Dr. Satender, 2026, 14:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

News Classification Using Machine Learning and Deep Learning: A Comparative Study

Dr. Satender[1], Dr. Brij Mohan[2], Dr. Raj Kumar[3]
[1]Dean Academic Quantum University
[2]Director QST Dept. Quantum University, Roorkee, India
[3]Assistant Professor Quantum University Roorkee

Abstract- In Natural Language Processing (NLP), news classification is a key task that involves automatic categorization of news articles into predefined topics, which allows for efficient content organization and information retrieval. This paper presents a news classification model that is based on machine learning. Collecting news datasets, pre-processing text, extracting features using TF-IDF, and classification using multiple algorithms, including Naive Bayes, Logistic Regression, and Deep Learning (LSTM), are the main tasks of the proposed method. Technology, Business, Sports, Education, and Entertainment are among the categories included in the dataset. The model's evaluation is based on metrics such as accuracy, precision, recall, and F1-score. The LSTM model has experimental results that demonstrate its highest accuracy of 91%, surpassing Logistic Regression (87%) and Naive Bayes (85%). Automated content categorization in the media sector is furthered by the study.

Keywords: news classification; natural language processing; machine learning; deep learning; LSTM; TF-IDF; Naive Bayes; Logistic Regression.

I. INTRODUCTION

Modern Natural Language Processing (NLP) has made automated classification an essential challenge due to the exponential growth of digital news content. The need for accurate, scalable, and interpretable classification systems is becoming more apparent as autonomous and AI-driven applications continue to revolutionize various sectors. Real-time sensing and categorization of articles is necessary for news classification systems to enable efficient content organization, personalized recommendations, and information retrieval.

Deep neural networks are utilized by modern text classification systems for these tasks. Pattern recognition and sequential modeling have been demonstrated to be strong by CNNs [8] and RNNs [3]. Document classification can benefit from the use of Long Short-Term Memory (LSTM) networks, which are a variant of RNNs and effective at capturing long-term dependencies in text. Recently, transformer-based architectures [13] have established new standards for NLP tasks, which include BERT [7] for pre-trained language comprehension and FastText [9] for lightweight, scalable classification. High performance alone cannot guarantee trust and practical deployment in critical applications because deep networks are often opaque and the internal decision process is not easily interpretable.

The purpose of this paper is to compare Naive Bayes [18], Logistic Regression [22], and LSTM [6] approaches for news classification, with TF-IDF [14] as the main feature extraction technique. The purpose of the work is to address the shortcomings in existing research [1], [12], which include a lack of diversity in datasets, insufficient real-time classification capabilities, and a lack of model interpretability. The goal of this study is to find the most effective approach for automated news categorization on a large scale by addressing these gaps.

A. Applications

- Real-time curation and delivery of personalized content on news platforms can be achieved by automating news aggregation by classifying articles into topics.
- The model for detecting fake news can be extended to social media monitoring to identify articles that are misleading or irrelevant.
- Advanced Driver Assistance Systems (ADAS) provide transparent reasoning alongside

predictions to inform safety-critical decision-making through interpretable models.
- In dynamic environments, intelligent analysis pipelines can benefit from explainable classification outputs for Smart Traffic and Surveillance.
- Classification systems are used by mobile robots to ensure safe navigation and contextual understanding in robotics and navigation.

II. LITERATURE REVIEW

For many years, news classification has been a well-known area of research in NLP. This task has been explored through various methods, including traditional machine learning algorithms and advanced deep learning models.

A. Traditional approaches to machine learning
Traditional machine learning techniques such as Support Vector Machines (SVM) [2], [22], Naive Bayes [18], and k-Nearest Neighbors (KNN) [1], [12] were used in early news classification works. Term Frequency-Inverse Document Frequency (TF-IDF) [14] is a typical method used by these models to convert text data into numerical features for classification. Although they are commonly used because of their simplicity and effectiveness, they are often unable to capture the deeper semantics of text and struggle with high-dimensional data.

B. Approaches to deep learning
News classification has been revolutionized by recent advances in deep learning. Complex patterns within text can be captured using CNNs [8] and RNNs [3]. Understanding local patterns, such as phrases and word combinations [8], is achieved by CNNs, while RNNs capture sequential information [11.] By using LSTM networks, RNNs can be extended to model long-range dependencies. GRU-based encoder-decoder architectures [19] are an alternative to LSTM that is computationally lighter and still offers competitive performance.

C. Approaches that combine hybrids and transformers
Hybrid models that combine traditional and deep learning techniques have been explored by several studies. Feature representation can be enhanced by integrating pre-trained word embeddings, such as Word2Vec [4] and GloVe [5], with LSTM networks. By paying attention to salient words and sentences, hierarchical attention networks [10] further enhanced document classification. Contextualized word representations have been utilized by transformer-based models such as BERT [7], ELMo [20], and FastText [9] to demonstrate state-of-the-art performance on text classification benchmarks. Vaswani et al. introduced a mechanism that focuses on self-attention. Many of these advances are based on [13]. LDA [15] has been utilized as an unsupervised pre-processing step to inform downstream classifiers in topic modeling.

D. There are gaps in the work that has already been done.
- Small or domain-specific datasets are often used in existing studies, which may not accurately represent the diversity of real-world news articles.
- Batch processing is the main focus of real-time classification research, but scaling real-time classification is still a challenge.
- Contextual understanding: Deep learning models still struggle with understanding the overall article context, especially when multiple topics are discussed.
- The practical deployment of deep learning models in applications that require transparent decision-making can be limited by their black box nature.

III. OBJECTIVES OF THE STUDY

The major aims of this study are as follows:
- To develop a machine learning news classification model that accurately categorizes news articles into predefined topics.
- The aim is to evaluate and contrast the performance of various algorithms, such as SVM, Naive Bayes, Logistic Regression, LSTM/GRU, and fine-tuned BERT.
- Standard metrics such as accuracy, precision, recall, and F1-score are used to analyze model efficacy.

- The objective is to find the algorithm that best balances accuracy and scalability for large-scale news classification.

IV. SCOPE AND LIMITATIONS

A. Scope

- The research focuses on categorizing news articles into predetermined categories such as politics, sports, business, entertainment, and technology.
- An openly available labeled news dataset is used to test the system.
- To find the most effective classification approach, the study compares traditional machine learning algorithms with modern deep learning methods.

B. Limitations

- The performance of a model is highly influenced by the diversity and quality of datasets. Performance may be significantly impacted with significantly different data.
- Optimizing the model for efficiency and speed may be necessary when applying it to real-time applications.
- Contextual comprehension may struggle with nuanced or ambiguous news stories, potentially resulting in misclassification.
- To accommodate previously unseen categories, the model requires retraining or fine-tuning.

V. IMPLEMENTATION METHODOLOGY

A. Methodology Design

The methodology utilizes a quantitative approach to evaluate the effectiveness of various machine learning and deep learning models for news classification. The main objective is to evaluate the accuracy and efficiency of different classification algorithms when categorizing news articles into predefined categories, like education, sports, entertainment, and technology.

B. Data Collection

The data is composed of a news dataset that is publicly available from Kaggle and contains labeled news articles in multiple categories such as business, sports, politics, and entertainment. A balanced and diverse set of articles is provided in the dataset to support robust model training and evaluation.

C. Data Preprocessing

Preparing raw text data for machine learning models requires data preprocessing to be a crucial step. The following steps were executed:

- Text cleaning involves the removal of special characters, punctuation, and numbers, as well as the conversion of text to lowercase.
- The process of splitting text into individual words or tokens is known as tokenization.
- Stopword Removal: Eliminating common words (e.g.The words 'the,' 'is,' and 'and' do not play a role in classification.
- Stemming and Lemmatization: Reducing words to their base or root form (e.g., "running" → "run").
- The extraction process includes TF-IDF vectorization for traditional models and sequence padding for LSTM models.

D. Tools and techniques

- Python is a programming language that can be used for machine learning and NLP, utilizing its robust ecosystem.
- Naive Bayes, SVM, and KNN classifiers can be implemented using Scikit-learn.
- For the purpose of manipulating, cleaning, and preparing data.
- Model performance metrics, including confusion matrices and accuracy graphs, are visualized using Matplotlib and Seaborn.
- Google Colab is a platform that is cloud-based and interactive to execute and experiment with code.

E. Metrics for evaluating models

- Accuracy: Measures the percentage of correctly classified news articles.
- Precision: Indicates how many articles predicted to be in a given category actually belong to that category.
- Recall: Measures how many actual articles in a category were correctly identified by the model.

- F1-Score: The harmonic mean of precision and recall, providing a balanced measure for classification tasks with imbalanced datasets.

VI. MODEL COMPARISON

A. Experimental Setup

Technology, Business, Sports, Education, and Entertainment are the five categories that make up the dataset. Traditional models require text cleaning, tokenization, and vectorization using TF-IDF for pre-processing, while LSTM models require sequence padding. The three models evaluated in this study are summarized in Table I.

TABLE I. COMPARISON OF CLASSIFICATION MODELS

Model	Description
Naïve Bayes	A probabilistic classifier based on Bayes' theorem with strong independence assumptions. Fast and interpretable.
Logistic Regression	A linear classifier that estimates probabilities using the logistic function. Useful for simple, linearly separable data.
LSTM (Deep Learning)	A type of RNN that captures long-term dependencies in text. Suitable for complex and context-heavy inputs.

B. Quantitative Comparison

Table II presents the quantitative comparison of the three models across all evaluation metrics.

Metric	Naive Bayes	Logistic Regression	LSTM	
Accuracy	85%	87%	91%	
Precision	0.84	0.86	0.92	
Recall	0.83	0.85	0.91	
F1-Score	0.83	0.85	0.91	
Training Time	Few seconds	A few seconds	~5 minutes	

C. Qualitative Analysis

The Naive Bayes model [18], while fast and easy to implement, underperformed slightly in terms of accuracy due to its assumption of feature independence [1]. Logistic Regression [22] improved classification accuracy by better modeling feature interactions. However, the LSTM model [6] demonstrated the highest performance across all metrics, benefiting from its ability to capture sequential dependencies and contextual patterns within text [11]. The trade-off is a significantly higher training time of approximately five minutes compared to the near-instantaneous training of the other two models. For contexts where performance must be balanced against speed, FastText [9] or lightweight transformer variants [16] may offer a practical middle ground.

VII. EXPECTED OUTCOMES AND SIGNIFICANCE

A. Expected Outcomes

- **Accurate News Classification System:** A model that accurately classifies news into predefined categories with high accuracy, precision, and recall.
- **Comparative Analysis:** A comprehensive comparison of traditional machine learning models versus deep learning approaches for text classification.
- **Optimized Feature Extraction:** A well-optimized TF-IDF feature extraction pipeline that enhances the model's ability to identify important terms and patterns.
- **Real-Time Classification Capability**: An optimized model suitable for processing news articles in real time, enabling use in dynamic news aggregators and recommendation systems.
- **Scalable Solution:** A classification model designed to handle large volumes of continuously generated content in practical applications.

B. Significance of the Project

- **Advancing Automated Categorization:** Reduces the time and effort required for manual sorting of articles, enabling media companies to manage content more effectively.

- **Enhanced User Experience:** Accurate classification improves content discovery and user engagement on news platforms.
- **Improved Personalization:** Classification systems enable more sophisticated recommendation engines that serve personalized content based on user preferences.
- **Fake News Detection**: The model can be extended to filter misleading or irrelevant news, contributing to the ongoing effort against misinformation.
- **Real-Time Trend Detection:** Enables media outlets and businesses to track emerging topics, popular stories, and public sentiment in rapidly changing news environments.

VIII. PROJECT TIMELINE

Table III presents the project timeline covering all major milestones from initial planning to final submission, structured to ensure each phase is completed methodically and on time.

TABLE III. TIMELINE FOR THE NEWS CLASSIFICATION PROJECT

Date	Activity / Task
Feb 2025	Week 1–2: Project Planning & Research - Define project scope and objectives. - Literature review and background research on existing models. - Identify and gather datasets for news classification.
Feb 2025	Week 3–4: Data Collection and Preprocessing - Collect and organize a diverse dataset of news articles. - Preprocess data: tokenization, stopwords removal, and stemming. - Explore feature extraction methods (TF-IDF, Word2Vec).
March 2025	Week 1: Model Development and Selection - Implement traditional machine learning models (Naive Bayes, SVM, KNN). - Begin training models on the prepared dataset.
March 2025	Week 2: Model Evaluation and Tuning - Evaluate models using accuracy, precision, recall, and F1-score. - Fine-tune hyperparameters using grid search and cross-validation. - Compare performance of traditional vs. deep learning models.
March 2025	Week 3: Real-Time Classification & Optimization - Work on optimizing models for real-time news classification. - Test scalability and performance under large datasets.
March 2025	Week 4 / March 12, 2025: Final Evaluation & Reporting - Final evaluation of model performance and validation. - Prepare results for the report and conclude the research. - Project Submission: Submit the final report and presentation.

IX. CONCLUSION

This paper presented a comparative study of Naive Bayes [18], Logistic Regression [22], and LSTM [6] approaches for news classification, using TF-IDF [14] feature extraction. Experiments on a labeled multi-category news dataset demonstrated that the LSTM model achieves the best performance (91% accuracy, 0.92 precision, 0.91 F1-score), outperforming Logistic Regression (87%) and Naive Bayes (85%), consistent with findings reported in the literature [3], [11]. While LSTM requires greater computational resources, the accuracy gains justify its use in applications where precision is paramount. Logistic Regression offers a strong balance between performance and speed for lightweight deployments.

Future work will focus on incorporating transformer-based models such as BERT [7] and ELMo [20] for improved contextual understanding, as well as exploring model interpretability techniques to make deep learning systems more transparent. Lightweight sentiment-aware embeddings [17] and efficient neural architectures [21] may further improve real-time deployment. Optimization for large-scale, real-time classification at scale—drawing on scalability insights from works such as [9] and [16]—remains a key direction for practical applications in news aggregation, content curation, and misinformation detection.

REFERENCES

1. F. Sebastiani, "Machine learning in automated text categorization," ACM Computing Surveys (CSUR), vol. 34, no. 1, pp. 1–47, 2002.
2. T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in Proc. European Conf. Machine Learning (ECML), pp. 137–142, 1998.
3. Y. Goldberg, "A primer on neural network models for natural language processing," J. Artificial Intelligence Research, vol. 57, pp. 345–420, 2016.
4. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in Neural Information Processing Systems (NIPS), vol. 26, pp. 3111–3119, 2013.
5. J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in Proc. 2014 Conf. Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543, 2014.
6. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
7. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, pp. 4171–4186, 2019.
8. Y. Kim, "Convolutional neural networks for sentence classification," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), pp. 1746–1751, 2014.
9. A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in Proc. 15th Conf. European Chapter of the Association for Computational Linguistics (EACL), vol. 2, pp. 427–431, 2017.
10. Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in Proc. NAACL-HLT, pp. 1480–1489, 2016.
11. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA: MIT Press, 2016.
12. C. C. Aggarwal and C. X. Zhai, "A survey of text classification algorithms," in Mining Text Data, C. C. Aggarwal and C. X. Zhai, Eds. New York: Springer, 2012, pp. 163–222.
13. A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems (NIPS), vol. 30, pp. 5998–6008, 2017.
14. K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," Journal of Documentation, vol. 28, no. 1, pp. 11–21, 1972.
15. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.
16. X. Liu, K. Duh, L. Liu, and J. Gao, "Very deep transformers for neural machine translation," arXiv preprint arXiv:2008.07772, 2020.
17. A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in Proc. 49th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 142–150, 2011.
18. T. Bayes, "An essay towards solving a problem in the doctrine of chances," Philosophical Transactions of the Royal Society of London, vol. 53, pp. 370–418, 1763.
19. K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), pp. 1724–1734, 2014.
20. M. E. Peters et al., "Deep contextualized word representations," in Proc. NAACL-HLT, pp. 2227–2237, 2018.
21. A. Howard et al., "Searching for MobileNetV3," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), pp. 1314–1324, 2019.
22. V. Vapnik, The Nature of Statistical Learning Theory. New York: Springer, 1995.