Duplicate Sentence Identifier

Our goal is:-

- To check duplication amongst any two sentences
- Analysing common beliefs through sentences
- Ease of searching in social media platforms, etc.

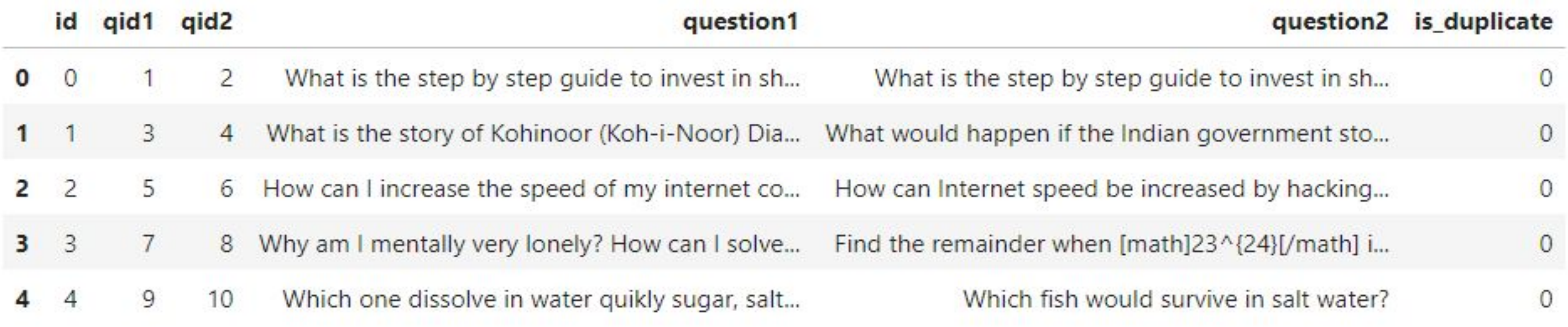
Dataset : Quora Question pairs

- The goal is to predict which of the provided pairs of questions contain two questions with the same meaning.
- Data fields
- **id** - the id of a training set question pair
- **qid1, qid2** - unique ids of each question (only available in train.csv)
- **question1, question2** - the full text of each question
- **is_duplicate** - the target variable, set to 1 if question1 and question2 have essentially the same meaning, and 0 otherwise.

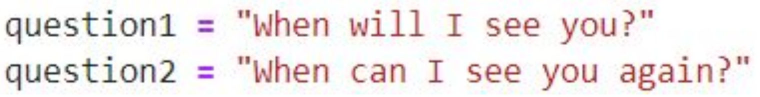
Ethical aspect of project :-

- In social media platform 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. Hence this model identifies similar question/sentences for ease of answering the question.
- While checking answer sheets it can find whether two answers are similar or not, can also act like a plagiarism checker.
- This model can analyze usage of similar keywords by group of people and understand beliefs and can act as a bridge between organization and common people.
- Product with duplicated brand names can also be checked with this model.

Samples :

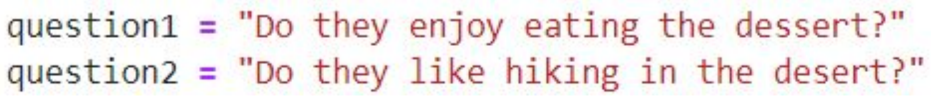


	id	qid1	qid2	question1	question2	is_duplicate
0	0	1	2	What is the step by step guide to invest in sh...	What is the step by step guide to invest in sh...	0
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0
2	2	5	6	How can I increase the speed 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 [math]23^{24}[/math] i...	0
4	4	9	10	Which one dissolve in water quikly sugar, salt...	Which fish would survive in salt water?	0



```
question1 = "When will I see you?"
question2 = "When can I see you again?"
```

True

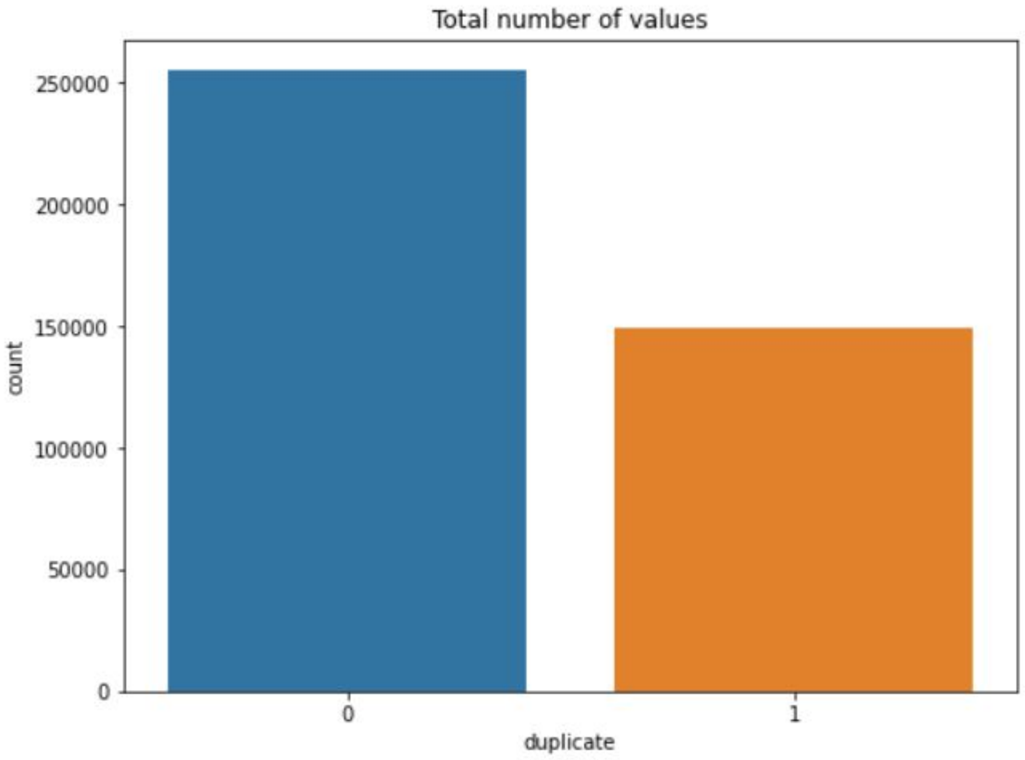


```
question1 = "Do they enjoy eating the dessert?"
question2 = "Do they like hiking in the desert?"
```

False

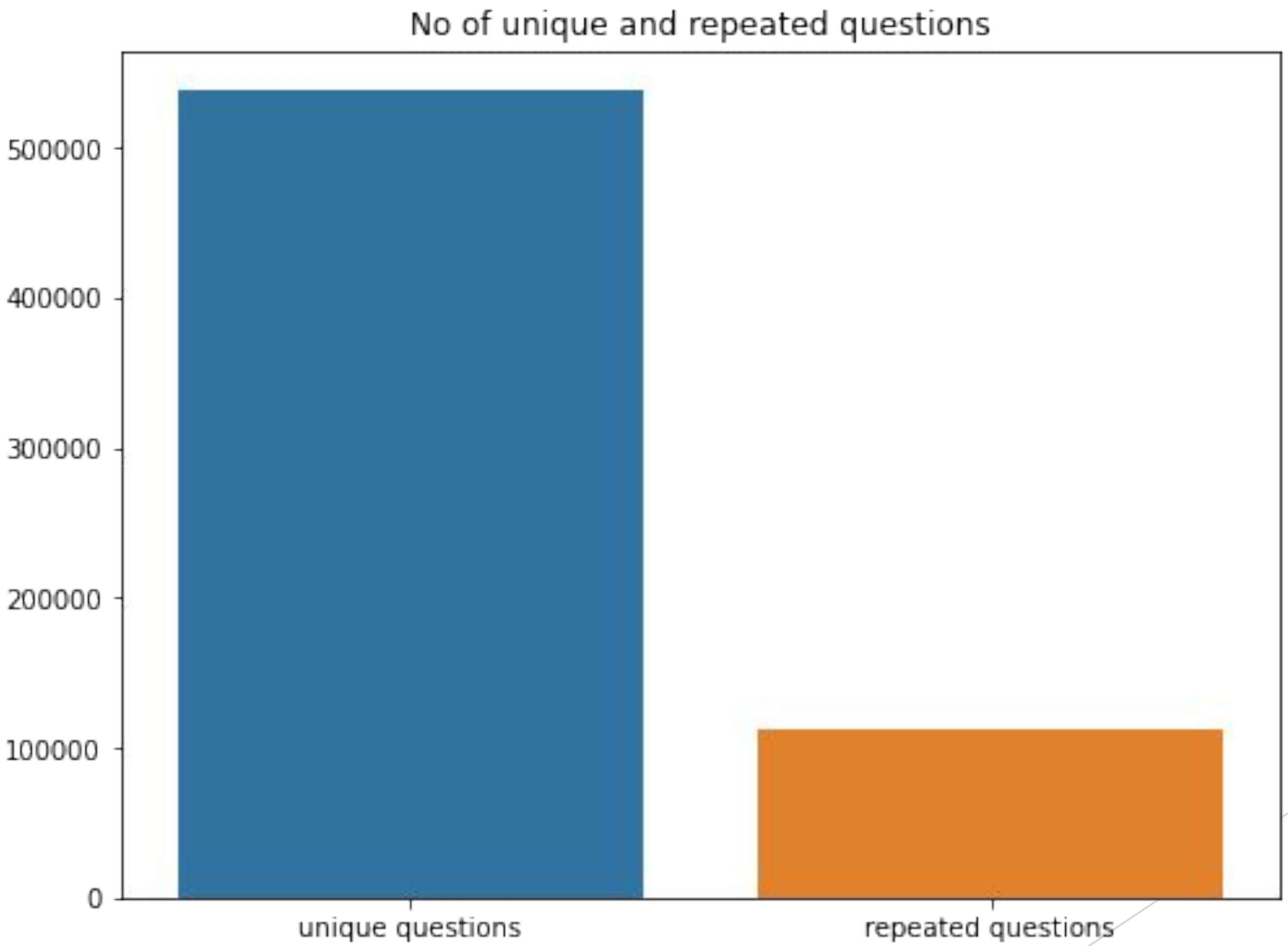
Exploratory Data Analysis

- Distribution of Target variable

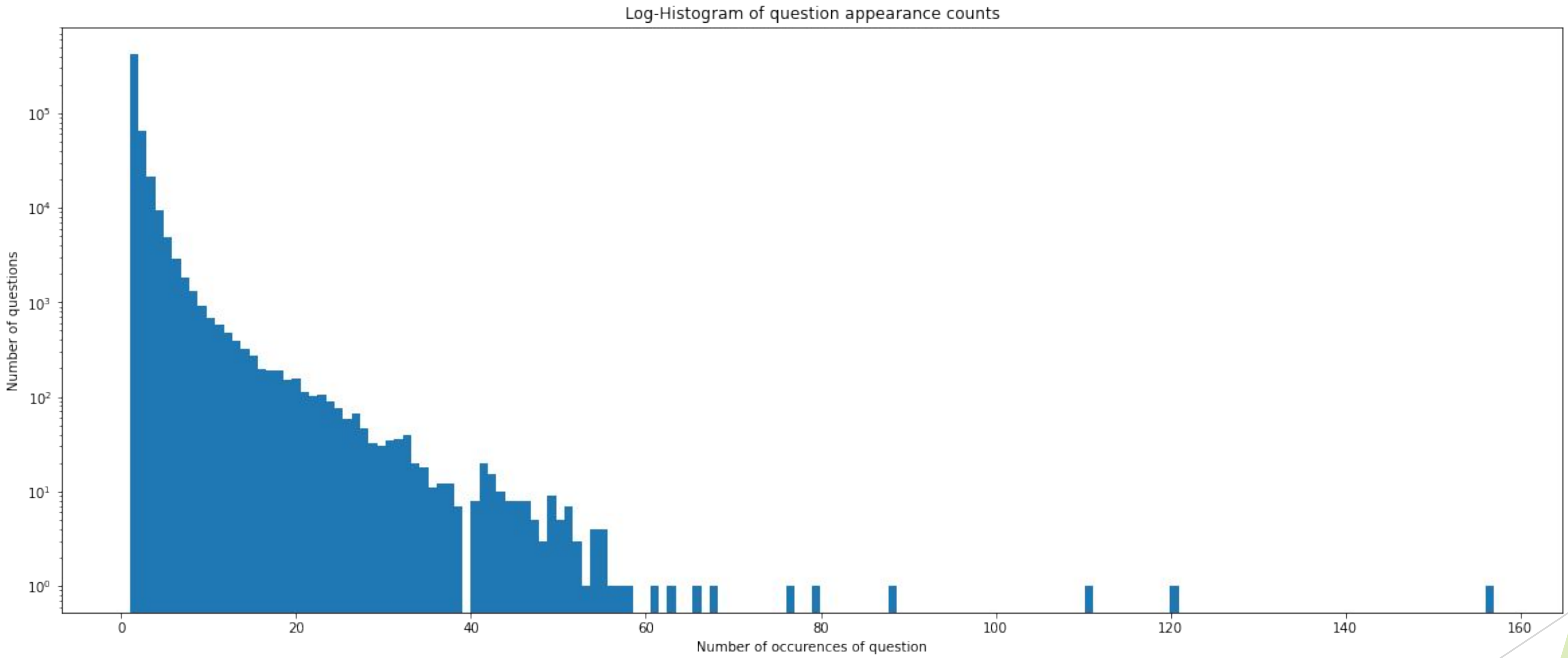


```
Total number of question pairs is 404287
Question pairs that are similar is 149263 which is 37 % of total
Question pairs that are not similar is 255024 which is 63 % of total
```

- Number of Unique and repeated questions
- Total unique questions : 537929
- Total repeated questions : 111778 i.e. 20.78 %
- Maximum no. of times question occurred : 157



Log – Histogram of question appearance counts



Feature Engineering

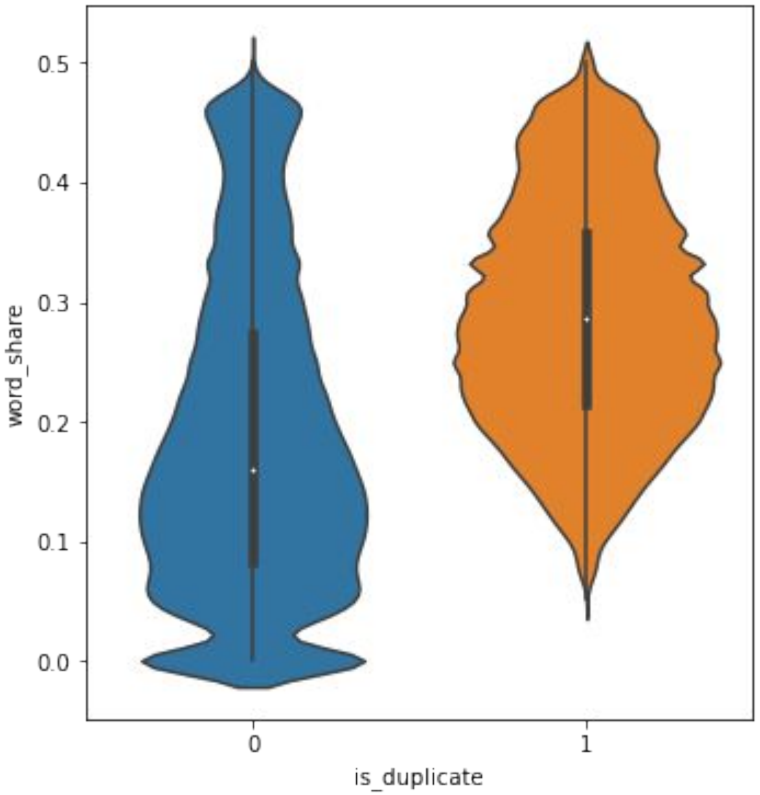
Adding some new features

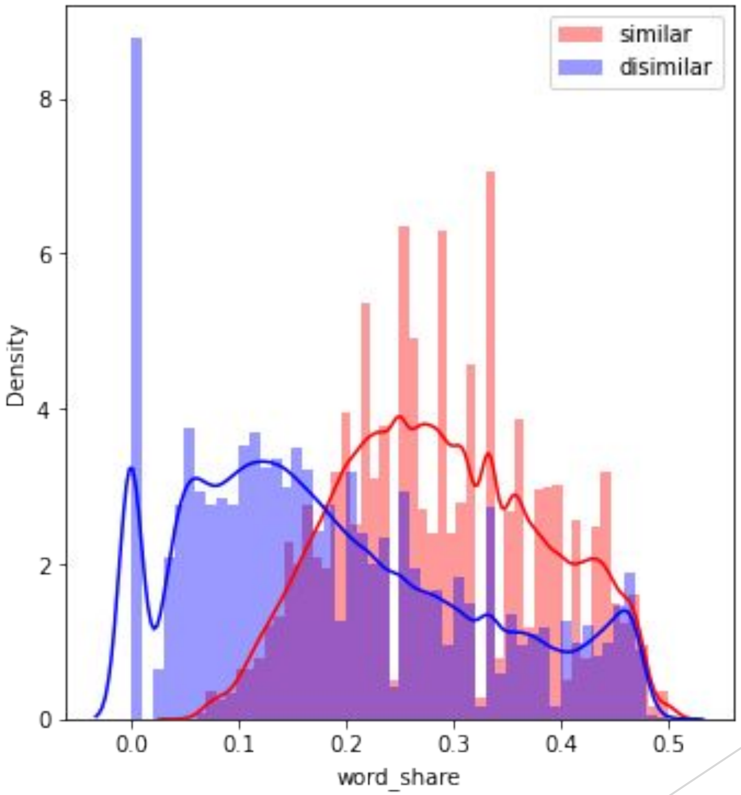
- **word_Common** = Number of common unique words in Question 1 and Question 2
- **word_Total** =Total num of words in Question 1 + Total num of words in Question 2
- **word_share** = word_common)/word_Total
- **freq_qid1** = Frequency of qid1's i.e, number of times question1 occur
- **freq_qid2** = Frequency of qid2's

Effect Of Shared words

As the word share increases there is a higher chance the questions are similar

Histogram word share has some information differentiating similar and dissimilar classes.

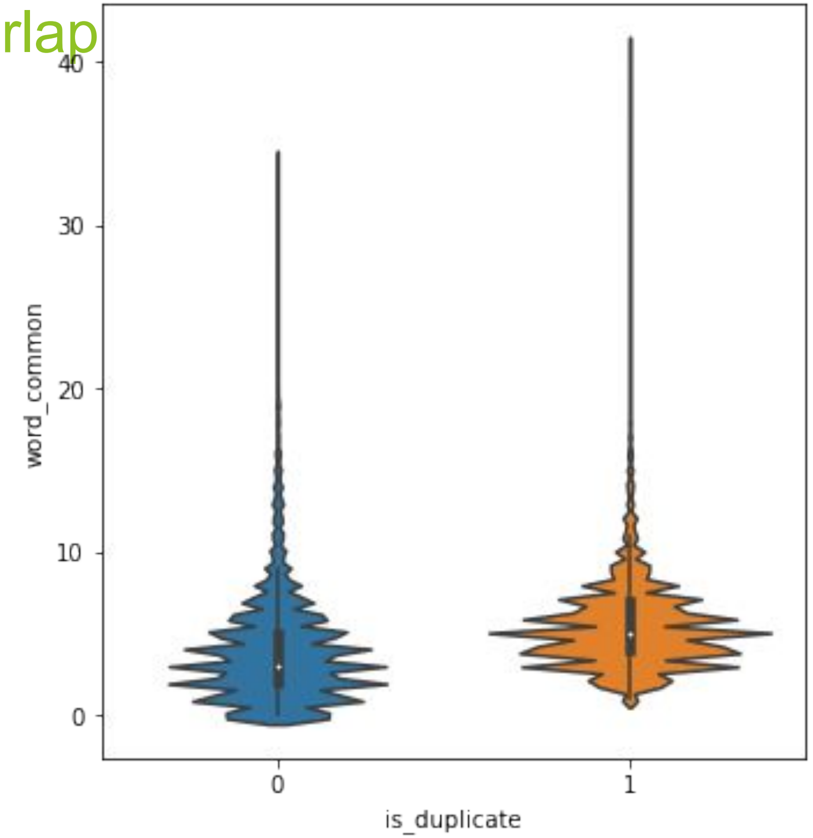


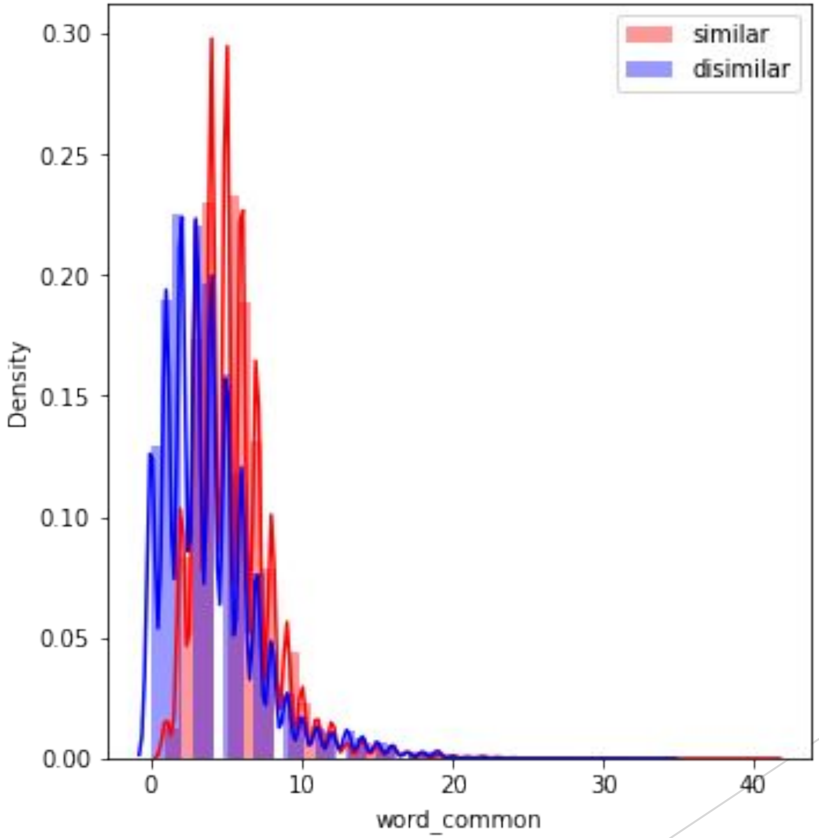


Effect of Common words

Common words doesn’t have enough information about differing classes

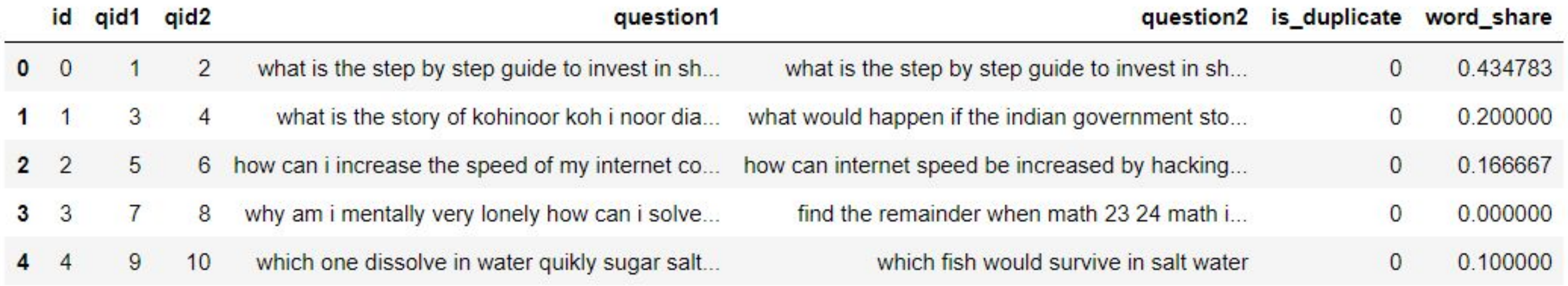
Histogram of duplicate and non-duplicate questions are overlap



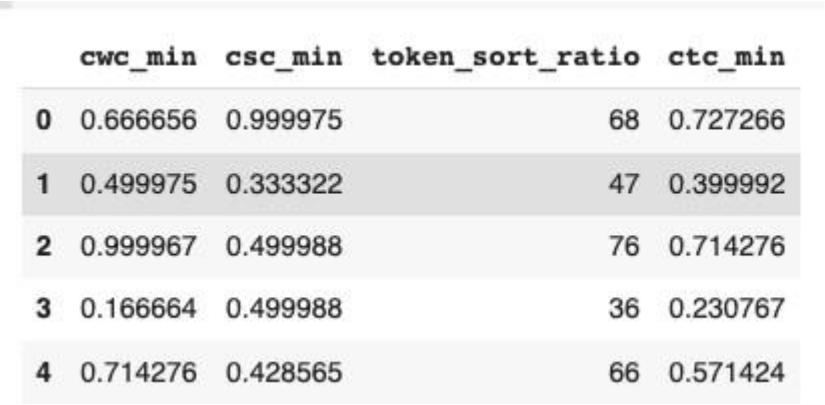


Text Pre-processing : cleaning

- Removing punctuations,
- performing stemming,
- Removing stop-words,
- Expanding contractions, replacing “can’t” with “can not”,“$” with" dollar” etc.



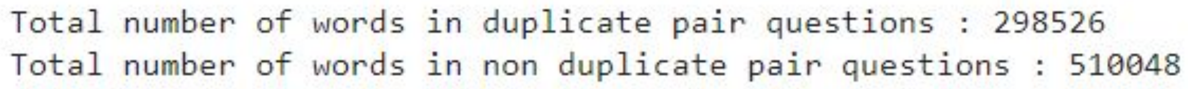
	id	qid1	qid2	question1	question2	is_duplicate	word_share
0	0	1	2	what is the step by step guide to invest in sh...	what is the step by step guide to invest in sh...	0	0.434783
1	1	3	4	what is the story of kohinoor koh i noor dia...	what would happen if the indian government sto...	0	0.200000
2	2	5	6	how can i increase the speed of my internet co...	how can internet speed be increased by hacking...	0	0.166667
3	3	7	8	why am i mentally very lonely how can i solve...	find the remainder when math 23 24 math i...	0	0.000000
4	4	9	10	which one dissolve in water quikly sugar salt...	which fish would survive in salt water	0	0.100000



	cwc_min	csc_min	token_sort_ratio	ctc_min
0	0.666656	0.999975	68	0.727266
1	0.499975	0.333322	47	0.399992
2	0.999967	0.499988	76	0.714276
3	0.166664	0.499988	36	0.230767
4	0.714276	0.428565	66	0.571424

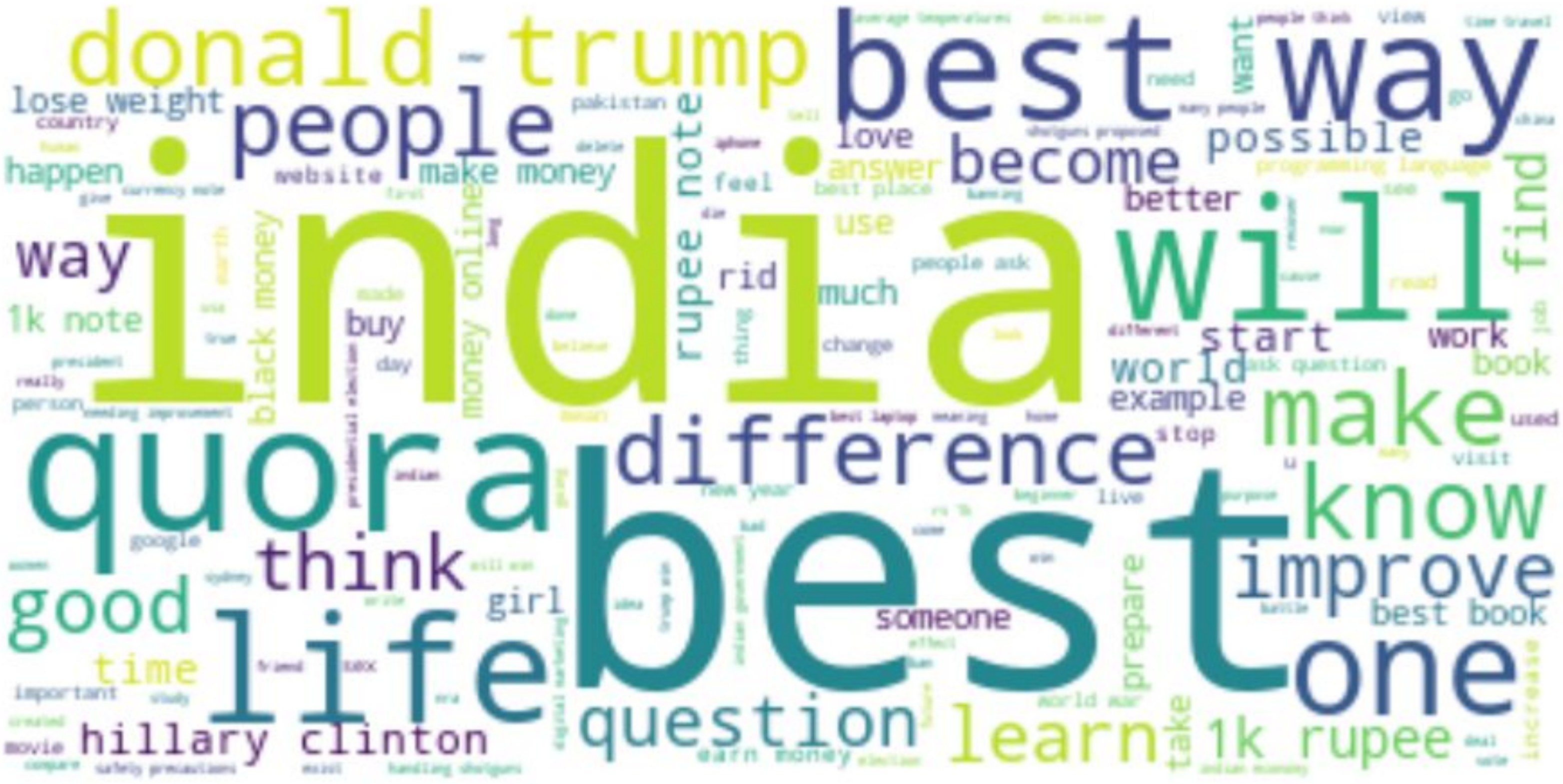
- cwc_min : Ratio of common_word_count to min length of word count of Q1 and Q2
  - common_word_count / (min(len(q1_words), len(q2_words))
- csc_min : Ratio of common_stop_count to min length of stop count of Q1 and Q2
  - common_stop_count / (min(len(q1_stops), len(q2_stops))
- ctc_min : Ratio of common_token_count to min lenghth of token count of Q1 and Q2
  - common_token_count / (min(len(q1_tokens), len(q2_tokens))
- token_sort_ratio: In some other cases even fuzz partial ratio will fail.
  - fuzz.partial_ratio("MI vs RCB","RCB vs MI") ⇒ 72 Actually both the sentence have the same meaning. But the fuzz ratio gives a low result. So a better approach is to sort the tokens and then apply fuzz ratio. fuzz.token_sort_ratio("MI vs RCB","RCB vs MI") ⇒ 100

Plotting the Word Clouds :

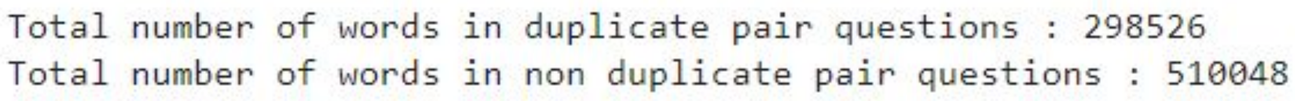


```
Total number of words in duplicate pair questions : 298526
Total number of words in non duplicate pair questions : 510048
```

Duplicate pair of questions :

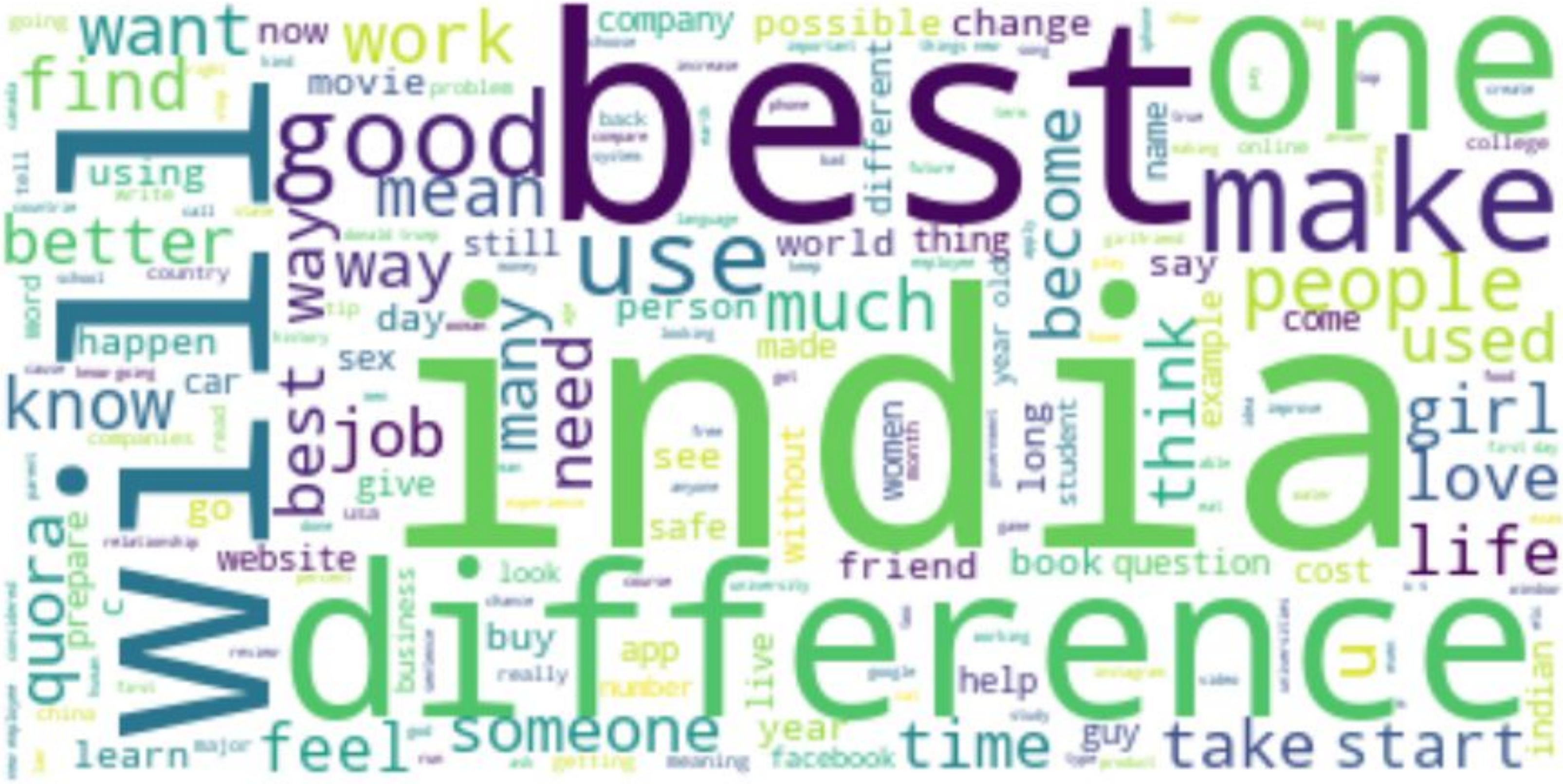


Plotting the Word Clouds :



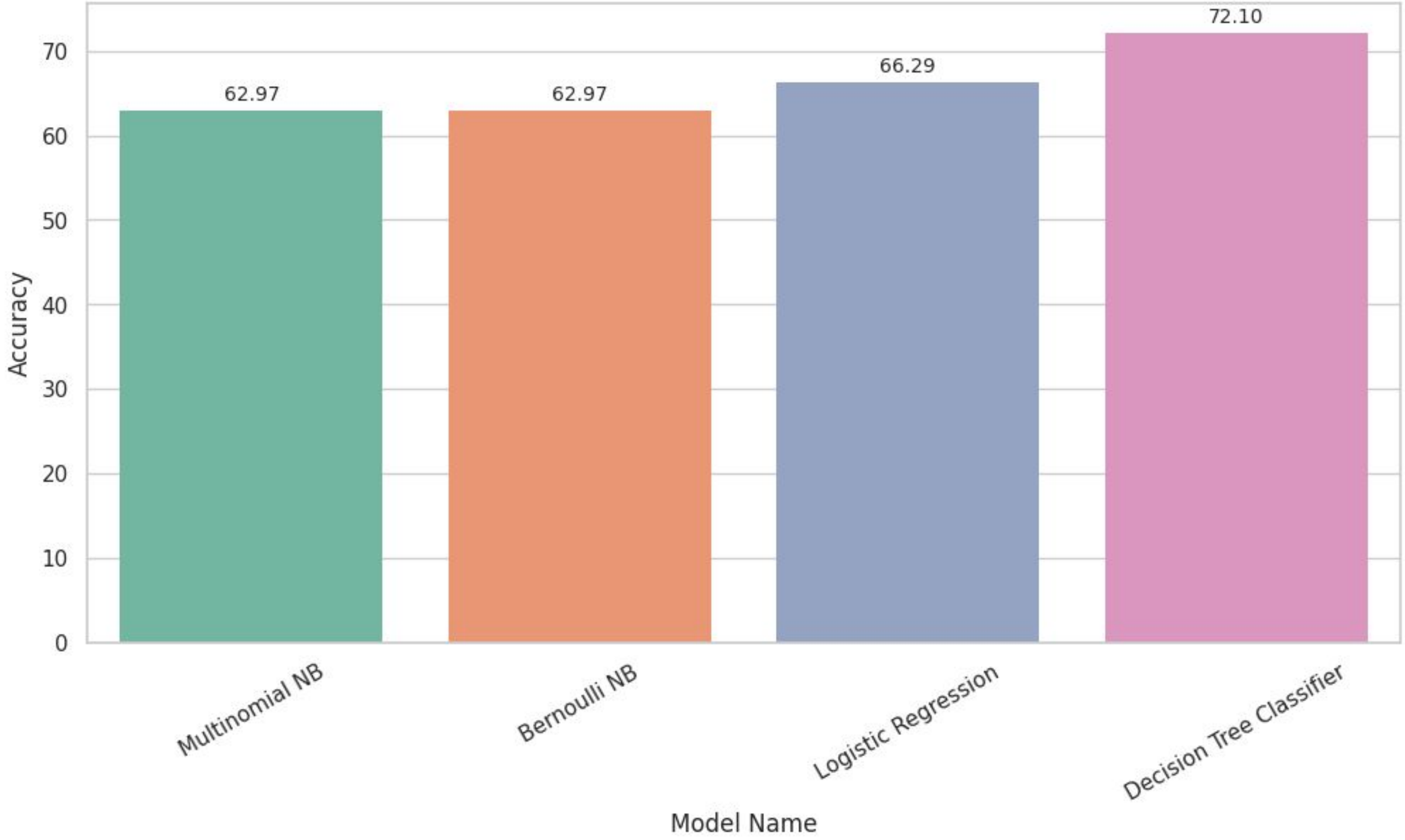
```
Total number of words in duplicate pair questions : 298526
Total number of words in non duplicate pair questions : 510048
```

- Non Duplicate Pair of Questions :

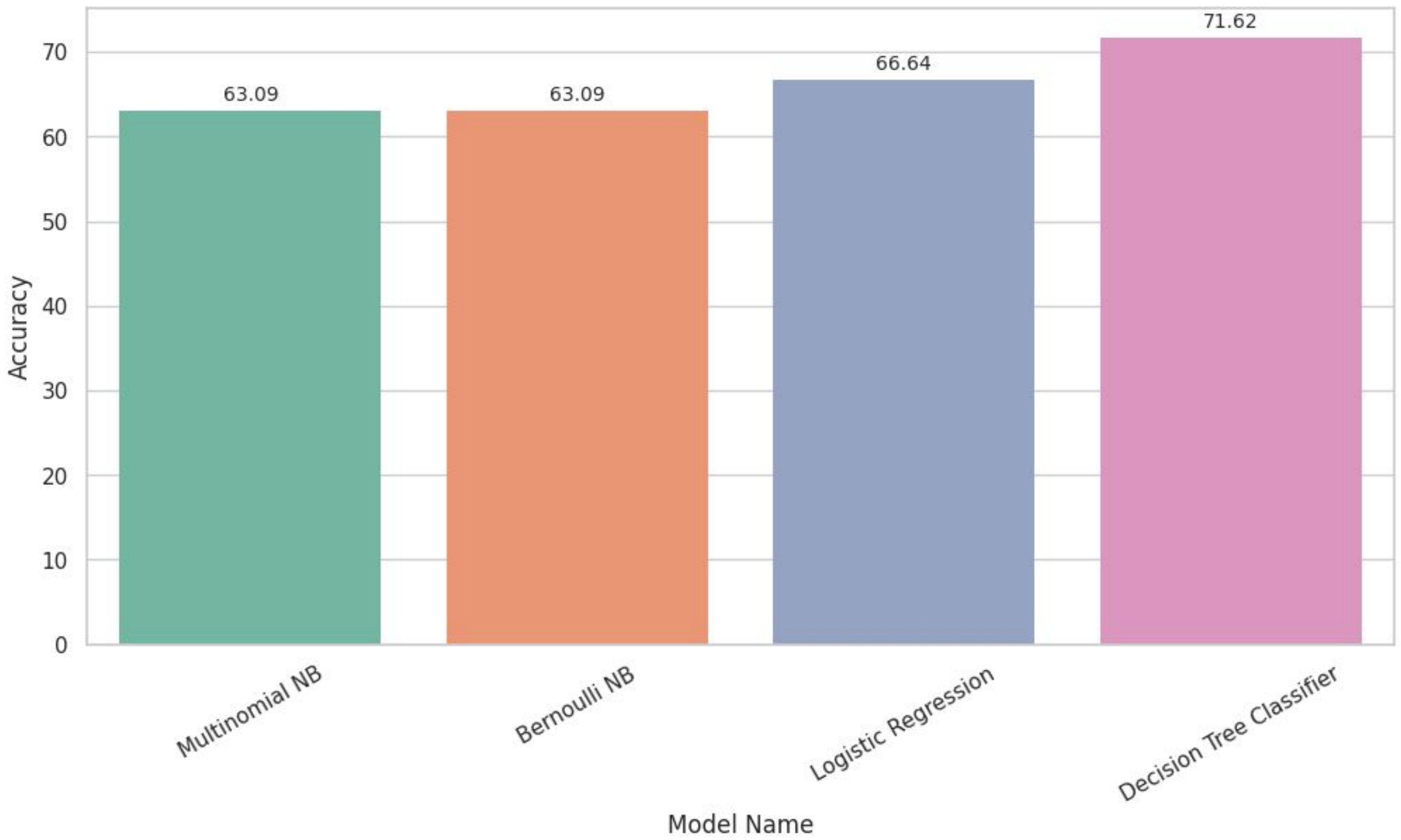


Machine Learning Models :

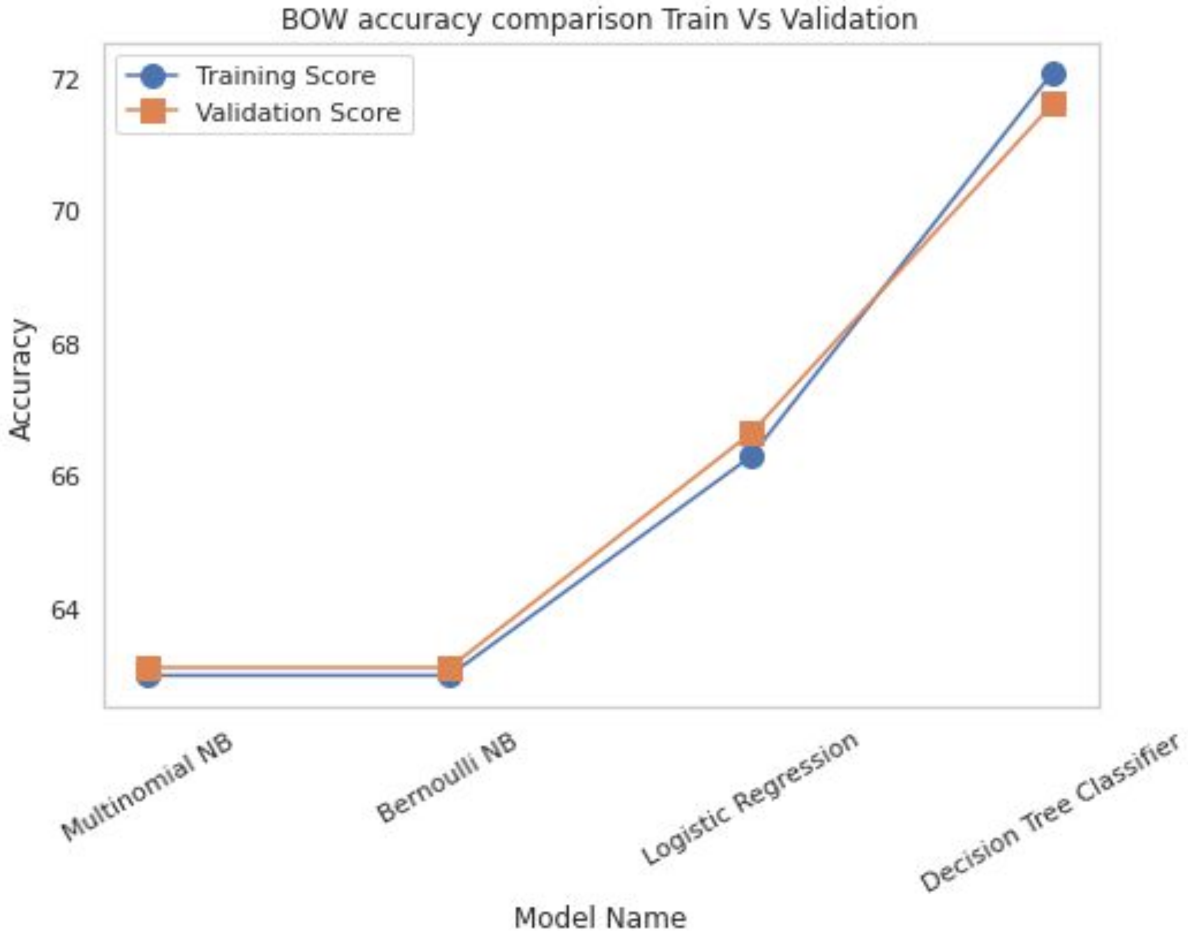
- Training Accuracy comparison of different models:



- Validation accuracy comparison of different models :



Training vs Validation score :


BOW accuracy comparison Train Vs Validation
Training Score
Validation Score
Accuracy
Multinomial NB
Bernoulli NB
Logistic Regression
Decision Tree Classifier
Model Name