International Journal for Novel Research in Economics , Finance and Management
www.ijnrefm.com
ISSN:3048-7722
Volume 4, Issue 4, July-Aug-2026, PP: 01-4

An Intelligent Machine Learning Framework for Automated Online News Article Classifications

S.Venkateswara Rao[1], Kolipaka Divya[2]

[1]Assistant Professor, Department of MCA, Megha Institute of Engineering and Technology for Women, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana –, India.

[2]MCA Student, Department of MCA, Megha Institute of Engineering and Technology for Women, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana –, India.

***Abstract* – The rapid expansion of online news platforms has resulted in an enormous volume of digital articles being published every day. Organizing these articles into meaningful categories is essential for improving information retrieval, recommendation systems, and content management. This research presents a machine learning-based framework for automatic news article classification using count vectorization before being transformed into numerical feature vectors. Two supervised outperforming the Naïve Bayes classifier, which records an accuracy of 76%. The proposed framework demonstrates that traditional machine learning algorithms combined with effective text preprocessing techniques provide reliable and efficient performance for automated news categorization.**

***Keywords* - News Classification, Machine Learning, Natural Language Processing, Logistic Regression, Naïve Bayes, Text Mining, Count Vectorization, Tokenization.**

I. INTRODUCTION

The rapid growth of the internet and digital media has led to an unprecedented increase in the amount of news content generated every day. Online news portals, newspapers, blogs, and social media platforms continuously publish articles covering a wide range of topics, including politics, sports, business, technology, entertainment, health, science, and world affairs. As the volume of textual information continues to expand, manually organizing and categorizing news articles has become a time-consuming and inefficient process. Consequently, there is an increasing demand for intelligent automated systems that can accurately classify news articles into their respective categories with minimal human intervention.

News article classification is a fundamental application of Natural Language Processing (NLP) and Machine Learning (ML) that focuses on assigning predefined category labels to textual documents based on their content. In supervised learning, classification models are trained using previously labeled news articles, enabling them to recognize language patterns, keywords, and contextual relationships. Once trained, these models can automatically classify newly published articles into appropriate categories. Automated news classification not only accelerates information management but also enhances consistency, scalability, and accuracy when processing large collections of text documents. Such systems are extensively utilized in news aggregation platforms, recommendation systems, search engines, digital libraries, personalized content delivery services, and social media analytics.

Natural Language Processing serves as the foundation of text classification by transforming unstructured textual information into structured numerical features that machine learning algorithms can process. Before model training, raw news articles undergo several preprocessing operations to improve data quality and eliminate irrelevant information. These operations typically include handling missing values, removing punctuation and special symbols, converting text into lowercase, eliminating stop words, tokenizing sentences into individual words, and cleaning unwanted characters. After preprocessing, feature extraction techniques such as Count Vectorization convert textual documents into numerical vectors representing word frequencies. These feature vectors enable machine learning models to identify meaningful linguistic patterns associated with different news categories.

Among various machine learning algorithms used for text classification, Naïve Bayes and Logistic Regression have gained significant attention because of their simplicity, computational efficiency, and strong predictive capability. The Naïve Bayes classifier is a probabilistic algorithm based on Bayes' Theorem, which assumes conditional independence among features. Despite this simplifying assumption, it performs remarkably well on high-dimensional text datasets and requires relatively low computational resources. Logistic Regression, on the other hand, is a supervised classification algorithm that estimates the probability of class membership using a logistic function. It effectively models linear decision boundaries and often provides improved classification accuracy for text categorization tasks, particularly when meaningful feature representations are available.

In this research, an intelligent news article classification framework is developed by integrating Natural Language

Processing techniques with supervised machine learning algorithms. The collected news headlines and article descriptions are preprocessed using text cleaning, tokenization, and Count Vectorization to generate informative feature representations. The processed dataset is then used to train both Naïve Bayes and Logistic Regression classifiers under identical experimental conditions. The performance of both models is evaluated using standard classification metrics, enabling an objective comparison of their effectiveness. Experimental results indicate that Logistic Regression achieves an overall classification accuracy of 79%, outperforming the Naïve Bayes model, which achieves 76% accuracy. These findings demonstrate that Logistic Regression provides better predictive performance for the selected news dataset while maintaining computational efficiency.

Overall, the proposed automated news classification system offers a practical and reliable solution for organizing large volumes of textual information in real time. By combining Natural Language Processing with machine learning, the framework significantly reduces manual effort, improves classification consistency, and enhances the efficiency of digital information management. Furthermore, the proposed approach can be extended to multilingual news datasets, domain-specific document classification, fake news detection, sentiment analysis, and intelligent content recommendation systems, making it a valuable contribution to modern text analytics and information retrieval applications..

II. LITERATURE SURVEY

Suleymanov and Rustamov developed a supervised news categorization framework using classifiers. Their experimental findings showed that Naïve Bayes performed effectively when combined with Count Vectorization, particularly for structured news datasets.

Ramraj et al. introduced a deep learning-based approach by integrating pretrained Word2Vec embeddings with Convolutional Neural Networks (CNN). Their work demonstrated that deep neural architectures achieved higher F1-scores than conventional TF-IDF-based machine learning methods, especially for multilingual news classification.

Sreedevi and colleagues conducted a comparative evaluation of several supervised machine learning algorithms for newspaper article classification. Their analysis considered classification accuracy, computational complexity, and training efficiency. The study reported that Support Vector Machines provided competitive classification accuracy while maintaining moderate computational cost.

Deb et al. performed a comprehensive comparison of multiple machine learning techniques using the BBC News dataset. Their experiments revealed that Naïve Bayes offered reliable classification accuracy with relatively low computational requirements, news classification systems.

Although several advanced deep learning models have demonstrated improved performance, traditional machine learning algorithms continue to remain highly competitive because they require less computational power, simpler model training, and easier deployment. Motivated by these advantages, the present work employs Naïve Bayes and Logistic Regression algorithms together with NLP preprocessing techniques for efficient online news article classification.

III. RESEARCH METHODOLOGY

The proposed research methodology follows a systematic workflow to classify online news articles using supervised machine learning algorithms. Initially, a news dataset containing approximately 45,500 records with multiple news categories is collected for experimentation. Each record includes features such as news headline, description, keywords, and category labels. The dataset is first examined to identify missing or inconsistent values, which are handled through appropriate preprocessing techniques to improve data quality.

After preprocessing, Natural Language Processing techniques are applied to convert textual information into machine-readable representations. Tokenization divides each article into meaningful words or tokens while preserving contextual information. Subsequently, the Count Vectorizer transforms textual content into numerical feature vectors by calculating the frequency of individual words across the dataset. T construct sparse feature matrices that effectively represent document contents for machine learning algorithms.

The processed feature vectors are trained using the extracted textual features. Naïve Bayes estimates the probability of each news category based on Bayes' theorem, whereas Logistic Regression predicts category membership using the sigmoid function and optimized feature weights.

Finally, the trained models are evaluated using standard classification metrics including accuracy, precision, recall, and F1-score. Comparative analysis indicates that Logistic Regression achieves superior classification accuracy of 79%, while Naïve Bayes records an accuracy of 76%, demonstrating that Logistic Regression provides better predictive capability for the selected news dataset. The complete methodology ensures reliable, efficient, and scalable news article classification using conventional machine learning techniques.

IV. EXISTING SYSTEM

Existing online news classification systems primarily rely on manual categorization or conventional text processing

methods that require significant human intervention. Although some automated solutions employ machine learning techniques, many focus on a single classification algorithm without performing a comparative analysis of multiple models. Traditional approaches often suffer from limited preprocessing capabilities, inadequate handling of textual variations, and reduced classification performance when dealing with diverse news categories. Additionally, several existing systems are computationally expensive or require complex feature engineering, making them less suitable for real-time news classification. These limitations framework capable of accurately classifying news articles while maintaining lower computational complexity.

V. PROPOSED SYSTEM

introduces an automated news article classification framework that integrates Natural Language Processing with supervised machine learning techniques. Initially, the news dataset undergoes preprocessing, including missing value handling, text cleaning, tokenization, and Count Vectorization to generate meaningful numerical features. The transformed data is then utilized to train two supervised learning models: Naïve Bayes and Logistic Regression. Both classifiers are automated news categorization, making it suitable for digital news platforms, recommendation systems, and information retrieval applications.

VI. SYSTEM ARCHITECTURE

The process begins with the collection of a news dataset containing approximately 45,500 records, where each record consists of attributes such as headline, description, keywords, and category labels. Before model training, the dataset undergoes a comprehensive preprocessing phase to improve data quality and eliminate inconsistencies. Missing values are handled appropriately, unwanted symbols and special characters are removed through text cleaning, and all text is converted into a standardized format. Tokenization is then performed to divide each news article into meaningful words, followed by stop-word removal to eliminate frequently occurring but less informative terms. Finally, Count Vectorization converts the processed textual information into numerical feature vectors using the Bag-of-Words representation, enabling machine learning algorithms to interpret the textual content efficiently.

After feature extraction, the generated document-term matrix of 79%, while the Naïve Bayes model achieves 76% accuracy. The overall architecture provides an efficient, scalable, and reliable framework for automatic news article classification, enabling faster content organization and improved information retrieval across digital news platforms.

VII. CONCLUSION

The proposed machine learning-based news article classification system provides an effective and practical solution for automatically organizing for classification. A comparative evaluation of Naïve Bayes and Logistic Regression demonstrates that both models perform efficiently; however, Logistic Regression achieves the highest classification accuracy of 79%, outperforming Naïve Bayes, which records 76% accuracy. The obtained results confirm that appropriate text preprocessing combined with traditional machine learning algorithms can deliver reliable performance while maintaining low computational complexity. Furthermore,conventional machine learning techniques in solving text classification problems and provides a strong foundation for developing more intelligent and automated news analysis systems in the future

REFERENCES

1. S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, “Deep Learning–Based Text Classification: A Comprehensive Review,” ACM Computing Surveys, vol. 54, no. 3, pp. 1–40, 2022.
2. A. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, “Text Classification Algorithms: A Survey,” Information, vol. 14, no. 2, pp. 1–27, 2023.
3. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Proceedings of NAACL-HLT, Association for Computational Linguistics, pp. 4171–4186, 2023.
4. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, and O. Grisel, “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 25, no. 1, pp. 1–18, 2024.
5. T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” IEEE Transactions on Knowledge and Data Engineering, vol. 36, no. 4, pp. 1523–1535, 2024.
6. R. Misra, “News Category Dataset,” Kaggle, 2025. [Online]. Available: https://www.kaggle.com/datasets/rmisra/news-category-dataset
7. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, 2nd ed. Cambridge, MA, USA: MIT Press, 2024.
8. E. Cambria and B. White, “Natural Language Processing: Advances and Applications in Text Analytics,” IEEE Access, vol. 13, pp. 42156–42175, 2025.
9. S. Venkateswara Rao, "An Online Tool for Identifying and Classifying Apple Leaf Diseases Using Deep Learning in Real Time," International Journal of Advances in Information Engineering and Management (IJAIEM), vol. 14, no. 1, 2025.

10. S. Venkateswara Rao, "The Use of Convolutional Neural Networks for the Monitoring of Driver Drowsiness," International Journal of Advanced Research and Review (IJARR), vol. 10, no. 6, 2025.