|  |  |  |  |
| --- | --- | --- | --- |
| Title | Strategies for enhancing the performance of news article classification in Bangla: Handling imbalance and interpretation | | |
| Author(s) Name | Khan Md Hasib, Nurul Akter Towhid, Kazi Omar Faruk, Jubayer Al Mahmud, MF Mridha | | |
| Contact Email(s) | firoz.mridha@aiub.edu | | |
| Published Journal Name | Engineering Applications of Artificial Intelligence | | |
| Type of Publication | Journal | | |
| Volume | 125 | Issue |  |
| Publisher | Pergamon | | |
| Publication Date | 2023/10/1 | | |
| ISSN |  | | |
| DOI | https://doi.org/10.1016/j.engappai.2023.106688 | | |
| URL | https://www.sciencedirect.com/science/article/abs/pii/S0952197623008722 | | |
| Other Related Info. |  | | |
|  | | | |

|  |  |
| --- | --- |
| Abstract |  |
| The rapid increase in obtainable online text data has made text categorization an important tool for data analysts to extract relevant information on the web. However, incorrect or incomplete classification of marginalized groups may result from using biased text data. In order to remedy the disparity in available data, this research suggests a system for classifying and analyzing Bangla news articles. The suggested approach first uses both Random Under-Sampling (RUS) and Synthetic Minority Oversampling Techniques to balance the massive unbalanced Bangla News dataset consisting of 4,37,948 instances (SMOTE). Secondly, the proposed system employs three machine learning models: Logistic Regression, Decision Tree, and Stochastic Gradient Descent along with three deep learning models: Artificial Neural Network (ANN), Convolutional Neural Network (CNN), and Bidirectional Encoder Representations from Transformers (BERT) for Bangla text categorization. The experimental results signify the superior performance of BERT to other classification models of the system as well as other existing methods in this domain. The proposed system achieves the maximum accuracy of 99.04% in balanced dataset and 72.23% in imbalanced dataset using BERT. K-fold cross validation with varied K values is used to determine the performance consistency of BERT. Finally, both LIME (Local Interpretable Model agnostic Explanations and SHAP (SHapley Additive exPlanations) techniques are applied for interpreting each prediction made by BERT. | |