International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

News Summarization for Civil Services using Sentiment Analysis

[1]Neeraj Ghosh, [2]Rohit Majnekar, [3]Nikhita Mangaonkar

[1,2]Student, Master of Computer Applications, Sardar Patel Institute of Technology
[3]Professor, Master of Computer Applications, Sardar Patel Institute of Technology

ABSTRACT

In the big data era, there has been an explosion in the amount of text data from a variety of sources. This volume of text is an inestimable source of information and knowledge which needs to be effectively text summarized to be useful. As we are aware of the vast usage of news articles and how they are being viewed by large audiences across the globe. During the Civil Service Exams period, time becomes the most essential component for the aspirants to work on and grasp as much information as they can. Reading large articles becomes time-consuming for the aspirants. Text summarization comes into this picture where we minimize the scale of information without losing its core knowledge written by those professionals and articulate the information in an easier to understand manner.

Keywords: Text Summarization, Civil Service Exam, Preserving, Articulate, Gist

I. INTRODUCTION

In today's world where time is essential for any aspirant to prepare for the examination, every detail and information gathered by the aspirant needs to be cross-checked before reaching any conclusion. As today's world is hugely affected with the amount of misinformation in context, it becomes hard for an individual to recognise facts. With the help of our machine learning project "Saransh" where we provide recently and

1. What is Summarization?

Text summarization is the task of producing a concise and fluent summary without any human help while preserving the meaning of the original text document. We can say that Text summarization is nothing but a way of gathering the foremost, distinguished or focal information from a provider to supply a pared-down version for a specific user or tasks

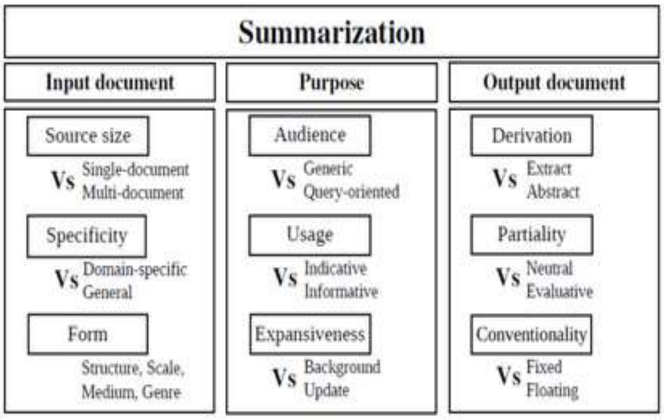


Fig 1.1: Summarization Overview

2. Abstractive Text Summarization

Utilizing linguistic methods to interpret and analyze the source text is the abstractive summarization method.. Usually, in the abstractive summarization process, advanced terminology generation and compression techniques are used to create an overview of information that is concisely conveyed.

3. Extractive Text Summarization

Extractive summarization extracts salient sentences or phrases from the source documents and groups them to produce a summary without changing the source text. Usually, sentences order is similar to the original document

II. OBJECTIVES

1. To web scrape through multiple renowned news websites.
2. Apply text summarizing techniques for requested articles and compare their accuracy using the ROUGE scoring algorithm.
3. To emphasize the highest scored article, applied sentiment analysis would also let us know about the perception of the dignitaries about the topic written and whether he/ she has a positive or negative outlook towards the context of the topic.

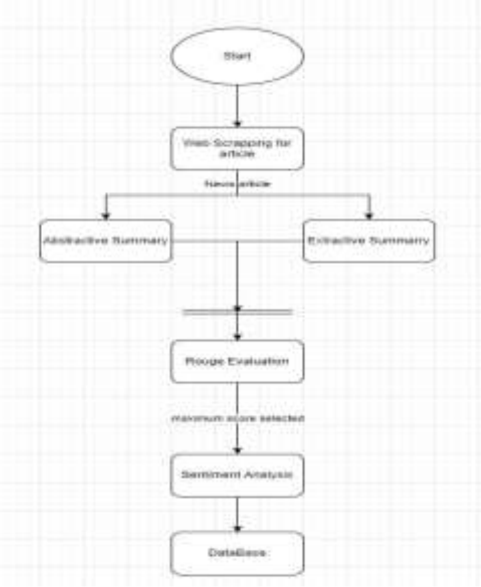


Fig 1.2: Flowchart of Process

1. INPUT:

We have used 5 news sources from the Internet with various topics such as technology, political, economical and world news to train the network.

1. OUTPUT:

Summary of a document or set of a document(s)

III. LITERATURE SURVEY

As new technologies arrive, a new age of newspaper reading and blogging articles has also become popular among the readers as well as the dignitaries of prominent newspaper houses to write articles about any political or apolitical social issues within the nation or globe. In similar terms news consumption and its spread of mouth have also become modernized as news

gets spread throughout the newspapers apps and chat applications. To reduce the widespread of factually incorrect information and misinformation amongst the general public increases the curiosity and ambition to read more about any certain trending topic, it becomes essential for us to provide factually right as well as minimal content rather than documented wide content to general readers to understand the context more easily.

Similar issues have been detected by other developers also and they made an effort to overcome such disastrous issues of misinformation amongst the general audience from hoax sources whose sole purpose is to blindfold newsreaders with manipulated information. Our key competitor in the market is the “In short” who happens to have been in the market for a long time now and has become a key player in the news application industry by providing daily news and articles to consumers in the simple summarized form of documentation. It also comprises a large audience of newsreaders which eventually transits to large capital support them.

As every application has shortcomings among them, in short also have them, as being a news application it becomes hard for readers to have an assumption whether the dignitary who wrote the articles comprises certain beliefs about the context (i.e . pro or against the law). Here, analyzing the sentiment and nuances of the articles play a vital role in news demonstration as articles could be misinterpreted if perception is not demonstrated rightly.

IV. METHODOLOGY

1. Web scraping

We created a pipeline to extract articles from various sources. From that pipeline, we take all the article data and clean it to keep the necessary data that'll be required to pass it to our model. Data like article text, dignitary name, data, and source from which it was referred. These data are stored in data to use further in our model.

2. Abstractive summarization methodology

Abstractive summarization is the task of generating a summary that captures the salient ideas of the source text. Synopses generated by automated summarization may contain sentences and phrases not included in the source text. Language generation capabilities are required for abstract summarization to generate summaries containing novel words and phrases not found in the source document.

The most recent and effective approach toward abstractive summarization is using transformer models fine-tuned specifically on a summarization dataset.

The model that we used is already pre-trained, the library we used is Transformers by Huggingface. Hugging Face provides two powerful summarization models to use: **BART** (bart-large-cnn) and **t5** (t5-small, t5-base, t5-large, t5–3b, t5–11b). We can use the min_length and max_length parameters to control the summary the model generates. When the model is in the experimentation phase, these two parameters can and should be changed to see if the model performance changes.

3. Extractive summarization methodology

Whereas, an extractive approach involves picking up the most important phrases and lines from the documents and then combines all the important lines to create the summary. So, in this case, every line and word of the summary belongs to the original document which is summarized. These news articles' topics consist of articles written by renowned dignitaries of renowned news houses. All these prime news topics are considered to analyze and understand the results of this research.

Traditionally, TF-IDF (Term Frequency-Inverse Data Frequency) is often used in info information retrieval and text mining to calculate the importance of a sentence for text summarization.

spaCy is a free and open-source library for NLP in Python with a lot of in-built capabilities. You can use spaCy to create a processed Doc object, which is a container for accessing linguistic annotations

During the process of extractive summarization following steps are followed to conquer summarization.

1. Tokenize the article using spaCy's language model.
2. Extract important keywords and calculate the normalized weight.
3. In order to extract information from unstructured text, rule-based matching is one of the steps. It's used to identify and extract tokens and phrases according to patterns and grammatical features.
4. Calculate the importance of each sentence in the article based on keyword appearance.
5. Sort the sentences based on the calculated importance.

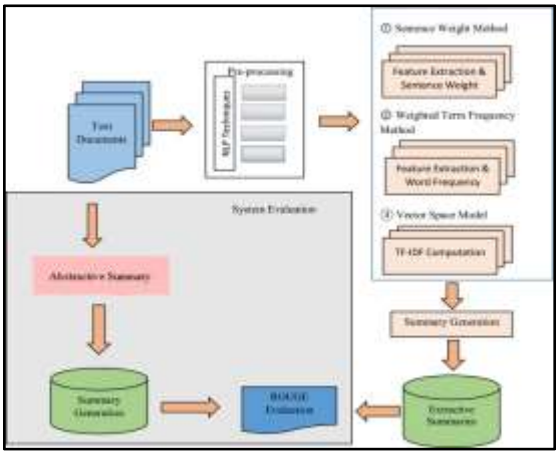


Fig:1.3: Flow Diagram of Rouge Evaluation

4. Comparing both Summarizations

The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) In a scoring algorithm, each candidate document is compared with a collection of reference documents. Use the ROUGE score to evaluate the quality of the summarized news article; it compares unigram (single-token) overlaps between the candidate document and the reference documents. Because the ROUGE score is a recall-based measure, if one of the reference documents is made up entirely of unigrams that appear in the candidate document, the resulting ROUGE score is one.

In the research community the ROUGE score is the accepted standard evaluation measure for summarization tasks; it is defined in the paper, ROUGE: A Package for Automatic Evaluation of Summaries.

$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \quad (1)$$

Fig 1.4: Rouge evaluation formula

ROUGE-

ROUGE-1 Precision and Recall compare the similarity of uni-grams between reference and candidate summaries. The word 'uni-gram' simply means a single word is considered during comparison.

$$\frac{number\_of\_overlapping\_words}{total\_words\_in\_reference\_summary}$$

Fig 1.5: Formula for ROUGE 1

ROUGE-2

ROUGE-2 Precision and Recall compare Bi-grams mean two consecutive words from the reference and candidate summaries when comparing the precision and recall comparisons with single-word. Uni-grams means a single word when comparing the precision and recall.

$$\frac{number\_of\_overlapping\_words}{total\_words\_in\_system\_summary}$$

Fig 1.6: Formula for ROUGE 2

ROUGE-L

ROUGE-L Precision and Recall measures the Longest Common Subsequence (LCS) words between reference and candidate summaries. Tokens that appear sequentially but do not necessarily have to be consecutive are known as LCS.

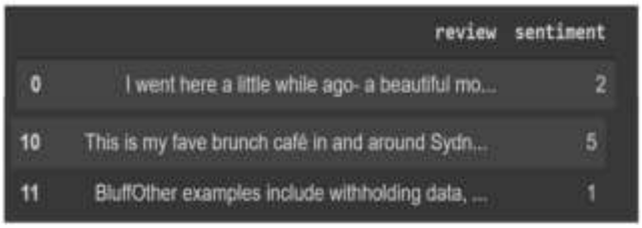
Together, A good measurement of model-generated versus golden-annotated summaries may be obtained from ROUGE-1, ROUGE-2, and ROUGE-L Precision/Recall. To make the scores even more concise, typically the F1 scores, which is the harmonic mean between Precision and Recall, are calculated for all the ROUGE scores.

Later, we perform sentiment analysis to demonstrate the ideology or viewpoint of the analyst.

BERT (Bidirectional Encoder Representations from Transformers) is a training strategy, not a new architecture design, published by researchers at Google AI Language. Comparatively, previous studies looked at text sequences either right-to-left or left-to-right. Based on the results of the study, it can be shown that a bi-directional language model has a deeper sense of context and flow than a single-direction language model..

Sentiment analysis focuses on the polarity of a text, to recognize whether the particular text has a positive or negative impact BERT is used here.

This is a bert-base-multilingual-uncased model fine-tuned for sentiment analysis on product reviews in six languages. It predicts the sentiment of the review as several stars (between 1 and 5).This model is intended for direct use as a sentiment analysis model for product reviews in any of the six languages or further fine-tuning on related sentiment analysis tasks.

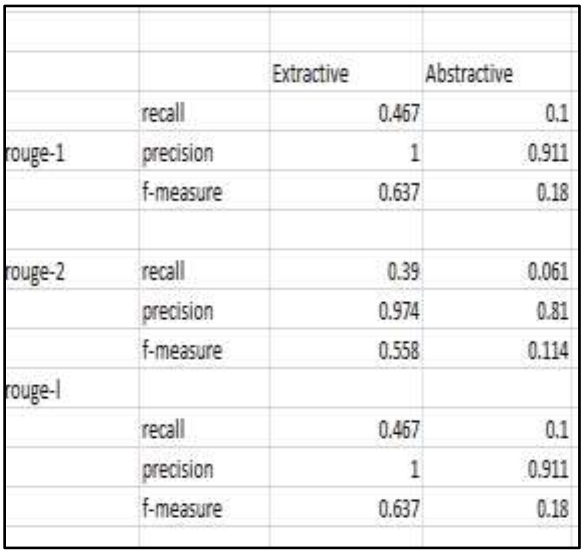


	review	sentiment
0	I went here a little while ago- a beautiful mo...	2
10	This is my fave brunch café in and around Sydn...	5
11	BluffOther examples include withholding data, ...	1

Fig 1.8 : Sentiment Analysis Reviews

The BERT framework was pre-trained using text from Wikipedia and can be fine-tuned with a question and answer datasets.

V. RESULTS



		Extractive	Abstractive
	recall	0.467	0.1
rouge-1	precision	1	0.911
	f-measure	0.637	0.18
rouge-2	recall	0.39	0.061
	precision	0.974	0.81
	f-measure	0.558	0.114
rouge-l			
	recall	0.467	0.1
	precision	1	0.911
	f-measure	0.637	0.18

Fig 1.7: ROUGE Metrics were used to compare both summarization techniques.

After performing the ROUGE metric on both the summarization techniques we compare both by implying the F-measure formula which is calculated by taking on the ratio between the product of precision and recall of text to the sum of them. This provides a clear perspective for us to decide which summarization is well suited for that article and should be further used to showcase in readers' feed.

According to the results based on figure 1.7, the f-measure scores are between 0 to 1. The abstractive summarization comes out as the least scoring compared to extractive summarization from the training model we performed.

VI. CONCLUSION

As we built a platform for such civil service aspirants and casual news enthusiasts, where they can consume large or crucial news articles within a few minutes of reading. To provide the most accurate information through the summaries various methodologies have been tested during the process. In most of these cases largely the extractive summarization of the articles comes out as close to accurate articles. Furthermore, sentiment analysis of the articles from the readers provides a bigger dataset to analyze them for future readers.

Potential conflicts of interest

None Declare

Research involving Human Participants and/or Animals

None Declare

Informed consent

None Declare

Funding

None Declare

Author contributions

As authors, our objective was to provide information in a most precise and accurate manner by gathering it through renowned news houses and dignitaries to cater for our specified audience at large. We curated source materials from prominent news houses and presented them to readers in a summarised format without disturbing the writing by the dignitary of the articles. In such a process, we tried to implement various methodologies and calculated their accuracy from the source text. Along with sentiment analysis of the article, we further tried to curate the reviews for the articles and the view of the writer.

VII. REFERENCES

1. Romain Paulus, Caiming Xiong, Richard Socher “A Deep Reinforced Model for Abstractive Summarization” 11 May 2017 v1 doi.org/10.48550/arXiv.1705.04304
2. J.N. Madhuri, R. Ganesh Kumar “Extractive Text Summarization Using Sentence Ranking” 1-2 March 2019 DOI: 10.1109/IconDSC.2019.8817040
3. Yuqi Wang, Qi Chen, Wei Wang “Multi-task BERT for Aspect-based Sentiment Analysis” 23-27 Aug. 2021, 10.1109/SMARTCOMP52413.2021.00077
4. Arpita Sahoo, Dr Ajit Kumar Nayak, Review Paper on Extractive Text Summarization, International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 5, Issue 4, April 2018.
5. Hwanhee Lee, Kang Min Yoo, Joonsuk Park, Hwaran Lee and Kyomin Jung, Masked Summarization to Generate Factually Inconsistent Summaries for Improved Factual Consistency Checking, Research in Seoul National University, 4 May 2022.
6. Shilpi Malhotra, Ashutosh Dixit, An Effective Approach for News Article Summarization, International Journal of Computer Applications (0975 – 8887)Volume 75– No.17, August 2018
7. Nidhi Patel ,Abstractive vs Extractive Text Summarization (Output based approach) - A Comparative Study ,2020 IEEE International Conference for Innovation in Technology (INOCON) ,978-1-7281-9744-9/20/$31.00 ©2020 IEEE
8. Alexander M. Rush, Sumit Chopra, and Jason Weston , A neural attention model for abstractive sentence summarization, CoRR abs/1509.00685.
9. BBC News Dataset - https://www.kaggle.com/pariza/bbc-news-summary
10. Chin-Yew Lin , ROUGE: A Package for Automatic Evaluation of Summaries
11. Zhang, Y. , Li, D. , Wang, Y. , Fang, Y. , Xiao, W. , Abstract Text Summarization with a Convolutional Seq2seq Model, Appl. Sci. 2019, 9, 1665.