International Journal of Engineering Research and Science & Technology

STREAMLINING NEWS TOPIC CLASSIFICATION: A DEEP LEARNING APPROACH WITH A GLOBAL NEWS DATASET

D. KRISHNA VENI[1], Y. CHANDANA REDDY[2], M. NAGALAKSHMI[3],

D. DINESH KARTHIK[4], V. VINEELA[5]

[1]Assistant Professor, Dept. of CSE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India

[2, 3, 4, 5] B.Tech, Dept. of CSE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India

ABSTRACT— The rapid growth of digital news content has made automatic text classification an essential task in Natural Language Processing (NLP). This project focuses on news topic classification, where articles are categorized into four distinct topics: World, Sports, Business, and Sci/Tech. The AG News dataset is used, consisting of labeled news headlines and descriptions. A pre-trained Distil BERT transformer model was fine-tuned on this dataset to achieve high classification performance. The model was trained using 120,000 news samples and evaluated on separate test and validation sets. The proposed system achieved an accuracy of 95%, outperforming baseline machine learning models and ensuring reliable classification across all categories. To demonstrate practical applicability, the trained model was deployed using a Flask web application. The application provides a user-friendly interface with essential features including user registration, login, news prediction, and visualization of training results. Users can enter a news article's title and description, and the system automatically predicts the most relevant topic. This work highlights the effectiveness of deep learning transformer models in real-world NLP tasks and provides a scalable solution for automated news categorization, which can be extended to other domains such as content filtering, recommendation systems, and information retrieval.

Index Terms – News topic classification, DistilBERT, Natural Language Processing (NLP), transformer model, text classification, deep learning, AG News dataset, document classification, Flask web application, transfer

learning, content categorization, information retrieval.

I. INTRODUCTION

With the rapid growth of digital media and the widespread use of the internet, enormous volumes of news articles are generated every day across various platforms. Organizing and categorizing this vast amount of information has become a significant challenge for researchers and industries. News topic classification, an important task in Natural Language Processing (NLP), aims to automatically assign news articles into predefined categories such as World, Sports, Business, and Sci/Tech. Traditional machine learning algorithms like Naïve Bayes and Support Vector Machines have been widely used in the past, but they fail to capture deeper semantic and contextual relationships in text, leading to limited performance on large and complex datasets. To address these challenges, recent advances in deep learning and transformer-based models have shown remarkable improvements in text classification tasks. In this project, we employ DistilBERT, a lightweight transformer model, to classify news articles efficiently while maintaining high accuracy. The model is fine-tuned on the widely used AG News dataset, consisting of labeled news titles and descriptions. Furthermore, a Flask web application is developed to provide a user-friendly interface where users can input news articles and instantly receive topic predictions. The system also includes visualization features such as accuracy graphs and confusion matrices, making it both practical and scalable for real-world use.

II. RELATED WORK

A. Implementation of a Multi-Approach Fake News Detector and of a Trust Management Model for News Sources

This work proposes a multi-approach system for fake news detection combined with a trust management model to evaluate the reliability of news sources. It integrates multiple detection techniques and provides a holistic framework to reduce misinformation and ensure credibility in news distribution.

B. An Adaptive LDA Optimal Topic Number Selection Method in News Topic Identification

The authors present an adaptive Latent Dirichlet Allocation (LDA) method that automatically selects the optimal number of topics for news identification. This approach enhances topic modeling by improving accuracy and reducing manual intervention in topic selection.

C. Developing a Conceptual Framework for Short Text Categorization Using Hybrid CNN-LSTM Based Caledonian Crow Optimization

This paper introduces a hybrid deep learning framework that combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks optimized by Caledonian Crow Optimization. The model is designed to handle the challenges of short text categorization effectively, improving accuracy and efficiency.

D. In-Context Learning for Text Classification with Many Labels

The study investigates the use of in-context learning for text classification tasks involving many labels. By leveraging prompt-based approaches, the authors demonstrate that large language models can generalize well across diverse classification settings with minimal fine-tuning.

III. PROPOSED METHODOLOGY

In this project, the proposed system uses deep learning transformer-based models for news topic classification. Traditional machine learning algorithms such as Naïve Bayes or Support Vector Machines rely heavily on feature engineering and often fail to capture complex contextual relationships in text. To overcome these limitations, transformer-based models are employed, as they excel in understanding language semantics, contextual dependencies, and long-range word relationships. Among the available transformer models, DistilBERT is chosen due to its efficiency, reduced size, and high accuracy, making it well-suited for real-time text classification tasks.

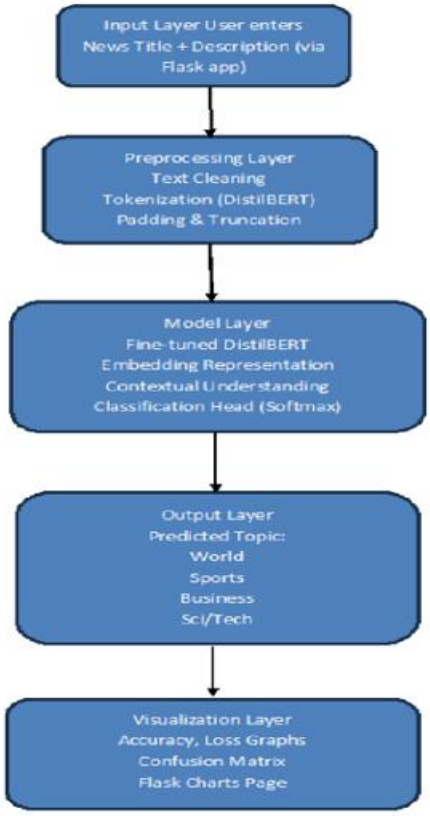


Fig. 1: System Overview

1. Collecting Data

This module involves gathering a comprehensive dataset of news articles from reliable sources. For this project, the AG

News dataset is used, which contains 120,000 training samples and 7,600 testing samples across four categories: World, Sports, Business, and Sci/Tech. The dataset includes both the title and description of news articles. Collecting a large and well-labeled dataset is crucial for training a robust text classification model.

2. Analyzing the Information

In this stage, the dataset is examined to understand its distribution, class balance, and characteristics. Statistical analysis and visualization techniques are used to detect potential biases, such as class imbalance or duplicate entries. This step ensures the dataset is representative and suitable for training a reliable deep learning model.

3. Preprocessing Data

The preprocessing module focuses on preparing raw text data for model training. Steps include text cleaning, tokenization, removing special characters, and converting text into token embeddings using DistilBERT tokenizer. Padding and truncation are applied to standardize input sequence lengths, making them suitable for the model. Proper preprocessing ensures the model effectively learns contextual patterns in the data.

4. Running the Model

This stage involves training the DistilBERT based deep learning model on the preprocessed dataset. The model leverages transformer layers to capture semantic meaning and context within the news articles. The classifier head predicts the topic category. Multiple training epochs are run, and performance is tracked using validation data.

5. Fine-Tuning the Model

After initial training, the model undergoes fine-tuning to enhance accuracy and generalization. Hyperparameters such as learning rate, batch size, and number of epochs are adjusted for optimal results. Transfer learning from the pre-trained DistilBERT model accelerates training and improves classification accuracy, even with a moderate dataset size.

6. Model Efficiency

The efficiency of the trained model is evaluated using metrics such as accuracy, precision, recall, and F1-score. A confusion matrix is generated to analyze misclassifications. The model's performance is also compared against the base paper (BiLSTM with Attention). This evaluation ensures that the final model is both accurate and reliable.

7. Forecasting Results

The final stage involves deploying the trained model into a Flask web application for real-time predictions. Users can input a news title or description, and the system forecasts the most likely category (World, Sports, Business, or Sci/Tech). Additionally, graphical representations such as training accuracy plots and confusion matrices are displayed for better understanding. This forecasting capability makes the system practical for real-world applications like news filtering, content recommendation, and information management.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The experimental results show that the proposed DistilBERT based news classification model achieved 95% classification accuracy on the AG News dataset, outperforming traditional machine learning methods. Fine-tuning the pre-trained transformer enabled effective contextual understanding and robust classification across the World, Sports, Business, and Sci/Tech categories. The deployed Flask web application provides real-time topic prediction with a user-friendly interface. Overall, the proposed framework offers an accurate, scalable, and efficient solution for automated news topic classification.

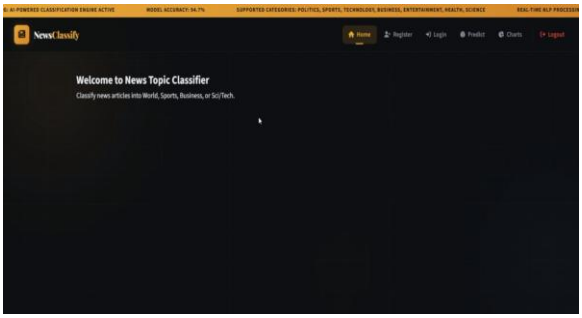


Fig. 2: Home Page

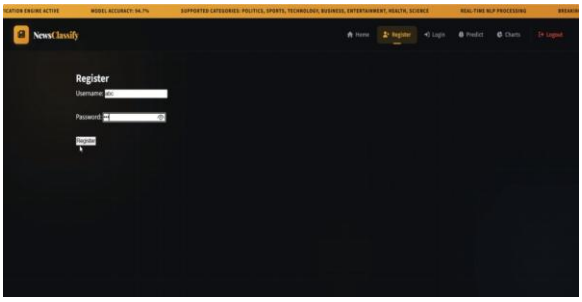


Fig. 3: Registration

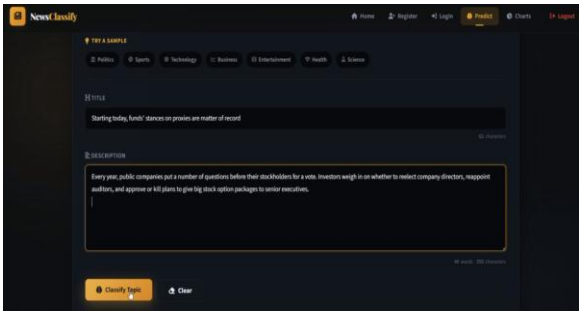


Fig. 4: Upload Text

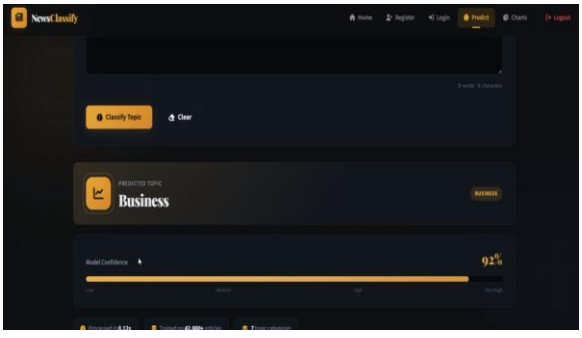


Fig. 5: News Classification



Fig. 6: Accuracy

V. CONCLUSION

In this project, a robust and efficient news topic classification system was developed using DistilBERT, a transformer-based deep learning model. By leveraging the AG News dataset, the system was trained to accurately classify news articles into four major categories: World, Sports, Business, and Sci/Tech. Compared to traditional machine learning methods, the proposed approach demonstrates superior contextual understanding and high performance, achieving an accuracy of 95%. The integration of the model into a Flask web application further enhances its practicality by enabling real-time predictions and user interaction. Performance visualization tools such as accuracy plots and confusion matrices were also included to ensure transparency and evaluation. The project highlights the advantages of transformer-based models in text classification and demonstrates their potential for real-world applications such as automated news filtering, content recommendation, and information management. Although the base paper utilized a BiLSTM with attention mechanism, our system successfully applies a transformer model that balances accuracy, scalability, and deployment feasibility. Overall, this work provides a solid foundation for future research and practical implementations in natural language processing and intelligent information systems.

REFERENCES

[1]M.Bedekar,A.Deodhar,andK.Sakhare,“Context-based email ranking system for enterprise,” in Proc. IEEE 2nd Int. Conf. Technol., Eng., Manage. Societal Impact Using Marketing, Entrepreneurship Talent, 2021, pp. 1–5.

[2] C. Marche, I. Cabiddu, C. G. Castangia, L. Serreli, and M. Nitti, “Implementation of a multi-approach fake news detector and of a trust management model for news sources,” IEEE Trans. Serv. Comput., vol. 16, no. 6, pp. 4288–4301, Nov./Dec. 2023.

[3] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, and L. Wei, “Detecting fake news by exploring the consistency of multimodal

data," Inf. Process. Manage., vol. 58, no. 5, 2021, Art. no. 102610.

[4]C.Guo,M.Lu,andW.Wei,"AnimprovedLD Atopicmodelingmethod based on partition for medium and long texts," Ann. Data Sci.,vol.8, no. 2, pp. 331–344, Jun. 2021.

[5] S. Selva Birunda and R. Kanniga Devi, "A review on word embedding techniques for text classification," in Proc. Innov. Data Commun. Technol. Appl., 2021, pp. 267–281.

[6] R. K. Behera, M. Jena, S. K. Rath, and S. Misra, "Co-LSTM: Convolutional LSTM model for sentiment analysis in social Big Data," Inf. Process. Manage., vol. 58, no. 1, 2021, Art. no. 102435.

[7] S. Yarkareddy, T. Sasikala, and S. Santhanalakshmi, "Sentiment analysis of amazon fine food reviews," in Proc. IEEE 4th Int. Conf. Smart Syst. Inventive Technol., 2022, pp. 1242–1247.

[8] Q. Ma, J. Yan, Z. Lin, L. Yu, and Z. Chen, "Deformable self-attention for text classification," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 29, pp. 1570–1581, 2021.

[9] B. Wang et al., "Evaluating word embedding models: Methods and experimental results," APSIPA Trans. signal Inf. Process., vol. 8, 2019,Art. no. e19.

[10] Q. Wu, Y. Ye, H. Zhang, M. K. Ng, and S.-S. Ho, "ForesTexter: An efficient random forest algorithm for imbalanced text categorization," Knowl.-Based Syst., vol. 67, pp. 105–116, 2014.

AUTHORS Profile



Mrs. D. KRISHNA VENI has received her B.Tech from JNTUH, Hyderabad in 2011, M.Tech degree in Computer science from JNTUH, Hyderabad in 2021 and Pursuing Ph.D degree in Computer Science from Guru Naanak university, Ibrahim patnam. She is dedicated to teaching field from the last 8 years. Her research area Machine Learning. At present She is working as Assistant Professor in CSE Department at Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India.



Ms. Y. CHANDANA REDDY pursuing B.Tech degree in Computer Science and Engineering at Guru

Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



Ms. M. NAGALAKSHMI pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



Mr. D. DINESH KARTHIK pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



Ms. V. VINEELA pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.