

Quora Question Pair Similarity

Divya Manoj(IN40057)

Harsha Vardhan(IN40062)

Suvin Koshy George(IN40193)

D.V.S.Rajkiran(IN40136)

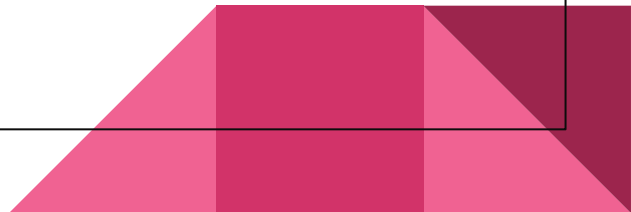
Under the Guidance of
Koorimi Kiran Kumar Sir.

INTRODUCTION

- World's biggest forum.
- Best place to share general knowledge.
- Topics are designed to only ask questions .

Problem

- People may ask similar questions
- Important interest to detect duplicated questions
- **Prediction problem : from a question pair, predict whether questions are the same or not.**

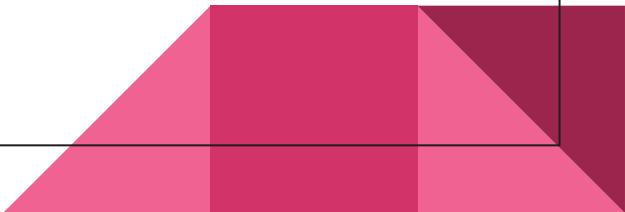


How Dataset Looks?

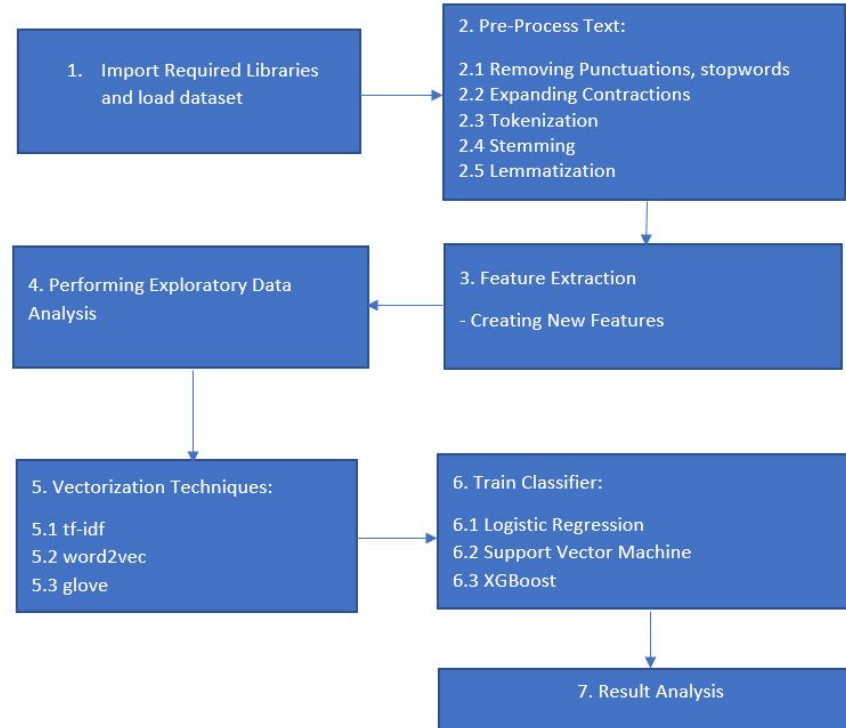
id	qid1	qid2	question1	question2	is_duplicate
0	0	1	2	What is the step by step guide to invest in sh... What is the step by step guide to invest in sh...	0
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia... What would happen if the Indian government sto...	0
2	2	5	6	How can I increase the speed of my internet co... How can Internet speed be increased by hacking...	0
3	3	7	8	Why am I mentally very lonely? How can I solve... Find the remainder when 23^{24} i...	0
4	4	9	10	Which one dissolve in water quikly sugar, salt... Which fish would survive in salt water?	0

DESCRIPTION OF THE DATASET

The dataset consists of 6 columns :

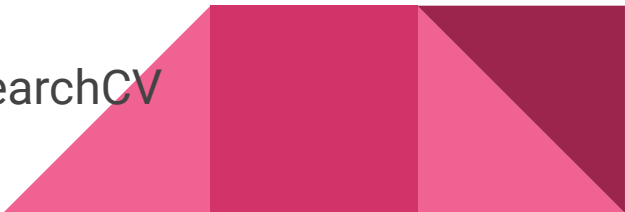
- id : Simple row ID.
 - qid1 & qid2 : Unique ID of each question in the pair.
 - question 1 & question 2 : The actual questions in the format of text.
 - is_duplicate : Target variable which we are trying to predict whether the two questions duplicate each other.
 - We have 404290 data points in which is_duplicate is the dependent variable with class labels 0 & 1.
 - It is a binary classification problem.
 - 3 Nan Values with in the data.
- 

Flow Diagram



Proposed Methodology

I. Importing Libraries and Loading Dataset

- a) Libraries for Mathematical Computation and Array :
 - Numpy, Pandas
 - b) Libraries for Visualization
 - Matplotlib, Seaborn, Plotly, WordCloud
 - c) Libraries for Pre-processing
 - Nltk, stopwords, WordNetLemmatizer, SnowballStemmer
 - d) Libraries for Vectorization Techniques
 - Sklearn, TfidfVectorizer, CountVectorizer
 - e) Libraries for Machine Learning Model
 - Sklearn, SGDClassifier, CalibratedClassifierCV, GridSearchCV
- 

II. PreProcessing Text

- Perform following Text Preprocessing to obtain cleaner texts:
 - Tokenization - Convert sentences to words.
 - Removing unnecessary punctuation, HTML tags.
 - Removing stop words - Frequent words such as "the", "is", etc. that do not have specific semantic.
 - POS Tagging - Categorizing words in a text (corpus) in correspondence with a particular part of speech, depending on the definition of the word and its context.
 - Lemmatization - Words are reduced to a root by removing inflection through dropping unnecessary characters, usually a suffix.

Before and After PreProcessing

What is the step by step guide to invest in share market in india?

What is the step by step guide to invest in share market?

step step guid invest share market india

step step guid invest share market

What is the story of Kohinoor (Koh-i-Noor) Diamond?

What would happen if the Indian government stole the Kohinoor (Koh-i-Noor) diamond back?

stori kohinoor kohinoor diamond

would happen indian govern stole kohinoor kohinoor diamond back

How can I increase the speed of my internet connection while using a VPN?

How can Internet speed be increased by hacking through DNS?

increas speed internet connect use vpn

internet speed increas hack dns

Why am I mentally very lonely? How can I solve it?

Find the remainder when 23^{24} is divided by 24,23?

mental lone solv

find remaind 23^{24} divid 2423

Which one dissolve in water quickly sugar, salt, methane and carbon di oxide?

Which fish would survive in salt water?

one dissolv water quik sugar salt methan carbon di oxid

fish would surviv salt water

Astrology: I am a Capricorn Sun Cap moon and cap rising...what does that say about me?

I'm a triple Capricorn (Sun, Moon and ascendant in Capricorn) What does this say about me?

astrolog capricorn sun cap moon cap risingwhat say

tripl capricorn sun moon ascend capricorn say

Should I buy tiago?

What keeps children active and far from phone and video games?

buy tiago

keep children activ far phone video game

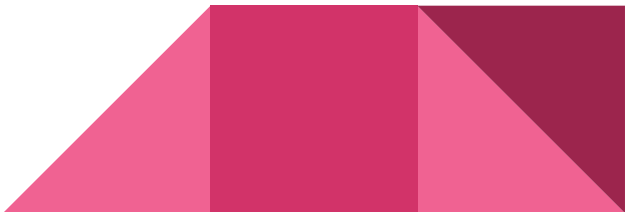
How can I be a good geologist?


What should I do to be a great geologist?

good geologist

great geologist

III. Feature Extraction

- ❑ **freq_qid1**: Frequency of qid1's
 - ❑ **freq_qid2**: Frequency of qid2's
 - ❑ **q1_char_len**: Characters count in question1
 - ❑ **q2_char_len**: Characters count in question2
 - ❑ **q1_word_len**: Words count in question1
 - ❑ **q2_word_len**: Words count in question2
 - ❑ **total_word_len**: Total words count in question 1 &2
 - ❑ **words_common**: Number of common unique words in Question 1 and Question 2
 - ❑ **word_share**: $(\text{word_common})/(\text{total_word_len})$
 - ❑ **freq_q1+freq_q2**: sum of the frequency of qid1 and qid2
 - ❑ **cwc_min**: This is the ratio of the number of common words to the length of the smaller question
 - ❑ **cwc_max**: This is the ratio of the number of common words to the length of the larger question
 - ❑ **csc_min**: This is the ratio of the number of common stop words to the smaller stop word count among the two questions
- 

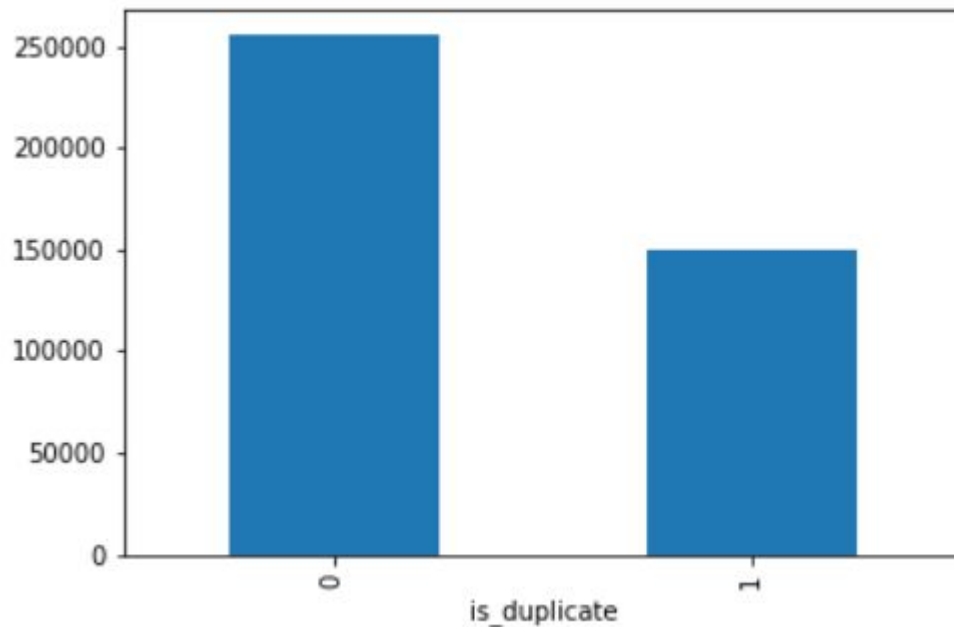
- ❑ **csc_max**: This is the ratio of the number of common stop words to the larger stop word count among the two questions
 - ❑ **ctc_min**: This is the ratio of the number of common tokens to the smaller token count among the two questions
 - ❑ **ctc_max**: This is the ratio of the number of common tokens to the larger token count among the two questions
 - ❑ **last_word_eq**: 1 if the last word in the two questions is same, 0 otherwise
 - ❑ **first_word_eq**: 1 if the first word in the two questions is same, 0 otherwise
 - ❑ **abs_len_diff**: Absolute difference between the length of the two questions (number of words)
 - ❑ **mean_len**: Mean of the length of the two questions (number of words)
 - ❑ **token_set_ratio**: token_set_ratio from fuzzywuzzy
 - ❑ **token_sort_ratio**: token_sort_ratio from fuzzywuzzy
 - ❑ **fuzz_ratio**: fuzz_ratio score from fuzzywuzzy
 - ❑ **fuzz_partial_ratio**: fuzz_partial_ratio from fuzzywuzzy
 - ❑ **longest_substr_ratio**: Ratio of the length of the longest substring among the two questions to the length of the smaller question
- 



Exploratory Data Analysis

Distribution of the target variable within the dataset

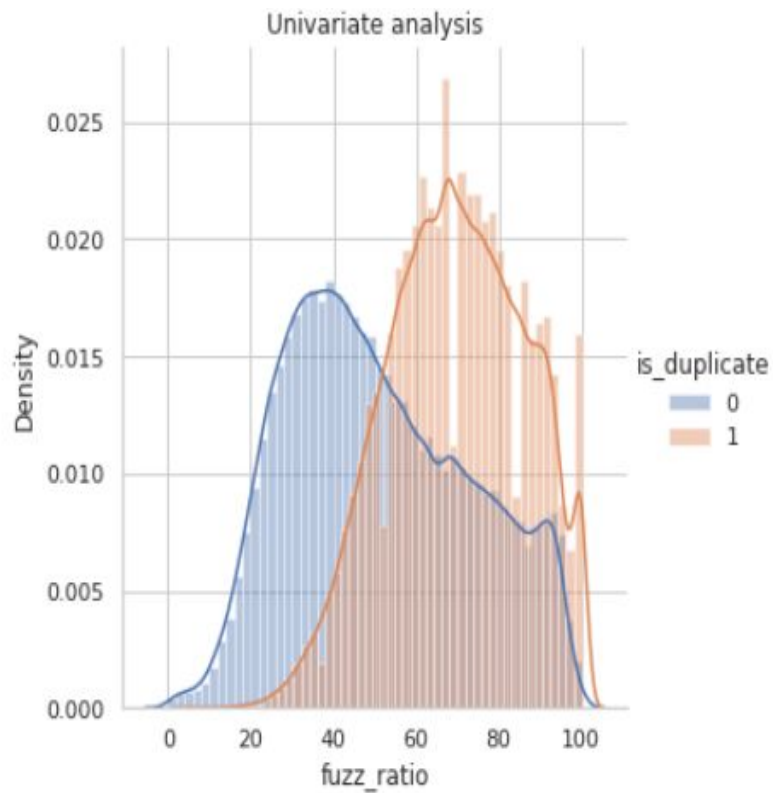
There are around 250000 non-similar questions and 150000 similar questions.



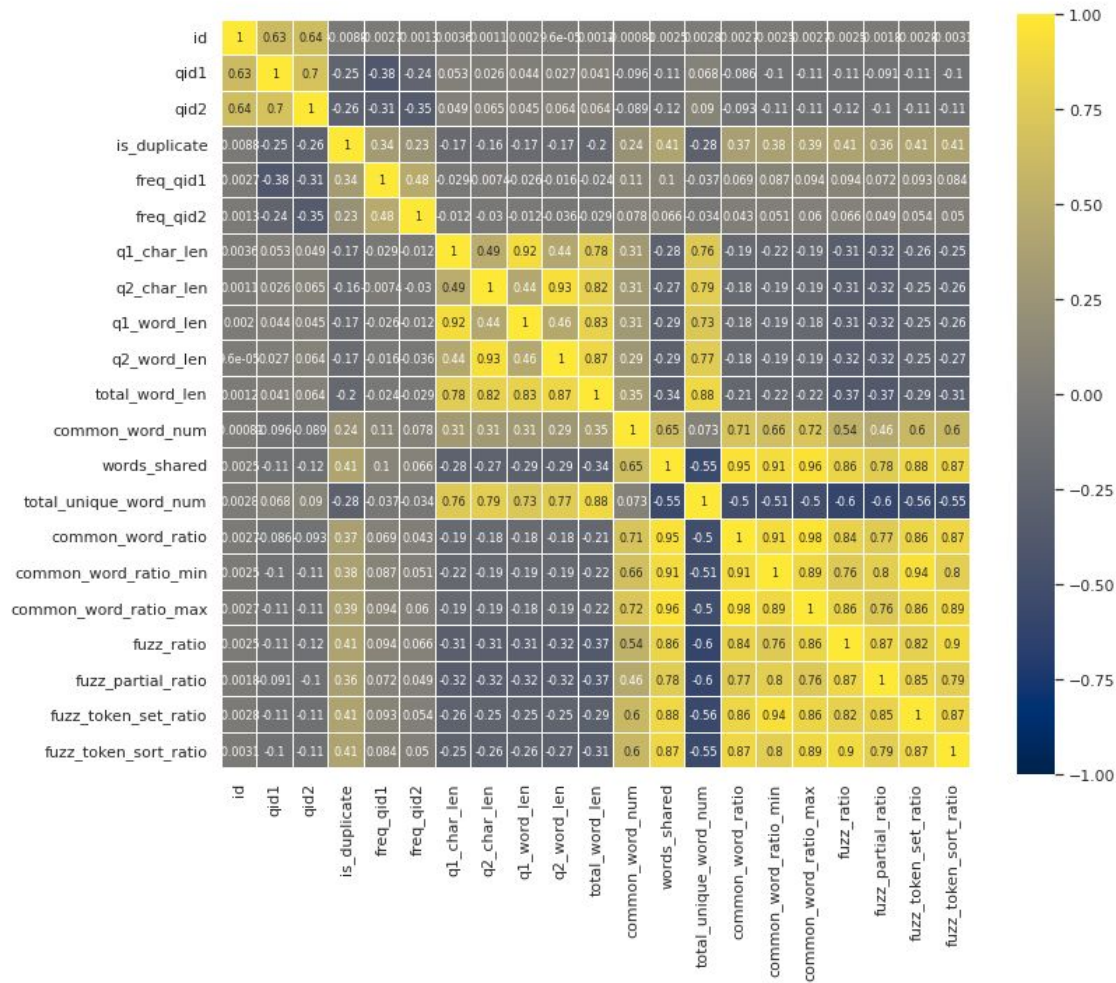


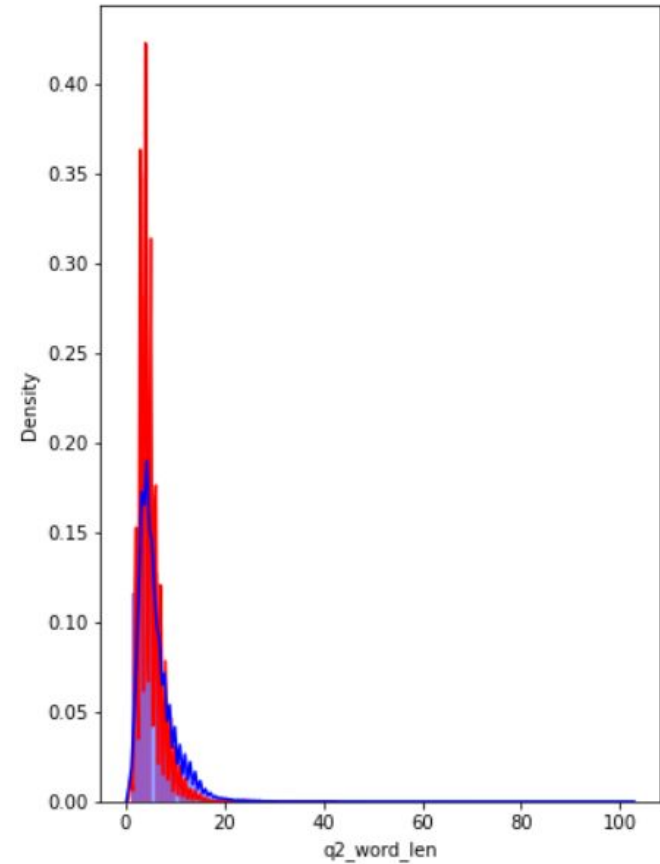
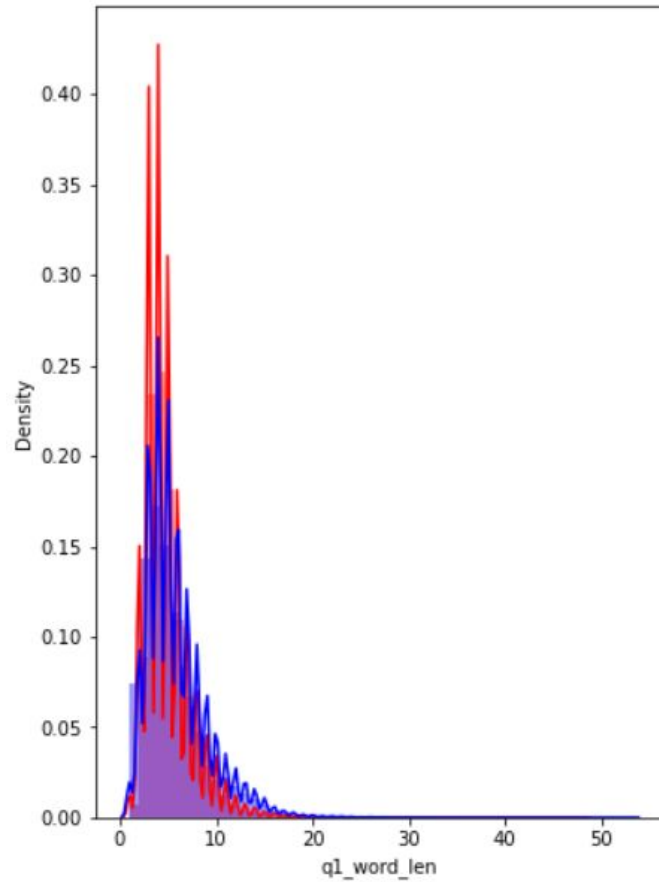
When fuzz_ratio is high, there are more similar questions

Fuzz_ratio is directly proportional to similarity of questions

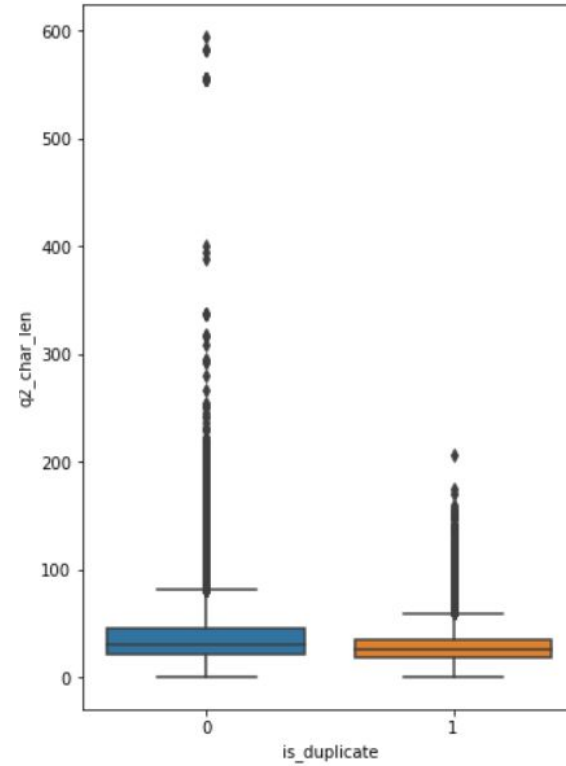
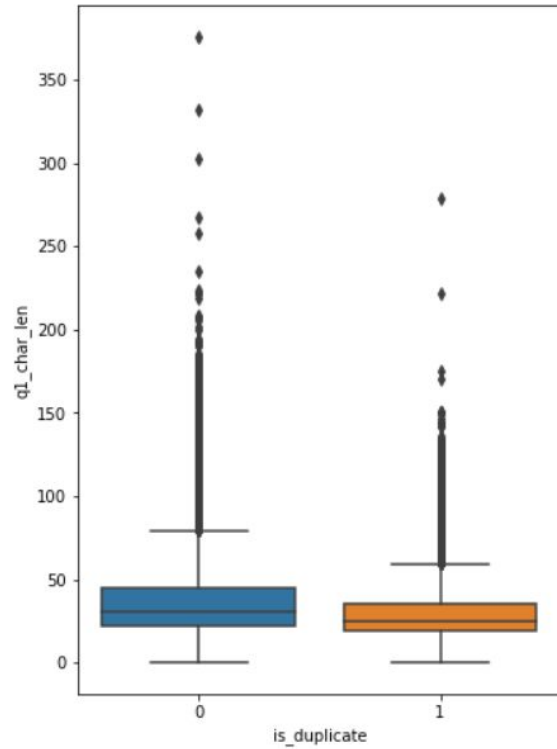


Correlation between different features



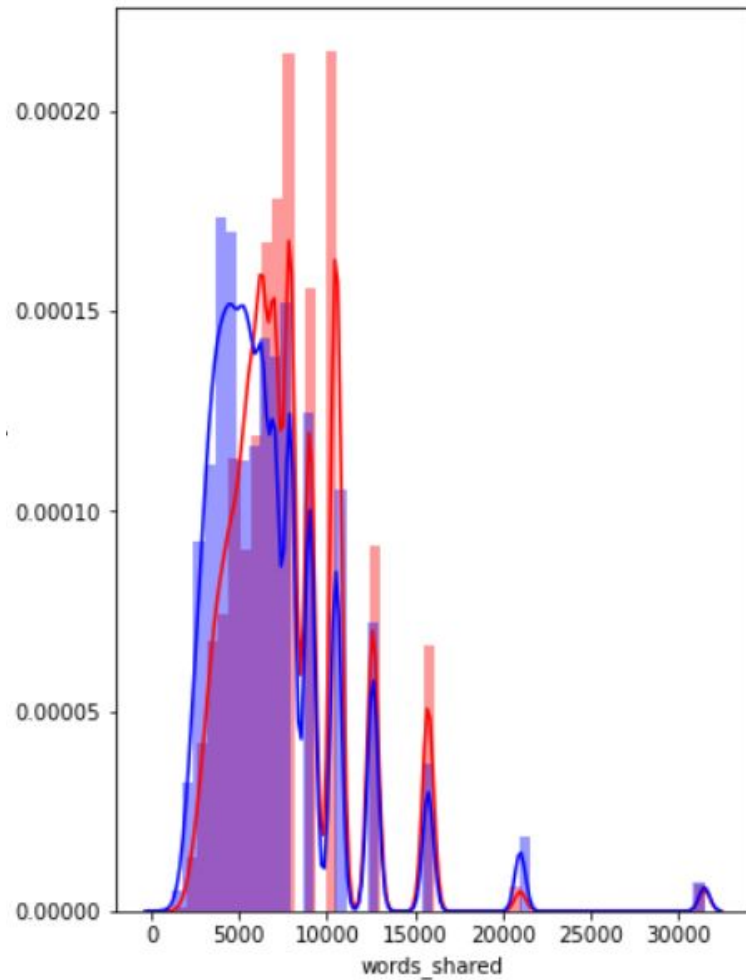


There are more no. of non-similar questions in question1 set than question2 set

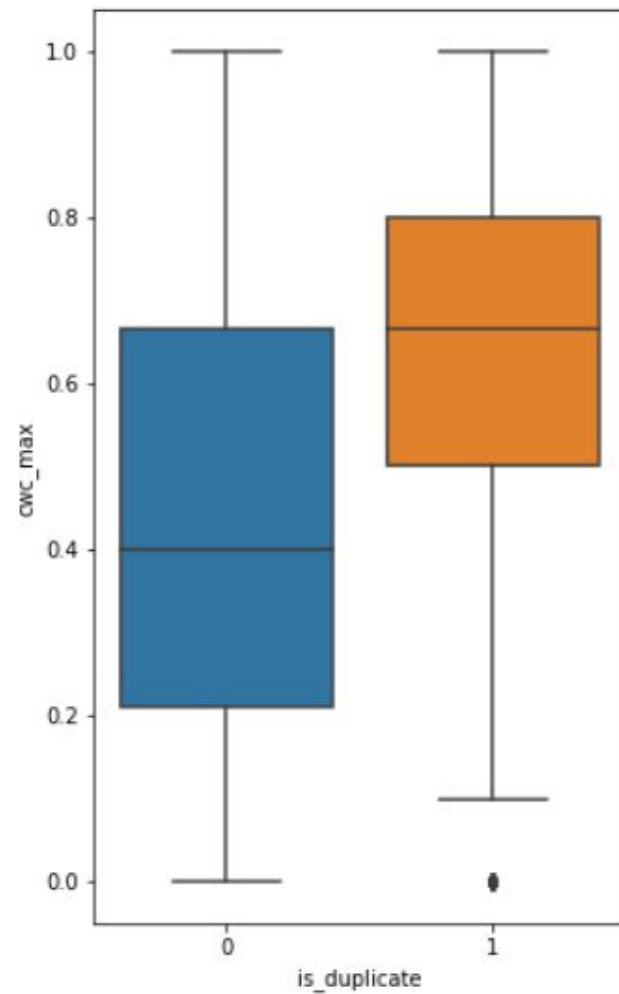


- No of characters in question1 is more than question2 for Non-Similar questions.
- Minimum characters of q2 is same for both Non-Similar questions & Similar questions.

- The distributions of the words_shared feature in Similar and Non-Similar questions are highly overlapping.



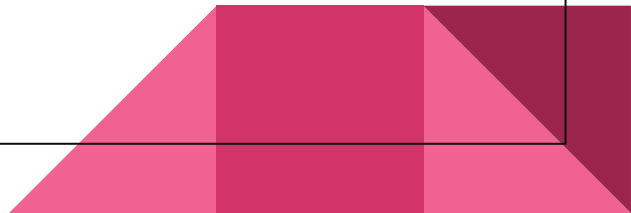
ratio of the number of
common words to the length
of the larger question is high
for similar questions



IV. Vectorization Techniques

Converting text to numerical conversion so that machine can understand easily.

- GloVe (Global and Vectors) – represents words in the form of vectors by mapping the words into space where the semantic similarity between the words is observed by the distance between the words.
- Word2Vec – a continuous bag of words architectures that take a text corpus as input and produces the word vectors of its constituent words as output.
- TF-IDF - intended to reflect how important a word is to a document in a collection or corpus.



MACHINE LEARNING MODELS

After successful cleaning and data pre-processing of the data, We introduce the data to the machine learning models.

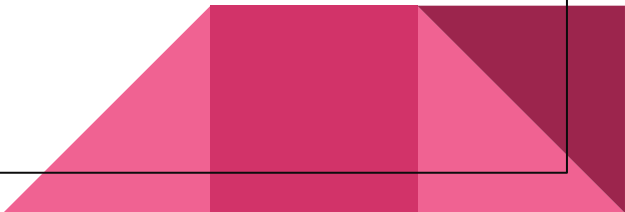
Logistic Regression:

- As it is binary classification problem we prefer to use logistic regression.
- It will predict the probability of an instance belonging to the default class.

Support Vector Machine:

- In SVM the hyperlane, segregates the classes into its own kind.
- Which makes it better model for the classification.

XGBoost:

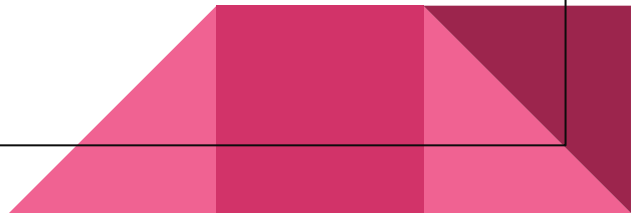
- fast learning through parallel and distributed computing and offers efficient memory usage.
 - an implementation of gradient boosted decision trees
- 

PERFORMANCE METRICS

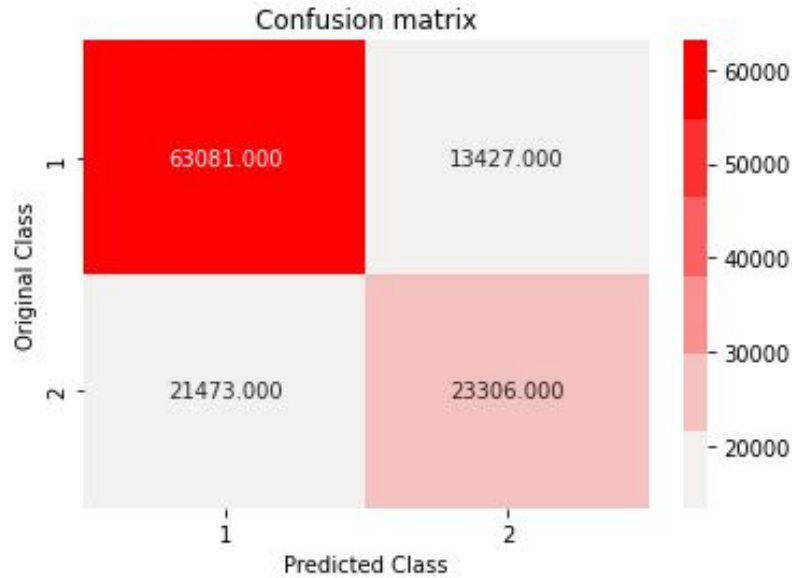
After the training of the model it is required to evaluate it's performance.

Confusion Matrix:

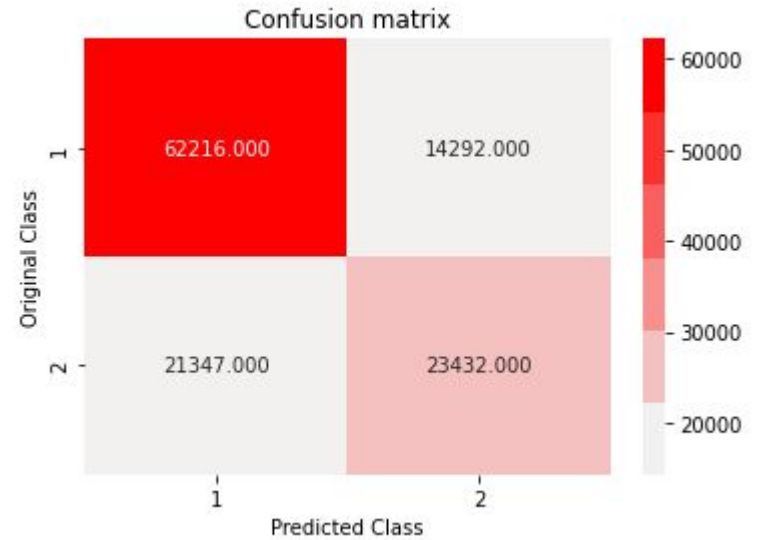
- In confusion matrix, we get the model's prediction and actual prediction displayed in a tabular form.



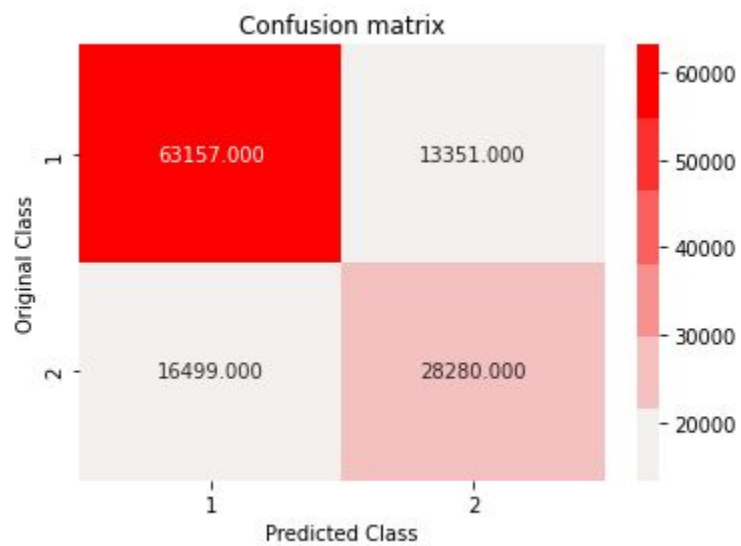
Results



Logistic Regression



SVM



XGBoost

Conclusion

- XGBoost performs better in classification compared with Logistic Regression and SVM.
- We can train the model using Bi-directional LSTM to get some accurate classification.





Thank You