Quora Question Pair Similarity Classification

By
Himanshu Patel
M.Tech CSE
IIT Guwahati

Quora is the place to gain and share knowledge about anything. It's a platform to ask questions and connect with people who contribute unique insights and quality answers. This empowers people to learn from each other and to better understand the world. Over 100 million people visit Quora every month, so it's no surprise that many people ask similarly worded questions. Multiple questions with the same intent can cause seekers to spend more time finding the best answer to their question and make writers feel they need to answer multiple versions of the same question. Quora values canonical questions because they provide a better experience to active seekers and writers, and offer more value to both of these groups in the long term.

Problem Statement

Identify which questions asked on Quora are duplicates of questions that have already been asked. This could be useful to instantly provide answers to questions that have already been answered. Our task is to predict whether a pair of questions are duplicates or not.

Real-world/Business objectives and constraints.

1.) The cost of a mis-classification can be **very high**.
2.) We would want a **probability** of a pair of questions to be duplicates so that we can choose any threshold of choice.
3.) **No strict latency** concerns.
4.) **Interpretability** is partially important.

Dataset

1.)Data will be in a file Train.csv
2.)Train.csv contains 5 columns : qid1, qid2, question1, question2, is_duplicate
3.)Size of Train.csv - 60MB
4.)Number of rows in Train.csv = **404,290**

Mapping the real-world problem to an ML Problem.

It is a **binary classification** problem, for a given pair of questions we need to predict if they are duplicates or not.

Performance Matrix used to quantify model performance

Metric(s):

- log-loss
- Binary Confusion Matrix

Train and Test construction

We build train and test by randomly splitting in the ratio of 70:30 or 80:20 whatever we choose as we have sufficient points to work with.

Final Results

Model	**Train Loss**	**Test Loss**
Logistic Regression	0.52	0.52
Linear SVM	0.53	0.53