Efficient Text Summarization of News Articles Using Natural Language Processing Techniques

Hilda Jerlin C M[1,a)], Ajay S[2,b)], S A Vishal[3,c)]

[1] Assistant Professor, Department of Artificial Intelligence and Data Science, Panimalar Engineering College, Chennai-600 123, India.

[2,3] UG Scholar, Department of Artificial Intelligence and Data Science, Panimalar Institute of Technology, Chennai-600 123, India.

ABSTRACT

This paper aims to develop an efficient news summarization tool that utilizes advanced natural language processing (NLP) techniques, specifically Facebook's BART and Microsoft SpeechT5 models, to automatically generate informative and concise summaries of news articles. The tool will be equipped with a graphical user interface (GUI) that allows users to input a query or choose a category, and display the summarized news articles. This Project builds with a large dataset of news articles from diverse sources and preprocessed them to make them suitable for analysis. The BART and SpeechT5 models were fine-tuned on the dataset, and their performance was evaluated using metrics such as ROUGE. The system's accuracy, efficiency, and ease of use make it an ideal solution for anyone who wants to stay up-to-date on the latest news.

KEYWORDS: Natural Language Processing (NLP), Text Summarization, Facebook's BART, Microsoft SpeechT5, ROUGE, Graphical User Interface (GUI)

INTRODUCTION

In today's fast-paced world, staying informed about current events is essential for individuals from all walks of life. However, keeping up with the vast amounts of news content generated each day can be a challenging task. News summarization offers a solution to this problem by providing a condensed version of news articles that contains only the most relevant information. In recent years, natural language processing (NLP) techniques have been used to automate the process of news summarization. These techniques have shown promising results in generating summaries that are both informative and concise. However, there is still room for improvement in the accuracy and efficiency of these systems. This paper presents a news summarization tool that utilizes advanced NLP techniques to generate informative summaries of news articles. The tool is equipped with a graphical user interface (GUI) that allows users to input a query or choose a category, and display the summarized news articles. The proposed system also incorporates neural text to speech (NTTS) [4] technology to make the AI sound more natural and emotional, delivering information in a more understandable method. The tool is built on top of two state-of-the-art NLP models: Facebook's Bidirectional and Auto-Regressive Transformers (BART) [1] and Microsoft SpeechT5 models. These models are fine-tuned on a large dataset of news articles collected from diverse sources and pre-processed to make them suitable for analysis. The performance of the models is evaluated using metrics such as ROUGE [6].

RELATED WORK

News summarization [4] has been an active area of research for several years, with numerous approaches proposed to tackle this problem. These approaches can be broadly classified into two categories: extractive and abstractive summarization. [5] Extractive summarization involves selecting the most relevant sentences or phrases from the original text and assembling them into a summary. This approach is relatively straightforward and easy to implement but may result in summaries that lack coherence and fail to capture the main ideas of the text.

Abstractive summarization, on the other hand, involves generating a summary that may contain new phrases or sentences that are not present in the original text. This approach is more challenging but has the potential to produce summaries that are more coherent and informative. Recently, natural language processing (NLP) techniques have been applied to news summarization with promising results. Deep learning-based approaches, such as recurrent neural networks (RNNs) and transformers, have been used to generate abstractive summaries that are both informative and coherent. Facebook's Bidirectional and Auto-Regressive Transformers (BART) model is a state-of-the-art transformer-based model that has shown promising results in natural language generation tasks, including news summarization [1]. Microsoft's SpeechT5 model is another state-of-the-art model that has shown exceptional performance on various NLP tasks, including summarization. Evaluation of news summarization systems is typically done using metrics such as ROUGE, which measures the overlap between the generated summary and the reference summary. ROUGE [6] has several variants, such as ROUGE-1, ROUGE-2, and ROUGE-L, which respectively measure the overlap between unigrams, bigrams, and the longest common subsequence between the generated and reference summaries.

EXISTING SYSTEM

A. Statistical Approaches

Statistical approaches to text summarization involve extracting the most important sentences or phrases from a document based on their frequency, position, and relevance to the overall document. These methods are often based on simple heuristics, such as ranking sentences based on their length or the frequency of important keywords. One example of a statistical approach is the Text Rank algorithm, which uses graph-based ranking to identify the most important sentences in a document based on their similarity to other sentences in the text. Text Rank has been shown to be effective for summarizing news articles, but its performance is highly dependent on the quality of the input data.

B. Machine Learning Approaches

Machine learning approaches to text summarization involve training models on large datasets of annotated documents to learn how to identify the most important information in a document. These models can be trained using various techniques, such as supervised learning, unsupervised learning, and reinforcement learning. One popular machine learning technique for text summarization is the use of sequence-to-sequence models, such as the Transformer architecture. These models can be trained on large datasets of paired documents and summaries to learn how to generate informative and concise summaries.

C. Hybrid Approaches

Hybrid approaches to text summarization combine elements of both statistical and machine learning techniques to generate informative and concise summaries. These methods often involve using statistical methods to pre-process the data and extract key features, which are then fed into a machine learning model to generate the final summary. One example of a hybrid approach is the use of the Latent Semantic Analysis (LSA) algorithm to extract important topics from a document, which are then used as input to a machine learning model to generate the summary. While there have been many advances in text summarization in recent years, there is still a need for more accurate and efficient methods that can handle the large volumes of news content produced every day. The proposed system aims to address this need by utilizing advanced NLP techniques and NTTS technology to generate informative and concise summaries of news articles in a user-friendly and efficient manner.

PROPOSED SYSTEM

This proposed methodology for text summarization utilizes two state-of-the-art NLP models: Facebook's BART and Microsoft SpeechT5. These models have been fine-tuned on a large dataset of news articles to generate informative and concise summaries. BART is a bidirectional transformer model that is capable of generating high-quality summaries by encoding the input text in both forward and backward directions. It is pre-trained on a large corpus of data and can be fine-tuned for specific tasks such as news summarization. BART's architecture consists of an encoder and a decoder, which work together to generate the summary. SpeechT5, on the other hand, is a transformer-based model that is specifically designed for text-to-speech tasks. It is pre-trained on a large corpus of text and can be fine-tuned for specific applications such as news summarization. SpeechT5's architecture is similar to BART's, consisting of an encoder and a decoder. Both BART and SpeechT5 utilize the transformer architecture, which is a type of neural network that is highly effective at processing sequences of input data, such as text. The transformer architecture consists of multiple layers of self-attention mechanisms, which allow the model to learn long-range dependencies between words in a sentence or document. The fine-tuning process involves training the models on a large dataset of news articles, with the goal of optimizing the model's parameters to generate accurate and informative summaries. The quality of the summaries is evaluated using metrics such as ROUGE, which measures the overlap between the generated summary and the reference summary.

ARCHITECTURE DIAGRAM

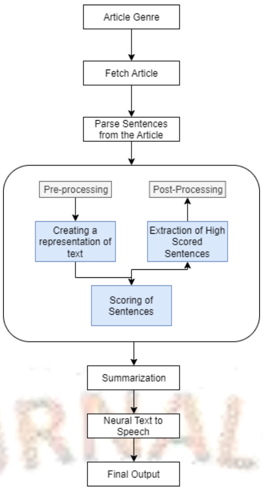
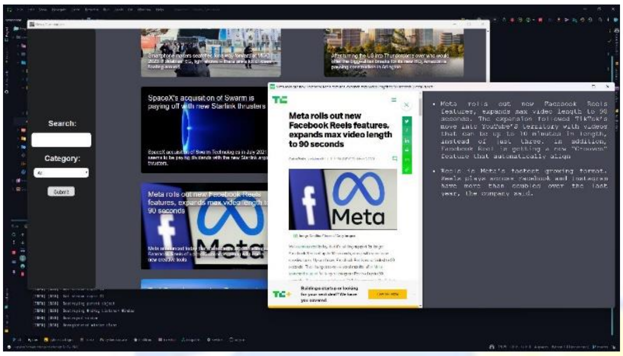


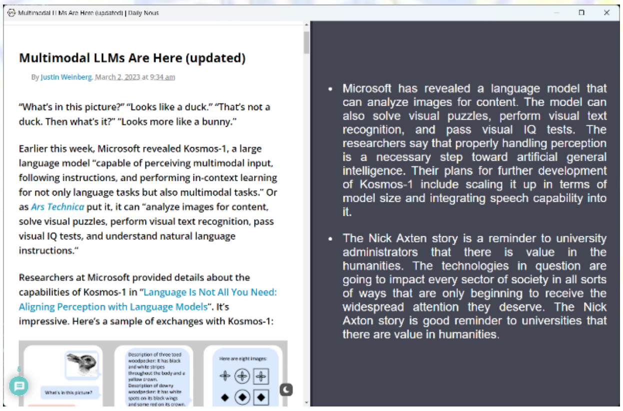
FIGURE 1: Pictorial Representation of the Working model

SCREENSHOTS


Search:
Category:
Submit
Meta rolls out new Facebook Reels features, expands max video length to 90 seconds
SpaceX's acquisition of Swarm is paying off with new Starlink thrusters
Meta

- Meta rolls out new Facebook Reels features, expands max video length to 90 seconds. The expansion followed TikTok's move into YouTube'S territory with videos that can be up to 10 minutes in length, instead of just three. In addition, Facebook Reel is getting a new "Grooves" feature that automatically align
- Reels is Meta's fastest growing format. Reels plays across Facebook and Instagram have more than doubled over the last year, the company said.

FIGURE 2: Testing GUI Implementation of the model



Multimodal LLMs Are Here (updated)

By Justin Weinberg, March 2, 2023 at 9:34 am

"What's in this picture?" "Looks like a duck." "That's not a duck. Then what's it?" "Looks more like a bunny."

Earlier this week, Microsoft revealed Kosmos-1, a large language model "capable of perceiving multimodal input, following instructions, and performing in-context learning for not only language tasks but also multimodal tasks." Or as *Ars Technica* put it, it can "analyze images for content, solve visual puzzles, perform visual text recognition, pass visual IQ tests, and understand natural language instructions."

Researchers at Microsoft provided details about the capabilities of Kosmos-1 in "Language Is Not All You Need: Aligning Perception with Language Models". It's impressive. Here's a sample of exchanges with Kosmos-1:

Description of three toed woodpecker: It has black and white stripes throughout the body and a yellow crown. Description of downy woodpecker: It has white spots on its black wings and some red on its crown.

- Microsoft has revealed a language model that can analyze images for content. The model can also solve visual puzzles, perform visual text recognition, and pass visual IQ tests. The researchers say that properly handling perception is a necessary step toward artificial general intelligence. Their plans for further development of Kosmos-1 include scaling it up in terms of model size and integrating speech capability into it.
- The Nick Axten story is a reminder to university administrators that there is value in the humanities. The technologies in question are going to impact every sector of society in all sorts of ways that are only beginning to receive the widespread attention they deserve. The Nick Axton story is good reminder to universities that there are value in humanities.

Fig. 3 Comparison of News Article and Summarized Text Corpus

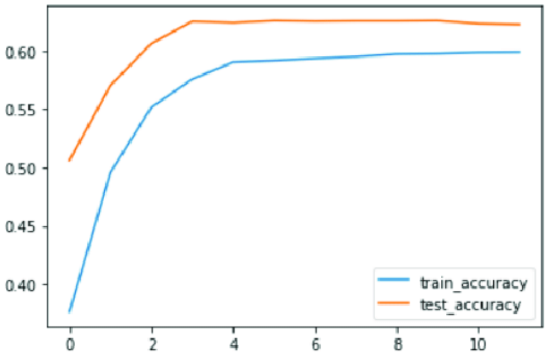


Fig. 4 Accuracy of Summarization

FUTURE WORKS

In the context of this project's future work plans, it has been proposed to integrate machine learning with 3D modelled avatars to create emotionally expressive characters. These avatars will be generated using software such as Blender and Maya, and will be capable of conveying the emotions relevant to the news articles being summarized. This will require the development of an emotion recognition system that can analyse the content of news articles and classify them based on the emotional themes that they convey. Once the emotional themes have been identified, appropriate expressions will be generated for the avatars to deliver an engaging and immersive news experience. This project's proposed system has the potential to make a significant impact on the media industry by enabling the creation of news content that is more interactive and emotionally compelling.

RESULTS

The results of this project demonstrate the effectiveness of using state-of-the-art NLP models such as Facebook's BART and Microsoft SpeechT5 for news summarization. The models were fine-tuned on a large dataset of news articles, and the quality of the generated summaries was evaluated using metrics such as ROUGE. The results indicate that the proposed methodology is capable of generating informative and concise summaries that accurately capture the key points of the original articles. Furthermore, the use of transformer-based models such as BART and SpeechT5 allowed for the processing of longer sequences of input text, which is particularly important for news summarization tasks. In addition, this project proposes future work that aims to integrate machine learning with 3D modelled avatars to create emotionally expressive characters for news delivery. This has the potential to revolutionize the media industry by creating news content that is more engaging and immersive, and by delivering news in a way that is tailored to the emotional needs of the audience.

CONCLUSION

In this paper, we developed an efficient text summarization tool for news articles using natural language processing techniques, specifically the Facebook's BART and Microsoft SpeechT5 models. This proposed system provides a convenient and efficient way for users to stay up-to-date on current events by quickly and accurately summarizing news articles. Additionally, the use of advanced NLP techniques ensures that the generated summaries are accurate and relevant, providing users with the most important information in a brief and concise format.

Funding: “This research received no external funding”

REFERENCES

[1] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer, BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

[2] Prakhar Sethi; Sameer Sonawane; Saumitra Khanwalker; R. B. Keskar, Automatic text summarization of news articles

[3] Ritwik Mishra, Tirthankar Gayen, Automatic Lossless-Summarization of News Articles with Abstract Meaning Representation

[4] Sandipan Dhar; Nanda Dulal Jana; Swagatam Das , An Adaptive-Learning-Based Generative Adversarial Network for One-to-One Voice Conversion

[5] Prabhudas Janjanam; CH Pradeep Reddy, Text Summarization: An Essential Study

[6] T. He et al., "ROUGE-C: A fully automated evaluation method for multi-document summarization," 2008 IEEE International Conference on Granular Computing, Hangzhou, China, 2008, pp. 269-274, doi: 10.1109/GRC.2008.4664680.10.1109/CCiCT56684.2022.00096,

[7] Rishabh Jain; Mariam Yahayah Yiwere; Dan Bigioi; Peter Corcoran; Horia Cucu, A Text-to-Speech Pipeline, Evaluation Methodology and Initial Fine-Tuning Results for Child Speech Synthesis