Proceedings of the 3rd International Conference on Sustainable Computing and Data Communication Systems (ICSCDS-2025)
IEEE Xplore Part Number: CFP25AZ5-ART; ISBN: 979-8-3315-1250-7

2025 3rd International Conference on Sustainable Computing and Data Communication Systems (ICSCDS) | 979-8-3315-1250-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICSCDS65426.2025.11167483

Multi-Model News Classification using BERT Features and Class-Balanced Resampling Techniques

Kallepalli Rohit Kumar
Department of CSE
Vignan's Institute of Information Technology
Visakhapatnam, Andra Pradesh, India
Krk542@gmail.com

Ranjith Kumar Chinnam
Assistant Professor, Department of AI&ML
Aditya University
Surampalem, India
ranjithkumarc@adityauniversity.in

Venkata Subbaiah Desanamukula
Department of CSE
Lakireddy Bali Reddy College of Engineering, Mylavaram, India
drdvs.2021@gmail.com

Ajahar Ismailkha Pathan
Assistant Professor,Dept of Computer Engineering
PSGVPM's D.N.Patel College of Engineering, Shahada Dist, Nandurbar, Maharashtra, India
pathan.ajaharkha@gmail.com

Naladi Ram Babu
Associate Professor,Department of EEE
Aditya University
Surampalem, India
rambabunaladii@gmail.com

Dr.G.Charles Babu
Professor, Dept of CSE
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, Telangana, India
charles1624@grietcollege.com

Abstract— **This paper presents a robust framework for automated news classification using transformer-based embeddings and deep neural models, enhanced through resampling techniques. The methodology begins with text preprocessing and feature extraction using BERT, followed by class balancing using both SMOTE and Random Oversampling. The balanced datasets are used to train a series of machine learning models, including Logistic Regression, Random Forest, and Gradient Boosting, as well as deep learning models such as 1D-CNN, LSTM, and a hybrid CNN-LSTM network. Experiments conducted on the BBC News dataset demonstrate that the hybrid CNN-LSTM model achieved the highest performance, with an accuracy of 94.10% and an F1-score of 94.00% on the SMOTE-balanced data. Traditional machine learning models also showed competitive performance, with Gradient Boosting reaching up to 94.00% accuracy. The comparison between resampling methods revealed that SMOTE consistently led to better classification outcomes. The proposed approach effectively combines contextual feature extraction and deep learning, making it a promising solution for scalable and accurate news classification.**

Keywords— News Classification, BERT, SMOTE, CNN-LSTM, Deep Learning, Text Categorization.

I. INTRODUCTION

The digital news ecosystem has boomed in recent years, spurred on by the popularity of publishing tools and social media. This explosion of information accessibility has resulted in major consequences for accessing information, but also poses challenges to this problem, such as how organized, categorized and get relevant news effectively. Automatic categorization of news articles, referred to as news classification, is of vital importance for powering content filtering, personalized recommendations and topic-based navigation in digital media systems.

Traditional approaches of news classification mainly involve ML models on the basic bag-of-words or TF-IDF information. The model, although computationally cheap, has limited ability to capture more complex semantic comparison as observed in natural languages. As a result they struggle in practical large scale scenarios where class overlap, fine-grained topic differences, and complex sentence semantics are often encountered. Yet another common challenge in news datasets is class imbalance, where some categories (e.g., politics, sports) are overrepresented relative to others (e.g., technology, entertainment). Such imbalance may result in predictions which are biased towards majority classes and this will harm both overall performance and fairness of the classifiers. A good control of this imbalance is critical to obtain sound and fair predictions for all the categories.

The advent of transformer-based models such as BERT has recently helped leaps and bounds in modeling semantics of the language. BERT does that by incorporating bidirectional information of surrounding words, which leads to a more complex representations of a piece of text. In this paper, a deep learning model proposed incorporating BERT embeddings and different types of classification model such as 1D-CNN, LSTM, and hybrid CNN- LSTM architecture. In order to deal with the class imbalance two resampling methods namely, SMOTE and Random Oversampling, are used. We train these models using classes balanced examples from the BBC News corpus and evaluate their classification using standard metrics. It endeavours to showcase how transformer-based feature extraction methods in combination with deep architectures and efficient resampling can boost the performance and fairness of automated news classification systems.

II. LITERATURE SURVEY

Widha Dwiyanti et al. [1] proposed that the news sentiment also influenced to predict Jakarta Composite Index (JCI) with Support Vector Regression model. They incorporated emotion features from ChatGPT based general large language models into a financial model. Their results demonstrated wide variation of task accuracy among different sentiment integration strategies and different settings of hyperparameters. Their results suggested that sentiment analysis contributed to stock price prediction

models but it depended on the approach for sentiment integration. F. Hamami et al. [2] established a system for news segmentation to solve the information overload issue in Indonesian news portals. They used the K-means algorithm (using the Elbow Method) for clustering the Kompas news articles. com using the similarity between contents. Daily varying segmentation results were made publicly available in a web application. S. K. Rao et al. [3] presented an approach to identify hyper partisan news articles that exacerbate media bias and public polarization. They did this using two models, fine-tuned DeBERTa-v3 and Bi-LSTM with Fast Text embeddings where they considered and compared both the models. Their experiments demonstrated that the DeBERTa-v3 model achieved superior performance even compared to Bi-LSTM, as evidence of the effectiveness of the transformer-based models for detecting political bias in news text.

Deepa Gupta et al. [4] proposed an automatic news classifier system for regional Indian languages such as Kannada and Telugu. In their work, the news were preprocessed and classified using natural language processing meta-features within pre-established categories, such as Entertainment, Sports or Business. They experimented with various ML models and proved that language-specific embeddings such as KannadaSBERT and TeluguSBERT increased classification accuracy. Their work focussed on issues in low-resource language processing and increased access to regional language content. X. Pu et al. [5] presented a news recommendation system NWT that unified word relevance and topic prediction through co-training. Their model by using a multi-head self-attention network on news title to learn better news title representations by considering the semantic relationships among words. Then topic specific features have been extracted by a topic perceptron. The user encoder additionally learned user preferences by emphasizing articles that best corresponded to the user preference. Experiments on the MIND dataset demonstrated that NWT achieved better results than some existing baseline methods in personalized news recommendation, which proved its effectiveness.

Evis Plaku et al. [6] authors presented a collection of 9,600 Albanian news titles and compared the various ML models for topic classification. Their research demonstrated that more complex models such as RNNs outperformed basic classifiers and ensemble models. A. Mathialagan et al. [7] introduced a fused ensemble model, that joints BERT and BLIP-2 for the multi-modal news summarization. Their method successfully solved the problem of combining text and image information and obtained better summarization quality comparing to the baseline models. D. S. Muthukumar et al. [8] built a chatbot that applied Latent Semantic Analysis and sub-topics of large language models for news aggregation across sources. The system generated concise, informative summaries which helped users to understand the given texts.

Ramdhani et al. [9] used Indonesian language specific CNN model for classifying news. They pre-processed the data by removing stopwords and stemming, and the model did well in predicting the corpus of multi-category. Lantian Zheng et al. [10] built an English news classification system based on machine learning with the set of 20 NewsGroups data. They trained a TF-IDF based model and their model did well in both training and testing stages, showing that it generalized well. Srinivas et al. [11] designed a model for the detection of fake job postings based on Deep Neural Network. They also compared affiliation results with other ML algorithms emphasizing the importance of automated detection for online job resources.

[12] explored various DL models for news classification like CNN, RNN (like LSTM) and Transformer style BERT. They found that BERT and its combined model with LSTM presented the best accuracy, highlighting the model's contextual sense for text classification. A. L. Rao et al. [13] dealt with fake news spread in social networks using MLmodels. Their Python based model has detected Fake news effectively and this experiment shows that NLP-based text classification methods can be used in the context of fake news detection. [14] introduced a ML based classification system of crime related news headlines using the live news data. Feature selection was based on TF-IDF, and we found that Decision Tree and Random Forest were the most effective classification algorithms among the models we assessed, which may be of value for crime news prediction.

III. RESEARCH METHODOLOGY

The proposed methodology is shown in figure 1 The proposed framework for news classification begins with the collection and preparation of the BBC News dataset, which includes news articles categorized under five distinct topics. Duplicate entries are removed to ensure data quality, and the remaining articles are subjected to standard text preprocessing steps. These include converting all text to lowercase, removing punctuation, digits, and common stop words, followed by tokenization using BERT's pre-trained tokenizer. This step transforms each article into a sequence of tokens compatible with the BERT model and prepares them for embedding generation.

To extract meaningful features from the text, the BERT-base-uncased model is used to generate contextual embeddings. Each news article is passed through the BERT model, and the embedding corresponding to the [CLS] token is extracted as a dense, fixed-length representation of the entire article. These embeddings capture the semantic and contextual nuances of the text, offering a more expressive input for classification compared to traditional vectorization methods. The resulting feature vectors serve as the foundation for training both ML, DL models.

Since the original dataset exhibits some imbalance across categories, two resampling techniques are applied to generate balanced versions of the training data. The first technique, SMOTE, creates synthetic samples for the minority classes by interpolating between existing instances, thereby improving class representation without duplication. The second approach, Random Oversampling, balances the dataset by duplicating existing samples in underrepresented classes. These two strategies result in two distinct training sets, each of which is used to train and evaluate the proposed classification models.

Classification consists of classical ML techniques and DL models. Baseline models We use Logistic Regression, Random Forest and Gradient Boosting models trained directly on the BERT embeddings. Meanwhile, several deep learning models, such as 1D-CNN and LSTM are applied and the performance of these models to capture the spatial and sequential properties is examined. The performance of a hybrid CNN-LSTM model is also implemented to exploit the

advantages of the two models. All models are trained using normal optimization with appropriate regularization and are tested on the SMOTE and Random Oversampling data sets in order to compare performances under different resampling.

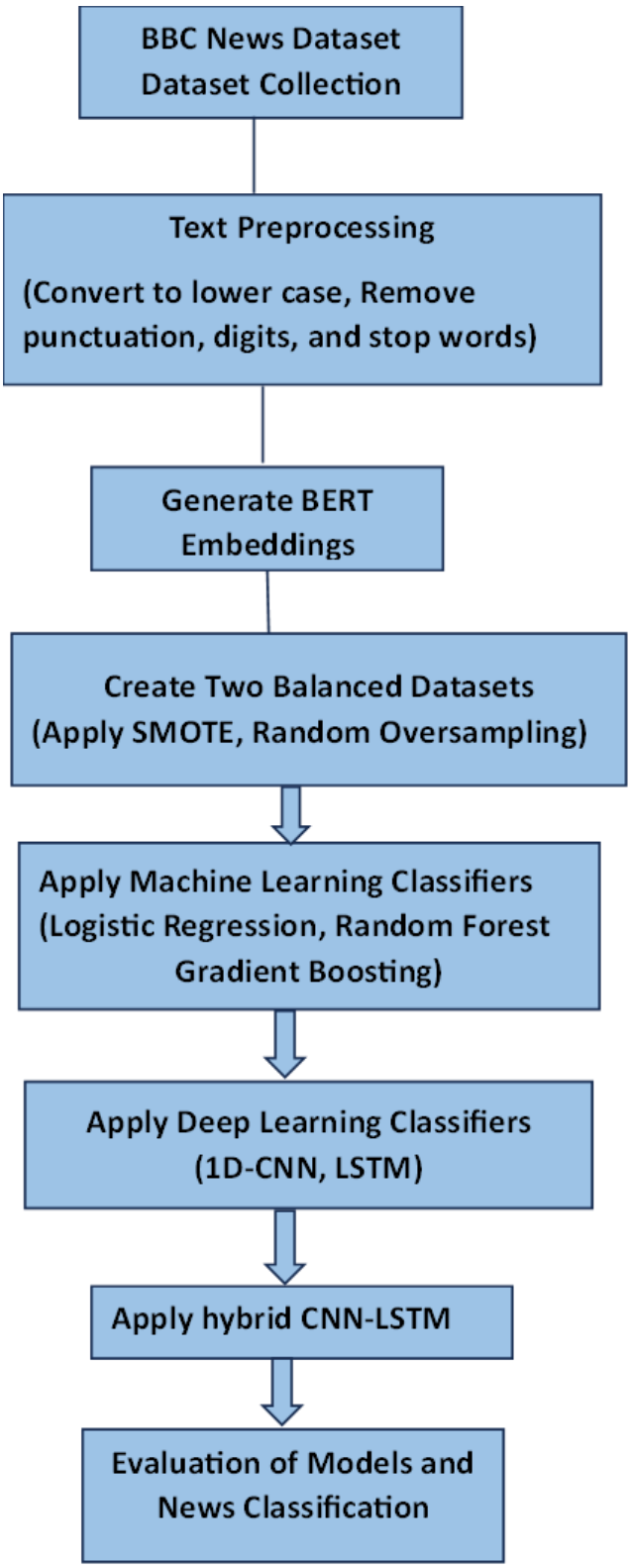

Fig. 1. Proposed Methodology

IV. Experimentation and Results

A. Data Collection

The data used in this study are news articles from the publicly available BBC News dataset in Kaggle [15]. This repository contains 1,490 articles, where each text is assigned to one of five categories: business, entertainment, politics, sport or tech. After duplicate had been removed, a number of duplicates were identified upon a first glance at the data, these were then removed to improve data quality, where only unique articles were selected for proceeding with the data process. After this cleaning process, the dataset consisted of 1,440 unique news articles. These articles were relatively well-distributed across the five categories, providing a suitable foundation for evaluating classification models. The cleaned dataset was then used for all subsequent stages, including text preprocessing, feature embedding using BERT, resampling for class balance, and model training. Its multi-class structure and real-world content made it appropriate for benchmarking advanced natural language processing techniques. Figure 2 shows class wise category of news.

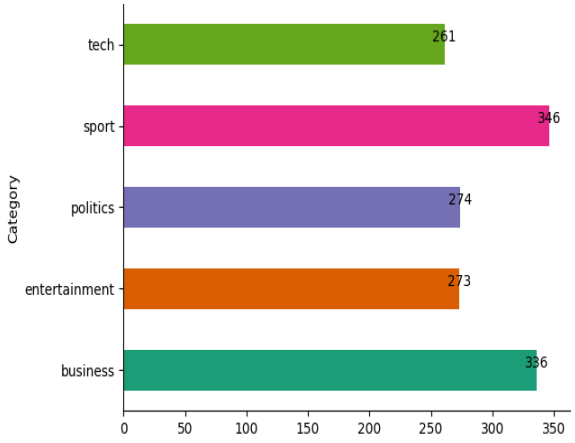

Fig. 2. Distribution of news categories in the dataset [14]

B. Data Preprocessing

Several preprocessing steps were applied to standardize and clean the news articles. Initially, all text content was converted to lowercase to maintain uniformity and reduce redundancy caused by case sensitivity. Punctuation marks, numerical digits, and special characters were then removed, as these elements typically do not contribute meaningful semantic information in the context of news categorization. Stop words, which are common words such as "the," "and," or "is" that often carry limited discriminative value, were also filtered out to minimize noise in the data. Following this, each article was tokenized into individual words using the tokenizer associated with the BERT model, enabling compatibility with the embedding process. The BERT tokenizer was specifically chosen due to its subword-based approach, which effectively handles out-of-vocabulary words and captures finer linguistic details. The tokenized text was converted into sequences of input IDs and attention masks – these are core elements that enable BERT to learn and process the input in a context-aware way such as to make predictions over long distance dependencies. These tokenized inputs formed the basis for high-dimensional embeddings, which were then used for training DL models. This pre-processing pipeline made it

possible to represent the raw textual data into a format that is structured and se- mantically rich and is appropriate for further processing of advanced neural networks.

C. BERT Embedding Generation

After pre-processing and tokenization of the textual data, we then generated dense semantics for each news article through a pre-trained transformer model. To this end, we used the BERT-base-uncased model, proposed in [23] and implemented in the HuggingFace Transformers library, which has been shown to be successful in modeling deep contextual signals from text. Each preprocessed article was fed into the BERT model and the output from the [CLS] token was sent into the next step. This token is a dense representation derived from the entire input sequence which is fed into a classifier. The output embeddings are 768-dimensional vectors which are invariant under the meaning of the news articles in context-sensItive way. These embeddings represented a richer and more informative feature space than traditional algorithms, such as TF-IDF or word2vec, since the meaning of the word is captured and combined with the information of the surrounding words within the article. These BERT embeddings were saved as numerical feature vectors that were used as input to all classification models. This method guaranteed our classifiers learned from more than just trivial word frequency information and instead rich, language-aware representations of the text.

D. Resampling Techniques

Following the generation of BERT embeddings, the issue of class imbalance within the dataset was addressed using two distinct resampling techniques: SMOTE and Random Oversampling. Although the dataset was relatively balanced across categories, slight disparities in class distribution could still impact model performance, particularly for minority classes. To mitigate this effect, two separate versions of the dataset were created. In the first version, SMOTE was applied to synthetically generate new data points for the underrepresented classes by interpolating between existing minority samples in the embedding space. This method helps create a more informative decision boundary by introducing synthetic but plausible samples. In the second version, Random Oversampling was used to replicate existing instances from the minority classes, thereby achieving balance by increasing their frequency in the training set. Both resampling techniques were applied only to the training data to preserve the integrity of the testing phase.

E. Applying Machine Learning Algorithms

To establish a comparative baseline before applying DL architectures, three classical ML algorithms were utilized: Logistic Regression, Random Forest, and Gradient Boosting. These models were selected for their popularity and proven effectiveness in text classification tasks.

The results obtained using the SMOTE-balanced dataset are summarized in Table I and figure 3. In Table I, Gradient Boosting achieved the highest classification accuracy of 94.0%, followed closely by Random Forest at 93.2%, while Logistic Regression lagged slightly behind at 91.0%. Similar trends were observed in precision, recall, and F1-score, where Gradient Boosting consistently performed best across all metrics, indicating its strong capability in leveraging the semantic richness of BERT embeddings. Random Forest also demonstrated competitive performance, significantly outperforming Logistic Regression, which showed comparatively lower recall and F1-score despite decent precision.

TABLE I. RESULTS ON SMOTE-BALANCED DATASET WITH ML MODELS

Model	Acc	Precision	Recall	F1
Logistic.Reg	91.0%	90.5%	89.8%	90.1%
Random Forest	93.2%	93.0%	92.5%	92.7%
GBC	94.0%	93.4%	94.1%	93.7%

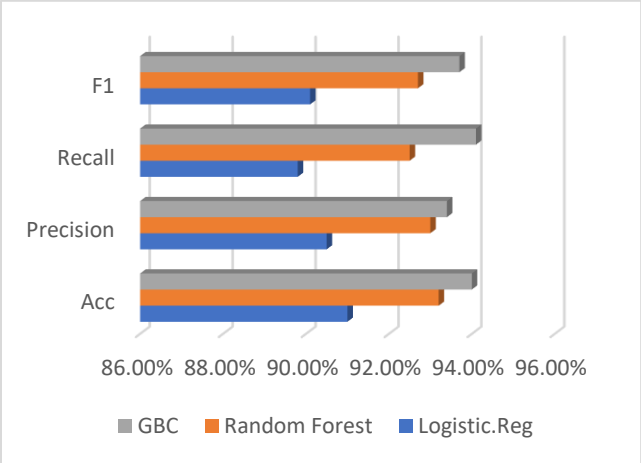

Fig. 3. Results of ML models with SMOTE data

The results obtained using the Random Oversampling dataset is presented in Table II and figure 4. As expected, the performance of all models was slightly lower than with SMOTE. Gradient Boosting still led with an accuracy of 92.3%, followed by Random Forest at 91.5% and Logistic Regression at 89.8%.

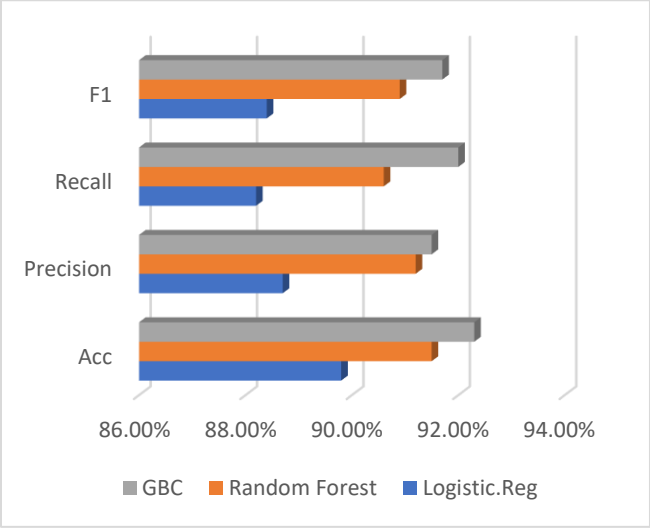

Fig. 4. Results of ML models with Random Oversampling data

Precision, recall, and F1-score values showed a similar pattern, reaffirming that Gradient Boosting is more effective

at exploiting the structure of balanced data. These findings highlight not only the superiority of Gradient Boosting among the three models but also the advantage of using SMOTE over simple random duplication for improving classification outcomes on imbalanced datasets.

TABLE II. RESULTS ON RANDOM OVERSAMPLING DATA WITH ML MODELS

Model	Acc	Precision	Recall	F1
Logistic.Reg	89.8%	88.7%	88.2%	88.4%
Random Forest	91.5%	91.2%	90.6%	90.9%
GBC	92.3%	91.5%	92.0%	91.7%

F. Applying Deep Learning Algorithms

To explore the effectiveness of deep neural networks in news classification, two distinct architectures (1D-CNN) and LSTM were implemented using BERT-based embeddings as input. The 1D-CNN model was designed to capture local semantic patterns in the high-dimensional embeddings using a single convolutional layer followed by max pooling, flattening, and dense layers. This architecture is particularly effective at extracting n-gram-like patterns in non-sequential input representations. In contrast, the LSTM model was employed to examine the sequential aspects of the BERT embeddings, treating the 768-dimensional vectors as sequences of features. Although BERT outputs are not inherently temporal, LSTMs can still capture long-range interactions and dependencies between different parts of the representation, often enhancing performance. Both models were trained and evaluated on the dataset balanced using SMOTE. The results for this configuration are presented in Table III and figure 5.

TABLE III. RESULTS ON SMOTE DATA WITH DL MODELS

Model	Acc	Precision	Recall	F1
1D-CNN	93.5%	93.2%	92.8%	93.0%
LSTM	92.3%	91.5%	91.0%	91.2%

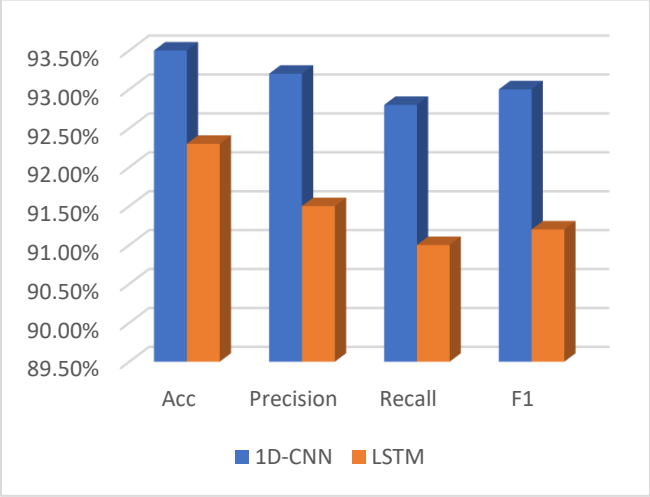

Fig. 5. Results with DL models on SMOTE-Balanced Dataset

The results in Table III demonstrate that 1D-CNN slightly outperformed LSTM across all evaluation metrics when trained on the SMOTE-balanced dataset. The 1D-CNN model achieved an accuracy of 93.5%, indicating its capability to effectively capture semantic cues embedded in BERT representations. Its precision and recall were also consistently higher, resulting in a strong F1-score of 93.0%. The LSTM model, while also showing robust performance with over 92% accuracy.

The models were further trained and tested on a different dataset generated from a Random Oversampling. The values obtained for this configuration are reported in Table IV and in fig.6. In Table IV, we observe that both models have a slightly diminished performance when they are trained in the Random Oversampled balanced dataset. However, even after that, the best results were still achieved using 1D-CNN model, with 91.8% accuracy and good F1-score of 90.8%. Looking at LSTM in comparison, it performed in the order of all parameters worse in terms of all different measures with an accuracy of 90.2% and an F1-score of 89.2%. These differences corroborate the insight that BERT embedding's structural properties are handled more effectively by 1D-CNN. And comparison between Tables III and IV shows that SMOTE resampling is more effective for training deep learning models on imbalanced data than Random Oversampling probably because it can generate more diverse and informative synthetic samples.

TABLE IV. RESULTS ON RANDOM OVERSAMPLING DATA WITH DL MODELS

Model	Acc	Precision	Recall	F1
1D-CNN	91.8%	91.0%	90.6%	90.8%
LSTM	90.2%	89.6%	88.9%	89.2%

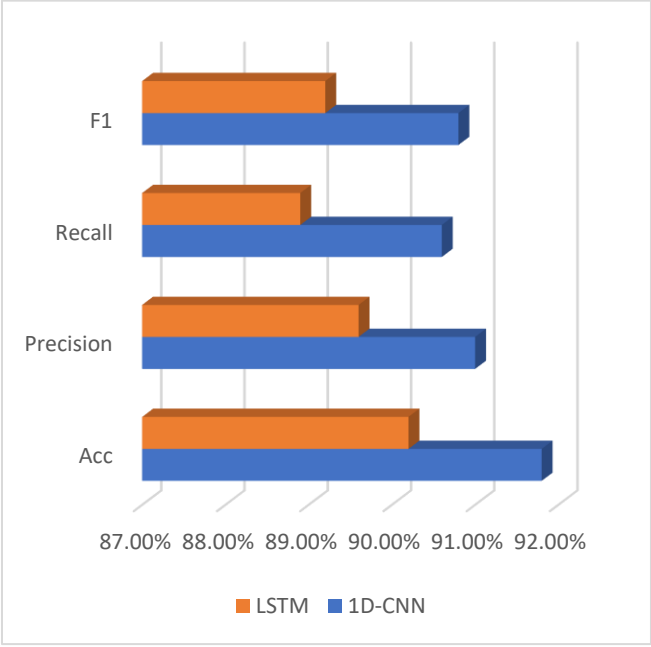

Fig. 6. Results with DL models on Random Oversampling Dataset

Table IV presents a slight reduction in performance for both models when trained on the dataset balanced using Random Oversampling. The 1D-CNN model still maintained superior results compared to LSTM, achieving 91.8% accuracy and a respectable F1-score of 90.8%. In contrast, LSTM scored lower across all metrics, with an accuracy of 90.2% and an F1-score of 89.2%. These differences reinforce the observation that 1D-CNN handles the structural characteristics of BERT embeddings more efficiently.

G. Applying Hybrid CNN-LSTM

In order to improve classification accuracy, a mixed deep learning model CNN-LSTM was applied. The results of hybrid CNN-LSTM on the above resampling methods are summarized in Table V and figure 7. Although the model performs well in prediction for all categories when trained on the SMOTE-balanced data set, which is indicated by an AUC of 0.94.10%, it indicates that the model is good at sorting into correct news categories. Precision and recall were also high, i.e. 93.70% and 94.30%, resulting in a balanced F1-score of 94.00%. These findings demonstrate that the hybrid model incorporated the ability to capture local and temporal features from BERT embeddings and took advantage of diversification introduced by SMOTE. For the random OVER sampled dataset, the performance of the model was somewhat lower across all measures as opposed to the good learning performance reported here. The accuracy decreased to 92.5% while the precision (92.10%), recall (91.50%), and F1-score (91.80%) all decreased. Even though to a much lesser extent, these findings have implications not only for its ability of creating informative variation when varying close neighbors, but also for its differentiation of original and new samples.

TABLE V. RESULTS WITH HYBRID MODEL

Measure	SMOTE-Data	Random Oversampling Data
Accuracy	94.10%	92.5%
Precision	93.70%	92.10%
Recall	94.30%	91.50%
F1	94.00%	91.80%

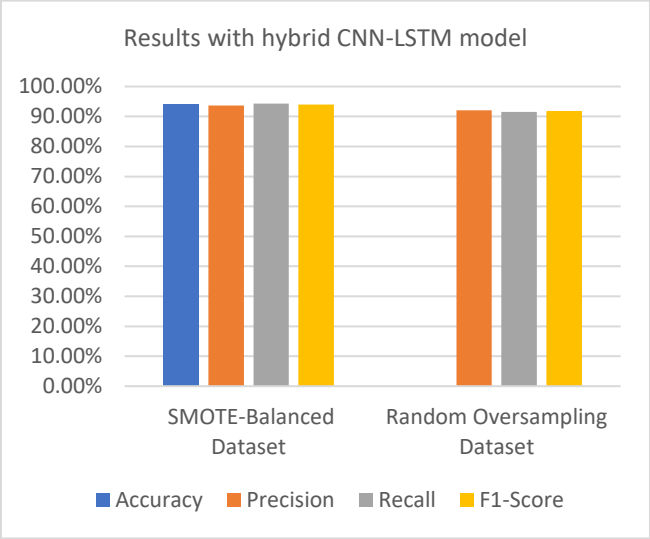

Fig. 7. Result with CNN+LSTM model

V. CONCLUSION

In this paper, a context-aware news classification approach was developed by integrating BERT-based embeddings with traditional ML and DL models, supported by hybrid resampling techniques. The BBC News dataset was used for experimentation, and two resampling methods—SMOTE and Random Oversampling—were applied to address class imbalance. Logistic Regression, Random Forest, and Gradient Boosting were implemented as baseline ML models, while 1D-CNN, LSTM, and CNN-LSTM were employed to evaluate the effectiveness of deep learning architectures on BERT-derived features. Experimental results showed that the CNN-LSTM model achieved the highest performance across all evaluation metrics, particularly when trained on the SMOTE-balanced dataset. Gradient Boosting also performed well among the ML models. The comparison between SMOTE and Random Oversampling further confirmed that SMOTE contributed to more stable and improved model performance. Overall, the findings demonstrate that combining transformer-based embeddings, deep neural models, and appropriate resampling techniques can significantly enhance the accuracy and robustness of news classification systems.

REFERENCES

[1] W. Dwiyanti et al.,"The Effect of News Sentiment on Jakarta Composite Index Prediction Using Support Vector Regression Method," 2025 International Conference on Advancement in Data Science, E-learning and Information System (ICADEIS), Bandung, Indonesia, 2025.

[2] F. Hamami e tal.,"Analyzing News Clustering with K-Means Algorithm: A Segmentation Perspective," 2024 4th International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS), Yogyakarta, Indonesia, 2024.

[3] S. K. Rao et al.,"Identifying Hyperpartisanship in News Media Using Bi-LSTM and Decoding-Enhanced BERT," 2024 Asian Conference on Intelligent Technologies (ACOIT), KOLAR, India, 2024, pp. 1-5.

[4] D. Gupta et al.,"Automated Regional Language News Article Classification System for Kannada and Telugu," 2025 Fifth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 2025, pp. 1-6.

[5] X. Pu et al.,"News Recommendation With Word-Related Joint Topic Prediction," in IEEE Access, vol. 12, pp. 72566-72577, 2024.

[6] E. Plaku et al., "A Machine Learning Framework for Automated News Article Title Classification in Albanian," 2024 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), Craiova, Romania, 2024, pp. 1-6.

[7] A. Mathialagan et al.,"A Fused Transformer-Based Ensemble Approach Using Pre-Trained LLMs for Multi-Modal News Summarization," 2025 International Conference on Inventive Computation Technologies (ICICT), Kirtipur, Nepal, 2025.

[8] D. S. Muthukumar et al.,"A Framework for Analyzing and Summarizing News and Articles using Large Language Model," 2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS), Gobichettipalayam, India, 2024.

[9] M. Ali Ramdhani, et al.,"Indonesian news classification using convolutional neural network," IJEECS, vol. 19, no. 2, p. 1000, Aug. 2020, doi: 10.11591/ijeecs.v19.i2.pp1000-1009.

[10] L. Zheng, "Classification and Prediction of English News Text Based on Machine Learning," 2024 International Conference on Power, Electrical Engineering, Electronics and Control (PEEEC), Athens, Greece, 2024.

[11] G. Srinivas et al.,"Fake News Detection Using ML and DL Approaches," 2023 International Conference on Circuit Power and Computing Technologie, Kollam, India, 2023, pp. 1322-1325.

[12] G. Geddam et al., "Exploring Deep Learning Approaches for News Classification with CNNs, RNNs and Transformers," 2024 First

International Conference on Innovations in Communications, Electrical and Computer Engineering , Davangere, India, 2024.

[13] A. Lakshmanarao et al., “An Effecient Fake News Detection System Using Machine Learning,” IJITEE, vol. 8, no. 10, pp. 3125–3129, Aug. 2019.

[14] Sreekala, Keshetti, Modugula Sivajyothi, Dukuru Chiranjevi, Gadipe Sunitha, and Mucha Swetha. "A Novel Integration of Hierarchical Clustering and Ensemble Classification Algorithms for News Classification." In *2024 International Conference on Cognitive Robotics and Intelligent Systems (ICC-ROBINS)*, pp. 524-528. IEEE, 2024.

[15] https://www.kaggle.com/competitions/learn-ai-bbc/data.