Summarizer:
A Deep Learning Based Approach to Text Summarization

Trishal Singh
Computing Science andEngineering Galgotias University, India
Anurag Sharma
School of Computing Science andEngineering Galgotias University, India
Abhist Kumar
School of Computing Science andEngineering Galgotias University, India

ABSTRACT–
As the information age progresses we will see a huge rise in unstructured data which we can structure and use it to train deep learning models. One of the ways to achieve that is to summarize the data. With the rise of transformer models in the recent years, they have become a goto framework to try to solve deep learning-based problems. In this project we will train several decoder-based transformers and try to find the most efficient model for the task of text summarization with respect to metrics such as accuracy, size, inference time etc. The most efficient model will then be deployed in the form of a web app for everyone to use.

Keywords: *Deep Learning, Natural Language Processing, Transformers.*

1. INTRODUCTION

In this project we will be utilizing various deep learning models for the task of text summarization. The deep learning models used will be mostly language models based on transformer architectures. We will first collect the data from various resources and organize it. Then we will collect the models used for summarization and train the models on the data. We will then deploy the model on the web in a form of web app.
A **Transformer** is a model architecture that eschews recurrence and instead relies entirely on attention mechanisms to draw global dependencies between input and output. Before Transformers, the dominant sequence transduction models were based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The Transformer also employs an encoder and decoder, but removing recurrence in favor of attention mechanisms allows for significantly more parallelization than methods like RNNs and CNNs.

Before going to the Text summarization, first we, have to know that what a summary is. A summary is a text that is produced from one or more texts, that conveys important information in the original text, and it is of a shorter form. The goal of automatic text summarization is presenting the source text into a shorter version with semantics.The most important advantage of using a summary is ,it reduces the reading time. Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. An Abstractive summarization is an understanding of the main concepts in a document and then express those concepts in clear natural language. There are two different groups of text summarization : indicative and informative.Inductive summarization only represent the main idea of the text to the user. The typical length of this type of summarization is 5 to 10 percent of the main text.On the other hand, the informative summarization systems gives concise information of the main text .The length of informative summary is 20 to 30 percent of the main text.

2. LITERATURE REVIEW

Chin-Yew Lin [16] In this paper author introduced RecallOriented Understudy for Gisting Evaluation ROUGE. That is an automatic evaluation package for text summarization. The paper also introduced four different measures of ROUGE: - ROUGE-N, ROUGE-L, ROUGE-
W and ROUGE-S. It measures the quality of summary by comparing the generated summary with other ideal summaries that are created by humans. These methods are efficient for automatic evaluation of single document summary as well as multi-document summaries. Akshil Kumar et al.
[17] In this paper author has analyzed and compared the performance of three different algorithms. Firstly, the different text summarization techniques explained. Extraction based techniques are used to extract important keywords to be included in the summary. For comparison three comparison three keyword extraction algorithms namely Text Rank, Lex

Rank, Latent Semantic Analysis (LSA) were used. Three algorithms are explained and implemented in python language. The ROUGE 1 is used to evaluate the effectiveness of the extracted keywords. The results of the algorithms compared with the handwritten summaries and evaluate the performance. In the end, the Text Rank Algorithm gives a better result than other two algorithms. Pankaj Gupta et al. [18] In this paper author has reviewed different techniques of Sentiment analysis and different techniques of text summarization. Sentiment analysis isa machine learning approach in which machine learns and analyze the sentiments, emotions present in the text. The machine learning methods like Naive Bayes Classifier and Support Machine Vectors (SVM) are used. These methods are used to determine the emotions and sentiments in the text data like reviews about movies or products. In Text summarization, uses the natural language processing (NPL) and linguistic features of sentences are used for checking the importance of the words and sentences that can be included in the final summary. In this paper, a survey has been done of previous research work related to text summarization and Sentiment analysis, so that new research area can be explored by considering the merits and demerits of the current techniques and strategies. Harsha Dave et al. [19] In this paper author has proposed a system to generate the abstractive summary from the extractive summary using WordNet ontology. The multiple documents had been used like text, pdf, word files etc. The author has discussed various text summarization techniques then author discussed step by step the multiple document text summarization approaches. The experiment result is compared with the existing online extractive tools as well as with human generated summaries and shows the proposed system gives good results. At last the author proposed for the future that the summarization accuracy can be increased by comparing this abstractive system with some other abstractive system. Yihong Gong et al. [20] In this research paper the author proposes two methods that create the generic text summaries by ranking and extracting sentences from the main text documents. The first method uses information retrieval (IR) methods that rank the sentence relevance and provides the relevance scores to sentences and the second method uses the latent semantic analysis (LSA) technique that based on latent semantic indexing (LSI) in order to identify the semantic importance of the sentences, for summary creations. The author uses the Singular Value Decomposition (SVD) to generate the text summary. Further, this paper author explains the SVD based methods step by step. The effect of different Weighted Schemes is also checked on the performance of the summaries. The purposed methods provide generic abstractive summaries. Finally, the results are compared with the human-generated summaries. It generates better human like abstractive summaries. For future author proposed to investigate various machine learning techniques so that quality of generic text summarization can be improved. Rada Mihalcea et al. [21] In this paper the author introduced the TextRank a graph-based ranking model for the processing of the text. it is an unsupervised method for keyword and sentence extraction. TextRank uses voting based weighting mechanism and provides the score to the sentence then finally determine the importance of the sentence. The nodes in the graph represent the sentences. The significance of the sentence based on incoming and outgoing edges from nodes. The weight of each is determined based on similarity score between the sentences. TextRank derived from the Google's Page Rank algorithm. TextRank provides extractive summaries of the text. Text Rank Provides the best results. Güneş Erkan et al. [22]In this paper the author introduces

graph-based method LexRank. In this, the sentence score is calculated based on Eigenvector Centrality. It is cosine transform weighting method. In this, the original text is split into sentences and a graph is built where sentences act as the nodes. The complete method is explained in the paper. The results show that LexRank outperforms the existing centroidbased methods. This method is also performed well in case of noisy data. This method generates an extractive summary of the text including the training set.

Kavita Ganesan et al. [23]In this research paper the author proposed graph-based text summarization framework Opinosis. It generates abstractive summaries. Opinosis works on redundant data like human reviews on movies or products and provides abstractive summaries. Firstly, it creates the direct Opinosis-Graph of the text. Where nodes represent the word units of the text. Three unique graph properties: Redundancy capture, Collapsible structures and Gapped subsequence capture is used to explore and explore different sub-paths that help in the creation of abstractive summaries of the text. The valid path is selected and marked with high redundancy score, collapsed path and summary generation. Then all paths ranked in descending order according to scores. The duplicate paths are removed using Jaccard measure the results are compared with human summaries. Results show Opinosis summaries has better agreement with human summaries. For future work author proposed to use a similar idea to overlay parse trees. Dharmendra Hinhu et al. [24] In this paper the author uses the extractive text summarization. The author gives the Wikipedia Articles as input to the system and identifies text scoring. Firstly, the sentences are Tokenized through pattern matching using regular expressions. Then we get data in form of set of words then stop words are removed from the set of words. The words are then stemmed. Then traditional methods are used for scoring of the sentences. Scoring helps in classifying the sentences if they included in summary or not. It is found that scoring sentences based on citation give better results. N. Moratanch et al. [25]In this paper the author presents an exhaustive survey on abstraction based text summarization techniques. The paper presents a survey on two broad abstractive summary approaches: Structured based abstractive summarization and Semantic-based abstractive summarization. The author presents the review of various researches on both approaches of abstractive summarization. The author also covered the various methodologies and challenges, in abstractive s summarization.

N. Moratanch et al. [26]In this paper the author presents the comprehensive review of extraction based text summarization techniques. In this paper the author provides survey on extractive summarization approach by categorized them in: Supervised learning approach and Unsupervised learning approach. Then different methodologies, the advantages are presented in the paper.

3. METHODS AND ALGORITHMS USED:

3.1. Machine Learning method

In this method,the training dataset is used for reference and the summarization process is modeled as a classification problem: sentences are classified as summary sentences and non-summary sentences based on the features that they possess. The classification probabilities

are learnt statistically from the training
data, using Bayes' rule: where, s is a sentence from the document collection, F1, F2...FN are features used in classification. S is the summary to be generated, and P (s∈< S | F1, F2, ..., FN) is the probability that sentence s will be chosen to form the summary given that it possesses features F1,F2...FN

3.2. **Text summarization with neural networks** In this method, each document is converted into a list of sentences. Each sentence is represented as a vector [f1,f2,...,f7], composed of 7 features.Seven Features of a Document

1) f1 Paragraph follows title
2) f2 Paragraph location in document
3) f3 Sentence location in paragraph
4) f4 First sentence in paragraph
5) f5 Sentence length
6) f6 Number of thematic words in the sentence
7) f7 Number of title words in the sentence

The first phase of the process involves training the neural networks to learn the types of sentences that should be included in the summary. Once the network has learned the features that must exist in summary sentences, we need to discover the trends and relationships among the features that are inherent in the majority of sentences. This is accomplished by the feature fusion phase, which consists of two steps: 1) eliminating uncommon features; and 2) collapsing the effects of common features.

3.3 Automatic text summarization using transformers

In this method we use decoder based transformer models such as

a) GPT
b) BART
c) T5

These method proved to be very efficient but had a vast amount of inference time.

4. Libraries used:

1.Pandas 2.torch 3.transformer
4. Streamlit

5. RESULTS

	rouge-1	rouge-2	rouge-l
f	0.459279	0.316947	0.446123
p	0.434198	0.307255	0.422018
r	0.508119	0.344722	0.493260

6. CONCLUSION

As the availability of data in the form of text increasing day by day. It becomes so difficult to read the whole textual data in order to find the required information which is both difficult as well as a time-consuming task for a human being. So, at that time ATS performs an important role by providing a summary of a whole text document by extracting only the useful information and sentences. There are different approaches of text summarization. The real-world applications of text summarization can be: documents summarization, news and articles summarization, review systems, recommendation systems, social media monitoring, survey responses systems. The paper provides a literature review of various research works in the field of automatic text summarization. This research area can be explored more by looking in existing systems and working on different and new techniques of NPL and Machine Learning

7. REFERRENCES

1. **https://arxiv.org/abs/1706.03762**
2. **https://arxiv.org/abs/2005.14165**
3. **https://arxiv.org/abs/1810.04805**