Question Pair Similarity Classification

Submitted by: Md. Mahmudul Islam, Email: shimulmahmud4097@gmail.com

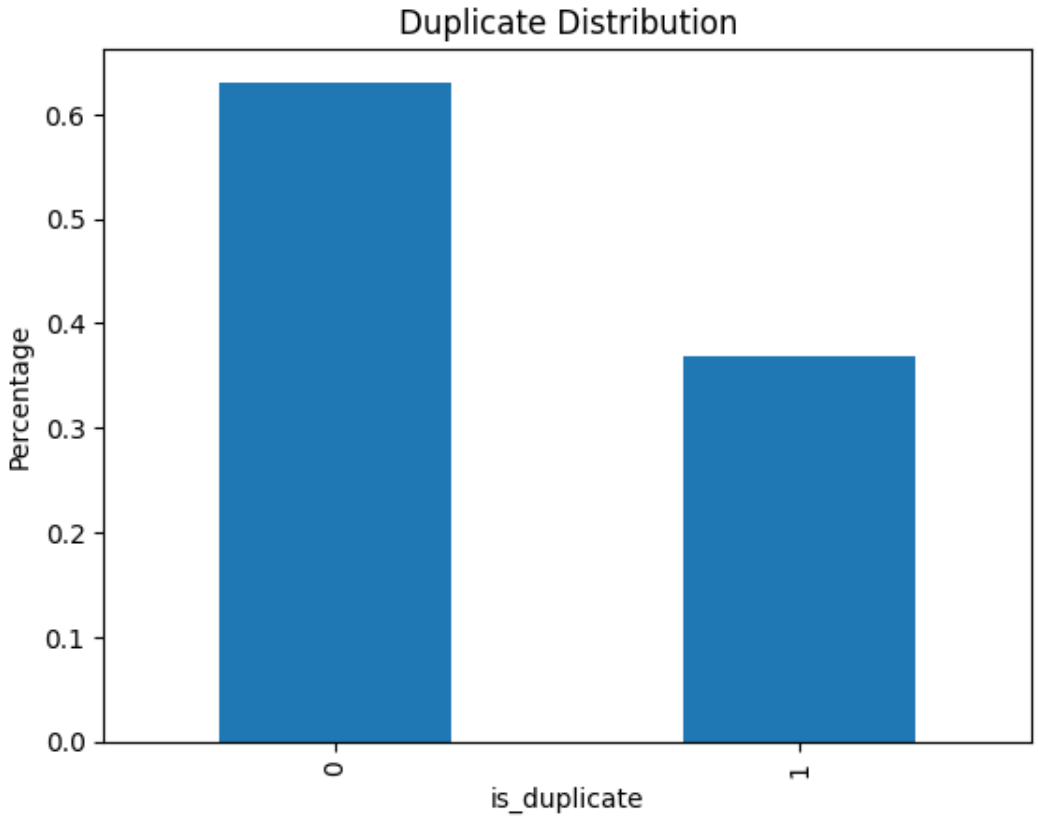
Contact: 01737422300
Date: May 10, 2025

Objective

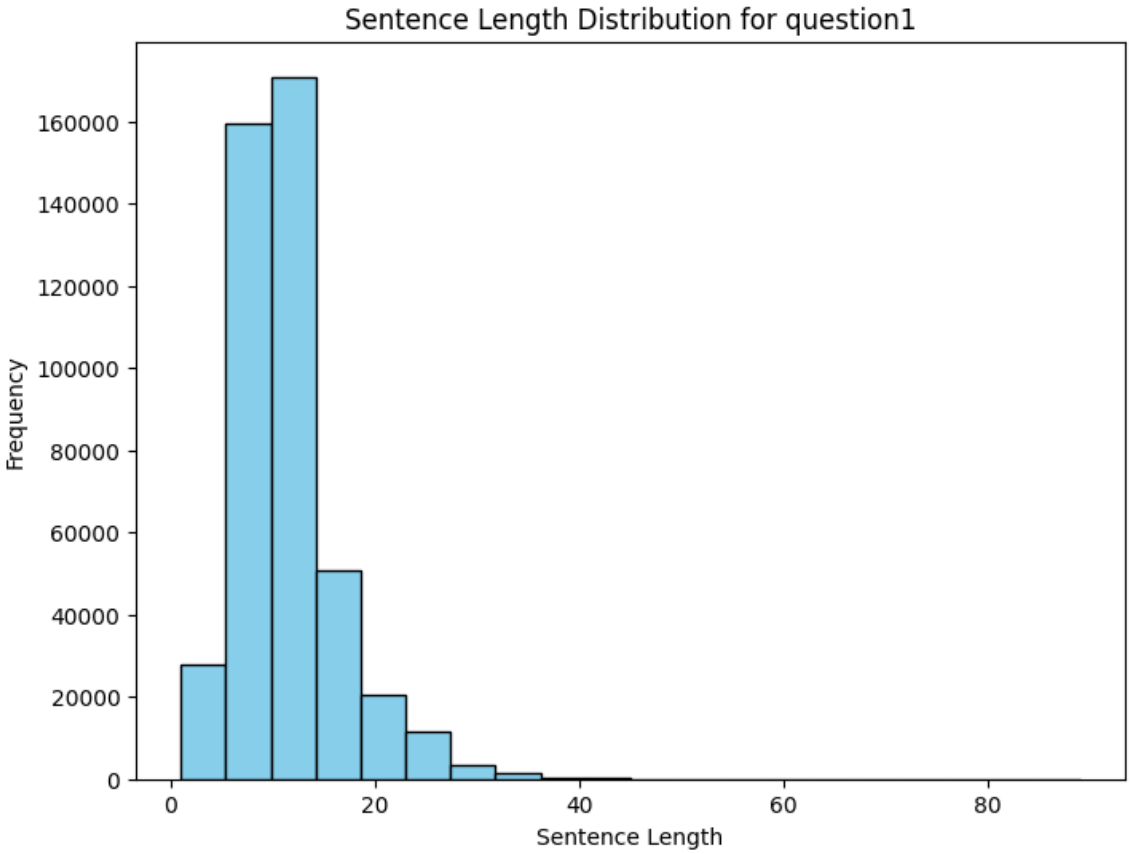
The goal of this project is to build a classification system that determines whether two questions from Quora are semantically similar (duplicates) or not. This task plays a vital role in improving the user experience by avoiding redundant content on Q&A platforms.

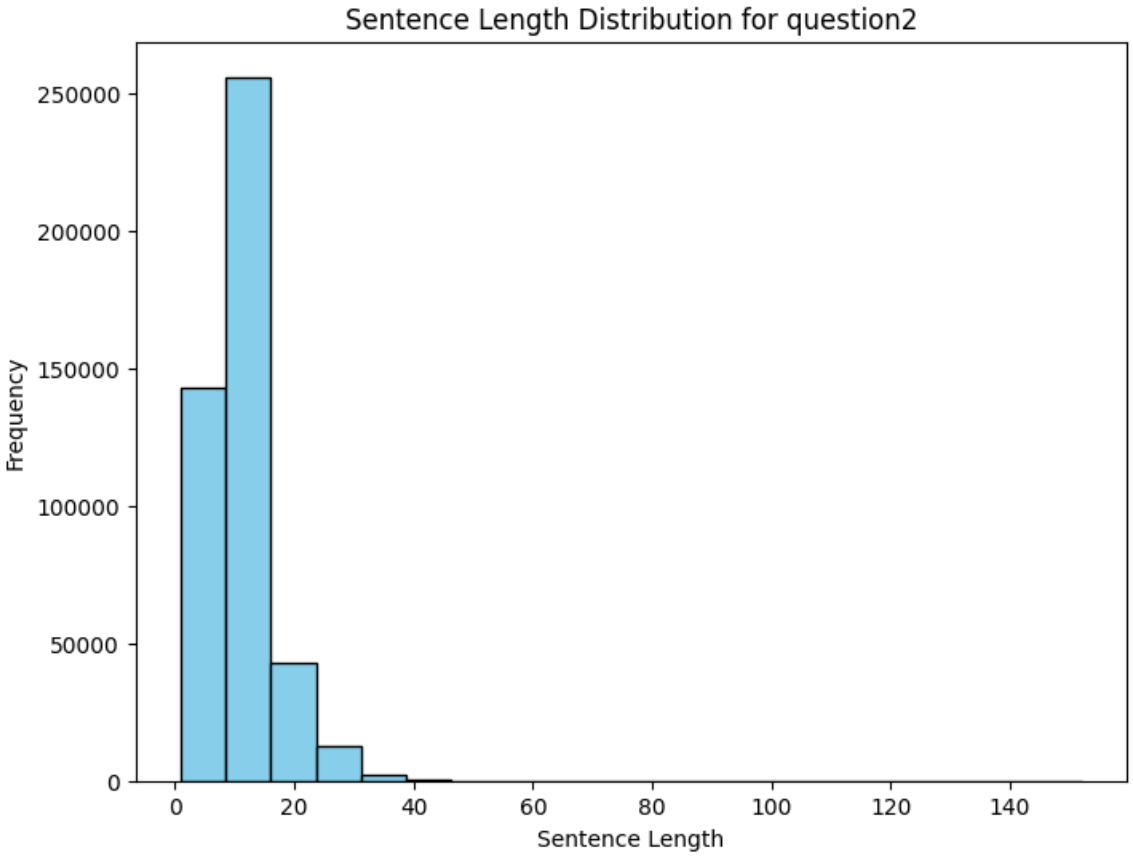
Exploratory Data Analysis (EDA)

- Analyzed class distribution of duplicates vs non-duplicates, Identified class imbalance (more non-duplicate samples)

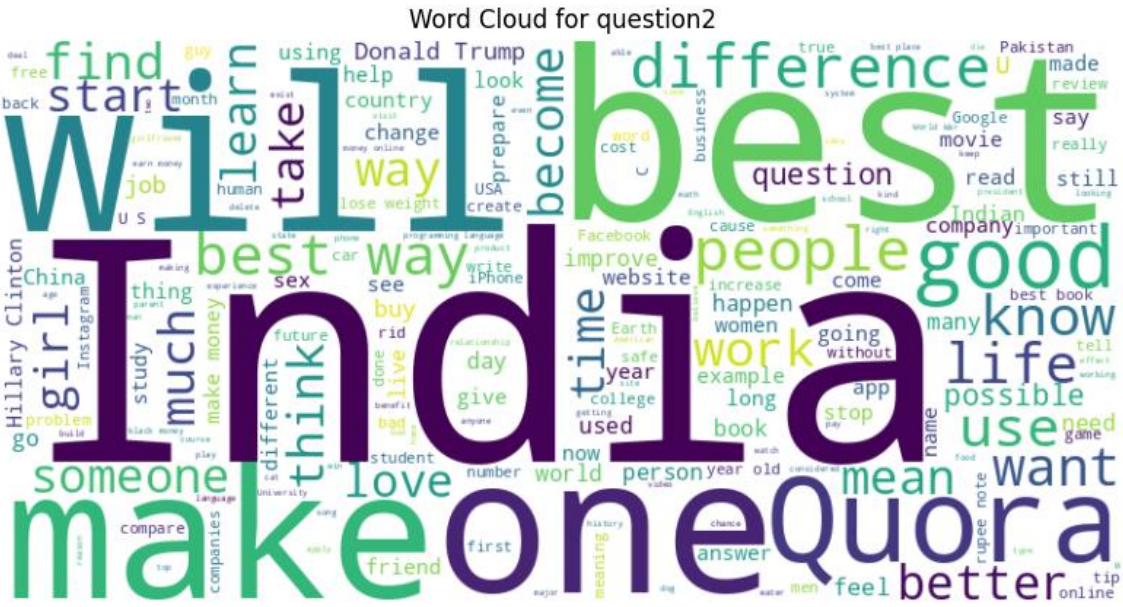


- Checked text lengths



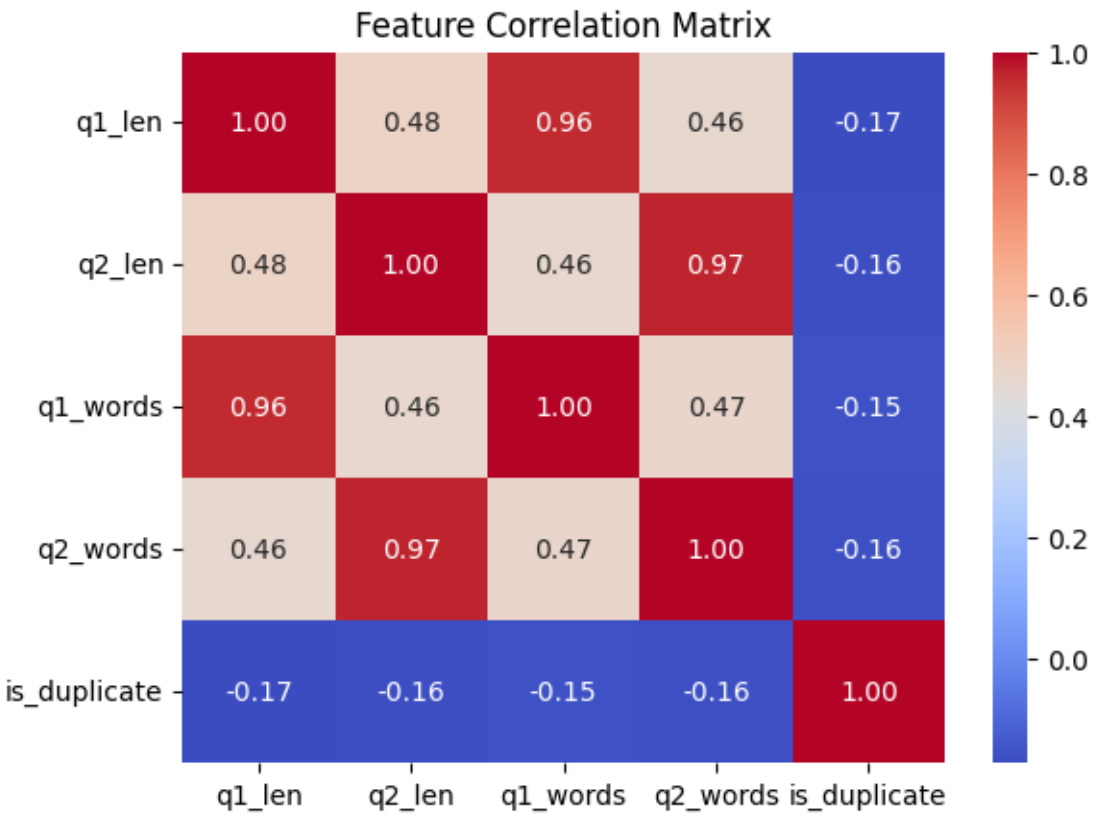


- Generated word clouds to visualize common terms in both question1 and question2



Text Preprocessing

- Converted text to lowercase
- Removed punctuation, special characters, and stopwords
- Tokenized text
- Applied lemmatization to normalize words
- Created new features: question length, question words count and visualize the correlation



- Used TF-IDF Vectorization to transform the questions into numerical representations.

Model Building

Implemented multiple models including:

- Logistic Regression (baseline)
- Artificial Neural Network (ANN) using Keras with Dense layers and dropout regularization
- LSTM (Long Short-Term Memory) model using sequential word embeddings

Model Evaluation

Models were evaluated using:

- Accuracy
- Precision
- Recall
- F1-score
- Confusion Matrix
- AUC-ROC Curve

Hyperparameter Tuning (RandomizedSearchCV)

- Optimized learning rate, batch size, and dropout
- Experimented with different activation functions (ReLU, sigmoid)
- Applied EarlyStopping and Model Checkpoint in training

The best performance was achieved with LSTM, which achieved 98% accuracy, due to its ability to capture sequence dependencies.