e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:05/May-2023 Impact Factor- 7.868 www.irjmets.com

AN AUTOMATIC MCQ & SUMMARY GENERATION BY USING NLP

Harshada Patil[*1], Aarti Gaikwad[*2], Mitali Dalave[*3], Sneha Surywanshi[*4], Prof. Pallavi Patil[*5]

[*1,2,3,4]Student, Department of Computer Science and Engineering. SETI, Panhala, Kolhapur, Maharashtra, India.

[*5]Assistant Professor, Department of Computer Science and Engineering, SETI, Panhala, Kolhapur, India

ABSTRACT

Assessments and Evaluations are going through a gigantic upset. Colleges, universities, and other instructive organizations are progressively moving towards on the web assessments. The example of appraisal is significantly moving towards the objective appraisal for example MCQ based, it is exceptionally hard to build and demands a lot of investment for setting various inquiries. There's a developing requirement for an expense viable and time-proficient computerized MCQ age framework. In this paper, the text is first summed up utilizing the BERT calculation, and likewise sentence planning is finished creating MCQs. To create decisions for the questions, distractors are created utilizing wordnet (A lexical data set for English). As the BERT calculation has much better execution over other inheritance techniques also as it can process a lot of information quicker than expected, it will upgrade the speed of creating MCQs from given text. Text summarization is defined as generating a short, accurate, and fluent summary. which is extremely valuable in a few certifiable applications. In this paper, we proposed. In this paper, we proposed an extractive synopsis model called ClinicalBertSum, which depends on BERT

Key Words: NLP, MCQs, BERT, Wordnet, Distractors generator, Summary.

I. INTRODUCTION

All organizations, universities, and schools have been changed to online learning. Appraisal is a fundamental apparatus to test the information on the understudies. Whats more, the example of the appraisal has changed from abstract based to objective based i.e. Multiple Choice Questions (MCQs). Automatic multiple-choice question generation (MCQG) is a useful still challenging task in Natural Language Processing (NLP). It is the task of automatic generation of correct and relevant questions from textual data. 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 web resources on the Internet (e.g. websites, user reviews, news, blogs, social media networks, etc.) are gigantic sources of textual data. Besides, there is a wealth of textual content on the various archives of news articles, novels, books, legal documents, biomedical documents, scientific papers, etc. The textual content on the Internet and other archives grow exponentially on a daily basis.

In summary generation there are many repeated or unimportant portions of the resulting texts. Therefore summarizing and condensing the text resources becomes urgent and much more important. Manual summarization is an expensive task and consumes a lot of time and effort. Practically, it is very difficult for humans to manually summarize this huge amount of textual The Automatic Text Summarization (ATS) is the key solution to this issue. This paper tells about a framework that produces questions automatically. In Automated MCQ Generator, questions are produced automatically with the assistance of NLP. The text of any area is given as contribution to the framework which is then summed up utilizing the BERT algorithm. Main task is generating relevant distractors. Distractors are produced utilizing the wordnet approach. Wordnet is a Programming interface used to get the right feeling of the word. So the great and appealing distractors are produced. This framework solves the problem of manual creation of Questions, Summary and reduces time consumption and cost.

II. PROPOSED SYSTEM DESIGN

In the Proposed System we are going to discuss detailed Knowledge about this project our project we use the python Flask for Backend for that we created python flask enviorment In our system we generate summary and Mcq by using the after that we generate Questions also we were able to create distractors for MCQ options. there are different stages involved in this process as shown in fig.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:05/May-2023 Impact Factor- 7.868 www.irjmets.com

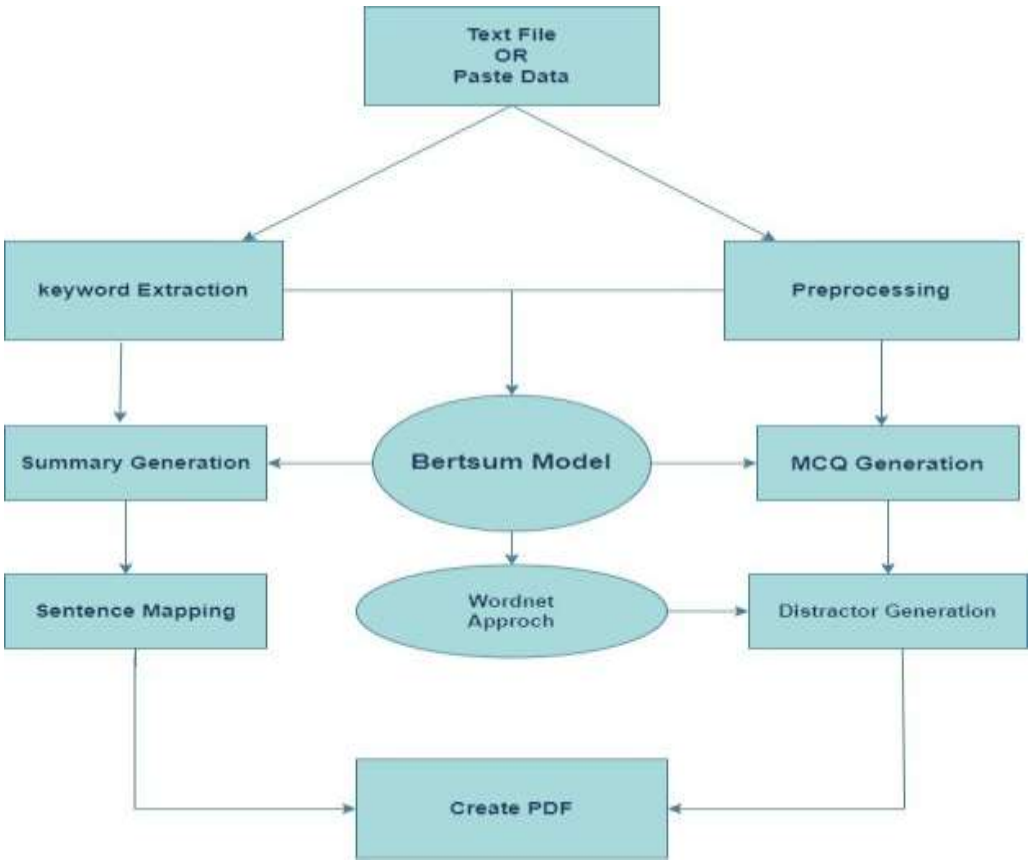


Figure 1: Proposed System Architecture

1.Text File OR Paste Data-The first step input text file or Paste Data i.e.input text of any domain for which the questions and mcq to be generated.

2. Bertsum Model-Summarizing the text, BERT Algorithm is used. Each sentence isn't fit for creating questions. As it were the sentences that contain a problematic truth can go about as a contender for making MCQs. BERT (Bidirectional Encoder Portrayals from Transformers) is a neural network-based method for nature language processing. It is a pre-prepared publicly released model from Google. It assists PCs with understanding the language somewhat more as people do. The information text is summed up utilizing the BERTSUM model, which is calibrated BERT for extractive outline. The design of BERTSUM is displayed in Fig-2.

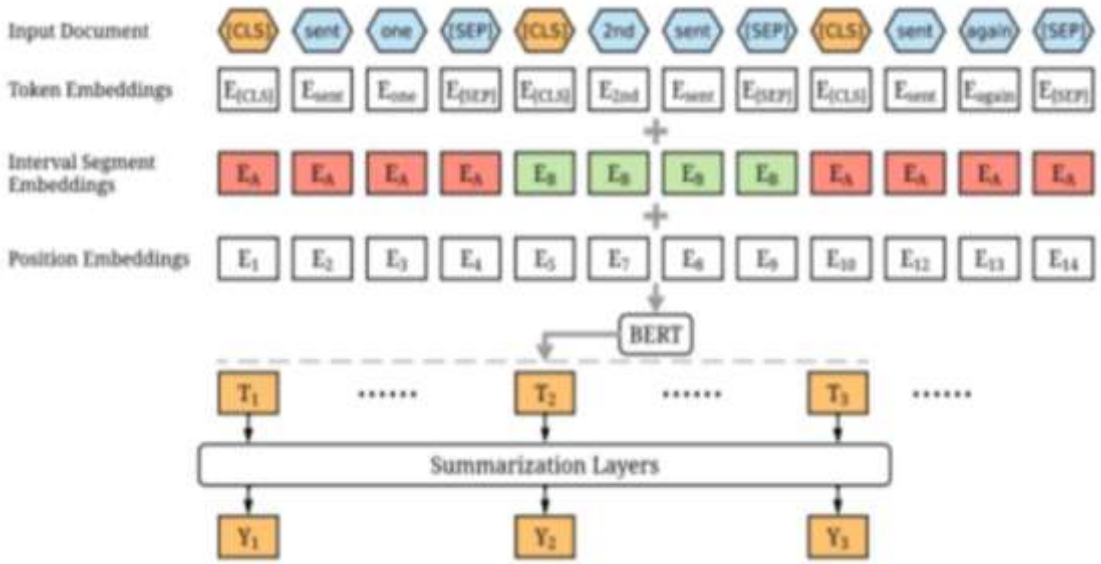


Fig.-2: Overview Architecture of BERTSUM mode

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:05/May-2023 Impact Factor- 7.868 www.irjmets.com

3.Keyword Extraction-

In keyword extraction we find out the most important and higher priority words as a keyword, keywords are nothing but tokens which can be collected by the using Vector matrix technique which can help to provides higher priority words for e.g: If there is sentence (Ram is Clever Boy) then the keywords are extracted from it are Ram, Clever, Boy and (is,a) are get eliminated by Rake algorithm.

The popular keyword extraction technique known as Rapid Automatic Keyword Extraction (RAKE) finds the most important words and phrases in a text by using a list of

stopwords and phrase delimiters. Three elements make up the majority of this algorithm.

- **Candidate Selection:**

All words, phrases, and terms that could be used as keywords are taken from the condensed text during candidate selection. Take a look at the text in the sample belowOne of such is quick automated keyword extraction.

This technique begins by dividing the text into a list of words and eliminating stopwords from that list. Following the division, two lists are created as follows:

Stopwords include: "is, not, that, there, are, can, you, with, of, those, after, all, one.a,an,it" Delimeters: [".", ","]

The words in the above list are ineligible for the candidate key. The remaining words will now be evaluated as candidates.

Eg:

Ram is a good boy

Candidates(keywords)	Stopwords
Ram	Is
Good	A
Boy	.

- **Word co-occurrence matrix**

After obtaining the candidate words, this method creates the word co-occurrence matrix. In addition to all other candidate words, each row in this matrix shows the frequency with which a certain candidate word appears in the candidate sentences.

Consider the following sentences:

Example Corpus: I like deep learning.

I like NLP.

I enjoy flying.

From the above corpus, the list of unique words present are as follows:

Dictionary: ['I', 'like', 'enjoy', 'deep', 'learning', 'NLP', 'flying']

The co-occurrence matrix for the above corpus becomes:

Counts	I	like	enjoy	deep	learning	NLP	flying
I	3	2	1	0	0	0	0
like	2	0	0	1	0	1	0
enjoy	1	0	0	0	0	0	0
deep	0	1	0	0	1	0	0
Learning	0	0	0	1	0	0	1
NLP	0	1	0	0	0	0	1
flying	0	0	1	0	0	0	1

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:05/May-2023 Impact Factor- 7.868 www.irjmets.com

- **Word selection and scoring:**

RAKE determines the keyword score after obtaining a word co-occurrence matrix. That possible word's score is the degree of a word in the matrix, or the entire Amount of co-occurrences the word has with any other content word in the text, is calculated using

K = deg(t)/freq(t)

which is the number of times the word appears in the text.

For example, the word "" has a degree of 8 and a frequency of 3, giving us the score

(K) of K=4/2 =2

The results would be as follows if we were to compute the keyword score for each word in our example:

If we were to calculate the phrase's keyword score, we have matrix for the text *sample* mentioned above:

words	Degree Score
I	3
like	2
enjoy	1
deep	0.5
Learning	0.5
NLP	0.5
flying	0.5

After the keyword is selected sentence is mapped for each keyword i.e. for each of the keyword corresponding sentences that has the word from the summarized text I s extracted.

4. Distractor Generation:

The construction of distractions is the most crucial stage in the production of automated multiple-choice questions. The level of difficulty of the MCQs is significantly influenced by the quality of the produced distractions. One that closely resembles the key but isn't the key works well as a deterrent. The Wordnet approach is therefore used to create diversion.

WordNet is a lexical database for the English language created at Princeton that is a part of the NLTK corpus. By use of linguistic linkages, words in the WordNet network are connected to one another. Meronym, holonym, sense to vectorget best sense(), and sense to vec get most similar() are a few examples of these linguistic connections. 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.

Example 1): If the provided data is the biography then there is some birth year and some other years are mentioned at that time if the question is generated like

What is birth year of Sachin Tendulkar……?

Option:

1) 1980

2) 1974

3) 1990

In this situation one option is include in our data but remaining options are mentioned in that data so the wordnet approach provides those remaining options .

Example 2): Bat are leaves on tree

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:05/May-2023 Impact Factor- 7.868 www.irjmets.com

This sentence produce disambiguation to avoid it wordnet approach provide actual options by using different libraries like

sense2vec.get-best-sense(): It is an extension of the infamous word2vec algorithm.Sense2vec creates embeddings for "senses" rather than tokens of words. A sense is a word combined with a label i.e the information that represents the context in which the word is used.

sense2vec.get-most-similare: It will help to remove disambiguation and provides most similar distractors.

5.Summary Generation:

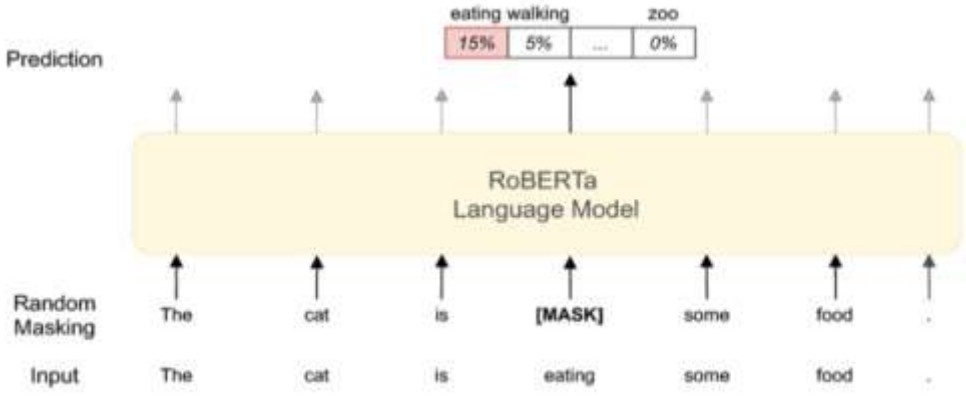
In summary generation we can use the above approach but some additional; contents require because for summary we need to minimize that data also we have to generate proper summary with the great arrangement of lines for that we use bertsum model. A quick and effective pre-trained model is BERT: Bidirectional Encoder Representations from Transformers (BERT) [312]. BERT is designed to modify the left and right context in all layers in order to pre-train deep bidirectional representations from the unlabeled text. Modern models for variety of tasks, including language inference and question-answering, are only available in one output layer of BERT. During fine-tuning, it does not need any significant task-specific architecture alterations.

The Transformer mechanism used by BERT teaches the contextual relationships between words in a text. A decoder that makes predictions and an encoder that receives text input are also included as two independent tools. Transformer encoder reads the full string of words at once, in contrast to directional models that read the text input sequentially (from right to left or left to right) [313].

A decoder that makes predictions and an encoder that receives text input are also included as two independent tools. Transformer encoder reads the full string of words at once, in contrast to directional models that read the text input sequentially (from right to left or left to right) [313]. as a result, it is regarded as bidirectional and, in some cases, non-directional to be more accurate. This capacity to integrate both sides significantly aids BERT in achieving better results. For that prediction bert provides the two pretrained techniques:

- **Masked Language model:**

Masked Language Modelling (MLM) is an unsupervised task performed as part of the BERT pre-train. The purpose of MLM is to assist BERT in comprehending deep bi- dimensional representations. 15% of the WordPiece tokens for the input sequence are randomly masked in MLM. The tokens are hidden by substituting them with [MASK] tokens, which BERT recognises and foresees.



RoBERTa—masked language modeling with the input sentence: The cat is eating some food

- **Next sentence Prediction:**

Another unsupervised task carried out as part of the BERT pre-train is next sentence prediction. This task's goal is to demonstrate the connection between two statements. BERT is pre-trained for a bilingual next sentence prediction that may be produced from any monolingual corpus [7] in order to capture sentence linkages. To do this, 50% of the inputs are converted to sentence pairs in which the second sentence is taken directly from the corpus. The second sentence is instead a randomly chosen sentence from the corpus in the remaining 50% of the sentence combinations. For instance, 50% of the time, if A is a sentence from the corpus, B is the sentence that follows A, and the remaining 50% is random from corpus

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:05/May-2023 Impact Factor- 7.868 www.irjmets.com

By using we can generate the abstractive+extractive summary.this summary can cover all those important points which includes in that document.all the saentence mapping is done by the burtsum model.

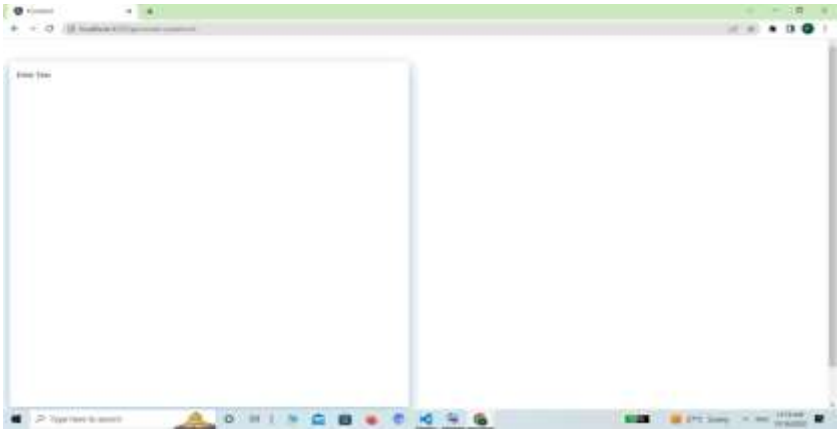
In our last module we are created a pdf of whole extracted that that includes questions, summary and distractors so the user can be able to carry it in their phone also because if anyone want the to extract same data repetedly for those that pdf is very useful also the teachers can set their question paper by using this project.

If sometime due to any technical issue if the distractor are not generated as user want so at that time user can edit that option(distractor)as well as they can make changes in question also. In this way our project is works and it can be useful for human being to minimize their work instead of reading whole document it can creates its summary.

III. RESULTS AND DISCUSSION

This are the snapshots of our output and we successfully implemented automatic mcq and summary generation as well as we provided there are different options for input data such as we can copy paste our data also we can upload textfile too. After the mcq generation we can edit then distractors(option) if sometime the wrong options are generated at that situation we can replace it by correct one We can generate the question paper also by using the print question option In this way our project is successfully implemented.

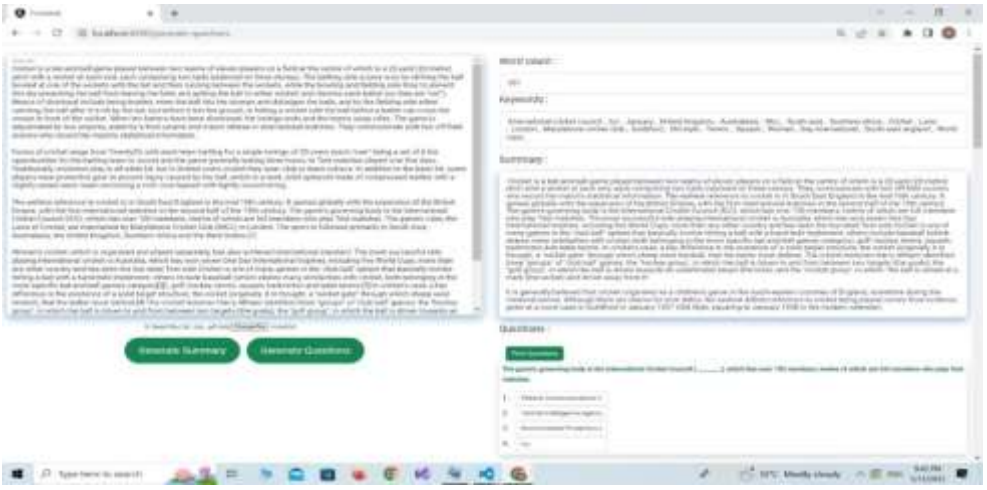
- Website Design:



e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:05/May-2023 Impact Factor- 7.868 www.irjmets.com

- Summary Generation:

- Question Generation:



IV. CONCLUSION

Multiple Choice Questions (MCQs) and summary are generated successfully. The problem of manually creating questions is solved with the proposed system. This system tends to overcome many flaws that the previous application Our proposed system depends on Google's BERT Model, the exactness of the framework will increment later on as the exhibition of the model is improved and as the research in the field of NLP is attempting to arrive at the human level consistently. The basic motivation of creating this system was examinations attempted during the pandemic which were inefficient.

V. REFERENCES

[1] Akhil Killawala, Igor Khokhlov, Leon Reznik – "Computational Intelligence Framework for Automatic Quiz Question Generation" - 2018 IEEE.

[2] A. Shirude, S. Totaia, S. Nikhar (Author), Dr. V. Attar (CoAuthor) and J. Ramanand (CoAuthor) - "Automated Question Generation Tool for Structured Data" - 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI).

[3] Rakesh Patra, Sujan Kumar Saha – "A hybrid approach for automatic generation of named entity distractors for multiple choice questions" – Springer Nature 2018.

[4] Dhaval Swali1, Jay Palan2, Ishita Shah3– "Automatic Question Generation from Paragraph"- International Journal of Advanced Engineering and Research Development.