JURNAL INFOTEL
VOL. 16, NO. 4, NOVEMBER 2024, PP. 729–739. DOI:10.20895/INFOTEL.V16I3.1224

RESEARCH ARTICLE

A Random Oversampling and BERT-based Model Approach for Handling Imbalanced Data in Essay Answer Correction

Dian Ahkam Sani[1,*] and M. Zoqi Sarwani[2]

[1,2]Department of Informatics, Merdeka University, Pasuruan 67129, Indonesia

*Corresponding email: dian.ahkam@unmerpas.ac.id

Received: August 08, 2024; Revised: September 11, 2024; Accepted: December 04, 2024.

Abstract: Automated essay scoring has long been plagued by the challenge of imbalanced datasets, where the distribution of scores or labels is skewed towards certain categories. This imbalance can lead to poor performance of machine learning models, as they tend to be biased towards the majority class. One potential solution to this problem is to use oversampling techniques that aim to balance the dataset by increasing the representation of the minority class. In this paper, we propose a novel approach that combines Random Oversampling (ROS) with a Bidirectional Encoder Representations from Transformers (BERT) base uncased model for essay answer correction. This research explores various scenarios of text pre-processing techniques to optimize model accuracy. Using a dataset of essay answers obtained from eighth-grade middle school students in the Indonesian language, our approach demonstrates good performance in terms of precision, recall, F1-Score and accuracy compared to traditional methods such as Backpropagation Neural Network, Naïve Bayes and Random Forest Classifier using FastText word embedding with Wikipedia 300 vector size pre-trained model. The best performance was obtained using the BERT-base uncased model with 3rd fold of the 5-fold cross-validation, 2e-5 learning rate, and a simplified pre-processing approach. By retaining punctuation, numbers, and stop words, the model achieved a precision of 94.63%, a recall of 93.77%, a F1-Score of 93.46%, and an accuracy of 94%.

Keywords: BERT-model, essay answer, imbalanced dataset, NLP, ROS

 ISSN: 2085-3688; e-ISSN: 2460-0997

1 Introduction

Natural Language Processing (NLP) has become an increasingly important field in the advancement of technology, particularly in tasks related to text analysis and understanding human language. Automated essay scoring has emerged as a significant topic within the realm of NLP [1], [2]. Automated scoring systems not only help reduce the workload for educators but also enhance consistency and efficiency in the assessment process. However, one of the primary challenges in developing these systems is data imbalance, where the class distribution within the dataset is uneven. For example, in an essay scoring dataset, there are often more high-scoring essays than low-scoring ones. This imbalance can lead to model bias [3] and potentially reduce the model's ability to provide accurate assessments [3], [4].

To address this issue, several approaches have been proposed, one of which is the Random Oversampling (ROS) technique. ROS works by balancing the class distribution by adding samples from the minority class, allowing the model to better learn from the available data [5]. This technique has proven effective in various classification applications, especially when faced with significant data imbalance problems. The use of ROS in this research is driven by the imbalance in the dataset. ROS duplicates existing instances of minority classes, avoiding the creation of synthetic data, which is a characteristic of Synthetic Minority Oversampling Technique (SMOTE). This is particularly helpful when working with textual data (such as the "Essay Answer") where creating synthetic samples might be difficult or introduce bias. Research [6] related to data imbalance using oversampling techniques compared five oversampling methods: SMOTE, SVM-SMOTE, Borderline SMOTE, K-means SMOTE, and Adaptive Synthetic Sampling (ADASYN). The performance of machine learning models such as Random Forest, SVM, KNN, AdaBoost, Logistic Regression, and Decision Tree was evaluated under these conditions. SVM with a linear kernel achieved the highest accuracy, 99.67%, on a dataset oversampled using ADASYN, and 99.57% on a dataset oversampled using SMOTE. Another research [7] applied several oversampling methods in machine learning, including ROS, ADASYN, SMOTE, and Borderline-SMOTE, to handle imbalanced data in binary classification with Naïve Bayes and SVM. Further research [8] utilized two data balancing techniques: Random Undersampling (RUS) and Random Oversampling (ROS). The results indicated that ROS was more effective in improving the accuracy of the model compared to RUS, with the highest accuracy reaching 85.55% in the "Gameplay and Story" aspect. ROS proved to enhance accuracy, precision, and recall better than RUS or imbalanced data. This study concluded that Multinomial Naïve Bayes is effective in aspect-based sentiment analysis, and ROS is the superior data balancing technique in this context.

On the other hand, the development of transformer-based models like Bidirectional Encoder Representations from Transformers (BERT) has led to significant breakthroughs in NLP tasks. BERT, with its bidirectional context understanding, can grasp deeper meanings from text, making it a powerful tool for automated essay scoring [9]. However, this model is also susceptible to biases caused by data imbalance, which can reduce the accuracy and fairness of assessments. To address this imbalance, researchers have explored the potential of using transfer learning approaches, such as fine-tuning multilingual language models such as BERT [10]. Some studies have referenced the use of BERT in handling data imbalance, [11] showing its effectiveness in managing class imbalance within BERT-based models. Other research [12], [13] these studies explored the issues of data imbalance

using BERT combined with oversampling techniques, resulting in improved model performance. Research related to essay scoring demonstrated that the resulting model was effective, achieving the best MAPE values of 5% for training and 8% for testing. The best combination of data used was 75% for training and 25% for testing [14]. This study indicates that this approach could be a valuable tool for automated scoring, offering high accuracy in essay correction.

Combining ROS and BERT approaches, it's anticipated that an automated essay scoring system can be developed that is not only accurate but also fair in its assessments, particularly in the context of data imbalance. This study contributes by examining the effectiveness of using the combination of ROS and BERT models and comparing them with several traditional machine learning methods (Random Forest Classifier, Naïve Bayes, and Backpropagation Neural Network) utilizing pretrained models. Various text preprocessing and pretrained data scenarios were explored to achieve higher accuracy results, as shown in Figure 1. This process aims to enhance the performance of automated essay scoring systems and explore how this approach can be applied in broader contexts.

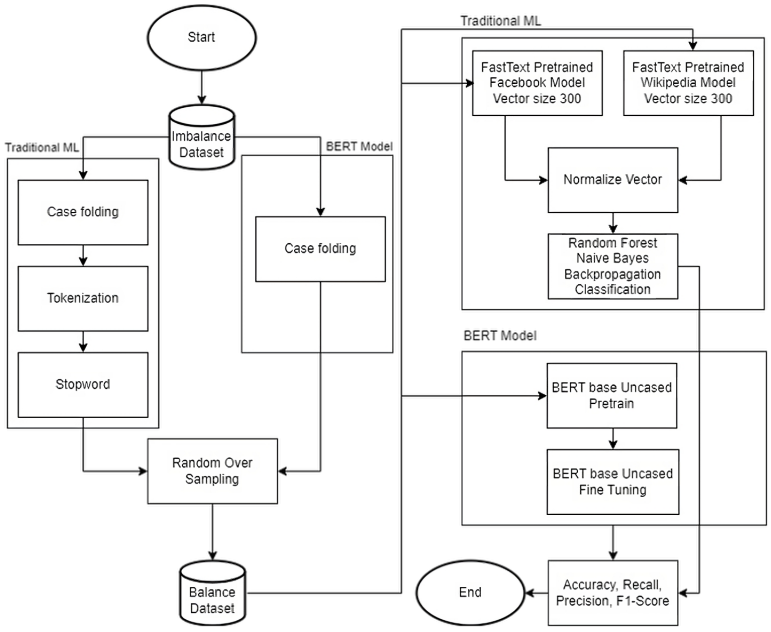

Figure 1: Flow system.

2 Research Method

This section provides an overview of the steps taken in this study, from text preprocessing to the evaluation phase, as shown in Figure 1. This research begins by inputting multiclass data derived from the essay responses of 8th-grade junior high school students in Indonesian language subjects, with scores ranging from 1 to 5. The data will be processed in two scenarios: one scenario employing traditional methods for data cleaning, which in-

cludes case folding, tokenization, and stopword removal; and another one is only case folding and tokenization using the BERT model. After the preprocessing stage is completed, the data will be analyzed to identify the number of minority and majority classes, followed by processing using the Random Over Sampling technique. For traditional model, the balanced data will be processed using pretrained models with FastText (Traditional Method). The pretrained model scenario in the traditional method is divided into two parts: one using the Facebook model with a vector size of 300, and the other using the Wikipedia model with a vector size of 300. After the process with pretrained data is finish, the next step is sentence vector normalization to capture the meaning of each sentence. The final step involves classification using the Random Forest Classifier, Naïve Bayes, and Backpropagation methods. Meanwhile, for the BERT model, the pretrained process is performed using only BERT base Uncased model. This dataset will be classified using the BERT Classifier with the BERT base Uncased model. The dataset used for classification is an oversampled dataset, which will then be analyzed to obtain the confusion matrix, including the calculation of Precision, Recall, F1-Score, and Accuracy. Based on the classification results and the values obtained from the confusion matrix, conclusions will be drawn regarding which method scenario demonstrates superior performance.

2.1 Dataset

The initial step in conducting this research involved collecting essay response data from 75 eighth-grade students for 10 questions in an Indonesian language subject. The total number of whole words in dataset 7773, the total of sentences in dataset 1502 with unique word 448. The obtained data is multiclass, consisting of 5 classes with scores ranging from 1 to 5. Table 1 presents the students' responses along with the scores assigned by the instructor for each question.

Table 1: Raw data

No	Question	Answer	Score
1	*Bagaimanakah menurutmu ciri-ciri iklan yang baik?*	*Teks iklan singkat, jelas, dan bertujuan untuk mengajak atau mempengaruhi orang.*	2
2	*Bagaimanakah menurutmu ciri-ciri iklan yang baik?*	*Iklan biasanya ada gambar dan teks yang menarik.*	5
3	*Bagaimanakah menurutmu ciri-ciri iklan yang baik?*	*teks iklan menggunakan kalimat yang singkat dan menarik perhatian*	3
...	...	...	...
750	*Membicarakan interaksi peserta didik difabel ... etc. Informasi penting apa yang disampaikan teks tersebut?*	*Informasi pentingnya adalah bahwa peserta didik difabel harus mendapat layanan interaksi khusus di sekolah inklusi.*	5

2.2 Text Pre-Processing

Text pre-processing typically involves the stages of case folding, tokenization, and stopword removal. The purpose of this step is to reduce dataset noise and enhance accuracy. In this study, the proposed BERT model utilized case folding and tokenizing steps from the BERT model [15] to ensure the data could be processed during the classification phase [16].

Table 2 presents the results of text pre-processing in the BERT model and tensor type tokenization.

Table 2: Pre-processing text

Answer	Case Folding	BERT-Model Tokenization
"Teks objektif menyampaikan informasi secara apa adanya tanpa pengaruh pendapat penulis"	*"teks objektif menyampaikan informasi secara apa adanya tanpa pengaruh pendapat penulis"*	`input_ids': tensor([[101, 107, 12008, 4616, 184, 1830, 5561, 21270, 8914, 1441, 17485, 4163, 7223, 1179, 12862, 17506, 14516, 8766, 1161, 170, 4163, 8050, 18266, 1161, 15925, 4163, 8228, 5526, 23698, 8228, 1810, 4163, 1204, 8228, 15818, 1116, 119, 107, 102]]), 'attention_mask': tensor([[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]])`

2.3 Random Oversampling

Data imbalance occurs when one or more classes in a dataset have significantly fewer instances compared than other classes. In this study, random oversampling (ROS) is employed as a technique to address the issue of data imbalance in essay answer correction by randomly selecting and duplicating data from the minority class as shown in Figure 2. In the prior research [17], it was acknowledged that random oversampling can enhance accuracy by addressing the issue of imbalanced datasets, as it creates identical replicas of the minority class samples.

2.4 BERT Model

Bidirectional Encoder Representation from Transformer (BERT) is a transformer model with bidirectional encoders. This model captures the context between layers and is capable of effectively representing the combination of left and right contexts [18]. Among its various configurations, BERT base uncased is a foundational model in NLP that has been widely adopted for various tasks due to its ability to understand context and semantics in text. Research indicates that BERT base uncased performs exceptionally well in tasks such as sentiment analysis and named entity recognition, leveraging its pre-training on a large corpus of text without case sensitivity, which enhances its robustness in handling diverse datasets [19], [20]. In previous research [18] proposed that the BERT model can be applied in two stages. The first stage involves pre-training, where the model learns to recognize the input text and its contextual relationships. The second stage involves fine-tuning, where the model assimilates and identifies the appropriate solutions. Fine-tuning a pre-trained

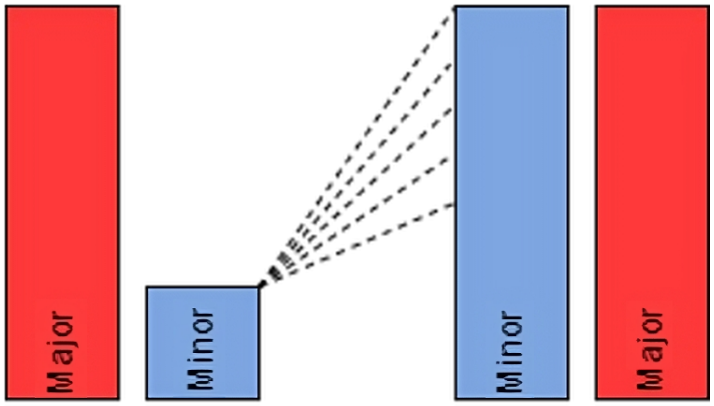

Figure 2: Random oversampling.

BERT model can be achieved by incorporating an additional layer to enhance contemporary performance.

2.5 K-Fold Cross Validation

K-Fold cross-validation is a statistical method used to evaluate the performance of a model being developed. In this research, the dataset will be divided using 5-fold cross-validation, resulting in 5 subsets of equal size. For each fold, one subset will be used as the testing data, while the remaining 4 subsets will be used as the training data during cross-validation.

2.6 Classification

In this research, the primary method is using BERT base uncased. We performed experiments that compared traditional machine learning methods such as Random Forest Classifier, Naïve Bayes, and Backpropagation. Each traditional machine learning method will utilize text representation with FastText word embeddings using pre-trained models from Facebook and Wikipedia with 300 dimensions of length vector.

2.7 Evaluation

Model evaluation is essential for assessing how well a model classifies a class. In this research, the evaluation method employed is the confusion matrix. The confusion matrix is a measurement tool used to evaluate the performance of a machine learning model in classification tasks, where the output consists of two or more classes. The parameters used for testing are True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). From this combination, precision, recall, F1-Score and accuracy are obtained, with the calculations as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

3 Results

This section discusses the results of the experiment conducted. The data obtained from the answers of 8$^{\text{th}}$-grade junior high school students in the Indonesian language subject were then analyzed to determine the distribution of minority and majority classes, as shown in Figure 3.

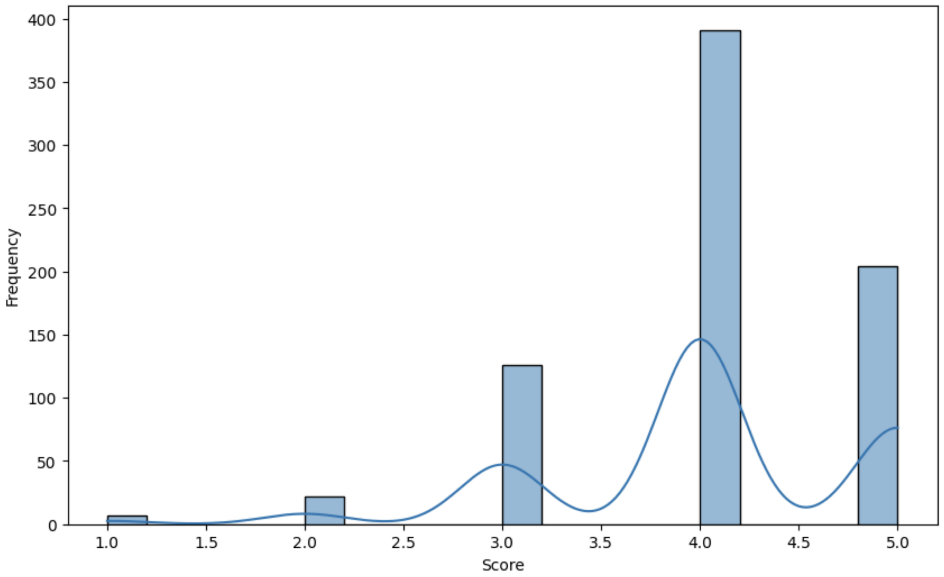

Figure 3: Distribution of score.

3.1 Result of Balancing Dataset

From Figure 3, the number of answers receiving a score of 1 is 7, a score of 2 is 22, a score of 3 is 126, a score of 4 is 391, and a score of 5 is 204. To solve this issue, this research addresses imbalanced data by oversampling minority classes to match or approach the majority class count based on the specified sampling strategy. The sampling strategy is to oversample classes 1 and 2 to 80% and 90% of the majority class count, respectively, and match the majority class count for class 3. This result is shown in Figure 4 where the minority class for a score 1 is 312, score of 2 is 351, and a score of 3 is 391. Finally, the

random oversampling technique results in the data having balanced values, bringing the data classes to an average level.

3.2 Result of Classification

We employed both deep learning using Transformers and machine learning-based approaches. The machine learning algorithms utilized include Random Forest Classifier, Naïve Bayes, and Backpropagation, while the deep learning model used is BERT base uncased. The classification results for BERT base uncased with various learning rate. The BERT base Uncased model is fine-tuned using the AdamW optimizer with 2e-5 learning rates. The result of cross validation with 5 K-Fold indicates that the 3rd fold yielded the highest average accuracy of 94% as shown as Table 3. Meanwhile the result of confusion matrix indicate that with learning rate of 2e-5 the average of precision score is 93%, recall of 94%, F1-Score of 94%, and an accuracy of 94%. Additionally, based on experiments using traditional ML methods with the application of text pre-processing and FastText word embedding, significant results were obtained for the Random Forest Classifier model in 4th fold with pre-trained Wikipedia, achieving a precision score of 85%, recall of 85%, F1-Score of 84%, and an accuracy of 85%. Meanwhile, with pre-trained Facebook data, the Backpropagation model outperformed other ML models in 4th fold, achieving a precision score of 83%, recall of 83%, F1-Score of 81%, and an accuracy of 83%. From these results, it can be concluded that the Traditional ML method using Random Forest with the pre-trained Wikipedia model achieved an average accuracy of 72%, while Backpropagation with the pre-trained Facebook model achieved 73%. Meanwhile, the BERT base uncased method achieved an average accuracy of 92%

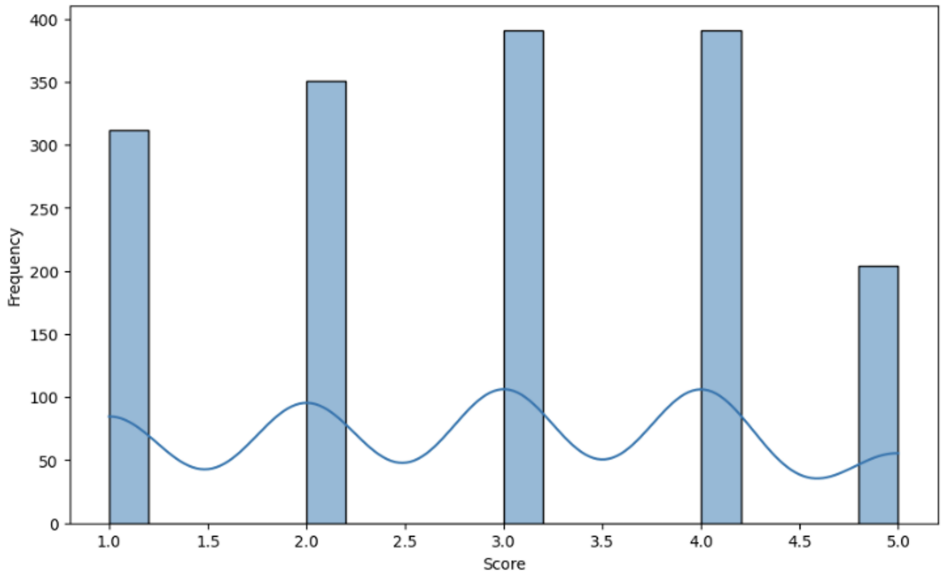

Figure 4: Dataset after random oversampling.

Table 3: Confusion matrix performance of variance learning rate

K-Fold	Learning Rate	Precision	Recall	F1-Score	Accuracy
1	2e-5	90.54%	90.78%	89.62%	90%
2	2e-5	93.54%	93.18%	94.37%	93%
3	**2e-5**	**93.68%**	**94.39%**	**94.79%**	**94%**
4	2e-5	92.32%	92.87%	92.18%	93%
5	2e-5	89.12%	89.62%	87.18%	89%
Mean	-	91.84%	92.16%	91.62%	92%

Table 4: Comparison model

	Accuracy Model						
Fold	Random Forest		Naïve Bayes		Backpropagation		BERT base uncased
	Wiki	FB	Wiki	FB	Wiki	FB	
1	67.8%	65.1%	44.5%	46.9%	72.7%	70.2%	90.3%
2	73.3%	70.1%	46.0%	42.4%	69.3%	77.2%	93.6%
3	68.7%	62.1%	48.7%	48.7%	67.5%	68.4%	**94.5%**
4	85.1%	75.1%	49.3%	53.0%	80.9%	83.6%	92.2%
5	65.6%	59.5%	46.8%	48.0%	63.2%	64.7%	89.3%

4 Discussion

The use of ROS successfully balanced a dataset that was previously skewed toward certain score categories. This rebalancing enabled machine learning models to more effectively learn from the data, particularly in identifying and accurately classifying the minority classes. Analyzing the results after oversampling, as evidenced by the confusion matrix, reveals a significant improvement in classification performance, especially in terms of precision and recall for the minority classes. These findings align with previous studies that underscore the value of ROS in boosting model accuracy by increasing the representation of underrepresented classes. Meanwhile, The BERT-base uncased model, fine-tuned with a learning rate of 2e-5 with 3rd fold, stood out as the top performer, achieving a precision of 93%, recall of 94%, and an F1-Score of 94%. These impressive results indicate that BERT, with its robust ability to grasp complex semantic relationships within text, surpasses traditional models such as Random Forest, Naïve Bayes, and Backpropagation, especially when used in conjunction with ROS. While traditional methods, particularly the Random Forest model paired with Facebook's pretrained FastText embeddings, demonstrated strong performance, they do have limitations. The primary challenge lies in FastText's ability to capture context-dependent meanings, which is less effective compared to the more advanced contextual understanding that BERT offers. The Random Forest Classifier achieved a notable accuracy of 90%, but this highlights the necessity of using more sophisticated models like BERT for tasks that demand deep linguistic analysis.

5 Conclusion

This study highlights the effectiveness of integrating ROS with a BERT-based model to tackle the issue of imbalanced data in automated essay scoring. The findings reveal that

ROS greatly enhances the model's ability to accurately classify minority classes, leading to improved overall performance metrics, including precision, recall, F1-Score, and accuracy. The BERT-base uncased model, fine-tuned with an optimal learning rate, surpassed traditional machine learning models such as Random Forest, Naïve Bayes, and Backpropagation, especially in dealing with complex semantic relationships in text. These results underscore the value of leveraging advanced NLP models such as BERT in combination with oversampling techniques to create fairer and more accurate automated scoring systems. This approach is particularly beneficial in educational settings where balanced and equitable assessment is essential. Future research might explore further enhancements in BERT fine-tuning or the application of other transformer-based models to continue advancing the performance of automated essay scoring system.

Acknowledgments

This research was funded by Kementerian Pendidikan, Kebudayaan, Riset dan Teknologi under Penelitian Dasar Scheme *(Penelitian Dosen Pemula)*.

References

[1] D. Ifenthaler, *Automated Essay Scoring Systems*. Singapore: Springer Nature Singapore, 2023, pp. 1057–1071. [Online]. Available: https://doi.org/10.1007/978-981-19-2080-6_59

[2] Z. Ke and V. Ng, "Automated essay scoring: A survey of the state of the art." in *IJCAI*, vol. 19, 2019, pp. 6300–6308.

[3] S. Datta and A. Arputharaj, "An analysis of several machine learning algorithms for imbalanced classes," in *2018 5th International Conference on Soft Computing & Machine Intelligence (ISCMI)*. IEEE, 2018, pp. 22–27.

[4] A. H. Filho, F. Concatto, J. Nau, H. A. d. Prado, D. O. Imhof, and E. Ferneda, "Imbalanced learning techniques for improving the performance of statistical models in automated essay scoring," 2019.

[5] Y.-g. Kim, Y. Kwon, and M. C. Paik, "Valid oversampling schemes to handle imbalance," *Pattern Recognition Letters*, vol. 125, pp. 661–667, 2019.

[6] M. Mujahid, E. Kına, F. Rustam, M. G. Villar, E. S. Alvarado, I. De La Torre Diez, and I. Ashraf, "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering," *Journal of Big Data*, vol. 11, no. 1, p. 87, 2024.

[7] T. Riston, S. N. Suherman, Y. Yonnatan, F. Indrayatna, A. A. Pravitasari, E. N. Sari, and T. Herawan, "Oversampling methods for handling imbalance data in binary classification," in *International Conference on Computational Science and Its Applications*. Springer, 2023, pp. 3–23.

[8] P. A. Perwira and N. I. Widiastuti, "Imbalance dataset in aspect-based sentiment analysis on game genshin impact review," *JURNAL INFOTEL*, vol. 16, no. 1, pp. 71–81, 2024.

[9] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[10] U. U. Acikalin, B. Bardak, and M. Kutlu, "Turkish sentiment analysis using bert," in *2020 28th Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2020, pp. 1–4.

[11] A. K. Jayaraman, A. Murugappan, T. E. Trueman, G. Ananthakrishnan, and A. Ghosh, "Imbalanced aspect categorization using bidirectional encoder representation from transformers," *Procedia Computer Science*, vol. 218, pp. 757–765, 2023.

[12] R. Mifsud, L. Deka, and I. Lahiri, "An optimised bert pretraining approach for identification of targeted offensive language: Data imbalance and potential solutions," in *2023 4th International Conference on Computing and Communication Systems (I3CS)*. IEEE, 2023, pp. 1–8.

[13] S. Wada, T. Takeda, K. Okada, S. Manabe, S. Konishi, J. Kamohara, and Y. Matsumura, "Oversampling effect in pretraining for bidirectional encoder representations from transformers (bert) to localize medical bert and enhance biomedical bert," *Artificial Intelligence in Medicine*, vol. 153, p. 102889, 2024.

[14] D. A. Sani and M. Z. Sarwani, "Koreksi jawaban esai berdasarkan persamaan makna menggunakan fasttext dan algoritma backpropagation," *Jurnal Nasional Pendidikan Teknik Informatika : JANAPATI*, vol. 11, no. 2, p. 92–111, Aug. 2022. [Online]. Available: https://ejournal.undiksha.ac.id/index.php/janapati/article/view/49192

[15] A. Jaiswal and E. Milios, "Breaking the token barrier: Chunking and convolution for efficient long text classification with bert," *arXiv preprint arXiv:2310.20558*, 2023.

[16] H. Saragih and J. Manurung, "Leveraging the bert model for enhanced sentiment analysis in multicontextual social media content," *Jurnal Teknik Informatika CIT Medicom*, vol. 16, no. 2, pp. 82–89, 2024.

[17] M. Hayaty, S. Muthmainah, and S. M. Ghufran, "Random and synthetic over-sampling approach to resolve data imbalance in classification," *International Journal of Artificial Intelligence Research*, vol. 4, no. 2, pp. 86–94, 2020.

[18] Z. Gao, A. Feng, X. Song, and X. Wu, "Target-dependent sentiment classification with bert," *Ieee Access*, vol. 7, pp. 154 290–154 299, 2019.

[19] S. K. Behera and R. Dash, "Fine-tuning of a bert-based uncased model for unbalanced text classification," in *Advances in Intelligent Computing and Communication: Proceedings of ICAC 2021*. Springer, 2022, pp. 377–384.

[20] M. Geetha and D. K. Renuka, "Improving the performance of aspect based sentiment analysis using fine-tuned bert base uncased model," *International Journal of Intelligent Networks*, vol. 2, pp. 64–69, 2021.