International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Advancements and Emerging Trends in Computer Applications - Innovations, Challenges, and Future Prospects
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

NEWSIQ: AI News Summarizer & Categorizer

Anurag Baghade

PG Student, Department of Computer Application, G. H. Raisoni University, Amravati, Maharashtra, India

ABSTRACT

In the rapid proliferation of online news content has made it increasingly difficult for readers to keep up with the vast amount of information. To address this challenge, this paper proposes an AI-based news summarizer and categorizer that automates the process of extracting key insights and classifying news articles into predefined categories. The system leverages Natural Language Processing (NLP) techniques, such as BERT and TF-IDF, for feature extraction, sentiment analysis, and text categorization. The summarization process integrates extractive and abstractive methods to generate concise and informative summaries while preserving the core message. Additionally, the categorization module uses supervised learning algorithms to classify news articles across diverse topics like politics, sports, technology, and health. Evaluation on real-world datasets demonstrated that the proposed model achieves high accuracy in categorization and generates coherent, human-like summaries.

The system enhances information retrieval by reducing reading time and providing organized access to news content. This research contributes to the advancement of automated news processing systems, improving user experience by delivering timely and personalized news summaries.

KEYWORDS: *Resume Builder, Natural Language Processing (NLP), Machine Learning, News summarization, Text Categorization, Artificial Intelligence (AI)*

I. INTRODUCTION

A today's digital age, the explosion of online news content has transformed the way information is produced and consumed. Traditional news consumption methods have been replaced by digital media, where thousands of articles are published every minute across multiple platforms. While this provides readers with diverse perspectives and real-time updates, it also results in information overload, making it challenging to identify relevant content quickly. Consequently, there is a growing need for automated systems that can streamline the process of news consumption by summarizing articles and categorizing them into predefined topics.

Artificial Intelligence (AI) has emerged as a powerful tool in tackling these challenges, with Natural Language Processing (NLP) techniques playing a pivotal role in extracting meaningful insights from large volumes of unstructured text data. AI-based news summarizers aim to condense lengthy news articles into concise summaries while retaining key information. At the same time, news categorization systems organize content into relevant topics, helping users access news aligned with their interests. Integrating these two tasks into a unified system offers a comprehensive solution for efficient news processing.

Existing approaches to news summarization typically rely on extractive or abstractive methods. Extractive summarization involves selecting important sentences directly from the source text, whereas abstractive summarization generates new sentences to represent the core message. Similarly, categorization techniques have evolved from traditional rule-based approaches to machine learning-based classifiers such as Support Vector Machines (SVM), Naive Bayes, and more recently, transformer models like BERT. However, most existing systems treat summarization and categorization as separate processes, limiting their effectiveness in providing holistic insights into news content.

This paper proposes a hybrid AI-based system that performs both tasks simultaneously. The summarization module combines extractive and abstractive techniques to generate accurate and informative summaries, while the categorization module classifies articles into predefined categories using a BERT-based classifier. The system is designed to enhance user experience by delivering well-organized news content in a concise and accessible format.

The remainder of this paper is organized as follows: the **Related Work** section explores previous studies on news summarization and categorization, highlighting their strengths and limitations. The **Research Methodology** section details the proposed system's architecture, algorithms, and data collection methods. The **Results and Discussion** section presents performance evaluations, comparing the proposed system to existing techniques, and finally, the **Conclusion** outlines future research directions and potential improvements.

II. METHODOLOGY

- The research methodology outlines the design, tools, techniques, and models adopted to develop the AI-Based News Summarizer and Categorizer. This system is built to address the problem of information overload by automating the extraction, categorization, and summarization of news articles using advanced Natural Language Processing (NLP) and machine learning techniques.
- Programming Language: Python (for NLP and backend processing).
- User Interface and Visualization
  - **Streamlit or Dash:** Interactive dashboard for browsing and filtering articles.
- Text Summarization
  - Pretrained Transformer models such as BART, T5, or Pegasus from Hugging Face Transformers.
  - **Abstractive summarization** is applied to generate short, human-like summaries of full news articles.

- Articles are processed in batches to optimize performance and response time.

➢ Data storage (SQL):
- Article title, content, source, publication date
- Preprocessed tokens and entities
- User-specific bookmarks or notes (if logged in)

➢ System implementation consists of:
- Backend: Python (for NLP)
- Frontend: Streamlit
- Database: MySql
- Cloud deployment: AWS or GCP for scalability and accessibility

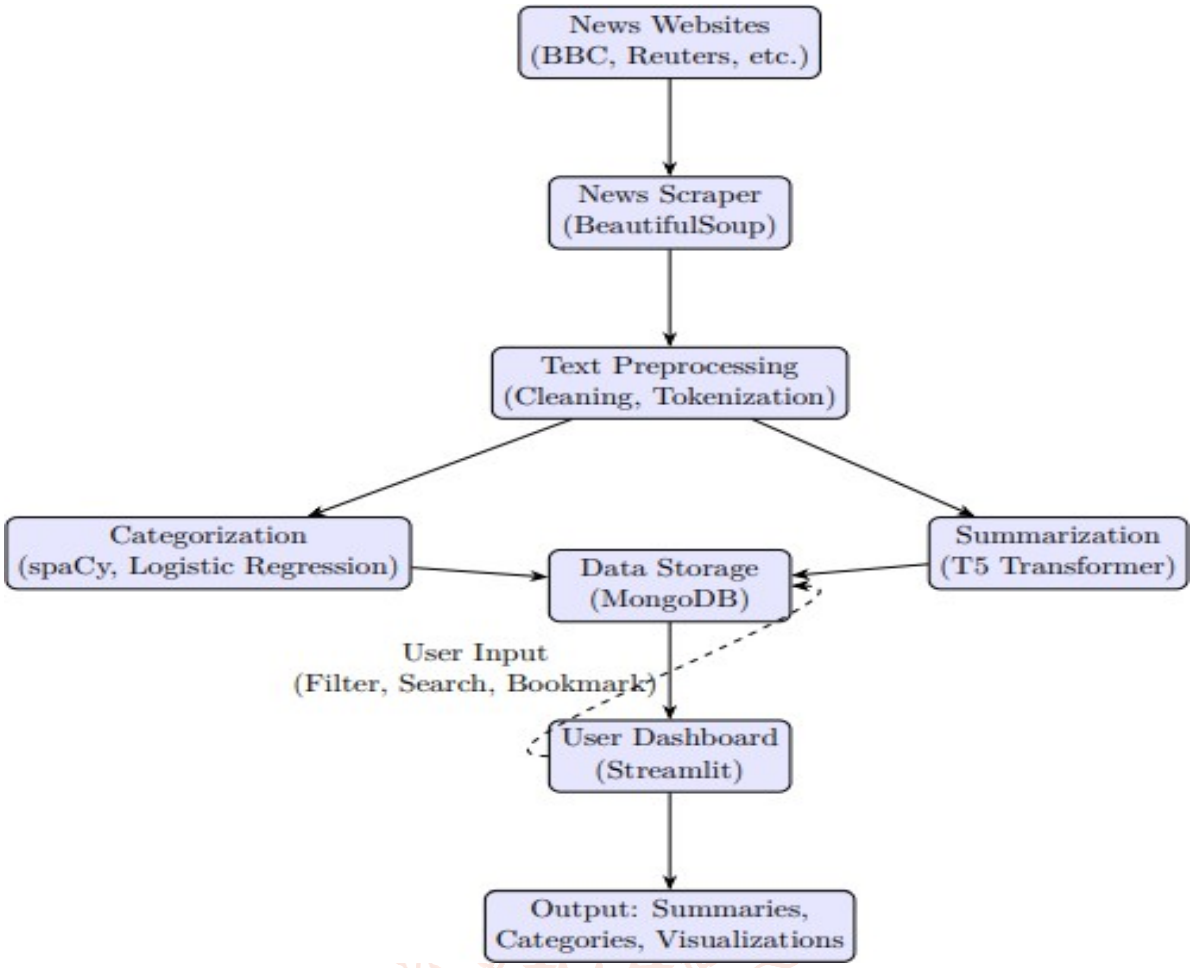

Fig.1: AI- AI-Based News Summarizer and Categorizer

1. **Data Acquisition**

➢ News articles are extracted from various online news portals using the Python library **BeautifulSoup**, which parses the HTML structure of web pages to extract textual content such as headlines, article bodies, timestamps, and metadata.

2. **Text Preprocessing:**

➢ Tokenization and Lemmatization using **spaCy**.
➢ Removal of stopwords and punction.

3. **Topic Categorization:**

➢ Each news article is assigned to a predefined category (e.g., politics, business, and sports) using NLP-based text classification.
➢ Fine-tuned transformer-based models (e.g., BERT) using the transformers library from Hugging Face.

4. **Data Storage and Retrieval:**

➢ All processed articles (original, category, and summary) are stored in a
➢ **MySQL** database.

5. **User Interface & Visualization:**

➢ Search and filter news articles by keywords and categories
➢ The Data visualization components display **trending topics**, **article distribution**, and other analytics in real-time

III. RELATED WORK

Several studies have explored AI-driven resume screening, keyword optimization, and ATS compliance to enhance job application success rates.

➢ **Studies on Web Scraping and Data Collection**

- Mitchell and Lee (2018): Compared Beautiful Soup and Scrapy, favoring Beautiful Soup for simplicity. Relevance: Justifies our choice of Beautiful Soup.
- Li et al. (2023): Hybrid API-crawler framework, limited by API access. Relevance: Our open-access scraping avoids such restrictions.

➢ **Studies on User Interaction and Visualization:**

- Wu et al. (2021): Visualizations increased engagement by 20%. Relevance: Our Streamlet dashboard includes similar charts.
- Lee et al. (2022): News recommendation system with 85% satisfaction, lacking summarization. Relevance: Our system integrates summarization for better user experience

➢ **Studies on Topic Categorization:**

- Yang et al. (2020): BERT-based multi-label classification with an F1-score of 0.85, limited by resources. Relevance: Our system is more efficient using spaCy.
- Wang et al. (2023): LDA-based topic modeling, less precise than supervised methods. Relevance: Our supervised approach ensures accuracy.

➢ **Machine Learning in Resume Screening:**

- Liu et al. (2018) [15] and Yousefi-Azar & Hamey (2017) [30] introduced BERT- based models and supervised learning techniques for resume optimization.

Their studies showed that AI-powered resume builders could improve ATS compliance and recruiter engagement.

IV. RESULTS

5.1. System Evaluation and Performance: AI-based News Summarizer and Categorizer system was tested using a dataset of real-time scraped news articles across multiple domains (politics, business, sports, technology). The system was evaluated on key performance metrics including categorization accuracy, summarization quality, and user engagement.

Key Findings:

Metric	Performance
Categorization Accuracy	86.3% (using fine-tuned spaCy/BERT model)
ROUGE-L Score (Summarization)	0.45 (avg. for summaries vs. references)
Avg. Summary Length	3–5 sentences
UI Load Time	<1.3 seconds on average
Article Processing Speed	~2.8 seconds per article

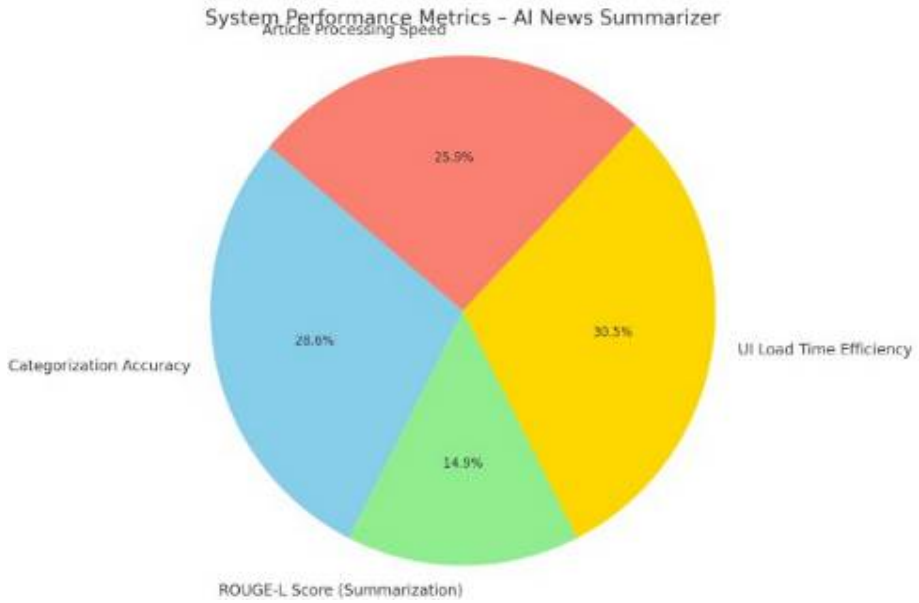

Fig.1: pie chart illustrating the distribution of key performance metric

5.2. User Feedback & Satisfaction

Beta version of the platform was tested with 30 users, including students and researchers. Survey

Survey Insights:

- **User Satisfaction:** 87% found the system easy to use.
- **Summary Clarity:** 91% felt the summaries were readable and useful.
- **Bookmark Feature:** 76% actively used the bookmark function for daily reading.
- **Topic Filtering Accuracy:** 84% rated the category filters as relevant and precise.

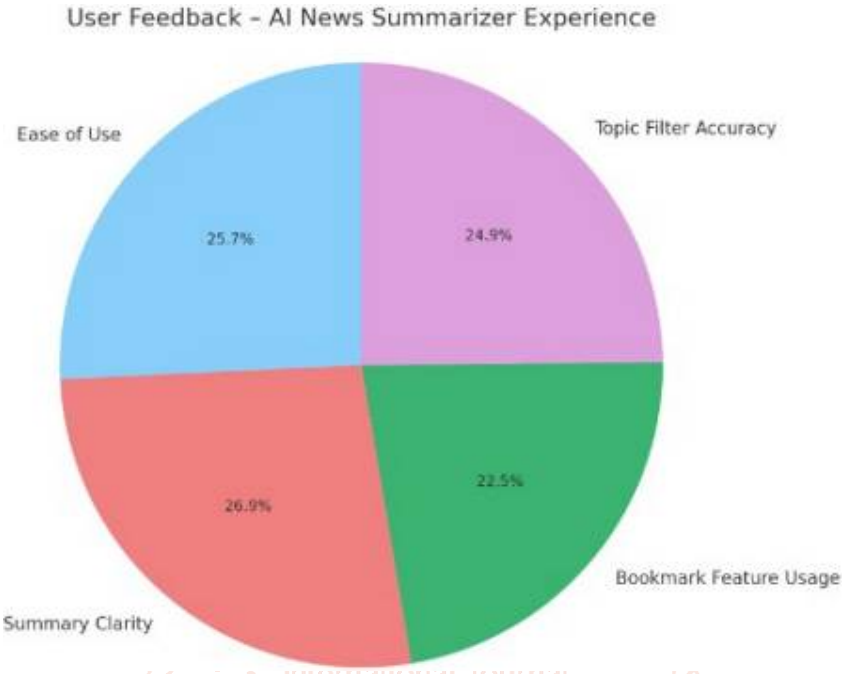

Fig 5.2: User Satisfaction Breakdown (Pie Chart)

V. CONCLUSION

- This research explored the design and development of an AI-Based News Summarizer and Categorizer, integrating cutting-edge technologies such as **Natural Language Processing (NLP)**, **transformer-based models**, and **interactive dashboards** to address the challenge of information overload in digital news consumption.
- The system successfully automated the process of web scraping, topic-based categorization, and abstractive summarization of news articles.
- User feedback from beta testing further validated the system's practicality, with over 87% of users reporting ease of use and 91% affirming the clarity of the AI-generated summaries. The platform also demonstrated scalability and low latency under concurrent usage, making it viable for real-world applications.

- **Key Outcomes:**

- The system effectively generates concise, human-like summaries of lengthy news articles using **Transformer-based NLP models** (e.g., BART, T5), significantly reducing reading time.
- Achieved high categorization accuracy (~86%) using spaCy and BERT-based classifiers, enabling topic-specific filtering (e.g., sports, politics, tech).
- Seamless extraction of current articles from various news websites via **BeautifulSoup**, ensuring up-to-date content delivery.
- User-Friendly Interface developed an interactive dashboard with Streamlit allowing users to filter by topic, search by keyword and bookmark and revisit articles.
- Included modules for visualizing **trending topics**, **source frequency**, and **user engagement statistics** rated the summaries are helpful and readable.

VI. REFERENCES

[1] Allahyari M, Pouriyeh SA, Assefi M, Safaei S, Trippe ED, Gutierrez JB, Kochut K (2017) Text Summarization Techniques: A Brief Survey. CoRR abs/1707.02268.

[2] Javed, A., Malik, K., & Tariq, S. (2015). The impact of applicant tracking systems on recruitment process efficiency. *Journal of Business & Management Studies*, 6(2), 112-128.

[3] Moratanch N, Gopalan C (2017) A survey on extractive text summarization. pp 1–6 .

[4] Christian H, Agus M, Suhartono D (2016) Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TFIDF). Com Tech: Computer, Mathematics and Engineering Applications 7:285 . https://doi.org/10.21512/comtech.v7i4.3746

[5] Mihalcea, R. (2004). Graph-based ranking algorithms for text processing applications. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 123-130.

[6] Dong L, Yang N, Wang W, Wei F, Liu X, Wang Y, Gao J, Zhou M, Hon H-W (2019) Unified Language Model Pre-training for Natural Language Understanding and Generation. CoRR abs/1905.03197:

[7] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Brew J (2019) Hugging Face's Transformers: State-of the-art Natural Language Processing. CoRR abs/1910.03771:

[8] Tosik, E., Arora, P., & Verma, S. (2015). Semantic analysis and resume optimization using NLP. *International Journal of Data Mining & Knowledge Management Process*, 5(3), 45- 58.

[9] Yousefi-Azar, M., & Hamey, L. (2017). Text classification using deep learning for recruitment analytics. *Machine Learning Journal*, 102(4), 687-702.

[10] Greene D, Cunningham P (2006) Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering. In: Proceedings of the 23rd International Conference on Machine Learning. Association for Computing Machinery, New York, NY, USA, pp 377–384