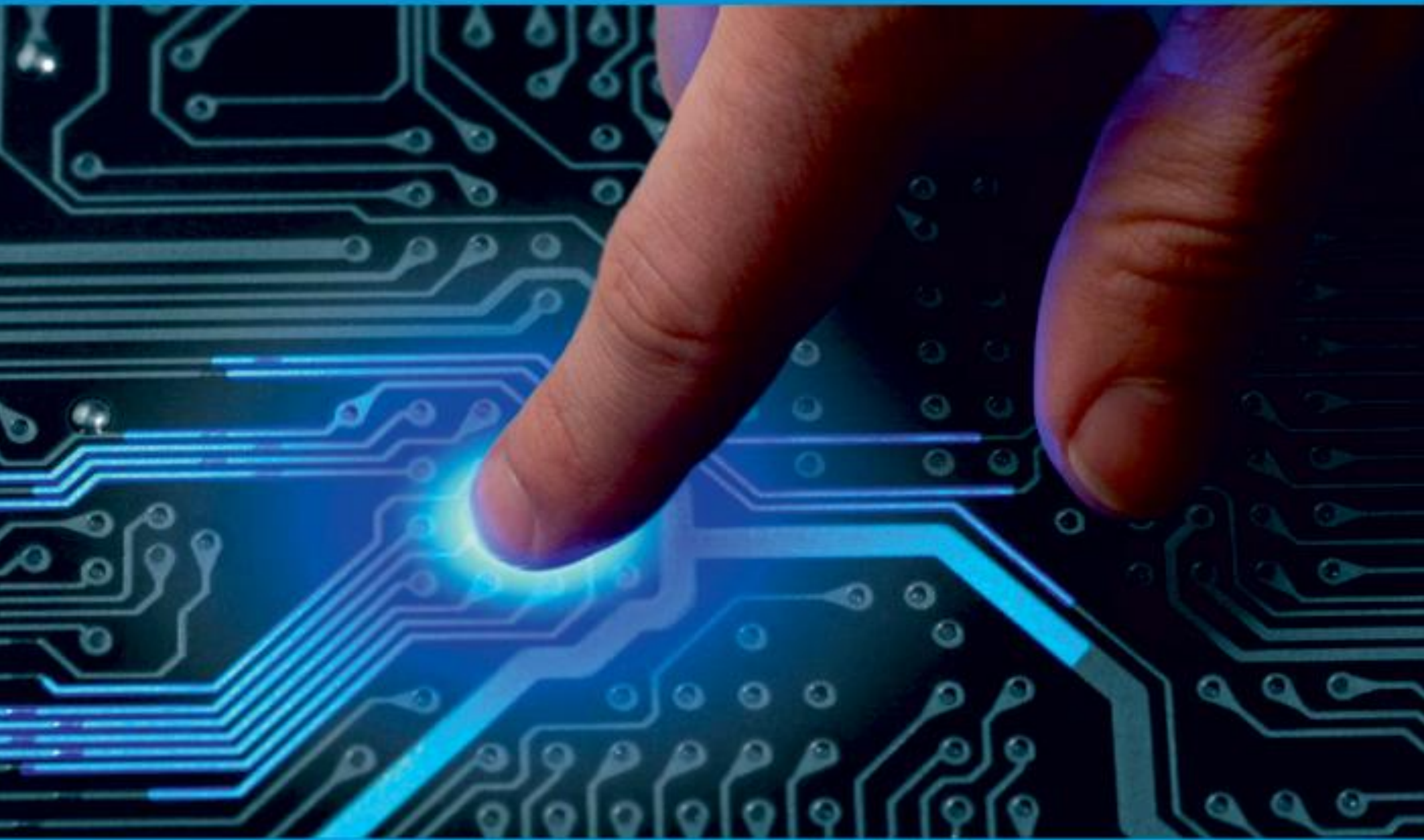
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 3, March 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 3, March 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1003076 |

Text Summarization of News Headline using Natural Language Processing

Ms. Bhavika Mulwani, Mr. Praveen Mirchandani, Mr. Sameer Israni, Prof. Dr.Dashrath Mane

Student, Department of Computer Science, Vivekanand Education Society Institute of Technology, Mumbai, India

Student, Department of Computer Science, Vivekanand Education Society Institute of Technology, Mumbai, India

Student, Department of Computer Science, Vivekanand Education Society Institute of Technology, Mumbai, India

Professor, Department of Computer Science, Vivekanand Education Society Institute of Technology, Mumbai, India

ABSTRACT: Text Summarization implies extracting texts and paragraphs into a smaller report, decreasing the content of the original text and at the same time keeping prime information and giving a vague description on what the article is. Text reading is a long and strenuous task, text summarization is becoming popular and thus the inclination for research. In this project, we will perform the project of Natural Language Processing to summarize text with Machine Learning algorithms .In our day to day life, there are various purposes for text summarization in different domains such as news synthesis, reviews of products on e-commerce websites, legal text synthesis, medical reports which can be accomplished with text summarization. The objective to summarize a text is to construct a factual and fluid summary containing only the important points expressed in the document..

KEYWORDS: Natural language processing, summarization, abstractive summarizer, extractive summarizer, information retrieval, PageRank, T5, BART, Pegasus.

I. INTRODUCTION

“I don't want to read the entire description of this product, I wish there was a summarized version of it”. We often found ourselves in these situations. We make a thorough report and the professor only has time to read the summary or we want to get our daily news but don't have time to read the whole newspaper or articles but get a short description on what has happened or we want to buy a product but don't want to read the whole description of that product. In this project we do exactly that and summarize the long documents for ease for someone who only wants the main features of the document rather than the whole document. There is a huge amount of data appearing digitally, so it is necessary to develop a unique procedure to immediately summarize long texts while keeping the main idea. Text summarization also makes it possible to shorten the reading time, speed up information searches and obtain as much information as possible on a subject. It is essential to learn what are the types of text summarization to understand how the process works.

A. TYPES OF SUMMARIZATION

Text Summarization is divided into: Extractive and Abstractive. Both these approaches are as follows

1. EXTRACTIVE APPROACH

The Extractive approach takes sentences directly from the document according to a function to form a connected summary. This approach operates by identifying the important sections of the text, sniping and collecting parts of the content to produce a concise version. Here, no new text is generated, only existing text is used. Although the methodologies for extractive summarization differ, they all have the same core goals:

1. Construct a representation of the input text (text to be summarized)
2. Based on the created representation, assign a score ‘m’ to the sentences.
3. Choose a summary that includes the top m most important sentences.

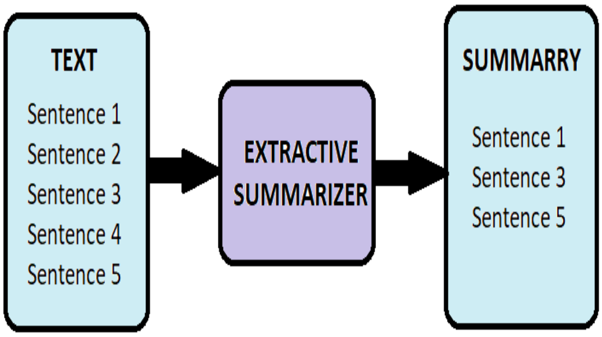


Fig 1. Working of Extractive Text Summarization

2. ABSTRACTIVE APPROACH

The Abstractive approach intends to manufacture a summary by interpreting the text using advanced NLP techniques to generate a new, smaller text –some content may not appear in the original document, which conveys the most information. It can be divided into two parts:

1. Structured Based Approach: The structure-based technique uses cognitive schemas like ontology, tree, lead, and body phrase structure to translate the most important information from the content.
2. Semantic Based Approach: The semantic representation of the document is processed by the natural language generation (NLG) system. By studying linguistic data, this approach targets noun and verb phrases.

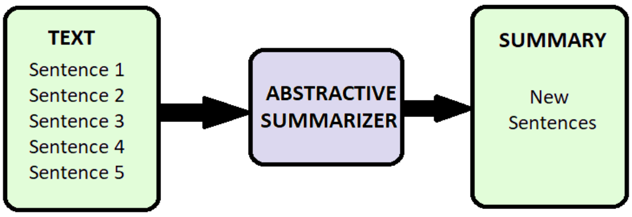


Fig 2. Working of Abstractive Text Summarization

II. LITERATURE SURVEY

Seq2seq model along with its advanced version LSTM is used for summarization in paper [2] to increase the accuracy of the model. The proposed model uses a seq2seq model with attention mechanism and LSTM and GRU cells are used to make accurate predictions for the summary. For classification model concept net Number batch word embedding model is used and during classification 1D convolution layer followed by max pooling layer. The results were satisfactory and the summary which model produced got better after every training.

Paper [3] proposes the Firefly algorithm for multiple text summarization. ROUGE score is used to depict the performance of the algorithm. Topic relation, cohesion and readability factor are used as fitness functions in the proposed algorithm. Fitness function calculates the score of every stanza and the stanza with the best score is used for summary. The performance of the algorithm surpasses particle swarm optimization (PSO) and genetic algorithm (GA).

Latent Semantics analysis (LSA) based summarization algorithms are proposed in paper [4] and the performances are characterized based on the ROUGE scores. In this paper the algorithm is tested on 2 different language documents and results for both were equally good. Cross method used worked better than other LSA based approaches but does not work well on shorter documents. These approaches are extractive and do not perform as successfully as ML algorithms.

References	**Algorithm**
Paper[2]	Seq2Seq model with LSTM
Paper[3]	Firefly
Paper[4]	Latent Semantic Analysis (LSA) based algorithms

Table 1. Algorithms used in reference papers

III. PROPOSED SYSTEM

A. MODULAR DIAGRAM

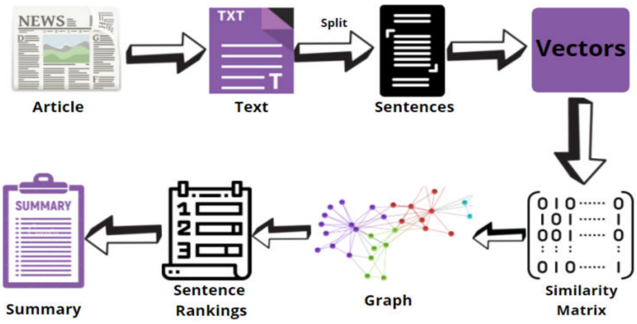


Fig 3. Working of proposed Extractive text summarizer

Through this modular diagram, Procedure for extractive summarization divided into 5 steps

1. One or many documents will be converted to a single text file.
2. Text file will be splitted into a list of sentences.
3. Pre-processing and Text Vectorization.
4. After vectorization, a similarity matrix is drawn out using cosine similarities between sentences.
5. A graph is drawn out which is used by the Text Rank algorithm to summarize the article.

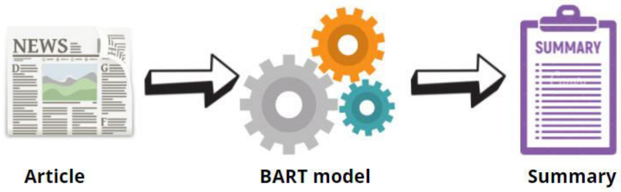


Fig 4. Working of proposed Abstractive text summarizer

A. TEXTRANK EXPLANATION

TextRank is an unsupervised graph-based text summarization technique. PageRank is an algorithm used to calculate rank of web pages, which is used by Google search engine. TextRank is based on the PageRank Algorithm. In place of web pages, text sentences are used for ranking. Similarity matrix is drawn by finding cosine similarities between sentences. Then using the PageRank algorithm, sentences are ranked and top N sentences are chosen for summary.

B. T5 EXPLANATION

T5 is a transformer model from Google. Text-to-Text Transfer Transformer is a transformer based architecture that uses a text-to-text approach. It conquers the latest outcomes on non-identical NLP tasks. It is an abstractive summarization algorithm.

C. PEGASUS EXPLANATION

PEGASUS uses an encoder-decoder model for sequence-to-sequence learning. Here, first the encoder will take the context of input and encode it into a vector (context vector), which is a numerical representation of input. Then, this context vector is given to the decoder which will decode it to generate gap sentences.

D. BERT EXPLANATION

Bart is a noise canceling auto encoder for pre-training seq2seq models. It is a constructive method for text generation, hence is good for abstractive summarization.

IV.EXPERIMENTAL SETUP AND OUTCOMES

A. EXPERIMENTAL SETUP

The dataset used is a public dataset, provided at the address given in reference [1]. The dataset was primarily cleaned and refined using numpy and pandas libraries. Data manipulation libraries like numpy, pandas were used for exploratory data analysis.

For abstractive summary, the dataset used is provided in reference [1]. Here three different models are used namely T5, Pegasus and Bart transformers with HuggingFace. Accuracy for all three models is generated, and we can compare that BART model has the best results among three models. For implementing the model, using the Django framework, a web application is made. The user is given features to enter text to summarize. The web app will then ask the user to choose between abstractive and extractive summarization or the user can compare both summaries.

B. DATASET

The publicly available news dataset [1], delivers short summaries of news from around the web. In this dataset, one will find headlines and summaries of news items along with their sources. .The dataset comprises 5 columns namely, Headline, Short, Source, Published date, time. From this we have only used Headline and short, where short depicts short news articles. Dataset consists of 55,104 rows.

C. WEB APPLICATION

To deploy the working of the text summarization, a web application was developed. This web application enables users to upload their text which is to be summarized and asks the user if they want an abstractive or extractive summary or compare both summaries. Fig 5. Shows the web application page where a user can paste their text. Fig 6. Shows the output for both, abstractive as well as extractive summaries.

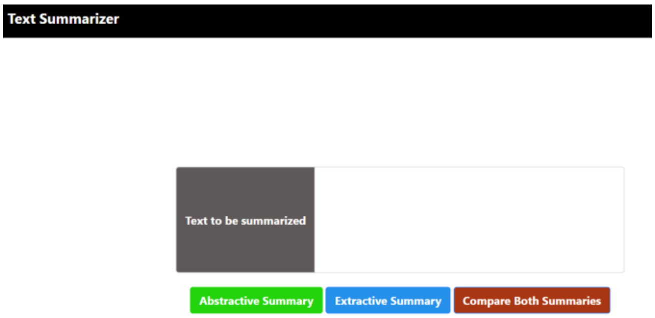
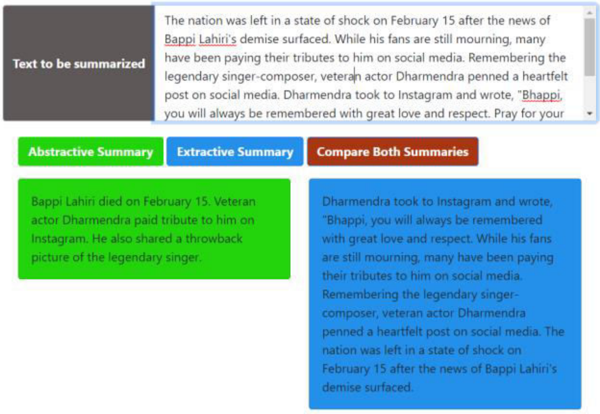


Fig 5. Home Page



The nation was left in a state of shock on February 15 after the news of Bappi Lahiri's demise surfaced. While his fans are still mourning, many have been paying their tributes to him on social media. Remembering the legendary singer-composer, veteran actor Dharmendra penned a heartfelt post on social media. Dharmendra took to Instagram and wrote, "Bappi, you will always be remembered with great love and respect. Pray for your

Bappi Lahiri died on February 15. Veteran actor Dharmendra paid tribute to him on Instagram. He also shared a throwback picture of the legendary singer.

Dharmendra took to Instagram and wrote, "Bappi, you will always be remembered with great love and respect. While his fans are still mourning, many have been paying their tributes to him on social media. Remembering the legendary singer-composer, veteran actor Dharmendra penned a heartfelt post on social media. The nation was left in a state of shock on February 15 after the news of Bappi Lahiri's demise surfaced.

Fig 6. Output of an image showing both summaries

D. PERFORMANCE EVALUATION

To evaluate the performance of abstractive summaries by different models, a dataset was apportioned. The total number of examples the model was tested on is 55000 and the performance was evaluated by the Rouge scoring algorithm, i.e., it assesses the likeness among a summary by model and a reference summary (given in dataset).

$$\text{ROUGE-N}_{\text{single}}(\text{summary\_by\_model}, \text{reference}) = \frac{\sum_{r\,i\,\in\,\text{reference}} \sum_{\text{n-gram}\in\, r\,i} \text{Count (n-gram, summary\_by\_model)}}{\sum_{r\,i\,\in\,\text{reference}} \text{numNgrams}\,(r_i)}$$

After evaluating performance of various algorithms, it was found that T5 outperformed the other models in terms of score. Despite the fact that the Pegasus model was implemented faster than BART, BART outperformed it in terms of performance.

V. CONCLUSIONS

Text summarization is an interesting area of ML that is gaining traction. It is the process of identifying the most important and meaningful information in a document. We are designing a project that will save users time by helping people get summary information from a document instead of reading the whole big document and finding meaningful points. This project tends to address this summary problem. This project will help users to quickly understand the main characteristics of a topic without having to read through the whole article and will save some time. It is used to summarize news articles in summary and extract format so that users can gain insight into the content of the news. For abstract methods, BART outperforms other models in terms of performance.

REFERENCES

1. https://www.kaggle.com/shashichander009/inshorts-news-data
2. Dutta, M., Das, A. K., Mallick, C., Sarkar, A., & Das, A. K. (2019). A graph based approach on extractive summarization. *Advances in Intelligent Systems and Computing*, *813*, 179–187. https://doi.org/10.1007/978-981-13-1498-8_16
3. Boorugu, R., Ramesh, G., & Madhavi, K. (2019). Summarizing product reviews using NLP based text summarization. *International Journal of Scientific and Technology Research, 8(10), 1127–1133.*
4. Tomer, M., & Kumar, M. (2021). Multi-document extractive text summarization based on firefly algorithm. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2021.04.004
5. Ozsoy, M. G., Alpaslan, F. N., & Cicekli, I. (2011). Text summarization using latent semantic analysis. *Journal of Information Science, 37(4), 405–417.* https://doi.org/10.1177/0165551511408848
6. Munot, N., & S. Govilkar, S. (2014). Comparative Study of Text Summarization Methods. *International Journal of Computer Applications*, *102*(12), 33–37. https://doi.org/10.5120/17870-8810
7. PadmaPriya, G., & Duraiswamy, K. (2014). An approach for text summarization using deep learning algorithms. *Journal of Computer Science*, *10*(1), 1–9. https://doi.org/10.3844/jcssp.2014.1.9
8. Haque, M. M., Pervin, S., & Begum, Z. (2013). Literature Review of Automatic Single Document Text Summarization Using NLP. In the International *Journal of Innovation and Applied Studies* (Vol. 3, Issue 3). http://www.issr-journals.org/ijias/
9. Steinberger, J., & Ježek, K. (2009). EVALUATION MEASURES FOR TEXT SUMMARIZATION. In *Computing and Informatics* (Vol. 28).
10. Kaestner, C., Neto, J. L., Freitas Celso, A. A., & Kaestner, A. A. (n.d.). *Automatic Text Summarization Using a Machine Learning Approach*. http://www.ppgia.pucpr.br/~alex
11. Saziyabegum, S. (n.d.). REVIEW ON TEXT SUMMARIZATION EVALUATION METHODS. */ Indian Journal of Computer Science and Engineering (IJCSE)*, *8*(4). http://www.pritisajja.info/
12. Yeasmin, S., Basak Tumpa, P., Mahjabin Nitu, A., Palash Uddin, M., Ali, E., & Ibn Afjal, M. (2017). Study of Abstractive Text Summarization Techniques. *American Journal of Engineering Research (AJER)*, *6*, 253–260. www.ajer.org
13. Yadav, A. K., Maurya, A. K., Ranvijay, & Yadav, R. S. (2021). Extractive text summarization using recent approaches: A survey. In *Ingenierie des Systemes d'Information* (Vol. 26, Issue 1, pp. 109–121). International Information and Engineering Technology Association. https://doi.org/10.18280/isi.260112
14. Andhale, N., & Bewoor, L. A. (2017, February 21). An overview of text summarization techniques. *Proceedings - 2nd International Conference on Computing, Communication, Control and Automation, ICCUBEA 2016*. https://doi.org/10.1109/ICCUBEA.2016.7860024
15. Verma, S., & Nidhi, V. (2017). *Extractive Summarization using Deep Learning*. http://arxiv.org/abs/1708.0443

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING