Quora Question Pair Duplicate Detection

Project Documentation

By
Omar Baig

June 20, 2024

Contents

1 Project Overview

This project aims to determine whether a pair of questions from Quora are duplicates. Given a dataset of question pairs, various machine learning models were trained and evaluated to achieve the highest accuracy in predicting duplicate questions. The final model chosen was a Random Forest classifier due to its superior performance compared to other models.

2 Dataset

The dataset used in this project consists of 404,000 rows and 5 columns. The data was split into training and test sets in an 80-20 ratio. The key columns in the dataset include:

- `question1`: The first question in the pair.
- `question2`: The second question in the pair.
- `is_duplicate`: A label indicating whether the questions are duplicates (1) or not (0).

3 Preprocessing

Data preprocessing involved several steps to clean and prepare the text data for model training:

1. **Text Cleaning**:
    - Lowercasing the text.
    - Replacing special characters and numbers with their string equivalents.
    - Decontracting words (e.g., converting "can't" to "cannot").
    - Removing HTML tags.
    - Removing punctuations.
2. **Tokenization**:
    - Tokenizing the text into words.
    - Removing stopwords.
3. **Feature Engineering**:
    - **Basic Features**:
        - `word_common`: The number of common words between the two questions.
        - `word_total`: The total number of words in both questions.

  - `word_share`: The ratio of common words to the total words.
- **Advanced Token Features**: Common non-stopwords and stopwords, common tokens, matching first and last words.
- **Fuzzy Features**: Fuzzy string matching ratios using the `fuzzywuzzy` library.

4 Feature Scaling

Feature scaling was performed to normalize the numerical features. The methods used included:

- **Word Embeddings**: Using Word2Vec embeddings with dimensions of 50, 100, and 300. The 300-dimensional embeddings performed the best and were chosen for the final model.
- **Standardization**: Standardizing features to have zero mean and unit variance.

5 Model Training

Multiple models were trained and evaluated using the preprocessed features:

1. **Neural Network**:
   - Test Loss: 0.4847
   - Test Accuracy: 0.7308
2. **XGBoost**:
   - Logloss (Train): 0.1759
   - Logloss (Validation): 0.4889
3. **Decision Tree**:
   - Test Accuracy: 0.6802
4. **Random Forest**:
   - Test Accuracy: 0.8372

6 RandomForest Report

```
Accuracy: 0.83723259552368
Confusion Matrix:
[[45048  6192]
 [ 6971 22659]]
```

```
Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.88      0.87     51240
           1       0.79      0.76      0.77     29630

    accuracy                           0.84     80870
   macro avg       0.83      0.82      0.82     80870
weighted avg       0.84      0.84      0.84     80870

F1 Score:  0.7749183495494263
Log-Loss: 0.3739482366573238
```

7 Conclusion

This project successfully implemented a machine learning solution to identify duplicate questions on Quora. The Random Forest model demonstrated the best performance and was deployed using a Flask web application, making it accessible for real-time predictions. The combination of extensive preprocessing, feature engineering, and model evaluation ensured the robustness and accuracy of the final solution.