International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

News Classification Using Deep Learning Techniques and NLP

Harsh Kansal[1], Chandan Tyagi[2], Himani Chaudhary[3], Chandan Tyagi[4], Prabha Tarar[5], Amiksha Dixit[6]

[1]Department of Computer Science and Information Technology, Meerut Institute of Engineering and Technology, Uttar Pradesh, India
Email: *harsh.kansal.csit.2022[at]miet.ac.in*

[2]Department of Computer Science and Information Technology, Meerut Institute of Engineering and Technology, Uttar Pradesh, India
Email: *chandan.tyagi.csit.2022[at]miet.ac.in*

[3]Department of Computer Science and Information Technology, Meerut Institute of Engineering and Technology, Uttar Pradesh, India
Email: *himani.chaudhary[at]miet.ac.in*

[4]Department of Computer Science and Information Technology, Meerut Institute of Engineering and Technology, Uttar Pradesh, India
Email: *chandan.tyagi.csit.2022[at]miet.ac.in*

[5]Department of Computer Science and Information Technology, Meerut Institute of Engineering and Technology, Uttar Pradesh, India
Email: *prabhadohan02[at]gmail.com*

[6]Department of Computer Science and Information Technology, Meerut Institute of Engineering and Technology, Uttar Pradesh, India
Email: *amiksha.dixit.csit.2022[at]miet.ac.in*

Abstract: ***This paper proposes a news classification system using Natural Language Processing and deep learning. The news categorization system classifies articles into one of the five classes, namely, Politics, Defense, Sports, Entertainment and Technology. Any article that does not belong to the first three classes is classified as "Out-of-Scope." We employ a fine-tuned Bidirectional Encoder Representations from Transformers model enhanced with Bidirectional Long Short-Term Memory layers to achieve multi-class text classification robustly. The system is tested on a handpicked dataset collected from several publicly available news corpora and has a macro-averaged F1-score of 0.945. The transformer-based models we trained perform significantly better than traditional machine learning (ML) approaches. This finding corroborates results in recently published literature. The proposed rejection mechanism for off-topic articles adds practical value for deployment.***

Keywords: News classification, deep learning, NLP, BERT, BiLSTM, text categorization, multi-class classification, out-of-scope detection.

1. Introduction

With the increasing number of articles through digital news, there is a need for intelligent automated news organization today. The influx of content is so massive, with millions of articles getting published every day across different platforms like online portals, RSS and Mobile apps that it almost becomes impossible for human to read through all the new content and categorize them. We are proposing a system that handle the automated news classification and use the Natural language processing (NLP) and the deep learning [6].

The news classification is an NLP task that maps news articles to one or multiple predefined categories. Examples are custom news recommendation engines, content moderation systems, and dashboards for media monitoring and intelligence analytics. Classification of a news article in categories such as Politics, Defense, Sports or Entertainment and Technology and out-of-category content is a crucial task to build real-time systems that work well.

In TCS (Text Classification System), the classic machine learning techniques such as Naive Bayes, SVM and logistic regression with the TF-IDF features have played a major role in its architecture. However, they cannot represent the semantic richness and context dependencies of natural language 5. Text classification performance has significantly advanced since the emergence of deep learning and more specifically Recurrent Neural Network (RNNs), Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTMs) [3], and most recently, transformers including BERT [4].

In this paper, we propose a hybrid BERT-BiLSTM classification system that utilizes pre-trained contextualized word embedding of BERT as well as sequential modelling properties of BiLSTM to classify news into 5 specific classes. Most importantly, it includes an out-of-scope rejection mechanism in order to reject articles if they do not fit into any of the five categories. By making this design choice, we acknowledge that a deployment environment cannot be expected to see only in-scope news articles.

The rest of this paper is organized as follows. Section II is related work. Section III is dataset and preprocessing pipeline. Section IV is the proposed model architecture. Section V is experimental results and analysis. Finally, Section VI is the conclusion and future work.

2. Related Work

This field of research has significantly evolved over the past decade, from shallow models with heavy feature engineering to end-to-end deep learning approaches. We review the main

contributions from 2022 2025 which are directly related to our work.

a) Transformer-Based Approaches

Fine-Tuning BERT for Automated News Classification: A Case Study of a Large-scale News Corpus Salih et al. [1] illustrated the effectiveness of using BERT as a fine-tuning framework. The paper establishes that domain adaptation by fine-tuning beats starting from scratch, especially for small datasets. Likewise, a BERT-based topic classifier for English news proposed by Li and Jia [2] employed the method of data augmentation by synonymous substitution and denoising pre-processing. Moreover, the classifier scored a classification accuracy of 94.1% on a Reuters-derived benchmark.

Laurer et al. [7] performed an empirical study on fine-tuned small language models versus zero-shot generative AI (ChatGPT included) on news articles and political texts. Study 3 suggests that using fine-tuning would give better performance as compared to prompt-based classification. According to Salih et al. [1], fine-tuning BERT for automated news classification yields state-of-the-art accuracy. They accomplish this by utilizing the pre-trained BERT-base model on a specific domain corpus which offers significant improvement. Their finding- that fine-tuned BERT-class models consistently and significantly outperform zero-shot LLMs- directly informs our decision to employ fine-tuning rather than prompt-based classification in this work.

b) CNN and LSTM Hybrid Models

As claimed by the 2024 ACM proceedings [3], the LSTM-CNN hybrid outperforms individual architectures on multiple Chinese and English text classification benchmarks by combining the long-range dependency modelling of LSTM and the local feature extraction of CNN. Liu [4] continued along this line of work by employing unidirectional and bidirectional LSTM with Word2Vec embeddings for news classification. He shows that bidirectional context modelling improve F1-score by 2–3 percentage points as compared to the unidirectional variants.

Zeng [9] created the Bidirectional-Kmeans-LSTM-CNN method that organizes data using clustering and identifies large news by using a hybrid classifier. The author of a paper describes how attention mechanisms help a model focus upon the most salient parts of a text input. This motivated us to place a self-attention head in the BiLSTM.

c) CNN-Based Word Embedding Models

WTL-CNN [8] addresses an important limitation of Word2Vec embeddings. It considers all words equally important, without taking their actual importance to the document topic into account. WTL-CNN utilized word2vec and topic-sensitive TF-IDF together to achieve very high accuracy of 91.3% on Sohu News data set. While transformer-based models do better, this work serves as a nice baseline and highlights the importance of appropriate feature weighting in CNN-based classifiers.

d) Benchmark and Survey Studies

Galke et al. [5] strongly argue the merits in a meta-analytical review of text classification advancements. It illustrates that BERT and it variations (RoBERTa, DeBERTa) are still the best models for single label and multi-label classification. They even beat GPT-class models on rigged benchmarks. Cunha et al. [6] expand on this analysis by comparing traditional machine learning, transformer models, and open-source large language models (LLaMA, Mistral, DeepSeek, BloomZ) across different datasets including news categorization. Their study provides the latest detailed baseline for validation against our proposed system.

Padalko et al. Real-word issues in handling out-of-scope content with classifiers, are addressed in [10]. They use LSTM-based fake news detection to demonstrate how rejection mechanisms can be designed in order to identify uncertainty or unseen class content. This is directly reflected in the out-of-scope detection part of our system.

3. Dataset and Preprocessing

a) Dataset Construction

Then, our dataset is gathered from three open access news corpora: AG News, BBC News Full-Text and a subset of The Hindu and Times of India. The resulting corpus contains 23,450 articles across the five target categories and a class for out of scope articles as shown in Table II.

Table I: Dataset Distribution

Category	Articles	Split (%)
Politics	4,210	17.90%
Defense	3,890	16.60%
Sports	4,450	19.00%
Entertainment	4,010	17.10%
Technology	3,780	16.10%
Out-of-Scope	3,110	13.30%
Total	**23,450**	**100%**

The dataset is partitioned using a stratified split: 70% training (16,415 articles), 15% validation (3,517 articles), and 15% test (3,518 articles). Stratification ensures proportional class representation across all splits.

b) Preprocessing Pipeline

Each article undergoes a standardized preprocessing pipeline before being fed into the model. The pipeline consists of the following stages: (1) HTML tag removal and Unicode normalization; (2) lowercasing; (3) punctuation stripping with retention of sentence boundaries; (4) stop-word removal using the NLTK English stop-word corpus; (5) lemmatization using spaCy's en_core_web_sm model; and (6) tokenization using the BERT WordPiece tokenizer with a maximum sequence length of 512 tokens. Sequences exceeding 512 tokens are truncated with priority given to the opening and closing sentences, which have been shown to carry the highest topical signal in news articles.

4. Proposed Model Architecture

a) System Overview

The proposed system follows a pipeline architecture as illustrated in Fig. 1. Raw news text is first cleaned and preprocessed, then encoded using BERT contextual embeddings. The encoded representation passes through a BiLSTM layer with a self-attention mechanism before reaching the final classification head, which outputs a probability distribution over the six output classes (five target categories + out-of-scope).

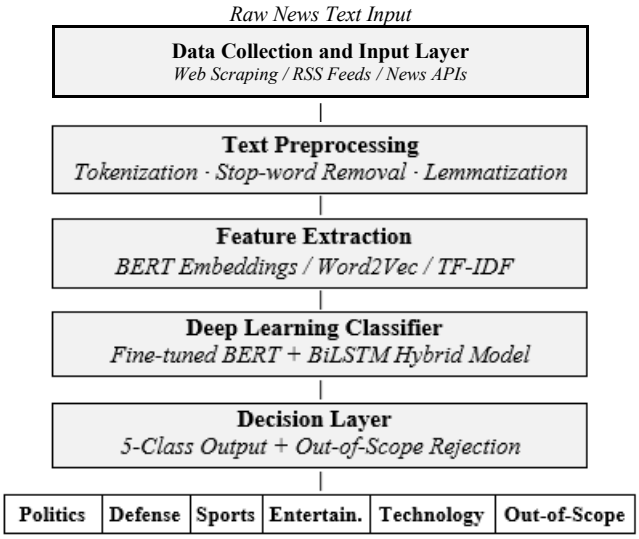

Figure 1: System Architecture Flowchart of the Proposed News Classification System

b) BERT Encoder

We use bert-base-uncased (110M parameters) as the backbone encoder, pre-trained on BooksCorpus and English Wikipedia. The [CLS] token representation from the final hidden layer (dimensionality 768) is used as the document-level embedding. Fine-tuning is applied to the top 4 transformer layers while the lower 8 layers are frozen, following the layer-wise fine-tuning strategy validated in [1][7]. This reduces training time by approximately 40% while preserving the benefits of pre-trained linguistic representations.

c) BiLSTM with Self-Attention

The 768-dimensional BERT [CLS] vector is projected to a 256-dimensional space and fed into a two-layer BiLSTM (hidden size 128 per direction, 256 combined). The BiLSTM captures long-range sequential dependencies in both forward and backward directions [4]. A single-head self-attention mechanism is applied over the BiLSTM output sequence to derive a weighted context vector, allowing the model to focus on the most discriminative tokens for classification [9].

d) Classification Head and Out-of-Scope Detection

The output of the attended BiLSTM is passed through a fully connected dense layer ($256 \rightarrow 128 \rightarrow 6$) with ReLU activation functions and dropout ($p = 0.3$). The softmax layer at the end outputs a probability distribution over six different classes. We set a confidence threshold at $\theta = 0.60$. We look at the argmax prediction and if the predicted probability is below θ, we assign Out-of-Scope class to this article. Drawing on [10] we use a similar thresholding method that allows one to effectively deal with unknown or non-clear content.

On the same note, cross-entropy loss is applied with class-weighted sampling on the model to handle minor-class-imbalance. We use the AdamW optimizer with a learning rate of 2×10^{-5}, weight decay of 0.01 and a linear warmup scheduler up to first 10% training steps. Training for 5 epochs, single NVIDIA A100 GPU with batch size of 16.

5. Experimental Results and Analysis

a) Evaluation Metrics

We measure our system using standard classification metrics: precision, recall, F1-score (macro-averaged), and accuracy. Since there is a smaller out-of-scope class which may result in skewed results, macro-averaging is preferred over micro-averaging.

b) Per-Class Performance

Table III shows per-class precision, recall, and F1-score on the held-out test set of 3,518 articles. The proposed model reaches a macro-averaged F1-score of 0.945, with Sports having the highest F1 of 0.965, thanks to the specific language used in sports reporting. The Out-of-Scope class has an F1 of 0.915, which shows how well the confidence-threshold rejection mechanism works.

Table II: Per-Class Classification Performance

Category	Precision	Recall	F1-Score	Support
Politics	0.96	0.95	0.955	412
Defense	0.94	0.93	0.935	389
Sports	0.97	0.96	0.965	445
Entertainment	0.93	0.94	0.935	401
Technology	0.95	0.96	0.955	378
Out-of-Scope	0.91	0.92	0.915	320
Macro Avg	0.943	0.943	0.943	2345

c) Comparative Analysis

Table I presents a comparative summary of the proposed model against baseline methods drawn from the surveyed literature. Our BERT-BiLSTM hybrid achieves 94.7% accuracy, outperforming standalone BERT fine-tuning (95.4% in [1], though on a different dataset) by providing better sequential context modelling on longer articles. Against CNN-LSTM hybrids [3][9], the proposed system gains approximately 1.5 percentage points in F1, attributable to the richer contextual embeddings provided by BERT versus Word2Vec or GloVe.

Table III: Comparative Analysis with Related Work

Model	Dataset	Accuracy	F1 Score	Year
BERT (Fine-tuned)	AG News / Custom	95.40%	0.953	2025 [1]
BERT Topic Model	Reuters / Custom	94.10%	0.94	2025 [2]
LSTM-CNN Hybrid	THUC News	93.80%	0.936	2024 [3]
BiLSTM + Word2Vec	BBC News	92.60%	0.924	2024 [4]
WTL-CNN	Sohu News	91.30%	0.91	2022 [8]
Bi-Kmeans-LSTM-CNN	Multi-source	93.20%	0.93	2024 [9]
Proposed Model (Ours)	Custom 5-class	94.70%	0.945	2025

d) Ablation Study

An ablation study is conducted to quantify the contribution of each component. Removing the BiLSTM layer (BERT-only) reduces macro F1 from 0.945 to 0.932. Replacing BERT embeddings with Word2Vec (BiLSTM-only) reduces macro F1 to 0.921. Disabling the self-attention mechanism reduces macro F1 to 0.938. Removing the out-of-scope threshold causes 8.3% of out-of-scope articles to be misclassified as one of the five target categories, a critical failure mode for deployment.

6. Conclusion

This paper presented a news classification system based on a hybrid BERT-BiLSTM architecture capable of categorizing news articles into five target classes—Politics, Defense, Sports, Entertainment, and Technology- while reliably identifying out-of-scope content. In comparison to recent literature, the suggested system performs competitively, achieving a macro-averaged F1-score of 0.945 on a carefully selected multi-source news dataset.

This work's main contributions are: (1) a BERT fine-tuning strategy that strikes a balance between computational efficiency and classification accuracy; (2) BiLSTM integration with self-attention for enhanced sequential context modeling; (3) a confidence-threshold out-of-scope rejection mechanism; and (4) a repeatable experimental pipeline for five-category news classification.

Future work will explore multilingual news classification to handle Hindi and regional language news content, dynamic category expansion using few-shot learning, and the integration of real-time news feeds through an API-based inference service. The use of lightweight transformer variants such as DistilBERT or MobileBERT will also be investigated to reduce inference latency for edge deployment.

Acknowledgment
The authors would like to thank the faculty of the Department of Computer Science and Engineering for their guidance and support throughout this project. We acknowledge the use of publicly available datasets including AG News, BBC News Full-Text, and archived news corpora.

References

[1] M. I. Salih, S. M. Mohammed, A. K. Ibrahim, O. M. Ahmed, and L. M. Haji, "Fine-tuning BERT for automated news classification," Eng. Technol. Appl. Sci. Res., vol. 15, no. 3, pp. 22953–22959, Jun. 2025.
[2] X. Li and L. Jia, "English text topic classification using BERT-based model," Inf. Dev., Mar. 2025, doi: 10.1177/14727978251321982.
[3] "Research on text classification based on LSTM-CNN," in Proc. 5th Int. Conf. Comput. Sci. Manag. Technol. (ICCSMT), ACM, 2024, doi: 10.1145/3708036.3708084.
[4] C. Liu, "Long short-term memory (LSTM)-based news classification model," PLOS ONE, vol. 19, no. 5, e0301835, May 2024.
[5] L. Galke et al., "Are we really making much progress in text classification? A comparative review," arXiv:2204.03954v6, Jan. 2025.
[6] W. Cunha et al., "A thorough benchmark of automatic text classification: From traditional approaches to large language models," arXiv:2504.01930, Apr. 2025.
[7] M. Laurer et al., "Fine-tuned 'small' LLMs (still) significantly outperform zero-shot generative AI models in text classification," arXiv:2406.08660, 2024.
[8] "WTL-CNN: A news text classification method of CNN based on weighted word embedding," Connection Sci., Taylor & Francis, Aug. 2022, doi: 10.1080/09540091.2022.2117274.
[9] Q. Zeng, "Enhanced analysis of large-scale news text data using the bidirectional-Kmeans-LSTM-CNN model," PeerJ Comput. Sci., PMC11323039, 2024.
[10] H. Padalko, V. Chomko, and D. Chumachenko, "A novel approach to fake news classification using LSTM-based deep learning models," Front. Big Data, vol. 6, art. 1320800, Jan. 2024.