Question Paper Generator

Submitted in partial fulfilment of the requirements

of the degree of

Bachelor of Engineering in

Artificial Intelligence and Data Science

by

Shubham Hadawle

Pranav Kotkar

Muhumaad Faayez

Mayur Pimpude

Under the guidance of

Mr. Amit Singh



**Department of Artificial Intelligence and Data Science**

**Vivekanand Education Society’s Institute of Technology**

**2022-2023**

**Department of Artificial Intelligence and Data Science**

**CERTIFICATE**

This is to certify that **Mr** Subham Hadavle**,Mr** Pranav Kotkar, **Mr** Muhummad Faayez,**Mr** Mayur Pimpudeof Second Year of Artificial Intelligence and Data Science studying under the University of Mumbai have satisfactorily presented the Mini Project entitled **Question Paper Question** as a part of the MINI-PROJECT for Semester-V under the guidance of **Mr** **Amit Singh** in the year 2022-2023.

Date:28/10/2022

**(Name and sign) (Name and sign)**

**Head of Department Supervisor/Guide**

**Department of Artificial Intelligence and Data Science**

**DECLARATION**

We, **Mr** Subham Hadavle, **Mr** Pranav Kotkar, **Mr** Muhummad Faayez, **Mr** Mayur Pimpude from , declare that this project represents our ideas in our own words without plagiarism and wherever others' ideas or words have been included, we have adequately cited and referenced the original sources.

We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our project work.

We declare that we have maintained a minimum 75% attendance, as per the University of Mumbai norms.

We understand that any violation of the above will be cause for disciplinary action by the Institute.

Yours Faithfully

1. \_\_\_\_\_\_\_\_\_\_\_

2. \_\_\_\_\_\_\_\_\_\_\_

3. \_\_\_\_\_\_\_\_\_\_\_

4. \_\_\_\_\_\_\_\_\_\_\_

**Acknowledgement**

We are thankful to Our HOD and Vice Principal Dr. Vijaylakshmi Madam for giving us a wonderful Topic : Question Paper Generator (NLP) with A.I. in it . As being the first batch of A.I. and D.S. in our college we are getting a great experience through A.I.

Our Mentor Mr. Amit Singh has also given us very good support in our project development . Sir also helped us throughout the semester and guided us through the right way and encouragement throughout the work. Without her guidance our project would not have taken shape.

We sincerely thank all the faculty members in the Department of Artificial Intelligence and Data Science for their advice and counseling from time to time .

Table of Contents

[**Abstract**](#_heading=h.px6ebou7cy8q) **7**

[Table Of Figure](#_heading=h.xt1liu0aolo) **9**

[**Chapter 1**](#_heading=h.bvtvxloiqfk1) **10**

[**Introduction**](#_heading=h.iw4tqnxletp6) **10**

[1.1 Introduction:](#_heading=h.byvlog4ybgq4) 10

[1.2 Problem Statement :](#_heading=h.fdsm6pkjatfw) 10

[1.3 Objective :](#_heading=h.v3izfit9ui1v) 10

[1.4 Scope :](#_heading=h.bxbzrhe3b8sw) 10

[**Chapter 2**](#_heading=h.hawp6345q5iy) **11**

[**Literature Survey**](#_heading=h.rljzjgn2s3m4) **11**

[2.1 Literature/Techniques :](#_heading=h.wd0fjftgkqu8) 11

[Transformer](#_heading=h.nn1igl1373jo) 12

[T5 Transformer](#_heading=h.vtlbmbtifina) 12

[2.2 Papers and Findings :](#_heading=h.bm995kpqtuw6) 13

[**Chapter 3**](#_heading=h.dxhfbb4pobd7) **14**

[**Data Set**](#_heading=h.ukac67qxxvw3) **14**

[3.1 Description of data set :](#_heading=h.moeroxtnklo9) 14

[3.2 Data collection methodology :](#_heading=h.7lw70a3fd0sb) 14

[3.3 Exploratory data analysis :](#_heading=h.i72rj8dcgbo2) 15

[3.4 Feature extraction :](#_heading=h.p7tjpvobdtz4) 15

[**Chapter 4**](#_heading=h.urja4wymxhfn) **16**

[**Analysis and Design**](#_heading=h.aq975v6bm9p3) **16**

[4.1 Analysis of the system :](#_heading=h.7ywkx1dj8aow) 16

[4.2 Proposed Solutions :](#_heading=h.1du5bp19rt55) 16

[4.3 Prototype design of the proposed system :](#_heading=h.6ir77p8rjepi) 17

[**Chapter 5**](#_heading=h.mr6i3x3oc2kc) **19**

[Results and Discussion](#_heading=h.j0kvth20o138) **19**

[**Chapter 6**](#_heading=h.epzyb3v1hgq) **20**

[Conclusion and Future Work](#_heading=h.aj79virwhupo) **20**

[**Reference :**](#_heading=h.467k5sv7gutk) **21**

# Abstract

In this project we are going to implement a question paper generator using a transformer in machine learning. The main motto is to take any kind of file from the user and extract the text from it and get an appropriate output with a set of questions from the given file.

The main ideology is to help parents and students understand the kind of questions that can be framed from a given paragraph which help them to improve their score. This could be also used by many teachers to help them conduct a short quiz with minimum efforts on preparation of questions.

Initially, we are going to extract text from OCR that is optical character recognition that converts that to text and feed it to the transformer for getting MCQ pattern questions.

# Table Of Figure

[Transformer](#_heading=h.bpzxi6dehh2u) 12

[T5 Transformer](#_heading=h.vtlbmbtifina) 13

[Prototype design of the proposed system](#_heading=h.hw7qzgj7sebm) 18

[Output](#_heading=h.999oksslinm6) 19

# 

# 

# 

# 

# 

# 

# 

# Chapter 1

# **Introduction**

### 1.1 Introduction:

In this project we are going to implement a question paper generator using a transformer in machine learning. The application will take any kind of file from the user and extract the text from it and get an appropriate output with a set of questions from the given file.

### 1.2 Problem Statement :

Converting any kind of file to extract text and generate a question paper from a given file in MCQ format.

### 1.3 Objective :

1. Extract Text from any given file (img,pdf,word).
2. Generate question paper with MCQ,True/false and short descriptive answers questions.

### 1.4 Scope :

This could be also used by many teachers to help them conduct a short quiz with minimum efforts on preparation of questions.

This will also help parents and students understand the kind of questions that can be framed from a given paragraph which help them to improve their score.

# 

# 

# 

# 

# 

# 

# 

# Chapter 2

# **Literature Survey**

### 2.1 Literature/Techniques :

A transformer model is a neural network that learns context and thus meaning by tracking relationships in sequential data like the words in this sentence.

Transformer models apply an evolving set of mathematical techniques, called attention or self-attention, to detect subtle ways even distant data elements in a series influence and depend on each other.

Transformers are translating text and speech in near real-time, opening meetings and classrooms to diverse and hearing-impaired attendees.

Transformers can detect trends and anomalies to prevent fraud, streamline manufacturing, make online recommendations or improve healthcare.

Before transformers arrived, users had to train neural networks with large, labeled datasets that were costly and time-consuming to produce. By finding patterns between elements mathematically, transformers eliminate that need, making available the trillions of images and petabytes of text data on the web and in corporate databases.

In addition, the math that transformers use lends itself to parallel processing, so these models can run fast.

#### Transformer



#### T5 Transformer

Google’s state of the art, T5 — Text-to-Text Transfer Transformer Model which was proposed earlier this year in the paper, “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”.

BERT-style architecture that is pre-trained on a Masked LM and Next Sentence Prediction objective and then fine-tuned on downstream tasks (for example predicting a class label in classification or the span of the input in QnA). Here, we separately fine-tune different instances of the pre-trained model on different downstream tasks.



The text-to-text framework, on the contrary, suggests using the same model, same loss function, and the same hyperparameters on all the NLP tasks. In this approach, the inputs are modeled in such a way that the model shall recognize a task, and the output is simply the “text” version of the expected outcome. Refer to the above animation to get a clearer view of this.

T5 as this is what makes the unified text-to-text approach possible. To avail the same model for all the downstream tasks, a task-specific text prefix is added to the original input that is fed to the model. This text prefix is also considered as a hyperparameter.

### 2.2 Papers and Findings :

[Link](https://towardsdatascience.com/t5-text-to-text-transfer-transformer-643f89e8905e) > The link tells about the transformer and its architecture and working.

[Link](https://amontgomerie.github.io/2020/07/30/question-generator.html) > From this website we learned about question generatoration using transformers and about multiple datasets.

[Link](https://towardsdatascience.com/questgen-an-open-source-nlp-library-for-question-generation-algorithms-1e18067fcdc6) > This website helped us understand an already available package in python called Questgen which can create a one word, MCQ, True false questions.

# 

# 

# 

# 

# 

# 

# 

# Chapter 3

# Data Set

### 3.1 Description of data set :

The IAM Handwritten Forms Dataset contains forms of handwritten English text which can be used to train and test handwritten text recognizers and to perform writer identification and verification experiments. The dataset contains complete forms of unconstrained handwritten text, which were scanned at a resolution of 300dpi and saved as PNG images with 256 gray levels. Forms are partitioned into separate directories such that all forms in each directory are written by the same person.

Size of the dataset is 4.42 GB .

### 3.2 Data collection methodology :

IAM Handwriting Dataset is a collection of handwritten passages by several writers. Generally, they use that data to classify writers according to their writing styles.

The dataset is extracted from kaggle which is an online community platform for data scientists and machine learning enthusiasts. Kaggle allows users to collaborate with other users, find and publish datasets and use GPU integrated notebooks.

Dataset Link :- <https://www.kaggle.com/datasets/naderabdalghani/iam-handwritten-forms-dataset>

### 3.3 Exploratory data analysis :

Input images are processed according to the chosen features extraction method, but whatever the chosen extraction method is, the image essentially goes through the same procedures. The preprocessing module simply crops the image to its useful content (i.e the handwritten text corpus) and returns a binarized version of the cropped image and a grayscale version depending on the type of the features extraction method.The dataset chosen has no anomalies.



### 3.4 Feature extraction :

The first is implemented where the image is segmented into word images which are then overlapped into a single image to represent the texture of a particular writer’s handwriting. Texture images are then segmented into texture blocks that are each represented as an input vector for the next phase using their local binary pattern histograms.

The other feature extraction method works by simply feeding the model in the next phase raw segmented images (e.g sentences, words) of the cropped image. Despite being simple, this method takes much more time to execute than the overlapping method mentioned above as it takes an average execution time of 10 seconds per iteration while the overlapping method takes around 3.5 seconds per iteration, in addition to that, the overlapping method yields better accuracy by a slight yet accountable margin making matter disadvantageous for the lines method.

# 

# 

# 

# 

# 

# 

# 

# Chapter 4

# Analysis and Design

### 4.1 Analysis of the system :

The system will be app as well as website based in which first the system will take the input from user in form of file in any file type and will extract text from the file provided the extracted text will be then passed to model which can generate the question in MCQ and one word form from which user has to decide and the quest with apt solution will be provided by model.

### 4.2 Proposed Solutions :

The solution to the problem is use of a new technology called Transformer which contains encoders and decoders that helps optimize and train the data as per requirement.

### 4.3 Prototype design of the proposed system :

#### 

 

The above images are of an android application that was developed by us and the app accepts the image file and extracts the text from that Img 1.2 as seen from above image.

The Gradio (Img 1.3) is used in Google colab notebook as UI which takes input as text from user and generates a one word answer based question from the given paragraph.

#### 



# 

# 

# 

# 

# 

# 

# Chapter 5

# Results and Discussion

We were able to build an app that can convert image files into text form which has been extracted. We also build a basic T5 model which use of Gradio as UI for now, the model takes input text and generates a Multiple Choice Question using Sense2Vec and Word2Vec.

In this semester we have accomplished the goal of initializing app generation and algorithm selection for given problem statements. Partial Model generation is also a [**colab link**](https://colab.research.google.com/drive/1Z4LgNLTm_8vtYvj-vVv8FpA_Fz7LWwIR?usp=sharing) in which we have been able to successfully generate a MCQ question.

# 

# 

# 

# 

# 

# Chapter 6

# Conclusion and Future Work

In this semester we learned about various algorithms that support our question generators such as Decoders, Encoders, BERT & Transformer. The OCR (optical character recognition) is implemented in an android app. We also implemented a T5 transformer model which generates questions in MCQ pattern for now.

In future work we would like to implement a website and integrate OCR with our T5 transformer model and will also accomplish the task that might not be foreseen for now.

# Reference :

**Journal Paper**

[1] Colin Raffel,Noam Shazeer, Exploring the Limits of Transfer Learning with a Unified

Text-to-Text Transformer, Journal of Machine Learning Research 21 (2020) .

[2] Mr. Aditya Sanjay Shendure , Mr. Aniruddh Dagadu Khamkar , Mr. Prasanna Premnath Vathare , Mr. Avadhut Salavi(Kumbhar), Question Paper Generator System , International Journal for Research in Applied Science & Engineering Technology (IJRASET) .

[3] Rohan Bhirangi, Smita Bhoir,Automated Question Paper Generation System, Computer Engineering Department, Ramrao Adik Institute of Technology,Navi Mumbai, Maharashtra, India.

[4] Surbhi Choudhary, Abdul Rais Abdul Waheed, Shrutika Gawandi and Kavita Joshi, “Question Paper Generator System,” International Journal of Computer Science Trends and Technology, vol. 3, issue 5, Sept – Oct 2015.

[5] Prita Patil and Kavita Shirsat, “An Integrated Automated Paperless Academic Module for Education Institutes,” International Journal of Engineering Science Invention Research and Development, vol. I, issue IX, March 2015.

[6] Ashok Immanuel and Tulasi.B, “Framework for Automatic Examination Paper Generation System,” International Journal of Computer Science Trends and Technology, vol. 6, issue 1, Jan - March 2015.