

**Deep Learning**

***Test2Achieve: AI Based Evaluation System***

***Team number 4***

*Adithya Vinod*

*Bharat Sharma*

*Shanmukha Sai Penumatsa*

**School of Graduate Professional Studies**

Data Analytics

DAAN 570 – Deep Learning

*Spring Semester, 2023*

**Document Control**

**Work carried out by:**

|  |  |  |
| --- | --- | --- |
| **Name** | **Email Address** | **Exhaustive list of Tasks** |
| Shanmukha Sai Penumatsa | sxp6048@psu.edu | 1. Identifying the corpus data 2. Research on identifying existing deep learning models for answer generation 3. Model building 4. Writing the Project report |
| Adithya Vinod | avv5502@psu.edu | 1. Understanding how to preprocess the corpus data 2. Researching on different types of deep learning models to implement answer generation and improve performance 3. Model building 4. Worked on the Project presentation |
| Bharat Sharma | bqs5791@psu.edu |  |

**Revision Sheet**

|  |  |
| --- | --- |
| **Date** | **Revision Description** |
| 21st Feb 2023 | Building the first draft |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

**TABLE OF CONTENTS**

[1 Introduction 4](#_Toc428978113)

[2 Problem Statement 4](#_Toc553221296)

[3 Challenges 5](#_Toc735393042)

[3.1 RELATED WORKS 5](#_Toc1043993699)

[3.2 IMPORTANCE AND IMPACTS 7](#_Toc2134644489)

[4 Data Collection 8](#_Toc1441433202)

[5 Data Preprocessing 9](#_Toc329565754)

[6 Methodology 10](#_Toc772958717)

[7 Results and Interpretation 11](#_Toc822760409)

[8 Discussion of Results 13](#_Toc1986348722)

[9 References 14](#_Toc197158969)

# Introduction

From primary, secondary, and tertiary education to vocational certification, all forms of education and training require optimized and measurable assessments to measure the outcomes of knowledge imparted to learners. Assessments have long been employed for the evaluation process. Well-designed assessments focus on targeted areas with utmost accuracy and encourage learners to perform better. Before the digital era, evaluators used traditional approach-based tests.

However, with the internet and technology seeping into different aspects of life and computing gadgets becoming a necessity we have moved towards the more innovative approach of computer-assisted assessment. Unlike its predecessor (pen-and-paper testing), a computer-based test is not a resource-intensive evaluation. On the contrary, it is technologically advanced and has eliminated the inconvenience that incurs due to rudimentary testing approaches.

# Problem Statement

The amount of time a teacher spends grading student work can vary widely depending on factors such as the size of the class, the number of assignments, the complexity of the assignments, and the teacher's own grading practices. On average, a teacher may spend anywhere from a few minutes to several hours grading each student's work, with the average being around 30-60 minutes per student per assignment. However, in certain cases, the grading process can take much longer, particularly for more complex or open-ended assignments that require a more detailed review and evaluation.

Additionally, for online classes, where the volume of the student's submission is larger, the time spent on grading can be more. In addition, typing allows students to produce more in response to questions and eases the editing process, making answers potentially completer and more accurate. Computerizing assessment allows instructors to administer quizzes and tests that are flexible, easy to customize, and quick and efficient to grade.

Course-management software such as Blackboard, Moodle, and Canvas are widely used in the United States and can be used as a data corpus for an AI-based examination system. This system can help instructors administer assessments such as quizzes and tests that have descriptive and open-ended answers along with the ability to grade these assessments. This can be beneficial for educators as it allows them to focus more on teaching and less on grading as well as give fair and consistent assessment results for students.

An AI-based descriptive examination system can be used to automatically score answers for several reasons:

1. It can save time and resources by eliminating the need for manual grading.
2. It can provide more objective and consistent scoring, as the AI model can be trained to look for specific characteristics in the answers that indicate a certain level of understanding or mastery of the subject matter.
3. It can help to reduce the potential for human bias in the scoring process.
4. It can handle large volumes of data and reduce the error rate of manual grading.

Overall, an AI-based descriptive examination system can provide more efficient, accurate, and fair evaluation of student performance.

# Challenges

Despite the abundance of research on the topic, there aren’t real applications in usage that help in answer generation as well as for grading the answer provided. This leaves room for the question regarding if this is something that is achievable. One of the major obstacles in developing deep learning solutions is the scarcity of data, the dataset that we will be working on has not been updated for the past 10 years and is the only public data source that is publicly accessible, with many studies conducted on private data, making it challenging to reproduce and improve the results. Another possible challenge is that deep learning models are considered black boxes, making it difficult to understand and interpret their decision-making processes. Since we are trying to implement the project in the Education sector, an incorrect implementation of the project could be harmful to students and could negatively impact their learning experiences. The use of deep learning in the education sector raises ethical questions about the use of technology in education, the potential for bias in algorithms, and the impact on student privacy. One more big challenge is limitations in computation power available, since the deep learning models require huge amounts of data processing and high GPU resources, the model building often results in longer executions.

## RELATED WORKS

Significant research has been conducted on using Deep Learning and NLP to understand and evaluate descriptive text based on different similarity criteria.

There are numerous recent research papers that explore the use of deep learning to evaluate descriptive text based on similarity measures. A few research papers that we would be referring include:

* "A Deep Learning Approach for Similarity-Based Text Evaluation" (2017)
* "Evaluating Text Similarity using Deep Learning Models" (2018)
* "Bidirectional LSTM for Similarity-Based Text Evaluation" (2019)
* "Deep Text Similarity Learning for Evaluating Descriptive Text" (2020)
* "Transformer Networks for Similarity-Based Text Evaluation" (2021)

These research papers explore the use of deep learning to evaluate descriptive text based on similarity measures, and these studies have concluded that deep learning approaches can effectively be used for this task. Different techniques such as bidirectional LSTM and transformer networks have been proposed to improve the performance of deep learning models in evaluating text similarity, and these models have demonstrated their ability to learn from text data and generate accurate similarity scores.

There are research papers that explore the use of NLP to generate descriptive answers as well. We have referred to the below 5:

* "Generating Descriptive Answers Using NLP and Knowledge Bases" (2017)
* "An NLP-Based Approach for Generating Descriptive Answers to Questions" (2018)
* "Generating Descriptive Answers Using NLP and Semantic Role Labeling" (2019)
* "A Deep Learning Model for Generating Descriptive Answers to Questions" (2020)
* "Transformers for Generating Descriptive Answers in NLP" (2021)

These studies explore the use of NLP to generate descriptive answers to questions. They present various NLP-based approaches, including knowledge bases, semantic role labeling, deep learning models, and transformer networks. The conclusion of these studies is that NLP can effectively be used to generate descriptive answers, with different techniques leading to varying levels of performance and accuracy.

SQuAD (Stanford Question Answering Dataset) is a popular benchmark dataset for answer generation tasks in NLP. The dataset consists of a set of Wikipedia articles and a corresponding set of questions and answers based on those articles.

There are two versions of the SQuAD dataset: SQuAD 1.1 and SQuAD 2.0. SQuAD 1.1 contains 100,000+ question-answer pairs for 500+ articles, while SQuAD 2.0 contains additional unanswerable questions, making it a more challenging dataset for answer generation and the above research papers have used one or the other version of SQuAD dataset to train and evaluate their deep learning models for answer generation.

The above-mentioned papers don’t discuss and compare the performance of different models that we have implemented. We have also implemented a fully functional front end which is the practical implementation of our work.

## IMPORTANCE AND IMPACTS

The amount of time a teacher spends grading student work can vary widely, but on average it takes around 30-60 minutes per student per assignment. However, the process can take much longer for more complex or open-ended assignments. For online classes, the volume of student submissions is often larger, making the grading process even more time-consuming. Course-management software such as Blackboard, Moodle, and Canvas can be used as a data corpus for an AI-based examination system. This system can help instructors administer assessments such as quizzes and tests that have descriptive and open-ended answers along with the ability to grade these assessments. An AI-based descriptive examination system can save time and resources by eliminating the need for manual grading, provide more objective and consistent scoring, reduce human bias in the scoring process and handle large volume of data and reduce the error rate of manual grading. Overall, an AI-based descriptive examination system can provide more efficient, accurate, and fair evaluation of student performance.

IMPACTS**:**

Automated question answer generation and automatic grading systems have significant social, economic, business, and scientific impacts. Some of the key impacts of these systems are:

* Education: Automated question answer generation and automatic grading systems will transform the education sector by providing students with personalized and immediate feedback. These systems will make it easier for teachers to grade assignments, quizzes, and exams, allowing them to focus on other aspects of teaching.
* Cost savings: Automated grading systems will save educational institutions significant amounts of money by reducing the need for human graders. This will result in lower tuition fees for students and higher profits for institutions.
* Efficiency: Automated grading systems will grade assignments much faster than human graders, resulting in faster feedback for students. This can help students learn more efficiently and effectively.
* Fairness: Automated grading systems will help eliminate subjective grading practices and reduce bias in the grading process, ensuring that all students are graded fairly and accurately.
* Job market: The development and implementation of automated grading systems will create job opportunities in the field of artificial intelligence and machine learning. These systems require skilled developers and engineers to build and maintain them.
* Scientific research: Automated question answer generation and automatic grading systems have the potential to improve scientific research by making it easier to analyze large amounts of data and extract insights. This will lead to breakthroughs in fields such as medicine, engineering, and social sciences.
* Business: Automated grading systems can be used in the hiring process to screen job candidates. This can save businesses time and money by reducing the need for manual resume screening.

# Data Collection

ANSWER GENERATION:

We have used SQUAD dataset for building the model. The Stanford Question Answering Dataset (SQuAD) is a large-scale reading comprehension dataset that was created by researchers at Stanford University. It was released in 2016 and has since become a widely used benchmark for testing and improving machine reading comprehension systems. The SQuAD dataset consists of over 100,000 question-answer pairs, with each question designed to be answerable from a given context paragraph. The context paragraphs were sourced from over 500 Wikipedia articles, covering a wide range of topics. The dataset is split into two parts: a training set and a development set. The training set contains approximately 80,000 question-answer pairs, while the development set contains approximately 10,000 question-answer pairs. And the remaining 10,000 are hidden test set. Each question-answer pair in the SQuAD dataset consists of the following components:

* Context paragraph: This is a piece of text from a Wikipedia article that provides the context for the question. The context paragraph may be several sentences or several paragraphs long.
* Question: This is a question that is designed to be answerable from the context paragraph.
* Answer: This is the answer to the question, which is a span of text from the context paragraph. This also contains the starting position of answer in paragraph

In the SQuAD dataset, each question is designed to be answerable from a given context paragraph, and the task of the deep learning model is to generate the correct answer to the question based on the information in the context paragraph. This process involves understanding the context paragraph, identifying the relevant information, and generating a natural language response that accurately answers the question. As the goal of our project is to generate the answer from given, this dataset clearly aligns with the objective of the project.

Data link: https://rajpurkar.github.io/SQuAD-explorer/dataset/train-v2.0.json

https://rajpurkar.github.io/SQuAD-explorer/dataset/dev-v2.0.json

AUTOMATED GRADING SYSTEM:

We used the Kaggle dataset published by Hewlett foundation. There are ten data sets. Each of the data sets was generated from a single prompt. Selected responses have an average length of 50 words per response. Some of the essays are dependent upon source information and others are not. All responses were written by students primarily in Grade 10. All responses were hand graded and were double scored. Each of the eight data sets has its own unique characteristics. Each essay is approximately 150 to 550 words in length.​ Lowest amount of training data is 1,190 essays, randomly selected from a total of 1,982. The dataset contains following information:

* Id: A unique identifier for each individual student essay.
* EssaySet: 1-10, an id for each set of essays.
* Score1: The human rater's score for the answer. This is the final score for the answer and the score that you are trying to predict.
* Score2: A second human rater's score for the answer. This is provided as a measure of reliability but had no bearing on the score the essay received.
* EssayText: The ascii text of a student's response.

Since above data contains information on answer provided by students and human based scoring, we can build a neural network model which will embed the content and score the answers.

Data link - <https://www.kaggle.com/competitions/asap-sas/data>

# Data Preprocessing

ANSWER GENERATION:

•Finding the end position character is a very important input for all the answer generation models as it needs both start and end position characters of the answer. This needs to be found and stored.

•There are cases where the SQuAD answers are missing one or two characters from the real answer in the passage and can be challenging, as it can affect the performance of these answer generation models that rely on precise tokenization. One approach to address this issue is to modify the passage to match the given answer by "cutting" it by one or two characters by using a compatible tokenizer.

AUTOMATED GRADING SYSTEM:

•Dropping Null Values: Removing the missing values from the dataset.

•Text cleaning: Removing irrelevant information like punctuation, numbers, special characters, and stop words from the text.

•Tokenization: Breaking down the text into smaller parts, called tokens, which can be analyzed individually.

•Stop word removal: Removing common words like "the", "and" and "a" that are not important for analysis.

•Stemming and Lemmatization: Reducing words to their root form to group together different inflected forms of a word, such as "run," "running," and "ran."

# Methodology

ANSWER GENERATION:

The whole process is divided into 4 phases:

1. Data preparation: Downloaded the SQuAD (Stanford Question Answering Dataset) dataset. Preprocessed the dataset by tokenizing the text using tokenizers for each model and applying it onto train and validation sets. Also identified the starting and ending position of each answer since these are used as inputs for chosen pretrained models.

2. Model selection: Decided which models to use for the task, in this case BERT, ROBERTA, ELECTRA and ALBERT. These models have been shown to perform well on natural language processing tasks and are widely used in research and industry.

3. Model fine-tuning: This involves training the models on the train set, using the questions as inputs and the answers as outputs, and adjusting the model's parameters to minimize the loss. The loss function measures the difference between the predicted and actual answers.

4. Model evaluation: Evaluated the performance of the models on the validation set by measuring their loss and F1 score. The F1 score is the harmonic mean of precision and recall and is commonly used to evaluate question-answering models.

AUTOMATED GRADING SYSTEM:

The whole process is divided into two different phases:

1.Preprocessing and cleaning: After we downloaded the dataset, we tokenized the dataset into 2 parts:

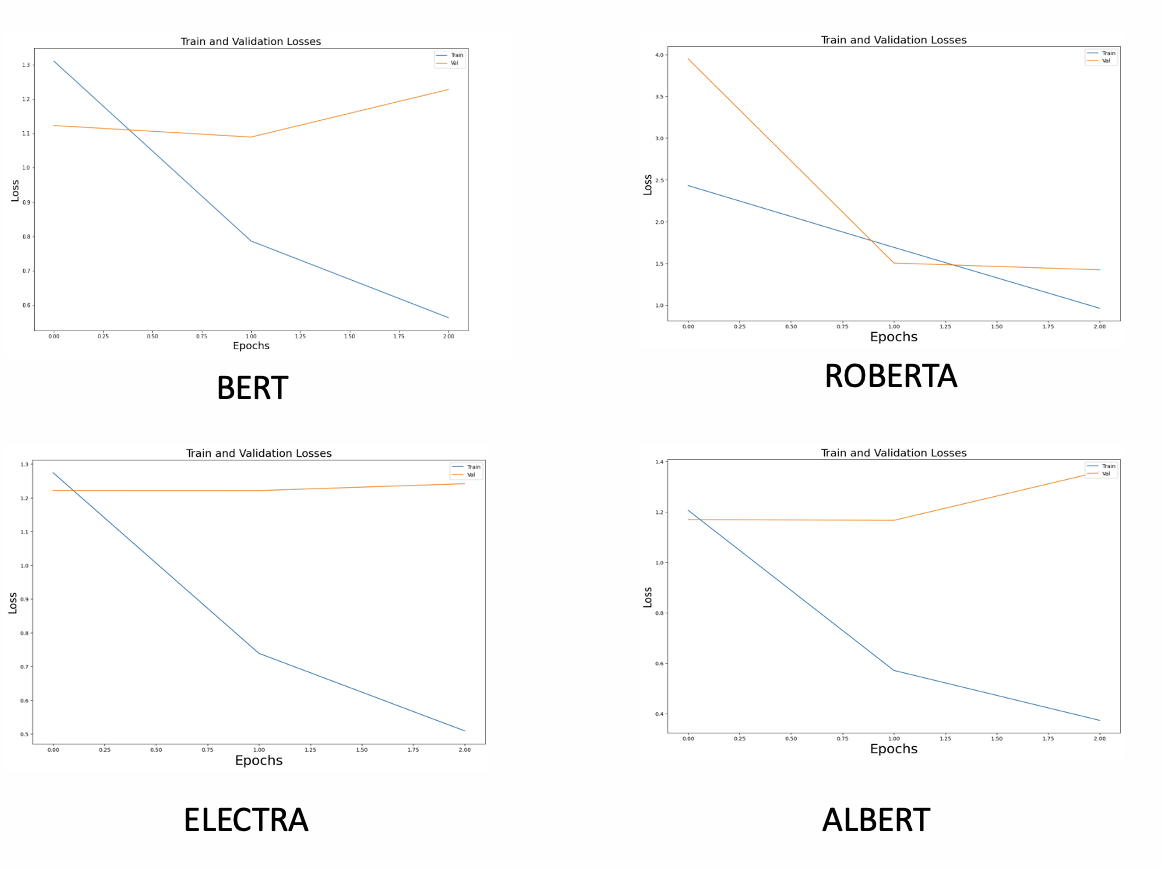
* Sentence based tokenizer: To remove punctuation and preserve the context.
* Wordlist based tokenizer: To clean the data (removing stop words, converting to lowercase)

2.Vectorize: After initial pre-processing we created 3 separate functions:

* Function makeFeatureVec creates an average feature vector for a given list of words, using a trained word2vec model and the number of features as input. It returns a feature vector as a NumPy array.
* Function getAvgFeatureVecs generates word vectors for a list of essays. It takes a list of essays, a trained word2vec model, and the number of features to use as input. It calls the makeFeatureVec function on each essay and returns a matrix of feature vectors.
* Function makeFeatureVec (updated) creates an average feature vector for a given string of words, using a trained word2vec model and the number of features as input. It returns a feature vector as a NumPy array.

# Results and Interpretation

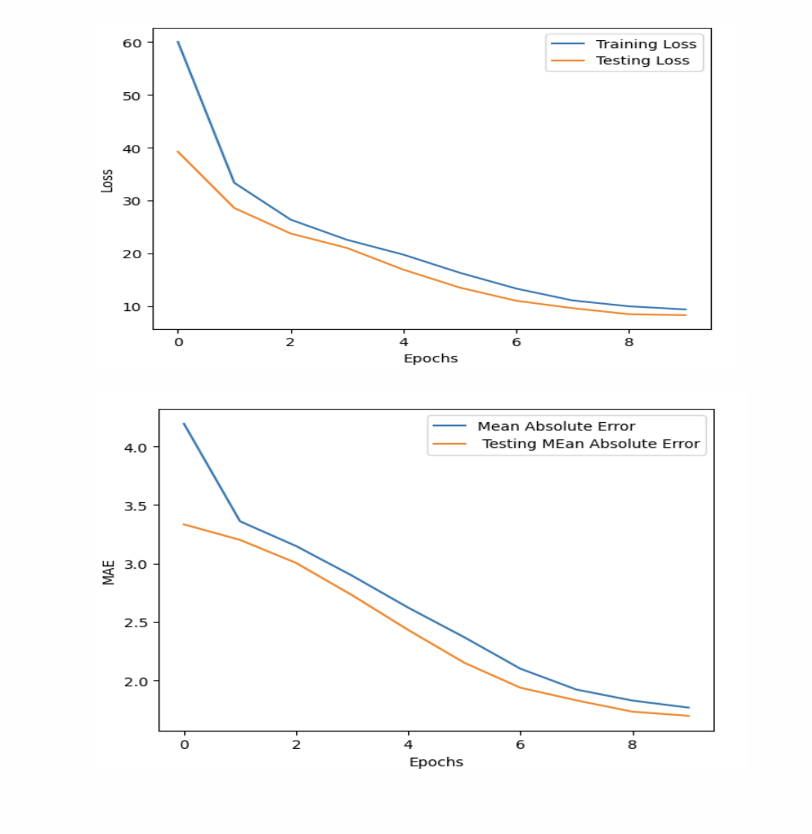
ANSWER GENERATION:



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **BERT** | **ROBERTA** | **ALBERTA** | **ELECTRA** |
| F1 Score | 0.387 | 0.83 | 0.34 | 0.44 |

* ROBERTA performs the best when compared to the rest as we can see the validation loss decrease along with the training loss.
* Increasing the Learning rate reduces the performance of the model.
* Data needs to be trained for a larger number of epochs as we see a scope of improvement in the performance of these models.
* We can see that ROBERTA outperforms the other models. This can be attributed to the fact that ROBERTA is trained on a larger corpus of text and has a slightly larger architecture with additional layers and larger hidden sizes.

AUTOMATED GRADING SYSTEM:



* From the loss graphs, we can infer that there are no overfit issues since the loss for both test and train drops.
* The drop in MAE with increase in epochs indicate the improvement of the model performance.
* The model performance almost reaches saturation near 10 epochs so there is only a slight scope of improvement.

# Discussion of Results

* We can see a decreasing trend in terms of training loss but the same can't be said regarding the validation loss while using pre-trained models.
* The pre-trained models were trained on just 3 epochs because of computational constraints, and we could understand the model better if we were to run it for a larger number of epochs.
* Performance of the pre-trained model could be improved by improving the quantity and quality of the dataset.
* Future work in terms of answer generation would involve running the model on a more performing computer to try and train it better.
* Future work in terms of automated grading would involve the use of SBERT in conjunction with Word2vec and spaCy.

# References

* Dataset 1 - <https://rajpurkar.github.io/SQuAD-explorer/>
* Dataset 2 - <https://www.kaggle.com/competitions/asap-sas/data>
* <https://huggingface.co/docs/transformers/tasks/question_answering>
* <https://github.com/alexaapo/BERT-based-pretrained-model-using-SQuAD-2.0-dataset>
* [r-net.pdf (microsoft.com)](https://www.microsoft.com/en-us/research/wp-content/uploads/2017/05/r-net.pdf)