Comparable to the Laroseda-specific model, showing that combining datasets does not influence performance.

4.2 BERT Model Results and Interpretation

```
Amazon Reviews with Amazon model results:
Accuracy: 0.7985
```

```
Precision: 0.6377
Recall: 0.7985
F1 Score: 0.7091

Amazon Reviews with Laroseda model results:
Accuracy: 0.3614
Precision: 0.7226
Recall: 0.3614
F1 Score: 0.4452

Amazon Reviews with Combined model results:
Accuracy: 0.8132
Precision: 0.7327
Recall: 0.8132
F1 Score: 0.7586

Amazon Reviews with Untrained model results:
Accuracy: 0.7660
Precision: 0.6717
Recall: 0.7660
F1 Score: 0.6990

Laroseda Reviews with Amazon model results:
Accuracy: 0.4187
Precision: 0.1753
Recall: 0.4187
F1 Score: 0.2471

Laroseda Reviews with Laroseda model results:
Accuracy: 0.7490
Precision: 0.7118
Recall: 0.7490
F1 Score: 0.7208

Laroseda Reviews with Combined model results:
Accuracy: 0.7477
Precision: 0.6924
Recall: 0.7477
F1 Score: 0.7041

Laroseda Reviews with Untrained model results:
Accuracy: 0.4087
Precision: 0.3313
Recall: 0.4087
F1 Score: 0.2936
```

- **Amazon Reviews with Amazon model:** High accuracy and F1 score indicate excellent performance on the same language dataset.

- **Amazon Reviews with Laroseda model:** Low accuracy and F1 score, suggesting poor transferability from Romanian to English.
- **Amazon Reviews with Combined model:** Best performance, demonstrating the benefit of multilingual training.
- **Amazon Reviews with Untrained model:** Surprisingly decent performance, indicating that even an untrained model has some inherent ability to classify text, likely due to its pretrained nature.
- **Laroseda Reviews with Amazon model:** Low accuracy and F1 score, again showing poor cross-language applicability, however better than Romanian to English likely due to English being trained on more data in the base model.
- **Laroseda Reviews with Laroseda model:** High accuracy and F1 score, confirming good performance on the same language dataset.
- **Laroseda Reviews with Combined model:** High performance, slightly less than the Laroseda-specific model, suggesting that multilingual training is beneficial.
- **Laroseda Reviews with Untrained model:** Low accuracy and F1 score, indicating that the untrained model's performance is not sufficient for practical use in a lower-resource language such as Romanian compared to the previous English performance.

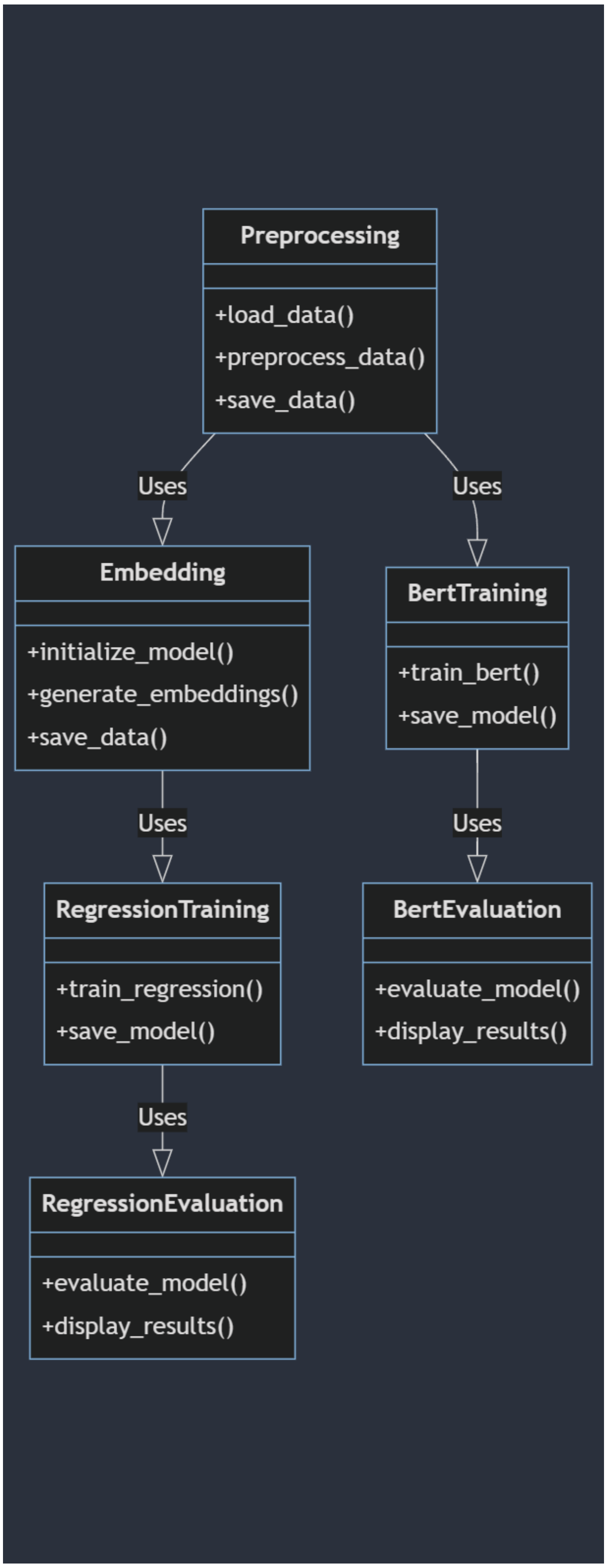

Figure 1: Application Diagram