News Summarization Articles by Using NLP

PRADEEP TRIPATHI[1], AMIT KUMAR PANDEY[2]
[1] *Senior Technical Architect, Coforg Ltd, NIIT Technologies, Sector Tech Zone, Greater Noida, Uttar Pradesh*
[2] *PG Student, Department of IT, Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai, Maharashtra, India*

Abstract- In this research endeavor, we delve into the realm of news summarization, harnessing the power of advanced natural language processing techniques for the automated condensation of BBC articles. Our dataset encompasses five diverse categories—business, entertainment, politics, sport, and tech—offering a rich tapestry of information. The central objective of our study is to craft a news summarization system that is both efficient and accurate, leveraging cutting-edge language models. We utilize the Hugging Face Transformers library to construct a summarization pipeline adept at distilling crucial insights from extensive news articles.

I. INTRODUCTION

The ever-expanding volume of digital content has made it increasingly difficult for readers to stay abreast of the vast information available. To address this challenge, automatic text summarization has emerged as a valuable solution, offering users concise and informative summaries of lengthy articles. Our research is dedicated to implementing a news summarization system specifically tailored for the diverse content within the BBC News dataset. We categorize articles into business, entertainment, politics, sport, and tech, aiming to tackle the unique challenges posed by each domain.

Utilizing state-of-the-art language models and the Hugging Face Transformers library, our goal is to craft a robust summarization pipeline. This system is intricately designed to analyze and distill key information from news articles, preserving their essence across different categories. Through our research, we aim to make a meaningful contribution to the fields of natural language processing and information retrieval, providing a comprehensive solution for summarizing news content from one of the world's most reputable news sources, the BBC.

II. LITERATURE REVIEW

The collaborative work of P. Sethi, S. Sonawane, S. Khanwalker, and R. B. Keskar, presented at the 2017 International Conference on Big Data, IoT, and Data Science in Pune, India, delves into the realm of automatic text summarization, an area of perpetual interest in academia. Despite the numerous techniques developed for this purpose, the pressing concern remains efficiency, particularly with the escalating size and quantity of online documents.

Their paper introduces a novel approach to text summarization, addressing the challenge of identifying the most crucial segments of a text to produce coherent and concise summaries. Instead of relying on full semantic interpretation, the proposed methodology employs a model of topic progression derived from lexical chains. This approach creates a summary that captures the essence of the text without requiring exhaustive semantic analysis.

The authors present an optimized and efficient algorithm for generating text summaries, utilizing lexical chains and incorporating the WordNet thesaurus. To enhance the quality of the summary, they tackle the limitations of the lexical chain approach by implementing pronoun resolution. Additionally, the paper introduces new scoring techniques that leverage the inherent structure of news articles, providing a comprehensive solution to the need for an efficient automatic news summarizer in the digital age.[1]

The collaborative effort of Y. Du and H. Huo, as documented in their 2020 publication in IEEE Access,

addresses the escalating significance of automatic text summarization over the past seven decades. With the exponential growth of Internet data, the extraction of valuable information and knowledge from texts has become crucial, serving diverse user needs. News text, being a ubiquitous form of communication in people's daily lives, stands out as a primary focus in their study.

The authors present an innovative automatic summarization model tailored for news text, incorporating fuzzy logic rules, multi-feature analysis, and Genetic Algorithm (GA). The model begins by emphasizing word features, wherein each word is scored, and those surpassing a predefined score are extracted as keywords. Given the unique elements present in news text, such as time, place, and characters, these specifics can also be directly extracted as keywords in certain instances.

Moving to sentence features, a linear combination of various features is employed to determine the importance of each sentence. Genetic Algorithm is employed to assign weights to each feature, contributing to the overall assessment. Finally, a fuzzy logic system is applied to calculate the final score, facilitating automatic summarization.

Comparative assessments using the ROUGE evaluation method on the DUC2002 dataset showcase the superior performance of the proposed method over other existing approaches, including Msword, System19, System21, System31, SDS-NNGA, GCD, SOM, and Ranking SVM. The results affirm the efficacy of their model in achieving more effective and accurate news text summarization.[2]
Yang Y, Tan Y, Min J, and Huang Z contribute to the field of automatic text summarization with their work published in The Journal of Supercomputing in August 2023. Focused on the specific domain of government news reports, their research aims to efficiently extract crucial information from the often detailed and normatively formatted content found in official government communications. In an era of information overload, the need for readers to quickly comprehend government news is paramount.

Their proposed automatic text summarization model relies on multiple features to tackle the challenge posed by the extensive length of government news reports. Initially, features are extracted through the TF–IDF algorithm and word vector embedding method based on bidirectional encoder representation from the transformers model. Subsequently, sentences are scored based on position, keywords, and similarity features. The highest-ranked sentence is then selected to compose the summarization.

To validate the effectiveness of their approach, the researchers employ the Edmundson and ROUGE evaluation criteria. The Edmundson evaluation reveals minimal score differences between the Automatic Text Summarization (ATS) based on their method and manual summarization, indicating high similarity across consistency, grammaticality, time sequence, conciseness, and readability.

Furthermore, the ROUGE evaluation criteria demonstrate the superiority of the proposed method over other models. Particularly, character-level ROUGE-1 scores for Precision, Recall, and F1 reach 0.84, 0.93, and 0.88, respectively. At the word-level ROUGE-1, Precision, Recall, and F1 scores are 0.81, 0.89, and 0.85, showcasing a significant enhancement compared to alternative models. Notably, the proposed method stands out in facilitating rapid information retrieval, offering advantages over manual methods in efficiently delivering important information to readers.[3]

Esmaeilzadeh, Peh, and Xu delve into the realm of abstractive text summarization in their 2019 work, as documented in the arXiv preprint arXiv:1904.00788. Their study involves a thorough exploration of various models, including LSTM-encoder-decoder with attention, pointer-generator networks, coverage mechanisms, and transformers. Through meticulous hyperparameter tuning, they conduct a comparative analysis of these proposed architectures for the task of abstractive text summarization.

The research goes beyond summarization, extending its application to the domain of fake news detection. In this context, the developed text summarization model serves as a feature extractor. The approach involves summarizing news articles before classification, and the results are then compared against the classification using only the original news text. This innovative extension demonstrates the

versatility of their summarization model, showcasing its potential utility in addressing broader challenges such as fake news detection.[4]

Y. Chen and Q. Song, in their 2021 work presented at the IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference, focus on addressing the challenge of accurately extracting valuable information from the rapidly growing data on the internet. They emphasize the crucial role of news reports in understanding societal affairs, highlighting the need for concise summaries to aid readers in comprehending content swiftly.

The paper introduces an enhanced strategy for abstractive text summarization, specifically targeting the issue of topic deviation. Their approach involves a combination of TextRank and the BART model. Initially, TextRank and BART are employed to extract and generate summarizations from news text. Subsequently, the results from both methods are merged to create a new text, elevating the significance of key sentences and ensuring thematic coherence. The process concludes by inputting the refined text into the BART model for the final summarization.

Experimental findings reveal notable improvements compared to a standalone BART model. The average recall scores for Rouge-1, Rouge-2, and Rouge-L exhibit enhancements of 1.5%, 0.5%, and 1.3%, respectively. This underscores the effectiveness of their proposed approach in mitigating topic deviation and enhancing the overall summarization quality.[5]
T. B. Mirani and S. Sasi, in their work presented at the 2017 International Conference on Networks & Advances in Computational Technologies, tackle the common practice of individuals reading multiple news articles to gather comprehensive information on a topic. Recognizing the time and energy involved, they employ an extractive-based approach to achieve a two-level text summarization from online news sources.

The research encompasses diverse news topics, including politics, sports, health, science, and movie reviews sourced from various news outlets like Fox News (USA), NZ Herald (New Zealand), Hindustan Times (India), and BBC (UK). The first-level summary is generated for each article across these topics, providing a condensed version. Sentiment analysis is then applied to the first-level summaries to discern variations in news articles from different agencies.

Moving to the second level, the research combines the first-level summaries of two or three related articles on a specific topic to generate a comprehensive summary. The evaluation of summarization performance employs the ROUGE metric.

In essence, their approach not only streamlines the information retrieval process by providing condensed first-level summaries for each article but also analyzes sentiments within these summaries. The subsequent generation of second-level summaries consolidates information from related articles, contributing to a more comprehensive and nuanced understanding of the chosen topics.[6]

J. Cheng, F. Zhang, and X. Guo present a noteworthy contribution to the realm of automatic text summarization in their work published in IEEE Access in 2020. Acknowledging the challenges of information overload in the information age, they propose an advanced text summarization model that extends the traditional sequence-to-sequence (Seq2Seq) neural approach.

The key innovation lies in the incorporation of a syntax-augmented encoder and a headline-aware decoder. The encoder is designed to encode both the syntactic structure and word information of a sentence into the sentence embedding. To enhance attention to syntactic units, the authors introduce a hierarchical attention mechanism. On the decoding side, a headline attention mechanism and a Dual-memory-cell LSTM network are integrated to elevate the quality of generated summaries.

Experiments comparing their proposed method with baseline models on the CNN/DM datasets demonstrate its superiority over abstractive baseline models, as evidenced by higher scores on ROUGE evaluation metrics. Impressively, the proposed method achieves a summary generation performance on par with the extractive baseline method. Qualitative analysis further validates the approach, revealing that the summaries generated by their model are not only more

readable but also less redundant, aligning well with intuitive expectations.

In essence, their syntax-augmented and headline-aware neural text summarization method represents a promising advancement in addressing information overload by producing more coherent and effective summaries.[7]

III. METHODOLOGY

In our study, we adopt a systematic methodology to develop a news summarization system, leveraging the capabilities of the Hugging Face Transformers library. The focus of our research is on the BBC News dataset, which is meticulously organized into five distinct categories: business, entertainment, politics, sport, and tech. Each category represents a unique domain, introducing specific challenges in the summarization process.

To process the textual data, we harness the summarization pipeline provided by the Transformers library. This involves configuring the pipeline with essential parameters like max_length, min_length, length_penalty, num_beams, and no_repeat_ngram_size. The fine-tuning of these parameters is crucial to strike a balance between the length and informativeness of the generated summaries.

For system evaluation, we select a subset of articles from each category to ensure a representative sample. The assessment is conducted using metrics such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation). This approach allows us to gauge the quality of the generated summaries by comparing them against reference summaries, providing valuable insights into the effectiveness of our news summarization system.

IV. RESULTS

Our findings underscore the efficacy of our news summarization system across a spectrum of categories. The ROUGE scores affirm the system's capability to efficiently capture essential information while upholding coherence and relevance. Notably, the system exhibits adaptability to the distinctive characteristics of each category, delivering succinct and informative summaries for articles in business, entertainment, politics, sport, and tech.

Through quantitative assessments, we observe the system's adeptness in generating summaries that closely align with the reference summaries. The flexibility in parameter settings allows for tailoring the system to different domains, ensuring optimal performance. The incorporation of the Hugging Face Transformers library emerges as a pivotal factor, playing a key role in attaining cutting-edge results in the domain of news summarization.

CONCLUSION

Our research project establishes a robust solution for automating news summarization, specifically designed for the complexities of the BBC News dataset. The successful deployment of our summarization pipeline across a range of categories highlights its adaptability and effectiveness, addressing the nuanced challenges presented by diverse news content.

The flexibility inherent in our approach, empowered by the Hugging Face Transformers library, ensures its customization for various datasets and domains. As we continuously refine and expand the system, its potential to enhance information retrieval and improve user experiences in navigating extensive news content becomes increasingly evident. This research serves as a foundational step for future advancements in automated summarization systems, paving the way for their integration into practical news consumption platforms.

REFERENCES

[1] P. Sethi, S. Sonawane, S. Khanwalker and R. B. Keskar, "Automatic text summarization of news articles," *2017 International Conference on Big Data, IoT and Data Science (BID)*, Pune, India, 2017, pp. 23-29, doi: 10.1109/BID.2017.8336568.

[2] Y. Du and H. Huo, "News Text Summarization Based on Multi-Feature and Fuzzy Logic,"

in *IEEE Access*, vol. 8, pp. 140261-140272, 2020, doi: 10.1109/ACCESS.2020.3007763.

[3] Yang Y, Tan Y, Min J, Huang Z. Automatic text summarization for government news reports based on multiple features. The Journal of Supercomputing. 2023 Aug 30:1-7.

[4] Esmaeilzadeh S, Peh GX, Xu A. Neural abstractive text summarization and fake news detection. arXiv preprint arXiv:1904.00788. 2019 Mar 24.

[5] Y. Chen and Q. Song, "News Text Summarization Method based on BART-TextRank Model," *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, Chongqing, China, 2021, pp. 2005-2010, doi: 10.1109/IAEAC50856.2021.9390683.

[6] 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)*, Thiruvananthapuram, India, 2017, pp. 19-24, doi: 10.1109/NETACT.2017.8076735.

[7] J. Cheng, F. Zhang and X. Guo, "A Syntax-Augmented and Headline-Aware Neural Text Summarization Method," in *IEEE Access*, vol. 8, pp. 218360-218371, 2020, doi: 10.1109/ACCESS.2020.3042886.