

Google News Summarization using Transformer based Model

Umesh J¹ and Krishna S²

^{1,2}Department Of Computer Science and Applications, Amrita School Of Computing, Amrita Vishwa Vidhyapeetham, Kollam, India

Email: amscp2csc21063@am.students.amrita.edu, krishnas@am.amrita.edu

Abstract—Automatic summarization of news articles plays a crucial role in information retrieval and consumption. This research paper presents an approach for summarizing Google News articles across four distinct classes: business, sports, entertainment, and technology. We collect a set of news items from these categories by utilizing the T5 Transformer model. The acquired articles are preprocessed and turned into a text-to-text format, where the complete article text is the input and a succinct summary is the desired output. The T5 model may be improved on this dataset to produce precise and useful summaries for each news category. Moreover, the generated summaries are accompanied by links that direct users to the original articles, providing an option for accessing the complete content. Experimental evaluation using established metrics demonstrates the effectiveness of our approach in generating high-quality summaries across different news domains. This research contributes to improving information retrieval and consumption by providing users with succinct summaries while retaining the option to explore the full article for more in-depth information.

Index Terms— Natural Language processing, Deep Learning, Summarization, Transformer (T5 Model).

I. INTRODUCTION

In today's rapidly evolving world, staying informed is of paramount importance. News articles provide valuable insights into current events, business trends, sports updates, entertainment news, and technological advancements. With the advent of digital platforms like Google News, accessing a vast array of news articles has become easier than ever before. However, the sheer volume of available news articles poses a significant challenge for users. It can be overwhelming to navigate through numerous articles, especially when time is limited. Additionally, the absence of concise and informative summaries makes it challenging to grasp the key points across different news categories. To address these challenges, To address these issues, we proposed a system that utilizes the transformative power of the T5 – Text-To-Text Transfer Transformer model to automatically summarize Google News articles. This system aims to generate coherent and concise summaries that capture the essential information present in the original articles. By providing accurate and informative summaries tailored to each news category, the system empowers users to quickly grasp the main points without the need to read the entire article. Furthermore, the system offers users the option to access the complete article through a link associated with each summary, providing a comprehensive and flexible news consumption experience. The main goal of the suggested method is to provide users with succinct and informative summaries

of Google News articles. By employing the T5 model, the system can analyze and comprehend the content of news articles, capturing the key points and important information. The generated summaries serve as condensed representations of the original articles, allowing users to quickly assess the relevance and significance of the content. The system's design involves several key components. First, a data collection mechanism is established to fetch real-time news articles from Google News using an API key from the Newsorg API. This ensures that the dataset is continuously updated and reflects the latest news content.

The collected articles, categorized into business, sports, entertainment, and technology, form the basis for training and evaluating the T5 model. To prepare the dataset for training, a preprocessing step is performed to clean the collected articles. This involves removing unnecessary characters, HTML tags, and formatting inconsistencies that may hinder the model's learning process. The preprocessed articles are then transformed into a suitable format for training the T5 model. This text-to-text format aligns the input as the full article text and the desired output as a concise summary. The training process involves fine-tuning the T5 model using the transformed dataset. During this phase, the model learns to generate summaries that capture the key points and important information from the original article.

II. LITERATURE SURVEY

The goal of news summarizing is to extract the most important information from a lengthy document and condense it into a brief summary, [1] This paper explores how text summarization is becoming more and more crucial as internet text data grows exponentially. Long articles must be manually summarized, which takes time. Therefore, automated approaches are needed to provide accurate and fluid summaries. The two approaches are examined in the paper: abstractive summary, which analyzes the text to produce summaries, and extractive summarization, which chooses pertinent lines from the original material. Using the BBC news dataset, the authors evaluate several pretrained models for text summarization based on transformer architecture[10]. They compare the automated summaries to those created by humans. The findings demonstrate that well-tuned transformer-based models, in particular the T5 model, perform better than others in producing reliable and well-organized summaries.

For text-based tasks in the field of natural language processing (NLP), [3] this research paper gives an in-depth overview on Transformer-based (TB) models. The authors' main goal is to provide an general overview of various TB models while comparing them with the Standard Transformer architecture as a whole. They categorize the models according to their design and training style, describing the benefits and drawbacks of well-liked methods. The report also identifies unexplored areas for investigation and suggests future lines of inquiry to overcome difficulties in TB applications in NLP. The objective is to offer the NLP research community a current overview, highlighting the crucial TB model application and architectural features. The authors look at several TB model groups, including auto-encoding Transformers, auto-regressive Transformers, and S2S Transformers, and describe their architectures.

The research paper [4] examines the performance of the T5 Transformer model on three datasets: CNNDM, MSMO, and XSUM, with an emphasis on abstractive text summarization. By creating summaries that are more fluid and human-like, the goal is to enhance the current summarizing algorithms. The ROUGE and BLEU scores, among other measures, are used in the study to assess the performance of the T5 model. The Author utilized the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) and Bilingual Evaluation Understudy (BLEU) metrics to assess the accuracy of the system generated summary. ROUGE score is a defacto measure for evaluating the automatically generated summaries, whereas BLEU score was initially developed for comparing a source and a reference translation and is now also used to evaluate machine-made summaries. The ROUGE and BLEU scores acquired by the MSMO dataset are higher, at 42.287 and 43.9, respectively, compared against the summaries produced from the CNNDM and MSMO databases.

The findings show that the T5 model excels at providing succinct and cohesive summaries for certain text contents. Comparative research demonstrates that given the MSMO dataset, the T5 model performs best. The research demonstrates that the T5 Transformer model works well for abstractive text summarization across a variety of domains.

The paper discusses the growing demand for text summarization as the amount of text data accessible keeps expanding. [9] Despite the widespread usage of extractive summarizing techniques, abstractive and real-time summarization are becoming more popular. The study on text summarizing is addressed in this work, with a focus on the usage of hybrid and abstract summarization approaches. The paper examines the methodologies, trends, and approaches utilized in text summarization research. It also discusses the challenges and future directions in this field of study. The summarization parameters are classified as follows: Abstractive and

Extractive Summarization, Multi-Lingual, Recommendation Based, Hybrid Method. The article comes to the conclusion that abstractive summarizing allows for advanced skills like paraphrase and generalization, whereas extractive summarization maintain grammar and correctness. Deep learning improvements made recently have showed potential for enhancing abstractive summarization. In general, automated text summarizing helps with information reduction and a number of downstream tasks, and has both academic and commercial uses.

III. PROPOSED SYSTEM

The suggested system uses the T5 (Text-To-Text Transfer) architecture to create an autonomous summarizing system in order to address the challenges brought on by the vast amount of articles published on Google News. The system includes both a backend summarization process and a user-friendly website as the frontend, providing users with a seamless and efficient news consumption experience.

A. Summarization techniques

Text summarizing techniques may be roughly divided into two categories:

Extractive summarization - Extractive summarizing is the process of producing a summary from a given text. The most important phrases and sentences in the text are selected and assessed using a score based on the terms in the sentence.

Abstractive summarization - In this procedure, an interpretation is initially produced by studying the text's content. Based on this comprehension, the machine proposes a summary. By paraphrasing portions of the original source, the content is changed.

B. Transformer-based models

Transformer-based models have become efficient tools for problems involving natural language processing. These models revolutionized the field by introducing self-attention mechanisms, allowing them to capture contextual dependencies efficiently. Transformers use an encoder-decoder architecture, in which the encoder decodes the input text and decoder produces the output text. Various transformer-based models have been developed, each with its unique characteristics and applications. Some notable transformer-based models include [8] BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer), and T5 (Text-To-Text Transfer Transformer).

C. T5 (Text-To-Text Transfer Transformer)

The T5 model, created by Google, is a flexible transformer-based paradigm that has gained a lot of attention in the natural language processing field. T5 uses a text-to-text system in which text strings are used as both the input and the output. This framework [13] allows the model to be fine-tuned on various tasks by framing them as text generation problems. T5 model which achieves the state of the art performance on various NLP tasks [11]; the model was trained on a largescale dataset that was finetuned for numerous downstream tasks. The transformer architecture provides a base for the T5 model. The model was initially trained using transfer learning on a task involving a large corpus before being refined on a downstream task to learn general-purpose tasks like summarization [6], machine translation, and caption generation. The five T5 varieties that have been proposed are T5-small, T5- base, T5-large, T5-3b, and T5-11b, in this paper, we have used the T5-base version. The T5 model's architecture consists of an encoder-decoder framework, where the encoder processes the input text, and the decoder generates the output text. It employs multi-head self-attention layers, enabling it to capture the relationships between words and generate coherent and contextually relevant summaries. The T5 model's extensive pre-training and fine-tuning make it a powerful tool for text generation tasks, including summarization.

IV. METHODOLOGY

The proposed system address the challenges posed by the immense amount of articles on Google News by developing an automatic summarization system [12] using the T5 (Text-To-Text Transfer) model. The system includes both a backend summarization process and a user-friendly website as the frontend, providing users with a seamless and efficient news consumption experience. The backend component focuses on collecting a diverse and representative dataset of Google News articles across four categories: business, sports, entertainment, and technology. These articles undergo preprocessing to ensure data quality, consistency, and appropriate formatting. The transformed dataset is then fine-tune by the T5 model, enabling it to generate accurate and informative summaries tailored to each news category. The core of the summarizing process is the trained model. On the frontend, the user-friendly website is developed to facilitate user interaction with the system. Users are presented

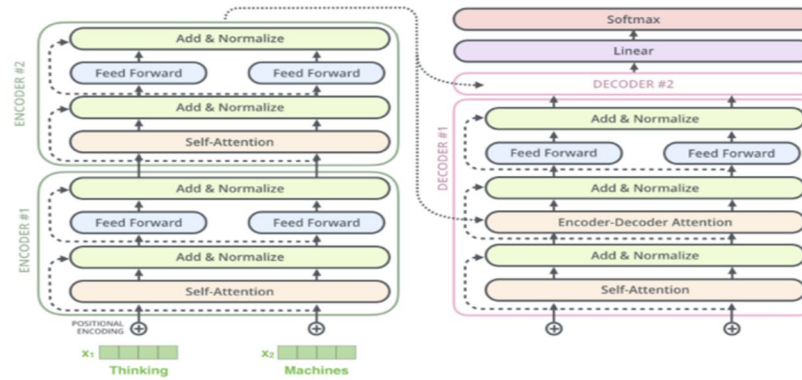


Fig. 1. Encoder-Decoder Architecture

with a visually appealing and intuitive interface that allows them to choose from the available news categories. When a specific category, such as business, is selected, the website displays a page showcasing a list of ten news articles relevant to that category. On clicking on a news article, the backend summarization process is triggered. The system retrieves the full article using the provided link or via an API integration with the source platform. The system consists of two main components: the User Interface and the Backend Processing. The User Interface encompasses the frontend components, including HTML, CSS, JavaScript, and web frameworks, which provide an interactive and user-friendly interface for users to interact with the system. The User Interface allows users to select news categories, browse articles, and view generated summaries.

The Backend Processing handles the core functionalities of the system. It includes several sub-components:

- **Data Collection Preprocessing:** This component is responsible for collecting real-time news articles from Google News using the Newsorg API. The collected articles undergo preprocessing, including data cleaning and formatting, to ensure consistency and quality for further processing.
- **T5 Model (Text To Text Transfer):** The T5 model, a transformer based model, is the key component for generating summaries. It takes the preprocessed news articles as input and generates concise and informative summaries. On the C4 dataset, T5 is first pre-trained using an encoder-decoder architecture for the denoising, corrupting span aim. C4 Dataset (Colossal Clean Crawled Corpus). The cleaned Common Crawl version dataset is around 700GB in size. The cleaned meaning that cleaning in the context of extracting just English content, eliminating code lines, deduplicating, and so on. It's a high quality pre-processed English language corpus. Using a supervised objective with appropriate input modeling for the text-to-text setting, the downstream tasks are then fine-tuned.
- **Generated Summaries:** The output from the model is the generated summaries for each news article. These summaries capture the main points and essential information from the original articles, enabling users to quickly grasp the key content without reading the entire article.
- **The Backend Processing communicates with the User Interface,** delivering the generated summaries and the associated full article links. The User Interface displays the summaries and provides options for users to view the full articles or navigate through different news categories. The system architecture ensures a seamless flow of data and interaction between the frontend and backend components. Users can easily browse articles, access summaries, and view full articles, enhancing their news consumption experience.
- **Dataset -** In this data collection is performed using the Google News API, offering access to a huge collection of articles from various sources. By leveraging the API, the system can retrieve articles based on specific criteria such as category, keywords, date ranges, and filters. The articles are collected across four distinct news categories: business, sports, entertainment, and technology. The collected articles undergo data preprocessing to clean and standardize the data. This involves removing noise such as HTML tags, special characters, and formatting inconsistencies. The methods of text normalization such as stemming and lower-casing may also be applied to promote consistency and reduce vocabulary variations within the dataset.

V. RESULT

Following Figures shows the result of summarizing Google News articles using the T5 model yielded promising results in terms of generating accurate and informative summaries. The system was able to provide concise

```

Not Found: /cdn-cgi/challenge-platform/h/g/scripts/jsd/19b997cb/invisible.js
[07/Jul/2023 12:15:30] "GET /cdn-cgi/challenge-platform/h/g/scripts/jsd/19b997cb/invisible.js HTTP/1.1" 404 3383
*** Business News ***
-----
Title: Griff's Hamburgers closing in Albuquerque - KOAT New Mexico
Source: KOAT New Mexico
-----
Title: Threads: Instagram launches app to rival Twitter - BBC
Source: BBC News
-----
Title: DoorDash driver seen cursing at customer for 25% tip - NBC News
Source: YouTube
-----
Title: BlackRock's Bitcoin spot ETF will 'democratize' crypto: CEO Larry Fink - Fox Business
Source: Fox Business
-----
Title: What Does An FAA Certification Mean For Flying Cars? - SlashGear
Source: SlashGear
-----
Title: The 20+ best Amazon device deals for early Prime Day: Save up to 75% on Fire TVs and more - Yahoo Life
Source: Yahoo Entertainment
-----
Title: Dow Jones Futures: Stocks Fall, Google, CRM Flash Buy Signals, Meta Launches Twitter Rival, Instagram Threads | Inve
siness Daily
Source: Investor's Business Daily

```

Fig. 2. Dataset sample

summaries that captured the main points of the original articles across different news categories, including business, sports, entertainment, and technology. The results showcased the effectiveness of the T5 model in capturing contextual dependencies and generating coherent summaries.

Our findings indicate that our Google News summarization system may efficiently recommend the summary of the articles to the user. The system is capable of producing good results. Our finding indicates that our suggested system can summarize the news articles extracted from google news and provide the relevant points summary, by that user, can understand the overall concept of the news article without going through the entire news article.

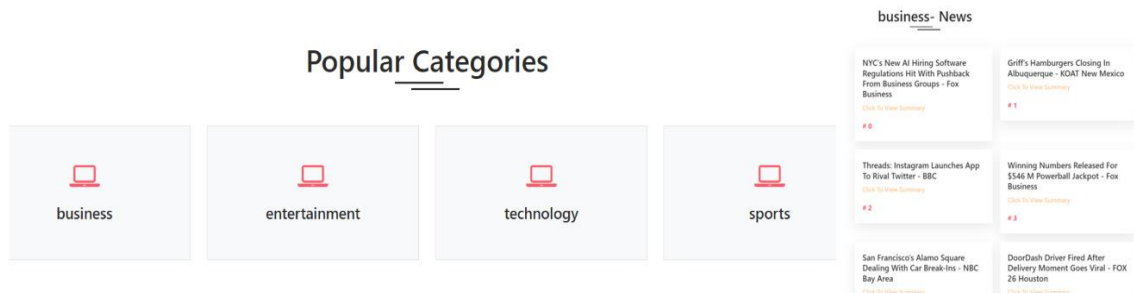


Fig. 3. Categories Of News and Classified News

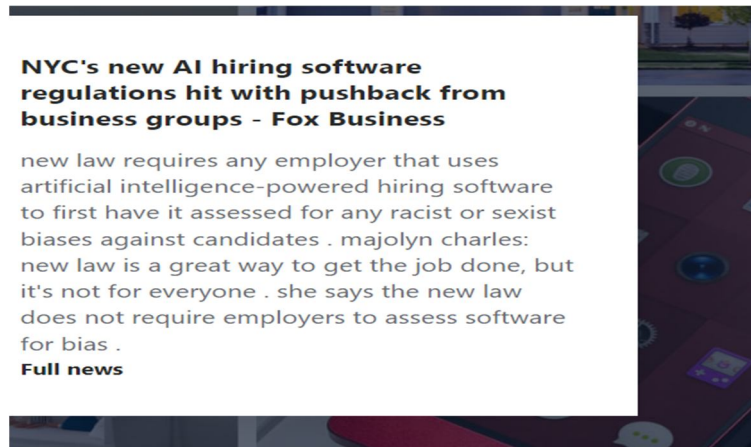


Fig. 4. Automated Summary Generated

VI. RESULT ANALYSIS

The system for summarizing Google News articles using the T5 model achieved successful results in terms of generating accurate summaries and providing a user-friendly news consumption experience. The system fetched

news articles in real-time from Google News, utilizing an API key from the Newsorg API. These articles were classified into different categories, such as business, sports, entertainment, and technology, enabling users to easily browse and access articles of their interest.

The summarization process implemented by the T5 model proved to be effective in generating concise and informative summaries for each article. The summaries captured the main points and key information from the original articles, allowing users to quickly grasp the essential content without having to read the entire article. This contributed to saving users' time and effort while still providing them with the necessary information for decision-making and staying up-to-date with the latest news trends. The ability to view the full news article when required added flexibility to the news consumption process. Users had the freedom to access additional information or explore different perspectives by accessing the original article directly from the system. This feature catered to users with varying levels of interest and allowed them to dive deeper into specific topics when desired.

In the Following Tables shows the News fetched from google news, The title publisher and the automated news summary generate by the system:

TABLE I: AUTOMATED SUMMARY GENERATED ON BUSINESS NEWS

News Article	Title-Publisher	Automated-Summary
Instagram's Threads app, a text-based social media platform poised to become Twitter's latest competitor, is now available to users in more than 100 countries. While most features mimic those of Twitter, its user interface design resembles Instagram's, with the same heart, comment and share buttons and similarly placed tabs	What You Need To Know About Threads, Instagram's New Twitter Competitor - NBC News	most features mimic those of Twitter, with the same heart, comment and share buttons . threads joins a growing list of ...
The U.S. saw an increase in auto sales during the second quarter despite higher prices for customers. Auto sales rose 16.8 to just over 4.1 million between April and June, primarily fueled by stifled demand from nearly two years of factories worldwide suffering from a computer chip shortage. Ivan Drury, director of insights for Edmunds.com, told the Associated Press that the trend in lower average prices and automaker discounts for most of the past year began to stall out in June as buyers paid an average of 45,978 for a vehicle, according to J.D. Power estimates. Inventory on dealer lots was expected to be a bit over 1.2 million in June, J.D. Power showed.	US new vehicle sales continue on upward trend during second quarter - Fox Business	auto sales rose 16.8 percentage to just over 4.1 million between April and June . the increase was primarily fueled by higher prices for customers

VII. CONCLUSION

In this study, we proposed a news summary system that can summarize the news article that is fetched from google news. The proposed system successfully addressed the challenges of information overload and the absence of concise summaries in Google News articles. By leveraging the power of the T5 model, accurate and informative summaries were generated for different news categories. The integration of the Newsorg API ensured real-time access to reliable news sources. The user-friendly interface facilitated seamless browsing, and the option to view the full article provided flexibility. Overall, the system showcased the potential of transformer-based models, specifically the T5 model, in automating the summarization of Google News articles. By leveraging the power of these models, users were empowered to efficiently retrieve and consume relevant news content tailored to their interests. The combination of advanced natural language processing techniques, a well-designed user interface, and the ability to view the full article when desired created a comprehensive news consumption experience. As for Further enhancements the summarize system could be made to personalize the summarization process based on user preferences and feedback. Additionally, exploring ways to expand the dataset and fine-tune the T5 model on a larger and more diverse corpus could lead to even more accurate and contextually relevant summaries.

REFERENCES

- [1] Gupta, Anushka Chugh, Diksha Anjum, Anjum Katarya, Rahul. (2021). Automated News Summarization Using Transformers.

- [2] Sirohi, Neeraj Rajshree, Mamta Rajan, Siddhi. (2021). Text Summarization Approaches Using Machine Learning LSTM. *Revista Gestao Inovac, ~ ao e Tecnologias*. 11. 5010-5026. 10.47059/revistagein-~tec.v11i4.2526.
- [3] Rahali, A.; Akhloufi, M.A. End-to-End Transformer-Based Models in Textual-Based NLP. *AI* 2023, 4, 54–110. <https://doi.org/10.3390/ai4010004>
- [4] T, Tawmo and Borah, Mrinmoi and Dadure, Pankaj and Pakray, Partha, Comparative analysis of T5 model for abstractive text summarization on different datasets (April 29, 2022). *Proceedings of the International Conference on Innovative Computing Communication (ICICC) 2022*
- [5] Zhang, J., Zhao, Y., Saleh, M., Liu, P. (2020, November). Pegasus: Pretraining with extracted gapsentences for abstractive summarization. In *International Conference on Machine Learning* (pp. 11328-11339). PMLR
- [6] Sunil, S., Sonu, P., "An Effective Approach for Detecting Acute Lymphoblastic Leukemia Using Deep Convolutional Neural Networks," In: Mandal, J.K., Hsiung, P.A., Sankar Dhar, R. (eds) *Topical Drifts in Intelligent Computing*. ICCTA 2021. *Lecture Notes in Networks and Systems*, vol 426. Springer, Singapore. <https://doi.org/10.1007/978-981-19-0745-6> Publisher : Springer, Singapore,(2022).
- [7] Jose, A., Harikumar, S. (2022). COVID-19 Semantic Search Engine Using Sentence-Transformer Models. In: Raman, I., Ganesan, P., Sureshkumar, V., Ranganathan, L. (eds) *Computational Intelligence, Cyber Security and Computational Models. Recent Trends in Computational Models, Intelligent and Secure Systems*. ICC3 2021. *Communications in Computer and Information Science*, vol 1631. Springer, Cham
- [8] S. R. K. Harinatha, B. T. Tasara and N. N. Qomariyah, "Evaluating Extractive Summarization Techniques on News Articles," 2021 International Seminar on Intelligent Technology and Its Applications (ISITIA), 2021, pp. 88-94, doi: 10.1109/ISITIA52817.2021.9502230.
- [9] Sara Tarannum¹, Piyush Sonar², Aashi Agrawal³ Krishnai Khairnar⁴, "NLP based Text Summarization Techniques for News Articles: Approaches and Challenges", *International Research Journal of Engineering and Technology (IRJET)* Volume: 08 Issue: 12 — Dec 2021
- [10] S. Thara and P. Poornachandran, "Transformer Based Language Identification for Malayalam-English Code-Mixed Text," in *IEEE Access*, vol. 9, pp. 118837-118850, 2021, doi: 10.1109/ACCESS.2021.3104106.
- [11] Dr. Sumathi, Dr. Suthir S, May 2022– "A comparative analysis of transformer based models for figurative language classification", *Computers Electrical Engineering* (0045-7906), Elsevier, <https://doi.org/10.1016/j.compeleceng.2022.108051>
- [12] Rahesh, R., V. V. Sajith Variyar, Ramesh Sivanpillai, V. Sowmya, K. P. Soman, and Gregory K. Brown. "Segmentation of Epiphytes in Grayscale Images Using a CNN-Transformer Hybrid Architecture." In *Data Engineering and Intelligent Computing*, pp. 119-129. Springer, Singapore, 2022.
- [13] Automated Relational Triple Extraction from Unstructured Text using Transformer", Akshay Hari and Priyanka Kumar presented at 3rd International Conference on Electrical and Electronics Engineering (ICEEE 2022).