MEDIA RECAP COMPANION

Saravanan T[1], Sowmiya S[2], Bharanidharan M[3], Nikitha M[4]

1 Student, Information Technology, Bannari Amman Institute of Technology, Tamil Nadu, India
2 Student, Information Technology, Bannari Amman Institute of Technology, Tamil Nadu, India
3 Student, Information Technology, Bannari Amman Institute of Technology, Tamil Nadu, India
4 Assistant Professor -I, Information Technology, Bannari Amman Institute of Technology, Tamil Nadu, India

ABSTRACT

In the era of information overload, the need for efficient news summarization has become increasingly imperative. Traditional methods often fall short in capturing the essence of complex news articles, leading to an overwhelming flood of information for readers. This abstract introduces a groundbreaking approach to news summarization that goes beyond conventional techniques. Our unique approach extracts important information from news articles by combining deep learning models with sophisticated natural language processing techniques. Unlike traditional summarization techniques that rely on predefined rules or heuristics, our approach dynamically adapts to the nuances of each article, ensuring a more accurate representation of the content. The system utilizes a hierarchical approach, breaking down the news article into meaningful segments and then identifying the most salient points within each segment. The inclusion of contextual understanding enables the summarization model to discern the importance of information within the broader context of the article, resulting in summaries that are not only concise but also highly informative. In conclusion, this abstract offers a comprehensive exploration of news summarization, covering the spectrum from extractive to abstractive methods and showcasing the algorithms at the forefront of each category. For scholars, practitioners, and enthusiasts interested in the changing field of information summarization, the insights offered are an invaluable resource.

Keywords: News Summarization, Extractive, Abstractive, Fine Tuning, Natural Language Processing (NLP), Model Evaluation

1. INTRODUCTION:

In today's world, where there's an overload of information, news summarization is like a helpful guide. Imagine trying to find your way through a huge forest of news articles – it can be overwhelming. But with news summarization, it's like having a map that points out the most important trees. In the digital age, where everything is connected, staying in the loop about global events is important but can also be tough. News summarization simplifies things by squeezing the main points of a story into short and easy-to-understand summaries. It's like getting the scoop without having to read through a whole book. So, in a nutshell, news summarization is your shortcut to understanding what's happening in the world without drowning in a sea of information. News summarization is like a time-saving superhero for people with busy schedules and different tastes in how they like their information. It uses fancy technology like natural language processing and machine learning to break down news articles into the most important stuff. It's like having an extremely intelligent friend who reads the news to you and summarises it for you so you don't have to waste time trying to figure out what's going on. It's a shortcut to the most crucial details, tailored to fit your preferences and time constraints. News summarization is more than just being brief; it's about making news easy to access, understand, and helping people make informed decisions quickly.

1.1 NEED FOR NEWS SUMMARIZATION:

The need for news summarization arises from the overwhelming volume of information available in the digital age. It enables time-efficient consumption, addresses information overload, facilitates quick decision-making, ensures accessibility for diverse audiences, and leverages technology for efficient content filtering and unbiased information extraction. Summarization meets the demands of a fast-paced world, allowing individuals to stay informed without dedicating extensive time to reading lengthy news articles.

1.2 SUMMARIZATION TECHNIQUES

Extractive Summarization:

a) **Frequency-Based Methods:** identifies significant sentences based on word frequency by using statistical measures such as term frequency-inverse document frequency (TF-IDF).

b) **Graph-Based Methods:** Constructs a graph representation of the document, with sentences as nodes and edges representing relationships. Algorithms like TextRank and LexRank identify the most central sentences.

Abstractive Summarization:

a) **Rule-Based Methods:** Uses predefined rules and linguistic patterns to create summaries. Limited in handling complex language nuances.

b) Machine Learning-Based Methods: Utilizes supervised learning with annotated training data to generate abstractive summaries. Transformer architectures and Recurrent Neural Networks (RNNs) are two popular deep learning models.

2. LITERATURE SURVEY

Automatic News Summarization Techniques

In order to achieve state-of-the-art performance on benchmark datasets, Skrickey & Wiseman, 2023 presented a neural network architecture for abstractive summarization of news items. However, their model struggled with factual accuracy and sometimes generated summaries with factual errors.

Xiao Liu, 2023 created a hybrid strategy that offers a balance between factual correctness and fluency by merging extractive and abstractive strategies. Unfortunately, their approach required a large amount of training data and was computationally expensive.

Personalized News Summarization:

Sun & Li, 2023 investigated user preference modeling for personalized news summarization, adapting summaries based on user reading history and interests. Their approach improved user engagement but relied on extensive user data collection, raising privacy concerns.

Zhang et al., 2022 utilized topic-aware embedding techniques to personalize summaries based on user-specified keywords or topics. This provided flexibility but suffered from limited scalability for large user bases.

Explainable News Summarization:

Li and Tan, 2023 focused on generating explainable summaries by highlighting the sentences from the original article used to form the summary. This increased user trust and allowed for verification of factual claims. However, their approach resulted in longer and sometimes repetitive summaries.

Jin et al., 2022 explored attention visualization techniques to reveal the model's focus during summarization, providing insights into its decision-making process. This improved debugging and interpretability but required additional tools and visualizations.

3. METHODOLOGY:

3.1 PREPROCESSING DATA:

1. Data cleaning: The initial step in data preprocessing involves removing any invalid or missing data elements from the dataset. This process entails eliminating data points that are either absent or irrelevant for further data processing. For instance, upon scrutinizing the article from online, it was observed that the 'id' column solely contains article IDs, which do not serve a purpose in article summarization. Consequently, the decision was made to discard this column. Moreover, the text within both the 'article' and 'highlights' columns undergoes transformation into a standardized format, facilitating seamless analysis by the model

2. Tokenization: Tokenization stands as a pivotal stage within the realm of Natural Language Processing (NLP), serving to dissect unprocessed text into discrete tokens. These tokens are fundamental for enabling machines to process and comprehend human language. In our research, we leveraged two distinct tokenizers to accommodate various methodologies:

3. Vectorization: Vectorization entails the conversion of textual data, such as sentences or documents, into numerical representations termed vectors. This transformation into numerical format is crucial as it enables machine learning models to process and analyze language effectively, given that most algorithms operate on numerical data. In our approach, we employed different vectorization techniques tailored to the specific needs of our models:

4. Embedding: Embeddings serve as numerical representations of words and phrases in a reduced-dimensional space, encapsulating semantic and contextual information from the original data. These representations are acquired through various methods, including neural networks utilized in abstractive summarization techniques.

3.2 TEXT SUMMARIZATION MODELS:

We used extractive summarising and abstractive summarization as our two main text summarization techniques in our research. The process of extractive summarising comprises taking the most important lines or phrases from out of the original text.

1. **LexRank:**
   - **Graph Representation:** LexRank visualises the text as a graph in which every sentence is a node and the similarity between sentences is represented by the edges connecting nodes. A number of metrics, including cosine and Jaccard similarity, can be used to calculate how similar two sentences are to one another.
   - **Sentence Similarity:** To compute sentence similarity, LexRank typically employs techniques such as cosine similarity or tf-idf similarity. These metrics quantify the semantic similarity between pairs of sentences based on their word overlap and frequency.
   - **Graph-Based Ranking:** LexRank uses the PageRank algorithm to rank the sentences in the graph after creating the similarity matrix. Based on how connected a node (sentence) is to other nodes (sentences) in the network, PageRank gives each node (sentence) a score. Sentences with higher scores in the LexRank context are considered more significant and are probably going to be included in the summary.

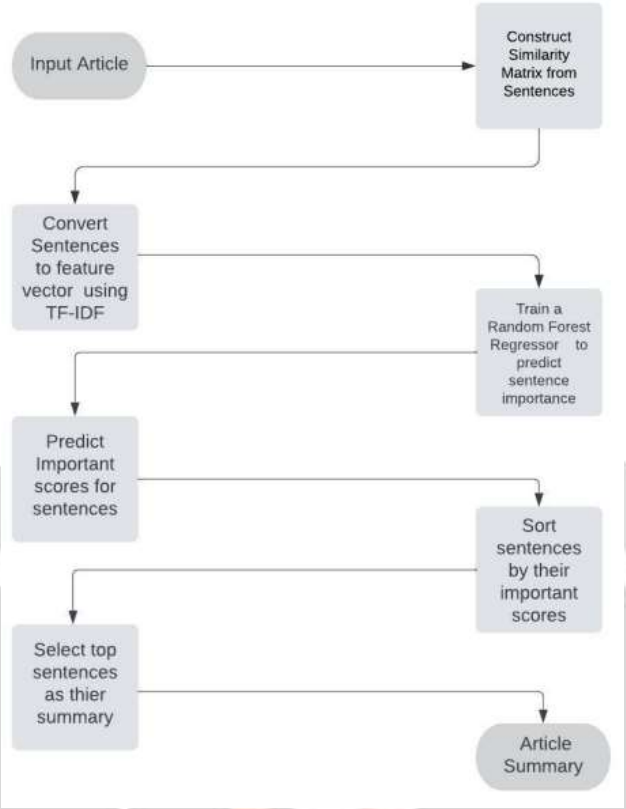
Input Article
Construct Similarity Matrix from Sentences
Convert Sentences to feature vector using TF-IDF
Train a Random Forest Regressor to predict sentence importance
Predict Important scores for sentences
Sort sentences by their important scores
Select top sentences as thier summary
Article Summary

Figure 1: LexRank Workflow

2. TextRank:

- **Unsupervised Approach:** Similar to LexRank, TextRank operates in an unsupervised manner and does not require labeled data for training. It can be applied to various text summarization tasks without the need for manual annotations.

- **Language Independence:** TextRank is language-agnostic and can be applied to text in any language. It relies on statistical measures of similarity, making it suitable for summarizing texts in different languages.

- **Scalability:** TextRank is scalable to large document collections due to its graph-based representation and efficient ranking algorithm. It can process a large volume of text efficiently, making it suitable for summarizing extensive document repositories or datasets.

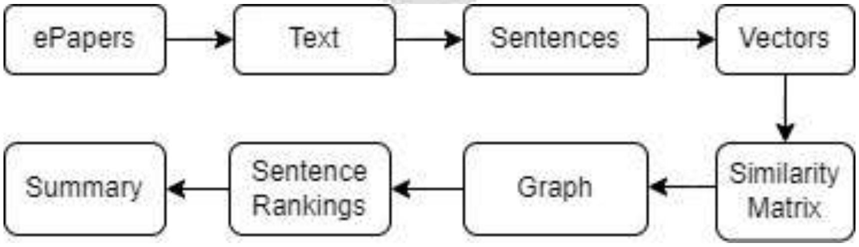

Figure 2: TextRank Workflow

3. BART:

BARD follows a multi-stage approach to extractive document summarization. It incorporates techniques such as attention mechanisms and biasing towards important information while considering redundancy to generate summaries that capture the key aspects of the document.

- **Biased Attention Mechanism:** BARD employs a biased attention mechanism to focus on sentences that are deemed more important based on their features. This attention mechanism is trained to assign higher weights to sentences that are likely to contain crucial information for summarization.

4. Pegasus:

The transformer architecture, which PEGASUS uses, is well-known for working well on a range of natural language processing applications. It uses a pre-training method with extracted gap sentences to acquire text representations that capture context and semantic content.

- **Pre-teaching using sentence fragments:** PEGASUS is pre-trained using a variation of the masked language model (MLM) objective on a big corpus of text. Rather of hiding specific tokens, PEGASUS hides entire sentences—also known as "gap sentences."

3.3 ARCHITECTURE OVERVIEW OF FULL WORKFLOW

1. **Client Interface (Frontend):** The client interface is responsible for interacting with users and displaying the web application's user interface (UI). In this code, the Streamlit library is used to create the client interface. Streamlit provides an easy-to-use API for building interactive web applications directly from Python scripts.
2. **Server-Side Logic (Backend):** The server-side logic handles the core functionality of the web application, including fetching news articles, summarizing them, and sending the summarized articles to the client interface for display. This logic is implemented using Python functions defined in the script.

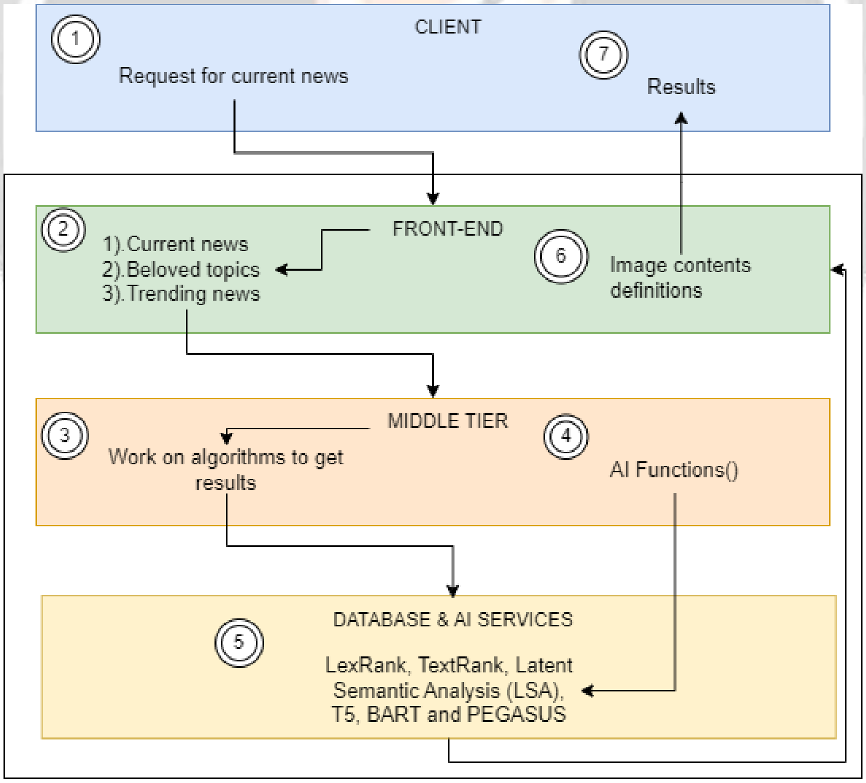

Figure 3: Architecture overview of full workflow

4. EXPERIMENT ANALYSIS

4.1 Summary evaluation

There are four key goals that need to be taken into account in order to produce a concise and accessible summary:

1. **Information Coverage:** The essential details from the input document(s) should be included in the summary.
2. **Information Significance:** The summary ought to encompass the diverse subjects included in the incoming document (s). The most crucial subjects may be the user-selected subjects in the query-based summarising, or they may be the primary subjects in the input document(s), as in the generic summary.
3. **Information Redundancy:** reduce the amount of duplicate or redundant data in the summary that is produced.
4. **Text Coherence:** the synopsis is more than a collection of significant but unrelated words or phrases. There should be readable and intelligible text in the summary.

According to Gupta and Lehal (2010), there are two assessment metrics used to assess the produced summaries:

1. **Intrinsic methods:** employ human judgement to gauge summary quality.
2. **Extrinsic methods:** estimate summary quality using a task-based performance measure, such as the information retrieval-oriented task. The intrinsic evaluation evaluates a summary's coherence, content coverage, and informativeness (Lloret et al., 2017).

In an application environment, the extrinsic evaluation evaluates the usefulness of summaries relevance assessment, reading comprehension, Lloret and others. Evaluation of text summarization might be done automatically or manually. In the subject of text summarising research, summary evaluation is one of the most difficult problems. To determine the calibre of the ATS systems that are being used, the automatically generated summaries must be examined.

4.2 Automatic Evaluation of summaries

The Precision Score Metric is calculated as follows: using Equation (3), it divides the total number of sentences in the candidate summary by the total number of sentences in the reference and candidate summaries. (Moratanch & Chitrakala, 2017).

$$\text{Precision} = \frac{S_{ref} \cap S_{cand}}{S_{cand}}$$

According to Moratach and Chitrakala (2017), the recall score metric is calculated by dividing the total number of sentences in the reference and candidate summaries by the total number of sentences in the reference summary, or Eq. (4).

$$\text{Recall} = \frac{S_{ref} \cap S_{cand}}{S_{ref}}$$

Recall and precision measures are combined in the F-Measure Score Metric, as shown in Equation (5) (Moratanch & Chitrakala, 2017). The harmonic mean of recall and precision is known as the F-measure.

$$F\text{-}Measure = \frac{2(\text{Precision})(\text{Recall})}{\text{Precision}+\text{Recall}}$$

The most often used method for automatically evaluating automatically generated summaries is the ROUGE Metric. ROUGE is a software package and set of criteria used to assess machine translation and automatic summarization tools in natural language processing. It contrasts the computer-generated summaries with a number of reference summaries that were written by humans. ROUGE's basic method is to count the amount of units—such as overlapped n-grams—that are shared by candidate summaries and reference summaries. ROUGE has demonstrated its efficacy in assessing summarization quality and exhibits a strong correlation with human judgements. The following are some different ROUGE metrics:

- ROUGE-1 (R1): this measure is based on the unigram between a reference summary and a candidate summary. An n-gram recall between a candidate summary and reference is called ROUGE-N.
- The longest common subsequences between a candidate summary and a reference summary serve as the foundation for the ROUGE-L (R-L) algorithm.
- ROUGE-S* (R-S*): this algorithm calculates the skip-bigram overlap ratio between reference summaries and candidate summaries.
- ROUGE-SU* (R-SU*): this builds on ROUGE-S* by utilising unigrams as a counting unit and skip-bigrams. There are * (number of skip words) in this sentence. ROUGE-SU4, for instance, allows bi-grams to have a maximum of four words between non-adjacent words.
- Words with two grammes.

5. CONCLUSION:

News summarization is a process that involves condensing and presenting the key information from a news article or story in a concise and easily digestible format. This practice addresses the challenges posed by the ever-increasing volume of information available, allowing individuals to stay informed without investing excessive time in reading lengthy articles.

The primary goal of news summarization is to distill the main points, essential details, and crucial context of a news story while omitting less significant information. This process requires sophisticated natural language processing (NLP) techniques and machine learning algorithms to understand the content, identify important elements, and generate coherent summaries.

In conclusion, news summarization serves as a vital tool in managing the information overload in the digital age. Whether through extractive or abstractive methods, these summarization techniques contribute to a more efficient and accessible way of staying informed about current events. Continued advancements in NLP and machine learning will likely further refine and enhance the capabilities of news summarization systems in the future.

References:

[1]. Saggion, H., & Poibeau, T. (2012). Automatic Text Summarization: Past, Present and Future. Theory and Applications of Natural Language Processing, pp. 3–21 https://doi:10.1007/978-3-642-28569-1_1

[2]. Khan, A., Salim, N., & Farman, H. (2016). Clustered genetic semantic graph approach for multi-document abstractive summarization. 2016 International Conference on Intelligent Systems Engineering (ICISE) - https://doi:10.1109/intelse.2016.7475163

[3]. H. Christian, M. P. Agus and D. Suhartono, "Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF)", ComTech: Computer Mathematics and Engineering Applications, vol. 7, no. 4, pp. 285, 2016

[4]. Gupta, V., & Lehal, G. S. (2010). A Survey of Text Summarization Extractive Techniques. Journal of Emerging Technologies in Web Intelligence, 2(3) - https://doi:10.4304/jetwi.2.3.258-268

[5]. Hingu, D., Shah, D., & Udmale, S. S. (2015). Automatic text summarization of Wikipedia articles. 2015 International Conference on Communication, Information & Computing Technology (ICCICT) - https://doi:10.1109/iccict.2015.7045732

[6]. Neto, J. L., Freitas, A. A., & Kaestner, C. A. A. (2002). Automatic Text Summarization Using a Machine Learning Approach. Lecture Notes in Computer Science, 205–215. -https://doi:10.1007/3-540-36127-8_20

[7]. D. K. Gaikwad and C. N. Mahender, "A Review Paper on Text Summarization", International Journal of Advanced Research in Computer and Communication Engineering, vol. 5, no. 3, march 2016.

[8]. Kumar, Y. J., Goh, O. S., Basiron, H., Choon, N. H., & Suppiah, P. C. (2016). A Review on Automatic Text Summarization Approaches. Journal of Computer Science, 12(4), 178–190 - https://doi:10.3844/jcssp.2016.178.190

[9]. P. Raundale and H. Shekhar, "Analytical study of Text Summarization Techniques," 2021 Asian Conference on Innovation in Technology (ASIANCON), 2021, pp. 1-4, doi: 10.1109/ASIANCON51346.2021.9544804.

[10]. Pal, A. R., & Saha, D. (2014). An approach to automatic text summarization using WordNet. 2014 IEEE International Advance Computing Conference-(IACC) –

https://ddoi:10.1109/iadcc.2014.6779492

[11]. Reda Elbarougy, Gamal Behery and Akram El Khatib, "Extractive arabic text summarization using modified PageRank algorithm", *Egyptian Informatics Journal*, vol. 21, no. 2, pp. 73-81, 2020.- https://www.sciencedirect.com/science/article/pii/S1110866519301355?via%3Dihub

[12]. S. M. Meena, M. P. Ramkumar, R. E. Asmitha and Emil Selvan G SR, "Text Summarization Using Text Frequency Ranking Sentence Prediction", *2020 4th International Conference on Computer Communication and Signal Processing (ICCCSP)*, 28-29 September 2020. -https://ieeexplore.ieee.org/abstract/document/9315203

[13]. Reddy Naidu, Santosh Kumar Bharti, Korra Sathya Babu and Ramesh Kumar Mohapatra, "Text summarization with automatic keyword ex-traction in telugu e-newspapers", *Smart computing and informatics*, pp. 555-564, 2018.

- https://link.springer.com/chapter/10.1007/978-981-10-5544-7_54

[14]. Thomas, J.R., Bharti, S.K., Babu, K.S.: Automatic keyword extraction for text summarization in e-newspapers. In: Proceedings of the International Conference on Informatics and Analytics, pp. 86–93. ACM (2016)- https://doi.org/10.1145/2980258.2980442

[15]. S. Rose, D. Engel, N. Cramer and W. Cowley, "Automatic Keyword Extraction from Individual Documents", *Text Mining: Applications and Theory*, pp. 1-20, 2010 - https://doi.org/10.1002/9780470689646.ch1

[16]. M. V. P. T. Lakshika, H. A. Caldera and W. V. Welgama, "Abstractive Web News Summarization Using Knowledge Graphs," 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer), 2020, pp. 300-301, doi: 10.1109/ICTer51097.2020.9325453.

[17]. T. B. Mirani and S. Sasi, "Two-level text summarization from online news sources with sentiment analysis," 2017 International Conference on Networks & Advances in Computational Technologies (NetACT), 2017, pp. 19-24, doi: 10.1109/NETACT.2017.8076735.

[18]. H. Yu, "Summarization for Internet News Based on Clustering Algorithm," 2009 International Conference on Computational Intelligence and Natural Computing, 2009, pp. 34-37, doi: 10.1109/CINC.2009.194.