*Digital Video Making and Smart Answering System*

***Abstract—Visual based learning is a highly accepted learning method in the modern world. The human brain captures videos faster than texts. kids can easily learn through videos rather than textbooks. Kids' learning ability increases when they learn through videos. In this paper, we proposed a system that can produce a video by using text content and kids can engage with video by asking questions to get the answers. Here, various complex technologies related to Natural Language Processing, Deep Learning, especially Sentimental Analysis, and Transfer Learning are used to carry out better outcomes.***

***Keywords — Deep learning, Lexical Simplification, Sentiment Analysis, Transfer Learning , Question answering***

###### INTRODUCTION

Visually telling a content has become an important factor in education, entertainment, and commerce. Many new products are developed and launched in this industry. The COVID- 19 pandemic has affected nearly all social areas, including the education domain. Teaching has become more difficult because of the distance-based learning systems. In the challenging situation of digital learning caused by the COVID-19 restrictions, learning outcomes of the students vary greatly. Most of the Schools and institutions still use old techniques to educate students. Most of the students consider this way of educating students as not effective learning because of the way it was presented. Education institutes should focus on other educational systems which provide better learning outcomes for the students. In this case, Self-learning should be encouraged. To mitigate these challenges, we must investigate other opportunities which will give solutions to these problems.

Children can more easily understand content through a video than by listening to the audio or reading a book. Because it is efficient, entertaining to watch and easily conveys the meaning of the content. So, most people make videos to explain their content. This process takes several steps like summarizing the content, searching images on the internet and other edits to sync the searched images with the audio. Also, it takes several weeks. People do not have much time to spend on video making. There are many AI-based Text-To-Video or Audio-To-Video products in the market. But those systems are not focused on educational purposes and for children. Also, there is not any system to produce video which is made through images from audio with question and answering features.

The growth of interest in data analysis as information retrieval in the education domain becomes unavoidable. In this research we have made a system for kids which can produce a video by combining a set of google images for a text content and a question and answering system. Also, kids can engage with the video and ask questions and get the answers. Anyone without proper training can easily make video content within minutes.

###### RELATED WORKS

To our knowledge, our study is a unique digital video making using google images with a question and answering system. As this research is divided into three major parts every part has a variety of related works.

Text Summarization is the process of reducing the source text into shorter versions to preserve its information content and overall meaning and lexical simplification used to identify the complexity of words and substitute with simple words. The summary should be short and accurate. Many techniques have been developed for summarization of text in various languages. Abstractive summarization techniques are broadly classified into structure based approach and semantic based approach. Structure based methods abstract the most important information through cognitive theories [1]. The sequence to sequence encoder-decoder architecture is the base for sequence transduction tasks. It essentially suggests encoding the complete sequence at once and then using this encoding as a context for the generation of decoded sequence or the target sequence.

The work on sequence to sequence models [3] and seq2seq with neural networks [2] opened up new possibilities for neural networks in natural language processing. In abstractive video summarization, models which incorporate variations of LSTM[13] and deep layered neural networks have become state of the art performers. In addition to textual inputs, recent research in multi-modal summarization incorporates visual and audio modalities into language models to generate summaries of video content. However, generating compelling summaries from conversational texts using transcripts or a combination of modalities is still challenging. The main issue with RNNs lies in their inability to provide parallelization while processing. The processing of RNN is sequential, i.e. we cannot compute the value of the next time step unless we have the output of the current. This makes RNN-based approaches slow. So, these problems are solved in Transformers [4].

Lexical Simplification (LS) [16] is an effective way of simplifying a text because some work shows that those who are familiar with the vocabulary of a text can often understand its meaning even if the grammatical constructs used are confusing to them. The popular lexical simplification approaches were rule-based, in which each rule contains a complex word and its simple synonyms [5], [6], [7].

Artificial intelligent methods and machine learning algorithms are used in automatic information extraction and it achieves high performance. The most basic techniques are syntactic rules and basic Nature Language Processing techniques under all used techniques. Find information from text is extracted by using some syntactic rules and patterns at the word level.

In our project, information extraction is used to find the relevant searching queries which are suitable for the sentence within the domain. To extract the information sentimental analysis, text analysis, semantic analysis is mainly used. A basic task in sentiment analysis is classifying the polarity, motion Detection, Intent Detection; Text Analysis is about parsing texts in order to extract machine-readable facts from them; Semantic analysis is the process of relating syntactic structures, from the levels of phrases, clauses, sentences and paragraphs to the level of the writing as a whole, to their language-independent meanings.

In question answering systems, we should always bear two different kinds of systems: closed-domain systems and open-domain QA systems. Open-domain systems take care of questions about nearly anything, and may solely think about general ontologies and world information. On the opposite hand, closed-domain systems deal with queries under a particular domain (for example, medicine or automotive maintenance), and may exploit domain-specific information by using a model that’s fitted to a unique-domain knowledge, over the years, question answering systems are additionally tailored towards domain-specific because of largely available dataset.

Long Short Term Memory (LSTM) is used widely for NLP tasks. It captures the context-dependent aspects of word meaning. Researchers obtained remarkable accuracy in using LSTM with pre-trained word embeddings like word2Vec, Glove, and contextualized word embeddings Elmo [8],[9],[10]. Even though LSTM is working fine they are doing sequential processing of text. To solve this problem nowadays researchers use Transformers which is the state of the art in NLP tasks. It overcomes the limitation by applying self-attention to computation in parallel. Based on transformers, pre-trained language models are built to fine-tune them for achieving better performance in NLP tasks. Since last year, the field of Natural Language Processing (NLP) has experienced a fast evolution thanks to the development in Deep Learning research and the advent of Transfer Learning techniques. Powerful pre-trained NLP models such as OpenAI-GPT, Elmo, BERT, and XLNet have been made available by the best researchers of the domain [11].

In recent years, BERT models perform very well on complex information extraction tasks. They can capture not only the meaning of words but also the context. To maximize accuracy for your application you’ll want to choose a benchmarking dataset representative of the questions, answers, and contexts you expect in your application.

###### EMPIRICAL STUDY

In our system, initially content modification is done using summarization with lexical simplification and it is responsible for getting the important information from the original content in the simplest form, then the second phase involves information extraction from sentences to produce a video which is a combination of google images using extracted information . Finally, a question answering and information retrieval system has developed. Children can interact with the system by asking questions to retrieve information. This technique is used for effective learning outcomes to identify relevant answers for the questions. It finally identifies the sentence having the right answer for a given question and is able to retrieve specific correct answers from the context.

1. *Content Modification using Summarization with Lexical Simplification and Topic Modeling*

This module is responsible for extracting the important information from the text content by summarizing and identifying the complex words from summarized content and substituting it with simple understandable words. Our main domain is kids. Kids find it difficult to understand complex words. So we have converted the complex words into the simplest form using lexical simplification. Here lexical simplification is done by predicting the complex words (hard vocabulary) and substituting it with simple words. So the output of this module is a simplified summary. So this module is mainly divided into two stages. In the first stage summarization is done to the text content. In the second stage summarized content will have passed to lexical simplifier to identify the complex words and substitute with simple words. So a simple form of summary will be generated at the end of this module.

An abstract summarization method has been used to summarize the content because it produces a summary in a human thinking way. News summary dataset has been used to train the model. News articles contain good vocabulary and grammar which gives greater insights. Summarization is done by transfer learning using neural network and optimizer with the self-attention model. It can be approached by using a model which is trained for a particular task to train our dataset and fine tune the base model for our dataset problem. In transfer learning, we first train a base network on a base dataset and task, and then we repurpose the learned features, or transfer them, to a second target network to be trained on a target dataset and task.

Initially text preprocessing was done in two phases. They are tokenization and dataset creation. The Dataset class was created to accept the Data frame as input and generate tokenized output that is used by the transfer learning model for training. We have used T5 tokenizer to tokenize the data into original text and summarized text columns of the data frame. The Custom Dataset class is used to create 2 datasets, for training and validation. Training Dataset is used to fine tune the model: 80% of the original data**.** Validation Dataset is used to evaluate the performance of the model. The model has not seen this data during training. Train and validation parameters are defined and passed to the pytorch data loader construct to create train and validation data loader. These data loaders will be passed to the train and validation function respectively for training and validation.

In the training phase we have used a self-attention model in transfer learning. After the training is completed, the validation step is initiated. We use the fine-tuned model to generate new summaries based on the article text. The original summary and generated summary are converted into a list and returned to the main function.

Then Lexical Simplification is done to substitute the complex word with simple words. This task has addressed in 3 steps:

**1. Identify complex words in a given sentence:** Create a model that can detect or identify possible complex words, it is called complex word identification (CWI). The Complex Word Identification (CWI) dataset has been used for lexical simplification. It contains additional information about the language level of the annotators for non-native English speakers. The data format is the same as the CWI shared task dataset plus 4 additional columns. The four columns show the number of advanced, intermediate, beginner, and (not-provided) number of annotators for the non-native speakers. The not-provided column shows where the worker does not provide their English language during the experiments.

**2. Generate candidates:** Used BERT´s masked language model to get possible words candidates. One of the tasks BERT was trained for, is to be able to predict the [MASK] word, so if we input into a BERT model, it will output the most probable word given the context. We are masking the complex words of each sentence and get the probability distribution of the vocabulary corresponding to the masked word. We have concatenated the original sequence and the sequence where we replace the complex word with mask token as a sentence pair, and feed the sentence pair into the Bert to obtain the probability distribution of the vocabulary corresponding to the mask word. By using the sentence pair approach,we not only consider the complex word itself, but also fit the contextual information of the complex word.

**3. Select the best candidates based on Zipf values:** Compute the zipf values of each candidate to select the simplest one.

Main objective of this module is to modify the original content. So finally we have integrated stage 1 and stage 2 in a sequential manner to do the content modification. Modified content will have passed to the next module to produce videos.

1. *Information extraction from sentences*

This is to extract the information from the relevant sentence or group of sentences and make a video using Google images. Sentimental analysis, Age categorizes, Text analysis, Semantic analysis help to do this task. The summarized sentences which are from the Content modification module will be analize.

So, make a video for the given audio with the suitable images which are for the particular sentences or words from Google. A basic task in sentiment analysis is classifying the polarity, motion Detection, Intent Detection; For sentiment analysis, going to analyze a sentence according to positive, neutral, negative sentiments[14] and additionally going to categorize the sentences age wise according to the children’s capability.

The first step is the collecting of the dataset. The dataset that we used was collected from Srilanka teachers handbook and students book for the selected age categories. The datasets are divided according to, training datasets and testing datasets into 80:20 ratios. Using this age categorizing can tell that the system made video is suitable for which age group. Most of the schools divide student’s knowledge capability according to Grade 1, 2; Grade 3, 4; Grade 5; Grade 6, 7, 8; Grade 9, 10 ,11; Above grade 11. So, categorize the video according to these age limits. This will help to enhance their knowledge, skills, and therefore enhance the standard of education. According to the majority age limit can get the applicable age. Here limit the age to five main contents like Grade 1, Grade 2, Grade 3, Grade 4, Grade 5. Above 14 years video will be mentioned, and adult contents and violence contents will be ignored from the video.

After that the system does semantic analysis. Semantic analysis is the process of relating syntactic structures, from the levels of phrases, clauses, sentences and paragraphs to the level of the writing as a whole, to their language independent meanings[14]. Through these techniques the system is going to find the relevant searching queries which are suitable for the sentence within the domain. In order to retrieve the image, using google image searcher.

Using selenium, the system can scrap for google images. Selenium is a portable framework for testing web applications. Selenium provides a playback tool for authoring functional tests without the need to learn a test scripting language. Python Imaging Library is a free and open source additional library for the Python programming language that adds support for opening, manipulating, and saving many different image file formats. To do the basic operations like cuts, concatenations, title insertions we can use the Python module MoviePy. It can be used for video composing, creating advanced effects, and video processing. It can read and write the most common video formats. The system will add the images according to the audio file time frame. So the children can easily learn information about the particular sentence with the images while watching the video.

1. *Question answering and Information retrieval System*

In the QA system, the first phase identifies sentences that have the answers in the context. Data was collected from the SQUAD [12] dataset and data was manually annotated using the QA data annotator tool. For predicting the sentence, first, we break the paragraph into multiple sentences using the text blob library. Then, we are using sentence embedding techniques to represent every sentence and question in semantic form. We created vocabulary from the training data and then we trained on an infersent model using that vocabulary. Then, we got the vector representation of each sentence and question using the infersent model. Based on this vector representation of the sentence, we derived the features based on Euclidean distance, cosine similarity by comparing sentence and question vector values. To improve model performance, we derived another feature based on the root match index which is based on comparing the root of the question and answer. We have created another feature that contains values ‘0’ or ‘1’.’1’ indicates that the roots of the sentence and question are equal and ‘0’ represents not equal. If the sentence and question have multiple roots, we compare every root and set the value either ‘1’ or ‘0’.

We consider all the paragraphs to have at least 10 sentences. So, if a paragraph does not have 10 sentences, we replace the cosine similarity value as ‘1’ and the Euclidean distance value as ‘60’, which indicates that it has a maximum deviation from the question vector value. Then using supervised machine learning techniques to predict the sentence which has the answer. Here we defined our target variable as a sentence index. In this research, we used supervised learning approaches to compare performance. We have used different supervised machine learning techniques like Logistic Regression, Gradient Boosting, and Random Forest. Here we split the data as 80% for training and 20% testing purposes.

In the second phase, retrieving answers from the context is implemented using transfer learning techniques. We used the manually annotated question answering dataset. Here we pre-trained the model to improve the performance. We used a distilbert-base-uncased pre-trained model for this task. We tokenized our dataset using hugging face Tokenizer to encode the question and context which was used to create each corresponding paragraph and question to train Bert. We used the PyTorch library to implement our finetuned model. We used a distilbert base model to fine-tune into our custom dataset. Once we trained the model, To estimate our model performance, we used an exact match matrix which is based on how many start and end tokens are predicted correctly.

1. TEST RESULTS OF SUMMARIZATION

| **Model** | **Method** | **RESULTS** |
| --- | --- | --- |
| Summarization | ROUGE-1 SCORE | 80.70% |
| Summarization | Manual evaluation | 74.72% |

1. TEST RESULTS OF LEXICAL SIMPLIFICATION

| **Model** | **Test accuracy** | **Validation accuracy** | **F1 score (weighted)** |
| --- | --- | --- | --- |
| Lexical  simplification | 99.20% | 98.02% | 0.75 |

1. PERFORMANCE EVALUATION OF ML ALGORITHMS

| ML Algorithm | Number of Features | Test accuracy | Validation accuracy | F1 score (weighted) |
| --- | --- | --- | --- | --- |
| Multiclass Logistic Regression | 2 | 56.79% | 58.34% | 0.5723 |
| Multiclass Logistic Regression | 3 | 63.83% | 65.12% | 0.6444 |
| Random Forest Regressor min samples leaf=8 no of estimators=60 | 3 | 66.34% | 68.96% | 0.6757 |
| XGBoost classifier max depth = 3 min child weight = 5 learning rate = 0.2 | 3 | 66.72% | 68.37% | 0.6749 |

###### RESULTS AND ANALYSIS

There are two main phases in the Content Modification module module. First phase outcome is summarized text and the second phase outcome is a simplified summary. Initially evaluation was done separately in two phases. For the summarization, the dataset has been divided into 80:20 ratios which mean 80% data is used for training and 20% data is used for testing purposes. For summarization, evaluation is done using two methods such as automatic evaluation and human evaluation. Automatic evaluation done using ROUGE matrices. ROUGE automatic evaluation compares an automatically generated summary with manual summary created by a professional summarizer. Also manual evaluation has been done using a professional lecturer to check the grammatical errors. The CWI dataset has been divided into training and testing dataset. Here accuracy, ROUGE Score and f1 score were evaluated. Overall evaluation has been done using kids.

In the information extraction module there are two outcomes. First outcome is the Sentimental analysis, Age categorizes, and the second outcome is Text analysis, Semantic analysis. In this project Sentiment is categorized into three groups as positive, neutral, negative and Age categorized into five groups as Grade 1, Grade 2, Grade 3, Grade 4, Grade 5. Evaluated the Sentiment analysis by dividing the training and testing into 80:20 ratios. The training accuracy is 0.9646 and the testing accuracy is 0.6499. And evaluated the final output video manually with the kids who are suitable for that video age group.

In the QA module, the first outcome was based on how many similar sentences were predicted correctly. Evaluation of the model is done through validation matrix scores and manual evaluation. Table III shows the evaluation results. Consider this outcome when we added a feature based on root match the accuracy improved around 5%. Out of these supervised learning techniques, Random Forest Regressor technique performed well. The second phase of the module was based on transfer learning. The model manages to predict the start and end index of the answer around 62.4883% in the exact match metric. The models we created are still not good enough to use in the daily learning activities of students, but with a domain-specific data set, We can increase the performance of each module.

###### CONCLUSION

In this era, Video-based learning is a widely accepted e-learning trend, and it’s gaining momentum. Visually telling a content has become an important factor in education, entertainment, and commerce. Children can more easily understand content through a video than by listening to the audio or reading a book. Because it is efficient, entertaining to watch and easily conveys the meaning of the content. So, most people make videos to

explain their content. This process takes several steps like summarizing the content, searching images on the internet and other edits to sync the searched images with the audio. Also, it takes several weeks. People do not have much time to spend on video making. So, this project proposes a system for kids which can produce a video by combining a set of images for a /text content and a question and answering system. Also, kids can engage with the video and ask questions and get the answers. To develop our system, we gather a dataset from educational sites and teachers who are teaching to children. After that we start our individual part separately, for that we want to get a proper idea about the project and our parts. So, we study the previous research and research papers. Gather good knowledge about the children's education. After that we study about the proper technology to do this research then we start our initial implementation.in this project. For every component we researched the existing literature review and came up with new solutions which enhance

the performance of the already available system. Every part in our system used a machine learning approach and was evaluated using different algorithms. The models we created are still not good enough to use in the daily learning activities of students, but with advanced machine learning approaches and domain specific data set We can increase the performance of each module. We planned to add more features to enhance the learning ability of kids. We planned to use some new techniques to achieve state of the art performance. This research is going to be a promising area in the future.

###### ACKNOWLEDGMENT

We would like to express sincere gratitude to the Faculty of Information Technology, University of Moratuwa, Sri Lanka for providing us guidance to complete this research successfully.

REFERENCES

| [1] | P.E. Genest and G. Lapalme, (2011) "Framework for abstractive summarization using text -to-text generation," in Proceedings of the Workshop on Monolingual Text- To-Text  Generation, pp. 64-73. |
| --- | --- |
| [2] | Kyunghyun Cho, Bart van Merrienboer, Çaglar Gül.ehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN  Encoder-Decoder for Statistical Machine Translation. CoRR abs/1406.1078 (2014).  arXiv:1406.1078 |
| [3] | Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. CoRR abs/1409.3215 (2014). arXiv:1409.3215 |
| [4] | Alamri, F., Kalkan, S., and Pugeault, N., “Transformer-Encoder Detector Module: Using Context to Improve Robustness to Adversarial Attacks on Object Detection”,2020 |
| [5] | M. Lesk, Automatic sense disambiguation using machine readable dictionaries: How to tell a pinecone from an ice cream cone, in: Proceedings of the 5th Annual International  Conference on Systems Documentation, 1986, pp. 24–26. |
| [6] | M. Maddela, W. Xu, A word-complexity lexicon and a neural readability ranking model for lexical simplification, in: EMNLP, 2018, pp. 3749– 3760. |
| [7] | E. Pavlick, C. Callison-Burch, Simple ppdb: A paraphrase database for simplification, in: ACL: Volume 2, Short Papers, 2016, pp. 143–148. |
| [8] | S. Devlin, J. Tait, The use of a psycholinguistic database in the simplification of text for aphasic readers, Linguistic Databases 1 (1998) 161173. |
| [9] | S. Petersen. Natural language processing tools for reading level assessment and text simplification for bilingual education. PhD thesis, University of Washington, 2007. |
| [10] | L. Feng, N. Elhadad, and M. Huenerfauth. Cognitively motivated features for readability assessment. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL), pages 229{237. Association for Computational Linguistics, 2009. |
| [11] | Martin Ponweiser (2012) Latent Dirichlet Allocation in R, Vienna University ofBusiness and Economics. |
| [12] | Rajpurkar, P. (2016, June 16). SQuAD: 100,000+ Questions for Machine Comprehension of Text. ArXiv.Org. https://arxiv.org/abs/1606.05250 |
| [13] | A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, “Learning word vectors for sentiment analysis,” in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, June 2011, pp. 142–150. |
| [14] | T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” CoRR, vol. abs/1301.3781, 2013 |
| [15] | M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” arXiv preprintarXiv:1802.05365, 2018. |
| [16] | W. Coster, D. Kauchak, Simple english wikipedia: a new text simplification task, in: ACL, 2011, pp. 665–669. |