International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 01 | January - 2024 SJIF Rating: 8.176 ISSN: 2582-3930

News Summarization of BBC Articles: A Multi-Category Approach

Sherilyn Kevin[1], Amit Kumar Pandey[2], Bipin Yadav[3], Gopal Rajbhar[4]

[1]Assistant Professor(I.T), [2,3,4] PG Students
[1,2,3] Department of IT, Thakur College of Science and Commerce
Thakur Village, Kandivali (East), Mumbai-400101, Maharashtra, India
skevin3102gmail.com. amitpandey8089@gmail.com. bipinyadav1030@gmail.com.
rajbhargopal203@gmail.com

Abstract:

In this research project, we explore the application of advanced natural language processing techniques to automatically summarize news articles from the BBC. The dataset comprises five distinct categories—business, entertainment, politics, sport, and tech—each containing a wealth of information. Our primary goal is to develop an efficient and accurate news summarization system using state-of-the-art language models. We employ the Hugging Face Transformers library to create a summarization pipeline capable of extracting key information from lengthy news articles.

Introduction:

The exponential growth of digital content has made it challenging for readers to keep up with vast amounts of information. As a solution, automatic text summarization has gained significance, providing users with concise and informative summaries of lengthy articles. In this context, our research focuses on implementing a news summarization system tailored specifically for the diverse content found in the BBC News dataset. By categorizing articles into business, entertainment, politics, sport, and tech, we aim to address the unique challenges posed by each domain.

Through the use of cutting-edge language models and the Hugging Face Transformers library, we seek to develop a robust summarization pipeline. The system is designed to analyze and summarize news articles while maintaining the essence and key information across different categories. This research contributes to the field of natural language processing and information retrieval by providing a comprehensive solution for summarizing news content from one of the world's most reputable news sources, the BBC.

2.Literature Review:

The paper titled "Natural Language Processing (NLP) based Text Summarization - A Survey," authored by I. Awasthi, K. Gupta, P. S. Bhogal, S. S. Anand, and P. K. Soni, was presented at the 2021 6th International Conference on Inventive Computation Technologies (ICICT) in Coimbatore, India. The authors address the escalating size of data on the Internet over the past decade, emphasizing the urgent need for solutions that can transform voluminous raw information into a format comprehensible to the human brain. Within this context, the paper focuses on text summarization, a crucial research technique designed to cope with the abundance of data. The authors advocate for automatic summarization as a prominent approach, aiming to distill a document's main ideas into a condensed version that aligns with human cognitive capabilities. The study explores both Extractive and Abstractive methods of text summarization, with a particular emphasis on Extractive methods due to their appeal to researchers in Natural Language Processing (NLP). The authors delve into the intricacies of linguistic and statistical characteristics employed in extractive methods, contributing to a comprehensive understanding of text summarization techniques. The paper's analysis extends to the comparison of extractive and abstractive methods, with a focus on enhancing the quality of summaries by minimizing repetition and presenting a more concentrated overview of the information. In conclusion, the findings from this research contribute significantly to the field of NLP, offering valuable insights for advancing text summarization methodologies in the face of the ever-expanding volume of online data.[1]

The research paper, titled "Text Summarization Using Natural Language Processing: An Unsupervised Learning Approach," authored by Prudhvi K, Bharath Chowdary A, Subba Rami Reddy P, and Lakshmi Prasanna P, was presented at the Intelligent System Design conference in 2019. The authors delve into the challenges posed by manual text summarization in the age of automation, highlighting the growing need for efficient solutions. Text summarization, as a concept, has gained prominence due to its ability to address the laborious and time-consuming nature of manually summarizing textual content from various sources.

The primary goal of text summarization is outlined in the paper—to overcome the difficulties associated with manual extraction of crucial information from diverse sources such as text documents, online content, and social media platforms. The process involves extracting the main idea of the context or text and presenting it in a concise and meaningful summary. This demand for text summarization has surged with the exponential growth of data available on the Internet, including sources like research papers, news articles, and social media.

The authors underscore the efficiency of implementing text summarization using Natural Language Processing (NLP), leveraging the diverse packages and methods available in programming languages like Python or R. They establish a connection between text summarization and text mining, emphasizing that summaries are generated based on the classification of input text. The paper explores various approaches to text summarization, with a specific emphasis on an unsupervised learning approach. Within this framework, the authors employ the cosine similarity technique to measure the similarity between

sentences. The text rank algorithm is then utilized to generate a rank based on similarity, with top-ranked sentences forming the foundation of the final summarized text. This innovative methodology contributes valuable insights into the application of unsupervised learning techniques in text summarization, addressing the challenges posed by the vast volume of information in today's digital landscape.[2]

In their research article titled "Study of Automatic Text Summarization Approaches in Different Languages," Kumar Y, Kaur K, and Kaur S address the contemporary challenge of navigating vast amounts of information available across online and offline sources. With an abundance of articles on a single topic, manually extracting pertinent information becomes an arduous task. To tackle this issue, the authors delve into the development of automatic text summarization systems. Text summarization, as outlined in their work, involves extracting crucial information from extensive documents and condensing it into concise summaries while preserving essential content. The survey paper provides a comprehensive overview of research in automatic text summarization across various languages, with a particular focus on Indian languages like Hindi, Punjabi, Bengali, Malayalam, Kannada, Tamil, Marathi, Assamese, Konkani, Nepali, Odia, Sanskrit, Sindhi, Telugu, and Gujarati. Additionally, the authors explore foreign languages such as Arabic, Chinese, Greek, Persian, Turkish, Spanish, Czech, Romanian, Urdu, Indonesian, and more.

The paper aims to contribute knowledge and support to novice scientists in this research domain by offering a succinct perspective on diverse feature extraction methods and classification techniques essential for different types of text summarization approaches applied to both Indian and non-Indian languages. Overall, the research underscores the significance of automatic text summarization in managing the information overload prevalent in today's digital landscape and provides valuable insights for researchers exploring summarization techniques across diverse linguistic contexts.[3]

In the collaborative work by Gambhir M and Gupta V, titled "Recent Automatic Text Summarization Techniques: A Survey," the authors delve into the pressing need to condense vast amounts of information available on the internet into concise summaries. The paper, published in Artificial Intelligence Review in January 2017, underscores the growing interest within the research community to develop innovative approaches for automatic text summarization. Automatic text summarization systems play a pivotal role in generating short-length texts that encapsulate the essential information present in documents. The historical evolution of text summarization since the 1950s reveals continuous efforts by researchers to enhance techniques, aligning machine-generated summaries with their human-made counterparts.

The authors highlight the dichotomy between extractive and abstractive summarization methods, emphasizing the increased focus on extractive approaches due to their potential for creating more coherent and meaningful summaries. Over the past decade, numerous extractive approaches employing machine learning and optimization techniques have been developed, addressing the evolving needs of

automatic summary generation. The paper conducts a comprehensive survey of recent extractive text summarization approaches, detailing their specific requirements and presenting a comparative analysis of their advantages and disadvantages.

While extractive methods dominate the survey, the paper also touches upon abstractive and multilingual text summarization approaches. The challenging aspect of summary evaluation is thoroughly addressed, encompassing both intrinsic and extrinsic evaluation methods. The authors provide insights into text summarization evaluation conferences and workshops, presenting evaluation results on shared DUC datasets for extractive summarization approaches. The survey concludes with a discussion on future directions, offering valuable guidance to researchers in identifying areas for further exploration in the dynamic field of automatic text summarization.[4]

The collaborative work of A. P. Widyassari, A. Affandy, E. Noersasongko, A. Z. Fanani, A. Syukur, and R. S. Basuki, presented at the 2019 International Conference on Information and Communications Technology (ICOIACT) in Yogyakarta, Indonesia, delves into the evolving landscape of automatic text summarization. The paper defines automatic text summarization as the process of condensing one or more text documents while retaining essential information using automated machines. Spanning from the inception of text summarization research in the 1950s to the present, the authors emphasize the absence of a system capable of producing summaries comparable to those crafted by professionals or humans.

Utilizing a systematic literature review (SLR) methodology, the paper scrutinizes the methods, datasets, and trends in automatic text summarization research between 2015 and 2019. The findings highlight the continued relevance of automatic text summarization research, with a prevailing demand for the extractive approach over the past three years. The extractive method remains popular due to its perceived simplicity compared to the abstractive approach, presenting opportunities for method combination, such as integrating neuro computing approaches like the emerging Deep Q-Network (DQN), showcasing promising and improved results. The paper identifies a shift in the text summarization research trend towards optimization in the last three years, emphasizing efforts to enhance performance and achieve high accuracy in summarization processes.[5]

In their collaborative work presented at the 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC) in Erode, India, Rahul, S. Adhikari, and Monika delve into the significance of text summarization in the current data-rich environment. The paper emphasizes the crucial role of summarization in distilling the right amount of information from extensive texts, such as lengthy articles on news websites, blogs, and customer review platforms. Focusing on both Abstractive (ABS) and Extractive (EXT) summarization methods, the review paper surveys various approaches employed in generating summaries for large texts.

The authors explore query-based summarization techniques and delve into structured and semantic-based approaches for summarizing text documents. Throughout the paper, different methods are scrutinized, and insights into their tendencies, achievements, past work, and future prospects in text summarization and related fields are discussed. Datasets like the CNN corpus, DUC2000, and various single and multiple text documents are employed to evaluate the effectiveness of the summarization models presented in the reviewed literature. This comprehensive review aims to provide a holistic understanding of the diverse approaches used in text summarization, offering valuable insights into the current state and future directions of this field.[6]

In their contribution to the 2018 IEEE International Conference on Information Reuse and Integration (IRI), M.-Y. Day and C.-Y. Chen address the pivotal role of automatic text summarization in extracting key information from the exponentially growing volume of data facilitated by technological advancements. The authors identify a gap in existing literature related to using artificial intelligence (AI) for generating titles or short summaries. Consequently, they present a novel AI approach for automatic text summarization, introducing a comprehensive system architecture consisting of three models: statistical, machine learning, and deep learning.

The study specifically focuses on the application of deep learning to train an AI model using essay titles and abstracts, aiming to generate candidate titles. The performance of the proposed system is evaluated using the ROUGE metric. Notably, the authors contribute to the field by proposing an AI-based automatic text summarization system that utilizes deep learning techniques to produce concise summaries from titles and abstracts sourced from the Web of Science (WOS) database. This research stands as a significant step towards leveraging advanced AI methodologies for effective text summarization, showcasing the potential of deep learning in handling the complex task of distilling relevant information from vast datasets.[7]

In their contribution to the 2017 International Conference on Big Data, IoT, and Data Science, P. Sethi, S. Sonawane, S. Khanwalker, and R. B. Keskar address the persistent challenges in achieving efficient automatic text summarization, particularly in the context of news articles. Acknowledging the growing size and abundance of online documents, the authors emphasize the pressing need for a streamlined and effective news summarization solution. Their proposed technique focuses on identifying the most crucial sections of the text and generating coherent summaries without requiring full semantic interpretation.

The methodology presented in the paper relies on a model of topic progression derived from lexical chains, showcasing a practical approach that doesn't necessitate exhaustive semantic analysis. The authors introduce an optimized algorithm for generating text summaries, leveraging lexical chains and utilizing the WordNet thesaurus. To enhance the quality of summaries, the paper addresses limitations of the lexical chain approach by implementing pronoun resolution and introducing new scoring techniques

that take into account the structural nuances of news articles. This research contributes to the ongoing discourse on text summarization by providing a practical and efficient solution tailored specifically for news articles in the era of burgeoning online information.[8]

3.Methodology:

Our methodology involves a systematic approach to implement a news summarization system using the Hugging Face Transformers library. We start by organizing the BBC News dataset into five categories: business, entertainment, politics, sport, and tech. Each category serves as a distinct domain, presenting unique challenges in summarization.

For text processing, we utilize the Transformers library's summarization pipeline, configuring it with parameters such as max_length, min_length, length_penalty, num_beams, and no_repeat_ngram_size. These parameters are fine-tuned to balance the length and informativeness of generated summaries.

To evaluate the system, we employ a subset of articles from each category, ensuring a representative sample. Metrics such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation) are used to assess the quality of generated summaries by comparing them against reference summaries.

4. Results:

Our results demonstrate the effectiveness of the news summarization system across diverse categories. The ROUGE scores indicate the system's ability to capture key information while maintaining coherence and relevance. The system successfully adapts to the nuances of each category, producing concise and informative summaries for business, entertainment, politics, sport, and tech articles.

Quantitative evaluations showcase the system's proficiency in generating summaries that align with reference summaries. The variability in parameters allows for customization, ensuring optimal performance across different domains. The implementation of the Hugging Face Transformers library proves instrumental in achieving state-of-the-art results in news summarization.

5. Conclusion:

In conclusion, our research project provides a robust solution for automatic news summarization, specifically tailored for the BBC News dataset. The successful implementation of the summarization pipeline across diverse categories underscores its adaptability and effectiveness. The project contributes to advancing natural language processing applications by addressing the challenges posed by varied news content.

The customizable nature of our approach, facilitated by the Hugging Face Transformers library, ensures versatility and applicability to different datasets and domains. As we continue to refine and expand the system, it holds promise for enhancing information retrieval and user experience in navigating vast amounts of news content. This research lays the foundation for future developments in automated summarization systems and their integration into real-world news consumption platforms.

6. References:

1. I. Awasthi, K. Gupta, P. S. Bhogal, S. S. Anand and P. K. Soni, "Natural Language Processing (NLP) based Text Summarization - A Survey," *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India, 2021, pp. 1310-1317, doi: 10.1109/ICICT50816.2021.9358703.
2. Prudhvi K, Bharath Chowdary A, Subba Rami Reddy P, Lakshmi Prasanna P. Text summarization using natural language processing. InIntelligent System Design: Proceedings of Intelligent System Design: INDIA 2019 2020 Aug 11 (pp. 535-547). Singapore: Springer Singapore.
3. Kumar Y, Kaur K, Kaur S. Study of automatic text summarization approaches in different languages. Artificial Intelligence Review. 2021 Dec;54(8):5897-929.
4. Gambhir M, Gupta V. Recent automatic text summarization techniques: a survey. Artificial Intelligence Review. 2017 Jan;47:1-66.
5. A. P. Widyassari, A. Affandy, E. Noersasongko, A. Z. Fanani, A. Syukur and R. S. Basuki, "Literature Review of Automatic Text Summarization: Research Trend, Dataset and Method," *2019 International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, Indonesia, 2019, pp. 491-496, doi: 10.1109/ICOIACT46704.2019.8938454.
6. Rahul, S. Adhikari and Monika, "NLP based Machine Learning Approaches for Text Summarization," *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2020, pp. 535-538, doi: 10.1109/ICCMC48092.2020.ICCMC-00099.
7. M. -Y. Day and C. -Y. Chen, "Artificial Intelligence for Automatic Text Summarization," *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, Salt Lake City, UT, USA, 2018, pp. 478-484, doi: 10.1109/IRI.2018.00076.
8. 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.