GENERATING MCQs FROM TEXT USING NLP

A Project Report in partial fulfillment of the degree

Bachelor of Technology

in

Computer Science & Engineering/Electrical & Electronics Engineering

By

19K41A05G2	M.YASHASHREE
19K41A05G1	B. SHRITHAJANI
20K45A0229	A. MAHESH BABU

Under the Guidance of

D. RAMESH

Submitted to

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
S R ENGINEERING COLLEGE(A), ANANTHASAGAR, WARANGAL
(Affiliated to JNTUH, Accredited by NBA)

NOV -2022

DEPARTMENT OF COMPUTER SCIENCE / ELECTRICAL & ELCTRONICS ENGINEERING

CERTIFICATE

This is to certify that the Project Report entitled **"**Generating MCQs from text Using NLP " is a record of bonafide work carried out by the student(s) M.Yashasree, M.Shrithajani, A. Mahesh Babu bearing Roll No(s) 19K41A05G2, 19K41A05G1, 20K45A0229 during the academic year 2022-2023 in partial fulfillment of the award of the degree of Bachelor of Technology in Electrical & Electronics /Computer Science Engineering by theJawaharlal Nehru Technological University, Hyderabad.

.

Supervisor **Head of the Department**

External Examiner

ABSTRACT

Automatic multiple-choice question generation (MCQG) is a useful yet challenging task in Natural Language Processing (NLP). Examinations and Assessments are undergoing a tremendous revolution. Universities, colleges, and other educational institutes are increasingly shifting towards online examinations. The pattern of assessment is majorly shifting towards the objective assessment i.e. MCQ based, it is very hard to construct and requires a considerable amount of time for setting numerous questions. There's a growing need for a cost-effective and time-efficient automated MCQ generation system. In this paper, the text is first summarized using the BERT algorithm, and accordingly sentence mapping is done for generating MCQs. In order to generate choices for the questions, distractors are generated using wordnet (A lexical database for English). As the BERT algorithm has much better performance over other legacy methods as well as it can process a large amount of data in less time, it will enhance the speed of generating mcq's from given text.

Table of Contents

1. INTRODUCTION

All institutes, colleges, and schools have been switched to online learning. Assessment is an essential tool to test the knowledge of the students. And the pattern of the assessment has changed from subjective based to objective based i.e., Multiple Choice Questions (MCQs).

So ,the problem is, it is very difficult for the teachers to set the questions as well as for the students who are preparing for competitive exams. The current method involves the setting of questions manually which requires a lot of human intervention and time. So there is a growing need for a system that can create questions with ease and less amount of time and requires less human effort.

Automatic multiple-choice question generation (MCQG) is a useful yet challenging task in Natural Language Processing (NLP). It is the task of automatic generation of correct and relevant questions from textual data. Despite its usefulness, manually creating sizeable, meaningful and relevant questions is a time-consuming and challenging task for teachers. In this project, we present an NLP-based system for automatic MCQ generation for computer-based testing examination (CBTE). We used NLP technique to extract keywords that are important words in a given lesson material.

This project tells about a system that generates questions automatically. In Automated MCQ Generator, questions are generated automatically with the help of NLP. The text of any domain is provided as input to the system which is then summarized using the BERT algorithm. BERT (Bidirectional Encoder Representation from Transformers) is a deep learning-based technique for natural language processing, a pre-trained model from Google. Now the keywords are selected from the summarized text using the python keyword extractor (PKE) and accordingly mapping of a keyword is done with a sentence. This keyword will be one of the options of MCQ. Now the main task is generating relevant distractors. Distractors are generated using the wordnet approach. Wordnet is an API used to get the correct sense of the word. So the good and relatable distractors are generated. This system solves the problem of manual creation of questions and reduces time consumption and cost.

Natural Language Processing (NLP) is a branch of artificial intelligence that deals with the interaction between computers and humans using the natural language. The ultimate objective of NLP is to read, decipher, understand, and make sense of the written or textual data in natural languages of human in a manner that is valuable. In this paper, we focus on the setting of MCQs through NLP to improve the method of setting MCQs and modification, and for creating a viable question bank for subsequent use by the academicians for their learners. This will ensure an MCQ that includes appropriate questions and +

options based on the learning objectives and importance of the topics discussed in a lesson material. We used NLP methods, Term Frequency-Inverse Document Frequency (TF-IDF) and N-grams, to extract the most significant words present in a lesson material and the selection does not depend on any vocabulary; these words are keywords that capture the main topics discussed in a lesson material, they also serve as an indication of document relevance for users in an information retrieval (IR) environment

2. LITERATURE REVIEW

We have studied several research papers on multiple choice question generation using different approaches. Santhanavijayan et al. have proposed a system of "Automatic generation of multiple-choice questions for e-assessment"

[1]. Chidinma A Nwafor and Ikechukwu onyenwe "Automatic multiple-choice question generation (MCQG) is a useful yet challenging task in Natural Language Processing (NLP). It is the task of automatic generation of correct and relevant questions from textual data. Despite its usefulness, manually creating sizeable, meaningful and relevant questions is a time-consuming and challenging task for teachers."

[2]. In their proposed system, they have used fireflies-based preference learning and ontology-based approach to generate MCQs. They have used a web corpus to make it feasible to create questions. The distractors are generated using similarity metrics such as hypernyms and hyponyms. The system also creates analogy questions to test the verbal ability of the students.

[3]. The Ayako Hoshino and Hiroshi Nakagawa "A real-time multiple-choice question generation for language testing: A preliminary study" [2] is based on machine learning to generate questions automatically. They implement machine learning algorithms, such as Naive Bayes and K-Nearest Neighbors, to create questions on English grammar and vocabulary from online news articles. They designed a system that can receive user input in the form of an HTML file and turns it into a quiz session.

[4]. Deepshree S. Vibhandik et al. have proposed a system, "Automatic / Smart Question Generation System for Academic Purpose" [4] in which the Automatic Question Generation system generates specific trigger questions and multiple-choice questions from student's literature review papers. To facilitate the generation of specific trigger questions, the system extracts key concepts from student's papers using the Lingo algorithm. Also, to bring out the generation of multiple-choice questions, the system pulling out abbreviations from student's review papers using the regular expression pattern matching techniques.

[5]. Ms. Nikeeta Patil, Ms. Kratika Kumari, Mr. Devendra Ingale have proposed a syatem ,"Efficient multiple-choice questions are produced withquality distractors. The necessity of human interventionfor generating question paper and answer is eliminated.The proposed system creates automated questions with the help of NLP that reduces human intervention and it is a cost and time-effective system also, the accuracy of the distractor generated is high."

[6]. Ebrahim Gabajiwala, Priyav Mehta, Ritik Singh have proposed a system,"The recent advancement in NLP techniques has shown a lot of promise. The proposed solution uses an NLP pipeline involving Bert and T5 transformers to extract keywords and gain insights from the text input. From the extracted keywords, different types of questions are generated such as fill in the blanks, true or false, Wh-type and multiple choice questions. Latest state-of-the-art models proved to perform better in all stages of our pipeline. The results from these models have shown a lot of promise".

[7]. Sonam Sonia, Praveen Kumarb, Amal Saha have proposed "generation (QG) is a very important yet challenging problem in NLP. It is defined as the task of generating syntactically sound, semantically correct and relevant questions from several input formats like text, a structured database or a knowledge base. Question generation can be naturally applied in many domainssuch as MOOC, automated help systems, search engines, chatbot systems (e.g. for customer interaction), and healthcare for analyzing mental health.

3. DESIGN:

3.1 Requirement Specifications (S/W & H/W)

Hardware Requirements

- ✓ **System** : Processor Intel(R) Core (TM) i5-8265U CPU @ 1.60GHz, 1800 MHz, 4 Cores, 8 Logical Processors
- ✓ **RAM** : 8 GB
- ✓ **Hard Disk** : 557 GB
- ✓ **Input** : Keyboard and Mouse
- ✓ **Output** : PC

Software Requirements

- ✓ **OS** : Windows 10
- ✓ **Platform** : Google Colaboratory / Jupyter Notebook
- ✓ **Deployment software** : Stream lit
- ✓ **Program Language** : Python

3.2 Flow chart

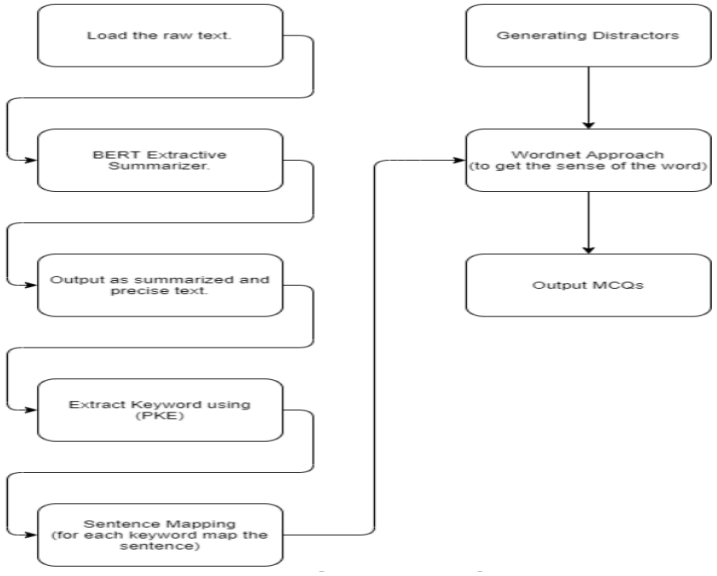
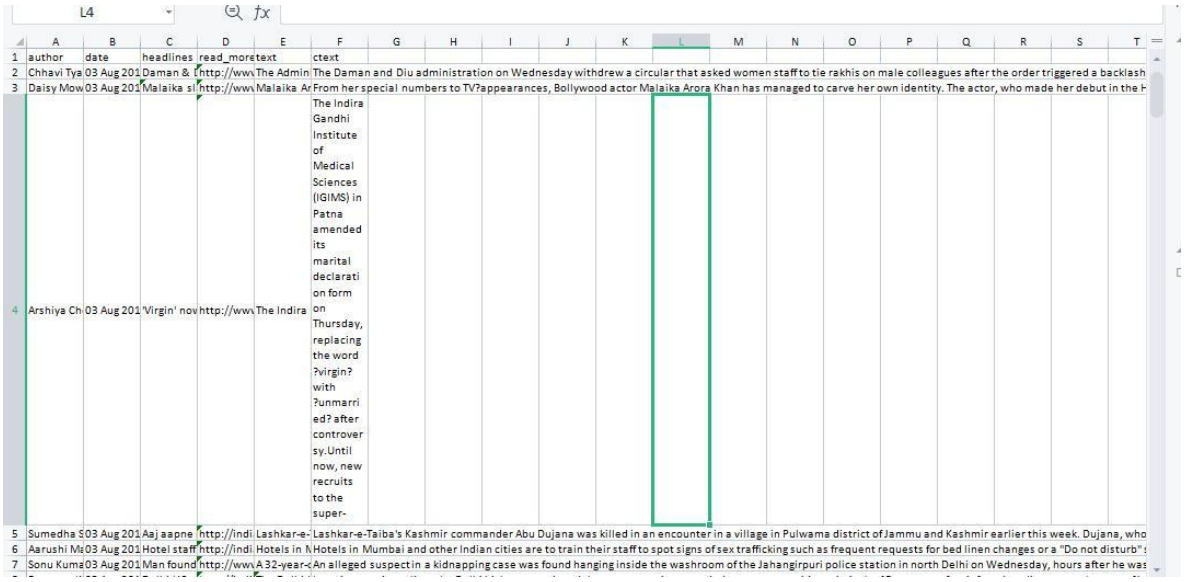


Fig 1: Flow chart representation

In the above flow chart we described the work flow of our project.

4. DATASET:

The data we used is from the “news article” which consists of 98,400 records for training and testing of data.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	author	date	headlines	read_more	text	ctext														
2	Chhavi Tya	03 Aug 201	Daman & D	http://ww	The Admin	The Daman and Diu administration on Wednesday withdrew a circular that asked women staff to tie rakhis on male colleagues after the order triggered a backlash														
3	Daisy Mow	03 Aug 201	Malaika sl	http://ww	Malaika Ar	From her special numbers to TV?appearances, Bollywood actor Malaika Arora Khan has managed to carve her own identity. The actor, who made her debut in the H														
4	Arshiya Ch	03 Aug 201	'Virgin' nov	http://ww	The Indira	The Indira Gandhi Institute of Medical Sciences (IGIMS) in Patna amended its marital declaration form on Thursday, replacing the word ?virgin? with ?unmarried? after controversy.Until now, new recruits to the super-														
5	Sumedha S	03 Aug 201	Aaj aapne	http://indi	Lashkar-e-	Lashkar-e-Taiba's Kashmir commander Abu Dujana was killed in an encounter in a village in Pulwama district of Jammu and Kashmir earlier this week. Dujana, who														
6	Aarushi Ma	03 Aug 201	Hotel staff	http://indi	Hotels in M	Hotels in Mumbai and other Indian cities are to train their staff to spot signs of sex trafficking such as frequent requests for bed linen changes or a "Do not disturb"														
7	Sonu Kuma	03 Aug 201	Man found	http://ww	A 32-year-(	An alleged suspect in a kidnapping case was found hanging inside the washroom of the Jahangirpuri police station in north Delhi on Wednesday, hours after he was														

Fig 2: Dataset of news article

Input features:

- Author
  - Value : Text
- Date
  - Value : 0 - 31 (Numeric input)
- Headlines
  - Value : Text
- Original Text
  - Value : text

Output feature:

- Summarized text
  - Value : text

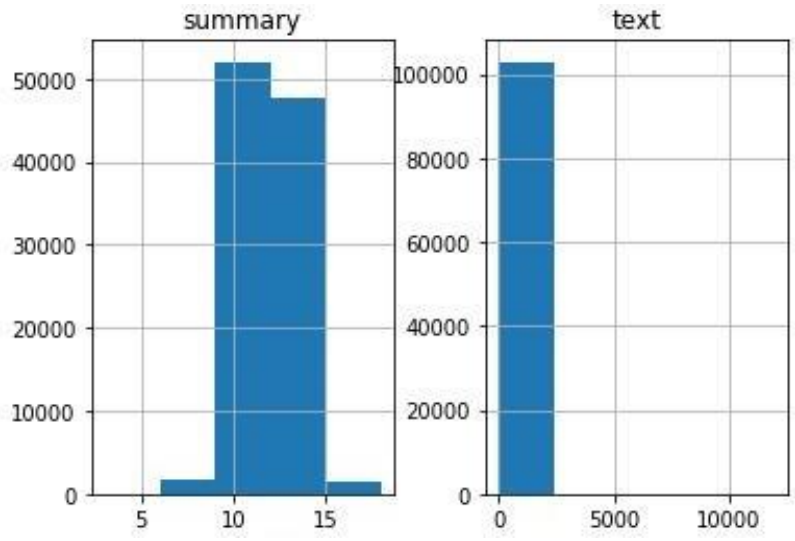


Figure 3: Visualizing attributes of the dataset

5. DATA PREPROCESSING:

Real-world data collection has its own set of problems. It is often very messy which includes missing data, presence of outliers, unstructured manner, etc. Before looking for any insights from the data, we have to first perform preprocessing tasks which then only allow us to use that data for further observation and train our machine learning model. Weuse missing values treatment, outliers detection, normalization and data split to process ourdata before feeding it to the machine learning model.

Data info:

```
In [2]: summary = pd.read_csv('/kaggle/input/news-summary/news_summary.csv', encoding='iso-8859-1')
        raw = pd.read_csv('/kaggle/input/news-summary/news_summary_more.csv', encoding='iso-8859-1')

In [3]: pre1 =  raw.iloc[:,0:2].copy()
        # pre1['head + text'] = pre1['headlines'].str.cat(pre1['text'], sep =" ")

        pre2 = summary.iloc[:,0:6].copy()
        pre2['text'] = pre2['author'].str.cat(pre2['date'].str.cat(pre2['read_more'].str.cat(pre2['text'].str.cat(pre2['ctext'], sep = " "), sep =" "),sep= "

In [4]: pre = pd.DataFrame()
        pre['text'] = pd.concat([pre1['text'], pre2['text']], ignore_index=True)
        pre['summary'] = pd.concat([pre1['headlines'],pre2['headlines']],ignore_index = True)

In [5]: pre.head(2)

Out[5]:
```

	text	summary
0	Saurav Kant, an alumnus of upGrad and IIIT-B's...	upGrad learner switches to career in ML & Al w...
1	Kunal Shah's credit card bill payment platform...	Delhi techie wins free food from Swiggy for on...

Fig 4 : Data preprocessing

Missing values treatment:

The real world's dataset often has many missing values which can be treated by using certain methods. But in our data, there are no missing values, because we collected the data manually through google forms survey and made sure not to miss out any data. To treat the missing values, we generally use the following strategies:

- Remove the entire row (If missing values are less in number)
- Replace the missing value with either mean or median
- Replace the missing value with most frequent value in the column (This is generally used only for large dataset)

Normalization*:*

Normalization is a technique for organizing data in a database. Data normalization is the process of rescaling one or more attributes to the range of 0 to 1. This means that the largest value for each attribute is 1 and the smallest value is 0. It is important that a database is normalized to ensure only related data is stored in each table and to avoid biasing towards huge values. When we normalize the data while feeding it to the model, we also have to de-normalize it. This process can be done using the formulas below:

- $= \frac{-}{-}$
- $= - (+ \quad)$

Data split:

To train any machine learning model irrespective what type of dataset is being used, we have to split the dataset into training and testing data. The reason to split the data is to give the machine learning model an effective mapping of input to outputs and to evaluate the model performance. We pass the training data to train our machine learning model and then test the model on testing data.We can do the data split using train_test_split module in python.

6. METHODOLOGY:

This section talks about the whole procedure and methods used for this project.

LOAD RAW DATA

The first step is to load raw text i.e. input text of any domain for which the questions to be generated.

SUMMARIZER

Each sentence is not capable of generating questions. Only the sentences that contain a questionable fact can act as a candidate for creating MCQs. Therefore, sentence selection plays a crucial role in the automatic MCQ generation task. Hence for summarizing the text, BERT Algorithm is used. BERT (Bidirectional Encoder Representations from Transformers) is a neural network-based technique for natural language processing. It is a pre-trained open-sourced model from Google. It helps computers to understand the language a bit more as humans do. The input text is summarized using the BERTSUM model, which is fine-tuned BERT for extractive summarization. The architecture of BERTSUM is shown in Fig-2.

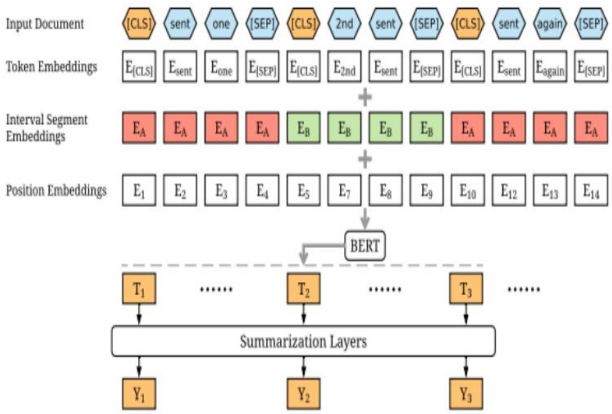


Fig.-2: Overview Architecture of BERTSUM model

KEYWORD EXTRACTION

After summarizing the text, keywords are selected from the sentence. This keyword will be the answer to the question. Since all the words in an informative sentence cannot serve as the key proper keyword selection is required. The extraction of keyword is done by python library RAKE.

Rapid Automatic Keyword Extraction (RAKE) is a well known keyword extraction method that uses a list of stop words and phrase delimiters to detect the most relevant words or phrases in a piece of text.

GENERATION OF DISTRACTERS

Generating distractors is the most crucial step in the generation of automated MCQs. The difficulty of MCQs highly relies on the quality of distractors produced. A good distractor is one that is very similar to the key but not the key itself. So, for generating distractors Wordnet approach is used.

WordNet is a lexical database for the English language, which was created by Princeton, and is part of the NLTK corpus. In the WordNet network, the words are connected by linguistic relations. These linguistic relations includes hypernym, hyponym, meronym, holonym, etc. WordNet stores synonyms in the form of synsets (Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms) where each word in the synset shares the same meaning. Basically, each synset is a group of synonyms. Each synset has a definition associated with it. Relations are stored between different synsets. This Lesk algorithm is based on the assumption that words in a given "neighborhood" (section of text) will tend to share a common topic. A simplified version of the Lesk algorithm compares the dictionary definition of an ambiguous word with the terms contained in its neighbourhood.

For example, if there is a sentence “The bat flew into the jungle and landed on a tree” and a keyword is “bat”, we automatically know that here we are talking about the mammal bat that has wings, not a cricket bat or baseball bat. Although we humans are good at it, the algorithms are not very good at distinguishing one from another. This is called word sense disambiguation (WSD). In the wordnet, “bat” may have several senses, one for a cricket bat, one for flying mammal etc. So the function get_wordsense tries to get the correct sense of the word given in the sentence. Once the sense of the word is identified the get_distractors_wordnet function is called to get the distractors. This function tries to get the distractors with the help of hypernyms and hyponyms of the key.

HYPERNYM:

Hypernym is a word that names a broad category that includes other words.

Ex: animal is hypernym of dog.

HYPONYM:

A word of more specific meaning than a superordinate term applicable to it.

Ex: car is a hyponym of vehicle.

Now all the possible hypernyms of the key and the corresponding hyponyms for each of the hypernym are found. These hyponyms are considered potential distractors. Then, the potential distractors are ranked, and the final distractors are chosen based on their rank. For the word 'bat' as the key, the hypernym for the bat is - an animal of which hyponyms can be an eagle and other birds which can be

good distractors for the key. The ranks given to the potential distractors are based on whether it occurs in the text extracted from the summarization. The potential distractors that exist in the extracted text and having the same part of speech structure as the key have a higher rank than those with a different structure. This is because distractors with the same structure as the key tend to confuse the test takers more. Any three potential distractors with higher ranks are chosen at random as the final distractors.

7. RESULTS:

We used an machine learning algorithms in this project, and the algorithm LSTM have high level accuracies for our dataset. We consider testing accuracy for final deployment of the project. Here, the testing accuracy of LSTM is highest with 87%. As this model has highest accuracy , we use that model to develop the web application for the project. Accuracies with training and testing data of the model is represented in graphical andtabular form below.

Training and Testing accuracy:

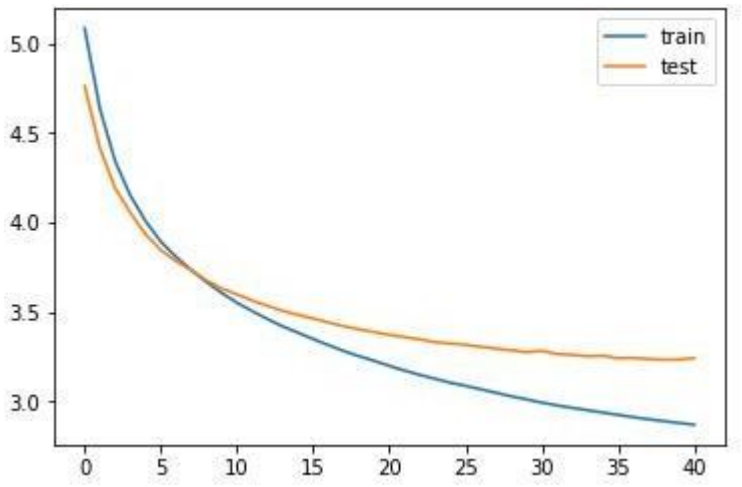


Fig 7: Training and Testing of the models

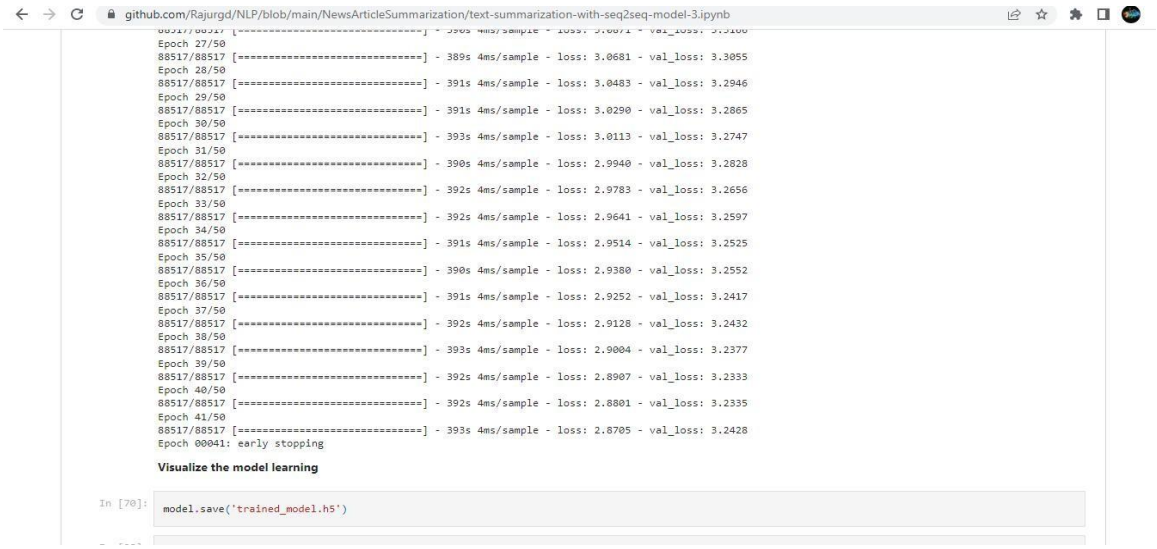


Fig 8: Model accuracy

Confusion matrix

- LSTM :

array([[3498, 4670, 6, ..., 0, 0, 0],

[279, 4, 47, ..., 0, 0, 0],

[389, 28, 1352, ..., 0, 0, 0],

...,

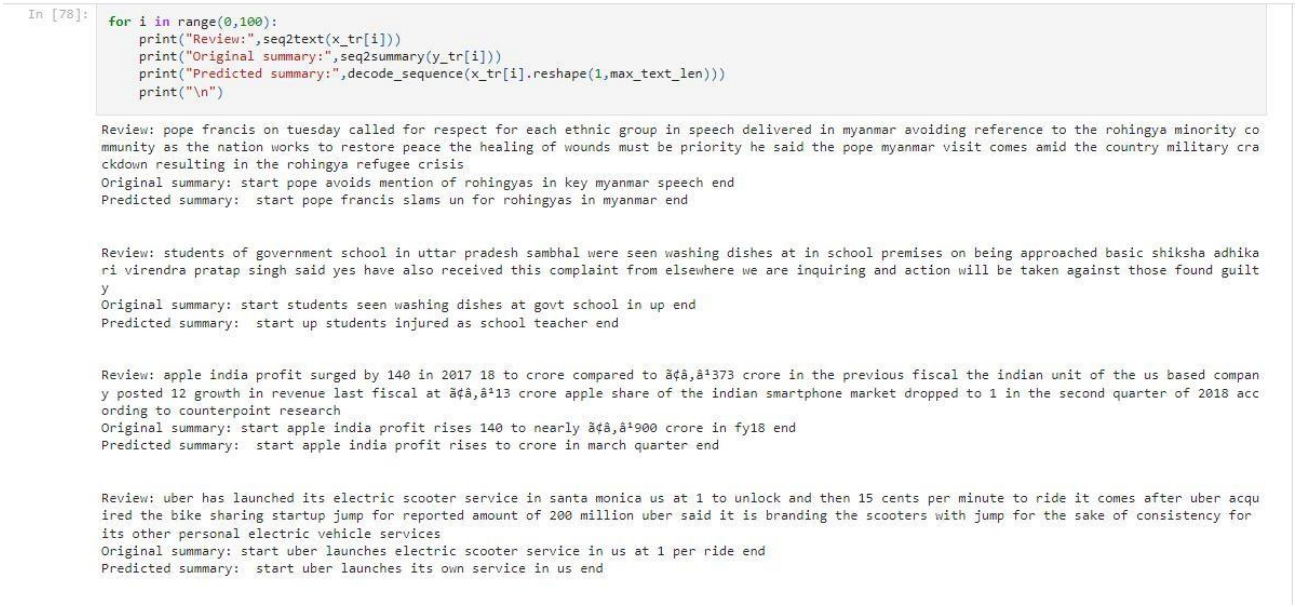
[16812, 73, 534, ..., 0, 0, 0],

[44, 10160, 1, ..., 0, 0, 0],

[145, 90, 490, ..., 0, 0, 0]], dtype=int32)

8. OUTPUT PREDICTION:

We have developed application through Kaggle Notebook by implementing LSTM and Seq2Seq as it has the highest testing accuracy)



```
In [78]: for i in range(0,100):
             print("Review:",seq2text(x_tr[i]))
             print("Original summary:",seq2summary(y_tr[i]))
             print("Predicted summary:",decode_sequence(x_tr[i].reshape(1,max_text_len)))
             print("\n")

Review: pope francis on tuesday called for respect for each ethnic group in speech delivered in myanmar avoiding reference to the rohingya minority co
mmunity as the nation works to restore peace the healing of wounds must be priority he said the pope myanmar visit comes amid the country military cra
ckdown resulting in the rohingya refugee crisis
Original summary: start pope avoids mention of rohingyas in key myanmar speech end
Predicted summary:  start pope francis slams un for rohingyas in myanmar end


Review: students of government school in uttar pradesh sambhal were seen washing dishes at in school premises on being approached basic shiksha adhika
ri virendra pratap singh said yes have also received this complaint from elsewhere we are inquiring and action will be taken against those found guilt
y
Original summary: start students seen washing dishes at govt school in up end
Predicted summary:  start up students injured as school teacher end


Review: apple india profit surged by 140 in 2017 18 to crore compared to ã¢â‚¬â„¢373 crore in the previous fiscal the indian unit of the us based compan
y posted 12 growth in revenue last fiscal at ã¢â‚¬â„¢13 crore apple share of the indian smartphone market dropped to 1 in the second quarter of 2018 acc
ording to counterpoint research
Original summary: start apple india profit rises 140 to nearly ã¢â‚¬â„¢900 crore in fy18 end
Predicted summary:  start apple india profit rises to crore in march quarter end


Review: uber has launched its electric scooter service in santa monica us at 1 to unlock and then 15 cents per minute to ride it comes after uber acqu
ired the bike sharing startup jump for reported amount of 200 million uber said it is branding the scooters with jump for the sake of consistency for
its other personal electric vehicle services
Original summary: start uber launches electric scooter service in us at 1 per ride end
Predicted summary:  start uber launches its own service in us end
```

Fig 9 : Output result

9. CONCLUSION:

Multiple Choice Questions (MCQs) are generated successfully. Efficient questions are produced with good quality distractors. The problem of manually creating questions is solved with the proposed system. The proposed system creates automated questions with the help of NLP that reduces human intervention and it is a cost and timeeffective system. And the accuracy of the distractor generated is reasonably high. This system not only helps teachers with E-assessments but also helps students who are preparing for competitive exams. Students can test their ability to solve the questions and can also check their understanding of the concepts.

10. REFERENCES:

[1] https://www.researchgate.net/publication/351665611_An_Automated_Multiple-Choice_Question_Generation_using_Natural_Language_Processing_Techniques

[2] https://www.researchgate.net/publication/317610512_Automatic_generation_of_multiple_choice_questions_for_e-assessment

[3] https://aclanthology.org/W05-0203.pdf

[4] https://www.ijettcs.org/Volume4Issue4/IJETTCS-2015-07-13-27.pdf

[5] https://ijsret.com/wp-content/uploads/2021/05/IJSRET_V7_issue3_470.pdf

[6] https://research.spit.ac.in/storage/290/8_Quiz-Maker--Automatic-Quiz-Generation-from-Text-using-NLP.pdf

[7] https://deliverypdf.ssrn.com/delivery.php?ID=740102120112125028008094012081087098016073055036039026094121111075069090031124025111057021116004010037021064027026011095099027041086005060042090101116006002122067089062011064008086029106068010127122096094086080121100067007027029071026093016074066069009&EXT=pdf&INDEX=TRUE

[8] https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

[9] Rahul, S. Adhikari and Monika, "NLP based Machine Learning Approaches for Text Summarization," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 2020, pp. 535-538