QUORA: IDENTIFYING QUESTIONS WITH THE SAME INTENT

CRTL-ALT-DESTRUCTION
MARIT BEEREPOOT, 10983430
JESSY BOSMAN, 11056045

THE CASE

Quora is a platform where questions can be asked and answered. Since Quora is a well known and widely used platform, a lot of questions are duplicates or have the same intent. Currently a Random Forest model is used to identify questions that are the same. The challenge of the kaggle competition was to use natural language processing to find duplicates.

Quora provided a labeled training set consisting of 317205 pairs of questions, that were labeled on whether the intent of the question was the same or not and a test set with 81126 pairs of questions. The goal was to get the highest accuracy on the test set.

FEATURES

The following features were selected to train the model on:

- Fuzzy Wuzzy simple ratio
- Fuzzy Wuzzy Partial ratio
- Fuzzy Wuzzy Sort ratio
- Fuzzy Wuzzy Set ratio
- Difflib Sequence Match ratio
- Jaccard distance
- The cosine similarity
- NLTK Countvectorizer features
- Average TF-IDF score of the sentence

EXPERIMENTS

The goal of the experiments is to find the highest accuracy. To find the highest accuracy, the right features, the right classifier and the right hyperparameters have to be found.

Experiment 1: Which features?
Different subsets of features were created to test which features should be trained on. The training data was splitted into a training set and a test set and k-fold cross validation (5 folds) was used. After splitting the data all combinations of features were tested and performance was measured using the accuracy, recall and a confusion matrix. The best scoring feature set can be found above.

Experiment 2: Which classifier?
There are a lot of different classifications methods. The following methods were compared:

- Neural network using Scikit learn
- Neural network using Keras
- Logistic regression using Scikit learn
- Gaussian Naive Bayes Classifier using Scikit learn
- Support Vector Classifier using Scikit learn
- K Neighbors Classifier using Scikit learn
- Decision Tree Classifier using Scikit learn
- Random Forest Classifier using Scikit learn

These methods were compared using k-fold cross validation, by comparing the accuracy, the recall and the confusion matrix.

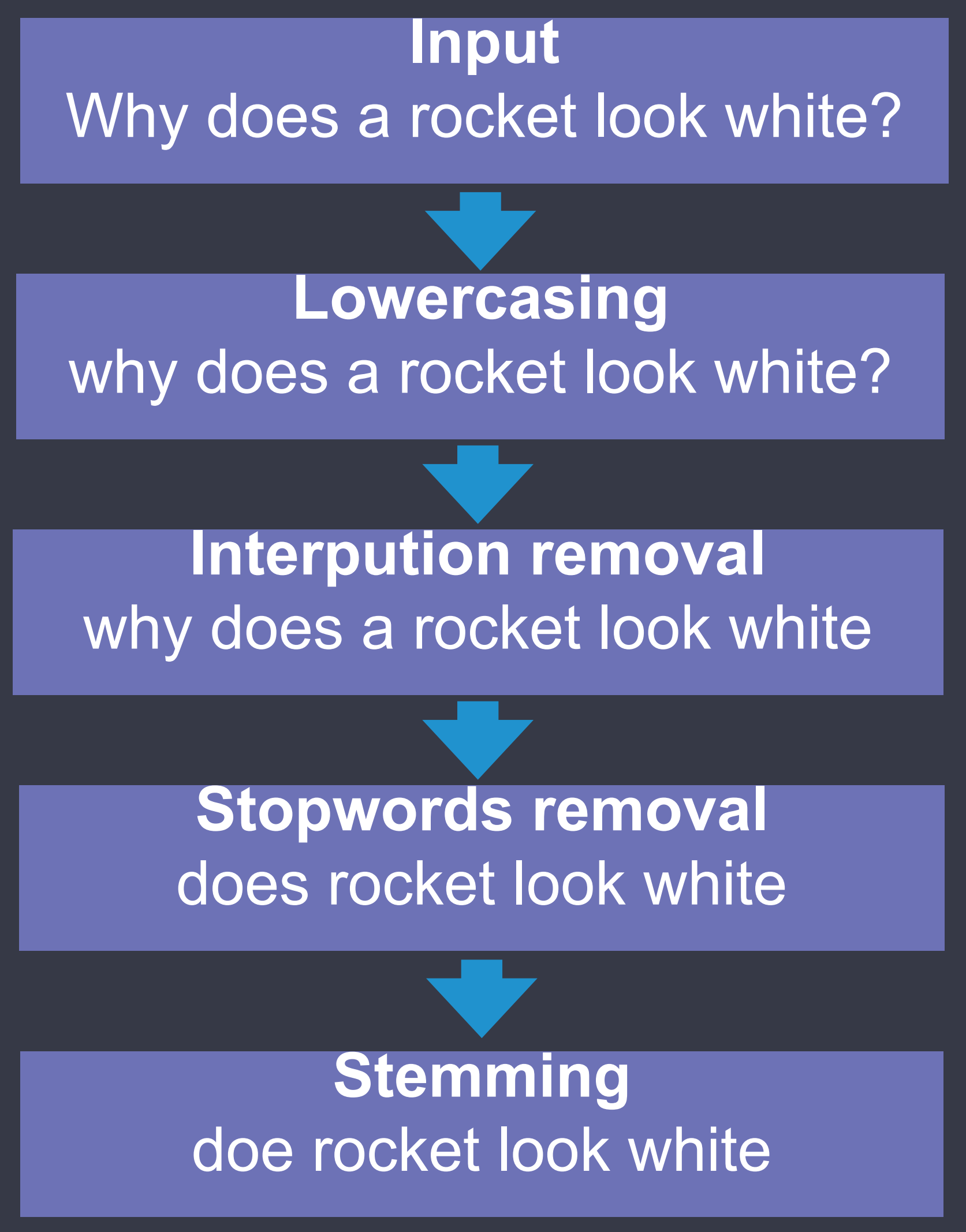
Experiment 3: Which parameters?
After the classifier was found, the optimal parameters of the classifier had to be found. There were primarily two outstanding parameters that had to be tuned:

- The number of features of the countvectorizer
- The number of estimators of the random forest classifier

Other parameters didn't seem to have an influence on the outcome.

PREPROCESSING

The data was loaded into a Pandas dataframe. After that the text data was processed by lowercasing the questions, the removal of interpunction, the removal of stopwords and by stemming the words.



Input
Why does a rocket look white?

Lowercasing
why does a rocket look white?

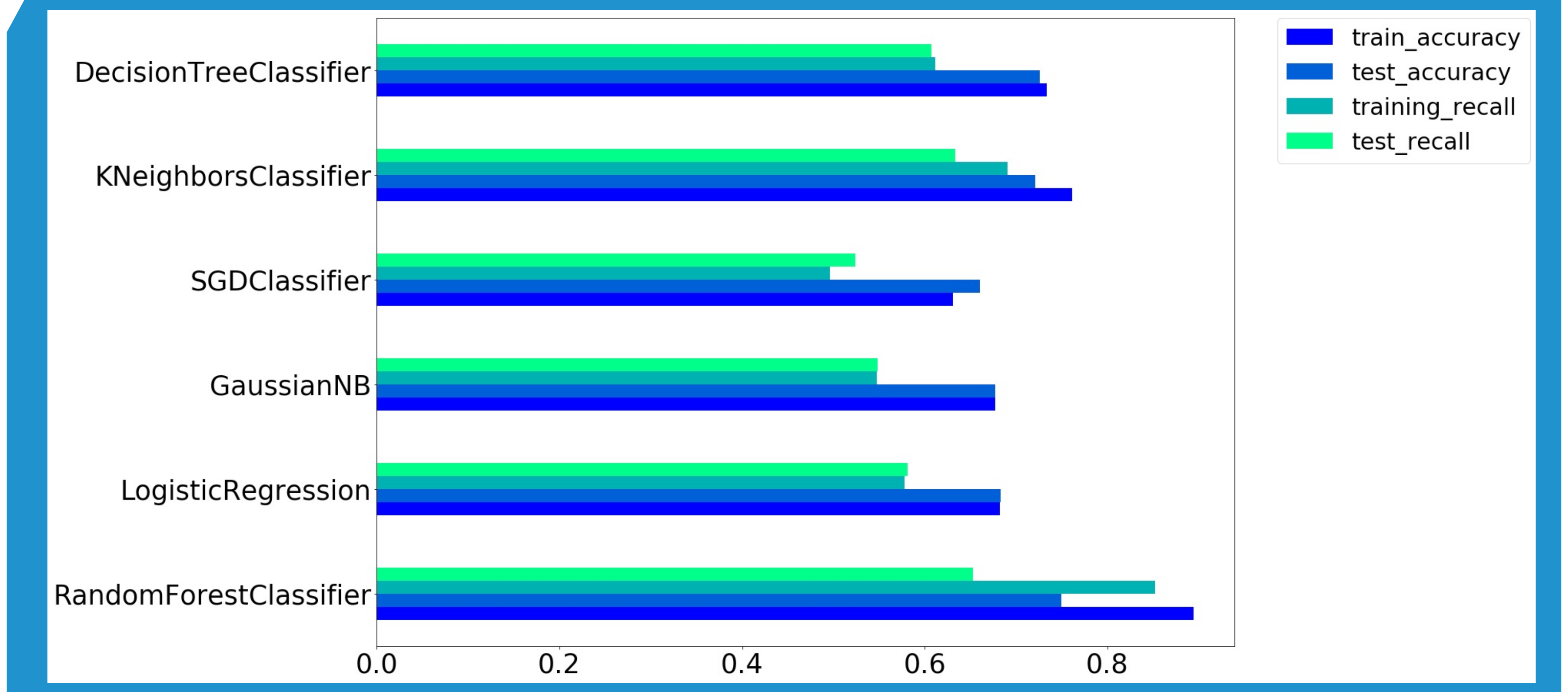
Interpution removal
why does a rocket look white

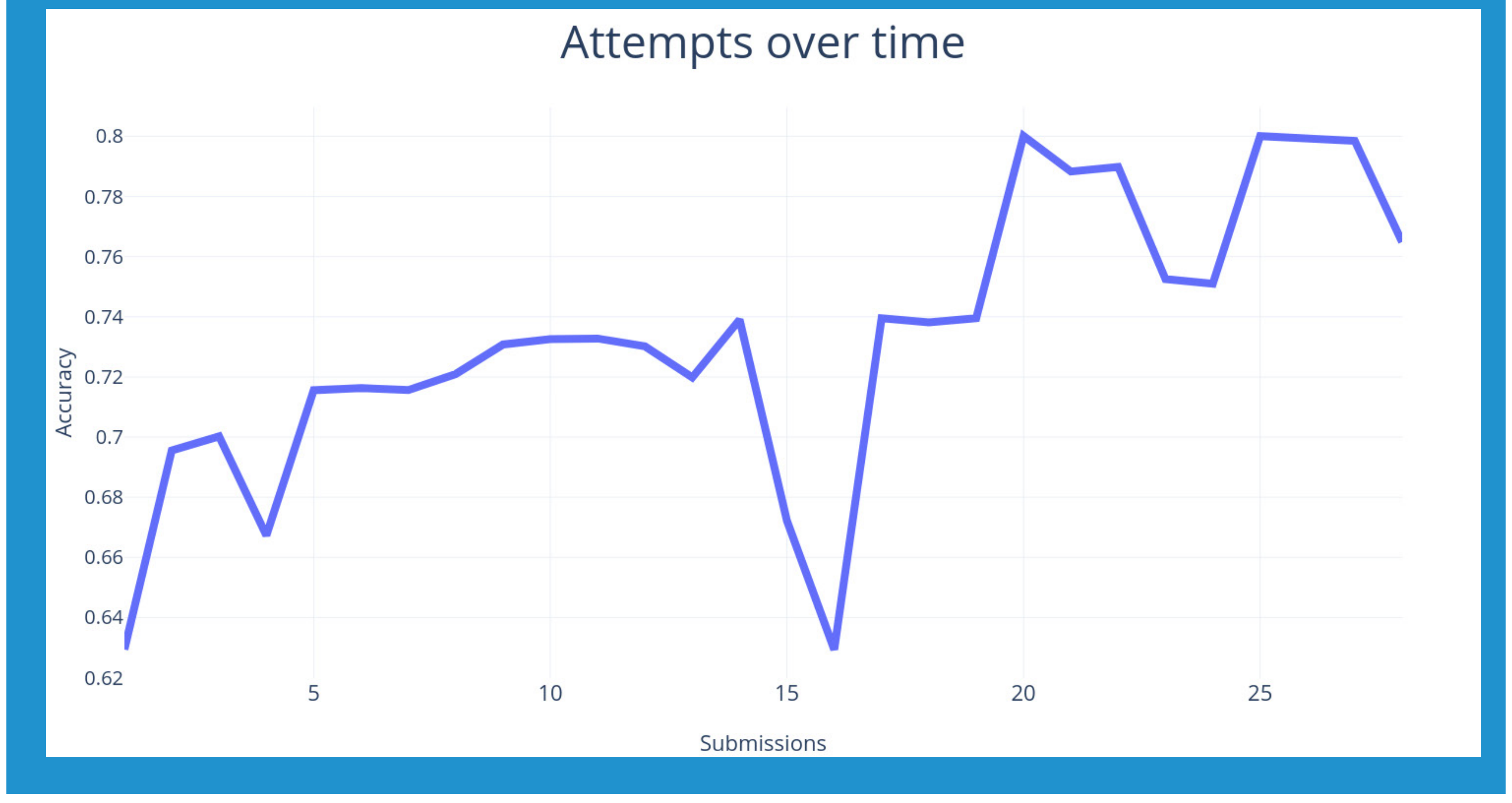
Stopwords removal
does rocket look white

Stemming
doe rocket look white

ANALYSIS

We found that the Random Forest classifier did the best job correctly labeling the test set. This is actually the same classifier that Quora currently uses.
An 80% accuracy was found while using the features on the left, 500 parameters in the countvectorizer and 50 estimators in the Random Forest Classifier. This setting also had the highest recall. Next to a high recall and high accuracy, this was also one of the fastest models from the ones we compared. This can be useful when Quora wants to compare a newly asked question for duplicates as soon as it's uploaded.




Attempts over time