Multiclass Text Classifications of Sindhi Newspaper Articles

Sanjai Kumar and Raja Vavekanand *

Posted Date: 22 January 2025

doi: 10.20944/preprints202501.1580.v1

Keywords: Sindhi Language; News Articles Classifications; Low Resource Languages

Article

Multiclass Text Classifications of Sindhi Newspaper Articles

Sanjai Kumar [1] and Raja Vavekanand [2]

[1] University of Sindh, Jamshoro76080, Pakistan
[2] Datalink Research and Technology Lab, Islamkot 69240, Pakistan
* Correspondence: bharwanivk@outlook.com

Abstract: The classification of newspaper articles into predefined categories is a key challenge in Natural Language Processing (NLP), particularly for underexplored languages like Sindhi, which present unique linguistic complexities. This study developed a custom-curated Sindhi newspaper dataset containing 6,156 articles categorized into entertainment, sports, and technology. Four deep learning models Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory Networks (LSTMs), and a hybrid CNN-LSTM model were trained using optimized hyperparameters and evaluated using metrics such as accuracy, precision, and recall. The dataset underwent rigorous preprocessing, including tokenization and normalization, to enhance model performance. Each model was trained using an 80-20 train-test split, and early stopping was employed to mitigate overfitting. The CNN and hybrid models achieved the highest accuracy of 96%, effectively capturing spatial and sequential patterns. LSTM closely followed with 95.85%, while the RNN lagged at 67%, highlighting its limitations with long-term dependencies. These results underline the potential of hybrid architectures and advanced sequence models for text classification tasks in low-resource languages like Sindhi. Source Code: https://github.com/rajavavek/Multiclass-Classification-of-Sindhi-Newspaper-Article

Keywords: Sindhi Language; News Articles Classifications; Low Resource Languages

1. Introduction

When it comes to organizing and retrieving textual content is a crucial task in today's digital age. Newspaper article classification plays a vital role in this process, enabling applications such as digital archiving and automated news categorization (Anitha et al., 2024; Setu et al., 2024). However, regional languages like Sindhi have been overlooked in natural language processing (NLP) research, making it challenging to develop efficient classification systems (Soomro et al., 2024). Sindhi, with its unique grammatical rules, rich vocabulary, and script, presents distinct challenges, particularly in tokenization and feature extraction. To address these challenges, our study explores state-of-the-art deep learning models for the multiclass classification of Sindhi newspaper articles into predefined categories, including entertainment, sports, and technology.

We evaluated the performance of several models, including CNN, RNN, LSTM, and Hybrid, to determine their ability to capture the linguistic nuances of Sindhi text. Our findings demonstrate the efficacy of deep learning approaches in handling complex text classification tasks, highlighting the potential for further exploration in Sindhi NLP. By focusing on this underexplored language, our research contributes to the development of Sindhi NLP and provides valuable insights applicable to other low-resource languages. Our study serves as a stepping stone for future research, paving the way for the creation of more efficient and accurate NLP systems for Sindhi and other regional languages.

2. Literature Review

The multiclass classification of Sindhi newspaper articles is an emerging area of natural language processing (NLP) leveraging deep learning techniques. Prior studies have explored various models for text classification in low-resource languages, emphasizing the challenges of data scarcity, morphological complexity, and the absence of pre-trained language models tailored (Magueresse et al., 2020; Ilyas et al., 2021; Aliyu et al., 2024; Song et al., 2024). Deep learning approaches, particularly Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and Bidirectional Encoder Representations from Transformers (BERT) have shown promise in multilingual and low-resource settings (Cruz & Cheng, 2020; Li et al., 2020; Fesseha et al., 2021; Maheen et al., 2022; Marreddy et al., 2022; Yohannes & Amagasa, 2022). For instance, studies applying CNNs and LSTMs to regional languages demonstrate notable accuracy improvements through semantic feature extraction (Ombabi et al., 2020). However, such methods often rely on transliteration or hybrid datasets, which may dilute the linguistic nuances of Sindhi (Rajan & Salgaonkar, 2021).

Sindhi-specific research remains limited, with most efforts focusing on basic NLP tasks like tokenization and sentiment analysis (Ali & Imdad, 2017; Nawaz et al., 2023). While transfer learning using multilingual BERT models has achieved moderate success, challenges persist in fine-tuning due to mismatched linguistic contexts and a lack of annotated datasets(Wadud et al., 2022; Pyysalo et al., 2020). A critical gap lies in the absence of large-scale labeled corpora and pre-trained Sindhi language models, hampering advancements in multiclass classification (Soomro et al., 2024). Existing studies also lack comprehensive evaluations of models’ robustness across diverse newspaper genres, limiting generalizability.

2.1. Methodology

This research utilizes a custom-curated Sindhi newspaper dataset to evaluate four deep-learning models: CNN, RNN, LSTM, and Hybrid. Optimized hyperparameters and performance metrics, including precision, recall, and accuracy, are used to compare the models' effectiveness in text classification (Fig 1), leveraging TensorFlow and Keras frameworks.

doi:10.20944/preprints202501.1580.v1

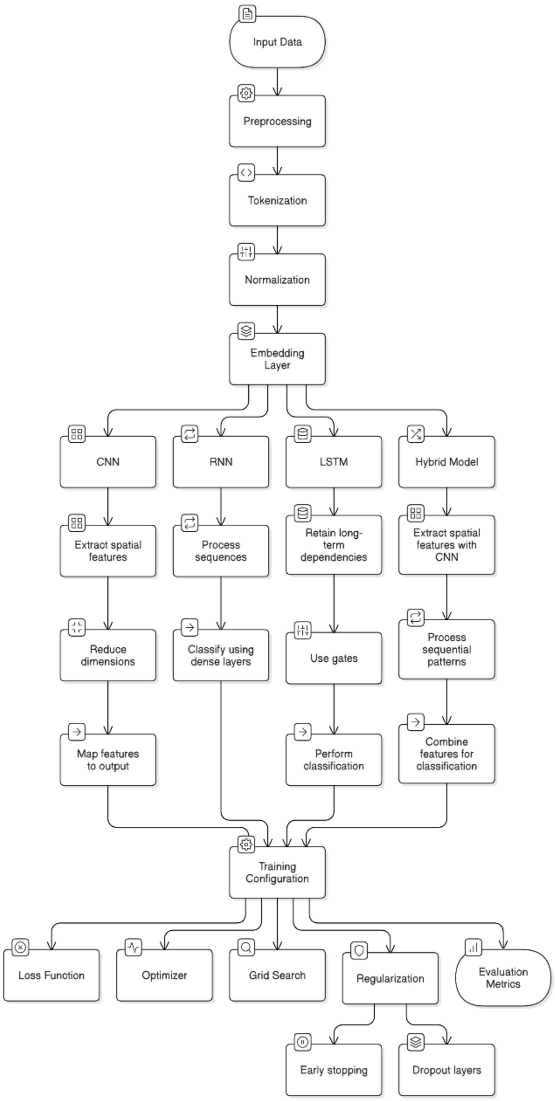

Figure 1. Overall methodology pipeline overview.

2.3. Dataset

This research utilizes a custom-curated dataset of Sindhi newspaper articles, containing 6156 articles.[1] Distributed across three categories: entertainment, sports, and technology. The dataset was prepared through meticulous preprocessing, including tokenization, normalization, and the removal of extraneous content (Fig 2). This ensured the input text was suitable for effective model training and evaluation.

[1] https://www.kaggle.com/datasets/owaisraza009/sindhi-articles-dataset-from-daily-kawish

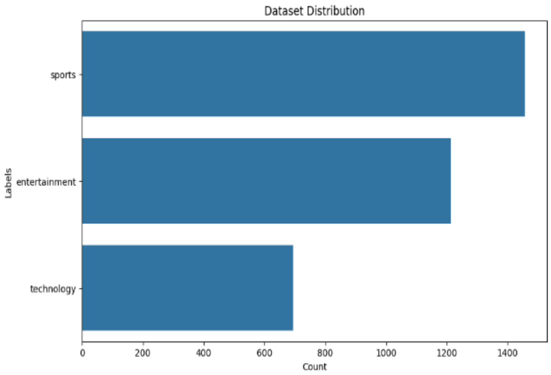

Figure 2. Visual representation of dataset distribution.

2.3. Models

- **Convolutional Neural Networks (CNNs):** Utilizes convolutional layers to extract spatial features from the text, followed by dense layers for classification. Its ability to capture localized patterns makes it effective for text data.

- **Recurrent Neural Networks (RNNs):** Employ sequential layers to model temporal dependencies. However, the architecture is prone to vanishing gradient issues, limiting its performance.

- **Long Short-Term Memory Networks (LSTMs):** Enhances sequential modeling by incorporating memory cells that retain information over longer sequences, overcoming RNN limitations.

- **Hybrid Model:** Combines the strengths of CNN and LSTM, leveraging spatial and sequential feature extraction for enhanced classification accuracy

2.4. Training and Evaluation

The models were trained using the categorical cross-entropy loss function and optimized with the Adam optimizer. The dataset was split into 80% training and 20% testing data. Early stopping was employed to mitigate overfitting, and hyperparameters, including batch size and learning rate, were optimized through grid search. Performance metrics such as precision, recall, F1-score, and accuracy were computed to compare the models comprehensively. The implementation leveraged TensorFlow and Keras frameworks, ensuring robustness and reproducibility.

3. Evaluation and Results

To evaluate model performance, precision, recall, F1-score, and accuracy metrics were calculated for each category—entertainment, sports, and technology—as well as overall averages. The models were tested on the 20% held-out testing dataset, ensuring the evaluation reflected generalizability to unseen data. Metrics were computed using standard Python libraries, including Scikit-learn, to ensure consistency and reproducibility.

The evaluation process revealed distinct strengths and limitations of the models. While CNN and Hybrid consistently outperformed others in capturing spatial and sequential patterns, RNN showed limitations in managing long-term dependencies. LSTM's robust handling of sequential data placed it close to the top-performing models. The detailed results are discussed in the subsequent section.

The experimental results demonstrate the comparative effectiveness of the models for multiclass classification of Sindhi newspaper articles:

3.1. CNN Model

Achieved an accuracy of 96%, with strong precision and recall across all categories (Fig 3, 4).

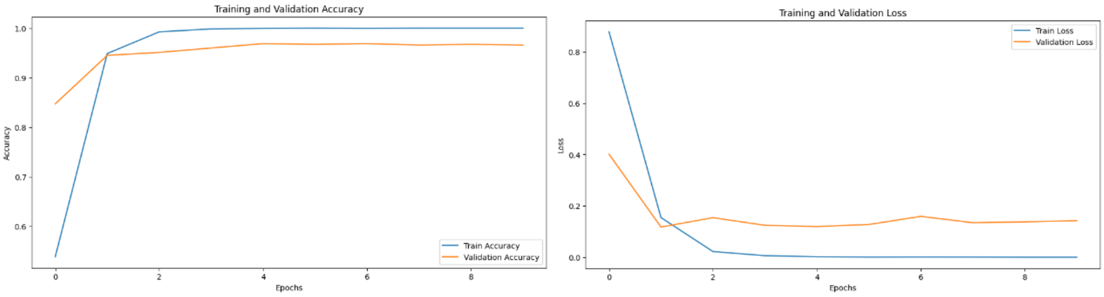

Figure 3. CNN model training and validation metrics: (a) Accuracy (b) Loss Over Epochs.

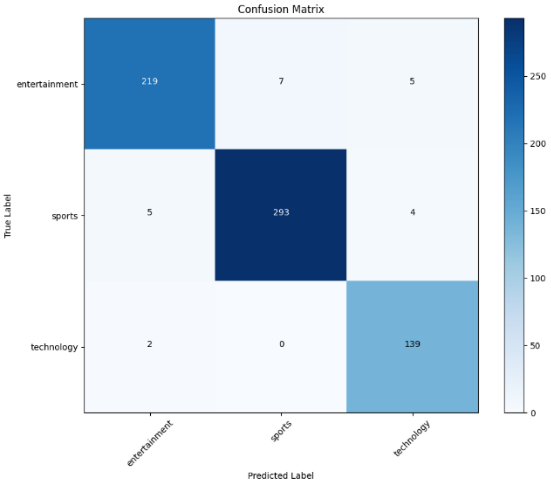

Figure 4. Visualization of a confusion matrix for CNN model evaluation.

3.2. RNN Model

Delivered a comparatively low accuracy of 67%, highlighting challenges with sequence learning (Fig 5, 6).

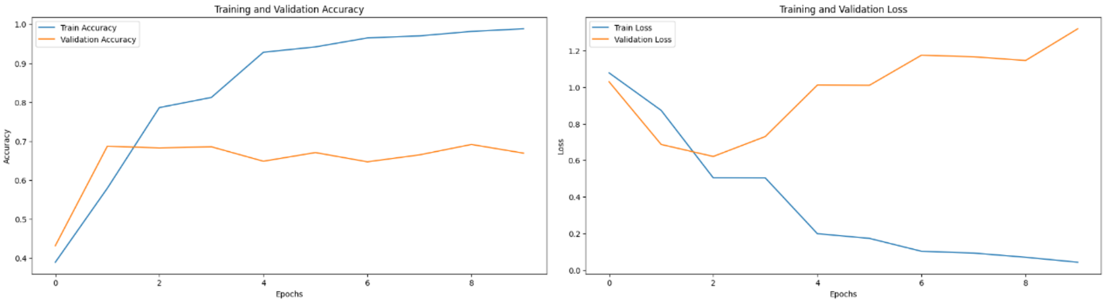

Figure 5. RNN model training and validation metrics: (a) Accuracy (b) Loss Over Epochs.

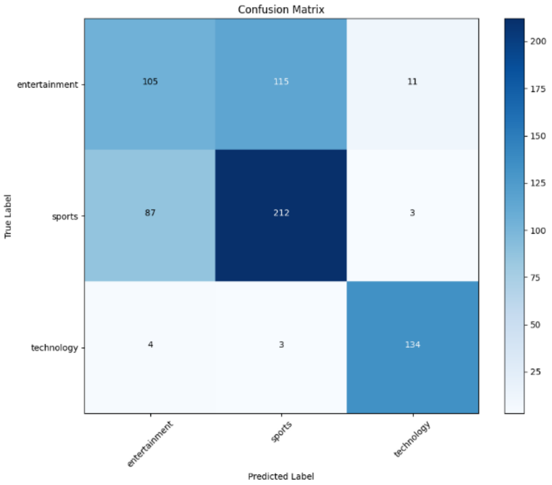

Figure 6. Visualization of a confusion matrix for RNN model evaluation.

3.3. LSTM Model

Performed commendably with an accuracy of 95.85%, effectively capturing sequential features (Fig 7, 8).

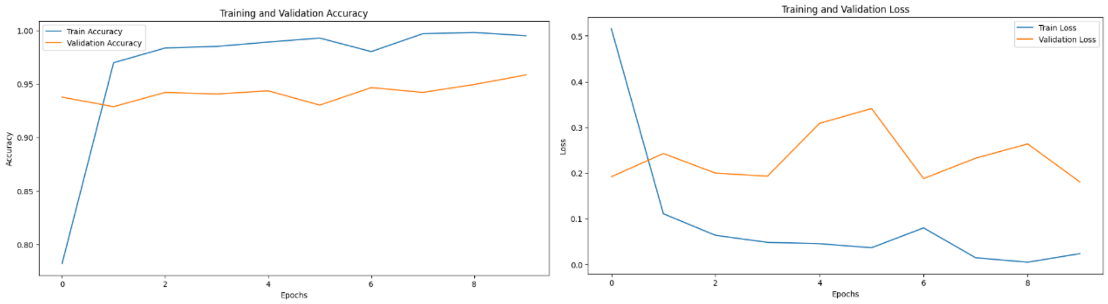

Figure 7. LSTM model training and validation metrics: (a) Accuracy (b) Loss Over Epochs.

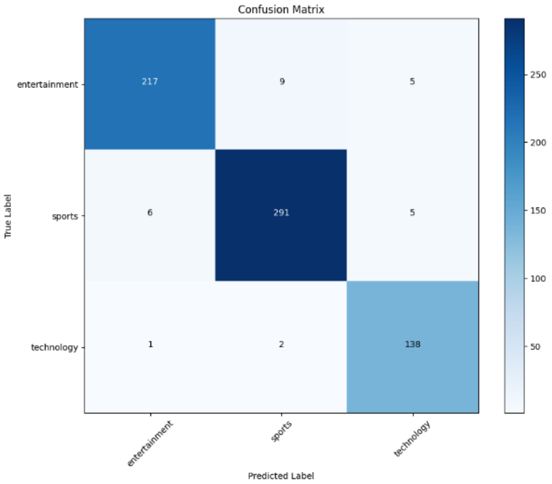

Figure 8. Visualization of a confusion matrix for LSTM model evaluation.

3.4. Hybrid Model

Matched CNN’s top performance, achieving 96% accuracy, showcasing the advantages of hybrid architectures(Fig 9, 10).

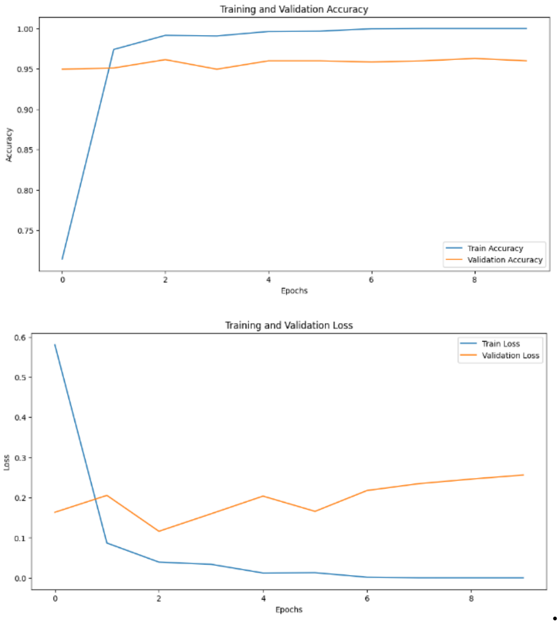

Figure 9. Hybrid model training and validation metrics: (a) Accuracy (b) Loss Over Epochs.

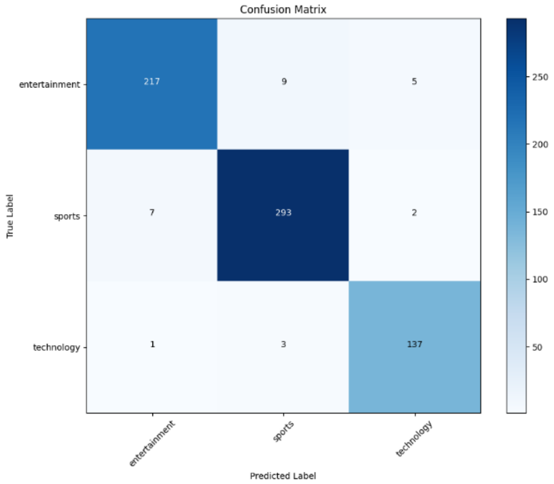

Figure 10. Visualization of a confusion matrix for Hybrid model evaluation.

The results underscore the importance of architecture choice in text classification tasks. CNN and Hybrid excelled due to their ability to extract complex patterns, while RNN highlighted the need for enhanced sequence handling. LSTM demonstrated near-competitive performance, affirming its utility in handling sequential data(Table 1). Future work could explore transformer-based models for even greater performance gains.

Table 1. Summarizes the evaluation metrics:.

Model	Precision	Recall	F1-Score	Accuracy
CNN	0.96	0.96	0.96	96%
RNN	0.67	0.67	0.67	67%
LSTM	0.95	0.96	0.96	95.85%
Hybrid	0.96	0.96	0.96	96%

4. Discussion

This study highlights the potential of deep learning models in addressing the challenges of Sindhi newspaper article classification. The high accuracy achieved by CNN and Hybrid models underscores their robustness in capturing both spatial and sequential patterns, critical for understanding the linguistic nuances of Sindhi text. LSTM's near-equivalent performance further confirms its suitability for sequential data processing. However, the comparatively low performance of the RNN model reveals the limitations of simpler sequential architectures, particularly in handling complex text data.

The findings emphasize the significance of data preprocessing, architecture selection, and hyperparameter tuning in achieving superior classification performance. This research not only contributes to Sindhi NLP but also serves as a framework for other low-resource languages. Future studies could integrate transformer-based models and explore larger datasets to push the boundaries of accuracy and efficiency in multiclass text classification.

5. Conclusions

This research successfully demonstrated the application of deep learning models for the multiclass classification of Sindhi newspaper articles. The CNN and Hybrid models achieved the highest accuracy of 96%, showcasing their efficacy in handling regional languages. LSTM also delivered robust performance, while RNN's limitations highlighted the need for architectural advancements. These findings contribute to Sindhi NLP and offer insights for similar applications in low-resource languages. Future work could incorporate transformer architectures and larger datasets to further enhance classification performance.

References

1. Ali, M., & Imdad, A. (2017). Sentiment Summarization and Analysis of Sindhi Text. International Journal of Advanced Computer Science and Applications, 8(10). https://doi.org/10.14569/ijacsa.2017.081038 Aliyu, Y.,
2. Sarlan, A., Danyaro, K. U., Rahman, A. S. B. A., & Abdullahi, M. (2024). Sentiment Analysis in Low-Resource Settings: A Comprehensive Review of Approaches, Languages, and Data Sources. IEEE Access, 12, 66883–66909. https://doi.org/10.1109/access.2024.3398635
3. Anitha, S., Varshini, E. K., Mahalakshmi, N. H., & Jishnu, S. (2024). Optimizing Multi-Class Text Classification Models for Imbalanced News Data. 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), 1–6. https://doi.org/10.1109/icccnt61001.2024.10724277
4. Cruz, J. C. B., & Cheng, C. (2020). Establishing Baselines for Text Classification in Low-Resource Languages. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2005.02068
5. Fesseha, A., Xiong, S., Emiru, E. D., Diallo, M., & Dahou, A. (2021). Text Classification Based on Convolutional Neural Networks and Word Embedding for Low-Resource Languages: Tigrinya. Information, 12(2), 52. https://doi.org/10.3390/info12020052
6. Ilyas, A., Obaid, S., & Bawany, N. Z. (2021). Multilevel Classification of Pakistani News using Machine Learning. 2021 22nd International Arab Conference on Information Technology (ACIT), 1–5. https://doi.org/10.1109/acit53391.2021.9677431
7. Li, X., Li, Z., Sheng, J., & Slamu, W. (2020). Low-Resource Text Classification via Cross-Lingual Language Model Fine-Tuning. In Lecture notes in computer science (pp. 231–246). https://doi.org/10.1007/978-3-030-63031-7_17
8. Magueresse, A., Carles, V., & Heetderks, E. (2020). Low-resource Languages: A Review of Past Work and Future Challenges. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2006.07264
9. Maheen, S. M., Faisal, M. R., & Karim, M. R. R. a. M. S. (2022, January 1). Alternative non-BERT model choices for the textual classification in low-resource languages and environments. In Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing. https://doi.org/10.18653/v1/2022.deeplo-1.20
10. Marreddy, M., Oota, S. R., Vakada, L. S., Chinni, V. C., & Mamidi, R. (2022). Multi-Task Text Classification using Graph Convolutional Networks for Large-Scale Low Resource Language. 2022 International Joint Conference on Neural Networks (IJCNN), 1–8. https://doi.org/10.1109/ijcnn55064.2022.9892105

11. Nawaz, A., Nawaz, M., Shaikh, N. A., Rajper, S., Baber, J., & Khalid, M. (2023). TPTS: Text pre-processing Techniques for Sindhi Language. Pakistan Journal of Emerging Science and Technologies (PJEST), 4(3), 1–12. https://doi.org/10.58619/pjest.v4i3.89
12. Ombabi, A. H., Ouarda, W., & Alimi, A. M. (2020). Deep learning CNN–LSTM framework for Arabic sentiment analysis using textual information shared in social networks. Social Network Analysis and Mining, 10(1). https://doi.org/10.1007/s13278-020-00668-1
13. Pyysalo, S., Kanerva, J., Virtanen, A., & Ginter, F. (2020). WikiBERT models: deep transfer learning for many languages. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2006.01538
14. Rajan, A., & Salgaonkar, A. (2021). Survey of NLP Resources in Low-Resource Languages Nepali, Sindhi and Konkani. In Lecture notes in networks and systems (pp. 121–132). https://doi.org/10.1007/978-981-16-0739-4_12
15. Setu, J. H., Halder, N., Sikder, S., Islam, A., & Alam, M. Z. (2024). Empowering Multiclass Classification and Data Augmentation of Arabic News Articles Through Transformer Model. 2022 International Joint Conference on Neural Networks (IJCNN), 101, 1–7. https://doi.org/10.1109/ijcnn60899.2024.10650716
16. Sindhi Articles Dataset From Daily Kawish. (2021). [Dataset]. In Kaggle. https://www.kaggle.com/datasets/owaisraza009/sindhi-articles-dataset-from-daily-kawish
17. Song, Y., Liu, X., & Zhou, Z. (2024). A Comprehensive Review of Text Classification Algorithms. Journal of Electronics and Information Science, 9(2). https://doi.org/10.23977/jeis.2024.090205
18. Soomro, S. A., Yuhaniz, S. S., Dootio, M. A., Murtaza, G., & Mughal, M. H. (2024). A Systematic Review on Sentiment Analysis for Sindhi Text. Baghdad Science Journal. https://doi.org/10.21123/bsj.2024.10954
19. Vavekanand, R., Sam, K., Kumar, S., & Kumar, T. (2024). CardiacNet: A Neural Networks Based Heartbeat Classifications using ECG Signals . Studies in Medical and Health Sciences, 1(2), 1–17. https://doi.org/10.48185/smhs.v1i2.1188
20. Wadud, M. a. H., Mridha, M. F., Shin, J., Nur, K., & Saha, A. K. (2022). Deep-BERT: Transfer Learning for Classifying Multilingual Offensive Texts on Social Media. Computer Systems Science and Engineering, 44(2), 1775–1791. https://doi.org/10.32604/csse.2023.027841
21. Yohannes, H. M., & Amagasa, T. (2022). A Scheme for News Article Classification in a Low-Resource Language. In Lecture notes in computer science (pp. 519–530). https://doi.org/10.1007/978-3-031-21047-1_47

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.