**Links used:**

[**https://qa.fastforwardlabs.com/no%20answer/null%20threshold/bert/distilbert/exact%20match/f1/robust%20predictions/2020/06/09/Evaluating\_BERT\_on\_SQuAD.html#Metrics-for-QA**](https://qa.fastforwardlabs.com/no%20answer/null%20threshold/bert/distilbert/exact%20match/f1/robust%20predictions/2020/06/09/Evaluating_BERT_on_SQuAD.html#Metrics-for-QA)

**https://towardsdatascience.com/the-quick-guide-to-squad-cae08047ebee**

**EVALUATING Q&A MODELS**

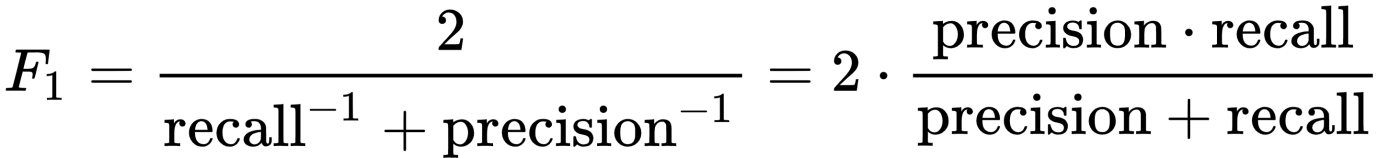
**Common Metrics for QA**

There are two dominant metrics used by many question answering datasets, including SQuAD: exact match (EM) and F1 score. These scores are computed on individual question+answer pairs. When multiple correct answers are possible for a given question, the maximum score over all possible correct answers is computed. Overall EM and F1 scores are computed for a model by averaging over the individual example scores.

### Exact Match

This metric is as simple as it sounds. For each question+answer pair, if the characters of the model's prediction exactly match the characters of (one of) the True Answer(s), EM = 1, otherwise EM = 0. This is a strict all-or-nothing metric; being off by a single character results in a score of 0. When assessing against a negative example, if the model predicts any text at all, it automatically receives a 0 for that example.

### F1

F1 score is a common metric for classification problems, and widely used in QA. It is appropriate when we care equally about precision and recall. In this case, it's computed over the individual words in the prediction against those in the True Answer. The number of shared words between the prediction and the truth is the basis of the F1 score: precision is the ratio of the number of shared words to the total number of words in the prediction, and recall is the ratio of the number of shared words to the total number of words in the ground truth. 

**Dataset Used**

The Stanford Question Answering Dataset (SQuAD) is a set of question and answer pairs that present a strong challenge for NLP models. Due to its size (100,000+ questions), its difficulty due to the model only has access to a single passage and the fact that its answers are more complex and thus require more-intensive reasoning, SQuAD is an excellent dataset to train NLP models on.

**SQuAD 1.1,** the previous version of the SQuAD dataset, contains 100,000+ question-answer pairs on 500+ articles. **SQuAD2.0** combines the 100,000 questions in SQuAD1.1 with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering.

We have decided to choose a model that has been pre-trained on SQuAD2.0 for improved performance.

**Our Evaluation**

We shortlisted 5 well performing pre-trained QA models from Hugging Face and analysed them using their f1 and EM values on the SQuAD2.0 dataset, which are as follows:

|  |  |  |
| --- | --- | --- |
| Model Name | F1 | EM |
| distilbert-base-cased-distilled-squad | 86.996 | 79.600 |
| deepset/roberta-base-squad2 | 82.950 | 79.931 |
| deepset/minilm-uncased-squad2 | 79.548 | 76.192 |
| bert-large-uncased-whole-word-masking-finetuned-squad | 83.876 | 80.885 |
| deepset/bert-base-cased-squad2 | 74.671 | 71.152 |

We also ran all models on a sample document provided by our client, asked the models questions related to our requirements, tabulated the results and asked our peers to rate the sensibility of the answers from 0-5 for each of the answers. The results are as follows: