loxz digital

# MLR: Semantic Scoring Methodology

**2022 Q2 Methodology Report**

**By Yiming Zhang**
**Lead Data Scientist, Loxz Digital**

Q2 2022
Methodology Report

loxz digital

## Abstract

In order to accurately evaluate students' answers from the ML Aptitude sub-score of our MLR diagnostic assessment we created an expert set of answers for each open-ended question called "Answer Corpus." The approach is to use a BERT tokenizer to tokenize the student's answer and and compare with the true answer, then feed those tokens into a pre-trained BERT model. Once the model is built and deployed we capture the dense vector embeddings from the last hidden state. Then we mean pool only the masked embeddings to get the vector representations of answers for calculating the cosine similarity score between the two. In our testing, we found that this semantic scoring method proved to be the most accurate way to assess a students ML Aptitude.

## Motivation

Loxz required additional granularity for the ML Aptitude portion of the diagnostic assessment. In order to have a better understanding of students' ML aptitude, only asking for multiple choice questions did not provide enough specificity for evaluating the aptitude in some prerequisite/foundations of ML. Providing students with the most accurate scoring provides an additional layer of certainly to the subscore. Adding text-input open-ended questions will provide additional insight to students for this purpose and answers could be used to assess the potency of a students career trajectory. Open-Ended questions/answers give us a deeper perspective of students overall performance in this sub-scoring category.

## Shortcoming of Classic Semantic Matrix

If we collect students' answers and aim to quantify them to get a "score" then how do we measure the quality of the input answer? For this kind of QA system, there are two classic metrics: Exact Match and F1 score.

**Exact Match:** the percentage of overlap between a student's answer and the true answer.
**F1 score :**= 2 * precision * recall / (precision + recall). Precision is the ratio of shared words in both answers to words in the candidate's answer and Recall is the ratio of shared words in both answers to words in the true answer. The two classic metrics are actually NOT that effective for answer quality evaluation. Think of a scenario that when asking about what is "2+2" then the student inputs an answer "4" which is absolutely correct.

However, what if our textbook answer is **"The answer is four"?** Then both metrics would give a 0 to this student who answered this question correctly. Another example is, asking about a name or an entity with a different alias. Looking for text exact match or f1 score often performed pretty badly in such scenarios with a lot of contextual understanding. In the subsequent pages, we examine these three different answers and provide you with reasons why we chose to add in a semantic scoring script to this sub-score.

## Methodology

The goal is to measure the semantic similarity between the survey input answer and our Answer Corpus. The answer corpus is provided by our Lead Data Scientist and four other members of the Loxz internal team. The approach is to use transfer learning with a pre-trained BERT tokenizer and a BERT model. We first tokenize the student answer and then each answer in the Answer Corpus then feed those tokens into the BERT model. Then we capture the dense vector embeddings from the last hidden state. To get the result, we use mean pooling on the masked embeddings to get the vector representations of answers for calculating the cosine similarity score between the student answer and answers in the Answer Corpus. We then take the average of all similarities between each answer in Answer Corpus and the student answer. The final similarity score will be averaged among all student-corpus answer pairs and normalized to 0 to 100.

## Experiment and Comparison

To demonstrate the power of the semantic scoring method, we use a real test case regarding Type I and Type II errors:

**Question:** What are Type I and Type II errors?

Textbook Answer: A Type I error is a false positive (claiming something has happened when it hasn't), and a Type II error is a false negative (claiming nothing has happened when it actually has). Here we take three answers from some students and use Exact Match (EM), F1-score, Semantic Similarity, and a human label (judged by a panel of five people).

**Student 1:** A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population.

**Student 2**: In statistical hypothesis testing, a type I error is the mistaken rejection of an actually true null hypothesis (also known as a "false positive" finding or conclusion; example: "an innocent person is convicted"), while a type II error is the mistaken acceptance of an actually false null hypothesis (also known as a "false negative" finding or conclusion; example: "a guilty person is not convicted").

**Student 3:** A Type I error is a false positive, and a Type II error is a false negative.

| | EM | F1-score | Semantic | Human |
|---|---|---|---|---|
| **Textbook** | 100.00 | 100.00 | **100.00** | 100.00 |
| **Student 1** | 14.63 | 36.36 | **81.38** | 86.6 |
| **Student 2** | 19.05 | 37.04 | **81.12** | 93 |
| **Student 3** | 93.75 | 68.97 | **89.17** | 96 |

We can see that the semantic scoring is way more consistent than EM and F1-score. EM and F1-score perform poorly when the IoU (intersection over union) of a given answer and the textbook answer is small though human highly rates those answers.

I

f you harken back to when your teacher were to grade the essay portion of your exam, she would have to measure your answer manually against the text book answer code, and provide you with an overall score, heavily weighted on the way the essay portion of the exam was answered. In this case, we're having our Bert Model score this section of the diagnostic assessment and include this input in the overall ML Aptitude portion of the exam.

## Constraints and Future Consideration

Though our semantic scoring method outperforms classic metrics, there is huge potential for further refinements and improvements. First, we used transfer learning with SentenceBert. Given only the recent history of our diagnostic assessment, the lack of a large trainable dataset leads to the potential inability to fine-tune the answer rating tasks. Given enough student answer data, one improvement could be adding a classification dense output layer with a binary Softmax activation function to judge whether an answer is semantically identical to the textbook answer (class 1) or not (class 0). Then the probability of that answer belonging to class 1 can be taken as a score when multiplied by 100. Also, a text ranking-based approach can also be applied if we consider student answers as different queries and textbook answers as retrievable documents.

loxz digital

## Summary

The ML Aptitude sub-score is weighted differently in the overall MLR score that students receive and is considered an important part of the diagnostic assessment. As more students rely heavily on the scoring and relay that scoring to potential interviewers or HR, it will be determined that a students ML Aptitude is accurately measured and provides confidence to both student an employer.

By transfer learning from SentenceBert, we created our own set of answers for each open-ended question called "Answer Corpus" to evaluate students' answers within a subsection (ML Aptitude) of our diagnostic assessment. The approach is to use a BERT tokenizer to tokenize the student's answer and the correct answer, and then feed those tokens into a pre-trained BERT model. The returned cosine similarity of answer embeddings will serve as the Semantic Score. This approach surpasses Exact Match and F1-score in terms of consistency especially when the word overlap is low.