International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(21s) 2026 pp. 674-687
https://doi.org/10.70917/ijcisim-2026-5336
Article

News Classification Using Hybrid Natural Language Processing Based Content Modelling with Machine Learning Methods

P. Malaiarasu[1], Dr. R. Kalaimagal[2]

[1]Research Scholar, PG & Research, Department of Computer Science, Government Arts College for Men(A), Nandanam
[2]Associate Professor, PG & Research, Department of Computer Science, Government Arts College for Men(A), Nandanam

Abstract: With the rapid expansion of digital media, automatic news classification has become essential for managing and filtering the overwhelming volume of online information. Accurate classification enables efficient access to relevant content and supports applications such as personalized recommendations, trend detection, and misinformation control. This study presents a scalable framework for news categorization on a big data platform, improving both effectiveness and efficiency in handling massive datasets. The method integrates Term Frequency–Inverse Document Frequency (TF-IDF) to capture statistical word importance with Bidirectional Encoder Representations from Transformers (BERT) for deep contextual semantics. Hybrid features are processed in a distributed environment using PySpark for fast and reliable computation. By combining TF-IDF and BERT for hybrid feature representation and processing them with an Extra Trees Classifier (ETC) Machine Learning (ML) model, the framework achieves 90.36% accuracy, demonstrating the effectiveness of integrating statistical, contextual, and ensemble-based learning for robust news classification. Experimental evaluation shows the hybrid model outperforms single-model baselines in both accuracy and robustness. By leveraging TF-IDF, BERT, and PySpark, the framework provides an effective NLP-based solution for real-time news classification and intelligent information management in Big Data.
Keywords: News Classification, TF-IDF, BERT, NLP, Big data, Extra Tree Classifier, ML Classifiers

1. Introduction

In today's digital age, the massive growth of online information has created both opportunities and challenges. On one hand, easy access to abundant data provides great convenience; on the other, it raises the difficulty of efficiently identifying and extracting valuable content from overwhelming volumes of text. To address this issue, researchers have increasingly turned to information retrieval and data mining techniques, which have driven significant advances in NLP. Among these developments, text classification has become a core method for organizing large-scale textual data. By extracting discriminative features from raw text and mapping them to predefined categories, it helps reduce the complexity of unstructured information and enables effective knowledge management [1].

The explosive growth of digital content has made automatic news categorization an essential requirement. With information spreading rapidly across diverse online platforms, there is a pressing demand for intelligent computational techniques that can reliably sort and predict news topics. Despite ongoing progress, many existing models for news classification and event detection still face limitations. They often struggle with large-scale, hierarchical categorization tasks and tend to produce lower levels of precision and accuracy, underscoring the need for more robust and scalable solutions [2].

News serves as a primary source of information about contemporary society. However, in today's fast-paced digital environment, the sheer volume of news makes it impossible for individuals to follow every event as it unfolds.

As a result, users often rely on categorized news feeds and personalized recommendations tailored to their interests. From the perspective of news organizations, timely categorization and systematic archiving of large volumes of articles are equally critical. These processes support efficient content management, enable rapid analysis of public opinion, and help track emerging social trends and hot topics [3].

Online news articles typically share several defining characteristics. First, they often adopt a conversational writing style. Second, the information provided is usually brief, presenting the essential facts of an event along with concise descriptions or, at times, the author's perspective. Third, the language tends to avoid complex vocabulary or grammar, making the content straightforward and easily understood. These stylistic features allow news articles to both inform and influence audiences through simple, direct communication. In most cases, online news topics form thematically consistent groups of documents, where categories are predefined by experts in information-analysis systems. Automatic classifiers are designed to recognize the subject matter of these documents and assign them to the appropriate categories [4].

Current studies on text categorization primarily emphasize two aspects: the representation of text and the design of classification models. In most cases, representation methods rely on the vector space model or its extensions, where documents are expressed as vectors whose elements correspond to semantic features. Despite its popularity, this approach often produces unsatisfactory results for short texts because of feature sparsity. To overcome this challenge, researchers have experimented with techniques such as incorporating association rules through external knowledge bases or applying semantic enrichment methods to expand the feature set. Nevertheless, the success of these solutions is largely constrained by the reliability of the knowledge base and the complexity of semantic modeling, which makes it difficult to capture the complete meaning of short textual content [5-7].

Recent developments in NLP have significantly transformed the way computers interpret and understand human language. NLP explores the interaction between language and computational systems, employing algorithmic techniques to analyze, represent, and process text or speech in a manner that mimics human understanding across diverse applications. This field has attracted considerable research interest and continues to present both promising opportunities and complex challenges [8]. In this study, NLP techniques were applied to preprocess and categorize news content. Among NLP tasks, text classification is one of the most widely utilized and serves multiple purposes, including the automatic categorization of news articles into predefined topics. The preprocessing pipeline implemented in this research involved several stages: converting text to a uniform case, cleaning the data, eliminating stop words, performing stemming, and tokenizing the text for further analysis [9].

The introduction of BERT represents a major revolution in the ability of NLP models to understand meaning in a robust, contextual way [10]. Unlike traditional models that analyze text word by word, BERT uses self-attention that considers every word at once, enabling bidirectional context. This means that a model can understand the meaning of a word means from the words around it, which has led BERT to achieve state-of-theart results on several NLP benchmarks [11]. We therefore use the Reuters dataset to illustrate the ability of a fine-tuned, pretrained BERT model for news classification. The Reuters dataset is commonly used in text classification research and consists of thousands of news articles spread across several categories. Using the pre-trained BERT model, we aim to determine whether transfer learning would yield significantly better classification performance than machine learning focused techniques. In NLP, one of the most prevalent methods for extracting features in classification of text, particularly in categorizing news articles, is TF-IDF. This approach determines how relevant a term is in a specific document compared to its overall occurrence across the corpus relative to its frequency across a corpus, helping to identify distinctive terms that can effectively represent news content. Recent studies have demonstrated the continued relevance and enhancement of TF-IDF in news classification.

The structure of this study is outlined as follows: In Section 2, an overview of prior research concerning news categorization and approaches to feature representation is provided. Section 3 introduces the adopted framework, which integrates hybrid feature extraction through TF-IDF and BERT along with ML classifiers. Section 4 explains

the datasets, evaluation measures, and experimental configuration. Section 5 interprets the findings and performance outcomes, while Section 6 offers the concluding remarks and suggests possible directions for future exploration.

2. Literature Review

Islam and Anas., have stated current digital era, the exponential growth of online news articles has posed significant challenges for users attempting to navigate the vast information landscape. To mitigate this issue, automated news classification systems have emerged as a critical research focus, aiming to categorize news content into predefined classes to enhance accessibility and information retrieval. Several studies have investigated the application of ML algorithms for this purpose, examining their efficiency, accuracy, and suitability for handling large-scale datasets. These systems leverage a range of feature extraction and classification techniques to optimize performance, highlighting the potential of ML-driven approaches in streamlining news consumption and supporting real-time information processing [12].

Zhao et al., have proposed a method for text classification that relies on analysing only portions of the available content. In the current digital landscape, financial documents often display uneven quality and missing details. By focusing on partial texts, classification techniques can better reflect these real-world conditions where information is frequently incomplete. However, categorizing documents based on fragments of text proves to be more complex and challenging compared to working with full-length content [13].

Loureiro et al., have suggested a model which built on the Transformer architecture, especially BERT, have been examined extensively for their capacity to resolve lexical ambiguity, and the findings highlight their strong ability in word sense disambiguation tasks. Nevertheless, investigating the performance of embedding models such as BERT for processing Indonesian texts, with a particular focus on classifying ambiguous sentences through binary classification methods, remains a crucial area of study [14].

Xu et al., have shown that text categorization is more widely used in NLP and information management. BERT is an improved model based on the Transformer encoder, which belongs to the bi-directional language model and can make full use of contextual information. This model is suitable for text categorization natural language understanding tasks. Traditional ML has limited representational capabilities in text categorization, with shallow semantic, structural and contextual understanding of text. Deep learning makes up for the shortcomings of traditional ML in text categorization and improves the ability to understand the context, but it also suffers from disadvantages such as poor model interpretability and difficulty in adjusting features [15-17].

Liang et al., have focused on enhancing automated text classification by integrating traditional statistical methods with deep learning approaches. One such approach involved modifying the TF-IDF algorithm and adapting the input structure of BiLSTM networks. In this method, the TF-IDF formula was redefined and combined with a sliding window technique to better capture contextual information. Additionally, word2vec embeddings were employed to represent semantic features of the text, which were then processed through a hybrid model consisting of BiLSTM and Text-CNN. The BiLSTM component was utilized to capture sequential dependencies across the entire text, while the Text-CNN effectively extracted local, sentence-level features. This combination resulted in improved classification accuracy, demonstrating the strength of hybrid models that integrate statistical feature extraction with deep learning architectures [18].

Li et al., have conducted a comprehensive review of text classification approaches spanning the years 1961 to 2021, grouping them into conventional methods and those based on deep learning. Their work examines advancements in techniques, commonly used benchmark datasets, and evaluation strategies, while also providing comparisons among different models. The study ends by highlighting major implications, outlining prospective research paths, and addressing unresolved challenges. In addition, it summarizes the progress achieved by both traditional and deep learning frameworks, analyzes datasets and evaluation practices, and reports quantitative findings of top-performing models while emphasizing future research issues [19].

Korobeinikov et al., have developed a predictive framework in Python for identifying fake news, employing deep learning methodologies. Text representation was achieved through semantic techniques that utilized inverse document frequency weighting in combination with term frequency statistics. For performance measurement, a diverse set of algorithms was applied. Among the ML models, Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbor (KNN), AdaBoost, Support Vector Machines (SVM), Decision Trees (DT), Neural Networks (NN), and Naïve Bayes (NB) were examined. Alongside these, deep learning architectures including Recurrent Neural Networks (RNN), Gated Recurrent Units (GRU), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) were also evaluated [20].

Although existing models have shown reasonable performance in news classification, they suffer from certain limitations such as feature sparsity, lack of semantic understanding, and difficulty in handling large-scale data efficiently. To overcome these drawbacks, this study proposes a novel hybrid approach that combines TF-IDF with BERT for effective text representation and employs ML classifiers for accurate categorization. This content-based hybrid model is expected to enhance classification performance, particularly in news domains, by leveraging both statistical and contextual features.

3. Research Methodology

The methodology adopted in this research focuses on a content-based approach for news classification within a big data environment to ensure scalability and efficiency. The process begins with the extraction of textual features using two complementary techniques: TF-IDF and BERT. While TF-IDF captures the statistical weight of terms across the dataset by emphasizing informative words and reducing the impact of common but less meaningful terms, BERT enriches this representation by providing deep contextual and semantic understanding of language patterns. The combination of these two feature representations creates a hybrid feature space that balances statistical significance with contextual relevance, thereby improving the overall quality of the input data. This hybridized feature set is then utilized by ML models to classify news articles into five predefined categories: Business, General, Technology, Science, and Environmental, Social, and Governance (ESG). By implementing this methodology on a big data platform, the approach leverages distributed computing to handle the volume, velocity, and variety of large-scale news datasets, ensuring that the framework is not only effective but also applicable to real-world scenarios where massive and continuously evolving data streams are involved.

Figure 1 begins with data collection, where raw news articles are gathered as input. These documents undergo NLP preprocessing to clean and normalize the text for further analysis. In the feature extraction stage, two complementary methods are applied: TF-IDF to capture statistical term-based features and BERT to generate rich contextual embeddings. These representations are then combined through hybrid feature construction, producing a unified feature vector [$F_{TFIDF}; F_{BERT}$]. To handle large-scale datasets efficiently, the hybrid features are processed in a distributed environment using PySpark, ensuring scalability and faster computation. The resulting features are fed into ML classifiers to build predictive models, followed by performance evaluation to assess their effectiveness. Finally, the system generates the news categorization output, classifying articles into their respective categories.

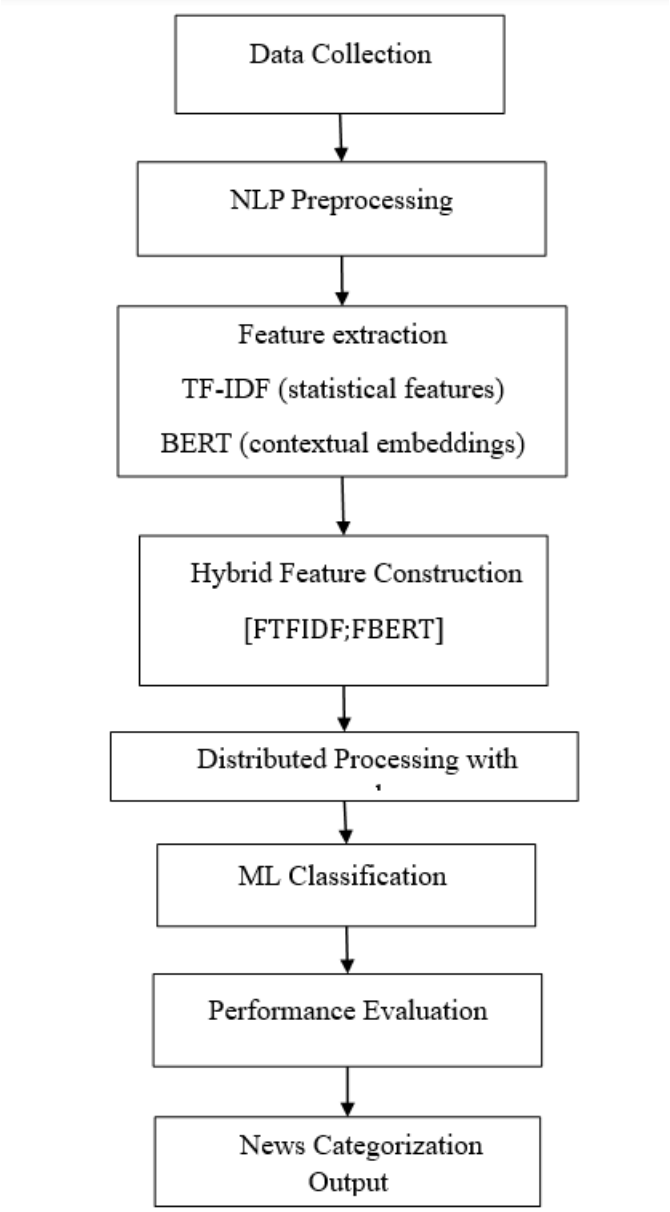
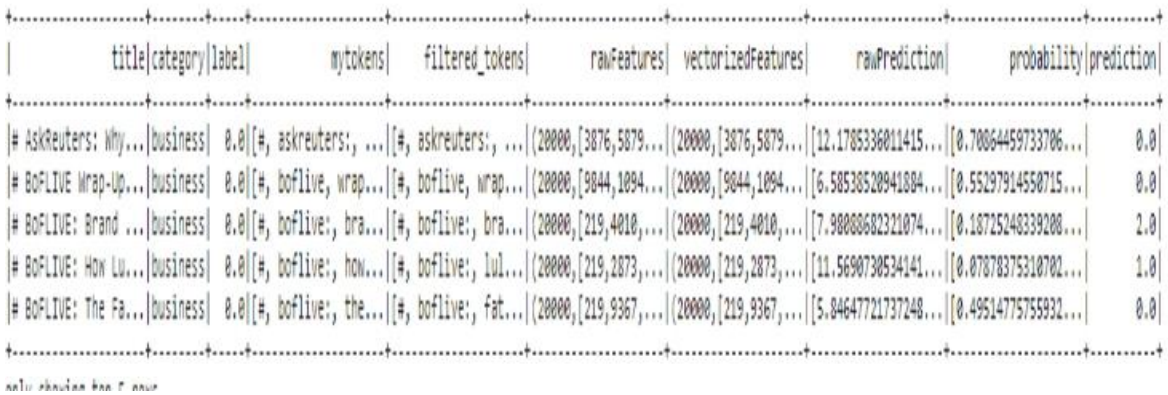


Figure 1 Overall Flowchart for the proposed model

Data Collection

In this study, a publicly available dataset of news articles was used for classification purposes. The dataset consists of a large corpus of text data categorized into five major classes: Business, General, Technology, Science, and ESG (Environmental, Social, and Governance). Each article includes a title and body text, which serve as the primary features for analysis. The dataset exhibits class imbalance, with certain categories containing significantly more samples than others. This characteristic was taken into account during preprocessing and model evaluation. The dataset's high volume and diversity make it appropriate for testing the scalability and effectiveness of big data processing frameworks like PySpark.

Data Preprocessing using NLP Techniques



title	category	label	mytokens	filtered_tokens	rawFeatures	vectorizedFeatures	rawPrediction	probability	prediction
# AskReuters: Why...	business	0.0	[#, askreuters:, ...	[#, askreuters:, ...	(20000,[3876,5879...	(20000,[3876,5879...	[12.1785336011415...	[0.7086445973370 6...	0.0
# BoFLIVE Wrap-Up...	business	0.0	[#, boflive, wrap...	[#, boflive, wrap...	(20000,[9844,1094...	(20000,[9844,1094...	[6.5853852094188 4...	[0.5529791455071 5...	0.0
# BoFLIVE: Brand ...	business	0.0	[#, boflive:, bra...	[#, boflive:, bra...	(20000,[219,4010,...	(20000,[219,4010,...	[7.9808868232107 4...	[0.1871524833920 8...	2.0
# BoFLIVE: How Lu...	business	0.0	[#, boflive:, how...	[#, boflive:, lul...	(20000,[219,2873,...	(20000,[219,2873,...	[11.569073053414 1...	[0.0787837531070 2...	1.0
# BoFLIVE: The Fa...	business	0.0	[#, boflive:, the...	[#, boflive:, fat...	(20000,[219,9367,...	(20000,[219,9367,...	[5.8464772173724 8...	[0.4951477575593 2...	0.0

only showing top 5 rows

Figure 2 Preprocessing and Model Prediction for News Samples

Figure 2shows the preprocessing and model prediction for news samples. In this research, data preprocessing is a critical step to transform raw and unstructured news articles into a clean and analyzable format suitable for classification. NLP techniques are systematically applied to ensure that the dataset is noise-free, consistent, and semantically meaningful. The preprocessing pipeline begins with text cleaning, where non-informative elements such as punctuation marks, special characters, hyperlinks, digits, and extra whitespace are removed. This is followed by tokenization, which segments the textual data into smaller units (tokens) that can be effectively processed by computational models. To reduce redundancy and improve feature quality, stop-word removal is carried out to eliminate high-frequency words that contribute little to semantic understanding. Lemmatization is then applied to convert words into their canonical form (e.g., *studies* → *study*), ensuring that different inflected forms of a word are treated uniformly. Finally, normalization techniques such as lowercasing and text standardization are employed to maintain uniformity across the corpus. These sequential steps refine the dataset by preserving only the most meaningful linguistic components, thereby establishing a robust foundation for feature extraction and subsequent classification tasks. Figure 1 shows the preprocessing and model prediction for news samples.

TF-IDF working principle

TF-IDF approach is a numerical technique frequently applied in NLP and text analytics to measure the significance of a word within one document compared to its presence in an entire collection. Its main function is to assign greater weight to words that appear often in a particular document but are relatively rare across the corpus.

The TF-IDF score for a term *t* in a document *d* can be mathematically represented as:

TF-IDF (*t, d*) = TF (*t, d*) × IDF (*t*)(1)

Term Frequency (TF):

TF refers to the number of times a word appears in a document, normalized by dividing by the total number of words in that document to avoid bias from longer texts.

$$TF(t, d) = \frac{Occurrences\ of\ t\ in\ d}{Total\ terms\ in\ d} \quad (2)$$

Inverse Document Frequency (IDF):

IDF indicates the rarity of a term within a dataset of documents. It reduces the importance of words that are widely distributed across many texts, as such terms provide minimal ability to differentiate between documents.

$$IDF(t) = log\frac{N}{DF(t)} \quad (3)$$

where

- N = The number count of documents contained in the corpus.,
- DF(t) = Number of documents that contain term t.

Formally:

$DF(t)=|\{d \in D: t \in d\}|$ (4)

Interpretation:

- High TF → the term is frequent within the document.
- Low DF → the term is rare across documents.
- High TF-IDF → the term is both locally frequent and globally distinctive.

This feature makes TF-IDF a valuable technique for generating attributes in tasks like automatic text categorization, grouping of documents, and retrieval of information. The value of Inverse Document Frequency (IDF) is obtained from Document Frequency (DF), which is computed by identifying the number of documents in which a particular term t is present. Unlike term frequency, DF simply checks the presence of a word across the document set, without considering how many times it is repeated within a single document.

BERT working principle

The BERT model is constructed on the principles of the Transformer encoder, forming a deep learning architecture for language representation, designed to generate contextual word embeddings for downstream NLP tasks. The model consists of multiple stacked encoder layers, where each layer applies the self-attention mechanism to capture bidirectional dependencies between words. The architecture BERT model is shown in figure 3.

Pre-training Phase

BERT undergoes pre-training on an extensive collection of unlabelled text by employing two unsupervised learning:

Masked Language Model (MLM): The approach consists of randomly masking certain tokens in the input sequence and training the model to recover these masked elements by relying on contextual signals from adjacent tokens.

LMLM=−t ∈MlogPωtω/t) (5)

Next Sentence Prediction (NSP): Predicts whether a sentence B logically follows sentence A.

LNSP=−[ylogPlabel+1−ylog(1−Plabel)] (6)

The combined pre-training objective is:

L=LMLM+LNSP (7)

Fine-tuning Phase

For downstream tasks (e.g., classification), the pre-trained model is fine-tuned on labeled data. Each input sequence is initialized with a classification marker and concluded with a separator marker. The latent vector linked to the classification marker position is interpreted as the comprehensive representation of the entire sequence.

A task-specific layer (Softmax classifier) is added on top of BERT to predict class probabilities:

$P(c|h) = softmax\ (Wh)$ (8)

where

- $h \in \mathbb{R}$ is the final hidden vector of the [CLS] token,

- W is the trainable task-specific parameter matrix,
- c is the predicted class label.

During fine-tuning, all BERT parameters and the classifier parameters (W) are updated jointly to maximize the log-likelihood of the correct label.

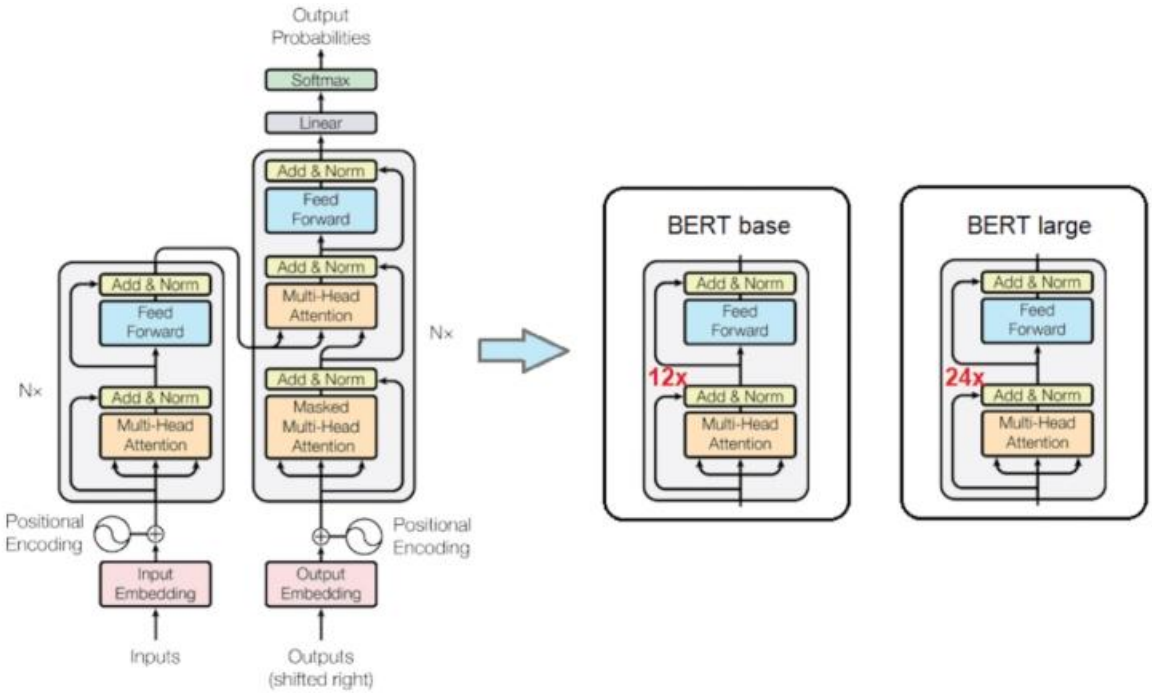


Figure 3 BERT Architecture

Hybrid Feature Fusion

To exploit the complementary strengths of statistical and contextual representations, this study integrates TF-IDF and BERT into a unified hybrid feature space. TF-IDF provides sparse vectors that emphasize the discriminative importance of terms within documents, thereby preserving keyword-level interpretability. In contrast, BERT generates dense contextual embeddings that capture semantic meaning and word dependencies within sentences. By concatenating these two representations, a hybrid feature vector is constructed:

$$F_{hybrid} = [F_{TFIDF}; F_{BERT}] \tag{9}$$

This fusion ensures that both lexical importance and semantic context are jointly preserved, resulting in a richer and more expressive representation. The hybrid features are then used as input to ML classifiers, enabling more accurate and robust news categorization compared to single-model baselines.

Algorithm of hybrid TF-IDF and BERT-based news classification in PySpark

Input: News dataset D, Labels L, Pre-trained BERT model MBERT

Output: Predicted categories $\hat{Y}$, optimized classification model, and evaluation metrics

1: Load dataset D into PySpark DataFrame

2: Preprocess D using NLP techniques:

→ Lowercasing, stopword removal, tokenization, lemmatization

3: Compute TF-IDF features FTFIDF using CountVectorizer + IDF in PySpark

4: Extract contextual embeddings FBERTfor each document using MBERT

5: Fuse features: F_{hybrid} = Concatenate $[F_{TFIDF}; F_{BERT}]$

6: Split FHybrid into training and testing sets

7: Train classifier C on training set

8: Predict categories $\hat{Y} = C$ (test set)

9: Evaluate model with Accuracy, Precision, Recall, and F1-score

10: Output final optimized classification model, predicted categories $\hat{Y}$, and performance report

4. Result and Discussion

This research is designed to build an efficient framework for classifying online news articles into five categories: Business, General, Technology, Science, and ESG. With the exponential growth of digital media, traditional keyword-matching techniques have proven inadequate for accurate categorization. To address this, the proposed approach integrates semantic feature extraction through BERT and statistical representation via TF-IDF, combined with ML classifiers to improve predictive performance. The framework also leverages Apache Spark's distributed environment, ensuring scalability and efficiency when processing massive datasets. It is expected that the hybrid model will achieve higher accuracy and robustness compared to single-model baselines, making it a practical solution for large-scale news classification.

Extra Tree Classifier

The ETC is an ensemble-based supervised learning algorithm that constructs a large number of DT and aggregates their predictions to achieve robust and accurate classification. Unlike conventional DT methods, at each node, instead of always identifying the most discriminative split, the Extra Trees algorithm introduces randomness by selecting thresholds randomly from the available features.This high degree of randomization reduces variance, enhances generalization capability, and accelerates training, while still maintaining strong predictive power. By leveraging ensemble averaging across multiple randomized trees, the ETC effectively mitigates overfitting and demonstrates superior performance, particularly in high-dimensional and complex classification tasks such as text and news categorization.

Classifier Performance Metrics

Classification across five categories is represented in figure 4 using a confusion matrix, which records the TF-IDF and BERT model with Extra Trees Classifier outcomes in terms of True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN). The diagonal values represent the TP, showing the correctly classified samples, such as 75,921 for class 0, 38,832 for class 1, 13,908 for class 2, 1,908 for class 3, and 549 for class 4. The off-diagonal values within each row indicate the FN, which are actual samples of a class that were misclassified into other categories—for example, 3,429 samples from class 0 were misclassified as class 1. Similarly, the off-diagonal values within each column correspond to FP, where samples from other classes were incorrectly predicted as belonging to that class, such as 3,517 class 1 samples being wrongly classified as class 0. All other values outside the corresponding row and column for a given class can be considered TN, representing correctly identified non-class samples. Overall, the model shows strong true positive rates for classes 0, 1, and 2, while classes 3 and 4 exhibit smaller TP counts and higher relative misclassifications, reflecting greater difficulty in accurate prediction for those categories.

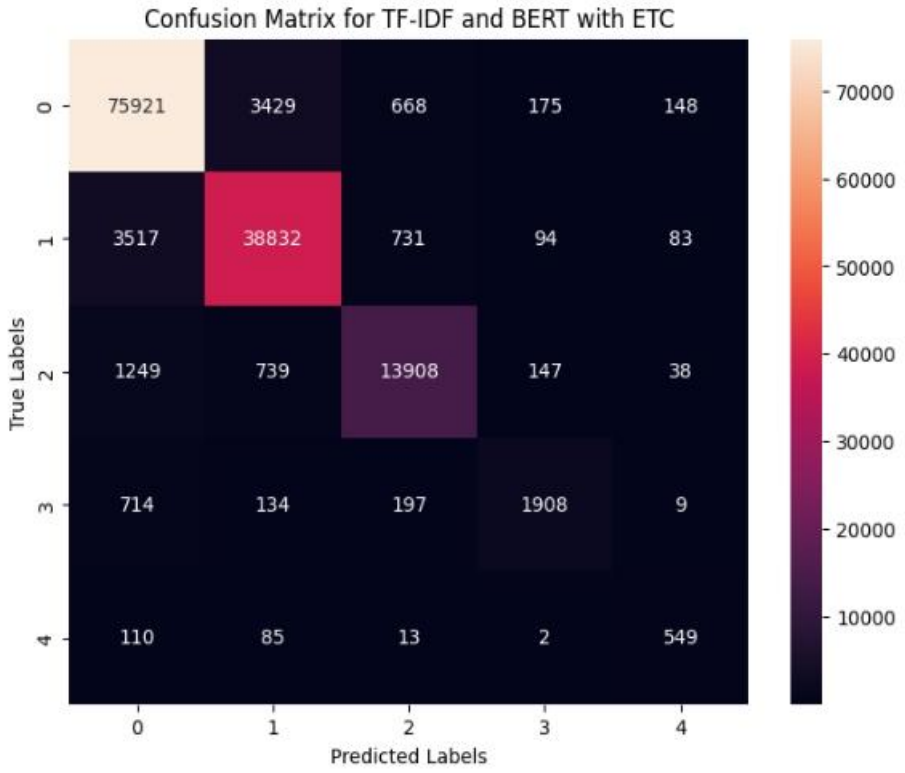


Figure 4 Confusion Matrix Heatmap for the Proposed Hybrid Model

Once the topic vectors are prepared, they are used to train and test ML models using Spark MLlib. The study utilizes three supervised classification algorithms ETC, RF, and XGBoost to evaluate the effectiveness of the semantic features. The motivation behind using these classifiers lies in their varying capabilities to handle linearity, non-linearity, and probabilistic structures in data. By combining the strengths of unsupervised topic modelling with supervised classification, the system aims to achieve robust categorization of news articles. The classification models are evaluated on key performance metrics such as Micro Precision, Micro Recall, and Micro F1-Score to ensure a comprehensive assessment.

Table 1 Performance comparison of hybrid Models (TF-IDF + BERT) with different classifiers

Classifier	**Micro Precision (%)**	**Micro Recall (%)**	**Micro F1-Score (%)**
TF-IDF + BERT + ETC	90.36	90.36	90.36
TF-IDF + BERT + RF	84.61	84.61	84.61
TF-IDF + BERT + XGBoost	83.92	83.92	83.92

The table 1 presents the performance comparison of three hybrid classifiers that integrate TF-IDF and BERT for feature representation with different machine learning algorithms for classification. Among the models, TF-IDF + BERT + ETC achieved the highest performance with a micro precision, recall, and F1-score of 90.36%, indicating better consistency and balanced accuracy across all classes. The RF model followed with 84.61%, showing slightly lower effectiveness compared to the ETC but still maintaining a competitive performance. On the other hand, XGBoost produced the lowest results with 83.92%, suggesting that its ensemble boosting mechanism was less effective for this dataset when combined with TF-IDF and BERT features. Overall, the results highlight that the ETC classifier is the most effective in this hybrid setup.

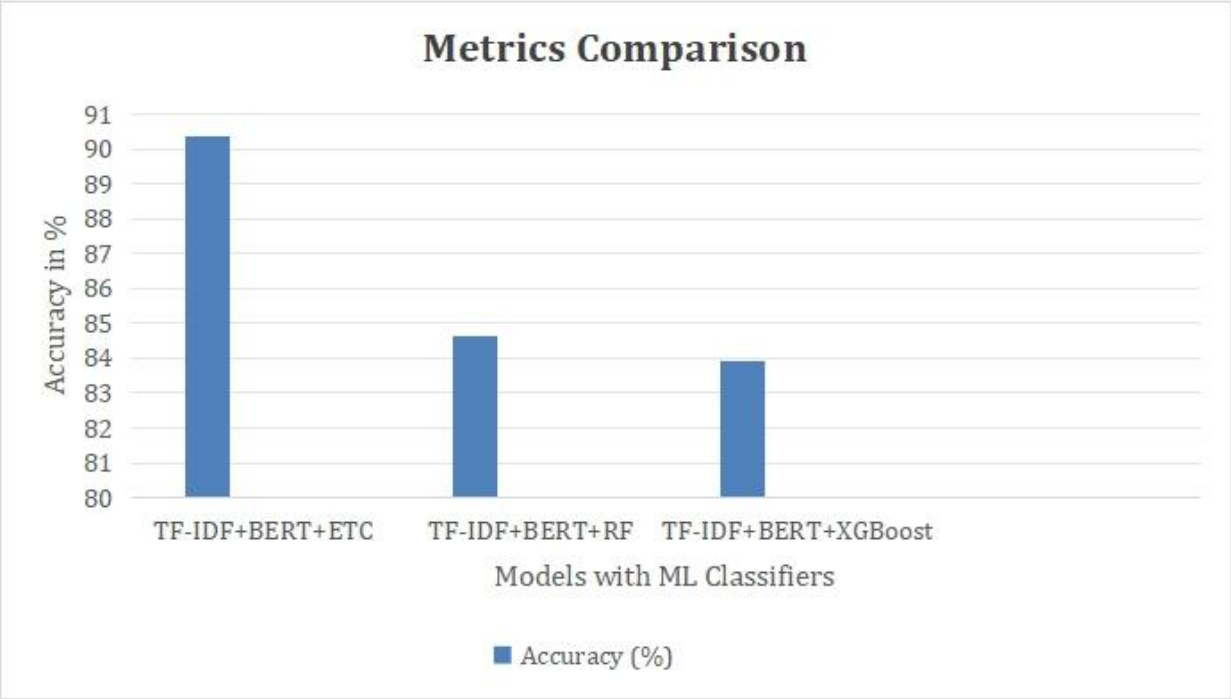


Figure 5 Performance metrics of ML classifiers for News categorisation

Figure 5 illustrates the micro precision (%) performance of three hybrid models that combine TF-IDF and BERT with different classifiers. The results clearly show that TF-IDF + BERT + ETCachieved the best precision at approximately 90.36%, outperforming the other approaches. The RF model comes next with a precision of about 84.61%, while the XGBoost model records the lowest precision at 83.92%. This indicates that, in this experimental setup, the ETC was most effective at correctly classifying instances across all classes, whereas XGBoost struggled comparatively, delivering the weakest precision among the three. Overall, the graph highlights that ETC, despite being a relatively simple ensemble tree-based method, can sometimes outperform more complex models when combined with TF-IDF and BERT features.

Output for Existing LDA Model

```
News Article:
The government has launched a new scheme to support startups in the technology sector.
It aims to create more jobs and boost innovation.

Detected Topic: Technology

Score Breakdown per Topic:
- Business: 0.20
- General: 0.10
- Technology: 0.60
- Science: 0.05
- ESG: 0.05
```

Figure 6 LDA model output for news classification

Output for proposed model

Figure 7 Output for the proposed hybrid model

Figures 6 and 7 depict the comparative performance of LDA and the hybrid TF-IDF + BERT + ETC model. LDA applies a topic-based approach that groups articles by latent topic distributions, which highlights general themes but does not fully capture semantic context. In contrast, the hybrid model follows a content-based framework where TF-IDF emphasizes the statistical importance of key terms such as technology, startups, and innovation, while BERT provides contextual embeddings that interpret words within their sentence structure, for example distinguishing that technology sector denotes an industry and boost innovation implies promotion of creativity. By integrating these representations, the ETC classifier evaluates enriched feature vectors against predefined categories (Business, General, Technology, Science, ESG) and assigns confidence scores to each class. The final classification is determined by the highest score, while the score distribution indicates alignment strength across categories. This approach demonstrates that the hybrid model generates content-based outputs with greater precision and reliability compared to LDA, which relies only on topic similarity.

Table 2 Accuracy Metrics for Various Text Representation Techniques

Model	**Accuracy (%)**
TF-IDF+ BERT+ ETC	90.36
LDA + LR	86.05
TF-IDF	83.89
Count Vectorization	65.96

The experimental evaluation in table 2 demonstrates clear variations in accuracy across different models. The hybrid TF-IDF + BERT + ETC outperformed all other models with an accuracy of 90.36%, highlighting the effectiveness of combining semantic embeddings from BERT with statistical features from TF-IDF. The LDA + Logistic Regression (LR) model followed with 86.05%, indicating that topic-based features also contribute to robust classification, though less effectively than hybrid approaches. Traditional TF-IDF alone achieved 83.89%, showing the strength of statistical word frequency methods but with limitations in capturing contextual meaning. In contrast, Count Vectorization recorded the lowest accuracy at 65.96%, suggesting that raw frequency counts without weighting or semantic representation are insufficient for handling complex news classification tasks.

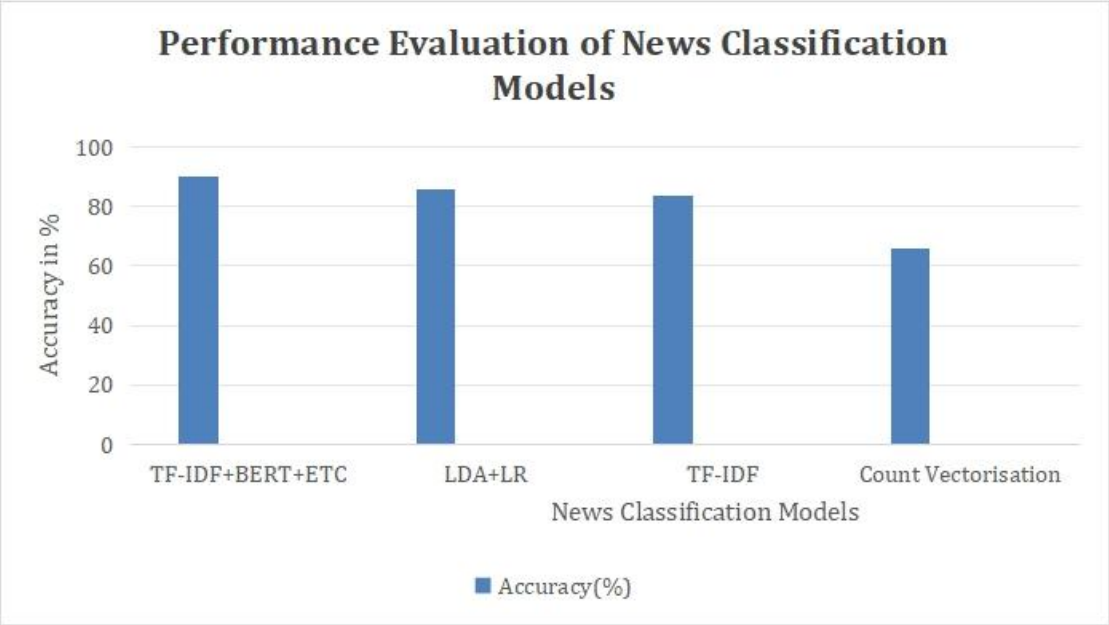


Figure 8 Accuracy comparison of Text Representation Techniques in News Classification

Figure 8 presents the performance evaluation of four news classification models in terms of accuracy. The TF-IDF + BERT + ETCmodel achieve the highest accuracy of 90.36%, demonstrating the effectiveness of combining traditional feature extraction with deep contextual embeddings and ensemble learning. LDA + LR follows with an accuracy of 86.05%, showing that topic modelling integrated with a linear classifier can also produce strong results. The standalone TF-IDF model attains an accuracy of 83.89%, reflecting its reliability but also indicating that hybrid approaches offer better performance. In contrast, Count Vectorization records the lowest accuracy of 65.96%, highlighting the limitations of simple frequency-based methods for complex news classification tasks.

5. Conclusion

This study demonstrates that integrating TF-IDF with BERT features and using the ETC classifier on the PySpark big data platform provides an effective and scalable framework for automated news classification. By combining statistical term importance TF-IDF with deep contextual embeddings BERT, the hybrid model captures both word-level and semantic patterns, enhancing classification performance over traditional methods. The hybrid TF-IDF + BERT + ETC model achieved an accuracy of 90.36%, highlighting its capability to accurately classify multi-class news articles. Designed for real-time deployment, the system can support applications such as live news filtering, financial analysis, and content recommendations. Overall, the proposed framework offers a robust and high-performing solution for managing large-scale, dynamic news data in contemporary big data environments.

REFERENCES

1. Asgarnezhad R., Soltanaghaei M., and Monadjemi S. A., An application of MOGW optimization for feature selection in text classification, *The Journal of Supercomputing*. (2020) 16, no. 15, 154–160.
2. A. Barua, O. Sharif, M.M. Hoque Multi-class sports news categorization using machine learning techniques: resource creation and evaluation Procedia Comput. Sci. (2021), pp. 112-121, 10.1016/j.procs.2021.11.002
3. K.M. Hasib, N.A. Towhid, K.O. Faruk, J. Al Mahmud, M.F. Mridha Strategies for enhancing the performance of news article classification in Bangla: handling imbalance and interpretation Eng. Appl. Artif. Intell., 125 (2023), 10.1016/j.engappai.2023.106688
4. Miao J.:Zhang Y.S.,Li J.L.:Classification of Chinese News Headlines Based on Bidirectional Encoder Representations from Transformers (BERT)[J]. Computer Engineering and Design, 43(08):2311-2316(2022)
5. Jing W., Bailong Y. News text classification and recommendation technology based on wide & deep-bert model, 2021 IEEE International Conference on Information Communication and Software Engineering (ICICSE). IEEE, 209-216 (2021).

6. Tan Z., Chen B., Fang W.:Analysis and Application of Financial News Text in Chinese Based on Bert Model, Proceedings of the 2020 Asia Service Sciences and Software Engineering Conference, 35-39 (2020).
7. Ambalavanan A K, Devarakonda M V.: Using the contextual language model Bidirectional Encoder Representations from Transformers (BERT) for multi-criteria classification of scientific articles, Journal of biomedical informatics,112, 103578 (2020)
8. D. Khurana, A. Koli, K. Khatter, and S. Singh, “Natural language processing: state of the art, current trends and challenges,” *Multimed. Tools Appl.*, vol. 82, no. 3, pp. 3713– 3744, Jan. 2023, doi: 10.1007/s11042-022-13428-4.
9. Sowmya V. B., B. Majumder, A. Gupta, and H. Surana, *Practical natural language processing: a comprehensive guide to building real-world NLP systems*, First edition. Sebastopol, CA: O’Reilly Media, 2020.
10. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, 2019, pp. 4171–4186, https://doi.org/10.18653/ v1/N19-1423.
11. H. Yuan, "Natural Language Processing for Chest X-Ray Reports in the Transformer Era: BERT-Like Encoders for Comprehension and GPT- Like Decoders for Generation," iRADIOLOGY, Jan. 2025, https:// doi.org/10.1002/ird3.115.
12. M. J. Islam and S. S. Anas, "News Classification Using ML Algorithms: An Exploration of Efficient Information Categorization,"*International Journal of Creative Research Thoughts (IJCRT)*, vol. 11, no. 6, Jun. 2023.
13. W. Zhao, G. Zhang, G. Yuan, J. Liu, H. Shan, and S. Zhang, "The Study on the Text Classification for Financial News Based on Partial Information," in IEEE Access, vol. 8, pp. 100426-100437, 2020, DOI: 10.1109/ACCESS.2020.2997969.
14. Loureiro, D., Rezaee, K., Pilehvar, M.T., Camacho-Collados, J. (2021). Analysis and evaluation of language models for word sense disambiguation. Computational Linguistics, 47(2): 387-443. https://doi.org/10.1162/coli_a_00405
15. Xu P.: Research on news text classification method based on deep learning,Nanjing University of Information Engineering, (2023).
16. Wu Y.: Research and Application of Deep Learning Based Text Classification of News Headlines, Yangzhou University, (2024).
17. Lu Z.B.: Research on Text Classification Algorithm for Chinese News Headlines Based on Deep Learning, Southwest University, (2023)
18. Minhui Liang, Tiansen Niu, Research on Text Classification Techniques Based on Improved TF-IDF Algorithm and LSTM Inputs, Procedia Computer Science, Volume 208,2022, Pages 460-470, bISSN 1877-0509, https://doi.org/10.1016/j.procs.2022.10.064.
19. Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P.S. Yu, L. He, A survey on text classification: from traditional to deep learning, ACM Trans. Intell. Syst. Technol. 13 (2) (2022) 31
20. Korobeinikov A, Bangyal WH, Qasim R, Rehman Nu, Ahmad Z, Dar H, et al. Detection of fake news text classification on COVID-19 using deep learning approaches. Comput Math Methods Med 2021;2021. http://dx.doi.org/10.1155/ 2021/5514220.