

Identification of Quora question pairs
with the same intent
(Project Report by Group-9, CDS-B3)

Contents

Project:	2
Problem statement.....	2
Background information	2
Dataset description and dataset source.....	2
Current benchmark	3
Proposed Plan	3
Feature Engineering	4
Exploratory Data Analysis	6
Model Training – Classical ML models	8
Model Training – Neural Networks and SBERT	10
Deploying Machine Learning Models	14
Edge case analysis and areas for model improvement	15
Conclusion	16
Team members' names	17

Identification of Quora question pairs with the same intent

(Project Report by Group-9, CDS-B3)

PROJECT:

Identification of Quora question pairs with the same intent

PROBLEM STATEMENT

Quora is a platform to gain and share knowledge on any topic. It allows people to ask questions and connect with people who contribute unique insights and quality answers. This equips people to learn from each other and to understand topics in diverse subjects. Over 100 million users visit Quora every month, and a lot of them ask similarly worded questions. Multiple questions with the same intent can cause seekers to spend more time finding the best answer to their question. This also leads to an inefficient system for writers, as they spend more time on answering multiple versions of the same question.

The **goal** of this project is

- To use NLP, ML algorithms and create an application to classify whether Quora question pairs that are already asked are **duplicate** or **not**.
- This could be useful to instantly provide answers to questions that have already been answered.

BACKGROUND INFORMATION

Recent years have witnessed a tremendous growth of community-based question answering (QA) forums, such as Quora, Reddit, StackOverflow etc where people can ask questions in the hopes of attracting high-quality answers. Often, questions that people submit have previously been asked.

Companies can improve user experience by identifying these duplicate entries. This would enable users to find questions that have already been answered and prevent community members from answering the same question multiple times.

In this project, we work on the Quora Question Pairs dataset and try to build a question similarity (binary) classifier using Machine Learning, Neural Networks and advanced NLP techniques.

The Sentence similarity classifier can be **applied** in Text Categorization, Automating CRM processes, Text consolidation for publishing, as plagiarism checker, Medical Records and Text Records Change Management.

DATASET DESCRIPTION AND DATASET SOURCE

The Quora Question Pairs dataset consists of a training set of **404,351** question pairs, and a test set of 2,345,795 question pairs, and is provided as part of a Kaggle competition. Since the test set provided does not contain labels for any question pair, the only measure of performance that can be obtained with this test set is accuracy (via online submission to Kaggle). We therefore felt it better to construct our own test set from the training set provided, since this would allow us to obtain performance metrics other than accuracy, and perform further parameter tuning of our classification models.

We have **404351** training data points. And only **36.92%** are positive. That means it is an imbalanced dataset.

- ✓ Total training data / No. of rows: **404351**
- ✓ No. of columns: **6**
- ✓ **is_duplicate** is the target/dependent variable.
- ✓ Class labels: **0, 1**

Identification of Quora question pairs
with the same intent
(Project Report by Group-9, CDS-B3)

- ✓ No. of non-duplicate data points is **255045**
- ✓ No. of duplicate data points is **149306**
- ✓ Available Columns:
 - **id**: unique ID of each pair
 - **qid1**: ID of first question
 - **qid2**: ID of second question
 - **question1**: text of first question
 - **question2**: text of second question
 - **is_duplicate**: are the questions duplicates of each other (0 indicates not duplicate, 1 indicates duplicate)

id	qid1	qid2	Question1	Question2	Is_duplicate
11	23	24	How do I read and find my YouTube comments?	How can I see all my Youtube comments?	1
21	43	44	What's causing someone to be jealous?	What can I do to avoid being jealous of someone?	0

Dataset source: - <https://www.kaggle.com/competitions/quora-question-pairs/data>

CURRENT BENCHMARK

Current accuracy benchmark is 92.3% with XL Net based model. Top 5 benchmarks as per Papers with Code are as below. Top 10 models in this list are deep neural networks based on BERT and T-5 architectures

Rank	Model	Accuracy ↑	Paper	Code	Result	Year	Tags
1	XLNet (single model)	92.3%	XLNet: Generalized Autoregressive Pretraining for Language Understanding			2019	
2	DeBERTa (large)	92.3%	DeBERTa: Decoding-enhanced BERT with Disentangled Attention			2020	
3	ALBERT	90.5%	ALBERT: A Lite BERT for Self-supervised Learning of Language Representations			2019	
4	T5-11B	90.4%	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer			2019	
5	MLM+ subs+ del-span	90.3%	CLEAR: Contrastive Learning for Sentence Representation			2020	

Source: [Quora Question Pairs Benchmark \(Question Answering\) | Papers With Code](#)

PROPOSED PLAN

Approach:

While the top performing models to solve this problem are based deep neural networks, we plan to take the sequential learning approach as mentioned below:

- a. Exploring classical ML models such as Logistic regression, Support Vector Machines, RandomForest, XG Boost, LGBM and iteratively improving accuracy scores through feature engineering and data cleaning methods
- b. Applying deep neural networks such as C-BOW, sentence BERT
- c. Deep dive on edge cases, identify areas for further improving the accuracy of the models

Model development will be on the below stages:

- I. **Text pre-processing:** Remove html tags, spaces, stop words, Porter stemming/ Lemmatize tokens

Identification of Quora question pairs
with the same intent
(Project Report by Group-9, CDS-B3)

- II. **Feature Extraction and Engineering:**
 - i. Construct NLP & Fuzzy Features from Question pairs. Convert Question pairs to Word vectors using TF-IDF, GloVe and other word (C-BOW uses Word2Vec) embedding techniques. Any one of these or merged features will be used in the model training.
 - ii. Feature engineering: Percentage of common tokens, length of questions, type of question (How or why question), n-grams and other similarity characteristics
- III. **Exploratory data analysis:** Visualization and understanding the importance of features
- IV. **Model Training:** Train ML (Logistic Regression, SVM & XGBoost), NN (C-BOW) and NLP (Sentence-BERT)
- V. **Model Evaluation & Hyperparameter Tuning:** Evaluate Model on metric “log-loss” and classification metric “F1-Score”, fine tune Hyper parameters to achieve. Predict on Test data and submit results on Kaggle competition.

Stages with defined deliverables:

Sr No	Stage	Deliverable
1	Feature Extraction and Engineering	Requisite features are identified for further analysis
2	Exploratory data analysis	Understanding importance of features through visualizations
3	Model Training – XG-Boost, Random forest, NN, Sentence BERTetc	Training outcomes using XG-Boost, Randomforest and other NLP models
4	Ensemble learning and fine tuning	Hyper parameter tuning, stacking and finalizing the model
5	Final report and data story	Present draft results and user story
6	Final presentation	Present final results

FEATURE ENGINEERING

228 features were basic and advance features were created through iterative data analysis and feature engineering process. Final list of features is as below:

Base features (13 features):

1. Length of question 1: q1lens
2. Length of question 2: q2lens
3. Same question type (what, why, when ,where, how): samequestion
4. Common words: commonwords, %commonwords
5. POS features:
 - a. 'q1_verb_count'
 - b. 'q2_verb_count'
 - c. 'q1_noun_count'
 - d. 'q2_noun_count'
 - e. 'q1_adj_count'
 - f. 'q2_adj_count'
 - g. 'q1_adv_count'
 - h. 'q2_adv_count'

Identification of Quora question pairs
with the same intent
(Project Report by Group-9, CDS-B3)

Fuzzy features (4 features)

([FuzzyWuzzy: Fuzzy String Matching in Python | Towards Data Science](#))

1. Fuzz Ratio (fuzz_ratio): This is perfect for strings with similar lengths and order
2. Fuzz Partial Ratio (fuzz_partial_ratio) - calculates the FuzzyWuzzy ratio for all substrings of the longer string with the length of the shorter one, and then returns the highest match – better for strings with differing lengths
3. Token set ratio (token_set_ratio) - This ratio separates each string into words, turns both lists into sets (discarding repeated words) and then sorts those before doing the ratio) – if strings have same meaning but order is different
4. Average Fuzz Score: Average of above three fuzzy features

Derived features (6 features):

1. Ratio of common word count with min question pair word count in denominator: cwc_min
2. Ratio of common word count with max question pair word count in denominator: cwc_max
3. Ratio of common non-stop words to min question pair word count in denominator: ctc_min
4. Ratio of common non-stop words to max question pair word count in denominator: ctc_max
5. Absolute length difference of q1lens and q2lens: abs_len_diff
6. Average lengths of the question pair: mean_len

100-dimension vectors TFIDF based weights and Glove Embeddings (200 features):

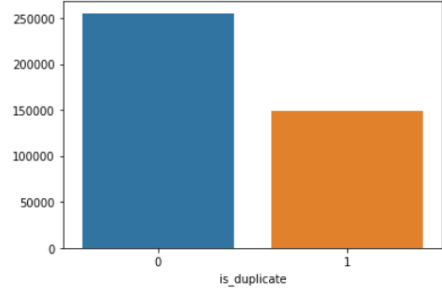
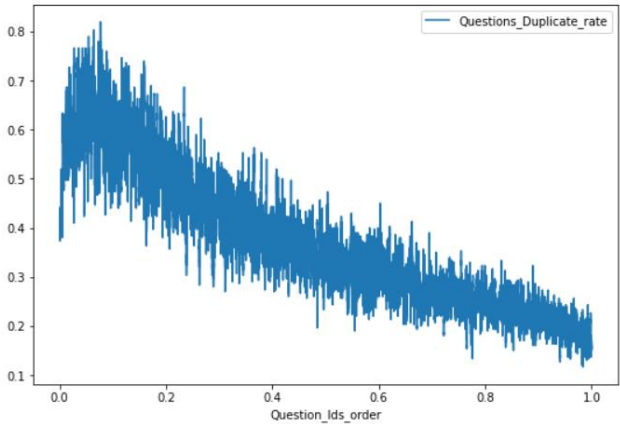
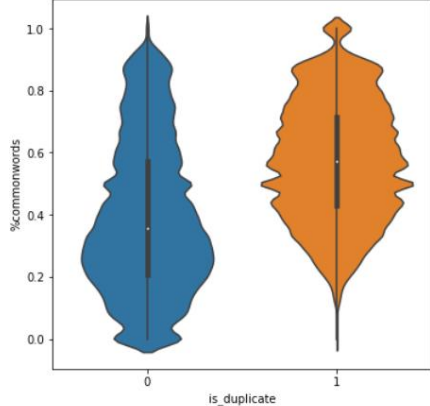
1. Vocabulary of words formed using train dataset and tfidf vectorizer
2. GLOVE 100d embeddings applied to TFIDF vocabulary
3. 100 features obtained for each question in the question pair by taking average of sumproduct of GLOVE embeddings and TFIDF weights. UNK token was used when applying this calculation

Distances between question pairs using the 100-dimension vectors (5 features)

1. Cosine distance
2. Cityblock distance
3. Canberra distance
4. Euclidean distance
5. Sequence Ratio

Identification of Quora question pairs
with the same intent
(Project Report by Group-9, CDS-B3)

EXPLORATORY DATA ANALYSIS

Sr	Observation	Outcomes
1	<ul style="list-style-type: none"> • 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. So only 36.92% are positive. That means it is an imbalanced dataset • 3 question pairs have null values 	<pre>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></pre> 
2	<ul style="list-style-type: none"> • Observed decreasing trend of Quora Question duplicates as the Question Id increases. Even though QID will be a good predictor but defeats the purpose of good model based on actual Questions text. • Train data is shuffled to overcome above issue and for better model learning. 	
3	<ul style="list-style-type: none"> • Repetition of questions: 99.7% of the questions are repeated less than 5 times • 138 questions are repeated more than 40 times 	<pre>bucket (0, 5] 806031 (5, 10] 1742 (10, 20] 644 (20, 30] 147 (30, 40] 0 (40, 50] 138 (50, 5000] 0</pre>
4	<ul style="list-style-type: none"> • Violin plot is thicker for duplicate questions when %common words are more indicating impact of this feature in identifying duplicate questions 	

Identification of Quora question pairs
with the same intent
(Project Report by Group-9, CDS-B3)

5	<ul style="list-style-type: none"> #52% of the question pairs are of same type (what, why, where, how, when), 47% are of different type #Probability of duplicate question increases when question type is same 	<pre> samequestion 0 1 is_duplicate 0 0.538690 0.461310 1 0.362753 0.637247 All 0.473692 0.526308 </pre>
6	<ul style="list-style-type: none"> Correlation of 0.35 to 0.4 observed between fuzzy features and the outcome metric As expected avgfuzzscore is correlated with other features. We will continue with four features as they are measuring impact due to different aspects 	
7	<ul style="list-style-type: none"> 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 	
8	<ul style="list-style-type: none"> Non-Stop word ratio (CTC metrics), common word count ratios and %commonwords have high correlations with duplication of question 	

Identification of Quora question pairs
with the same intent
(Project Report by Group-9, CDS-B3)

MODEL TRAINING – CLASSICAL ML MODELS

Iteration 1:

Stage 1: Data Cleaning

Covert to lower → remove non ascii characters → remove punctuation → remove stop words → remove numbers → lemmatize

Stage 2: Features:

q1lens, q2lens, commonwords, samequestion, verb-noun-adj-adv counts, token_set_ratio, fuzz ratio, fuzz partial ratio, tfidf based glove embeddings (200 vector features)

Stage 3: Model development:

	Logistic Regression	Linear SVM	XG Boost
Accuracy	65.56%	65.59%	71.30%

Stage 4: Edge case analysis:

Observed misclassifications due to the following reasons:

- Stop words impacted the intent of question, hence decided not to remove but add two stop word related features
- Numbers impacted the intent of question, hence decided to transform numbers to their text equivalents and not remove them

Iteration 2:

Stage 1: Data Cleaning

Covert to lower → remove non ascii characters → remove punctuation → lemmatize

Stage 2: Features:

- Iteration 1 features
- Derived features
- Distance features
- Fuzzy features – Added average fuzz ratio

Stage 3: Model development:

	Logistic Regression	Linear SVM	XG Boost
Accuracy	72.60%	67.30%	82.35%

Stage 4: Edge case analysis

- Impact of non-special characters, numbers
- Impact of words not in tfidf/ Glove dictionary unidentified words
- Impact of sequence of tokens

Iteration 3:

Stage 1: Data Cleaning

1. Covert to lower
2. Replace special characters with their string equivalents like % with percent, \$ with dollar
3. Replace numbers with billion, million, k etc
4. De-contracting words: ain't --> am not etc

Identification of Quora question pairs with the same intent (Project Report by Group-9, CDS-B3)

5. Removed html tags
6. Removed punctuations
7. Removed non ascii characters
8. Lemmatization
9. Do not remove stop words

Stage 2: Features:

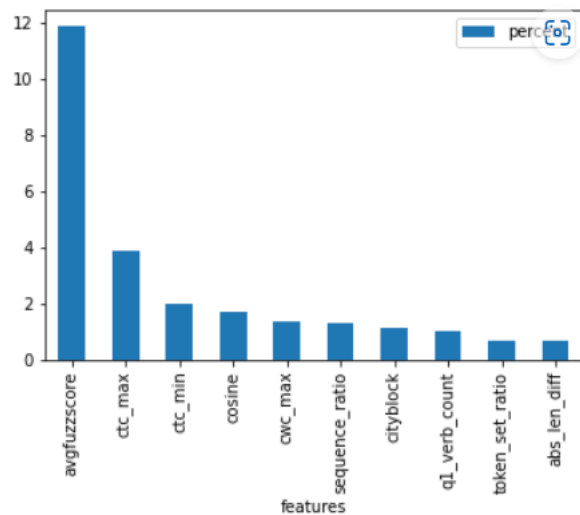
- Iteration 1 & Iteration 2 features
- Added weights for unknown words using UNK token in GLOVE
- Feature – Sequence Ratio added to address weights for sequencing of tokens

Stage 3: Model development:

	Logistic Regression	Linear SVM	XG Boost	LGBM
Accuracy	71.20%	71.90%	84.00%	82.63%

Stage 4: 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



Logistic Regression:

```
Accuracy: 0.7051286279683378
F1 score: 0.6666542383624467
Recall: 0.7968979810135045
Precision: 0.5730034610947314
```

clasification report:

	precision	recall	f1-score	support
0	0.85	0.65	0.74	38203
1	0.57	0.80	0.67	22437

Linear SVM:

Identification of Quora question pairs
with the same intent
(Project Report by Group-9, CDS-B3)

```
Accuracy: 0.7129782321899736
F1 score: 0.6676341971088663
Recall: 0.7791148549271293
Precision: 0.5840628132308721
```

clasification report:

	precision	recall	f1-score	support
0	0.84	0.67	0.75	38203
1	0.58	0.78	0.67	22437

LGBM

```
Accuracy: 0.8236147757255937
F1 score: 0.7602331315848464
Recall: 0.7557605740517894
Precision: 0.7647589410544355
```

clasification report:

	precision	recall	f1-score	support
0	0.86	0.86	0.86	38203
1	0.76	0.76	0.76	22437

XG Boost

MODEL TRAINING – NEURAL NETWORKS AND SBERT

We decided to start with the feed forward neural network model based on C-BOW (Word2Vec) and gradually try more complicated models. For these models, questions text has been cleaned, removed all punctuation, lower case, de contracted abbreviations, didn't remove stop words and contracted big numbers to k, m and b.

We ultimately tested the following 3 models,

1. Continuous Bag of Words (CBOW-Word2Vec)
2. Long Short Term Memory Recurrent Neural Network (LSTM)
3. Sentence-BERT

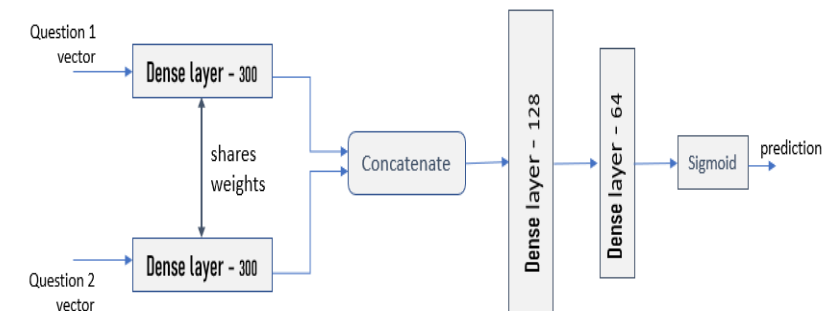
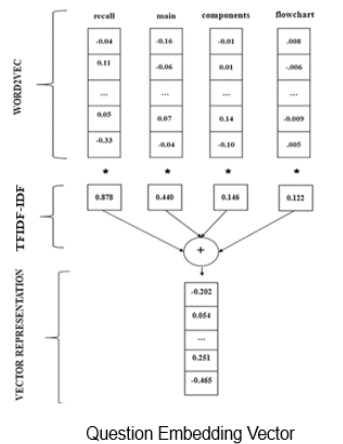
C-BOW Word2Vec & TFIDF Weighted Word2Vec:

C-BOW model uses a single vector representation for each question, and use both question's representations to compute the label prediction. The vector representation (of dimension 300) for a question was simple sum of WORD2VEC (C-BOW) embedding representations of its words. Here we trained Gensim model for WORD2VEC embeddings on QQP corpus but the word embeddings obtained from Google News Corpus gave better results.

Each Question embedding vector was passed on to Siamese Dense layer (300 nodes, shared layer between Q1 and Q2 embeddings) and the concatenated output was fed to a deep, 2-layer [128, 64 nodes] feedforward neural network and then followed by a sigmoid classifier.

Identification of Quora question pairs with the same intent (Project Report by Group-9, CDS-B3)

With the same above model, tried TFIDF's Inverse Document Frequency (IDF) weighted WORD2VEC embeddings (IDF * WORD2VEC) for each word in Question text. The Question's vector is again simple sum of IDF weighted WORD2VEC embeddings of its words. We didn't observe significant difference in Accuracy or F1 score between the models using above mentioned embeddings.



Quora Questions Pairs Classifier Neural Network based on Word2Vec

The below table lists Test-set Accuracies and F1-Scores on QQP test-set.

FNN - Question Word embeddings	Accuracy	F1 Score	Classification Report			
Word2Vec	81.47	74.50		pre	rec	F1
			0	85	86	85
			1	76	73	75
TFIDF-IDF weighted Word2Vec	81.00	74.73		pre	rec	F1
			0	86	84	85
			1	73	76	75

This Model performs well if similar words or synonyms present in Questions and doesn't consider semantics or sequence of words, this helps when LSTM or SBERT fails to predict duplicates (around 4% of data falls in this area).

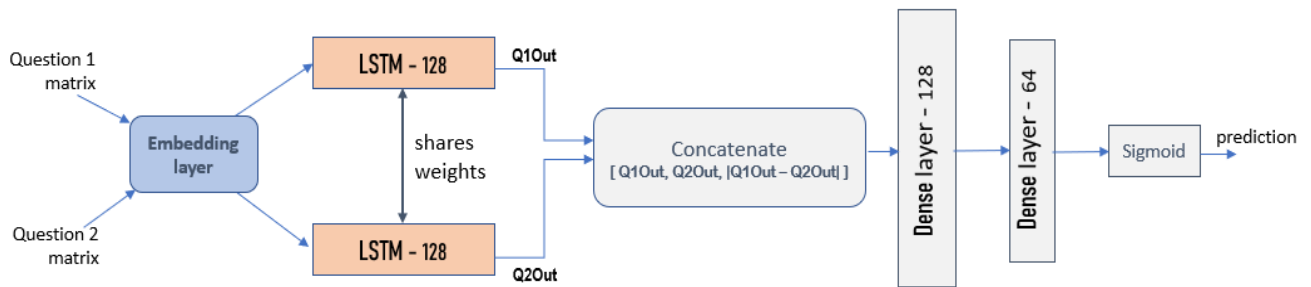
Siamese LSTM with Word2Vec & TFIDF Weighted Word2Vec:

Siamese LSTM model uses an Embedding layer comprising either Word2Vec Or TFIDF-IDF weighted Word2Vec embeddings for all the Words in the QQP corpus. Each Question is represented by a Matrix (of dimension 216x300), and use both question's matrices to compute the label prediction. The matrix representation for a Question was collection of column vectors of Word2Vec embedding representations of its words.

Each Question embedding matrix was passed on to Siamese LSTM layer (128 nodes – shares weights, shared layer between Q1 and Q2 matrix embeddings) and the concatenated output of Q1Out, Q2Out and absolute difference of Q1Out & Q2Out was fed to a deep, 2-layer [128, 64 nodes] feedforward network, followed by a sigmoid classifier.

Identification of Quora question pairs with the same intent (Project Report by Group-9, CDS-B3)

With the same above model, tried TFIDF's Inverse Document Frequency (IDF) weighted Word2Vec embeddings ($IDF * WORD2VEC$) for each word in Question text. The Question's matrix is again collection of column vectors of IDF weighted Word2Vec embeddings of its words. We didn't observe significant difference in Accuracy Or F1 score between the models using above said embeddings.



Siamese LSTM QQP Classifier with Word2Vec & TFIDF-IDF weighted Word2Vec Embeddings

The below table lists Test-set Accuracies and F1-Scores on QQP test-set.

LSTM - Question Matrix embeddings	Accuracy	F1 Score	Classification Report			
Word2Vec	84.39	78.86		pre	rec	F1
			0	87	88	88
			1	79	79	79
TFIDF-IDF weighted Word2Vec	84.00	78.80		pre	rec	F1
			0	88	86	87
			1	77	80	79

This Model performs well if there is matching sequence of words present in question pair and it does consider word similarity and the sequence, this helps when FNN(Word2Vec) or SBERT fails to predict duplicates (around 3% of data falls in this area).

Sentence-BERT

Sentence-BERT uses a pre-trained BERT model as an encoder and Siamese architecture to generate sentence embeddings, fed to forward layers for final prediction. we made some modifications on SBERT by replacing single SoftMax layer with a dense (128 nodes, fully-connected) layer and a Sigmoid layer for final prediction. So, by using pre-trained BERT base encoder, we have fine-tuned the model on Quora Question pairs dataset. This model is based on paper http://cs230.stanford.edu/projects_fall_2021/reports/102673633.pdf by Lynette Gao and Yujing Zhang.

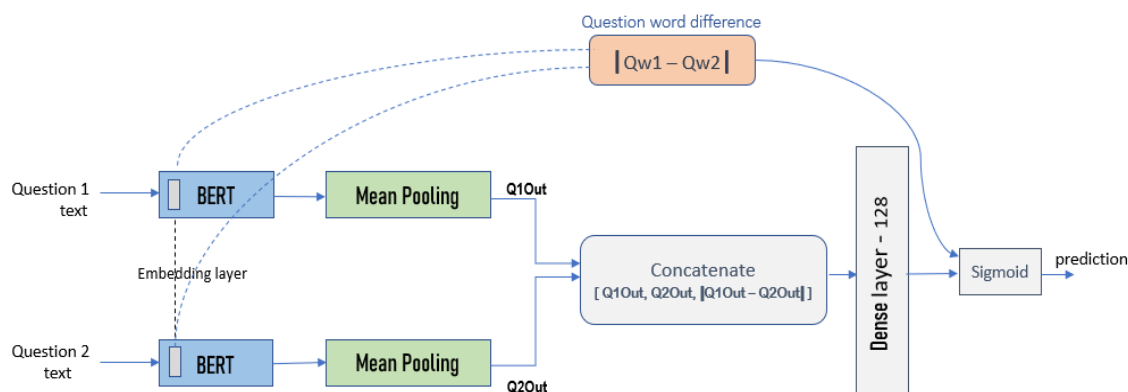
This model was developed with pre-trained BERT base encoder (12 - Encoding layers, 12 - Attention heads, vocabulary size = 30522, 110 million parameters) for Question's sentence encodings, mean pooling on the output from BERT encoder, concatenating pooled output of Q1, Q2 and $|Q1 - Q2|$ was fed to

Identification of Quora question pairs with the same intent (Project Report by Group-9, CDS-B3)

Dense layer, also skip connection of Question words encoding (e.g., when, who, how, why, where) difference to last classification (Sigmoid) layer as they play an important role in the semantic of a question sentence.

SBERT outperforms previous models in both training and test Accuracy, precision, recall and F1 score. Trained this model for 20 epochs using TPU distribution strategy, with Adam optimizer, learning rate $3e-5$, mini batch size = 32, dynamic masking and binary cross entropy as the loss function.

During evaluation of predictions, we experimented with various threshold for predictions. If the similarity/probability score exceeds the threshold, we consider the two sentences as duplicate. We got a better Accuracy and Recall rate for duplicates with threshold = 0.7. With increasing threshold from 0.7 to 0.8, we got better Precision at the cost of Recall for Duplicates.



Sentence BERT with a Dense layer and Sigmoid Classification layer with Question word difference skip connection

The below table lists SBERT Test-set Accuracies and F1-Scores on QQP test-set, with Similarity score\Probability threshold of 70%, 75% and 80% for identifying duplicates

Sentence BERT	Accuracy	F1 Score	Classification Report			
70% Similarity score threshold	85.57	81.00		pre	rec	F1
			0	90	87	88
			1	79	84	81
75% Similarity score threshold	85.69	81.03		pre	rec	F1
			0	90	87	89
			1	79	83	81
80% Similarity score threshold	85.78	81.02		pre	rec	F1
			0	90	88	89
			1	80	82	81

This Model outperforms other models and does well if there is semantic match between Question pair, can match complex semantics and provide wide range of similarity score (even though evaluation cut off is 0.70). Also due to strict semantic match, it does fail to predict duplicates if or, and clauses present in question pair, in this particular scenario LSTM fares better.

Identification of Quora question pairs
with the same intent
(Project Report by Group-9, CDS-B3)

Interpretation of Results

	FNN – WORD2VEC	Siamese LSTM (Word2Vec)	Sentence-BERT
Pros	- Performs well if similar words or synonyms present in Questions	- Performed well if 60% of sequence of words match between Question sentences	- Performed well with even complex Semantics between Question pair - Gives wide range of Similarity score covering semantic match of Question sentences
Cons	- Doesn't consider Sequence, Context or Semantics, named nouns - Miss classifies if < 80% of words are not similar or synonyms	- Doesn't consider Context and Semantics, named nouns - Miss classifies if 80% of sequence of words doesn't match	- Strict Semantic match results in missing some duplicates (like OR, AND clauses)

Consolidated test predictions of FNN(Word2Vec) -- LSTM -- SBERT models									
True Label	Not Duplicate								
	0	18960	1422	1502	753	996	414	709	789
Duplicate	1	629	505	413	1058	586	1057	825	9812
Predictions →		0-0-0	0-0-1	0-1-0	0-1-1	1-0-0	1-0-1	1-1-0	1-1-1
True Label	Not Duplicate								
	0	74.22	5.57	5.88	2.95	3.9	1.62	2.78	3.09
Duplicate	1	4.23	3.39	2.77	7.11	3.94	7.1	5.54	65.92
Predictions →		0-0-0	0-0-1	0-1-0	0-1-1	1-0-0	1-0-1	1-1-0	1-1-1
<p>Observations</p> <ul style="list-style-type: none"> ➤ Around 66% of the times all Three models correctly predicted duplicate Questions ➤ 74% of the times all Three models correctly predicted not duplicate Questions ➤ For predicting duplicates pairing SBERT with FNN and LSTM increases correct predictions by 10% ➤ All three models predicted duplicate as not duplicate 4.23% times. ➤ Stacking of these models will better Accuracy by 10% points. 									
<p>Note: Read predictions as <FNN>-<LSTM>-<SBERT> (E.g: 1-0-1)</p> <p>For E.g: Read 0-0-1 as only SBERT predicted as Duplicate, FNN and LSTM predicted as Not Duplicate</p>									

DEPLOYING MACHINE LEARNING MODELS

- Deploying ML models as API using FastAPI and ASGI server for production such as [uvicorn](#).
- Imported Pydantic package for data parsing and validation. We declare a class inherited from BaseModel.
- Loaded ML- XGBoost classification model pickle file.
- Defined a function for endpoint “/” with POST request method to predict.

Identification of Quora question pairs
with the same intent
(Project Report by Group-9, CDS-B3)

Not secure | b4b4-104-196-235-154.ngrok.io

Quora Question Based Classification

Enter Your Query

Question1 : how to open an account on quora site
Question2 : how to close an account on Quora site
Quora Question Result : **Not Duplicate**

EDGE CASE ANALYSIS AND AREAS FOR MODEL IMPROVEMENT

In order to improve the performance of our models, we compared the outcomes from two of our best performing models: XG Boost, and TFIDF weighted Word-to-Vec LSTM models. In all we observed that XG Boost outcomes matched with that of LSTM model outcomes for 82% of the cases. However, there were 1080 or 18% of the cases where these outcomes were conflicting. We took a sample of these cases and attempted to identify reasons for these. This could help us in inferring areas of improvement for our future iterations.

Below are the findings in brief. Detailed report with supporting examples for each of the observations is as per the attached excel.

1. Among the several features we engineered, one of the key impacting feature was cityblock distance. We observed that there were multiple question pairs for which cityblock distance was ZERO, but the question pair was identified as NOT DUPLICATE. On further analysing approx. 200 such pairs, we observe there were about 25 question pairs that were identical in text, but were incorrectly classified (Refer sheet named “Duplicates classified as Not”). For example, consider the below question pair marked as Not-Duplicate
 - a. What are the differences between Github and Bitbucket?
 - b. What are the differences between BitBucket and GitHub?
2. A set of questions were identified that were marked as duplicate, but intent of the question was otherwise. These questions were detected as duplicate by XG Boost model but otherwise by LSTM model. For example, consider the below question pair:
 - a. What is the exercise to remove belly fat for girls?
 - b. How do I remove belly fat?Though both these pairs are talking about removal of body fat, first question is more specific about belly fat for girls. However, this question is marked as duplicate. When we review more questions, we find that there are several instances, where such pairs are marked as not-duplicate. This again raises the question of gage R&R
3. 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. We observed that LSTM based networks performed better in identifying duplication in such scenarios. Example question pair as below:
 - a. What is the best gift for my boyfriend on Valentine's Day?
 - b. What are the Best Gifts for men on Valentine's day?
4. There was a set of question pairs, where 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

Identification of Quora question pairs
with the same intent
(Project Report by Group-9, CDS-B3)

duplicates. It is observed that much of these questions were identified correctly by LSTM model than with XG boost model. Example question pair as below:

- a. What is the specific heat capacity of steam?
 - b. What is specific heat capacity?
5. As a corollary of the above, there was a set of question pairs, where 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. Example question pair as below:
- a. What do you want to be remembered for?
 - b. What do you want to be remembered for when you die?
6. Another challenge that the models faced in prediction were the question pairs where numbers were involved. We observe randomness in prediction by both the models. There is need to further explore how such differences can be correctly classified. Below is a sample question pair classified as “Not Duplicate” as per the ground truth
- a. Is it safe to use castor oil at 36 weeks to induce labor?
 - b. Is it safe to use castor oil at 39 weeks to induce labor?

CONCLUSION

Following is the summary of outcomes:

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	FNN - Word2Vec	81.47	74.5
8		FNN-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

Best results were achieved through Sentence BERT modeling with 85.8% accuracy and 81% F1 Score.

Interpretation of Results

	FNN – WORD2VEC	Siamese LSTM (Word2Vec embeddings)	Sentence-BERT
Pros	Performs well if similar words or synonyms present in Questions	Performed well if 60% of sequence of words match between Question sentences	Performed well with even complex Semantics between Questions Gives wide range of Similarity score covering

Identification of Quora question pairs
with the same intent
(Project Report by Group-9, CDS-B3)

			semantic match of Question sentences
Cons	Doesn't consider Sequence, Context or Semantics, named nouns Miss classifies if < 80% of words are similar or synonyms	Doesn't consider Context and Semantics, named nouns Miss classifies if 80% of sequence of words match	Strict Semantic match results in missing some duplicates (like OR, AND clauses)

TEAM MEMBERS' NAMES

Sr	Name	Email	Phone No.
1	Sneh Lata	16.slata@gmail.com	9341649826
2	Sourav Bose	sourav2007bose@gmail.com	88610018865
3	Nitin Garg	Gargnitin759@gmail.com	7722027848
4	Goutham Reddy	goutham.n@gmail.com	9611122227