To Be Equal or Not To Be Equal

That is the question

Xavier Rubiés Cullell
Dànae Canillas Sánchez

QUORA DATASET

	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
5	5	11	12	Astrology: I am a Capricorn Sun Cap moon and c...	I'm a triple Capricorn (Sun, Moon and ascendan...	1
6	6	13	14	Should I buy tiago?	What keeps childern active and far from phone ...	0
7	7	15	16	How can I be a good geologist?	What should I do to be a great geologist?	1
8	8	17	18	When do you use シ instead of し?	When do you use "&" instead of "and"?	0
9	9	19	20	Motorola (company): Can I hack my Charter Moto...	How do I hack Motorola DCX3400 for free internet?	0
10	10	21	22	Method to find separation of slits using fresn...	What are some of the things technicians can te...	0
11	11	23	24	How do I read and find my YouTube comments?	How can I see all my Youtube comments?	1
12	12	25	26	What can make Physics easy to learn?	How can you make physics easy to learn?	1
13	13	27	28	What was your first sexual experience like?	What was your first sexual experience?	1
14	14	29	30	What are the laws to change your status from a...	What are the laws to change your status from a...	0
15	15	31	32	What would a Trump presidency mean for current...	How will a Trump presidency affect the student...	1
16	16	33	34	What does manipulation mean?	What does manipulation means?	1
17	17	35	36	Why do girls want to be friends with the guy t...	How do guys feel after rejecting a girl?	0
18	18	37	38	Why are so many Quora users posting questions ...	Why do people ask Quora questions which can be...	1
19	19	39	40	Which is the best digital marketing institutio...	Which is the best digital marketing institute ...	0
20	20	41	42	Why do rockets look white?	Why are rockets and boosters painted white?	1

404290

Total of question pairs for training

36.92%

Duplicate pairs

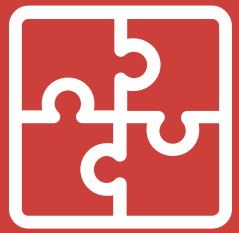
MAIN GOAL

Predict whether a pair of questions are duplicates or not

- Add useless load to the website system
- Spend more time finding the best answer to their question
- Make writers feel they need to answer multiple versions of the same question.

WHY ?

SUMMARY



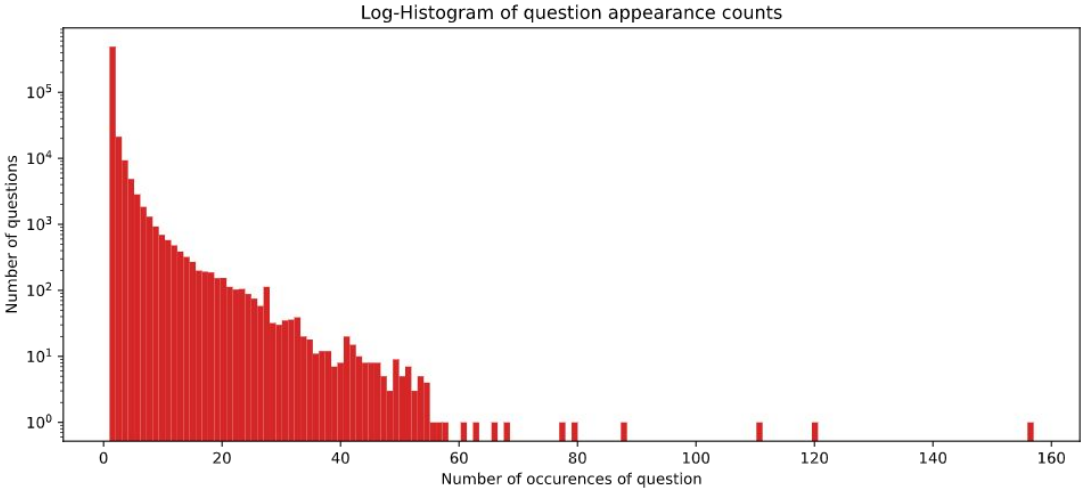
Data Analysis

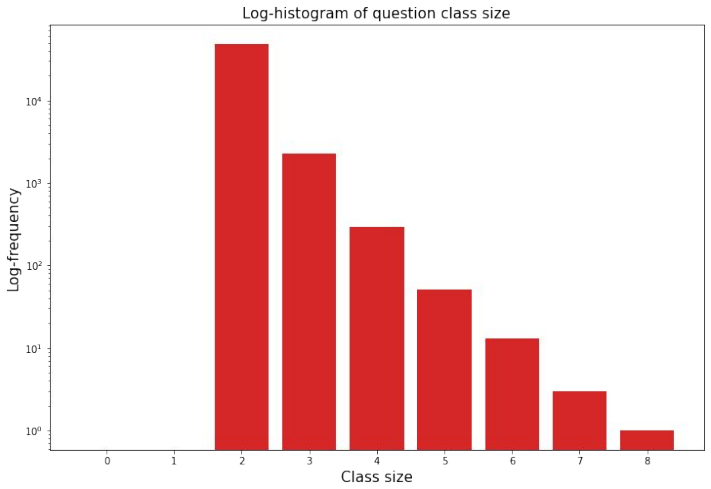
Structure

Results

Next Steps

DATA ANALYSIS

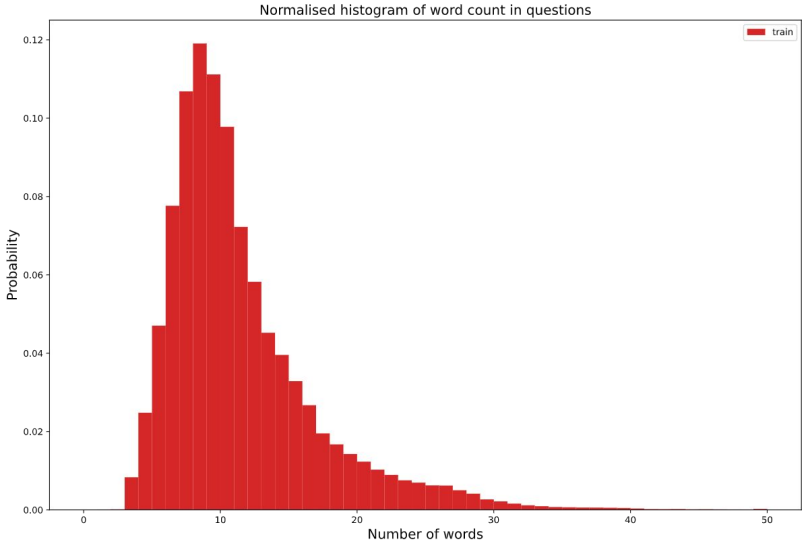




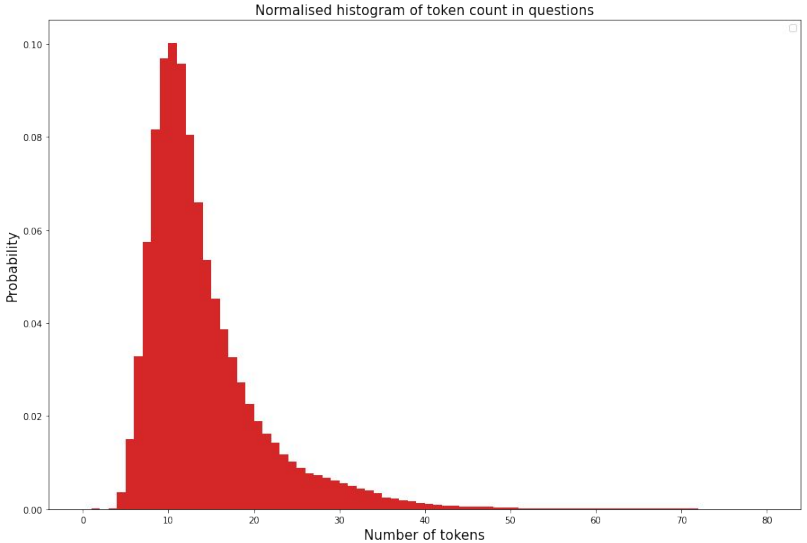
537933 different sentences

51071 different classes

Length distribution



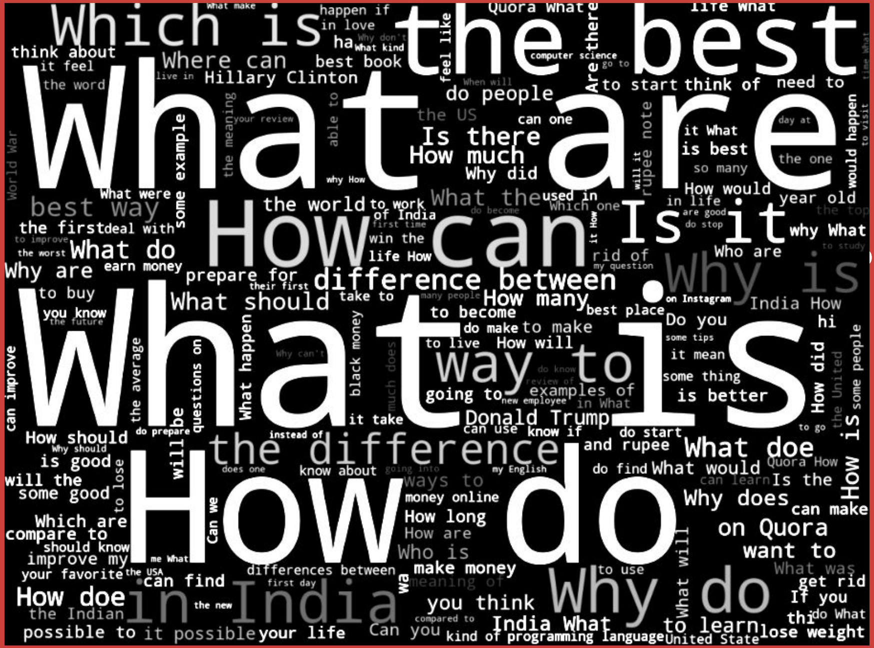
Number of words



Number of tokens

WORD FREQUENCY

(The , 372316),
(What , 292887),
(is , 216832),
(I , 212601),
(a , 208748),
(to , 204709),
(How , 202001),
(in , 191594),
(of , 159425),
(do , 145554)



STRUCTURE

Fine-Tuning: use pre-trained models

Quicker Development

Pre-trained models weights encode a lot of information.

Takes much less time to train our fine-tuned model.

Less Data

Fine-tuning our task on a much smaller dataset than would be required in a model that is built from scratch.

Better Results

State of the art results with minimal task-specific adjustments

MODEL CLASSES

	MODEL	Tokenizer	Pretrained config
'albert':	AlbertForSequenceClassification,	AlbertTokenizer,	'albert-large-v2'
'bart':	BartForSequenceClassification,	BartTokenizer,	'bart-large'
'bert':	BertForSequenceClassification,	BertTokenizer,	'bert-base-uncased'
'camembert':	CamembertForSequenceClassification,	CamembertTokenizer,	'camembert-base'
'distilbert':	DistilBertForSequenceClassification,	DistilBertTokenizer,	'distilbert-base-uncased'
'flaubert':	FlaubertForSequenceClassification,	FlaubertTokenizer,	'flaubert-base-uncased'
'roberta':	RobertaForSequenceClassification,	RobertaTokenizer,	'roberta-base'
'xlm':	XLMForSequenceClassification,	XLMTokenizer,	'xlm-mlm-en-2048'
'xlm_roberta':	XLMRobertaForSequenceClassification,	XLMRobertaTokenizer,	'xlm-roberta-base'
'xlnet':	XLNetForSequenceClassification,	XLNetTokenizer,	'xlnet-base-cased'

Input Formatting

Question A	Question B	Question A	Question B
Did Ben Affleck shine more than Christian Bale as Batman?	No fanboys please, but who was the true batman, Christian Bale or Ben Affleck?	How does 3D printing work ?	How do 3D printing work ?

Tokenize the sequence

['did', 'ben', 'af', '##fle', '##ck', 'shine', 'more', 'than', 'christian', 'bal', '##e', 'as', 'batman', '?', 'no', 'fan', '##boys', 'please', ',', 'but', 'who', 'was', 'the', 'true', 'batman', ',', 'christian', 'bal', '##e', 'or', 'ben', 'af', '##fle', '##ck', '?']	['how', 'does', '3d', 'printing', 'work', '?', 'how', 'do', '3d', 'printing', 'work','?']

Truncate the input

['did', 'ben', 'af', '##fle', '##ck', 'shine', 'more', 'than', 'christian', 'bal', '##e', 'no', 'fan', '##boys', 'please', ',', 'but', 'who', 'was', 'the', 'true', 'batman', ',', 'christian', 'bal', '##e']	

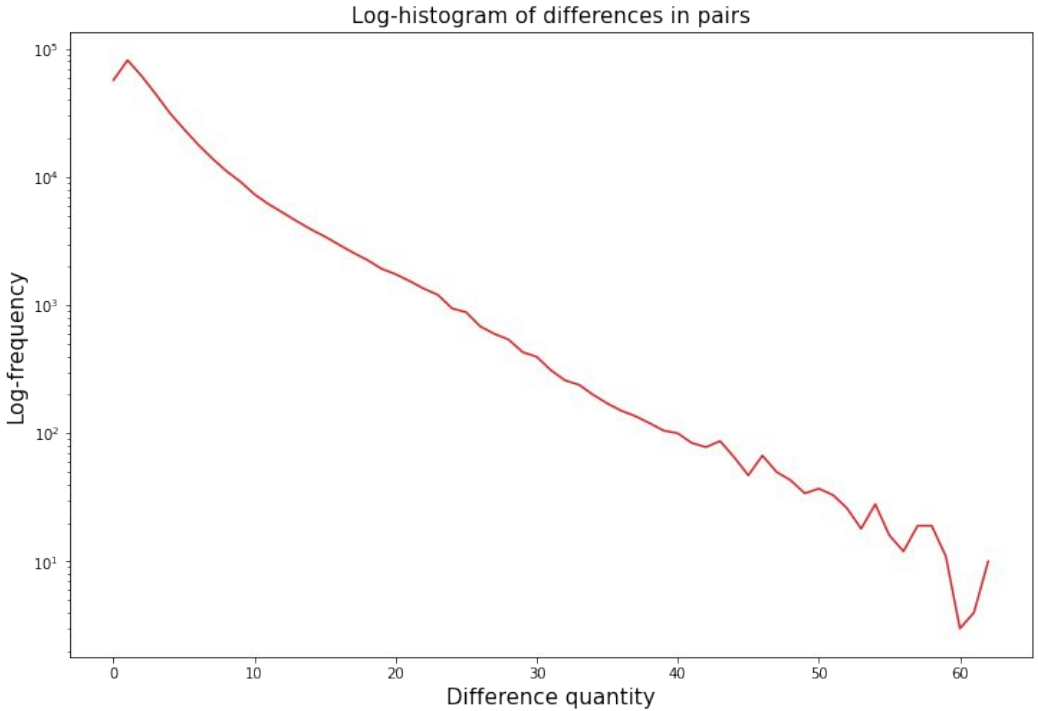
Add [CLS] and [SEP] tokens

[[CLS], 'did', 'ben', 'af', '##fle', '##ck', 'shine', 'more', 'than', 'christian', 'bal', '##e', [SEP], 'no', 'fan', '##boys', 'please', ',', 'but', 'who', 'was', 'the', 'true', 'batman', ',', 'christian', 'bal', '##e', [SEP]]	[[CLS], 'how', 'does', '3d', 'printing', 'work', '?', [SEP] 'how', 'do', '3d', 'printing', 'work','?', [SEP]]

Padding the input

	[[CLS], 'how', 'does', '3d', 'printing', 'work', '?', [SEP] 'how', 'do', '3d', 'printing', 'work','?', [SEP], [PAD], [PAD], [PAD], [PAD], [PAD], [PAD], [PAD], [PAD], [PAD], [PAD], [PAD], [PAD]]

Truncate function motivation



Input tokens

101, 2129, 2515, 7605, 8021, 2147, 1029, 102, 2129, 2079, 7605, 8021, 2147, 1029, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0

Attention Mask

1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0

Segment tokens

0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0

Hyperparameters

'max_seq_length': 70,

'batch_size': 16,

'epochs': 2,

'learning_rate': 4e-5,

'epsilon': 1e-8,

'num_training_steps': 1000,

'num_warmup_steps': 10,

'max_grad_norm': 1.0

RESULTS

	XLM	XLM_Roberta	BERT	DistilBERT
Training Accuracy	61.32%	85.39%	89.99%	89.88%
Training Loss	0.6751	0.28116	0.5764	0.2968

Troubles...

COMPUTATIONAL RESOURCES

NEXT STEPS

Our next goals

EQUIVALENCES STUDY

$A \equiv B$

$B \equiv C$ $\Rightarrow$ $A \equiv C$

CLUSTERING

Similar sentences clustering

PREDICTION

Duplicate searching

OTHER ARCHITECTURES

1. No pre-train
2. Siamese network

Thanks for your attention

If there is any doubt, now is the time to ask!