BalBERT: A New Approach to Improving Dataset Balancing for Text Classification

Laouni Mahmoudi*, Mohammed Salem

Department of Computer Science, University of MASCARA, Mascara 29000, Algeria

Corresponding Author Email: laouni.mahmoudi@univ-mascara.dz

ABSTRACT

Text classification has seen a lot of research, especially after social media platforms came into existence. It involves categorizing text into a variety of classes, such as positive, negative, neutral, or any other class label. The primary issue is that the usual text datasets collected from social media platforms are skewed (i.e., imbalanced). Consequently, classifiers become less effective. However, this paper addresses the issue by developing a new technique called BalBERT, which is based on the BERT scheme and includes a new sublayer to improve performance when dealing with imbalanced text classification. This new sublayer introduces balancing techniques that take advantage of the BERT representation step, which is context-based. The results demonstrate the effectiveness of BalBERT as measured by two metrics: AVG-Recall and F1-score. It outperforms both the BERT baseline and the state-of-the-art results.

1. INTRODUCTION

Text classification is a well-known topic in natural language processing and one of the most active study areas, particularly with the rise of social media platforms, where users post massive amount of text on a variety of subjects in a matter of seconds [1]. Text classification is the process of classifying text into categories. It is also known as "text labeling" or "text categorization." Text classifiers can analyze text automatically and then assign it to one of a set of predetermined labels or categories depending on its content [2].

Many techniques have been proposed in the literature to deal with this issue. Deep learning and related algorithms are gaining popularity. Among them, LSTM [2], BiLSTM [3, 4], CNN [2] and others. All of these algorithms surpassed the conventional machine learning approaches such as SVM [5], LR [5], and others. The publication of the Transformer algorithm [6] in 2017 revolutionized the field of natural language processing, particularly text classification, and several techniques were inspired due to the scalability of its design such as BERT [7] and GPT-3 [8]. The bidirectional encoder representation from the transformer (BERT) technique and its variants surpassed all previously constructed models [9].

Although the BERT approach towards text classification is approved when the datasets fed are balanced (i.e., the distribution of samples in classes is equal or has a slight skewing), its effectiveness is limited when tested on imbalanced datasets. Further, in real life, text datasets are usually imbalanced, which poses a new challenge for BERT, where the results given were ineffective.

The skewing dataset issue reduces the efficacy of all machine learning algorithms, including deep learning, particularly in classification [10]. However, balancing techniques such as undersampling, oversampling, or hybrid sampling [11] are needed to resample the imbalanced datasets in order to enhance the efficiency of the classifiers. These balancing techniques were evaluated on a numerical value dataset and demonstrated their effectiveness, but they are still difficult to implement in text datasets due to their complicated feature representation.

In this paper, we introduce the BalBERT approach based on BERT scheme to deal with imbalanced dataset for text classification. The new approach's originality is the enhancement of the BERT architecture by adding a new layer, that carry out the sampling of the dataset. The overall concept is to add a new layer that will resample the representation obtained by BERT fine-tuning, taking advantage of the context-based feature space generated in this phase. Tested with metrics appropriate for imbalanced classification AVG-Recall and F1-score [12], the new approach gives a prominent results outperformed BERT baseline and the state-of-the-art.

The remainder of the paper is structured as follows: The second section of the study investigates the existing literature. Section 3 offers a description of the BERT baseline for text classification architecture. Section 4 presents balancing techniques used to deal with skewing. Our proposed BalBERT approach is described in Section 5, whereas Section 6 presents the experiments, results, and discussion. Finally, a conclusion is presented.

2. RELATED WORK

Text classification is a crucial task in natural language processing (NLP) that entails categorizing text into predetermined categories. Since large-scale text data has become more commonly available [1], deep learning-based algorithms have become the state-of-the-art for text classification challenges. In recent years, transformer-based models, like as BERT, have exhibited remarkable performance in a range of NLP tasks, including text classification. In this literature review, we evaluate the current state-of-the-art in text categorization utilizing deep learning and transformer-

based algorithms. We also discuss the advantages and disadvantages of various methodologies, as well as some of the concerns that will need to be addressed in future research.

The use of deep neural network architectures for detecting hate speech on social media platforms has been a subject of recent research. In a study published in the study [13], the authors proposed the use of different deep neural network architectures for detecting hate speech on Twitter. The use of deep learning techniques has shown great potential in addressing these challenges and improving the accuracy of hate speech detection systems.

In addition to the aforementioned studies, LSTM and RNN-based approaches have also been applied to various real-world applications of text classification. In the study [14], the authors proposed a spam message classification system based on an LSTM network. The system achieved a high accuracy rate in detecting spam messages. In the study [15], the authors of this paper propose a novel approach to sentiment analysis that combines neural networks and ensemble learning. They integrate these models using stacking ensemble learning, which leads to a significant improvement in the accuracy of text sentiment analysis.

The transformer architecture has significantly transformed the field of natural language processing (NLP), especially in the area of text classification. Among the various algorithms that are based on the transformer architecture, BERT (Bidirectional Encoder Representations from Transformers) and its variants have gained immense popularity in recent years. BERT has shown impressive performance in a wide range of NLP tasks, including text classification. For instance, in the study [16], the authors proposed the BAE: BERT architecture based on BERT to enhance the baseline and evaluated it on various balanced datasets, such as AMASON, IMDB, and MR. Similarly, In the study [17], the authors highlight the difficulty of accurately classifying tweets. They utilized the Bidirectional Encoder Representations from Transformers (BERT) language model pre-trained on plain text instead of tweets using BERT Transfer Learning, which they refer to as modified (M-BERT). The modified model achieved an average F1-score of 97.63%.

Imbalanced datasets are a common challenge in deep learning classifiers, including BERT and its variants. In their study, the authors addressed this issue by utilizing numerical values datasets and achieved good performance. However, it is important to note that this approach may not generalize to other types of imbalanced datasets, and further research is needed to explore the effectiveness of this approach in other domains. Additionally, there are various techniques available to address the issue of imbalanced datasets in deep learning, such as oversampling, undersampling, and data augmentation, which can be considered depending on the specific characteristics of the dataset and the research question at hand.

The authors [18], employed the oversampling SMOTE algorithm and the Easy Ensemble concept to cope with imbalanced data., The authors [19] conduct a comparative experimental investigation to detect credit card fraud and to address the imbalance classification problem using several machine learning techniques for dealing with unbalanced datasets.

Due to the transition of text into numerical feature space, text-imbalanced classification is more challenging than numerical space in nature. To overcome this issue, we developed BalBERT, a novel architecture that uses BERT's performance in context-based sentence representation to

resample these outcomes.

The following sections describe and clarify the approaches.

3. BERT TEXT CLASSIFICATION

The Bidirectional Encoder Representation from Transformer (BERT) framework [7] is based on the revolutionary transformer architecture. This approach is divided into two steps: pre-training, and fine-tuning (see Figure 1). The model is trained on unlabeled data across several pre-training tasks in the first phase. For the fine-tuning, the BERT model is first started with the pre-trained parameters, and then all of the parameters are fine-tuned using labeled data from the downstream task. As thus, applied to balanced datasets, a BERT model fine-tuned to text classification task outperforms the state-of-the-art. The descriptions of the two steps are described below.
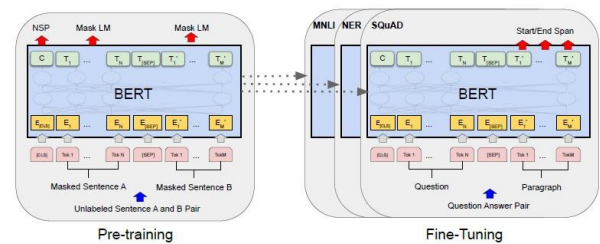

Figure 1. BERT architecture [7]

3.1 BERT pre-training

The pre-training step is a deep bi-directional model, where a shallow concatenation of a left-to-right and a right-to-left modeling is applied.

3.1.1 Task 1: Masked Language Modeling (MLM)
To do this, a random 15 percent of the input tokens are masked and then predicted to train a deep bidirectional representation. Finally, the mask tokens' final hidden vectors are fed into an output softmax over the vocabulary. For prediction, the training data generator selects 15 percent of the token placements at random. If the i-th token is picked, we replace it 80 percent of the time with the mask token, 10 percent with a random token, and 10 percent with the unmodified i-th token. Then, cross entropy loss and T_i are used to predict the original token.

3.1.2 Task 2: Next Sentence Predicting (NSP)
The purpose of this task is to predict association between two sentences. The model tests if a sentence is next or not for another. A monolingual corpus is utilized to train this type of binarized next sentence challenge model, which detects sentence connections.

After BERT pre-training, all parameters are transferred to initialize the downstream task (fine-tuning).

3.2 BERT fine-tuning

Fine-tuning a pre-trained model for many downstream tasks is a simple process using BERT's basic and yet flexible architecture, which involves just putting in the relevant inputs and outputs and fine-tuning the entire model's parameters end-to-end. Text classification and sequence tagging (e.g., POS,

NER), use concatenated sentences with one blank. as a single input phrase.

The input into the BERT baseline is a set of tokens that represent sentences, which result in a set of token representations when processed. REP<cls> is an aggregated representation of the text, and a representation for each token is generated (see Figure 2). Finally, models are built by fitting a classifier to the output of the representation data, using REP<cls>.

Although the BERT architecture performs well for textual classification tasks, it is inefficient for imbalanced classification. This issue is due to the skewing distribution of examples in text classes. However, imbalanced classification still poses challenges for predictive modeling in text even with this powerful approach.

In this study, we propose an enhancement to the BERT scheme by incorporating a new layer to deal with imbalanced problems by applying well-known numerical balancing techniques: oversampling, undersampling, and hybrid.
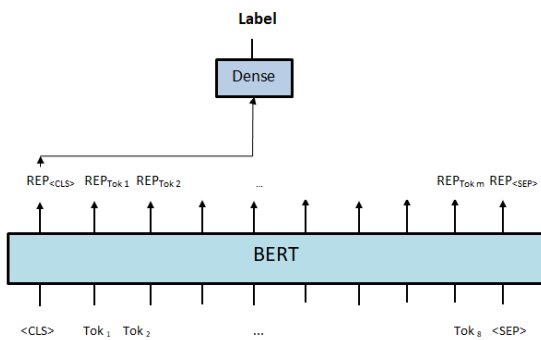

Figure 2. BERT baseline approach

4. BALANCING TECHNIQUES

Imbalanced classification is the process of developing prediction models for classification datasets with a significant degree of class imbalance. Working with imbalanced datasets offers the challenge that most machine learning algorithms will neglect, and so perform badly on, the minority class, despite the fact that the minority class is frequently the most significant.

To deal with this issue, in the literature, three well-known balancing techniques: oversampling, undersampling, and hybrid sampling have been tested and approved for their performance on numerical value datasets. The basis of these techniques is detailed below.

4.1 Balancing

Balancing is the modification of training data to achieve a more balanced class distribution that allows classifiers to perform similarly to regular classification.

Let dataset d have n instances consisting of pairs (X_i, y_i), i=1, ..., n, where X_i provides an input vector of attributes and y_i defines the corresponding class label. Each of the n instances has m input characteristics and should be assigned to a single class.

Balancing (i.e., sampling) is the process of taking a dataset d as input and transforming it into a new dataset d', where all classes will have an equal number of samples in Eq. (1). The number of n' is depending of the function f (i.e., technique

used, oversampling, undersampling or hybrid sampling). If new instances will be added the n' will increase in contrast it will be decreased.

$$f: d \rightarrow d'$$
$$d(n, k) \rightarrow d'(n', k) \tag{1}$$

4.2 Oversampling

Oversampling is the process of increasing the amount of samples from the minority class by duplicating certain instances or producing new instances from existing ones until a specific balance ratio R, which is determined using Equation 2, while X"$minority$" is the minority class and X"$majority$" is the majority [18]. The range of ratios used in oversampling can vary widely depending on the specific dataset and task. Common ratios used in oversampling are 1:1, 2:1, 3:1, or higher, depending on the degree of imbalance.

$$R = \frac{X"minority"}{X"majority"} \tag{2}$$

The objective is to generate new samples out of old ones.

Balancing using oversampling can be performed using one of two methods: random oversampling or synthetic oversampling. The first approach involves taking random samples from the minority group and duplicating them until they achieve a specified ratio as compared to the dominant group.

The issue with this strategy is that it is prone to overfitting because it uses the same texts for balancing.

We can reduce the risk of overfitting by lowering the ratio. The second approach is called as Synthetic Minority Over-Sampling Technique (SMOTE) [18], augmentation is used to produce samples, which creates new data points within the region of known points.

4.3 Undersampling

The objective of the undersampling approach is to lower the amount of samples by removing instances from the majority class.

This method is based on two principles: random undersampling and boundary oversampling [16].

The first is a simple method that involves randomly eliminating samples until a specific ratio R (See Eq. (2)) is obtained. The range of ratios used in undersampling can also vary widely depending on the specific dataset and task. Common ratios used in undersampling are 1:1 or lower, with lower ratios resulting in more aggressive undersampling.

The class is fully balanced if R equals one.

The problem of this approach is that we may miss critical features that might help us develop an effective model. The second approach collects samples at the intersection of two or more classes. It uses the Condensed Nearest Neighbors algorithm (CNN). The algorithm generally prefers to choose points around the class boundaries and move them to the new group.

4.4 Hybrid Sampling

The hybrid method combines the benefits of undersampling and oversampling strategies, allowing us to accentuate the minority class while filtering out any noisy observations.

5. BALBERT PROPOSED APPROACH

As aforementioned, the BalBERT approach is original in that a new layer is added to the original BERT architecture. The baseline technique serves two functions: representation and classification, but the new strategy build on it and consist of three tasks: representation, balancing, and classification.

5.1 BERT representation

In this step, preprocessing is used, cleaning the data, followed by tokenization. The tokens of every sentence are converted to its embedding representation using the BERT Embedding block. This embedding's sentences are then processed through a fully connected feed-forward network layer to obtain a context-based representation, REP<cls> (i.e. the aggregated representation of sentences), and each sentence's tokens (See Figure 3).
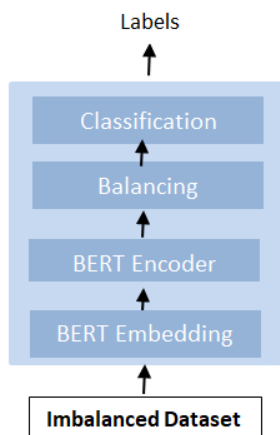

Figure 3. BalBERT approach

5.2 BERT balancing

The output of the first (i.e., Context-Based Feature Space) step is used in the balancing block, using the various mathematical techniques to tackle the problem of imbalance, which are the techniques of oversampling, subsampling and hybrid.

Mathematically, this involves randomly selecting a subset of the majority class with size equal to that of the minority class, or a user-defined ratio. Oversampling, on the other hand, involves increasing the size of the minority class by replicating samples or generating new synthetic samples using various techniques such as SMOTE or ADASYN. Mathematically, this involves randomly selecting a subset of the minority class and applying the oversampling technique until the desired ratio is achieved, which can range from a user-defined ratio to 1:1. The choice of oversampling or undersampling technique, as well as the ratio to use, depends on the specific problem and the data at hand.

Architecturally, BERT's is built on blocks, and balancing techniques are based around this concept.

5.3 BERT classification

After the BERT sampling step, the classifier is trained with their output (i.e., balanced dataset), using the labeling of each sample, to create the models, then, it will be tested on test dataset to select the best sampling block.

The next section investigates an ensemble of experiments to highlight the BalBERT approach on a well-known Twitter imbalanced dataset.

6. EXPERIMENTS

6.1 Experiment processes and settings

In this section, we describe the experiments that were conducted and the results that were obtained of our models regarding a specific metrics for imbalanced classification compared to the baseline BERT model.

6.1.1 Dataset settings and parameters
We used two particular unbalanced datasets with many classes. These datasets are gathered to deal with sentiment analysis, a subfield of text classification.

The first is known as ASAD (Twitter-based Benchmark Arabic Sentiment Analysis dataset) [20], it was published for a KAUST-sponsored sentiment analysis competition (See Table 1).

Table 1. ASAD dataset

	Tweet's Number	Positive	Negative	Neutral
ASAD Dataset	55,000	8,200	8,821	37,359

Table 2. Review dataset

	Tweet's Number	1	2	3	4	5
Review Dataset	63,257	2,939	5,285	12,001	19,054	23,778

The second dataset is a severely skewed Review dataset [21] containing five classes. It is being launched for the CERIST NLP Challenge 2022. (See Table 2).

6.1.2 Experiments description
All experiments follow the same preparation procedure, which involves removing noise, hashtags, URLs, and other symbols like imogis and non-alphabetic letters from sentences. When evaluating the performance of a multi-class imbalanced text classification model, two commonly used metrics were the average recall (AVG recall) and F1-score. AVG recall measured the average proportion of relevant instances that were correctly classified across all classes, taking into account the imbalanced nature of the dataset. F1-score, on the other hand, was the harmonic mean of precision and recall, providing an overall measure of the model's accuracy that considered both the false positives and false negatives. By using these two metrics, a more comprehensive evaluation of the model's performance on imbalanced datasets was obtained.

AraBERT [22] is an effective Arabic-specific variant of BERT. Before applying the new BalBERT technique, the AraBERT approach is directly fitting to the unbalanced

datasets, it is necessary to assess the default outcome and create a baseline in performance (see Figure 2). In the other experiments, the BalBERT approach is used each time, with a different balancing technique (see Figure 3).

6.2 Experiment results

The results of the experiments are presented in Tables 3-5. Table 3 shows the worst outcome compared to the others, indicating that the baseline BERT technique is inadequate in addressing the skewing issues. For oversampling, SMOTE outperforms the Random approach in both datasets. Both strategies also outperform the baseline BERT in the two metrics, with a minimum increase of 17% and 20%, respectively. Tables 4 and 5 demonstrate the effectiveness of the boundary technique when fitted to the two imbalanced datasets using undersampling. It achieves a novel improvement, with the greatest result reaching 79% for AVG-Recall. Furthermore, Tables 4 and 5 show a significant result of 80% and more in the two datasets.

Table 3. BERT baseline fit to ASAD & review datasets

Metrics	ASAD Dataset	Metrics	Review dataset
F1-pos	0.54	F1-1	0.43
F1-Neg	0.55	F1-2	0.48
F1-Neu	0.85	F1-3	0.52
		F1-4	0.60
		F1-5	0.65
Positive-Recall	0.46	1-Recall	0.40
Negative-Recall	0.45	2-Recall	0.45
Neutral-Recall	0.89	3-Recall	0.44
		4-Recall	0.59
		5-Recall	0.82
AVG-Recall	**0.60**	AVG-Recall	**0.54**
F1-PN	**0.54**	123-Recall	**0.47**

Table 4. BalBERT fit to ASAD dataset

Metrics	Over Ran	Over SMO	Under Ran	Under Boun	Ran & Ran	Boun & SMO
F1-pos	0.68	0.70	0.69	0.70	0.75	0.76
F1-Neg	0.70	0.71	0.66	0.72	0.76	0.77
F1-Neu	0.80	0.82	0.85	0.87	0.89	0.88
Pos-Recall	0.67	0.71	0.66	0.69	0.76	0.75
Neg-Recall	0.68	0.72	0.69	0.67	0.74	0.72
Neu-Recal	0.81	0.83	**0.87**	**0.86**	**0.98**	**0.87**
AVG-Recall	**0.68**	0.75	**0.74**	**0.79**	**0.80**	**0.77**
F1-PN	**0.69**	0.70	0.68	**0.71**	**0.77**	**0.76**

Table 5. BalBERT fit to review dataset

Metrics	Over Ran	Over SMO	Under Ran	Under Boun	Ran & Ran	Boun & SMO
F1-1	0.61	0.68	0.61	0.62	0.65	0.68
F1-2	0.63	0.68	0.65	0.65	0.69	0.70
F1-3	0.67	0.74	0.72	0.71	0.73	0.74
F1-4	0.74	0.78	0.77	0.71	0.78	0.79
F1-5	0.77	0.79	0.82	0.75	0.81	0.83
1-Recall	0.61	0.66	0.60	0.62	0.63	0.68
2-Recall	0.62	0.68	0.67	0.62	0.67	0.71
3-Recall	0.68	0.72	0.70	0.71	0.72	0.73
4-Recall	0.71	0.74	0.78	0.79	0.76	0.78
5-Recall	0.78	0.78	0.80	0.81	**0.88**	0.81
AVG-Recall	**0.68**	**0.72**	**0.71**	**0.71**	**0.73**	**0.74**
123-PN	0.64	0.70	0.66	0.66	0.69	**0.71**

6.3 Discussion

The findings presented in the Result section demonstrate the BalBERT approach efficiency. Based on BERT representation step, this conducts to resample the dataset in contrast of the embedding step which cannot give a context-based feature space. Although results of BalBERT are considerable taking account of the baseline BERT and results from [18], it can be enhanced with other balancing techniques using the same BalBERT approach. Applying SMOTE approach added new features that helped to discriminate classes in datasets which push the AVG-Recall from 60 percent to 75

percentAdditionally, it is important to note that the undersampling approach used in BalBERT removes interacting features, resulting in improved model performance. This is evident in Tables 2 and 3, where BalBERT achieved an AVG-Recall of 79 percent compared to only 60 percent achieved by the baseline BERT, despite the presence of dataset skewing. These results are a significant improvement over both the baseline and state-of-the-art techniques [23].

The hybrid approach that combines SMOTE and Boundary achieved a performance of 80%, which is a reasonable outcome given the complementary roles of each technique. Although the hybrid approach that combines random

oversampling and random undersampling produced the best result at 81%, the former hybrid approach is considered more rational. After conducting a more in-depth analysis of the experimental results, it was discovered that the adoption of novel balancing techniques led to a remarkable enhancement in the performance of the minority classes, with an increase of no less than 30%. This finding is particularly significant and highlights the efficacy of these newly implemented approaches in tackling the challenge of imbalanced datasets. It underscores the fact that such techniques have the potential to greatly improve the overall performance of machine learning models and provide more accurate results when dealing with datasets that contain an uneven distribution of classes. The observed improvement in the performance of the minority classes demonstrates that the balancing techniques have effectively addressed the issue of underrepresented classes, resulting in a more comprehensive and reliable model evaluation.

Upon comparing the experimental results of the two datasets, it was observed that BalBERT demonstrated superior performance on the ASAD dataset in comparison to the Review dataset. As a result, the imbalanced nature of the Review dataset and the presence of a larger number of classes posed a greater challenge for the BalBERT model in achieving higher accuracy compared to the ASAD dataset.

7. CONCLUSION

We addressed the issue of imbalanced text classification in this paper, which is a well-known task in natural language processing and where many challenges are launched. We used two datasets published to Challenge for testing our novel based on BERT.

The intuition of this new approach, BalBERT, is the efficacy of BERT in the text classification domain and balancing techniques applied in mathematical feature space. So, we add a new layer to the BERT fine-tuning after the representation step, taking into account that it is a context-based space.

Three balancing techniques were used: oversampling, undersampling, and hybrid sampling. For each strategy, we investigated two well-known and efficient approaches.

Finally, hybrid methods were used.

The new architecture outperforms the baseline BERT and the state-of-the-art results in all approaches, especially, where hybrid sampling are used (SMOTE and Boundary). Hence, results approved our intuition in dealing with imbalanced text classification issue. Moreover, in future work, we want to do additional experiments on more datasets with more classes as well as tests on multi-domain datasets.

REFERENCES

[1] Yao, T., Zhai, Z., Gao, B. (2020). Text classification model based on fasttext. In 2020 IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS), pp. 154-157. https://doi.org/10.1109/ICAIIS49377.2020.9194939

[2] Zhang, C., Li, Q., Cheng, X. (2020). Text sentiment classification based on feature fusion. Revue d'Intelligence Artificielle, 34(4): 515-520. https://doi.org/10.18280/ria.340418

[3] Luo, S.Y., Gu, Y.J., Yao, X.X., Fan, W. (2021). Research on text sentiment analysis based on neural network and ensemble learning. Revue d'Intelligence Artificielle, 35(1): 63-70. https://doi.org/10.18280/ria.350107

[4] Pei, P. (2022). Short texts classification based an improved BiLSTM. In 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), pp. 297-300. https://doi.org/10.1109/AEMCSE55572.2022.00067

[5] Palacharla, R.K., Vatsavayi, V.K. (2022). A new supervised term weight measure based approach for text classification. Revue d'Intelligence Artificielle, 36(3): 395-407. https://doi.org/10.18280/ria.360307

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 5998-6008, https://doi.org/10.48550/arXiv.1706.03762

[7] Devlin, J. Chang, M.W., Lee, K., Toutanova, K. (2019). Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the (2019) Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1: 4171-4186. https://doi.org/10.48550/arXiv.1810.04805

[8] Floridi, L., Chiriatti, M. (2020). GPT-3: Its Nature, scope, limits, and consequences. Minds and Machines 30: 681-694 https://doi.org/10.1007/s11023-020-09548-1

[9] Deping, L., Hongjuan, W., Mengyang, L., Pei, L. (2021). News text classification based on Bidirectional Encoder Representation from Transformers. In 2021 International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA) Xi'an, China, pp. 137-140. https://doi.org/10.1109/CAIBDA53561.2021.00036

[10] Keya, A.J., Wadud, M. A. H., Mridha, M. F., Alatiyyah, M., Hamid, M. A. (2022). AugFake-BERT: Handling imbalance through augmentation of fake news using Bert to enhance the performance of fake news classification. Applied Sciences, 12(17): 8398. https://doi.org/10.3390/app12178398

[11] Bellinger, C., Drummond, C., Japkowicz, N. (2016). Beyond the boundaries of smote: A framework for manifold-based synthetically oversampling. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Springer, pp. 248-263. https://doi.org/10.1007/978-3-319-46128-1_16

[12] Sundarkumar, G.G., Ravi, V. (2015). A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance. Engineering Applications of Artificial Intelligence, 37: 368-377. https://doi.org/10.1016/j.engappai.2014.09.019

[13] Wong, T.T. (2020). Linear approximation of F-measure for the performance evaluation of classification algorithms on imbalanced data sets. IEEE Transactions on Knowledge and Data Engineering, 34(2): 753-763. https://doi.org/10.1109/TKDE.2020.2986749

[14] Chirra, V.R., Maddiboyina, H.D., Dasari, Y., Aluru, R. (2020). Performance evaluation of email spam text classification using deep neural networks. Review of Computer Engineering Studies, 7(4): 91-95. https://doi.org/10.18280/rces.070403

[15] Alharbi, L.M., Qamar, A.M. (2022). Arabic sentiment analysis of eateries' reviews using deep learning.

Ingénierie des Systèmes d'Information, 27(3): 503-508. https://doi.org/10.18280/isi.270318

[16] Garg, S., Ramakrishnan, G. (2020). Bae: Bert-based adversarial examples for text classification. arXiv preprint arXiv:2004.01970. https://doi.org/10.18653/v1/2020.emnlp-main.498

[17] Kannan, E., Kothamasu, L.A. (2022). Fine-tuning BERT based approach for multi-class sentiment analysis on Twitter emotion data. Ingénierie des Systèmes d'Information, 27(1): 93-100. https://doi.org/10.18280/isi.270111

[18] Yuan, Z., Zhao, P. (2019). An improved ensemble learning for imbalanced data classification. In 2019 IEEE 8th joint international information technology and artificial intelligence conference (ITAIC), Chongqing, China, pp. 408-411. https://doi.org/10.1109/ITAIC.2019.8785887

[19] Baker, M.R., Mahmood, Z.N., Shaker, E.H. (2022). Ensemble learning with supervised machine learning models to predict credit card fraud transactions. Revue d'Intelligence Artificielle, 36(4): 509-518. https://doi.org/10.18280/ria.360401

[20] Alharbi, B., Alamro, H., Alshehri, M., Khayyat, Z., Kalkatawi, M., Jaber, I.I., Zhang, X. (2020). ASAD: A twitter-based benchmark arabic sentiment analysis dataset. arXiv preprint arXiv:2011.00578. http://doi.org/10.48550/arXiv.2011.00578

[21] Aly, M.A., Atiya, A.F. (2013). LABR: A Large Scale Arabic Book Reviews Dataset. ArXiv, abs/1411.6718. https://doi.org/10.48550/arXiv.1411.6718

[22] Antoun, W., Baly, F., Hajj, H. (2020). Arabert: Transformer-based model for arabic language understanding. arXiv preprint arXiv:2003.00104. https://doi.org/10.48550/arXiv.2003.0010

[23] Mahmoudi, L., Salem, M. (2023). Improving Multi-class Text Classification Using Balancing Techniques. In Artificial Intelligence: Theories and Applications: First International Conference, ICAITA 2022, Mascara, Algeria, November 7–8, 2022, Revised Selected Papers, Mascara, Algeria, pp. 264-275. https://doi.org/10.1007/978-3-031-28540-0_21