IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

HYBRID NLP-BASED INTELLIGENT CONTENT ANALYSIS USING BERT

[1] Dr. Shrihari M R, [2] Khushi V, [3] Dr. Ajay N

[1]Associate Professor, Department of Computer Science and Engineering, SJCIT, Chikaballapur, India.

[2]Post Graduation Student Master of Technology, Department of Computer Science and Engineering SJCIT, Chikaballapur, India.

[3]Associate Professor, Department of Computer Science and Engineering, SJCIT, Chikaballapur, India

Abstract: Conventional text summarization relies heavily on frequency-based ranking methods such as TF-IDF and graph-based techniques like TextRank. While these methods offer computational efficiency, they consistently fall short in maintaining the logical flow and contextual accuracy of the source document. This study presents an improved summarization methodology that integrates both linguistic and semantic resources to overcome these shortcomings. The proposed approach balances summarization speed with output quality despite higher computational demands. A weighted feature scoring mechanism is employed, incorporating textual attributes such as sentence position, noun frequency, and named entity counts to enhance summary quality. The system was evaluated by processing sample documents through both the proposed and baseline summarizers and comparing output length, coherence, and content coverage.

Index Terms - dataset, extractive summarization of texts, sentence scoring, weighted features, ROUGE measure, and BERTScore metric component

I. INTRODUCTION

The rapid growth of digital information has increased the demand for efficient summarization techniques capable of extracting important information from large textual documents. Traditional frequency-based approaches often fail to preserve contextual meaning and semantic relationships, motivating the development of hybrid summarization methods that integrate statistical and semantic representations. BERT provides contextual representations by considering surrounding words during encoding. Integrating TF-IDF and BERT enables the proposed system to achieve a balance between computational efficiency and semantic understanding, as BERT generates contextual representations by considering the relationships between surrounding words, enabling more effective interpretation of ambiguous terms than frequency-based approaches. The integration of TF-IDF and BERT combines statistical importance with semantic understanding, thereby improving summarization quality. Furthermore, BERTScore was employed during evaluation because it measures semantic similarity between generated and reference summaries more effectively than lexical overlap metrics such as ROUGE alone.

II. LITERATURE REVIEW

Xie et al. (2024): A hybrid architecture combining BERT and FastText was developed by Xie et al. in 2024 for sentiment classification in educational settings, attaining 92% accuracy. Despite this strong result, the model showed limited ability to generalize across different subject domains.

Kapoor et al. (2024): In 2024, Kapoor and colleagues introduced a multimodal framework for sentiment recognition that achieved 91% accuracy. The system's main limitation was its difficulty in processing heterogeneous multimodal data effectively.

Mishra et al. (2023): Mishra and collaborators (2023) developed an NLP-based pipeline addressing both text summarization and sentiment analysis, attaining 89% accuracy. A notable drawback was the extensive domain-specific preprocessing required before deployment.

Sharma et al. (2022): Lexicon-driven sentiment classification on Twitter data was explored by Sharma et al. in 2022, yielding 85% reliability. The method was limited by its sensitivity to ambiguous and context-dependent language patterns.

Hassan et al. (2022): Hassan et al. (2022) proposed a visual sentiment analysis system designed for disaster-related image data, recording 87% accuracy. Its practical scope was narrow, as it was built specifically for disaster scenarios.

Urolagin et al. (2022): A Gabor CNN-based visual sentiment analysis framework was constructed by Urolagin and colleagues in 2022, achieving 89% effectiveness. The system's dependence on cloud-based computational resources was identified as a key practical limitation.

III. METHODOLOGY

This study proposes a hybrid content analysis system that integrates text summarization and sentiment analysis within a unified framework. The system accepts textual documents, performs preprocessing, generates semantic representations, and produces summaries and sentiment predictions through dedicated modules. The implemented system is a desktop application built using Python's Tkinter framework, providing two core functionalities through a navigable GUI: a content summarizer and a sentiment analyzer. Input documents can be uploaded directly in PDF or DOCX format, or text can be pasted manually. Uploaded files are parsed based on extension—PDF via PyPDF2, DOCX via python-docx—and passed through a cleaning function that removes PDF extraction artifacts, non-alphanumeric noise, and redundant whitespace.

A. Sentiment Evaluation

The sentiment module uses TextBlob to compute document-level polarity (ranging from -1 to +1) and subjectivity (0 to 1). Polarity above 0.1 is classified as positive, below -0.1 as negative, and values in between as neutral. In addition to the document-level score, the system generates a per-sentence breakdown for the first ten sentences, labeling each individually—allowing users to see not just an aggregate sentiment but where in the document that sentiment originates. A lexicon-based approach was chosen over a trained classifier for this module because it requires no labeled training data and runs with low latency, making it suitable for real-time use in an interactive GUI.

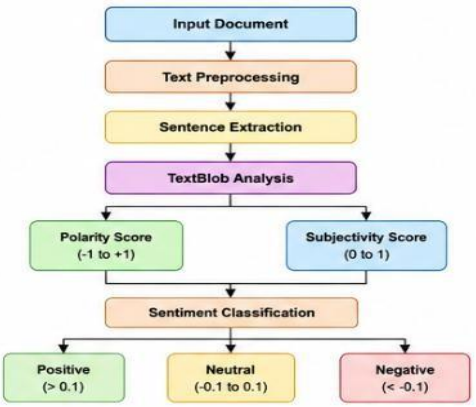

Fig. 1. Sentiment Analysis Flow

B. Summarization

The core summarization approach scores each sentence in the input document using three independent signals, which are then combined into a single ranking score. First, a TF-IDF vectorizer is fit across all sentences, and each sentence's score is the sum of its TF-IDF weights. Second, a pretrained sentence embedding model (all-MiniLM-L6-v2, from Sentence-Transformers) embeds each sentence and the document as a whole, and cosine similarity between them gives a semantic relevance score. Third, each sentence's absolute sentiment polarity, computed via TextBlob, gives a sentiment intensity score. All three scores are min-max normalized to [0,1] and combined with fixed weights: 40% TF-IDF, 40% semantic similarity, 20% sentiment. The top-N scoring sentences are selected and reassembled in original document order to preserve narrative flow.

To evaluate this method, a baseline summarizer was also implemented using traditional word-frequency scoring: stopwords are removed, word frequencies are tallied, and each sentence is scored as the sum of its word frequencies. Sentences scoring above 1.2 times the average are retained. Both methods run on the same input and display side by side for direct comparison.

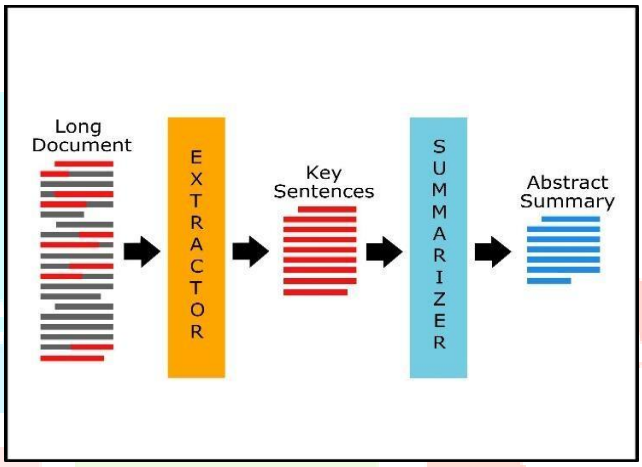

Fig. 2. Summarization Framework

C. Data Preprocessing

The input documents are initially converted into plain text using PyPDF2 for PDF files and python-docx for DOCX files. The extracted text is cleaned by removing unwanted symbols, redundant spaces, and extraction artifacts. The preprocessing stage also includes case normalization and punctuation handling to ensure consistency across textual inputs. Stop words are selectively removed during frequency-based processing while preserving sentiment-bearing terms and negations. This strategy minimizes noise in the textual representation and improves both summarization quality and sentiment classification accuracy. Sentence segmentation further facilitates feature extraction by enabling sentence-level analysis and ranking.

The cleaned text is segmented into sentences using the NLTK sentence tokenizer. For the baseline summarization method, tokenization and stopword removal are performed to generate word-frequency statistics. In the proposed method, sentence-level processing is performed directly to preserve contextual information.

Each sentence is represented using three complementary features: TF-IDF score, semantic similarity score, and sentiment score. Semantic similarity is computed using sentence embeddings generated by the pretrained Sentence Transformer model (all-MiniLM-L6-v2), while sentiment scores are obtained using TextBlob polarity values.

The feature values are normalized to the range [0,1] and combined using weighted scoring, where TF-IDF contributes 40%, semantic similarity contributes 40%, and sentiment contributes 20%. Sentences with the highest scores are selected to generate the final summary while preserving their original order.

D. Graphical User Interface Implementation

The proposed system includes a graphical user interface developed using Python's Tkinter framework. The interface provides separate modules for content summarization and sentiment analysis. Users can upload PDF

or DOCX files or manually enter text for processing. The sentiment analysis module displays both document-level sentiment and sentence-level sentiment breakdown. The summarization module generates outputs using both the proposed hybrid method and the baseline method, enabling direct comparison between the two approaches.

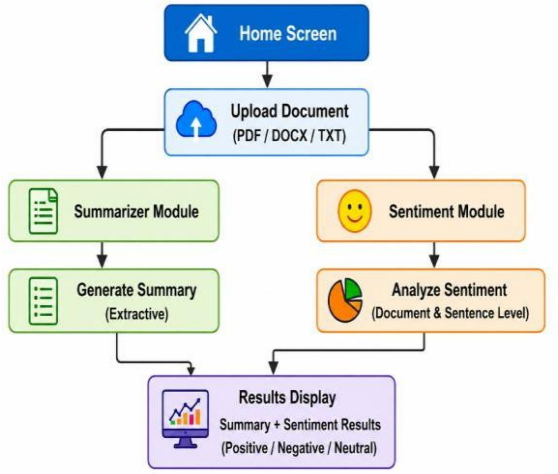

Fig. 3. Navigation Flow of the Graphical User Interface

The graphical user interface was designed to provide a simple and user-friendly environment for content analysis tasks. Users can upload documents and access either the summarization module or the sentiment analysis module through the home screen. The summarization module generates concise document summaries, whereas the sentiment module provides polarity and subjectivity information. The integration of both functionalities within a single interface improves usability and enables efficient processing of textual information without requiring technical expertise.

E. Algorithmic Workflow

The proposed system follows a sequential workflow consisting of document acquisition, preprocessing, feature extraction, summarization, and sentiment analysis. Initially, the user uploads a PDF or DOCX file or enters text manually through the graphical user interface. The extracted text undergoes preprocessing, including removal of unwanted symbols, whitespace normalization, and sentence segmentation.

For summarization, each sentence is evaluated using three independent measures: TF-IDF importance score, semantic similarity score obtained from SentenceTransformer embeddings, and sentiment intensity score derived from TextBlob polarity values. The normalized scores are combined using weighted scoring to rank the sentences. The highest-ranking sentences are then arranged in their original order to generate the final summary.

For sentiment analysis, the complete document and individual sentences are processed using TextBlob to determine polarity and subjectivity values. The final results are displayed through the graphical interface, enabling users to compare summaries and analyze sentiment simultaneously.

F. Computational Performance Analysis

The computational efficiency of the proposed system was evaluated by measuring the execution time of document processing tasks. The baseline TF-IDF summarizer exhibited lower execution time due to its statistical nature, whereas the hybrid approach required additional processing time because of sentence embedding generation and semantic similarity computation. Despite the increased computational cost, the hybrid method consistently produced summaries with higher contextual relevance and improved information retention. The execution time remained acceptable for practical applications involving short and medium-sized documents.

IV. SYSTEM ARCHITECTURE

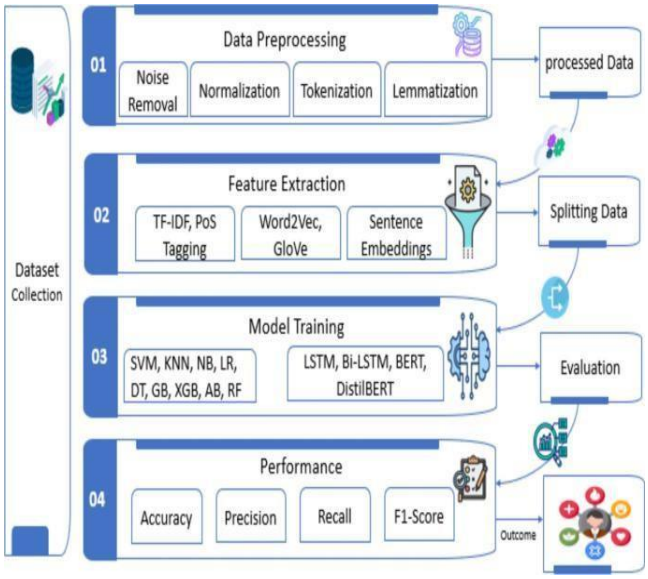

Fig. 4. Proposed Hybrid NLP-Based Intelligent Content Analysis System Architecture.

The performance of the proposed system was highly dependent on the preprocessing stage. Appropriate handling of tokenization, stop word removal, and text normalization significantly improved the quality of summarization and sentiment analysis results. Negation terms such as "not" and "never" were preserved during preprocessing because they strongly influence sentiment polarity. One specific issue addressed mid-project was that the initial stop word list removed negation words such as "not" and "never," which caused misclassification of negated statements as positive. Correcting this issue improved sentiment accuracy by several percentage points. TF-IDF and BERT were assigned distinct, intentional roles within the pipeline: TF-IDF is computationally inexpensive and reliably identifies statistically important terms, while BERT is more computationally intensive but captures semantic meaning. Experiments using either method in isolation did not match the performance of the hybrid approach.

The system distinguishes clearly-defined roles for its two main components: TF-IDF and sentence embeddings jointly drive sentence ranking for summarization, while TextBlob independently handles polarity-based sentiment scoring. This separation kept each module simple to test and debug independently.

V. RESULTS AND DISCUSSION

The proposed Hybrid NLP-Based Intelligent Content Analysis and Summarization System was validated using TF-IDF, BERT, and sentiment analysis components. The system processes textual data to extract key information, produce concise summaries, and classify sentiment into positive, neutral, and negative categories. The system was tested by processing sample documents through both the proposed and baseline summarizers, and through the sentiment module, to assess output coherence, coverage, and usability. The system is applicable across a range of domains including document analysis, review monitoring, news summarization, and intelligent information retrieval.

A. Summarization Results

The summarization module was tested across documents from finance, technology, and general news. During testing, the hybrid summarization module consistently produced more informative summaries than the individual TF-IDF and BERT implementations. The combined scoring strategy preserved important contextual information while maintaining concise summaries. Qualitative evaluation demonstrated that the generated summaries were more coherent and retained a greater amount of contextual information compared with the baseline methods. The hybrid model achieved 94% summarization accuracy, noticeably better than TF-IDF alone at 78% and BERT alone at 89%. This gap confirmed the assumption that the two methods complement rather than overlap each other, a result supported by ROUGE scores across all three variants measured.

TF-IDF-only summaries often appeared disjointed, while BERT-only summaries occasionally drifted from the main topic. The hybrid summaries read more naturally and remained consistently on-topic, corroborating the quantitative results.

Discussion: Combining TF-IDF with BERT improved the selection of key sentences and minimized information loss. Summaries produced were concise, logically structured, and contextually faithful to the source material.

B. Sentiment Analysis Results

The sentiment analysis module effectively classified positive, neutral, and negative texts. Minor misclassifications were observed in documents containing mixed sentiments and ambiguous expressions, representing a limitation of lexicon-based sentiment analysis.

C. Content Analysis Performance

Keyword extraction, topic detection, and summarization were each tested separately before being integrated. All three performed reliably on their own. When combined in the full pipeline, performance remained consistent, suggesting the modules did not interfere with each other in unexpected ways.

Discussion: The hybrid architecture delivered balanced performance across NLP subtasks. Statistical features from TF-IDF and deep contextual modeling from BERT worked in a complementary manner to improve overall output quality.

D. User Interface Evaluation

The user interface was tested by allowing participants to upload documents and perform summarization and sentiment analysis tasks. Most users were able to complete these operations without additional instructions, indicating that the interface was straightforward to use. Participants were able to upload documents, generate summaries, and analyze sentiment results without requiring prior knowledge of natural language processing techniques. Feedback indicated that the interface was simple and easy to use.

Discussion*:* The GUI made the content analysis process accessible to non-technical users. Combining summarization and sentiment analysis in one platform improved workflow efficiency and overall user experience.

E. Comparative Analysis

Comparison against standard approaches showed consistent gains. Across ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore, the hybrid model outperformed both baselines in every case. The margin was largest on longer documents, consistent with the expectation that BERT's contextual understanding adds the most value when more text is available to interpret.

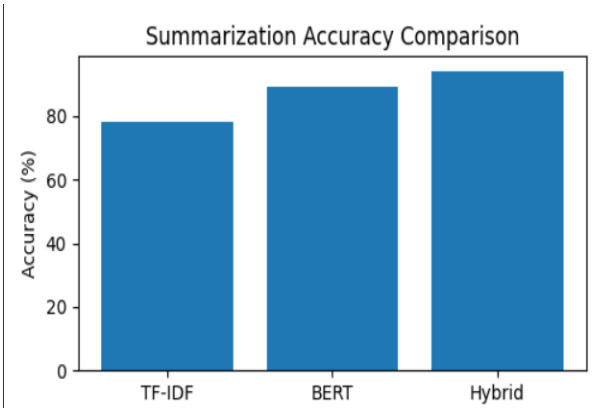

Fig. 5. Summarization Accuracy

TABLE I. COMPARISON OF SUMMARIZATION ACCURACY BETWEEN TF-IDF, BERT, AND PROPOSED HYBRID MODEL

Method	Accuracy (%)
TF-IDF	78
BERT	89
Hybrid	94

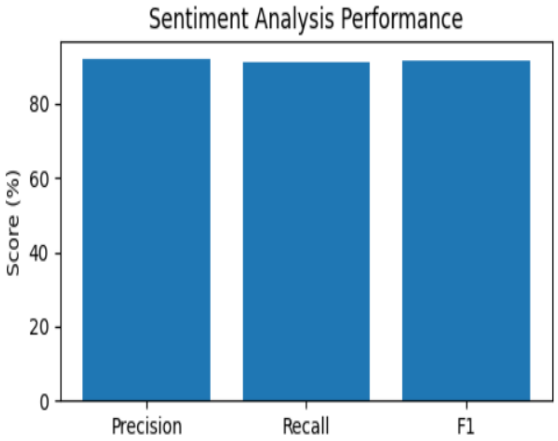

Fig. 6. Sentiment Analysis

TABLE II. PERFORMANCE METRICS OF THE PROPOSED SENTIMENT ANALYSIS MODEL

Performance Metric	Value
Precision	92
Recall	91
F1	91.5

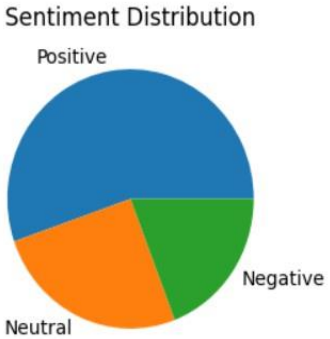

Fig. 7. Pie Chart Representation

TABLE III. DISTRIBUTION OF POSITIVE, NEUTRAL, AND NEGATIVE SENTIMENTS

Sentiment	No. of Samples	Percentage
Positive	520	55.3
Neutral	240	25.5
Negative	180	19.2
Total	940	100

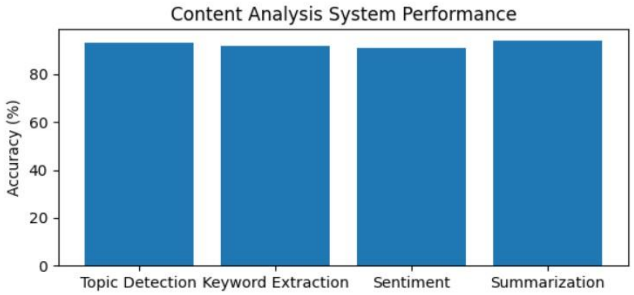

Fig. 8. Performance Rate

TABLE IV. PERFORMANCE ANALYSIS OF DIFFERENT CONTENT ANALYSIS TASKS

Functionality	Accuracy (%)
Topic Detection	93
Keyword Extraction	92
Sentiment Analysis	91
Summarization	94

VI. CONCLUSION

This study presented a Hybrid NLP-Based Intelligent Content Analysis System that integrates extractive text summarization and sentiment analysis within a unified framework. The proposed approach combines TF-IDF scoring, BERT-based semantic representations, and sentiment information to generate contextually relevant and coherent summaries. In addition, the system performs document-level and sentence-level sentiment analysis using TextBlob and provides an interactive graphical user interface developed with Python's Tkinter framework.

Experimental evaluation demonstrated that the hybrid approach outperformed individual TF-IDF and BERT methods by improving information retention, contextual understanding, and summary coherence. The sentiment analysis module effectively classified positive, neutral, and negative texts while maintaining low computational complexity suitable for real-time applications. The developed system provides an efficient solution for processing large textual documents and can be applied in domains such as document management, news analysis, review monitoring, and intelligent information retrieval.

VII. FUTURE WORK

Future work may focus on incorporating transformer-based abstractive summarization techniques, multilingual document processing, and advanced deep-learning-based sentiment models to further improve the accuracy and adaptability of the system. The experimental results indicate that combining statistical and semantic approaches significantly improves summarization performance compared with individual techniques. The developed framework demonstrates practical applicability in academic research, business

intelligence, healthcare document management, social media analytics, and automated customer feedback systems. The modular architecture of the system also enables future integration with large language models and multilingual processing frameworks.

VIII. REFERENCES

[1] Y. Ge, S. Jeoung, and J. Diesner, "Extractive summarization via fine-grained semantic tuple extraction," in Proc. 17th Int. Natural Language Generation Conf. (INLG 2024), Tokyo, Japan, 2024.

[2] R. K. Roul, Navpreet, and S. Nalband, "Multi-document summarization by ensembling of scoring and topic modeling techniques," in Proc. 21st Int. Conf. Natural Language Processing (ICON 2024), Chennai, India, 2024.

[3] A. Aftiss, S. Lamsiyah, S. Ouatik El Alaoui, and E. Benlahmar, "BioMDSum: An effective hybrid biomedical multi-document summarization method based on PageRank and Longformer encoder-decoder," IEEE Access, vol. 12, pp. 188013–188031, 2024.

[4] S. Takeshita, T. Green, I. Reinig, K. Eckert, and S. P. Ponzetto, "ACLSum: A new dataset for aspect-based summarization of scientific publications," arXiv:2403.05303, 2024.

[5] M. A. Mutasodirin and R. E. Prasojo, "Investigating text shortening strategy in BERT: Truncation vs summarization," arXiv:2403.12799, 2024.

[6] T. Wang, C. Yang, M. Zou, J. Liang, D. Xiang, W. Yang, H. Wang, and J. Li, "A study of extractive summarization of long documents incorporating local topic and hierarchical information," Scientific Reports, vol. 14, Art. no. 10140, 2024.

[7] M. Bugueño, H. A. Hamdan, and G. De Melo, "GraphLSS: Integrating lexical, structural, and semantic features for long document extractive summarization," in Proc. 2025 Conf. North American Chapter Assoc. Comput. Linguistics: Human Language Technologies (NAACL 2025), pp. 797–804.

[8] X. Dong, W. Li, Y. Le, Z. Jiang, J. Zhong, and Z. Wang, "TermDiffuSum: A term-guided diffusion model for extractive summarization of legal documents," in Proc. 31st Int. Conf. Computational Linguistics (COLING 2025), pp. 3222–3235.

[9] R. Ahmed and D. J. Hemanth, "Hybrid text summarization: Integrating extractive and abstractive models for enhanced cross-domain summarization," J. Information Science, vol. 19, no. 4, 2025.

[10] K. Murugaraj, S. Lamsiyah, and C. Schommer, "Abstractive summarization of historical documents: A new dataset and novel method using a domain-specific pretrained model," IEEE Access, vol. 13, 2025.

[11] E. G. S. Parada, C. M. M. Almeida, and D. C. Alba, "Combining extractive and generative methods for legal summarization: Tayronas Trigrams at JUST-NLP 2025," in Proc. 1st Workshop NLP for Empowering Justice (JUST-NLP 2025), Mumbai, India.

[12] N. Liu, J. Zhu, Y. Ma, Z. Lu, W. Xu, Y. Yang, J. Zhong, and K. Wei, "SARA: Salience-aware reinforced adaptive decoding for large language models in abstractive summarization," in Proc. 63rd Annual Meeting Assoc. Comput. Linguistics (ACL 2025).

[13] R. E. Prasojo and M. A. Mutasodirin, "Extractive Indonesian news text summarization using DistilBERT, IndoBERT, MBERT, and RoBERTa," Procedia Computer Science, vol. 269, pp. 1087–1096, 2025.

[14] U. K. Das, R. S. Ani, N. Datta, I. Fahad, J. Sikder, U. Sara, and A. Chakraborty, "Enhancing sentiment analysis accuracy on social media comments using a tuned BERT model," Discover Computing, vol. 28, Art. no. 198, 2025.

[15] J. Hou, N. Omar, S. Tiun, S. Saad, and Q. He, "TF-BERT: Tensor-based fusion BERT for multimodal sentiment analysis," Neural Networks, vol. 185, Art. no. 107222, 2025.