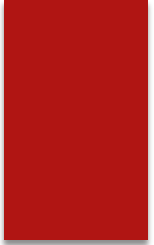# AUTO GRADING OF ANSWERS FOR TEXT-BASED OPEN-ENDED QUESTIONS USING NATURAL LANGUAGE PROCESSING

## BURAK *KESKİN*

AKDENİZ ÜNİVERSİTESİ
"Eğitimde ve Bilimde Öncü Üniversite"

# Objective

In text-based exams with open-ended questions, it takes a long time to read and evaluate the answers.

In this study, the objective is to read and successfully evaluate these exams in computer environment using natural language processing methods.

# Overview

What is CBA?

What is NLP?

Literature Review.
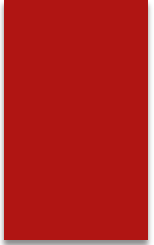
Which methods are suitable for sentence analysis in NLP?

What are the results achieved in this study?

# Computer Based Assessment

Computer Based Assessment(CBA) is the technique that is used to assess the students with the help of computer environment.

Usually used for multiple choice exams but rarely used for open-ended text based exams.
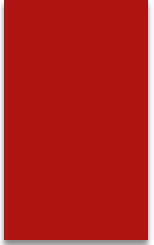
It is not easy to use in exams with math content.

On the other hand, it can be used for text-only exams.

Worldwide known exams like TOEFL, GMAT, GRE are examples of CBA.

# Advantages of CBA

Immediate score reporting: According to the study Daniels and Gierl 2017, anxiety of the students are dramatically reduced when they learn their grades immediately after the exam is finished.

Time efficiency: Compared to reading answer sheets manually, CBA saves huge amount of time by quickly assessing the sheets.

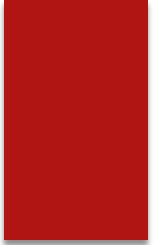Storage space: CBA systems do not require large physical space in the real world like paper based assessment.

Feedback: It is easier to analyze the results and give feedback to provide students the room for improvement.

# Disadvantages of CBA

Cost: CBA requires computers, smart phones, tablets etc. to be able to make and assess the exam. Which is considered expensive in both hardware and deployment of the required software compared to the paper based exams.

User Interaction: Although it is common knowledge to interact with tech devices in present, there may be bugs, issues, incompatibilities in the system that limits the performance.

Availability: Each student should have access to the CBA environment and in rare cases this may become a problem.

# Natural Language Processing

Natural language processing(NLP) is a field of computer science to understand the interaction between human and computers.

Content extraction, text summarization, question answering, next sentence prediction, word-sentence similarity detection are possible by the libraries provided with NLP.

# Literature Review

Objective is accurate evaluation of sentence similarity.

TF-IDF, T5 and BERT variations are the most popular methods to be used for this objective.

In order to measure sentence similarity, three aspects have been used in different studies.

These aspects are Word based, Structure Based, Vector Based.

# Aspects of Sentence Similarity

Word Based: Compare the word similarity between the student's answer and the solution.

Structure Based: Compare the structure of the student's answer to solution.

Vector Based: Transform both student's answer and the solution into vectors and compare the similarity of the vectors.

# Word based

Word based assessment process is not optimal because the solutions of the questions asked in the exam differs in number of words.

An irrelevant answer containing the required words may have high grade which is not the desired result.

'This is a pencil.' and 'This is.' results in %89 similarity when word based methods are used.

TF-IDF is one of the popular techniques used in word based assessment.

# Structure based

Structure based assessment process is also not optimal because there are many grammar errors, misspelling, use of other languages.

There are models that pre-trained for multilingual assessment of the sentences but performance is poor.

Results change due to the order of words in the answers which may lead to false evaluations.

T5(Text to text transformer model (Haller 2020)) is a popular technique used in structure based assessment.

# VECTOR BASED

Vector based assessment is found more suitable for assessment technique. Process of the vector based assessment can be summarized as:

Take the sentences and transform them into high dimensional vectors that can be used for semantic similarity, text classification and other natural language tasks.

Compare the vectors of student's answer and the solution to find similarity between them.

# Transformers

- Transformers are the deep learning models that proposed in the paper (Vaswani et al. 2017) developed to solve the problems encountered in previous approaches like LSTM.

- An attention mechanism is used to process all input at once which makes parallelization possible. Popular models are:

- BERT(Bidirectional Transformer Encoder Representation)

- DistilBert

- Roberta-Bert

- T5
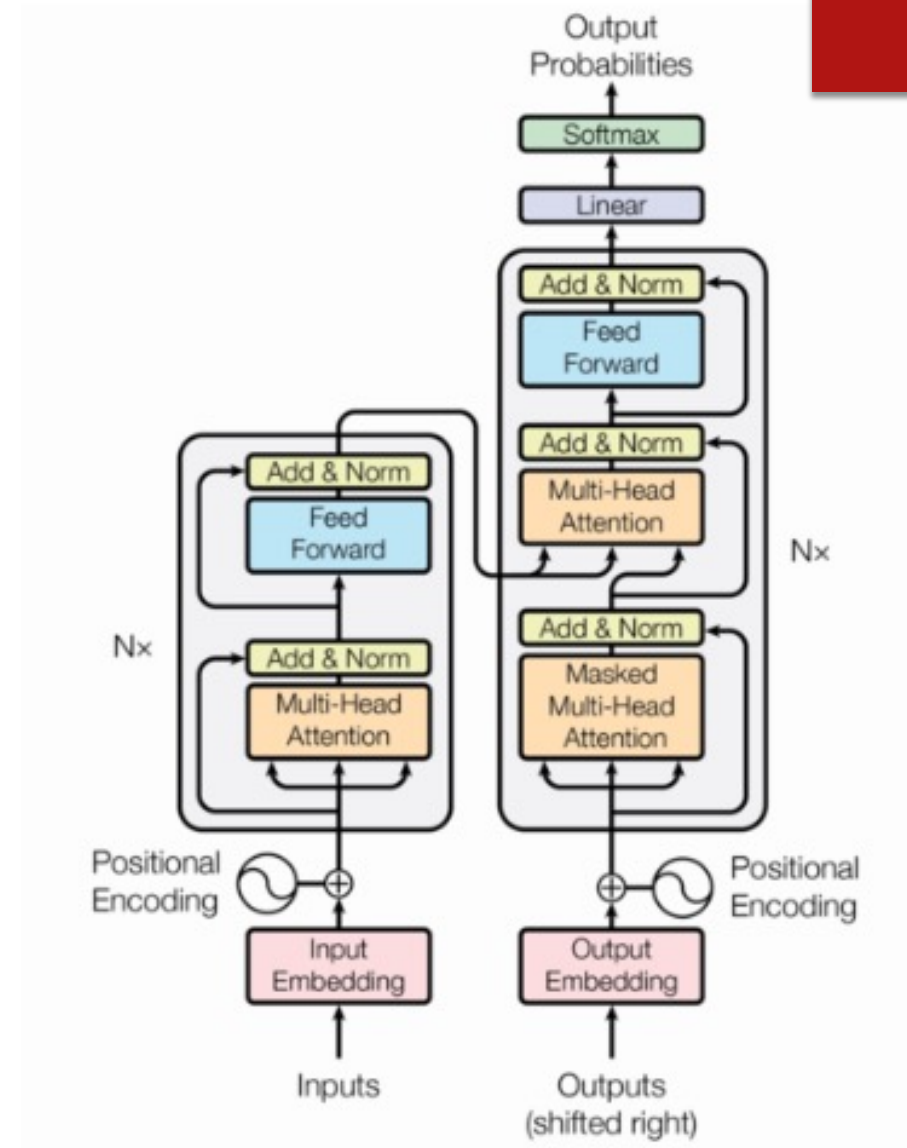
# T5

- ► Aim of this model is to have an all-in-one model for all the natural language processing tasks.

- ► It has state of art performance in generic datasets like SciEntsBank (Haller 2020).
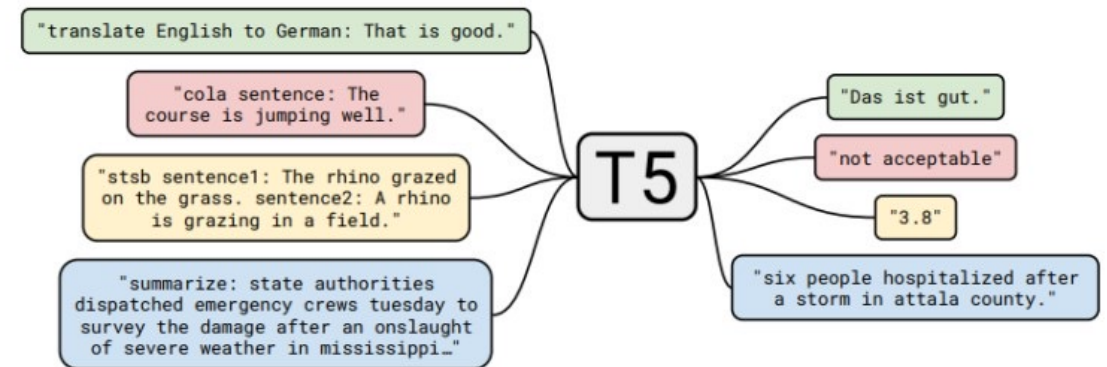
# BERT

- BERT was created by Jacob Devlin and his colleagues in 2018.

- Unlike previous models, BERT use bidirectional encoding mechanism hence provides better results compared to the previous models in well known datasets as seen in the figure.
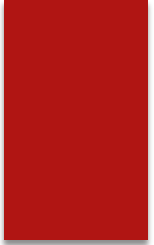
| Tasks | Dev Set | | | | |
| --- | --- | --- | --- | --- | --- |
| | MNLI-m (Acc) | QNLI (Acc) | MRPC (Acc) | SST-2 (Acc) | SQuAD (F1) |
| BERT$_{BASE}$ | 84.4 | 88.4 | 86.7 | 92.7 | 88.5 |
| No NSP | 83.9 | 84.9 | 86.5 | 92.6 | 87.9 |
| LTR & No NSP | 82.1 | 84.3 | 77.5 | 92.1 | 77.8 |
| + BiLSTM | 82.1 | 84.1 | 75.7 | 91.6 | 84.9 |

# Methodology

## Pre-processing

In order to avoid errors pre-processing the answers was necessary.

Stopwords have been removed, NaN values have been fixed as empty.

# Process

Ask the question to ChatGPT and get the answers.

Get the students' answers from the excel file.

Create dataframes for each question to be processed.

Vectorize both the correct answers and the students' answer as a tensor list.

From the tensor list, compare the similarity of the answer and the solution and evaluate the score using cosine similarity.

# Cosine Similarity

▶ Vector based assessment is selected in the scope of this study.

▶ Cosine similarity is the technique to calculate the similarity between vectors.
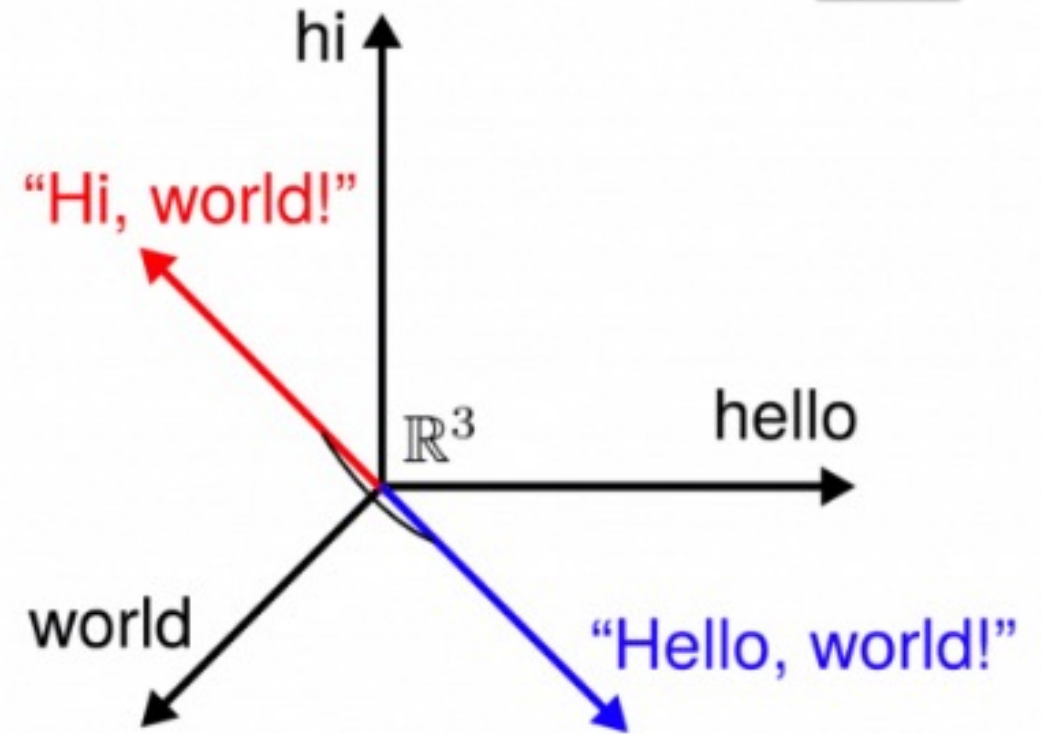


Cosine Similarity

# Dataset

For this study, dataset from the Natural Science course was used which includes the answers of 133 students for five different questions.

There are questions with with open-ended answers in the dataset.

For the questions with open-ended answers, grades given were lower compared to the original grades.

# Dataset

Example of a question: What are the contributions of the Roman Empire to the legal system?

Answer given by ChatGPT was : 'The Roman Empire's contributions to the legal system include the development of the Twelve Tables, which laid the foundation for Roman law, influencing modern legal systems with concepts of democracy, division of power, legal rights in trade and conflict, and the basis for evidence and proof in judicial proceedings.'.

As seen in the figure, instructor gave scores depending on the content of the answer.

| | StudentID | Answer | Score |
|---|---|---|---|
| 1 | | | |
| 2 | 20190808014 | advanced modern legal system civil engineering military engineer | 17 |
| 3 | 20190808021 | main battlefield surgeries bound books 12 tables julian calendar a | 18 |
| 4 | 20190808035 | roads highways geometry military engineering science changes life | 11 |
| 5 | 20200808003 | | 0 |
| 6 | 20200808008 | achieved civil army engineering republic government system func | 17 |
| 7 | 20200808015 | contribute law world rules strict | 0 |
| 8 | 20200808020 | great war government didn t choose king choose consuls senator a | 18 |
| 9 | 20200808026 | good engineering predictable improving invention practical infor | 11 |
| 10 | 20200808033 | - theater pop ancient greece like birthplace theater focused mytho | 17 |

```python
query_q2 = "The Roman Empire's contributions to the legal system include the development of the Twelve Tables, which laid the foundation for Roman law, influencing modern legal systems with concepts of democracy, division of power, legal rights in trade and conflict, and the basis for evidence and proof in judicial proceedings."

# ST modeli ile vektörleri hesapla
roberta_vectors2 = roberta_model.encode(answers_q2 + [query_q2], convert_to_tensor=True)
roberta_query_vector2 = roberta_vectors2[-1]
roberta_cosine_similarities2 = util.pytorch_cos_sim(roberta_vectors2[:-1], roberta_query_vector2)

# ST modeli 2 ile vektörleri hesapla
bert_vectors2 = bert_model.encode(answers_q2 + [query_q2], convert_to_tensor=True)
bert_query_vector2 = bert_vectors2[-1]
bert_cosine_similarities2 = util.pytorch_cos_sim(bert_vectors2[:-1], bert_query_vector2)

# DistilBERT modeli ile vektörleri hesapla
distilbert_vectors2 = distilbert_model.encode(answers_q2 + [query_q2], convert_to_tensor=True)
distilbert_query_vector2 = distilbert_vectors2[-1]
distilbert_cosine_similarities2 = util.pytorch_cos_sim(distilbert_vectors2[:-1], distilbert_query_vector2)

df_q2['Roberta Score'] = np.round(roberta_cosine_similarities2.numpy().flatten() * 20).astype(int)  # ST skoru, tamsayı olarak
df_q2['Bert Score'] = np.round(bert_cosine_similarities2.numpy().flatten() * 20).astype(int)  # ST2 skoru, tamsayı olarak
df_q2['DistilBert Score'] = np.round(distilbert_cosine_similarities2.numpy().flatten() * 20).astype(int)  # DistilBERT skoru, tamsayı olarak
```

# Code Sample

# Results and discussion

BERT-Base, RoBERTa-Large and DistilBert-Base and T5-XL models have been used.

Results for each model are explained in the following slides.

Original grades given by the instructor are used for comparison of the models.

In all models, results were divided into five parts as Very Low, Low, Medium, High and Very High Success.

# Original Scores

- Original distribution of the grades is shown in the figure.

- In the original scores 28 students graded as Very Low, 36 students were graded as Low, 42 students were graded as Medium, 24 students were graded as High and 3 students were graded as Very High.
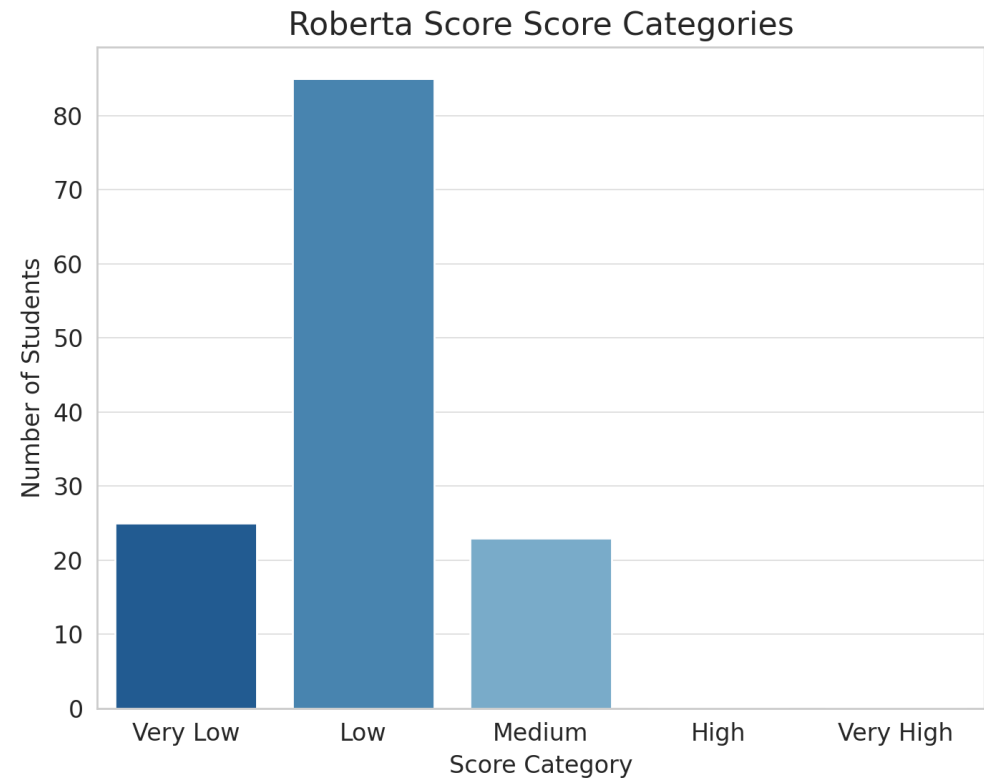


Original Score Score Categories
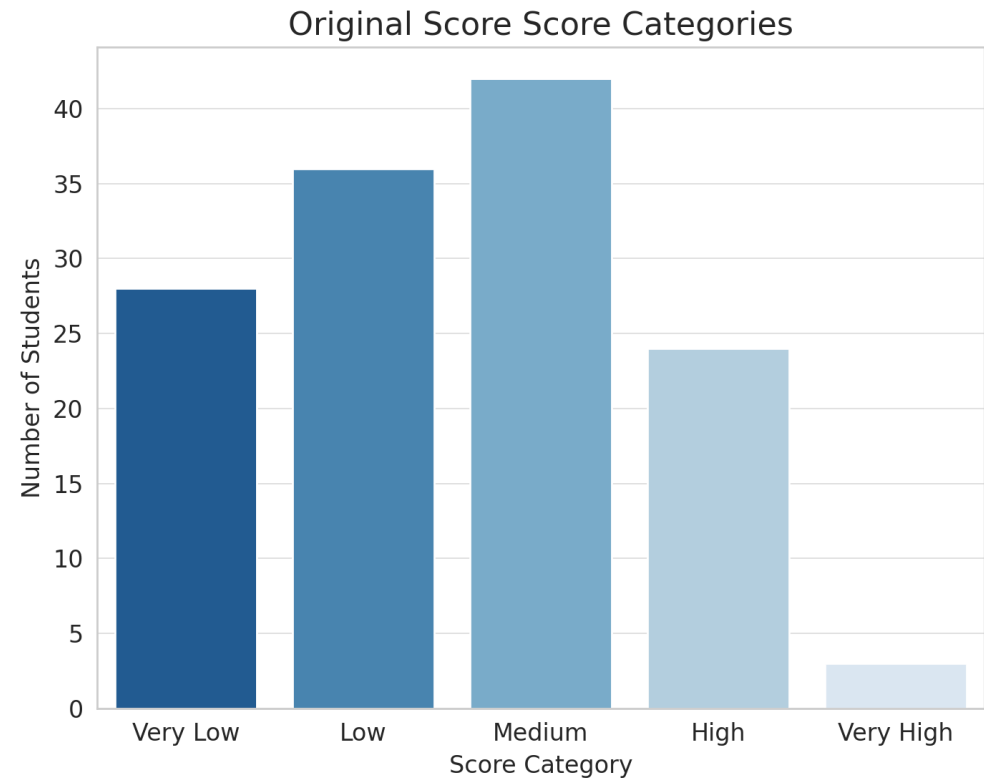
# Bert-Base

- Distribution of the grades assessed with BERT-Base model is shown in the figure.

- In this model, 11 students were graded as Very Low, 70 students were graded as Low and 52 students were graded as Medium.



Bert Score Score Categories
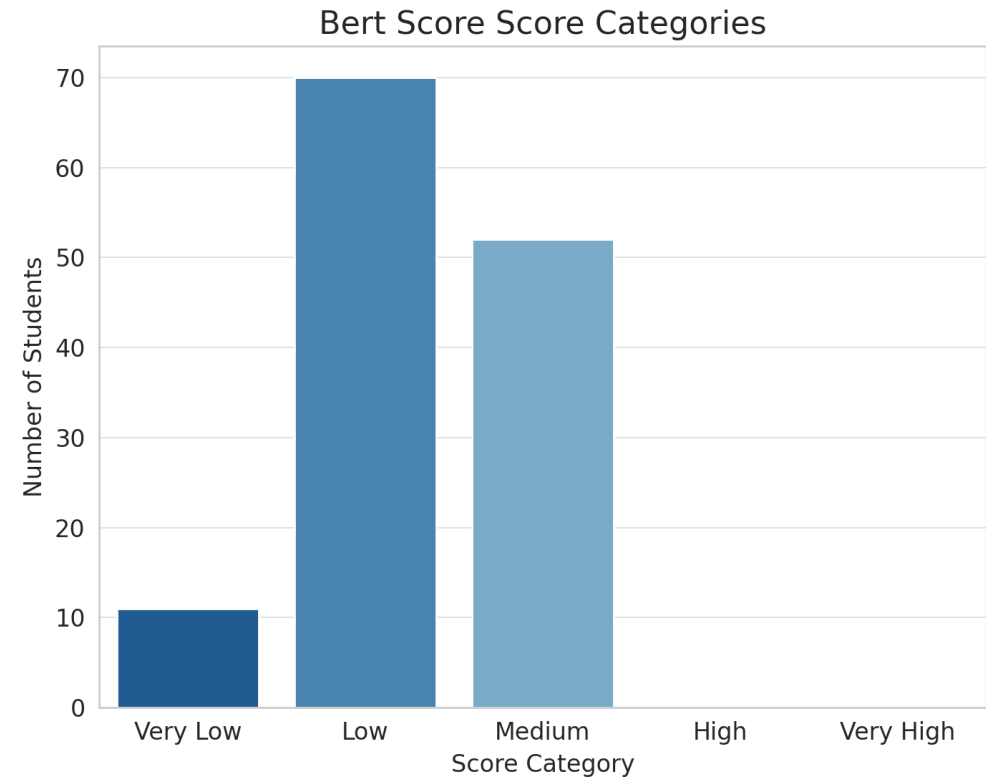
# Roberta-large
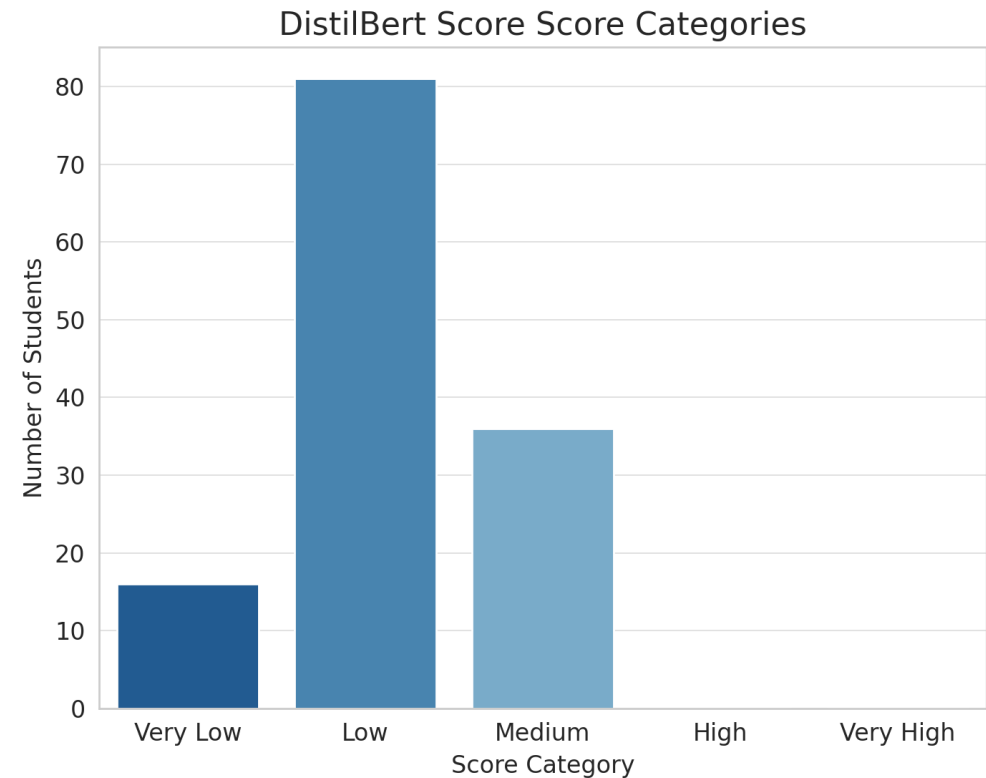
- Distribution of the grades assessed with RoBERTa-Large model is shown in the figure.

- In this model, 25 students were graded as Very Low, 85 students were graded as Low and 23 students were graded as Medium.
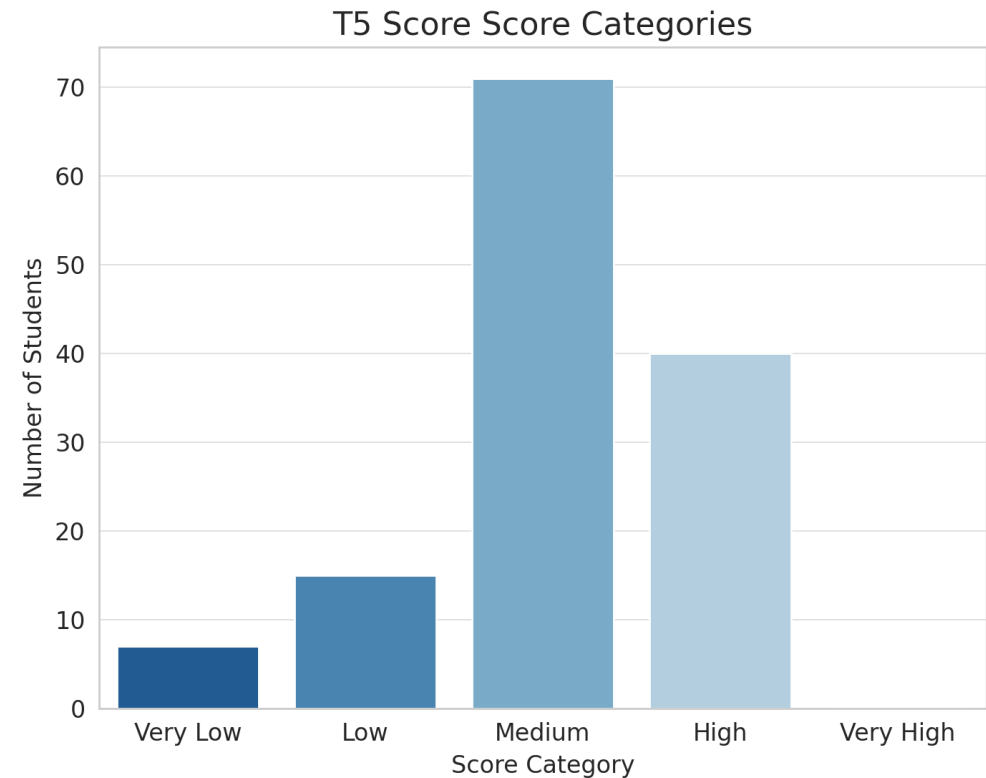


Roberta Score Score Categories

# DistilBert

▶ Distribution of the grades assessed with DistilBert model is shown in the figure.

▶ In this model, 16 students were graded as Very Low, 81 students were graded as Low and 36 students were graded as Medium.



DistilBert Score Score Categories

# T5-XL

- Distribution of the grades assessed with T5-XL model is shown in the figure.

- In the original scores 7 students graded as Very Low, 15 students were graded as Low, 71 students were graded as Medium, 40 students were graded as High.

# Model Performance Based on Correct Classifications

▸ RoBERTa excels in assessing lower-performance students, indicated by its better scores in "Low" and "Very Low" categories.

▸ BERT shows versatility with strong performance in "Low" and "Medium" and fair in "Very Low," suitable for a broad range of students.

▸ DistilBERT is reliable for "Low" to "Medium" levels, suggesting a focus on lower to mid-performing students.

▸ T5 stands out in "High" and "Medium" evaluations, making it ideal for higher-achieving students.

| Model | Very Low | Low | Medium | High |
|---|---|---|---|---|
| Roberta | 18 | 28 | 5 | 0 |
| Bert | 10 | 30 | 22 | 0 |
| DistilBert | 14 | 32 | 12 | 0 |
| T5 | 7 | 3 | 28 | 20 |

# Conclusion

In literature, evaluation of university level exams with custom datasets were missing. Each study was made on the popular large sized datasets.

Each model used in this study evaluated the students in different categories from original scores given by the instructor.

In this study, it is shown that using vector based models for evaluating the similarity of sentences in custom datasets with small size can be considered as a viable option.

No model outperformed other models used in this study in terms of comparison with the original grades but it is observed that depending on the expected score range of the students, different models can be used in similar small sized datasets.

# THANK YOU FOR LISTENING