Duplicate Question Pairs Detection using Machine Learning

Presented by:
Shruti Sureshan (M21CS015)

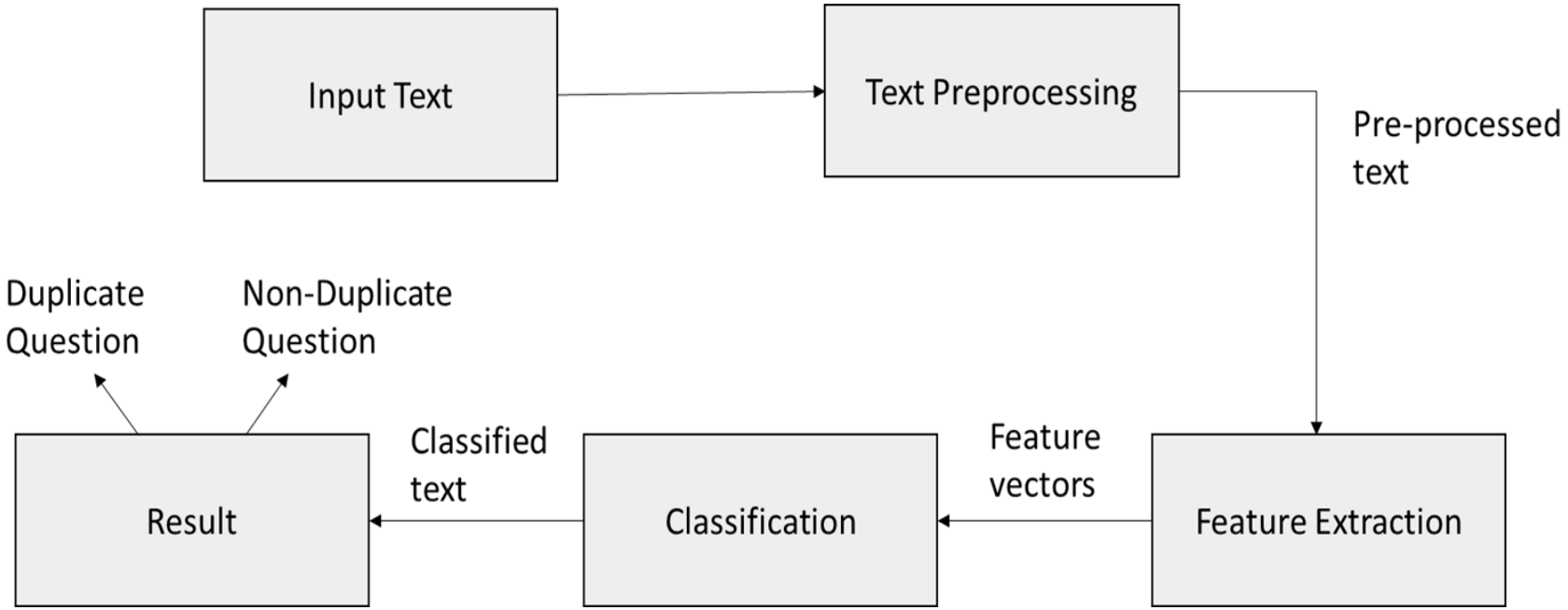
Introduction

Social Q&A forums like Quora is a growing platform comprising of a user-generated collection of questions and answers. It receives 300 million unique visitors every month, which raises the problem of different users asking similar questions with same intent but in different words. This can cause readers to spend more time to find the best answer, and make writers answer multiple versions of the same question. With the growing database, there is a need for Quora to preserve the trust of the users by maintaining quality content by discarding duplicate information.

Problem Statement

Identifying semantically similar questions on Quora is challenging as natural language is very expressive. In this project, our approach primarily focuses on Machine learning techniques for detecting whether a pair of questions asked on Quora is duplicate or not by taking the text input and preprocessing it, followed by feature extraction using Bag of Words and TF-IDF algorithm. Further, the questions are classified using machine learning classifiers like Logistic Regression, XGBoost and Support Vector Machines.

Workflow of the system



Text Preprocessing

- Removal of HTML Tags, punctuation marks, comma between numbers, removal extra whitespace, lowercase all texts, removal of special characters, expand contractions are performed first.
- This is followed by stop words removal. Stop words are a set of commonly used words in a language. They carry very little useful information.
- Finally stemming is performed which is the process of reducing a word to its word stem.

Feature Extraction

- Feature Extraction phase allows the data to be observed to extract the basic features from the data.
- The extracted features are length of question in question pairs, number of words present in question 1 and question 2, difference between the number of characters in question 1 and question 2, difference between the number of words in question 1 and question 2, number of common words, common words ratio.
- Also we have used different string comparison techniques from FuzzyWuzzy library to extract fuzzy features. Fuzzy String Matching is also known as approximate string matching. It is the process of finding strings that approximately match a given pattern.

Bag of Words Technique

- The machine learning models work with numerical data rather than textual data. Hence, by using the bag-of-words (BoW) technique, we convert a text into its equivalent vector of numbers.
- The “Bag of words” representation is a Natural Language Processing technique of text modelling. It is a common way to represent a text document.
- Here the words that are present in the document are assigned their frequency of occurrence. Vocabulary words which are not present in the document are assigned a frequency 0.
- This way a text document can be represented using a sparse and multi-dimensional feature vector which is then passed as an input to the different machine learning classifiers.

TF-IDF Algorithm

- TF-IDF(Term frequency–inverse document frequency) is the ratio of the number of times a word occurs in a particular question, to the number of times the word occurs in all the questions (our entire corpus).
- The TF–IDF value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general. A higher TF-IDF value indicates a more important word.
- The model learns the inverse frequency of words from the set of combined unique question1 and question2 character set. The corresponding TF-IDF, word level feature, obtained for each of the questions in the pair is then passed as an input to the different machine learning classifiers.

Logistic Regression Classifier

- Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more independent variables.
- The sigmoid function returns a probability value between 0 and 1. This probability value is then mapped to a discrete class which is either “0” or “1”.
- In order to map this probability value to a discrete class, we select a threshold value. This threshold value is called Decision boundary. Above this threshold value, we will map the probability values into class 1 and below which we will map values into class 0.
- Mathematically, it can be expressed as follows:-

  $p \geq 0.5 \Rightarrow class = 1$

  $p < 0.5 \Rightarrow class = 0$

XGBoost

- Xgboost is an ensemble technique for machine learning. Sometimes, it may not be sufficient to rely upon the results of just one machine learning model. Ensemble learning offers a systematic solution to combine the predictive power of multiple learners. Xgboost uses gradient boosting and it converts weak learners into strong learners.
- Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.
- XGBoost is an extension to gradient boosted decision trees and specially designed to improve speed and performance.

Support Vector Machines

- SVMs are one of the powerful machine learning algorithms for classification. An SVM classifier builds a model that assigns new data points to one of the given categories. Thus, it can be viewed as a non-probabilistic binary linear classifier.
- Our objective in SVM is to find a hyperplane that separates positive and negative examples with the largest margin while keeping the misclassification as low as possible.
- SVM searches for the maximum margin hyperplane in the following 2 step process –

1. Generate hyperplanes which segregates the classes in the best possible way. There are many hyperplanes that might classify the data. We should look for the best hyperplane that represents the largest separation, or margin, between the two classes.
2. So, we choose the hyperplane so that distance from it to the support vectors on each side is maximized. If such a hyperplane exists, it is known as the maximum margin hyperplane and the linear classifier it defines is known as a maximum margin classifier.

Results

Classifier	Feature Extraction	Accuracy
Logistic Regression	Bag of Words	69.02%
Logistic Regression	TF-IDF	67.36%
XGBoost	Bag of Words	75.38%
XGBoost	TF-IDF	75.31%
SVM	Bag of Words	77.52%
SVM	TF-IDF	75.29%