DUPLICATE QUESTION PAIR DETECTION(QUORA)

TEAM 9

MOTIVATION

Online Q&A platforms like Quora, Stack Overflow, and Reddit experience a high influx of questions, many of which are semantically similar.

Problem:

Duplicate questions can clutter the platform, making it harder for users to find relevant answers and for moderators to manage content.

Impact:

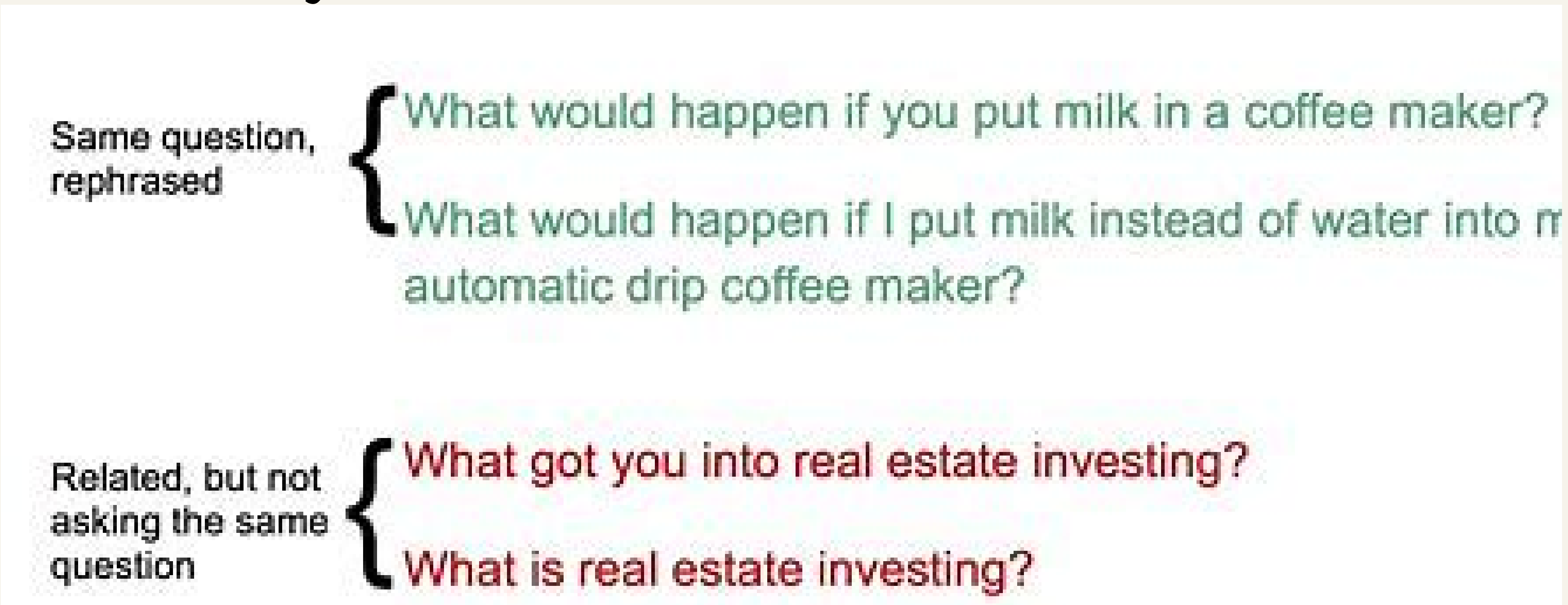
Detecting duplicates improves efficiency, reduces search time, and increases overall user satisfaction.

Reduces redundancy by ensuring similar questions are linked or merged.

PROBLEM STATEMENT

In large-scale Q&A platforms like Quora,Stack Overflow, and Reddit, users often post questions that are semantically similar but phrased differently. This results in a cluttered repository of redundant content, which makes it difficult for users to find relevant answers efficiently. Detecting the duplicate question pairs can significantly improve the user experience by reducing redundancy, improving search accuracy.

Example:



Same question, rephrased
- What would happen if you put milk in a coffee maker?
- What would happen if I put milk instead of water into m automatic drip coffee maker?

Related, but not asking the same question
- What got you into real estate investing?
- What is real estate investing?

PROPOSED PIPELINE

1. Data Collection:

- We will use the Quora Question Pairs dataset, where each pair is labeled as 'duplicate' or 'non-duplicate.'

2. Preprocessing:

- We aim to clean text, tokenize, remove stopwords, and lemmatize to normalize the data.

3. Feature Extraction:

- We will convert text to vectors using TF-IDF and embeddings like Word2Vec or BERT.

4. Similarity Calculation:

- To compare question pairs, we want to calculate semantic similarity using techniques like Cosine Similarity or Euclidean distance.

5. Model Training:

- We will train models such as Logistic Regression, SVM, or deep learning models like LSTM or BERT.

6. Evaluation:

- We will evaluate our model using Accuracy, Precision, Recall, F1 Score, and cross-validation techniques.

TOOLS AND TECHNOLOGIES

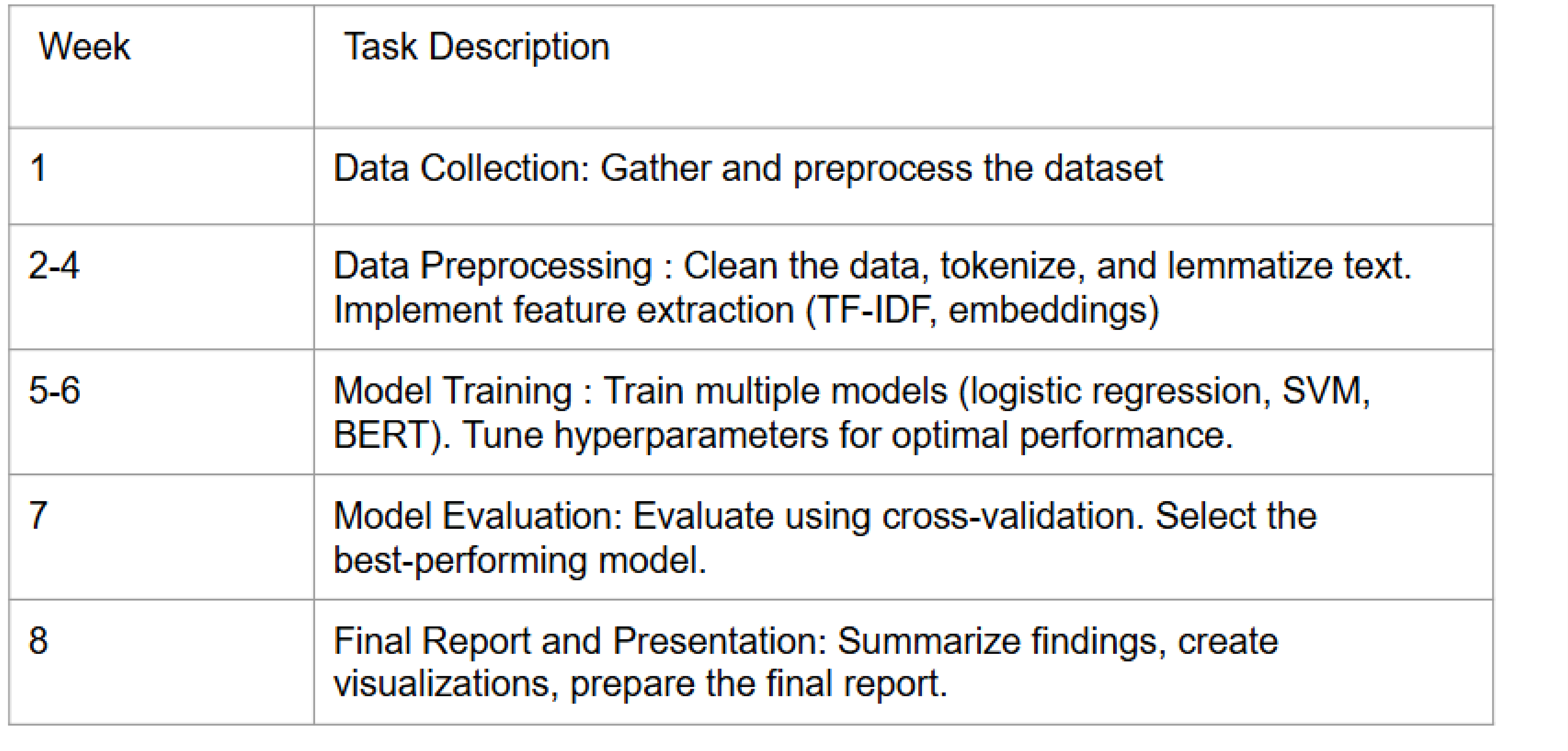
Libraries:

- spaCy & NLTK: For text preprocessing tasks such as tokenization and lemmatization.
- Scikit-learn: Used for classical machine learning models and evaluation metrics.
- TensorFlow/PyTorch: To implement deep learning models like LSTM and BERT.
- Hugging Face Transformers: For using pre-trained models (e.g., BERT) to generate sentence embeddings.

Data Sources:

- Quora Question Pairs Dataset: A dataset with over 400,000 pairs of labeled questions, used to train and evaluate the model.

TIMELINE



Week	Task Description
1	Data Collection: Gather and preprocess the dataset
2-4	Data Preprocessing : Clean the data, tokenize, and lemmatize text. Implement feature extraction (TF-IDF, embeddings)
5-6	Model Training : Train multiple models (logistic regression, SVM, BERT). Tune hyperparameters for optimal performance.
7	Model Evaluation: Evaluate using cross-validation. Select the best-performing model.
8	Final Report and Presentation: Summarize findings, create visualizations, prepare the final report.

EXPECTED OUTCOME

- Our model will accurately predict whether two questions are semantically identical.
- It will handle variations in sentence structures, synonyms, and paraphrasing.
- Q&A Platforms: Our output will automate the detection of duplicate questions, minimizing redundancy.
- Customer Support: Our system will help identify repetitive queries for better support management.

THANK YOU

Team Members

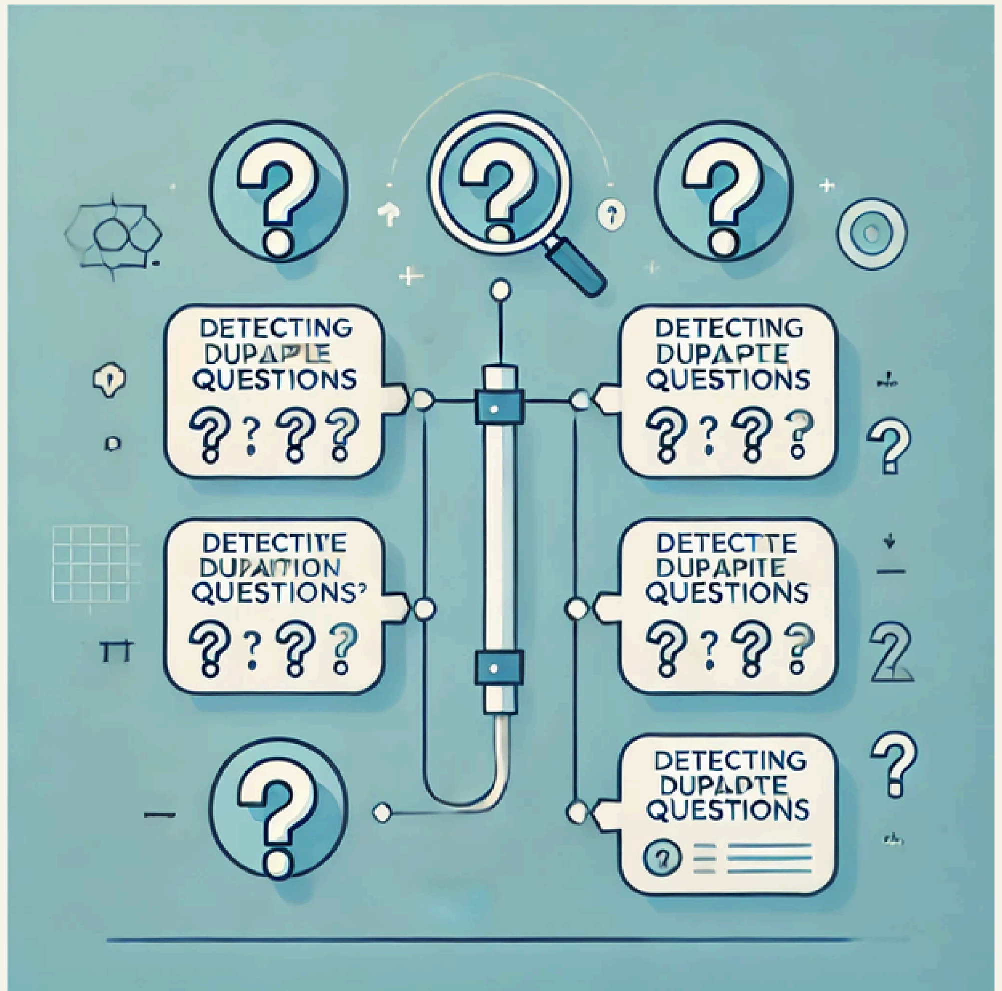
M.GunaShritha-SE22UARI102

M.HarshithaReddy-SE22UARI097

S.Mrudula-SE22UARI163

T.Harshitha-SE22UARI157

Sree Harshitha-SE22UARI204


DETECTING DUPAPLE QUESTIONS
DETECTING DUPAPTE QUESTIONS
DETECTITE DUPAITION QUESTIONS?
DETECTTE DUPAPITE QUESTIONS
DETECTING DUPAPTE QUESTIONS