Identification of Quora question pairs with same intent

Group 9

Goutham Reddy
Nitin Garg
Sneh Lata
Sourav Bose

Quora Question Pairs (QQP)

Objective

Build a question similarity classifier to detect duplicate questions

Why is this problem important?

- Growth in community based question answering such as Quora, Reddit etc. As number of questions grow, so do the duplicates
- Treating duplicate questions independently will restrict presenting high quality responses to the users
- Reduces answering burden for the responders, thereby improving overall user experience
- Improving user experience

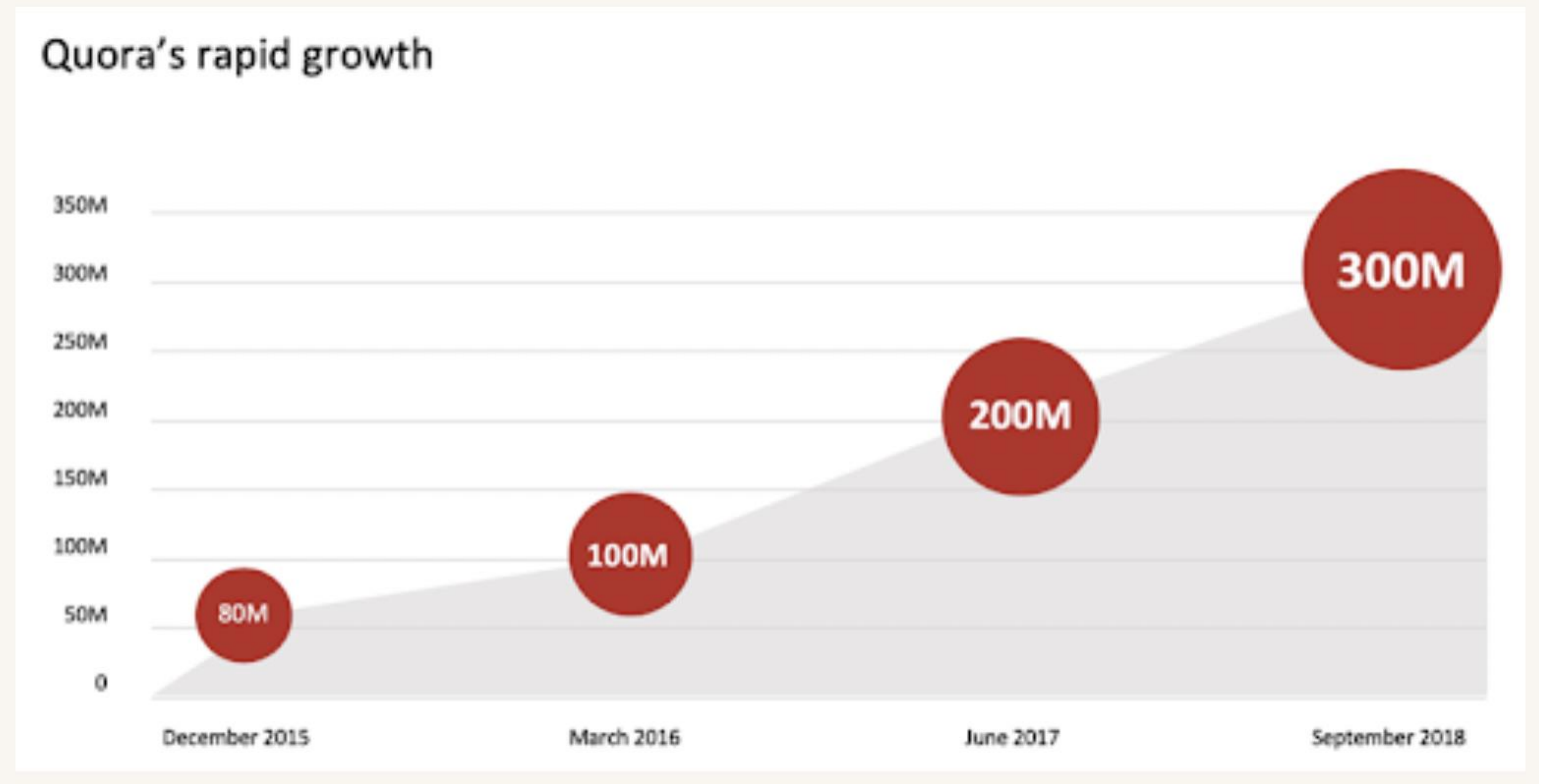
Extension of these models @ work

- IT companies have a huge compendium of problems and their solutions. Many of these problems are duplicate. Application of these models will help in presenting coherent set of solutions to the users, thereby helping in cross deployment of knowledge

About Quora

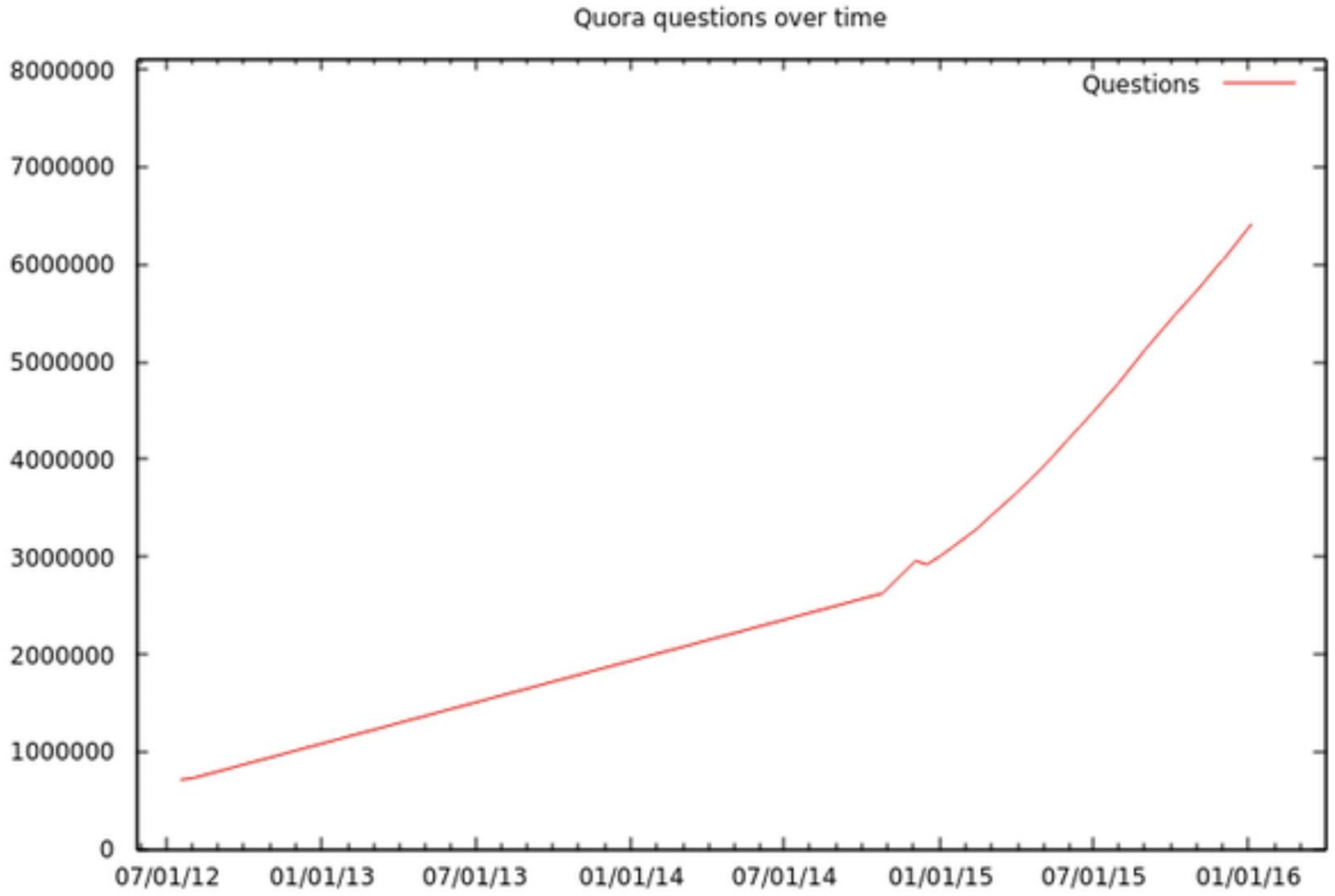
Quora Statistics

- Quora has 300+ million monthly active users
- The majority of Quora users are from 18 to 24
- The average time spent by a user on the platform is around 8-10 minutes
- 3000 to 5000 Questions are asked on the platform every day



Question Statistics

- As of 2021, Quora gets an average of 3,000 – 5,000 questions daily
- 99% of questions asked on Quora get answered
- 20% of questions get not less than 3 answers
- 90% of questions on Quora have less than ten answers
- Over 30% of Quora users haven't responded to any question



Benchmarks & Approach

Quora question pair problem has been extensively researched and is part of the GLUE benchmarks under similarity and paraphrase type of NLP tasks

Benchmarks

- **GLUE Benchmark**[1]: Accuracy → 86.5; F1Score → 66.1 with BiLSTM Model + Attn + ELMO
- **Current Benchmark**: Accuracy → 92.3% with XLNet

Approach:

❖ **Classical ML models** such as Logistic regression, Support Vector Machines, Random Forest, XG Boost, LGBM, Voting Classifier and iteratively improving accuracy scores through feature engineering and data cleaning methods

❖ Applying **deep neural networks** such as C-BOW, sentence BERT

❖ Deep dive on edge cases, identify areas for further improving the accuracy of the models

1. pdf (openreview.net)
2. ***Quora Question Pairs Benchmark (Question Answering) | Papers With Code***

Feature Engineering

Base Features

- Length of question 1
- Length of question 2
- Same question type (what, why, when, where, how)
- % Common words
- POS features:
  - 'q1_verb_count'
  - 'q2_verb_count'
  - 'q1_noun_count'
  - 'q2_noun_count'
  - 'q1_adj_count'
  - 'q2_adj_count'
  - 'q1_adv_count'
  - 'q2_adv_count'

Derived Features

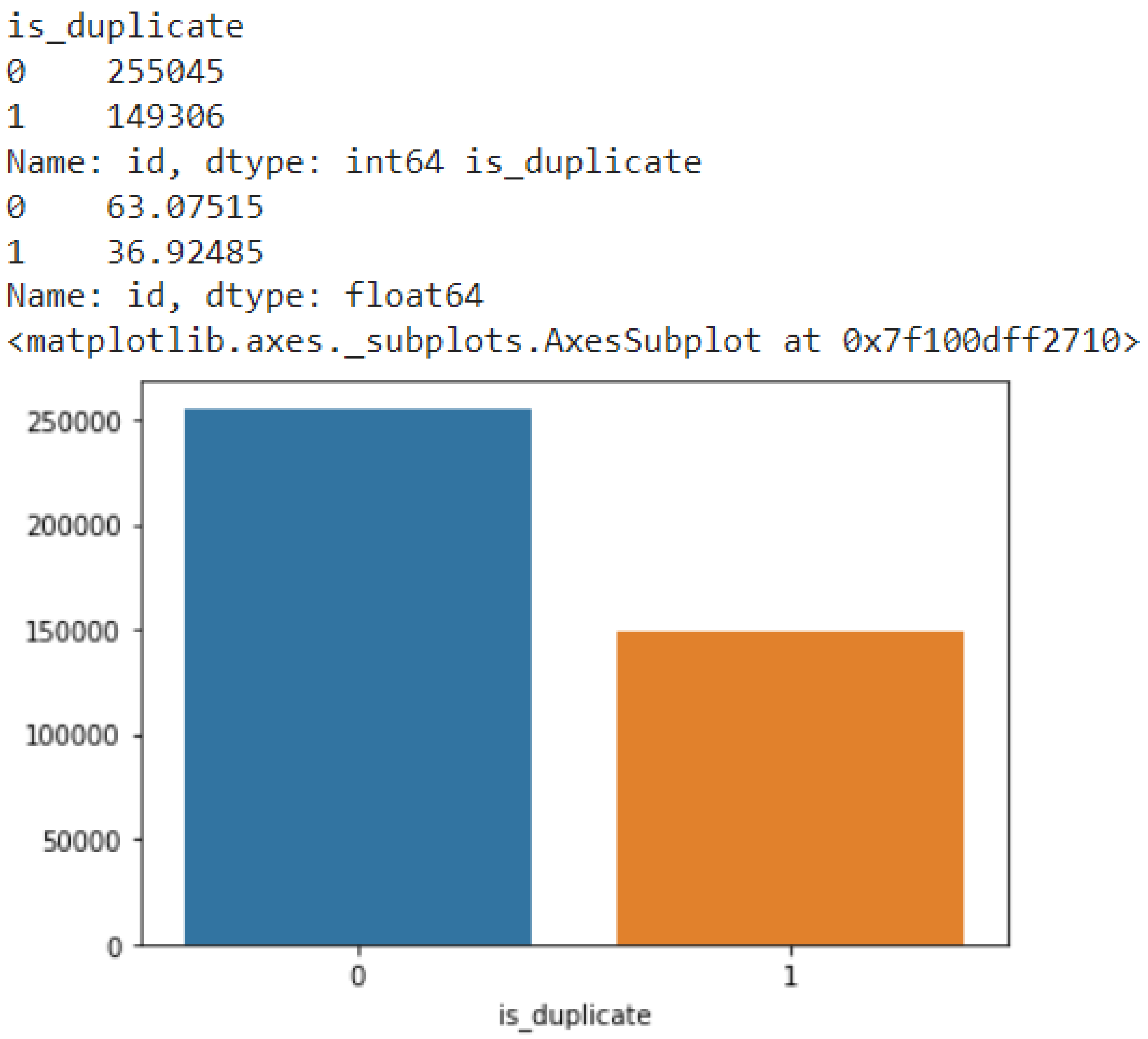
- Fuzz Ratio
- Fuzz Partial Ratio
- Token Set Ratio
- Average Fuzz Score
- Ratio of common word count with word count (min/ max)
- Ratio of common non-stop words to word count (min/ max)
- Absolute length difference of question pair
- Average lengths of the question pair

Embeddings & Distances

- TFIDF based Weighting
- Mean weights with GLOVE 100 Embeddings
- Distances between two vectors
- Cityblock distance
- Canberra distance
- Cosine distance
- Euclidean distance
- Sequence Ratio

Exploratory Data Analysis

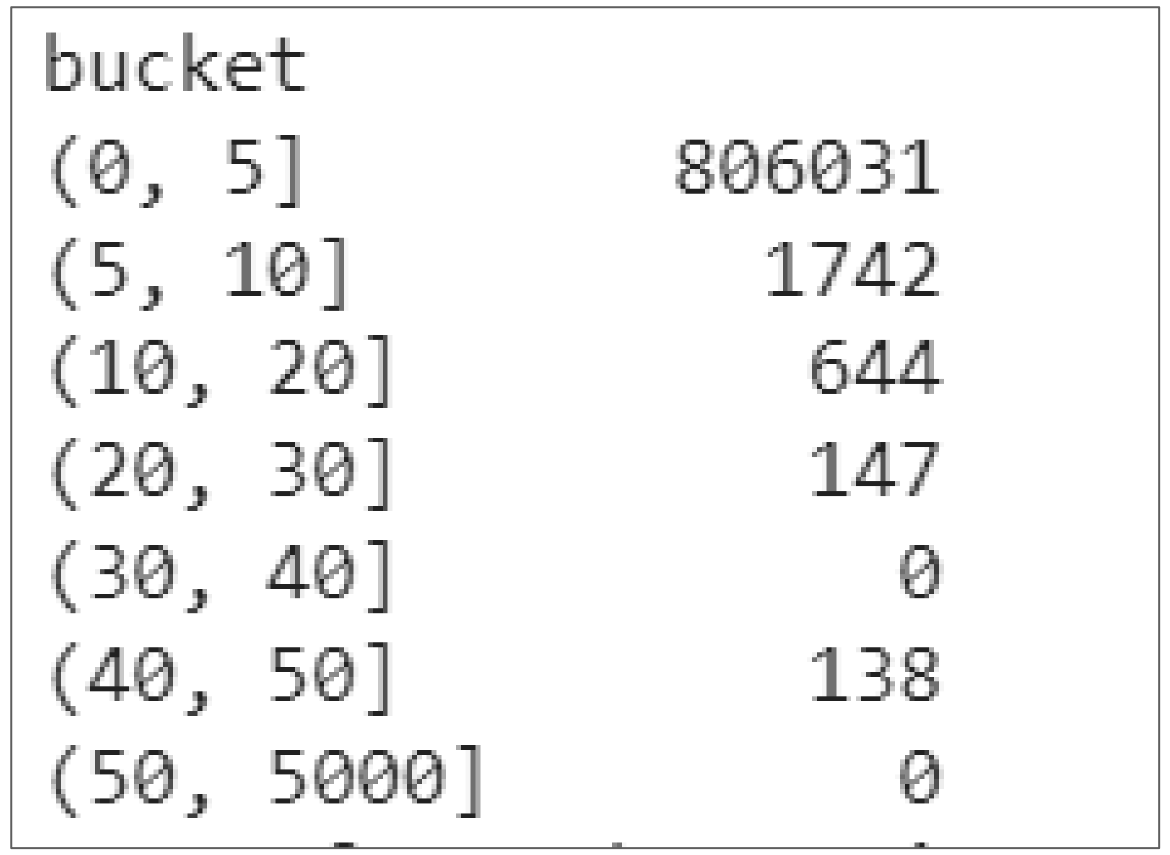
Data Distribution



```
is_duplicate
0    255045
1    149306
Name: id, dtype: int64 is_duplicate
0    63.07515
1    36.92485
Name: id, dtype: float64
<matplotlib.axes._subplots.AxesSubplot at 0x7f100dff2710>
```

- 404351 training data points. No. of non-duplicate data points is 255045 and duplicate data points is 149306
- 36.92% of question pairs are duplicates and 63.08% of questions pair non-duplicate

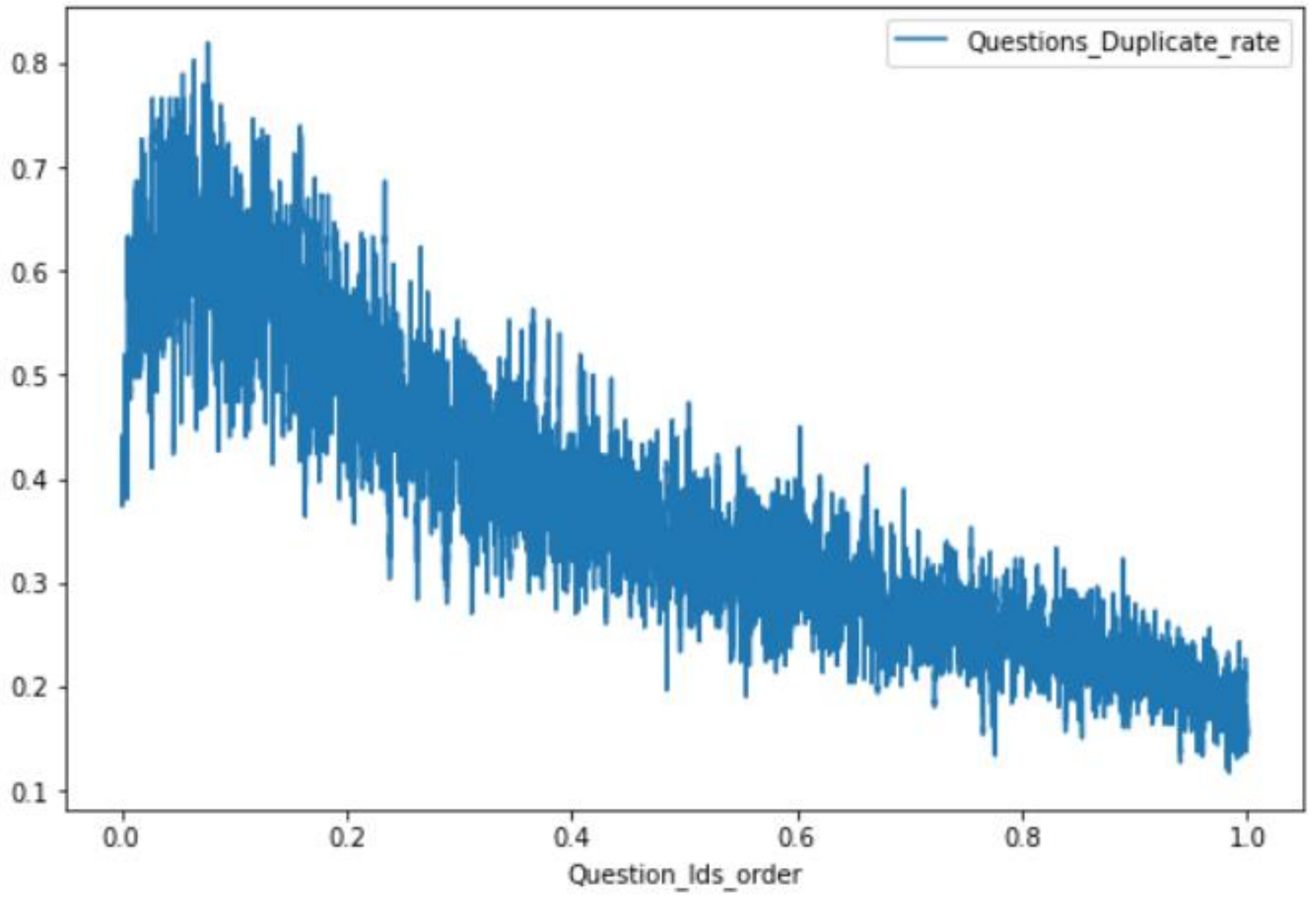
Repetition of Questions



```
bucket
(0, 5]         806031
(5, 10]          1742
(10, 20]          644
(20, 30]          147
(30, 40]            0
(40, 50]          138
(50, 5000]          0
```

- Repetition of questions: 99.7% of the questions are repeated less than 5 times
- 138 questions are repeated more than 40 times

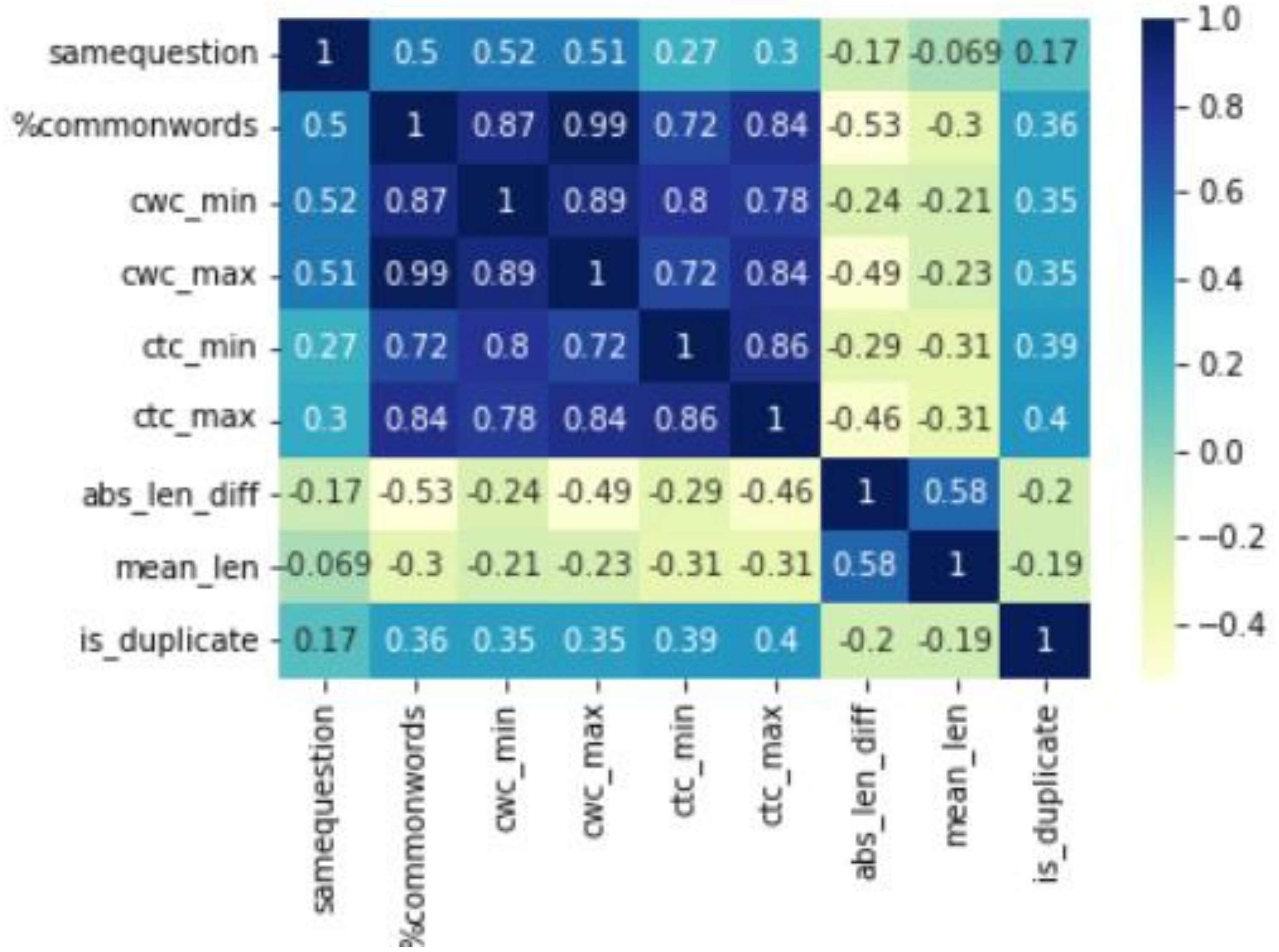
Temporal Nature



- Decreasing trend of Quora Question duplicates as the Question Id increases
- Train data is shuffled to overcome above issue and for better model learning

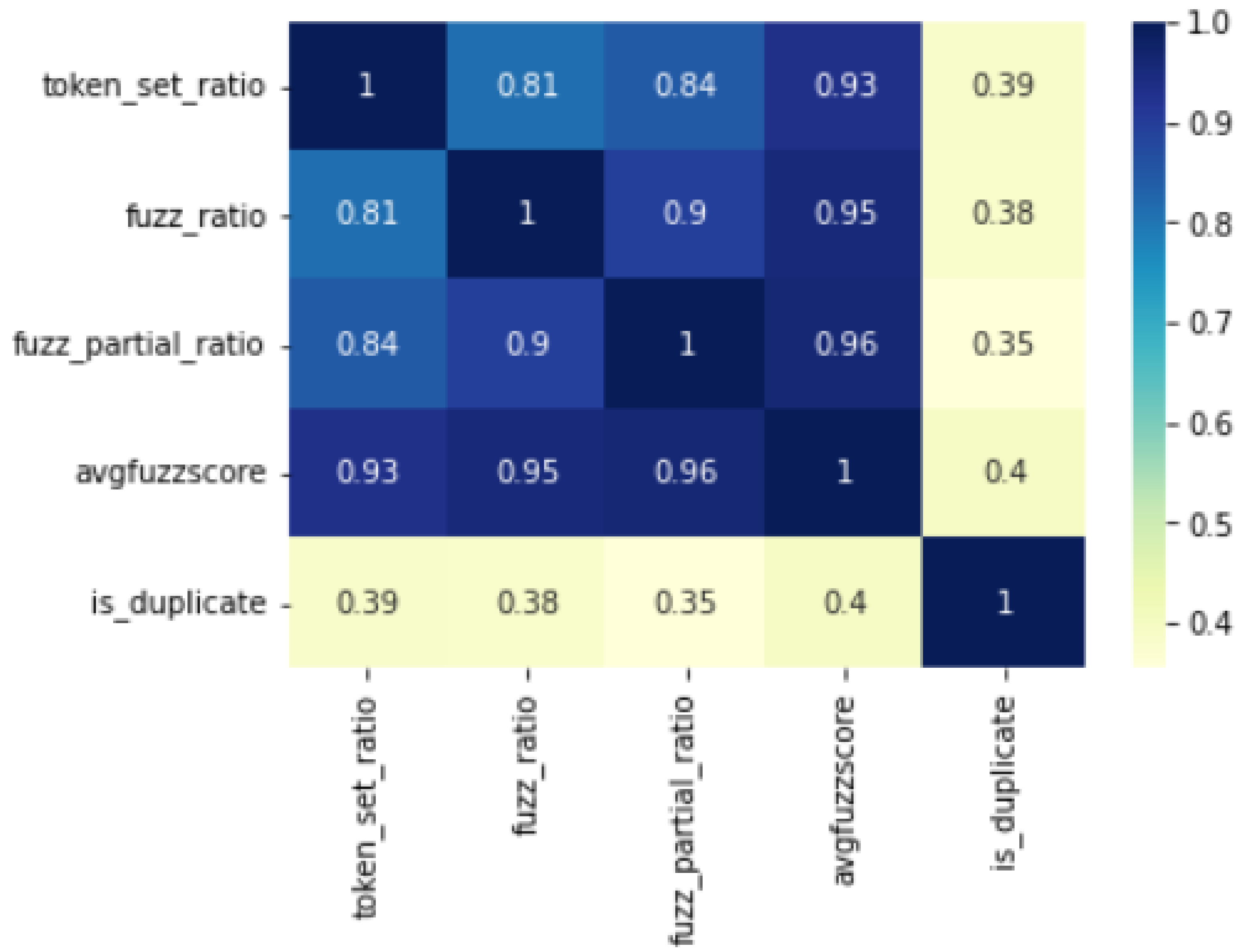
Exploratory Data Analysis

Base Features



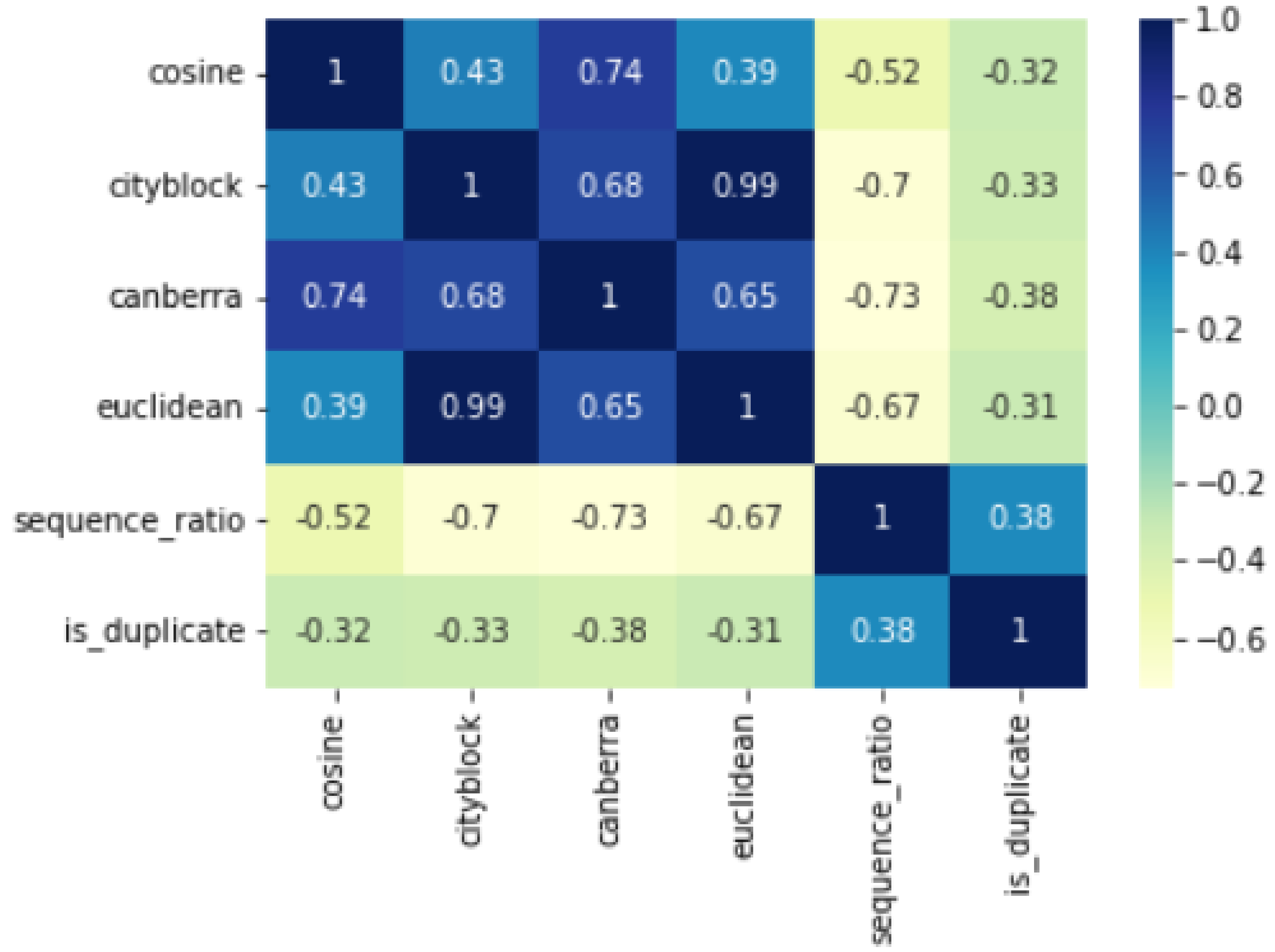
- Non-Stop word ratio (CTC metrics), common word count ratios and % commonwords have high correlations with duplication of question

Derived Features



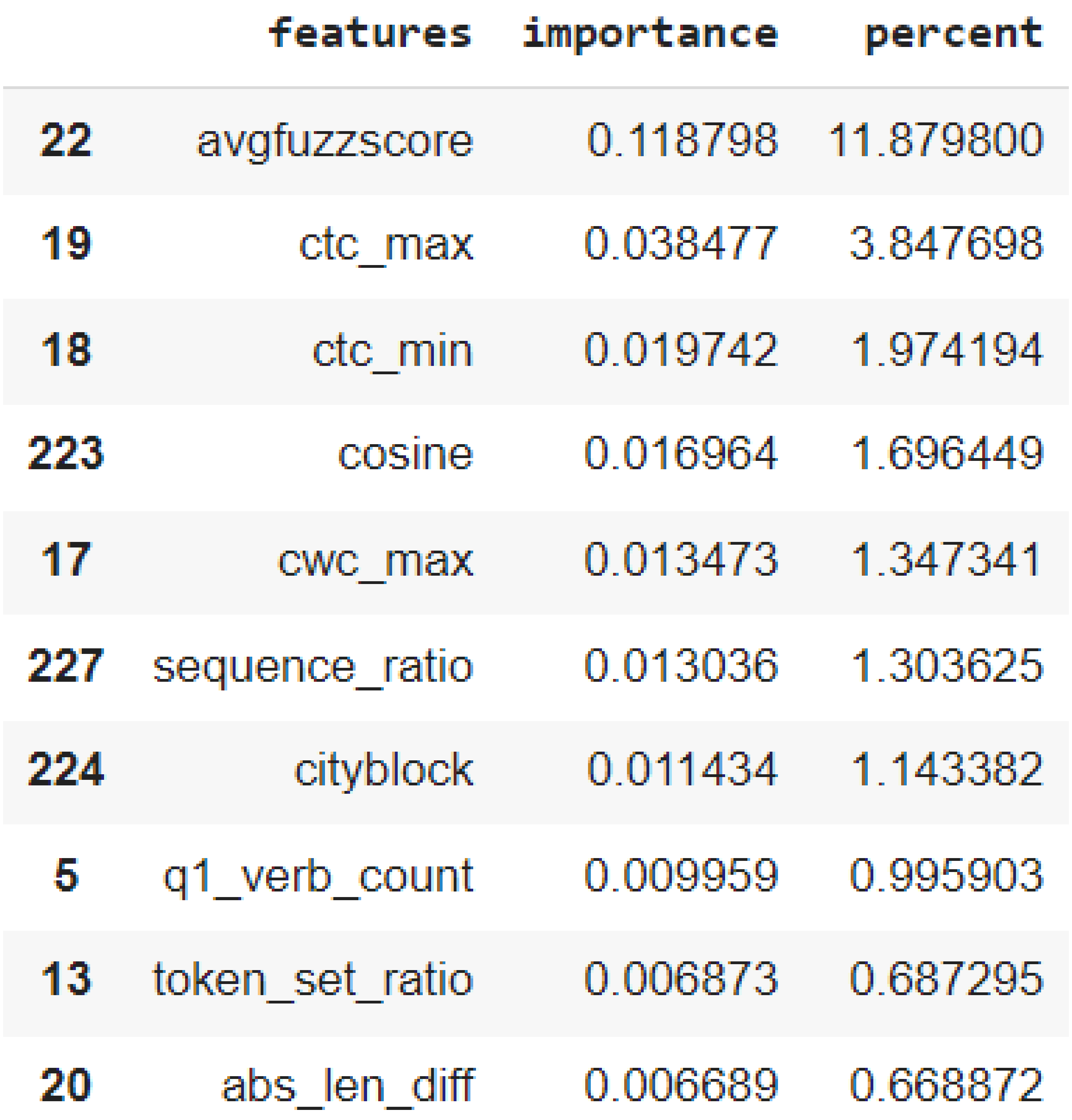
- Correlation of 0.35 to 0.4 observed between fuzzy features and the outcome metric

Embeddings & Distances

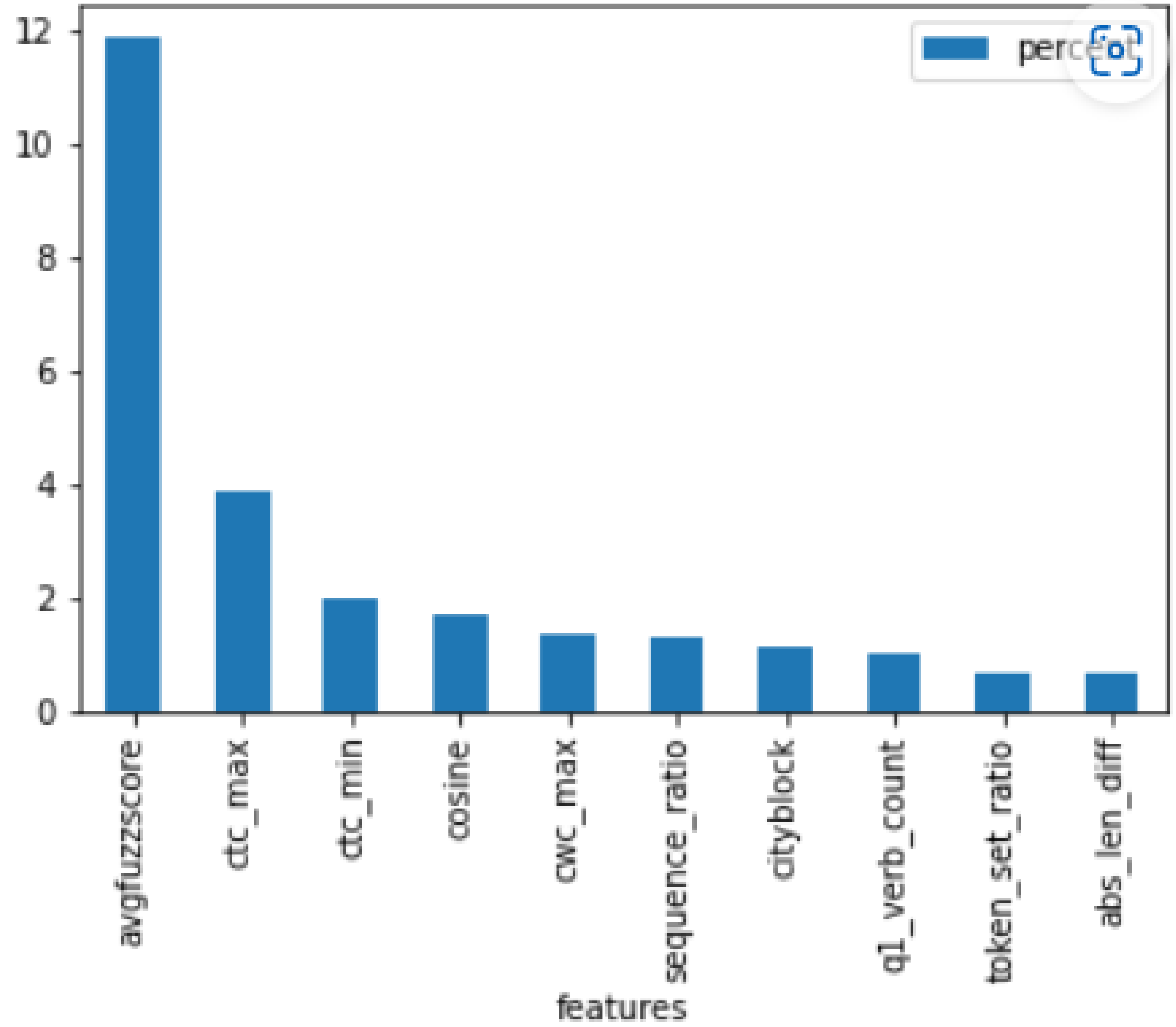


- Negative correlation between distances and duplication of questions indicating smaller distance with higher probability of duplication
- Sequence ratio and Canberra distances have highest correlation among these features

Feature Importance



	features	importance	percent
22	avgfuzzscore	0.118798	11.879800
19	ctc_max	0.038477	3.847698
18	ctc_min	0.019742	1.974194
223	cosine	0.016964	1.696449
17	cwc_max	0.013473	1.347341
227	sequence_ratio	0.013036	1.303625
224	cityblock	0.011434	1.143382
5	q1_verb_count	0.009959	0.995903
13	token_set_ratio	0.006873	0.687295
20	abs_len_diff	0.006689	0.668872



- Most impacting features are: Average Fuzz Ratio, Ratio of Non stop words with total words, Cosine and Cityblock distances, and Sequence Ratio

Edge Cases

Gage R&R/ Misclassifications

- Observed question pairs that were identical in text, but were incorrectly classified as Not Duplicate
  - What are the differences between Github and Bitbucket?
  - What are the differences between BitBucket and GitHub?

- Questions were marked as duplicate, but intent of the question was otherwise
  - What is the exercise to remove belly fat for girls?
  - How do I remove belly fat?

Changes in words but the intent remains the same

We observed that there were question pairs, wherein the second question was changed slightly by using synonyms or similar words. Its observed that LSTM based networks performed better in identifying duplication in such scenarios

- What is the best gift for my boyfriend on Valentine's Day?
- What are the Best Gifts for men on Valentine's day?

Edge Cases

Differing length of Questions

- One question is small (in length) while second part is longer. However more text in one of the question **changes the intent of the question** hence they are not duplicates.
- It is observed that much of these questions were **identified correctly by LSTM and SBERT model** than with XG boost model
    - What is the specific heat capacity of steam?
    - What is specific heat capacity?

- One question is small (in length) while second part is longer. However more text in one of the question **does not change the intent of the question** hence they are duplicates.
- We observed **randomness in correctly identification** of such question pairs by both the models.
    - What do you want to be remembered for?
    - What do you want to be remembered for when you die?

Impact of Numbers

Question pairs where numbers were involved and change intent of the question. We observe randomness in prediction by both the models

- Is it safe to use castor oil at 36 weeks to induce labor?
- Is it safe to use castor oil at 39 weeks to induce labor?

Results

Sr	Model Category	Model	Accuracy	F1-Score
1	**Classical ML Models**	Logistic Regression	70.5	66.7
2		Linear SVM	71.3	66.7
3		Random Forest	TBD	TBD
4		XG Boost	83.7	76.5
5		LGBM	82.4	76
6		Voting Classifier	TBD	TBD
7	**Neural Networks**	C-BOW – Word2Vec	81.47	74.5
8		C-BOW – TFIDF-IDF weighted Word2Vec	81	74.7
9		LSTM – Word2Vec	84.4	78.9
10		LSTM – TFIDF-IDF weighted Word2Vec	84	78.8
11		**Sentence BERT**	**85.8**	**81**

Conclusion

- Best results were achieved through Sentence BERT modelling with 85.8% accuracy and 81% F1 Score.
- Observed few question pairs that were identical in text, but were incorrectly classified as Not Duplicate
- Observed few questions were marked as duplicate, but intent of the question was otherwise

Future work :
Stacking model could achieve further accuracy. We have created a stacking models but evaluation is in progress.