Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.6712

Intelligent News Aggregation System with Automatic Classification, Clustering, and Summarization

Ihsan Ghozi Zulfikar[1*], Yudi Wibisono[2], Asep Wahyudin[3]
[1,2,3]Universitas Pendidikan Indonesia, Indonesia
[1]ihsanghozizulfikar@upi.edu, [2]yudi@upi.edu, [3]away@upi.edu

open access

***Corresponding Author**

Article History:
Submitted: 29-07-2025
Accepted: 03-08-2025
Published: 11-08-2025

Keywords:
Artificial Intelligence; BART; BLSTM-2DCNN; Clustering; K-means; News Aggregation; Summarization; Text Classification.

ABSTRACT

The rapid growth of online news content has made it increasingly difficult for users to access relevant information efficiently. This study presents the development of an intelligent web-based news aggregation system that performs automatic classification, clustering, and summarization of Indonesian-language news articles. The system aims to enhance the news reading experience by organizing articles by category and topic, and by providing concise summaries. The system was built using the ADDIE development model, with each AI component trained and evaluated separately. News classification is handled by a BLSTM-2DCNN model trained on the IndoSum dataset, achieving 86% accuracy and an F1-score of 0.85. This model was also applied to classify 37,187 real-world articles scraped from Kompas and TribunNews during June 2025. Topic clustering is performed using K-means with entropy-weighted Bag-of-Words features over 5-day sliding windows. The clustering quality, evaluated using the Calinski-Harabasz Index, ranged from 5.21 to 525.44 with an average of 80.53, indicating varying cluster cohesion. For summarization, a fine-tuned BART model was used to summarize the article closest to each cluster's centroid. The model achieved ROUGE scores of 0.6389 (ROUGE-1), 0.5458 (ROUGE-2), and 0.6017 (ROUGE-L). The integrated system automatically scrapes news, classifies and clusters articles, and displays generated summaries through a user-friendly web interface. The results show that combining deep learning and natural language processing offers an effective approach for intelligent news aggregation, helping users consume news faster and more meaningfully.

INTRODUCTION

News is a vital part of daily life and plays a crucial role in shaping public opinion on socio-political issues. In 2024, the Indonesian Press Council reported over 1,800 verified media outlets in Indonesia, with more than 1,000 operating online. This rapid growth of digital news sources has led to information overload, where users struggle to find relevant news efficiently. This often results in news avoidance behavior, especially when faced with redundant, irrelevant, or excessive information (Tian, 2022).

News aggregator platforms aim to address this problem by collecting articles from multiple sources into one centralized interface. These platforms benefit users by streamlining access and support publishers by increasing visibility and user engagement. For instance, the shutdown of Google News in Spain in 2014 led to a significant decline in news traffic, illustrating the importance of aggregation systems for both readers and media outlets (Athey et al. 2021).

Despite their advantages, many existing news aggregator platforms do not incorporate personalization or content curation mechanisms. Without classification or filtering, these platforms may still present irrelevant or duplicated articles. To optimize the user experience, intelligent content processing is required, including news categorization, topic-based clustering, and automatic summarization.

Recent studies show that machine learning algorithms such as SVM, logistic regression, and neural networks can achieve high accuracy in news classification (Yeasmin et al., 2021). Additionally, clustering methods like K-means are effective in grouping unlabeled data based on topic similarity. Lastly, recent advancements in natural language processing, particularly the use of large language models (LLMs), have enabled automatic summarization that outperforms human-written summaries in both coherence and content retention (Liu et al., 2023).

In this study, we propose a comprehensive system for intelligent news aggregation, integrating three core components: automatic news classification using the BLSTM-2DCNN architecture, topic-based clustering using the K-means algorithm, and text summarization using a fine-tuned BART model. This system is built as a web-based application to facilitate personalized news consumption. The integration of these components is expected to significantly reduce users' cognitive load and improve engagement by providing relevant, clustered, and concise news content.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.6712

LITERATURE REVIEW

The development of intelligent news aggregation systems integrates multiple fields of study, including text classification, unsupervised clustering, and natural language generation. Each component has been extensively studied, and this section highlights key research supporting the methods used in this study.

News Aggregation Platforms

News aggregation platforms collect articles from various sources and present them in a unified interface. Studies have shown that such platforms increase user engagement and streamline content consumption (Hoe et al., 2024). Athey et al. (2021) reported that the removal of Google News in Spain led to a significant decline in site traffic for major news outlets. In Indonesia, platforms like Google News and Opera News aggregate content from popular media such as Kompas, Detik, and TribunNews.

Web Scraping

Web scraping is a widely used method for the automated retrieval of online content. It enables applications to extract large volumes of information from websites in a structured manner without manual intervention. Khder (2021) divides the web scraping process into three main stages: fetching, extraction, and transformation. In the fetching stage, the target website is accessed using the HTTP protocol via libraries such as curl or wget, which send HTTP GET requests and receive HTML content as responses. The extraction stage involves parsing the HTML document to retrieve relevant data using tools like regular expressions (regex), HTML parsers, or XPath queries. Finally, in the transformation stage, the extracted data is normalized into a consistent structure to facilitate storage or direct usage in downstream systems.

Text Classification

Text classification is the process of assigning predefined categories to textual data. It has been widely used for spam detection, sentiment analysis, and news categorization. Recent approaches leverage deep learning models such as Bidirectional Long Short-Term Memory (BLSTM) for capturing sequential context and Convolutional Neural Networks (CNN) for extracting local features. Peng et al. (2016) introduced the BLSTM-2DCNN architecture, which achieved state-of-the-art accuracy in the 20NG dataset, outperforming traditional models like SVM and logistic regression.

Text Clustering

Text clustering is an unsupervised learning technique used to group similar documents based on their content without relying on predefined labels. In the context of news aggregation, clustering plays a vital role in identifying emerging topics and organizing articles with similar themes into coherent groups. This allows users to navigate large volumes of information more effectively and enables systems to generate topic-based summaries.

One of the most widely used algorithms for document clustering is K-means, primarily due to its simplicity and computational efficiency. The algorithm works by partitioning the dataset into clusters, each represented by a centroid, and iteratively minimizing the variance within clusters. Rashid et al. (2020) showed that K-means consistently outperformed other popular methods such as Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) when used to group news articles based on topic similarity.

To assess clustering performance, this study employs the Calinski-Harabasz Index (CH-Index). The CH-Index evaluates the quality of clustering by analyzing the dispersion ratio between and within clusters (Effendie, 2023). It is calculated using the between-cluster sum of squares (BCSS), which quantifies how far apart the centroids of different clusters are, and the within-cluster sum of squares (WCSS), which measures how close the data points are to the centroid of their respective cluster. A higher CH-Index value indicates better clustering performance, reflecting well-separated and internally coherent clusters.

Automatic Summarization

Automatic summarization is the task of condensing a long piece of text into a shorter version while preserving its core meaning and key information. There are two main approaches to summarization: extractive and abstractive (El-Kassas et al., 2021). Extractive summarization identifies and selects important sentences directly from the source text, whereas abstractive summarization generates new sentences that may not exist in the original text but are semantically equivalent.

With the emergence of transformer-based architectures, abstractive summarization has seen significant advancements. One of the most effective models in this area is BART (Bidirectional and Auto-Regressive Transformer). BART combines the strengths of encoder-decoder architectures with denoising autoencoding pretraining to produce fluent, high-quality summaries. According to Chen et al. (2023), BART achieved superior ROUGE scores compared to other advanced models like ASGARD and MultiBART-GAT on the CNN/DailyMail dataset, highlighting its effectiveness in complex summarization tasks.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.6712

The performance of summarization models is commonly evaluated using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric developed by Lin (2004). ROUGE compares the model-generated summaries with human-written reference summaries using overlapping units such as n-grams, word sequences, and word pairs. ROUGE-1 focuses on unigram overlap, ROUGE-2 on bigram overlap, and ROUGE-L on the longest common subsequence between texts. A higher ROUGE score indicates greater similarity to human summaries, and thus better summarization quality.

Dataset

For this research, the IndoSum dataset was used as the primary benchmark for training and evaluating the classification and summarization model. IndoSum was introduced by Kurniawan and Louvan (2018) and consists of approximately 20,000 Indonesian-language news articles, each paired with a human-written summary. The articles are sourced from reputable local media such as CNN Indonesia and Kumparan. The dataset is well-structured and contains no missing or null values, making it suitable for both classification and summarization tasks.

METHOD

This research follows the ADDIE instructional design framework, which consists of five stages: Analyze, Design, Develop, Implement, and Evaluate. ADDIE was selected as the underlying framework to ensure a systematic approach in developing educational and interactive systems. The use of ADDIE allows for iterative refinement and evaluation throughout the development cycle.

The software development process for the news aggregation system was carried out using the Waterfall model, which provides a structured and linear sequence of stages including requirements analysis, system design, implementation, testing, and deployment. This model was chosen for its clarity in phase separation and ease of documentation, particularly suited for research-based software projects with well-defined goals, as outlined by Saravanos and Curinga (2023). Each AI component was constructed and evaluated separately.

System Architecture

The proposed system integrates four key components: web scraping, news classification, topic clustering, and summarization, into an automated pipeline for Indonesian-language news aggregation. The system is implemented as a web-based application, allowing users to access clustered and summarized news content directly from a browser. The architecture is illustrated in Figure 1.

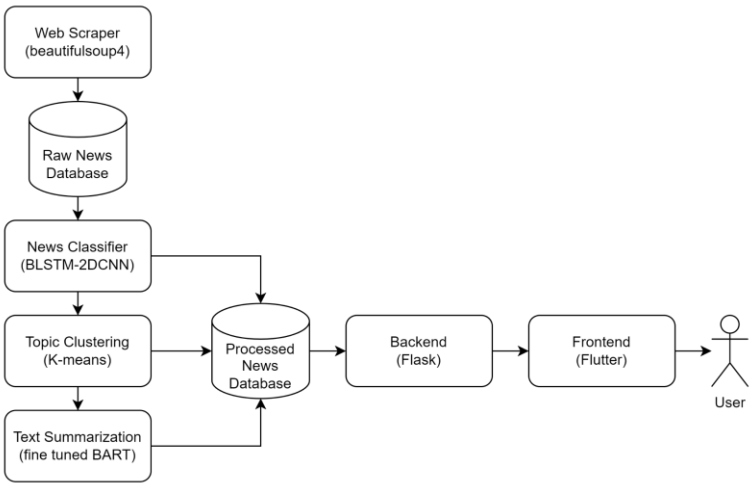

Fig. 1 System Architecture Diagram

Data Collection

News data were collected through web scraping from two major Indonesian news websites, Kompas and TribunNews, over a period of June 1 to June 30, 2025. The scraping process focused on extracting only the main body text of each article, omitting multimedia content, advertisements, and user comments. This real-time dataset was used for testing the deployed news aggregation system, including its classification, clustering, and summarization components.

For model training and evaluation, additional labeled data were obtained from the IndoSum dataset, a publicly available benchmark for Indonesian news summarization and classification tasks. The dataset contains 18,774 articles, each accompanied by human-written summaries and category labels. The distribution of articles across six categories in IndoSum is shown in Table 1.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.6712

Table 1. The Distribution of Articles Categories in Indosum

Metric	Value
Headline (Tajuk Utama)	7,192
Sports (Olahraga)	4,768
Showbiz	2,578
Technology (Teknologi)	2,303
Entertainment (Hiburan)	1,803
Inspiration (Inspirasi)	130

For supervised learning, the IndoSum dataset was divided into three subsets: 70% of the data (13,141 articles) was used for training, 20% (3,755 articles) for validation, and the remaining 10% (1,878 articles) for testing. This structured split allowed for effective model fine-tuning while ensuring an unbiased evaluation of the classification and summarization performance under realistic conditions.

News Classification with BLSTM-2DCNN

The news classification component is implemented using a BLSTM-2DCNN model adapted from Peng et al. (2016), combining the temporal capabilities of Bidirectional Long Short-Term Memory (BLSTM) networks with the spatial feature extraction power of two-dimensional convolutional layers (Conv2D). This hybrid architecture allows the model to capture both sequential patterns and local dependencies in textual data.

The model was trained on the IndoSum dataset to classify news articles into six predefined categories: headline, technology, sports, entertainment, inspiration, and showbiz. The architecture is outlined in Table 2, which details the structure and output shapes of each layer.

Table 2. BLSTM-2DCNN Classification Architecture

Layer Type	Output Shape
Input	(442)
Embedding	(442, 100)
SpatialDropout1D	(442, 100)
Bidirectional LSTM	(442, 64)
Reshape	(442, 64, 1)
Conv2D	(442, 64, 32)
BatchNormalization	(442, 64, 32)
MaxPooling2D	(442, 32, 32)
GlobalMaxPooling2D	(32)
Dense	(256)
Dropout	(256)
Output (Softmax)	(6)

The model was evaluated using accuracy and F1-score metrics on the IndoSum test set to assess its ability to handle imbalanced class distributions and multi-class classification.

News Clustering with K-means

To group news articles into topic-based clusters, the K-means clustering algorithm was applied separately to articles within each predefined category. Prior to clustering, each article was converted into a numerical representation using the Bag of Words (BoW) model, enhanced by global entropy-based weighting. This approach prioritized terms that were more informative and topic-specific across the corpus.

To capture temporal variation and ensure contextual relevance, clustering was performed using a 5-day sliding window strategy. For each sliding window, such as June 1-5, June 2-6, and so on, the system aggregated all articles within that time frame and performed clustering independently. The optimal number of clusters for each window was selected dynamically using the Elbow method based on within-cluster sum of squares.

Clustering quality was assessed using the Calinski-Harabasz (CH) Index, which measures the ratio of between-cluster dispersion to within-cluster cohesion. A higher CH score indicates more distinct and well-separated clusters.

News Summarization with BART

The summarization component utilizes a BART model fine-tuned on the IndoSum dataset, which consists of Indonesian news summaries. Rather than summarizing all articles within each cluster, only the article closest to the cluster centroid, determined during the K-means clustering process, is selected as the representative input for summarization. This strategy ensures that the most central and relevant article for each topic is condensed into a single,

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.6712

coherent summary. The model's performance is evaluated using standard ROUGE metrics: ROUGE-1, ROUGE-2, and ROUGE-L.

RESULT

This section presents the outcomes of each component in the proposed intelligent news aggregation system: classification, clustering, and summarization. All experiments were conducted using a machine with GPU acceleration on Kaggle.

News Classification Results

The trained BLSTM-2DCNN model achieved strong performance when evaluated on the IndoSum test set, as shown below:

Table 3. Performance of BLSTM-2DCNN on Test Set

Metric	Value
Accuracy	86%
F1 score	0.85

These results indicate the model's robustness in identifying various categories, even in the presence of class imbalance. The high F1-score particularly suggests reliable precision and recall across most categories.

In addition to the evaluation on the IndoSum dataset, the model was also applied to classify real-world news articles collected through web scraping. During the period of June 1-30, 2025, a total of 37,187 articles were collected from Kompas and TribunNews and passed through the classification model. The distribution of predicted categories is shown in Table 4.

Table 4. Distribution of Predicted News Categories

Category	Number of Articles	Percentage (%)
Headline (Tajuk Utama)	22,591	60.75
Sports (Olahraga)	3,592	9.66
Showbiz	4,685	12.60
Technology (Teknologi)	715	1.92
Entertainment (Hiburan)	5,604	15.07
Inspiration (Inspirasi)	0	0

The classification results reveal that the "Headline" category dominated the prediction outcomes, followed by "Entertainment" and "Showbiz". Notably, no articles were classified under "Inspiration", which may reflect the very limited number of training samples in this category, only 130 articles in the entire IndoSum dataset. This likely caused the model to underfit and fail to generalize effectively for this class.

These results highlight both the strengths and current limitations of the classification component. Future work may address class imbalance through data augmentation or alternative modeling approaches such as focal loss or ensemble classifiers.

News Clustering Results

Following classification, articles in each category were clustered into topic groups using the K-means algorithm. Each article was represented using Bag of Words (BoW) vectors with entropy-based term weighting to enhance semantic differentiation. The clustering process was repeated across a sequence of 5-day sliding windows throughout the month of June 2025 to capture temporal topic dynamics. The CH Index values observed across all sliding windows and categories ranged from 5.21 to 525.44, with an average of 80.53 and a median of 38.93.

This wide distribution, particularly the gap between the mean and median, indicates a right-skewed distribution of cluster quality. In other words, while a small subset of clusterings achieved excellent separation and compactness, the majority of windows yielded moderate to low clustering performance. This variability can be attributed to: the homogeneity of topics within certain short timeframes, making it difficult to form distinct groups, and a limited number of articles in some windows, reducing the statistical strength of the clustering process.

These findings suggest that while the method performs well under certain conditions, the system may benefit from adaptive strategies for window size, richer semantic embeddings, or hybrid clustering approaches in future enhancements.

Summarization Results

To assess the quality of the generated summaries, we used ROUGE metrics, a widely adopted evaluation standard in automatic summarization. The table below compares the performance of the pre-trained BART model and its fine-tuned version on the IndoSum dataset.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.6712

Table 5. Summarization Result Evaluated Using ROUGE

Metric	BART (pretrained)	BART (fine-tuned)
ROUGE-1	0.6102	0.6389
ROUGE-2	0.5303	0.5458
ROUGE-L	0.5788	0.6017

These results indicate that fine-tuning enables the model to better capture key information (ROUGE-1), sequential phrasing (ROUGE-2), and overall structural coherence (ROUGE-L). The fine-tuned model is able to generate summaries that are more fluent, concise, and semantically aligned with the human-written references. An example comparison between reference summaries and generated outputs from the test dataset is presented in Table 6.

Table 6. Example of Reference vs. Generated Summary

No	Reference Summary	Generated Summary (Fine-tuned BART)
1	Mantan Ketua Komisi Pemberantasan Korupsi (KPK) Antasari Azhar mengaku tak punya masalah dengan Presiden keenam Republik Indonesia Susilo Bambang Yudhoyono. Dirinya menegaskan, acara syukuran ini bukan diselenggarakan untuk menghadirkan polemik baru dan tak ingin menciptakan suasana gaduh. Ia pun meminta awak media tidak mempertanyakan perihal tokoh-tokoh yang ia undang. Antasari mengaku hanya ingin tenang sambil menikmati masa-masa kebebasannya.	Mantan Ketua Komisi Pemberantasan Korupsi (KPK) Antasari Azhar mengaku tak punya masalah dengan Presiden keenam Republik Indonesia Susilo Bambang Yudhoyono (SBY). Hal itu diungkapkannya saat menjawab pertanyaan terkait alasan SBY tak diundang dalam acara syukuran atas pembebasan bersyaratnya di salah satu hotel berbintang, Tangerang Selatan, Banten, Sabtu (26/11). Ia pun meminta awak media tidak mempertanyakan perihal tokoh-tokoh yang ia undang.
2	Menteri Keuangan Sri Mulyani Indrawati menegaskan pemerintah akan terus mengejar konglomerat yang tidak taat pajak yang salah satunya ditunjukkan tidak dimilikinya Nomor Pokok Wajib Pajak (NPWP). Dalam Undang-Undang Nomor 11/2016 tentang Pengampunan Pajak atau Tax Amnesty, kata dia, memang menyebutkan yang belum memiliki NPWP diperbolehkan mendaftarkan untuk mendapatkan NPWP dan tahap pertama sudah berjalan.	Menteri Keuangan Sri Mulyani Indrawati menegaskan pemerintah akan terus mengejar konglomerat yang tidak taat pajak yang salah satunya ditunjukkan tidak dimilikinya Nomor Pokok Wajib Pajak (NPWP). Pernyataan itu disampaikan Menkeu ketika menjawab pertanyaan peserta kuliah umum yang berlangsung di Gedung Prof Soedarto Undip itu mengenai apakah ada kongsomerat di Indonesia yang belum memiliki NPWP.

These examples demonstrate that the generated summaries are generally faithful to the original content and closely resemble human-written summaries in both structure and meaning. However, in some edge cases, generated outputs may omit finer contextual details, particularly when the original article is highly condensed or lacks clear topic focus.

In addition to evaluation on benchmark data, the summarization model was applied to articles obtained through real-time scraping. For each topic cluster produced by K-means, the article closest to the centroid was summarized using the fine-tuned BART model.

System Implementation

In addition to the machine learning models, the research includes the development of a web-based application that automates the entire news aggregation process. The system scrapes news articles from various Indonesian online media sources, classifies them by category, clusters them into topical groups, and presents summaries for the most central articles within each cluster.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.6712

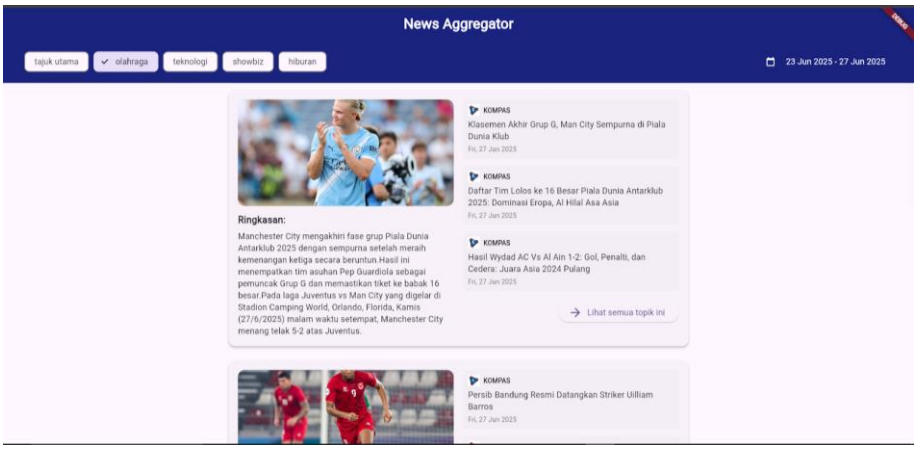

Fig. 2 Screenshot of the News Aggregation Website

The web interface provides a clean and user-friendly layout that displays a categorized feed of news clusters, each cluster's top summary with expandable details, and article sources, publication times, and optional redirection to the original content. The integration of deep learning with a live web frontend demonstrates the practical viability of the system, showing potential for real-world deployment and future personalization features.

DISCUSSION

The implementation of the proposed intelligent news aggregation system demonstrates the effectiveness of combining deep learning and natural language processing techniques in solving real-world problems related to news overload and content personalization. The classification results using the BLSTM-2DCNN model show that the architecture is capable of learning discriminative features from Indonesian news text, as evidenced by an accuracy of 86% and an F1-score of 0.85. However, the model failed to recognize underrepresented categories such as "Inspiration," indicating a need for improved data balancing or augmentation strategies.

The clustering results revealed that while certain 5-day time windows produced high-quality topic clusters, as reflected in CH Index scores exceeding 500, many clusters showed moderate to poor cohesion. This variability highlights the importance of content diversity and volume in clustering performance. The sliding window approach helped capture evolving topics, but clustering may be further improved by incorporating semantic embeddings (e.g., SBERT) instead of Bag of Words.

Summarization using the fine-tuned BART model yielded improved ROUGE scores compared to the pre-trained baseline, demonstrating that domain adaptation significantly enhances performance. The summaries were generally coherent and informative, especially for well-structured news topics. Nevertheless, some generated summaries lacked context when the source articles were short or weakly structured. Future refinements may include multi-document summarization strategies or reinforcement learning with human feedback.

The successful integration of these components into a functioning web application emphasizes the system's practical applicability. It provides a streamlined reading experience for users while maintaining the integrity of news content through automatic classification, clustering, and summarization. This research thus contributes both a methodological framework and a functional prototype that can be extended for real-world deployment.

CONCLUSION

This study presents the design and implementation of a web-based intelligent news aggregation system for Indonesian-language news. By integrating a BLSTM-2DCNN classification model, K-means topic clustering, and BART-based abstractive summarization into a single pipeline, the system addresses the challenges of redundant news content and lack of topic coherence in existing aggregator platforms.

The classification model achieved high performance on the IndoSum dataset, while topic clustering using 5-day sliding windows revealed both strong and weak grouping quality depending on article diversity and quantity. Summarization using a fine-tuned BART model produced coherent summaries, as confirmed by ROUGE evaluations and application to real-world scraped data.

Despite its promising performance, the system has limitations. It struggled to classify underrepresented categories and showed sensitivity to article volume in clustering. Summarization also depended heavily on input quality. Therefore, future research should focus on enhancing model generalization, introducing semantic representations for clustering, and incorporating user feedback mechanisms to refine summaries.

The proposed system offers practical benefits for news consumers, media platforms, and developers of AI-powered applications. It demonstrates the feasibility of using deep learning models for large-scale, automated news

E-ISSN : 2807-9035
Volume 5, **Number** 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.6712

curation and paves the way for further development of personalized news recommendation systems in Indonesia.

ACKNOWLEDGMENT

The authors would like to thank Universitas Pendidikan Indonesia and the Department of Computer Science for providing support, guidance, and resources throughout this research.

REFERENCES

Athey, S., Mobius, M., & Pal, J. (2021). *The impact of aggregators on internet news consumption* (No. w28746). National Bureau of Economic Research.

Chen, Tong, Xuewei Wang, Tianwei Yue, Xiaoyu Bai, Cindy X. Le, and Wenping Wang. 2023. "Enhancing Abstractive Summarization with Extracted Knowledge Graphs and Multi-Source Transformers." *Applied Sciences* (Switzerland) 13(13). doi: 10.3390/app13137753.

Dewan Pers. 2024. "SURVEI LANSKAP MEDIA PERS INDONESIA." Retrieved (https://dewanpers.or.id/berita/detail/2554/survei-lanskap-media-pers-indonesia).

Effendie, A. R. (2023). A medoid-based deviation ratio index to determine the number of clusters in a dataset. *MethodsX*, 10, 102084.

El-Kassas, Wafaa S., Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed. 2021. "Automatic Text Summarization: A Comprehensive Survey." *Expert Systems with Applications* 165:1–46. doi: 10.1016/j.eswa.2020.113679.

Hoe, C. W., Raheem, M., & Abubacker, N. F. (2024). News Aggregation And Summarisation. *Journal of Applied Technology and Innovation (e-ISSN: 2600-7304)*, 8(4), 51.

Khder, M. A. (2021). Web scraping or web crawling: State of art, techniques, approaches and application. *International Journal of Advances in Soft Computing & Its Applications*, 13(3).

Kurniawan, Kemal, and Samuel Louvan. 2018. "IndoSum: A New Benchmark Dataset for Indonesian Text Summarization." *Proceedings of the 2018 International Conference on Asian Language Processing*, IALP 2018 215–20. doi: 10.1109/IALP.2018.8629109.

Lin, Chin-Yew. 2004. "ROUGE: A Package for Automatic Evaluation of Summaries." *Japanese Circulation Journal*. doi: 10.1253/jcj.34.1213.

Liu, Yixin, Kejian Shi, Katherine S. He, Longtian Ye, Alexander R. Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2023. "On Learning to Summarize with Large Language Models as References."

Peng, Zhou, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Xu Bo. 2016. "Text Classification Improved by Integrating Bidirectional LSTM with Two-Dimensional Max Pooling.Pdf." 3485–3495.

Rashid, Junaid, Syed Muhammad Adnan Shah, and Aun Irtaza. 2020. "An Efficient Topic Modeling Approach for Text Mining and Information Retrieval through K-Means Clustering." *Mehran University Research Journal of Engineering and Technology* 39(1):213–22. doi: 10.22581/muet1982.2001.20.

Saravanos, A., & Curinga, M. X. (2023). Simulating the software development lifecycle: The waterfall model. *Applied System Innovation*, 6(6), 108.

Tian, Qiuxia. 2022. "Impact of Social Media News Overload on Social Media News Avoidance and Filtering: Moderating Effect of Media Literacy." F*rontiers in Psychology* 13(March). doi: 10.3389/fpsyg.2022.862626.

Yeasmin, Sharmin, Ratnadip Kuri, A. R. M. Mahamudul Hasan Rana, Ashraf Uddin, A. Q. M. Sala Uddin Pathan, and Hasnat Riaz. 2021. "Multi-Category Bangla News Classification Using Machine Learning Classifiers and Multi-Layer Dense Neural Network." *International Journal of Advanced Computer Science and Applications* 12(5):757–67. doi: 10.14569/IJACSA.2021.0120588.