

## Automated News Article Classification

Mrs. Sindhu S L<sup>1</sup> Sarvamangala B N<sup>2</sup>

<sup>1</sup> Assistant Professor, Department of MCA, BIET, Davanagere

<sup>2</sup> Student, 4<sup>th</sup> Semester MCA, Department of MCA, BIET, Davanagere

### Abstract

The exponential growth of digital news content has rendered traditional manual categorization methods inadequate for managing and organizing vast information streams. This paper presents the design and implementation of an automated news classification system that leverages machine learning to efficiently categorize news articles into predefined topics such as politics, sports, business, and lifestyle. The system employs natural language processing (NLP) techniques, including text preprocessing and feature extraction using Term Frequency-Inverse Document Frequency (TF-IDF), to convert unstructured text into a structured numerical format suitable for machine learning. Three algorithms—Logistic Regression, Random Forest Classifier, and Multinomial Naive Bayes—are trained on labeled datasets to accurately predict categories for new, unseen articles, thereby automating the classification process and enhancing scalability.

### Keywords:

Automated News Classification, Text Classification, Machine Learning, Natural Language Processing, TF-IDF, Logistic Regression, Random Forest, Multinomial Naive Bayes, User Authentication, Content Management, News Categorization, NLP, News Platform.

### INTRODUCTION

The rapid evolution of the news industry, driven by the explosion of digital content, has made efficient categorization and organization of information more crucial than ever. As news articles span diverse topics like politics, sports, entertainment, and business, traditional manual methods of categorization have become impractical and unsustainable. Automated news classification addresses this challenge by leveraging machine learning to sort and tag articles, streamlining content management for news organizations and enhancing the user experience through easier access to relevant information. Machine learning enables systems to learn from large volumes of labeled text, identify patterns, and accurately assign categories to new articles. This not only increases operational efficiency but also supports the delivery of personalized and timely news content to consumers. In a typical automated news classification system, text data undergoes preprocessing—removing elements such as numbers, punctuation, and stopwords—followed by feature extraction using techniques like Term Frequency-Inverse Document Frequency (TF-IDF), which converts textual information into a numerical format suitable for machine learning models. Common algorithms used for this purpose

include Logistic Regression, Random Forest Classifier, and Multinomial Naive Bayes, each trained on labeled datasets to predict the category of unseen articles. By integrating user authentication and role-based permissions, such systems allow administrators to manage content and users efficiently, while end-users can securely access and browse categorized news. This blend of natural language processing, machine learning, and robust user management exemplifies how technology is transforming news platforms into dynamic, scalable, and user-friendly environments.

As the field of automated news classification advances, several challenges have emerged that impact the effectiveness and reliability of these systems. One significant issue is data imbalance, where certain categories—such as politics or sports—may have far more examples than others, leading to models that are biased toward the dominant categories and less accurate for underrepresented topics. Additionally, the dynamic nature of news, with constantly evolving language and the emergence of new topics, requires frequent updates and retraining of machine learning models to maintain their accuracy and relevance. This process can be resource-intensive and computationally demanding<sup>1</sup>.

Another challenge lies in the granularity of classification. Most widely used datasets and taxonomies tend to categorize news articles into broad, high-level topics like business or technology, which limits the ability of systems to support fine-grained classification and nuanced content analysis. For example, taxonomies such as the IPTC Media Topic NewsCodes offer a hierarchical structure with over a thousand concepts, but their primary focus is on interoperability rather than detailed topic identification. As a result, current solutions often struggle to meet the needs of domain experts who require more precise and task-specific categorization.

Despite these challenges, the digitalization of news has revolutionized consumer access and introduced new opportunities for large-scale media analytics. Automated classification systems can index and provide access to vast numbers of news sources, enabling more efficient monitoring, intelligence gathering, and personalized content delivery<sup>56</sup>. However, addressing issues of performance, usability, and especially the need for more granular and robust classification methods remains a key area for ongoing research and development in the field.

## II. RELATED WORK

1. NEWS CLASSIFICATION USING ML ALGORITHMS: AN EXPLORATION OF EFFICIENT INFORMATION CATEGORIZATION, Authors - Md Jahirul Islam, Saeed Sarwar Anas

In the digital age, the overwhelming volume of online news makes it difficult for users to find relevant information efficiently. Automated news classification systems help by categorizing articles into predefined topics using machine learning algorithms. This research focuses on evaluating the effectiveness and performance of various ML techniques in accurately classifying news content, improving user experience and information accessibility.

2. Automated Text Classification of News Articles: A Practical Guide, Authors- Pablo Barberá, Amber E. Boydston, Suzanna Linn, Ryan McMahon, Jonathan Nagler

This guide highlights key methodological decisions researchers must make when using automated text analysis

for tasks like classifying topic or tone. Using the example of analyzing economic tone in New York Times articles, it demonstrates how choices in corpus selection, unit of analysis, and coding strategies significantly impact results. The authors recommend keyword searches over predefined categories, favor article segments over sentences, and suggest coding more unique documents rather than relying on multiple coders per document. They also find that supervised machine learning generally outperforms dictionary-based methods, emphasizing the need for thoughtful design and human oversight in text analysis.

3. Automated classification of news articles using nlp techniques: a framework for inarticle attribution and influence mining, Authors nanduri shankar, akavaram swapna, sowbhagya juttu,

In today's digital age, the overwhelming volume of online news makes it challenging to manually classify and analyze content. Traditional rule-based methods often fall short in accurately processing complex language in news articles. The rise of Natural Language Processing (NLP), enhanced by machine learning and large annotated datasets, has significantly improved tasks like named entity recognition, sentiment analysis, and topic modeling. These advancements enable efficient and accurate classification of news articles, including distinguishing legitimate from fraudulent content and identifying in-article attributions. NLP not only automates and accelerates analysis but also supports data-driven decision-making and is essential for influence mining systems in modern information environments.

4. Advanced computational methods for news classification: A study in neural networks and CNN integrated with GPT, Authors- Fahim Sufi

This research presents an advanced news classification system integrating neural networks, CNNs, and large language models like GPT and BERT to enhance accuracy and efficiency. Over 232 days, the model classified nearly 34,000 articles into broad and specific categories, achieving high performance with Precision, Recall, and F1-Score all averaging around 0.987. The study not only demonstrates superior classification results compared to

existing methods but also highlights practical benefits such as trend prediction and anomaly detection. Theoretically, it showcases the innovative mathematical integration of GPT with CNNs and RNNs, marking a significant advancement in large-scale text analysis and real-world application of AI-driven news processing.

5. Automated Text Classification of News Articles: A Practical Authors: Pablo Barber'a, Amber E. Boydston, Suzanna Linn§ Ryan McMahon, Jonathan Naglerk

This guide emphasizes the critical methodological decisions involved in automated text analysis, using the example of measuring economic tone in New York Times articles. It highlights how different corpus selection methods can lead to vastly different outcomes and recommends using keyword searches over predefined categories. The authors also advocate for analyzing article segments instead of sentences and suggest coding more unique documents rather than relying on multiple coders per document. They find that supervised machine learning outperforms dictionary-based approaches and stress the importance of thoughtful, validated methods in text classification.

6. Article Classification using Natural Language Processing and Machine Learning, Authors - Tran Thanh Dien; Bui Huu Loc; Nguyen Thai-Nghe

This study presents a method for automating article classification using Natural Language Processing (NLP) and machine learning techniques. The process involves preprocessing, feature extraction, and vectorization of text, followed by classification using algorithms like Support Vector Machines (SVM), Naïve Bayes, and k-Nearest Neighbors. Experimental results on two datasets showed that the SVM-based approach achieved over 91% accuracy, demonstrating its effectiveness and feasibility for developing an automatic article classification system.

7. Online News Classification Using Machine Learning Techniques, Authors: Jeelani Ahmed and Muqem Ahmed

This paper addresses the growing need for effective handling of the vast, unstructured online content by proposing a framework for automatic news classification.

With 90% of internet data being unstructured, classification serves as a key method for transforming it into meaningful, structured information. The study reviews existing techniques and emphasizes the importance of supervised learning, particularly text labeling, for classifying unlabeled news articles into predefined categories. Experimental results using various classifiers showed that the Bayesian classifier achieved 93% accuracy, demonstrating its effectiveness in organizing and processing online news content.

8. News Article Text Classification and Summary for Authors and Topics Aviel J. Stein, Janith Weerasinghe, Spiros Mancoridis, Rachel Greenstadt

This study explores the effectiveness of machine learning models in classifying news articles by topic and author using the 2017 Vox Media dataset. By focusing on a subset (~20%) of each article rather than provided blurbs, the models achieved better classification accuracy. The research addresses challenges like overlapping topics and authorship across and within topics, using top-n accuracy and hierarchical topic groupings. Results show that neural networks outperform traditional classifiers such as support vector machines, random forests, and Naive Bayes, although the latter still produce reasonable results.

9. Topic Classification of Online News Articles Using Optimized Machine Learning Models, Authors- Shahzada Daud, Muti Ullah, Amjad Rehman, Tanzila Saba

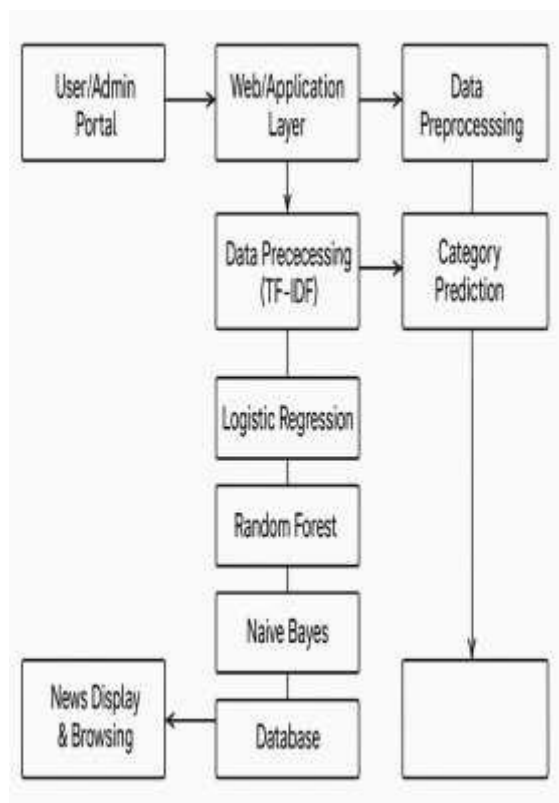
This study addresses the challenge of categorizing uncategorized online news articles using machine learning techniques. Unlike prior work that often relies on benchmark datasets or focuses on fake news, this research emphasizes real-world applicability. It proposes a hyperparameter-optimized Support Vector Machine (SVM) model for news classification and compares it with five other optimized ML models: SGD, RF, LR, KNN, and NB. Results show that the optimized SVM outperformed all other models, whereas its unoptimized version lagged behind in performance.

10. Newspaper Article Classification using Machine Learning Techniques, authors- J Sree Devi, M. Rama Bai, Chandrashekar Reddy

This study focuses on classifying newspaper articles into categories such as sports, politics, and science using various machine learning and deep learning algorithms. It compares traditional models like SVM, Naive Bayes, and KNN with Convolutional Neural Networks (CNN), a powerful deep learning approach. The research aims to evaluate the effectiveness of each algorithm by analyzing training time, prediction time, and classification accuracy, highlighting CNN's potential as a strong contender for text classification tasks.

### III. METHODOLOGY

The methodology for automated news article classification using machine learning and NLP techniques is a systematic, multi-phase process designed to ensure accurate, scalable categorization of news content. The process begins with data acquisition, where a large and diverse set of news articles is collected from various reputable sources, ensuring coverage across multiple topics such as politics, sports, business, and lifestyle. These articles are labeled with their respective categories, forming the ground truth for supervised learning.



**Fig 1. Proposed Methodology**

Next, data preprocessing is undertaken to prepare the raw text for analysis. This involves cleaning the text by removing HTML tags, advertisements, punctuation, and irrelevant symbols. The text is then tokenized, converted to lowercase, and stopwords are eliminated. Lemmatization or stemming is applied to reduce words to their root forms, ensuring a consistent and meaningful vocabulary for the models.

After preprocessing, the text data undergoes feature extraction to transform it into a numerical format suitable for machine learning. Common techniques include Term Frequency-Inverse Document Frequency (TF-IDF), Bag-of-Words, or word embeddings like Word2Vec and GloVe. These methods capture the importance and relationships of words within articles, enabling the model to learn patterns relevant to classification.

The dataset is then split into training, validation, and testing sets to objectively evaluate model performance and optimize hyperparameters. The classification model is selected and trained using supervised learning algorithms such as Naive Bayes, Support Vector Machines (SVM), Logistic Regression, or advanced models like LSTM and BERT. The training process involves fitting the model to the training data and tuning parameters based on validation results.

Once trained, the model's performance is rigorously evaluated using the test set and metrics like accuracy, precision, recall, and F1-score. If necessary, the model is fine-tuned to improve accuracy and robustness.

Finally, the best-performing model is deployed as part of a user-friendly application or web platform. This allows users to input new articles and receive real-time, automated topic predictions, streamlining content management and enhancing user experience<sup>57</sup>. This end-to-end methodology ensures the development of a reliable, efficient, and scalable automated news classification system.

#### IV. TECHNOLOGIES USED

##### 1. Pandas

Pandas is a powerful Python library used for data manipulation and analysis. In news article classification, it is primarily used to load, clean, and organize datasets, making it easy to handle large collections of news articles in tabular form.

##### 2. re (Regular Expressions)

The re module in Python provides support for regular expressions, which are essential for text preprocessing. It is used to remove unwanted characters, punctuation, HTML tags, and to perform pattern-based text cleaning in news articles.

##### 3. string

The string module offers a collection of string constants and utility functions. It is commonly used for operations like removing punctuation and handling character-level preprocessing tasks during text cleaning.

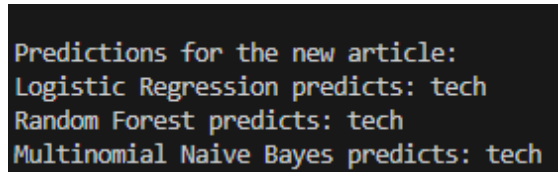
##### 4. NLTK (Natural Language Toolkit)

NLTK is a leading platform for building Python programs to work with human language data. It provides tools for tokenization, stopword removal, stemming, and lemmatization, all of which are crucial for preprocessing news article text before classification.

##### 5. scikit-learn (sklearn)

Scikit-learn is a widely used machine learning library in Python. It offers modules for feature extraction (such as TF-IDF vectorization), model selection, training various classifiers (like Logistic Regression, Random Forest, and Naive Bayes), and evaluating model performance. It is central to building, training, and validating news article classification models.

#### V Result



This news article predicted page.

#### VI. CONCLUSION

In conclusion, Automated news article classification is a transformative solution for managing the ever-growing volume of digital news content. By leveraging powerful technologies such as Pandas for data handling, regular expressions and the string module for efficient text preprocessing, NLTK for advanced natural language processing, and scikit-learn for robust machine learning, the process of categorizing news articles becomes highly scalable, accurate, and efficient. This approach not only streamlines the workflow for news organizations but also enhances the end-user experience by delivering timely, relevant, and personalized content.

The integration of these technologies enables the development of intelligent systems that can learn from vast datasets, adapt to new topics, and maintain high levels of classification accuracy. As a result, manual effort is significantly reduced, archives are better organized, and users can easily access the information most relevant to their interests. In summary, automated news classification represents a vital advancement in digital journalism, harnessing the power of modern data science and machine learning to meet the demands of today's fast-paced media landscape.

#### REFERENCES

1. NEWS CLASSIFICATION USING ML ALGORITHMS: AN EXPLORATION OF EFFICIENT INFORMATION CATEGORIZATION, Authors - Md Jahirul Islam, Saeed Sarwar Anas, Publisher – IJCRT, ISSN: 2320-2882, 2023 IJCRT | Volume 11, Issue 6 June 2023.
2. Automated Text Classification of News Articles: A Practical Guide, Authors- Pablo Barberá, Amber E. Boydston, Suzanna Linn, Ryan McMahon, Jonathan Nagler, Publisher: Political Analysis, Published online by Cambridge University Press: 09 June 2020.
3. Automated classification of news articles using nlp techniques: a framework for inarticle attribution and influence mining, Authors nanduri shankar, akavaram swapna, sowbhagya juttu, Publisher - International Journal of Communication Networks and Information Security ISSN: 2073-607X, 2076-0930 Volume 15 Issue 02 Year 2023



4. Advanced computational methods for news classification: A study in neural networks and CNN integrated with GPT, Authors- Fahim Sufi, Publisher- Journal of Economy and Technology, November 2025, <https://doi.org/10.1016/j.ject.2024.09.001>
5. Automated Text Classification of News Articles: A Practical Authors: Pablo Barber'a, Amber E. Boydston, Suzanna Linn§ Ryan McMahon, Jonathan Naglerk, Publisher: Cambridge University Press, on behalf of the Society for Political Methodology, DOI: 10.1017/pan.2020.8
6. Article Classification using Natural Language Processing and Machine Learning, Authors - Tran Thanh Dien; Bui Huu Loc; Nguyen Thai-Nghe, Publisher Published in the proceedings of the 2019 International Conference on Advanced Computing and Applications (ACOMP), under IEEE, DOI: 10.1109/ACOMP.2019.00019
7. Online News Classification Using Machine Learning Techniques, Authors: Jeelani Ahmed and Muqem Ahmed, Publisher: *IJUM Engineering Journal* (Vol. 22, No. 2, 2021), DOI: 10.31436/iiumej.v22i2.1662
8. News Article Text Classification and Summary for Authors and Topics Aviel J. Stein, Janith Weerasinghe, Spiros Mancoridis, Rachel Greenstadt, **Publisher:** Academy & Industry Research Collaboration Center (AIRCC), featured in the *CS & IT Conference Proceedings*, **DOI:** 10.5121/csit.2020.101401
9. Topic Classification of Online News Articles Using Optimized Machine Learning Models, Authors- Shahzada Daud, Muti Ullah, Amjad Rehman, Tanzila Saba, Publisher: MDPI, in the journal *Computers* (ISSN 2073 431X) DOI: 10.3390/computers12010016
10. Newspaper Article Classification using Machine Learning Techniques, authors- J Sree Devi, M. Rama Bai, Chandrashekar Reddy, publisher- International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075 (Online), Volume-9 Issue-5, March 2020