NEWS TOPIC CLASSIFICATION

[1] Yamini Chouhan, [2] Meesala Sameera, [3] Swargam Srilatha, [4] Abhishek Lingampally
[1]AssistantProfessor, [234]Students
Department of Computer Engineering(Software Engineering)
Siddhartha Institute of Technology & Sciences, Narapally
yaminichouhan_cse@siddhartha.co.in, 23tq1a5641@siddhartha.co.in, 23tq1a5620@siddhartha.co.in, 23tq1a5625@siddhartha.co.in

Abstract

The rapid expansion of digital media has led to an overwhelming volume of news articles being generated daily, making it challenging to efficiently organize, filter, and access relevant information. Automatic news topic classification has become essential for managing and structuring this vast amount of textual data. This project aims to develop a machine learning-based system that automatically classifies news articles into predefined categories based on their content.

A balanced dataset consisting of multiple news topics is used, and various Natural Language Processing (NLP) techniques are applied, including text preprocessing, tokenization, stop-word removal, and feature extraction. The processed text is transformed into numerical representations to enable effective model training. Multiple classification algorithms are evaluated, among which the Random Forest classifier demonstrates strong performance, achieving high accuracy, precision, and recall on the test dataset.

Additionally, feature importance analysis is conducted to identify the most influential words contributing to accurate classification. The developed model provides an efficient and scalable solution for automatic news categorization. It can be integrated into media platforms, search engines, and digital libraries to improve content organization, enhance user experience, and enable faster information retrieval.

I. Introduction

The rapid expansion of digital news platforms has resulted in an enormous volume of news content being generated and published every day. With thousands of articles covering diverse topics such as politics, sports, business, technology, and entertainment, it has become increasingly challenging for readers and news organizations to efficiently organize, manage, and retrieve relevant information. This information overload makes it difficult to access content quickly and accurately.

Traditionally, news articles are categorized manually or using simple keyword-based methods. However, manual classification is time-consuming, labor-intensive, and prone to human errors, especially when dealing with large-scale datasets. On the other hand, traditional keyword-based approaches often fail to capture the contextual meaning and semantic relationships within text, resulting in inaccurate classification.

To address these challenges, there is a growing need for automated and intelligent systems that can efficiently classify news articles. Machine Learning (ML) and Natural Language Processing (NLP) techniques provide powerful tools for analyzing textual data and identifying patterns within it. By leveraging these techniques, it is

possible to build models that can automatically categorize news articles based on their content.

II. Literature Survey

Existing research demonstrates that Machine Learning (ML) and Natural Language Processing (NLP) techniques are highly effective for news topic classification. Several studies have utilized popular datasets such as the BBC News dataset and various Kaggle news datasets to build and evaluate classification models. These datasets provide a diverse collection of news articles across multiple categories, making them suitable for training and testing classification algorithms.

Researchers have applied algorithms such as Naive Bayes, Support Vector Machine (SVM), Random Forest, and Logistic Regression to classify news articles into predefined categories. Among these, Naive Bayes and SVM are widely used due to their efficiency in handling high-dimensional text data, while Random Forest offers improved accuracy through ensemble learning. These models have achieved high performance by effectively learning patterns from textual features.

A key component of successful classification models is the use of text preprocessing techniques, including tokenization, stop-word removal, stemming, and TF-IDF feature extraction. These techniques help convert raw text into meaningful numerical representations, improving model accuracy and efficiency.

Recent advancements in the field have explored **deep learning models** such as Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), and transformer-based approaches, which further enhance classification performance by capturing contextual and semantic information within text.

III. System Analysis

The rapid increase in digital news content requires an efficient system capable of handling large-scale textual data. Traditional systems are not designed to process such high volumes of information in real time, leading to delays and inefficiencies in content organization. Therefore, an automated system becomes essential to manage, classify, and retrieve news articles effectively.

Another important aspect of system analysis is understanding the nature of textual data. News articles contain unstructured data, which makes it difficult to process using conventional methods. The proposed system addresses this challenge by converting unstructured text into structured numerical representations using NLP techniques, enabling better analysis and classification.

The system also focuses on improving classification accuracy by learning patterns from historical data. Unlike rule-based systems, machine learning models continuously improve as more data becomes available. This adaptability ensures that the system remains effective even when new topics or trends emerge in the news domain. Furthermore, the proposed system is designed to be scalable and efficient, allowing it to process large datasets without significant performance degradation.

Existing System

In many traditional news platforms, classification is performed manually by editors who assign categories based on their understanding of the content. While this approach can be accurate for small volumes of data, it becomes impractical as the number of news articles increases. Manual classification also introduces inconsistencies, as different individuals may interpret the same content differently.

Some systems rely on basic keyword-matching techniques, where predefined words are associated with specific categories. However, this method lacks the ability to understand context, synonyms, or semantic meaning. For example, an article about "stock market growth" may not be correctly classified if the exact keyword is missing, even though it clearly belongs to the business category.

Disadvantages of Existing System

· Time-consuming and labor-intensive manual process
· High chances of human error
· Inaccurate classification due to lack of context understanding
· Inefficient for large-scale data
· Limited adaptability to new topics or patterns
· Poor performance with ambiguous or complex text

Proposed System

The proposed system introduces an automated and intelligent approach to news classification using **Machine Learning (ML)** and **Natural Language Processing (NLP)** techniques. Unlike traditional methods, the system is capable of understanding the context and semantic meaning of text, enabling more accurate categorization of news articles. It processes large volumes of data efficiently and reduces the need for manual intervention.

The system begins with **data preprocessing**, where raw text is cleaned by removing punctuation, stop words, and irrelevant characters. Tokenization and normalization techniques are applied to convert the text into a standardized format. This step ensures that the input data is consistent and suitable for further processing.

Following preprocessing, **feature extraction** is performed using techniques such as **TF-IDF**, which converts textual data into numerical vectors by assigning importance to words based on their frequency and relevance. This allows the model to focus on meaningful terms that contribute to classification.

Advantages of Proposed System

· High accuracy in classification
· Automated and time-efficient process
· Handles large volumes of data effectively
· Better understanding of text context
· Scalable and adaptable to new data
· Improves content organization and retrieval

IV. Methodology

The system follows a structured machine learning pipeline for automatic news topic classification. Initially, data collection is performed by loading a dataset containing news articles along with their respective topic labels such as politics, sports, business, technology, and entertainment.

Next, Exploratory Data Analysis (EDA) is conducted to understand the distribution of categories and basic characteristics of the text data. This helps in identifying patterns and potential issues in the dataset.

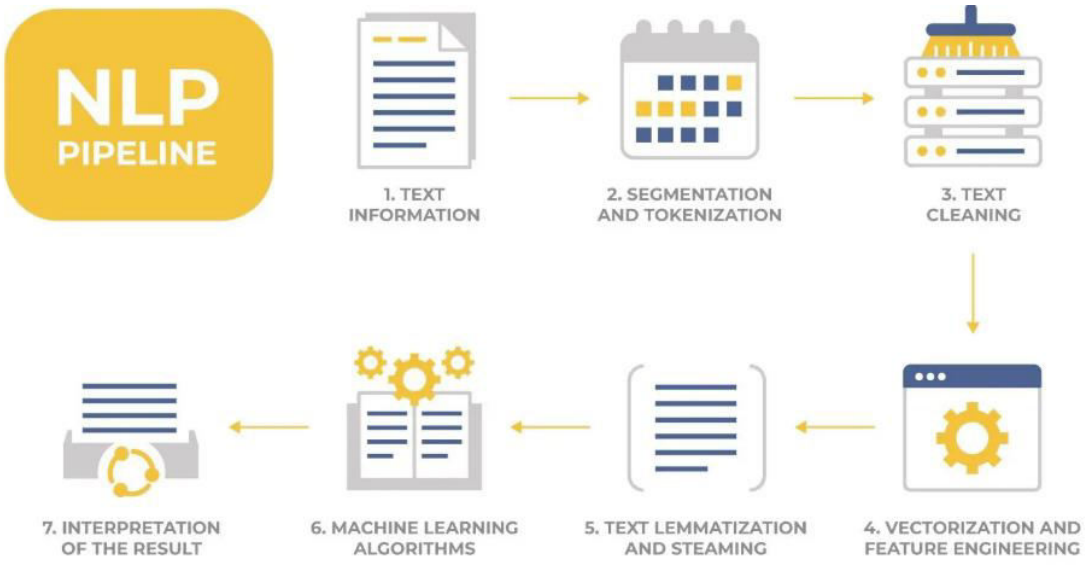
In the preprocessing stage, the text is cleaned by removing punctuation, special characters, and stop words. The cleaned text is then converted into numerical form using techniques such as TF-IDF or Count Vectorizer. The dataset is divided into input features (X) and target labels (y).

During model training, the dataset is split into training (80%) and testing (20%) sets. Machine learning algorithms such as Naive Bayes, Random Forest, and Logistic Regression are trained on the data to learn classification patterns.

The models are then evaluated using metrics such as accuracy, precision, recall, and confusion matrix to assess performance.

System Architecture

· Input Layer: News articles dataset
· EDA Module: Analyze data distribution and patterns
· Preprocessing Module: Cleaning, tokenization, TF-IDF
· Model Training: ML algorithms (NB, RF, LR)
· Evaluation Module: Accuracy, precision, recall
· Prediction Module: Classifies new articles
· Output Layer: Displays predicted topic



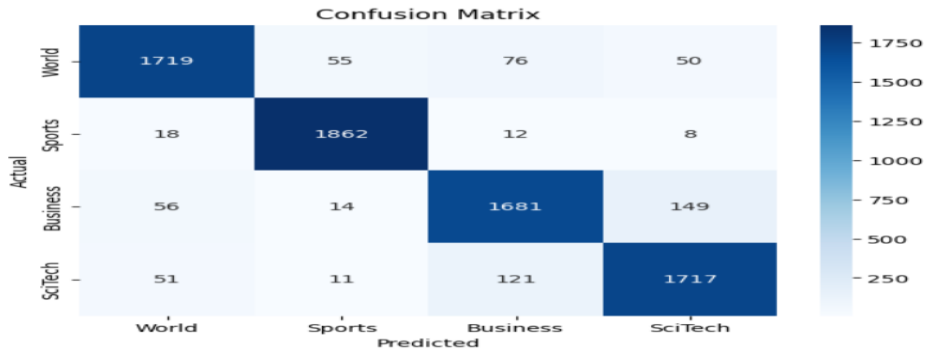
V. Result and Output

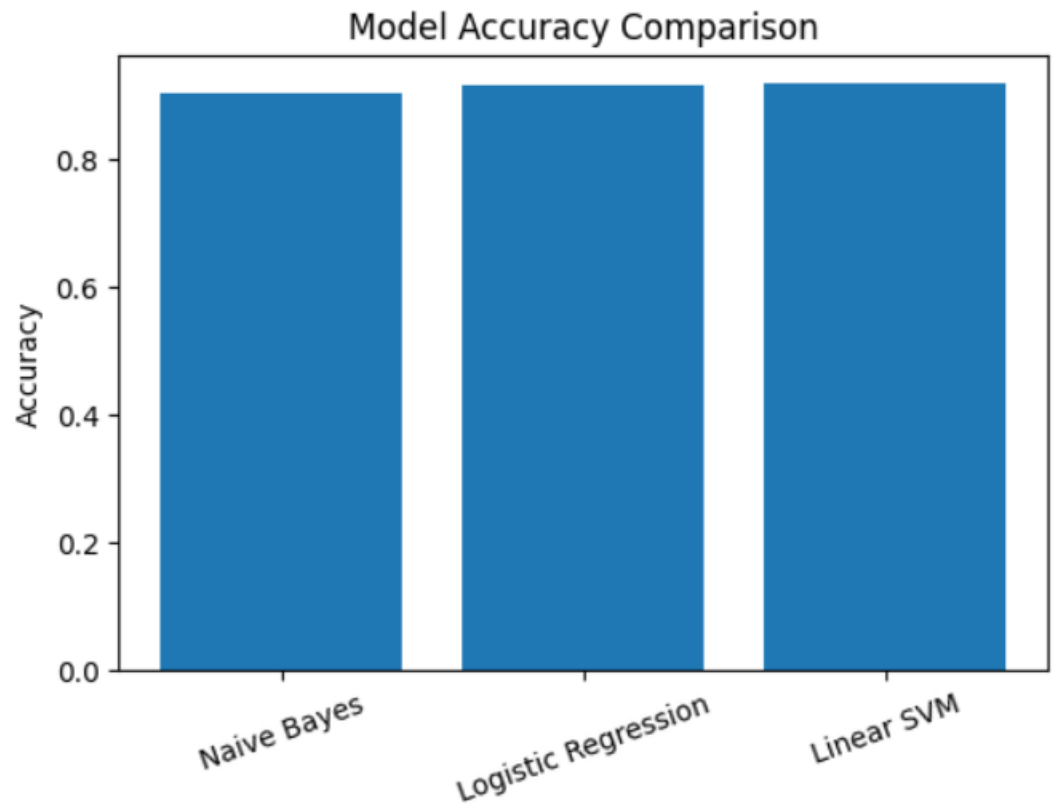
```
Train size: (120000, 4)
Test size: (7600, 4)
```

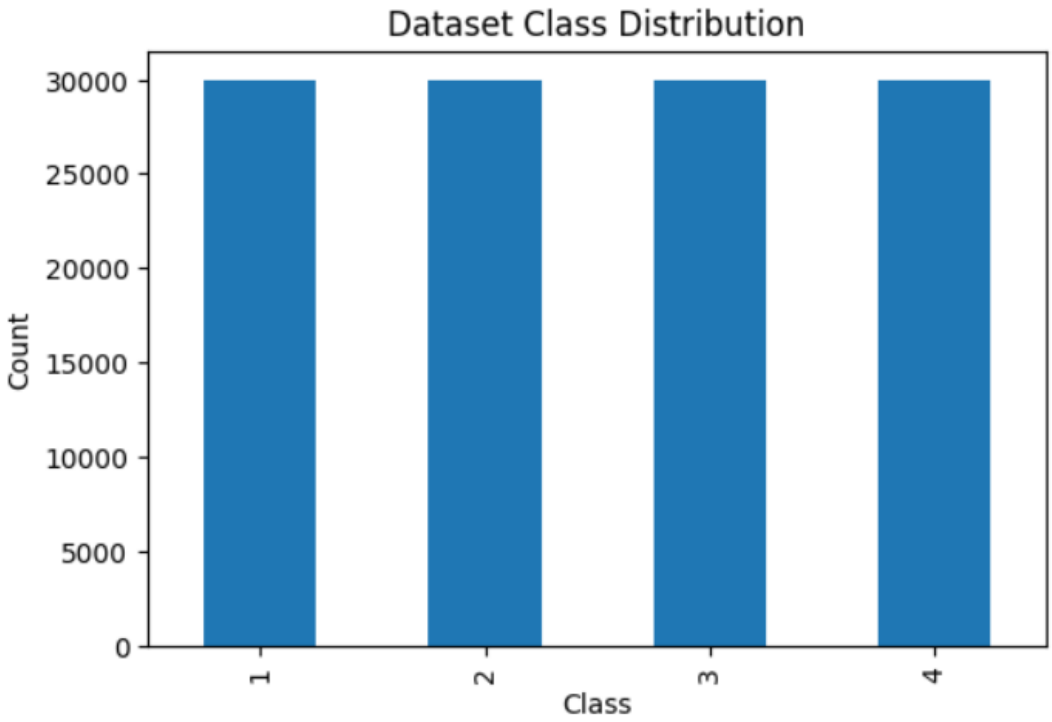
	Class Index	Title	Description	text
0	3	Wall St. Bears Claw Back Into the Black (Reuters)	Reuters - Short-sellers, Wall Street's dwindli...	Wall St. Bears Claw Back Into the Black (Reute...
1	3	Carlyle Looks Toward Commercial Aerospace (Reu...	Reuters - Private investment firm Carlyle Grou...	Carlyle Looks Toward Commercial Aerospace (Reu...
2	3	Oil and Economy Cloud Stocks' Outlook (Reuters)	Reuters - Soaring crude prices plus worries\ab...	Oil and Economy Cloud Stocks' Outlook (Reuters...
3	3	Iraq Halts Oil Exports from Main Southern Pipe...	Reuters - Authorities have halted oil export\f...	Iraq Halts Oil Exports from Main Southern Pipe...
4	3	Oil prices soar to all-time record, posing new...	AFP - Tearaway world oil prices, toppling reco...	Oil prices soar to all-time record, posing new...

Class	Category	Precision	Recall	F1-Score	Support
1	World	0.93	0.90	0.92	1900
2	Sports	0.96	0.98	0.97	1900
3	Business	0.89	0.88	0.89	1900
4	Sci/Tech	0.89	0.90	0.90	1900
Overall Accuracy	—	—	—	**0.92**	7600

Average Type	Precision	Recall	F1-Score	Support
Macro Average	0.92	0.92	0.92	7600
Weighted Average	0.92	0.92	0.92	7600




Model Accuracy Comparison
Accuracy
0.0
0.2
0.4
0.6
0.8
Naive Bayes
Logistic Regression
Linear SVM


Dataset Class Distribution
Count
0
5000
10000
15000
20000
25000
30000
1
2
3
4
Class

VI. Conclusion

This project successfully developed a machine learning-based system for automatic news topic classification using a labeled dataset of news articles. The model effectively categorizes news into various topics such as politics, sports, business, and technology, reducing the need for manual classification.

By applying text preprocessing techniques and machine learning algorithms, the system achieves improved accuracy and efficiency in handling large volumes of textual data. The use of Natural Language Processing (NLP) enables the model to understand patterns and extract meaningful features from news content.

Overall, the project demonstrates the significant potential of machine learning and NLP in managing and organizing digital news data. The developed system can assist news platforms in better content organization, faster information retrieval, and delivering more relevant and personalized information to users.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring
COVID-19 through intensive investigation with supervised machine learning algorithm. In
Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.
[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International
Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer
Nature Switzerland.
[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on
colorectal cancer with SMOTE on machine learning Algorithm
[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve
1Professor, Department of computer Science & engineering, Anurag University, TS, India.
2Student, Department of computer Science & engineering, Anurag University, TS, India.
[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, “Real-Time
Object Detection in Drone Surveillance Using YOLOv5,” in Proc. 2025 3rd Int. Conf. IoT,
Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6,
doi: 10.1109/ICICAT68430.2025.11414670.
[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, “Contextual Analysis
of Big Data Analytics in Intelligent Transportation Frameworks,” in Intersection of Artificial

Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications
in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer,
2025, doi: 10.1007/978-3-031-88304-0_79.
[7] R. D. Kumar, V. N. S.Manaswini, "Applications of blockchain in smart cities: detecting
fake documents from land records using blockchain technology," in Blockchain for Smart
Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.
[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie
recommender using collaborative filtering and sentiment analysis," *International Research
Journal of Modernization in Engineering Technology and Science*, vol. 7, no. 7, July 2025,
doi: 10.56726/IRJMETS81618.
[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system
using LightGBM and decision tree machine learning models," *Journal of Machine and
Computing*, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.
[10] Ravi Kumar Banoth, **Dr. B.V. Ramana Murthy**, "Smart agriculture through IoT and
machine learning for analyzing carbon footprints," in *Proc. Int. Conf. Computer Science and
Communication Engineering (ICCSCE)*, Apr. 2025.**[11]** Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer
learning approach: MobileNetV2 with CNN," *SN Computer Science*, vol. 5, art. no. 199,
2024, doi: 10.1007/s42979-023-02500-x.