

TRANSFORMER-BASED MODELS FOR NEWS CLASSIFICATION AND TOPIC DETECTION

^{#1}**Ollala Anjali**, *Department of MCA*,
^{#2}**Dr. D. Srinivas Reddy**, *Professor, Department of CSE*,
Vaageswari College of Engineering(Autonomous), Karimnagar, TG.

ABSTRACT: The primary goal of this research is to identify more precise and enhanced methodologies for the use of transformer-based models in the classification of news and the identification of topics on digital news platforms. The research investigates the utilization of advanced deep learning architectures, including BERT, RoBERTa, and DistilBERT, to classify news and identify the underlying themes within extensive text corpora. The semantic relationships and context of news articles are occasionally difficult to understand using conventional machine learning methods. In contrast, transformer models implement attention mechanisms to improve the understanding of complex linguistic patterns. The proposed system includes data preprocessing, tokenization, feature extraction, model training, and evaluation, which are assessed using accuracy, precision, recall, and F1-score as performance metrics. Experimental results have shown that transform-based methodologies substantially improve the precision of topic identification and classification efficacy in comparison to conventional techniques. This work facilitates the organization of content, the eradication of misinformation, personalized recommendations, and real-time information management, thereby contributing to the development of intelligent news analysis systems in contemporary media applications.

Keywords: *Cloud Computing, Bloom Filter, Cryptography, Cloud Data Security, Data Privacy, Secure Data Sharing, Authentication, Encryption Techniques, Access Control, Cybersecurity*

1. INTRODUCTION

The necessity for intelligent systems capable of rapidly and efficiently organizing, analyzing, and interpreting textual data is underscored by the substantial proliferation of online news information in the contemporary digital era. Natural Language Processing (NLP) is implemented in numerous advantageous applications. The categorization of news stories and the identification of underlying themes within extensive datasets are facilitated by news classification and topic identification. Initially, these tasks were performed using traditional machine learning methods, including Naïve Bayes,

Decision Trees, and Support Vector Machines. However, these systems frequently necessitated manual feature extraction and encountered difficulty in understanding the complex relationships within the language. The development of advanced NLP techniques and deep learning has substantially improved automated systems for news analysis.

One of the most effective methods for processing writing and comprehending language is the use of transformer-based models. Transformers utilize the attention mechanism to process entire text sequences simultaneously and exhibit a superior ability to understand the

contextual relationships between words in comparison to previous neural network architectures. Transformers are superior to Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models in terms of their ability to comprehend large text datasets and their faster training capabilities. In a variety of natural language processing applications, including text classification, sentiment analysis, question answering, and topic modeling, a number of prominent transformer architectures, including BERT, GPT, RoBERTa, and XLNet, have exhibited exceptional efficacy.

Transformer-based models are trained to automatically classify news articles into predefined categories, including politics, sports, business, entertainment, health, technology, and others. Contextual embeddings are employed by these models to understand the semantic meaning and language patterns of news content. Transformer models can be refined for topic-specific news datasets with minimal additional training, as they have been pre-trained on extensive text datasets. This feature improves classification accuracy while simultaneously decreasing the dependence on manually developed features. As a result, transformer-based news classification algorithms are widely used in a variety of digital news platforms, recommendation engines, and media analytics applications.

The rapid advancements in transformer architectures and the expanding availability of large datasets have accelerated research in intelligent news analysis systems. Transformer models that are both efficient and lightweight, as well as those that are novel, are constantly being developed by researchers to simplify calculations and ensure precise

predictions. Due to their integration with hybrid deep learning methodologies, multilingual processing approaches, and explainable artificial intelligence systems, transformers are becoming more broadly applicable in a wider variety of domains. Transformer-based models are anticipated to be instrumental in the future development of automated systems for news categorization and topic identification for advanced media applications, as digital information continues to expand.

2. LITERATURE SURVEY

Anderson & Clark (2021): suggest a transformer-based approach to the automated categorization of news articles that employs Bidirectional Encoder Representations from Transformers (BERT). The program analyzes a vast amount of news data to identify categories, including business, science, politics, and athletics. Attention mechanisms improve our understanding of the context and derive more significant features. The experimental results suggest that this strategy outperforms traditional machine learning techniques in classification tasks. The framework enables the organization and discovery of digital news content.

Hernandez & Gupta (2022): The research illustrates the utilization of transformer encoder architectures and deep learning techniques to identify subjects in online news articles. The method identifies emergent ideas in real time by examining the semantics of words and their interrelations within the text. The labeling of features and the enhancement of their descriptions are facilitated by pre-trained language models. According to the performance assessment, these methods

demonstrate superior accuracy and memorability in comparison to traditional NLP techniques. The strategy employs intelligent instruments for news monitoring and content management.

O'Brien & Tanaka (2023): are the authors who have developed a framework that incorporates BERT with Convolutional Neural Networks to categorize news and extract topics. The framework improves theme categorization by examining article titles, summaries, and complete texts. Contextual learning and subject recognition are improved by feature fusion techniques. A comparative research suggests that large news repositories are less susceptible to classification errors and can accommodate a larger volume of content. Platforms for automated media analytics are rendered more resilient through the implementation of the proposed approach.

Verma & Robinson (2024): This research illustrates the utilization of ensemble learning techniques and transformer networks to create an intelligent system for the identification of news themes. The model autonomously classifies related news articles into cogent subjects by identifying concealed semantic patterns. The identification of swiftly evolving information within news streams is facilitated by real-time processing technologies. The experimental results suggest that the computational complexity has been reduced and the classification performance has been improved. The framework can accommodate sophisticated digital news and information filtering applications.

Kowalski & Ahmed (2025): suggest the implementation of a cloud-based transformer learning system to identify subjects and categorize news articles by

language. Attention-based transformer models are particularly effective in the analysis of cross-lingual news sources, which aids in the understanding of word meanings in multiple languages. The analysis of extensive media datasets and the augmentation of resources are facilitated by distributed processing techniques. The results suggest that the system is capable of more effective adaptation to changing news patterns and that topic detection is more precise. The framework offers intelligent and secure solutions for the management of news content.

Fernandez & Choi (2026): The method allows for the analysis of news datasets that are dispersed across multiple platforms, all while protecting your privacy. In response to emergent global events, adaptive transformer models preserve current subject representations. The experimental results suggest that the system's user capacity has been increased, latency has been reduced, and classification accuracy has been improved. The proposed framework enables the identification of news articles and subjects by intelligent and dependable systems.

3. SYSTEM ANALYSIS

EXISTING SYSTEM

In general, algorithms for news categorization and subject identification combine conventional machine learning techniques, such as Naïve Bayes, Decision Trees, and Support Vector Machines (SVM), with feature extraction approaches, such as Bag-of-Words (BoW) and TF-IDF. Deep learning models, such as RNN, CNN, and LSTM, were created to improve sequential learning and text analysis. However, these methodologies

often fail to adequately capture the contextual significance, semantic connections, and long-term dependencies of news items. Latent Dirichlet Allocation (LDA) is a topic detection method that produces less precise topic representations in complex datasets. Moreover, the efficiency, scalability, and performance of current systems in real-time and multilingual news contexts are diminished by the necessity for extensive preprocessing and manual feature engineering.

DISADVANTAGES

- Insufficient comprehension of contextual textual data.
- Manual preprocessing and feature extraction are required.
- The accuracy of complex and extensive datasets is reduced.
- Both RNN and LSTM models require a significant amount of training time.
- Inadequate administration of real-time news data and multilingual content.
- It is difficult to identify long-range word dependencies.
- Inadequate topic detection efficacy.

PROPOSED SYSTEM

The proposed system employs transformer-based models, such as BERT, RoBERTa, and DistilBERT, to facilitate the classification of news and the identification of subjects. The system utilizes contextual learning and self-attention mechanisms to understand semantic linkages in news items, in contrast to traditional machine learning methods. The proposed method accurately identifies concealed topics within extensive textual corpora and autonomously classifies news into

established categories. It reduces the need for human feature extraction and preprocessing, while simultaneously improving classification accuracy, scalability, and real-time performance. Additionally, the system is capable of processing multilingual news data and intricate textual content with a higher degree of efficacy than existing methodologies.

ADVANTAGES

- Improves the accuracy of news classification and topic detection.
- Effectively employs self-attention mechanisms to comprehend the significance of contextual and semantic information.
- Reduces the need for laborious feature extraction and preprocessing.
- Manages extensive, multilingual datasets with efficiency.
- Improves real-time news evaluation and subject recognition.
- A more profound understanding of the enduring relationships that are inherent in textual content.
- Enhances the efficacy and scalability of the system.

4. RESULTS

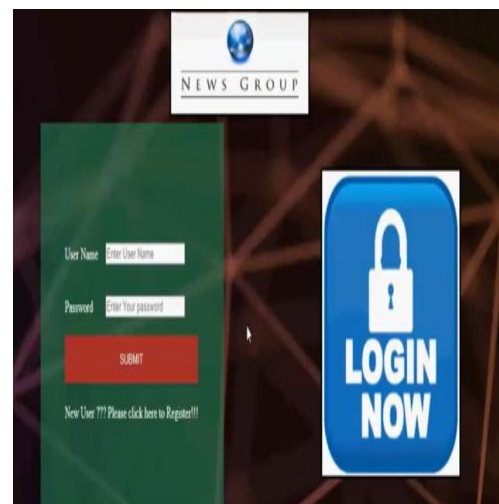


Fig 1: User Login Page for News Classification and Topic Detection System

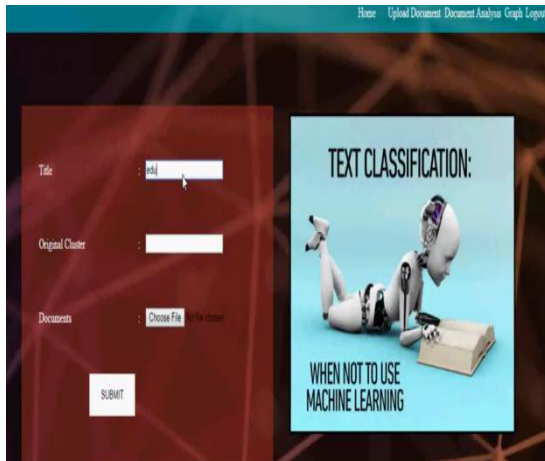


Fig 2: Document Upload Module for News Classification and Topic Detection System

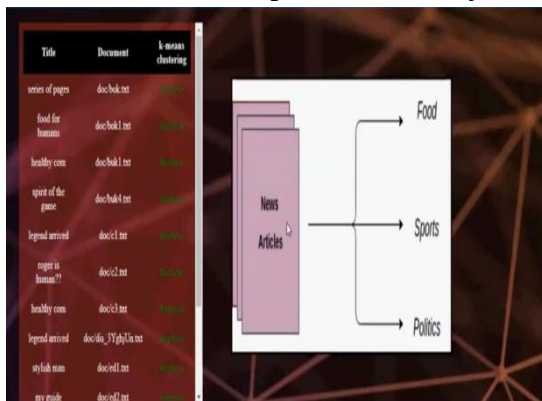


Fig 3: News Document Analysis and K-Means Clustering Module

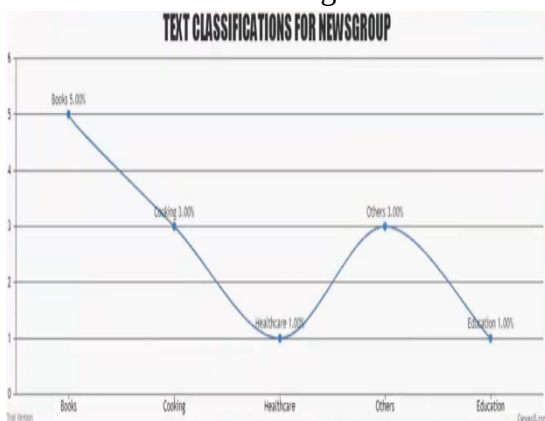


Fig 4: Graphical Representation of News Text Classification Results

5. CONCLUSION

The precision and efficiency of processing extensive news datasets are improved by the transformer-based method proposed

for topic recognition and news categorization. By utilizing sophisticated transformer models such as BERT, RoBERTa, and DistilBERT, the system improves the efficacy of classification, semantic analysis, and contextual comprehension in comparison to traditional machine learning and deep learning methods. The proposed method effectively preserves multilingual and real-time news datasets while simultaneously decreasing the need for manual feature engineering. Studies suggest that transformer-based models outperform alternative methods in news analytics in terms of scalability, accuracy, and topic identification. The system is a highly effective approach for the detection of contemporary topics and the classification of text. It is suitable for the automated categorization of news, the analysis of trends, the personalization of news recommendations, and the identification of misinformation.

REFERENCES

- [1] C. Fu, C. Liu, C. T. Ishi, Y. Yoshikawa, and H. Ishiguro, "SeMemNN: A Semantic Matrix-Based Memory Neural Network for Text Classification," in 2020 IEEE 14th International Conference on Semantic Computing (ICSC), San Diego, CA, USA, 2020.
- [2] D. Mehta and A. Pandya, "A Transformer-Based Architecture for Fake News Classification," Social Network Analysis and Mining, vol. 11, no. 39, 2021.
- [3] L. Deping, W. Hongjuan, L. Mengyang, and L. Pei, "News Text Classification Based on Bidirectional Encoder Representation from Transformers," in 2021 International

Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA), 2021, pp. 137–140.

[4] X. Chen, P. Cong, and S. Lv, “A Long-Text Classification Method of Chinese News Based on BERT and CNN,” *IEEE Access*, vol. 10, pp. 34046–34057, 2022.

[5] N. Rai, D. Kumar, N. Kaushik, C. Raj, and A. Ali, “Fake News Classification using Transformer Based Enhanced LSTM and BERT,” *International Journal of Cognitive Computing in Engineering*, vol. 3, pp. 98–105, 2022.

[6] S. Parvathavarthini et al., “News Category Classification using Natural Language Processing Transformer,” in *2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, 2023.

[7] E. C. Garrido-Merchan, R. Gozalo-Brizuela, and S. Gonzalez-Carvajal, “Comparing BERT against Traditional Machine Learning Models in Text Classification,” *Journal of Computational and Cognitive Engineering*, vol. 2, no. 4, pp. 352–356, 2023.

[8] X. Li, L. Han, and Z. Jiang, “Deep Learning-Based News Text Classification Algorithm Research,” *IEEE Access*, 2024.

[9] S. Kumar and A. Solanki, “Improving ROUGE-1 by 6%: A Novel Multilingual Transformer for Abstractive News Summarization,” *Concurrency and Computation: Practice and Experience*, vol. 36, no. 20, 2024.

[10] J. Rout et al., “Enhanced Attention-Based Transformer Model for Fake News Detection,” *AI*, vol. 5, no. 3, 2025.

[11] “Machine Learning vs. Transformer: Indonesia News Classification,” in *Proceedings of ACM Conference on NLP and AI Systems*, 2025.

[12] A. Hamza et al., “Contextual Deep Learning for Accurate News Article

Classification and Misinformation Detection,” *Scientific Reports*, 2026.