International Journal of Research Publication and Reviews

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

A Research Based Project on News Summarizer or News Summary Generator

Ranjeeta Krishan[1]*, Vaishnavi Kahar*[2]*, Atul Prasad Sharma*[3]*, Divya Mishra*[4]

[1]*Assistant Professor, Bhilai Institute of Technology, Raipur, Chhattisgarh, India*
[2,3,4]*Student, Bhilai Institute of Technology, Raipur, Chhattisgarh, India*

ABSTRACT

We present our project, an Automated News Summarizer that makes use of machine learning and natural language processing. The system creates succinct summaries from long news items by utilizing ML models like TextRank or BERT in conjunction with NLP techniques like Named Entity Recognition and Sentence Embeddings. This method may change the way people consume news in today's information-rich environment because it saves time, makes information more accessible, and lessens cognitive strain on consumers.

Keywords: Information retrieval, text summarization, NLP, machine learning, and news summarizer.

1. Introduction

At the forefront of transforming the way we consume news is our News Summarizer initiative. Our method uses machine learning algorithms along with sophisticated Natural Language Processing (NLP) techniques to automatically produce brief summaries from long news items. We extract important information and condense it into readable summaries using techniques like Named Entity Recognition and advanced machine learning models like TextRank or BERT, saving consumers a great deal of time and mental strain.

This effort improves accessibility to important news items in addition to addressing the deluge of available information. Users may keep informed in today's fast-paced world by using our News Summarizer to rapidly understand the important aspects of a story without having to trawl through lengthy pieces. By providing a simple and effective remedy for summarizing news articles.

1.1 Summarization Methods

STATISTICAL BASED	• Sentence similarity, sentence length, etc.
GRAPH BASED	• TextRank, LexRank
DISCOURSE BASED	• Maximum marginal information
MACHINE LEARNING BASED	• Neural network techniques
OPTIMIZATION TECHNIQUES	• Particle swarm optimization

FIG. 1- Summarization methods

- Statistically Based Methods: These techniques provide summaries based on statistical characteristics of the document, such as sentence length, sentence position, sentence centrality, sentence numeric data, title similarity, etc [1]. These methods are independent of language and don't require a lot of storage or quick processors.
- Graph-Based Methods: In this method, words or sentences are represented by the vertices of a graph of (v, e), and the similarity relationship between these nodes is represented by the edges. By navigating the graph and choosing the sentences with a similarity index above the predetermined threshold, the sentences to be included in the extracted summary are located. Graph-based techniques such as TextRank and LexRank are well-known [4].
- Discourse Based Methods: These methods require understanding the textual structure. It becomes complex to use as they consider the sentence connections and their parts within a document [1].

- Topic-Based Methods: In this method, the summary is generated by firstly identifying the subject or theme of the document. Then this is used to extract the sentences which are related to the subject [1].
- Machine Learning-Based Methods: These include approaches where the machine learns to produce the summary from data provided to it. These data can either be supervised where the training data is provided with the summary and the machine can learn how to produce the summary from this training data or unsupervised where the machine learns by analyzing the document since no training data is available. Some of the machine learning techniques are neural network, SVM, Generic algorithm and fuzzy logic[4].
- Optimization Techniques: These use nature-inspired algorithms such as swarm algorithms and are usually used in combination with other techniques[3].

1.2 Key Assignments of summarization

The following are the main tasks for summarization:

- As a result, use a keyword to retrieve online articles from news websites.
- Divide entire articles into a collection of sentences that serves as the dataset for processing in advance.
- Speaking in a way that is understandable and justified to machines.
- Finding semantic similarities between sentences to eliminate actual excess in a summary.
- Grouping similar sentences together to distinguish between sentences with different semantic meanings.
- Selecting sentences from each cluster that relate to the information that the comparative cluster has displayed.
- Arranging the phrases in a chronological manner to demonstrate the developments as they occurred

1.3 Proposed System

Using an extractive summarizer, we can tackle this issue. Natural language processing (NLP) aims to summarize articles by selecting a set of words that contain the most important information. In order to generate a summary, this method extracts a substantial section of a sentence. Several algorithms and techniques are used to define sentence verbs and then rank them according to their importance and similarity.

In order to assist individuals rapidly identify and use the right information, news summary techniques are much needed given the volume of text material available online. The use of text summaries also shortens reading passages, expedites information research, and expands the pool of potentially unrelated material.

2. Approach

The frequency-based method for news summarizing is the main topic of this research study. Sentence and word tokenization, followed by the calculation of sentence score based on TF-IDF score—which is used to identify the most significant sentence, retain the information, and combine it to make a summary—are the procedures involved in creating a news summarizer.

Step 1: Bring in all required libraries

When working with text in Python, one commonly used library is called NLTK (Natural Language Toolkit). Stop words comprise an English stop word list that must be eliminated in the pre-processing stage depicted in Figure 2.

Step 2: Produce Orderly Sentences

The most crucial stage in getting a consistent and fruitful approach outcome is text processing. As seen in Figure 3, the processing phases eliminate special characters, words, and numerals.

Step 3: Produce a matrix and compute TF-IDF

We will determine each word's TF and IDF within a paragraph.

TF (t) is equal to (total_no. of t in the document) / (frequency of t from the document).

IDF (t) is equal to log_e (total_number_of_documents/number_of_documents−t it) [4].

The computed TF and IDF values will now be multiplied, as indicated in Figure 4, to create a new matrix.

Step 4: Give each sentence a score

Here, a paragraph is given weight by the usage of TF-IDF word points in a sentence. Sentence rating fluctuates depending on the methodology, though.

Step 5: Produce the synopsis

The process of news summarizing ends here. Finally, a summary is generated by including the top sentences, which are determined by the user's recall rate and score.

3. Conclusions

In order to give users rapid access to important information, the news summarizer project seeks to reduce vast amounts of news stories into succinct summaries. By utilizing natural language processing (NLP) techniques, the system recognizes significant sentences and extracts pertinent data, allowing users to remain informed without having to read through full articles. Adding machine learning algorithms can also improve the process of summarizing by continuously raising the relevance and accuracy of the summaries that are produced. All things considered, the news summarizer project provides an invaluable means of effectively processing enormous volumes of news material, sparing users time and energy while ensuring they are informed.

References

[1] Kumar, Akshi, Aditi Sharma, and Anand Nayyar. "Fuzzy Logic based Hybrid Model for Automatic Extractive Text Summarization." In Proceedings of the 2020 5th International Conference on Intelligent Information Technology, pp. 7-15. 2020.

[2] Gambhir, Mahak, and Vishal Gupta. "Recent automatic text summarization techniques: a survey." Artificial Intelligence Review 47, no. 1 (2017): 1-66.

[3] VS, Raj Kumar, and D. Chandrakala. "A survey on text summarization using optimization algorithm." ELK Asia Pacific Journal of Computer Science and Information Syatems 2, no. 1 (2016): 31-40.

[4] Sharma, Richa, and Prachi Sharma. "A survey on extractive text summarization." International Journal 6, no. 4 (2016).

[5] Mihalcea, Rada, and Paul Tarau. "Textrank: Bringing order into text." In Proceedings of the 2004 conference on empirical methods in natural language processing, pp. 404-411. 2004.

[6] Sethi, Prakhar, Sameer Sonawane, Saumitra Khanwalker, and R. B. Keskar. "Automatic text summarization of news articles." In 2017 International Conference on Big Data, IoT and Data Science (BID), pp. 23-29. IEEE, 2017.

[7] Kosmajac, Dijana, and Vlado Kešelj. "Automatic Text Summarization of News Articles in Serbian Language." In 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH), pp. 1-6. IEEE, 2019.

[8] Allahyari, Mehdi, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. "Text summarization techniques: a brief survey." (2017).

[9] Gaikwad, Deepali K., and C. Namrata Mahender. "A review paper on text summarization." International Journal of Advanced Research in Computer and Communication Engineering 5, no. 3 (2016): 154-160.