2024 3rd International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE 2024), 25-27 April, 2024, Gazipur, Bangladesh

2024 3rd International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE) | 979-8-3503-8828-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICAEEE62219.2024.10561678

Bangla Sentiment Analysis On Highly Imbalanced Data Using Hybrid CNN-LSTM & Bangla BERT

Fahmida Khanam
Computer Science and Engineering
Bangladesh University
Dhaka, Bangladesh
fahmidakhannisa@gmail.com

Anik Chakraborty
Computer Science and Engineering
Bangladesh University
Dhaka, Bangladesh
arunabhoanik@gmail.com

Md. Ahsan Habib
Computer Science and Engineering
Bangladesh University
Dhaka, Bangladesh
ahsan.habib@bu.edu.bd

Md. Sadiq Iqbal
Computer Science and Engineering
Bangladesh University
Dhaka, Bangladesh
sadiq.iqbal@bu.edu.bd

***Abstract*—Sentiment analysis is a technique that combines machine learning and natural language processing to identify the emotional attitude of a text. This is a very active research area in recent years. Bengali is the fifth most spoken Indo-European language in the world. Many people in Bangladesh use news portals and social media to gather information on various topics. We used a publicly available dataset from Kaggle. This data set consists of more negative reviews than positive reviews. We try to experiment with this dataset with different models, such as traditional ML models and deep learning models like CNN, LSTM, and the transformer model (Bangla-BERT-base). The Bangla-BERT-base achieved a notable 96% accuracy through 10-fold cross-validation. Several other performance measures are also used to evaluate our model.**

***Index Terms*—Sentiment Analysis, Preprocessing, Cross Validation (K-Fold), Hybrid CNN-LSTM, SMOTE, BERT, Text Classification, Word Embedding, Deep Learning.**

I. Introduction

Sentiment analysis(SA) is a popular research topic in Natural Language Processing(NLP) [1]. Most of the major approaches to sentiment analysis are based on machine learning techniques [2], [3]. This is a field study to determine the polarity and intensity of text containing human emotions, sentiments, and views regarding an entity [4]. Entities can be products, subjects, individuals, and services [5], [6]. Sentiment Analysis is also known as Opinion Mining, but some researchers say they are different because Opinion Mining extracts and analyzes people's opinions, while Sentiment Analysis identifies sentiment in the text, then analyzes [7]. Sentiment analysis emerged as an analytical and predictive process [8]. Sentiments of the Internet, the repository of information where social networks, websites, web forums, and blogs are the means by which people express their opinions. Anyone can collect huge amounts of data from them. In the age of technology, correct information is worth more than millions of tons of gold. Since we are working on Bangla sentiment analysis, collecting various Bangali data is not such an easy task. In order to improve the quality of the data, it is important to be able to analyze how people feel about it. A machine trained by labeled data will give it an edge for future analysis [9]. Many works in this field, specifically Bengali texts, criticisms, and songs, have been made in the last few years. Our research is enhanced by approaching it with a combined technique of deep learning. The benefit of our research to publishers will be to reflect on their work when it comes to making the right decisions or publishing quality information to audiences based on analytics. The fact that this method is real and can purify more things shows that it is useful. Extracting text from different websites, classifying these texts, and polarizing them will give the best guess. Classification beyond polarization is also a key factor.

Multiple studies have shown that using rule-based machine learning techniques yields average accuracy, but deep learning techniques stand out when it comes to the highest accuracy and refinement [10] [11]. Research demonstrates that as recent datasets get more diverse, these deeper networks perform better. Our main motivation is therefore to apply a strong combination of machine learning and deep learning architectures to our work.

In this study, we systematically investigated diverse methodologies to tackle sentiment analysis on imbalanced data. Our exploration included traditional models such as SVM, Multinomial Naïve Bayes, K-NN, Logistic Regression, and Decision Tree, each incorporating SMOTE oversampling. Additionally, we introduced a hybrid CNN-LSTM model that leverages the strengths of both, where CNN focuses on feature learning through pooling, and LSTM captures contextual information. Recognizing the challenges posed by imbalanced data, we fine-tuned the Bangla-bert-base and introduced a BERT-CNN-LSTM hybrid, combining the contextual understanding of BERT with the complementary capabilities of CNN and LSTM. This comprehensive approach aims to yield improved results in the context of sentiment analysis on imbalanced

979-8-3503-8828-2/24/$31.00 ©2024 IEEE

datasets.

Our paper's structure is as follows: Section II reviews Bangla sentiment analysis, Section III covers preprocessing and our dataset, it discusses our proposed model and hyper-parameters, Section IV presents findings of our model, and Section V concludes with future directions.

II. Related Works

A significant number of research works have been done within this scope. Despite these developments in English Sentiment analysis, Bengali Sentiment analysis has received relatively little attention.

In the realm of Bangla sentiment analysis, several datasets have been curated to facilitate research. Hasmot [12] introduced "BanglaSenti," a lexicon-based dataset comprising 61,582 Bangla words. Another comprehensive dataset, "SentiGOLD," was developed by multiple authors [13], encompassing 70,000 samples labeled across five classes from 30 domains.

Various studies have leveraged deep learning and machine learning techniques in Bangla sentiment analysis. In [14], sentiment analysis on KN95 mask reviews was conducted using TF-IDF vectorization and classifiers (Support Vector Machine, Gaussian Naïve Bayes, and Multinomial Naïve Bayes), with Gaussian Naïve Bayes demonstrating superior accuracy, recall, and F1-score. Paper [15] introduced a five-layered GRU neural network model surpassing the state-of-the-art Bidirectional LSTM (BLSTM) result. The research by author Naimul Hossain [16] has accrued 94.22% accuracy using a combined CNN-LSTM model. They used the Bangla w2v model for word embedding and trained the model using a lower learning rate. When their validation loss wasn't improving, they used an early stopping callback to stop the training.

In [17], employing Glove word embedding and CNN-based classifiers, achieving an impressive 99.43% accuracy. Additionally, [18]presents a novel CNN and LSTM-based classifier for multi-class sentiment analysis of Bengali social media comments, achieving 85.8% accuracy and 0.86 F1 scores.

The utilization of transformer-based models is prevalent in recent works. In [19], abusive Bangla comments on Facebook were swiftly identified using BERT and ELECTRA, yielding notable test accuracies of 85.00% and 84.92%, respectively. Author Bhowmik [20] introduced deep learning models for Bangla sentiment analysis, incorporating an extended lexicon data dictionary and employing rule-based sentiment score generation. Their experiment showcased high accuracy, with the proposed BERT-LSTM model achieving 84.18%.

Author Khondoker [21]introduced 2 classes and 3 classes of datasets, both manually tagged. They used the BERT model with end-to-end deep learning models like GRU, LSTM, and CNN. They compared the performance with other state of the art embedding models such as Word2Vec and fastText. From their research, they got 71% accuracy for 2 class classification and 68% for 3 class.

These diverse contributions collectively advance the field of Bangla sentiment analysis and provide valuable insights for future research.

III. Methodology

We meticulously selected a series of baseline methods to thoroughly evaluate our dataset. In this section, we delve into the intricate details of our implementation and model architecture. The Fig. 1 gives an overview of our model.

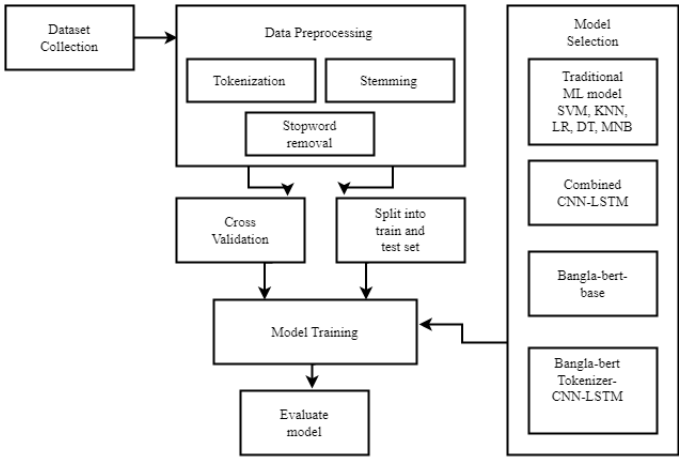

Fig. 1. Flowchart of the sequential explanatory mixed methods design.

A. Dataset

In our paper, we used publicly available dataset named as 'Bengali News Headline Sentiment[1]' from Kaggle. This dataset, encompassing a variety of Bangla reviews and texts, features sentiments labeled as '1' for positive and '0' for negative. Each sentence in our dataset has been manually labeled to ensure precise sentiment annotations. The dataset comprises 2757 reviews in total, with 754 labeled as positive and 2,003 as negative. Given the inherent imbalance, we applied an oversampling technique, resulting in a dataset with a total of 4,006 samples

B. Data Preprocessing

Our initial endeavor involved preprocessing the data, and transforming raw information into a clean and structured format suitable for model training. Our preprocessing tasks encompassed stop word removal, stemming, and punctuation elimination. Stop word removal in Bangla was a crucial preprocessing step aimed at eliminating common words that frequently occur but hold little substantive meaning in the text data. We used the BanglaStemmer, a specific stemming algorithm to stem the words, which reduces words to their base form by removing affixes and capturing the core meaning [22]. Lastly, we eliminated punctuation marks, digits, URLs, and emoticons from the text. While stopword removal and stemming were crucial for traditional models, for the BERT model,

[1]https://www.kaggle.com/datasets/kaisermasum24/bengali-news-headline-sentiment

we specifically retained stopwords and avoided stemming to preserve context, we only removed URLs.

TABLE I
PREPROCESSING A SAMPLE FROM THE DATABASE.

Preprocessing	Text
Original sentence	জরিমানা করা হউক। ৩ মাসের বেতন কর্তন।(Let there be a penalty.Deduct 3 months' salary.)
Removing stopwords	জরিমানা হউক। ৩ মাসের বেতন কর্তন।
Stemming bengali words	জরিমানা হউক। ৩ মাস বেতন কর্তন।
Removing unnecessary punctuation and number	জরিমানা হউক মাস বেতন কর্তন

For our transformer model, Bangla BERT, we retained the original context of the text without applying traditional preprocessing techniques. The paper [23]shows that transformer models, by design, are adept at capturing contextual information, making certain preprocessing steps unnecessary.

C. Baseline model & setup

In this section, we present implementation details of our experimental setup. In our paper we have employed five traditional machine learning algorithms. They are:1)Support Vector Machine (SVM), 2)K-Nearest Neighbors (KNN), 3)Multinomial Naive Bayes (MNB), 4)Decision Tree and 5)Logistic Regression. We also use two deep learning methods. Such As: 1)A hybrid CNN-LSTM model and 2)The integration of the BERT model(Bangla-BERT).

For the hybrid CNN-LSTM model, we employed a combination of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) layers. The process involves tokenizing words using a vocabulary size of 892 and padding sequences to a maximum length of 59. The model consists of sequential layers, beginning with an Embedding layer. It employs 32 CNN filters with a kernel size of 3 and ReLU activation, followed by max-pooling. The output is fed into an LSTM layer with 100 hidden units, supplemented by a dropout rate of 0.45 for regularization. An additional Bidirectional LSTM layer with 64 units is used. Hyperparameter optimization is achieved using Optuna [24]. Finally, a dense layer with a sigmoid activation function accommodates the binary classification task.

In addition to the traditional ML algorithms and the hybrid CNN-LSTM model, we also experimented with the BERT model. We utilized the "bangla-bert-base" model and its corresponding tokenizer provided by the "sagorsarker[2]" library [25]. Tokenization involved padding and truncating sequences to a maximum length of 128 tokens. Fine-tuning of the BERT model utilized an Adam optimizer with a learning rate of 1e-5 and Sparse Categorical Cross-Entropy loss. Training occurred over 10 epochs with a batch size of 16. The model was compiled, trained, and evaluated on the test set. Similarly, we implemented another method BERT-CNN-LSTM, that utilizes the Bert tokenizer provided by the "sagorsarker" library and used our proposed hybrid CNN-LSTM model to classify the sentiment.

To tackle the challenge of imbalanced data, we introduced SMOTE (Synthetic Minority Over-sampling Technique). It mitigates bias toward the majority class [26]. It creates synthetic samples by randomly selecting one of the k nearest neighbors and creating a new instance along the line connecting the selected instance and the chosen neighbor. SMOTE is used to oversample the training data after it has been tokenized and before training the model.

The dataset underwent a split into training and testing sets using an 80-20 split for all ML models, including CNN-LSTM, Bangla-BERT-base, and BERT-CNN-LSTM. In addition, we employed stratified k-fold with 5 and 10 splits for the hybrid CNN-LSTM model, Bangla BERT, and BERT-CNN-LSTM.

TABLE II
DATA BALANCING TECHNIQUE SMOTE OVERSAMPLING PERFORMANCE WITH ML AND DL MODELS IN HOLD OUT METHOD

Method		Accuracy	Precision	Recall	F1
SVM	Without SMOTE	0.72	0.50	0.19	0.28
	SMOTE	0.75	0.71	0.84	0.77
KNN	Without SMOTE	0.72	0.51	0.17	0.26
	SMOTE	0.55	0.52	0.96	0.68
MNB	Without SMOTE	0.73	0.54	0.18	0.27
	SMOTE	0.53	0.54	0.44	0.48
LR	Without SMOTE	0.72	0.48	0.11	0.18
	SMOTE	0.75	0.71	0.83	0.77
DT	Without SMOTE	0.69	0.43	0.39	0.41
	SMOTE	0.69	0.66	0.80	0.72
CNN-LSTM	Without SMOTE	0.69	0.23	0.50	0.31
	SMOTE	0.62	0.40	0.73	0.51
Bangla-bert-base	Without SMOTE	0.90	0.83	0.89	0.85
	SMOTE	0.91	0.86	0.80	0.82
BERT-CNN-LSTM	Without SMOTE	0.79	0.72	0.74	0.78
	SMOTE	0.78	0.70	0.70	0.75

IV. RESULT AND DISCUSSION

In this study, we investigate the impact of data balancing techniques on the performance of both traditional machine learning (ML) models and sophisticated deep learning models. We examine the influence of the Synthetic Minority Oversampling Technique (SMOTE) oversampling on an imbalanced dataset. Table II shows how SMOTE impacts performance on our imbalanced dataset. Our experimentation with SMOTE oversampling reveals an interesting dichotomy between traditional ML models and deep learning models. The f1 score, a robust metric for imbalanced datasets, demonstrates noticeable improvement across all traditional ML models. This is due to the inherent bias of traditional models towards the majority

[2]https://huggingface.co/sagorsarker/bangla-bert-base

class. Among these, the Support Vector Machine (SVM) emerges as the top performer, achieving an impressive f1 score of 77.20% and an accuracy of 75%.

On the other hand, deep learning models such as CNN-LSTM and BERT exhibit a distinctive behavior. Unlike traditional models, they exhibit the ability to automatically learn complex hierarchical features from the data [27]. Because of this, the effect of SMOTE oversampling is not as noticeable when using these models with unbalanced datasets. Table II shows how our model performs suing SMOTE.

Utilizing 5-fold and 10-fold cross-validation, our combined CNN-LSTM model demonstrates substantial improvement, achieving an average accuracy of 74%. The performance of the Bangla BERT model is particularly noteworthy. Fine-tuning the model with 10-fold cross-validation results in a remarkable average accuracy of 96%, underscoring its effectiveness in handling complex language understanding tasks. The standard deviation accuracy score for the 10-fold bangla-bert-base model is 6.4%, which indicates that there is some variability in its performance for different folds but the model's accuracy stays consistent across these folds. The BERT-CNN-LSTM model also performs well. It got an average accuracy of 77%. Table III shows the mean and IV standard deviation score of deep learning models.

TABLE III
THE MEAN EVALUATION SCORE OF CNN-LSTM, BANGLA-BERT-BASE, BERT-CNN-LSTM MODELS WITH 10 AND 5 FOLD CROSS-VALIDATION .

K-value	Method	F1	Accuracy	ROC AUC
10	CNN-LSTM	0.40	0.68	0.59
	Bangla-bert-base	0.90	0.96	0.93
	BERT-CNN-LSTM	0.70	0.77	0.79
5	CNN-LSTM	0.52	0.74	0.67
	Bangla bert-base	0.82	0.91	0.88
	BERT-CNN-LSTM	0.71	0.77	0.79

TABLE IV
THE STANDARD DEVIATION SCORE OF CNN-LSTM, BANGLA-BERT-BASE, BERT-CNN-LSTM MODELS WITH 10 AND 5 FOLD CROSS-VALIDATION .

K-value	Method	F1	Accuracy	ROC AUC
10	CNN-LSTM	0.038	1.667	0.025
	Bangla-bert-base	0.885	0.063	0.103
	BERT-CNN-LSTM	0.033	0.024	0.032
5	CNN-LSTM	0.031	1.375	0.020
	Bangla bert-base	0.158	0.083	0.109
	BERT-CNN-LSTM	0.034	0.028	0.020

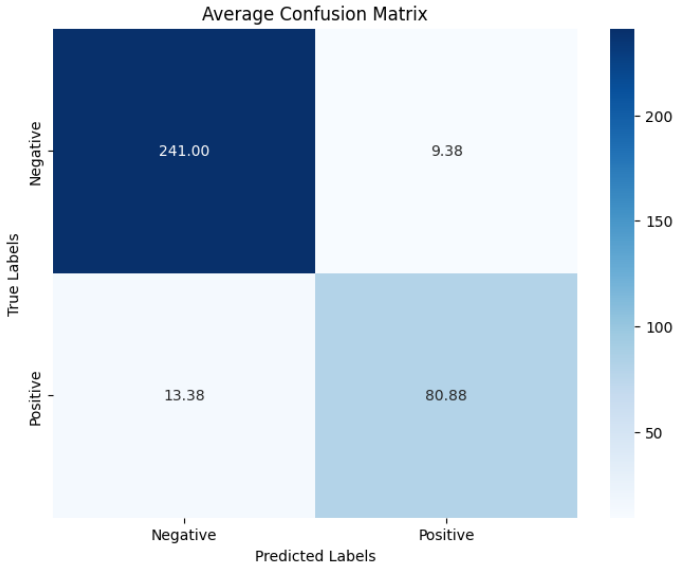

Fig. 2. Confusion matrix of bangla-bert-base with 10 fold.

In summary, our study showcases the dual impact of data balancing techniques on traditional ML models and deep learning models. While traditional models benefit from SMOTE, deep learning models inherently possess the capacity to handle imbalanced data. The performance gains achieved through 10-fold cross-validation and the exceptional performance of the Bangla BERT model highlight the potential of advanced techniques. Fig 2 shows the average confusion matrix of bangla-bert-base with 10 fold

V. CONCLUSION AND FUTURE WORK

In our study, a hybrid architecture is used to implement sentiment analysis. This is the first time we are developing this research paper based on various kinds of Bengali news & information. Although we have not achieved a hundred percent accuracy with this model, we can still give near-satisfactory accuracy when evaluating Bengali sentences. This model can also be reused for any kind of text perspective or other types of Sentiment Analysis. However, with this research, we are still working to get better results and will be regularly updated with new insights. To achieve better results, we will continue to expand the size of the SA datasets in Bengali and investigate the use of additional deep-learning models. We can also further improve emotion detection and spam detection. Future research directions could involve investigating more advanced data balancing techniques tailored to deep learning models and exploring ensemble methods to combine the strengths of various approaches.

REFERENCES

[1] Tetsuya Nasukawa and Jeonghee Yi. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 70–77, 2003.

[2] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*, 2002.

[3] Oscar Araque, Ignacio Corcuera-Platas, J Fernando Sánchez-Rada, and Carlos A Iglesias. Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, 77:236–246, 2017.

[4] Bing Liu. *Sentiment analysis and opinion mining*. Springer Nature, 2022.

[5] Kumar Ravi and Vadlamani Ravi. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-based systems*, 89:14–46, 2015.

[6] Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135, 2008.

[7] Lei Zhang and Bing Liu. Sentiment analysis and opinion mining. *Encyclopedia of machine learning and data mining*, 1:1152–1161, 2017.

[8] S Naveen Balaji, P Victer Paul, and R Saravanan. Survey on sentiment analysis based stock prediction using big data analytics. In *2017 Innovations in Power and Advanced Computing Technologies (i-PACT)*, pages 1–5. IEEE, 2017.

[9] Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780, 2022.

[10] Shumaiya Akter Shammi, Sajal Das, Narayan Ranjan Chakraborty, Sumit Kumar Banshal, and Nishu Nath. A comprehensive roadmap on bangla text-based sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–29, 2023.

[11] Nayan Banik, Md Hasan Hafizur Rahman, Shima Chakraborty, Hanif Seddiqui, and Muhammad Anwarul Azim. Survey on text-based sentiment analysis of bengali language. In *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, pages 1–6. IEEE, 2019.

[12] Hasmot Ali, Md Fahad Hossain, Shaon Bhatta Shuvo, and Ahmed Al Marouf. Banglasenti: A dataset of bangla words for sentiment analysis. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–4. IEEE, 2020.

[13] Md Ekramul Islam, Labib Chowdhury, Faisal Ahamed Khan, Shazzad Hossain, Md Sourave Hossain, Mohammad Mamun Or Rashid, Nabeel Mohammed, and Mohammad Ruhul Amin. Sentigold: A large bangla gold standard multi-domain sentiment analysis dataset and its evaluation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4207–4218, 2023.

[14] Dipta Roy Karmakar, Shirina Akter Mukta, Busrat Jahan, and Jony Karmakar. Sentiment analysis of customers' review in bangla using machine learning approaches. In *Innovations in Computer Science and Engineering: Proceedings of the Ninth ICICSE, 2021*, pages 373–384. Springer, 2022.

[15] Nasif Alvi, Kamrul Hasan Talukder, and Abdul Hasib Uddin. Sentiment analysis of bangla text using gated recurrent neural network. In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021, Volume 2*, pages 77–86. Springer, 2022.

[16] Naimul Hossain, Md Rafiuzzaman Bhuiyan, Zerin Nasrin Tumpa, and Syed Akhter Hossain. Sentiment analysis of restaurant reviews using combined cnn-lstm. In *2020 11th International conference on computing, communication and networking technologies (ICCCNT)*, pages 1–5. IEEE, 2020.

[17] Md Shihab Mahmud, Md Touhidul Islam, Afrin Jaman Bonny, Rokeya Khatun Shorna, Jasia Hossain Omi, and Md Sadekur Rahman. Deep learning based sentiment analysis from bangla text using glove word embedding along with convolutional neural network. In *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE, 2022.

[18] Rezaul Haque, Naimul Islam, Mayisha Tasneem, and Amit Kumar Das. Multi-class sentiment classification on bengali social media comments using machine learning. *International Journal of Cognitive Computing in Engineering*, 4:21–35, 2023.

[19] Tanjim Taharat Aurpa, Rifat Sadik, and Md Shoaib Ahmed. Abusive bangla comments detection on facebook using transformer-based deep learning models. *Social Network Analysis and Mining*, 12(1):24, 2022.

[20] Nitish Ranjan Bhowmik, Mohammad Arifuzzaman, and M Rubaiyat Hossain Mondal. Sentiment analysis on bangla text using extended lexicon dictionary and deep learning algorithms. *Array*, 13:100123, 2022.

[21] Khondoker Ittehadul Islam, Md Saiful Islam, and Md Ruhul Amin. Sentiment analysis in bengali via transfer learning using multi-lingual bert. In *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pages 1–5. IEEE, 2020.

[22] Tapashee Tabassum Urmi, Jasmine Jahan Jammy, and Sabir Ismail. A corpus based unsupervised bangla word stemming using n-gram language model. In *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, pages 824–828. IEEE, 2016.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[24] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

[25] Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Kazi Samin, Md Saiful Islam, Anindya Iqbal, M Sohel Rahman, and Rifat Shahriyar. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. *arXiv preprint arXiv:2101.00204*, 2021.

[26] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[27] Maryam M Najafabadi, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall Wald, and Edin Muharemagic. Deep learning applications and challenges in big data analytics. *Journal of big data*, 2(1):1–21, 2015.