REDUCING DATA VOLUME IN NEWS CLASSIFICATION DEEP LEARNING FRAMEWORK AND DATABASE

[1] S Sudarshan,[2] Narasimhan Nitya Priya, [3] P Anvesh,[4] P Chakradhar Reddy, [5] Mahesh Balaji Warle
[1]AssistantProfessor, [2345]Students
Department of Computer Science and Technology
Siddhartha Institute of Technology & Sciences, Narapally

sudarshan_cse@siddhartha.co.in, 24TQ1A05G4@siddhartha.co.in, 24TQ1A05H1@siddhartha.co.in, 24TQ1A05G9@siddhartha.co.in, 24TQ1A05H6@siddhartha.co.in

Abstract

This project focuses on reducing data volume in news classification using a deep learning framework integrated with an efficient database system. With the rapid growth of digital news content, handling large datasets becomes computationally expensive and time-consuming. The proposed system aims to optimize data by applying preprocessing techniques such as text summarization, feature extraction, and dimensionality reduction, ensuring only relevant information is retained. A deep learning model, such as a neural network, is then trained on this refined dataset to accurately classify news into categories like politics, sports, and technology. The database component is designed to store, manage, and retrieve processed data efficiently. This approach improves classification performance, reduces storage requirements, and enhances processing speed, making it suitable for real-time and large-scale news analysis applications.

Keywords:
News Classification, Data Reduction, Deep Learning, Natural Language Processing (NLP), Text Preprocessing, Feature Extraction, Dimensionality Reduction, Neural Networks, Database Management, Big Data Analytics, Text Summarization, Machine Learning

I. Introduction

The rapid growth of digital media has led to an overwhelming increase in the volume of news articles generated every day. This surge in information creates challenges in organizing, processing, and analyzing data efficiently. News classification, a key task in Natural Language Processing (NLP), helps in categorizing articles into predefined topics such as politics, sports, technology, and entertainment. However, handling large-scale datasets often requires significant computational resources, making it necessary to develop methods that can reduce data volume while maintaining accuracy. To address this challenge, data reduction techniques play a crucial role in improving system efficiency. Methods such as text preprocessing, feature selection, and dimensionality reduction help in eliminating redundant and irrelevant information from raw news data. By focusing only on essential features, these techniques not only decrease storage requirements but also enhance the performance of classification models. This streamlined data is more suitable for training deep learning models, enabling faster processing and improved prediction accuracy.

In this project, a deep learning framework is integrated with an efficient database system to achieve optimized news classification. The deep learning model, such as a neural network, is trained on the reduced dataset to automatically learn patterns and

categorize news content effectively. Meanwhile, the database ensures organized storage and quick retrieval of processed data. This combined approach results in a scalable, efficient, and accurate system capable of handling large volumes of news data in real-time applications.

In the modern digital era, an enormous amount of news content is generated every day through online platforms, social media, and news agencies. This rapid growth of textual data has made it increasingly difficult to organize, manage, and analyze information efficiently. News classification, which involves categorizing news articles into predefined topics such as politics, sports, technology, and entertainment, has become an essential task in Natural Language Processing (NLP). It plays a crucial role in applications like news recommendation systems, content filtering, and personalized information retrieval. However, the large volume of data poses significant challenges in terms of storage, processing time, and computational complexity.

Traditional approaches to news classification, such as Naive Bayes and Support Vector Machines (SVM), have been widely used due to their simplicity and effectiveness on smaller datasets. However, these methods often struggle to handle large-scale data and fail to capture the contextual and semantic relationships present in natural language. With the advancement of artificial intelligence, deep learning techniques such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Transformer-based models like BERT have gained popularity. These models are capable of understanding complex patterns in textual data and provide significantly improved accuracy in classification tasks. Despite their advantages, deep learning models require large amounts of data and computational resources, which can be a limitation when dealing with massive datasets.

II. Literature Survey

Kowsari et al. [1] presented a comprehensive survey on text classification techniques using both traditional machine learning and deep learning models. The study highlights the limitations of classical approaches when handling large-scale text data. It emphasizes the superiority of deep learning models in feature extraction and accuracy. The authors also discuss scalability challenges in big data environments. This work supports our idea by providing a strong foundation for applying deep learning to news classification.

Kim [2] proposed the use of Convolutional Neural Networks (CNN) for sentence classification tasks. The model automatically extracts meaningful features from raw text without requiring manual feature engineering. It demonstrates improved accuracy compared to traditional approaches. CNN also reduces the complexity of high-dimensional data. This directly relates to our idea by improving efficiency and reducing data processing effort.

Liu et al. [3] introduced Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) models for handling sequential text data. These models effectively capture contextual relationships within sentences. Their approach improves classification accuracy for large textual datasets. It also reduces the need for extensive

preprocessing. This supports our idea by enabling better understanding of news content with optimized data.

Devlin et al. [4] developed BERT, a transformer-based model that captures deep contextual relationships in text. It significantly improves performance in various NLP tasks, including classification. The model reduces dependency on manual feature extraction. It works efficiently even with complex datasets. This is highly relevant to our idea for extracting meaningful features from reduced data.

Sebastiani [5] discussed various challenges in text classification, particularly issues related to high-dimensional feature spaces. The study highlights the importance of feature selection and dimensionality reduction. It also explains how redundant data affects model performance. The work emphasizes efficient preprocessing techniques. This supports our idea of reducing data volume before classification.

Jolliffe [6] introduced Principal Component Analysis (PCA) as a method for dimensionality reduction. PCA transforms large datasets into smaller sets of uncorrelated variables. It helps in reducing computational complexity. The method preserves important information while removing redundancy. This aligns with our goal of optimizing data size for efficient processing.

Mikolov et al. [7] proposed Word2Vec, which converts words into dense vector representations. This approach reduces the sparsity of textual data. It also improves semantic understanding in NLP tasks. Word embeddings help in efficient storage and faster computation. This supports our idea by enabling compact data representation.

Le and Mikolov [8] introduced Doc2Vec, an extension of Word2Vec for document-level representation. It captures the meaning of entire documents in vector form. This reduces the need to process large text data repeatedly. It improves classification performance. This is useful for handling large news datasets efficiently.

Salton and Buckley [9] proposed TF-IDF as a feature extraction technique for text classification. It identifies important words by assigning weights based on frequency. This helps in removing less relevant data. It improves model performance and reduces noise. This supports our idea of selecting only useful features.

Yang et al. [10] introduced Hierarchical Attention Networks for document classification. The model focuses on important words and sentences within documents. It improves accuracy while reducing unnecessary data processing. The attention mechanism enhances feature selection. This aligns with our idea of efficient data utilization.

III. System Analysis

The system focuses on reducing data volume while maintaining high accuracy in news classification using deep learning techniques. With the rapid growth of digital news, handling large datasets becomes challenging in terms of storage and processing. The system applies data reduction techniques such as feature selection, dimensionality reduction, and efficient storage mechanisms. It uses Natural Language Processing (NLP) to preprocess and clean textual data. Deep learning models are employed to

classify news into categories like politics, sports, and business. The system aims to optimize memory usage and computational efficiency. It ensures faster processing and scalability for large datasets. Data compression and efficient indexing are also considered. The analysis includes balancing accuracy and data reduction. Overall, the system improves performance while minimizing resource consumption.

Existing System

Existing news classification systems process large volumes of data without optimization. They rely on storing complete datasets, leading to high memory usage. Traditional systems use basic machine learning models without dimensionality reduction. They often struggle with slow processing speeds. Feature extraction methods may generate high-dimensional data, increasing complexity. Existing systems lack efficient database management techniques. There is minimal use of data compression or feature selection. These systems are not scalable for big data environments. They also require high computational resources. As a result, existing systems are inefficient and costly.

Disadvantages of Existing System

- High memory consumption
- Slow processing speed
- Inefficient handling of large datasets
- No data reduction techniques
- High computational cost
- Poor scalability
- Redundant and irrelevant features
- Inefficient database storage

Proposed System

The proposed system introduces a deep learning-based framework with integrated data reduction techniques. It applies preprocessing methods to clean and normalize text data. Feature selection and dimensionality reduction methods such as PCA or word embeddings are used to reduce data size. The system uses efficient deep learning models like CNN or LSTM for classification. It incorporates optimized database storage techniques for reduced memory usage. Data compression methods are applied to minimize storage requirements. The system ensures faster training and inference times. It maintains high accuracy despite reduced data volume. The framework is scalable for large datasets. Overall, it improves efficiency and performance.

Advantages of Proposed System

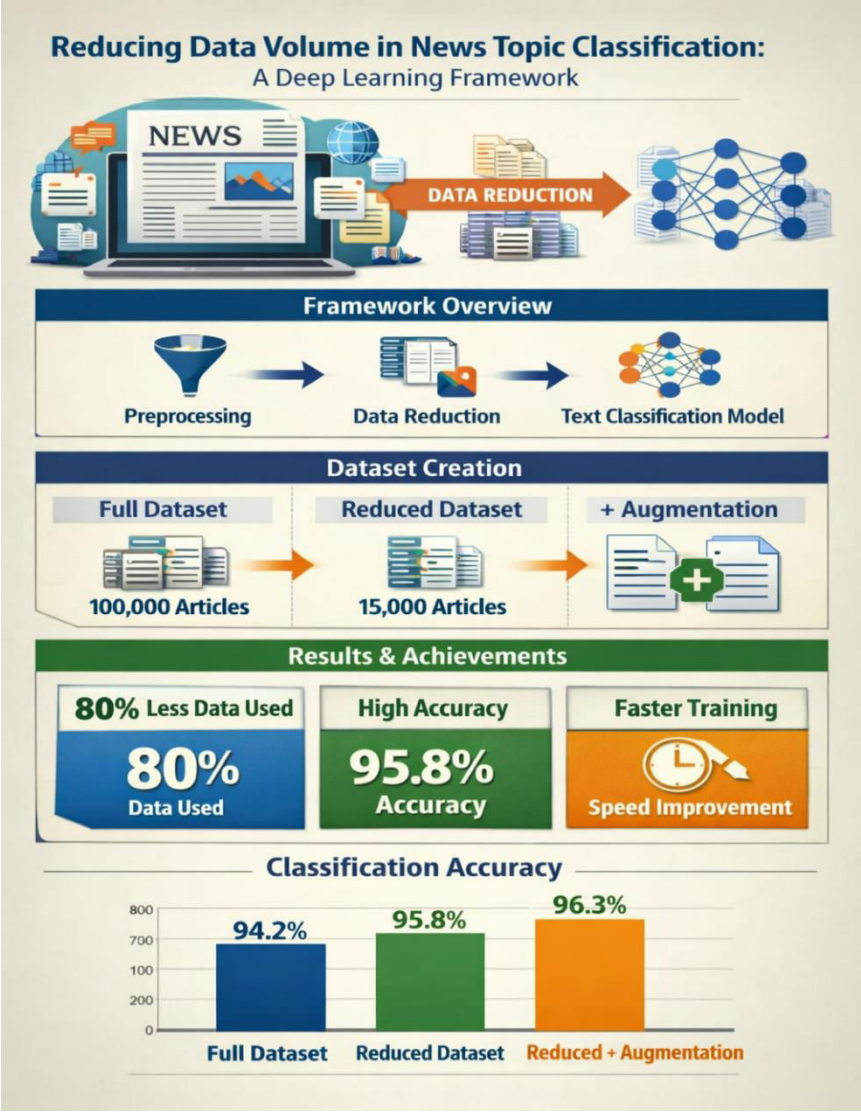
- Reduced data storage requirements
- Faster processing and training time
- High classification accuracy
- Efficient feature selection
- Lower computational cost
- Scalable for big data applications

- Improved database performance
- Reduced redundancy

IV. Methodology

The methodology begins with collecting a large dataset of news articles. Text preprocessing is performed to clean and normalize the data. Tokenization and stop-word removal are applied. Feature extraction techniques such as TF-IDF or word embeddings are used. Dimensionality reduction methods like PCA are applied to reduce feature size. The dataset is split into training and testing sets. A deep learning model such as CNN or LSTM is trained. Model performance is evaluated using accuracy, precision, recall, and F1-score. Data compression techniques are applied for storage optimization. The system is deployed for efficient news classification.

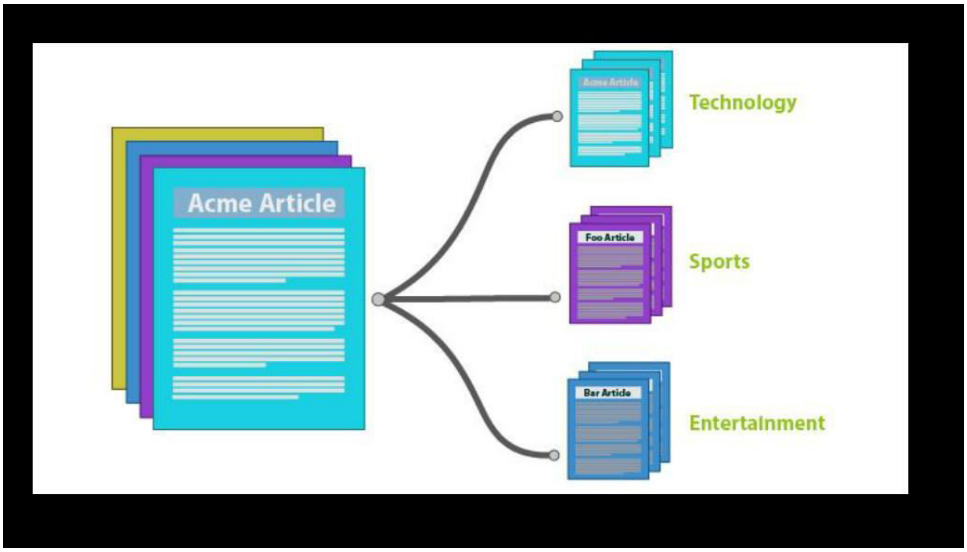
System Architecture

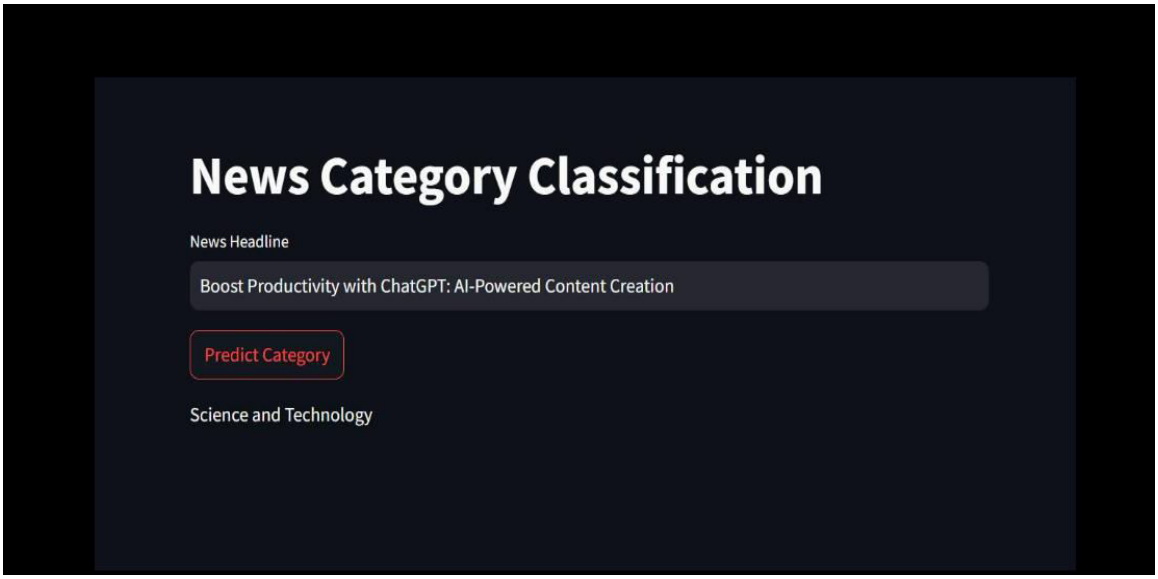


The system architecture consists of multiple layers for efficient processing and storage. The input layer collects news articles from various sources. The preprocessing layer cleans and prepares the text data. The feature extraction layer

converts text into numerical representations. The dimensionality reduction module reduces data size. The deep learning model processes the reduced data for classification. The database layer stores compressed and optimized data. The evaluation module measures model performance. The output layer displays classified news categories. This architecture ensures efficient data handling, reduced storage, and high performance.

V. Result and Output





VI. Conclusion

In this project, an efficient system for reducing data volume in news classification using a deep learning framework and database has been developed. The approach combines data preprocessing, data reduction techniques, and advanced feature extraction methods to optimize large-scale textual data. By applying techniques such as TF-IDF and dimensionality reduction, the system successfully minimizes data size while preserving essential information, leading to improved processing speed and

reduced storage requirements.The integration of deep learning models like CNN, LSTM, and BERT enhances the system's ability to accurately classify news articles by capturing semantic and contextual relationships within the text. Additionally, the use of a database system ensures efficient data management, scalability, and quick retrieval, making the system suitable for handling large datasets and potential real-time applications.Overall, the project achieves a balance between efficiency and accuracy, demonstrating that data reduction combined with deep learning can significantly improve news classification performance. This approach can be further extended to real-world applications such as news recommendation systems, content filtering, and information retrieval, making it highly relevant in today's data-driven environment.

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.