A Cost-Sensitive Hybrid Model of ALBERT Model and Convolutional Neural Network for Personality Classification

Rilo Chandra Pradana[1]* and Derwin Suhartono[2]
[1]Mathematics Department, School of Computer Science, Bina Nusantara University
Jakarta, Indonesia 11480
[2]Computer Science Department, School of Computer Science, Bina Nusantara University
Jakarta, Indonesia 11480
Email: [1]rilo.pradana@binus.ac.id, [2]dsuhartono@binus.edu

***Abstract*—A tremendous amount of text data from social media activity can be used to extract information about a user's personality, including the Myers-Briggs Type Indicator (MBTI). The MBTI personality type is extensively used to identify individual traits, which helps to solve problems in human resources and mental health awareness. Nonetheless, constructing an effective model for classifying MBTI types that are insensitive to unbalanced data remains a major challenge, as certain types dominate the social media environment. The research proposes a hybrid classification model that combines the transformer-based language model A Lite Bidirectional Encoder Representations from Transformers (ALBERT) with a Convolutional Neural Network (CNN), leveraging cost-sensitive learning to address class imbalance. The model is trained on the PersonalityCafe dataset and evaluated across the four MBTI dimensions. Experimental results show that the proposed ALBERT+CNN model achieves an overall F1-score of 77.67%, outperforming baseline models such as Bidirectional Encoder Representations from Transformers (BERT), Bidirectional Long Short-Term Memory (BiLSTM), and traditional CNN. When integrated with cost-sensitive learning, the model reaches an improved F1-score of 80.50%, surpassing the performance of oversampling techniques like Random Oversampling (ROS) and Synthetic Minority Oversampling Technique (SMOTE). The exponential cost function proves to be the most effective in weighting misclassifications for minority classes. In addition to higher accuracy, the proposed model demonstrates balanced prediction performance across personality dimensions, reducing bias toward dominant classes. These findings highlight the potential of hybrid deep learning and cost-sensitive strategies for personality classification in imbalanced textual data.**

***Index Terms*—Cost-Sensitive Hybrid Model, A Lite Bidirectional Encoder Representations from Transformers (ALBERT), Convolutional Neural Network (CNN), Personality Classification**

I. Introduction

IN April 2023, around 59.9% of the global population or 4.8 billion individuals were involved in social media activities, with a 3.2% user growth rate [1]. This large user base creates enormous volumes of text data that may be used to extract useful information, such as defining a social media user's personality type. This information may be used to tailor the user's social media timeline [2], thus eliminating the need for standard personality tests, which are frequently laborious and time-consuming.

Several psychological tests, like the Myers-Briggs Type Indicator (MBTI) [3], have been used to evaluate personality characteristics. The validity and reliability of MBTI have been called into question if it is compared to other models [4] like the Big Five [5] or Dominance, Influence, Steadiness, and Compliance (DISC) [6]. However, it is still commonly used to deal with issues of mental health and human resources [7–9]. Moroever, obstacles exist in autonomous personality recognition using text data, such as identifying appropriate classification models and managing data imbalance caused by the distribution of personality types in the online environment.

Previous research has used different machine learning [2, 10–15], deep learning [10, 13, 16–18], and transformer-based models, such as Bidirectional Encoder Representations from Transformers (BERT) [14, 19], and A Lite Bidirectional Encoder Representations from Transformers (ALBERT) [20], for MBTI categorization. BERT, with its capacity to gather contextual information [21], faces challenges related to training

Received: July 02, 2024; received in revised form: Jan. 13, 2025; accepted: Jan. 13, 2025; available online: April 28, 2025.
*Corresponding Author

speed and scalability owing to the number of variables involved. Hence, ALBERT addresses this issue by significantly reducing parameters while maintaining performance through parameter sharing and factorized embedding parameters [22].

The unequal distribution of MBTI types both geographically [23] and online [11, 24] presents issues, since classifiers tend to overfit the majority class [25]. It results in poor prediction of some types despite their high accuracy [26]. Oversampling approaches such as Random Oversampling (ROS) [16] and Synthetic Minority Oversampling Technique (SMOTE) [2] have been used to address this issue. However, they can introduce duplication and fake data, affecting class distribution and potentially lowering classifier performance [26]. Traditional machine learning models such as Support Vector Machine (SVM), Logistic Regression, and Extreme Gradient Boosting (XGBoost), as well as Standard Text Representations, have demonstrated different degrees of performance. These models produce F1-scores ranging from 67.2% [11] to 92% [12]. But, they struggle with unbalanced data and biased predictions toward specific MBTI dimensions or types.

Deep learning techniques, such as Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) models, have also been applied for MBTI categorization, with results that are better than those classical models. For example, CNN achieves an accuracy of 81.4% [13], whereas LSTM obtains 77.8% [18]. Then, CNN outperforms LSTM in multiclass MBTI classification, achieving greater accuracy and F1-scores [17]. Transformer models, such as BERT, when employed as feature extractors and integrated with machine learning models, exhibit enhanced accuracy of more than 90% without data balancing operations [14, 19]. ALBERT, an upgraded version of BERT, also performs better, with a macro F1-score of 68.5% for the Big Five personality assessment [20].

Then, data balancing strategies like ROS, SMOTE, and Borderline-SMOTE have been employed to address unbalanced datasets, effectively enhancing the F1-score of models [2, 16]. Nonetheless, these strategies may adversely affect the performance of transformer models, as modifications to the training distribution may result in a decline in accuracy [27, 28]. However, cost-sensitive learning, an alternative approach, has shown significant positive effects in other domains. Cost-sensitive EfficientNetV2 [29], cost-sensitive Bayesian Neural Network (BNN) [30], and cost-sensitive XGBoost [31] are some of the suggested cost-sensitive deep learning models that have outperformed their non-cost-sensitive counterparts in terms of accuracy. For instance, cost-sensitive boosts the CNN-based model performance from the accuracy of 93% to 99%, improves the accuracy of the BNN model from 86.72% to 93.76%, and increases the Area Under the Curve (AUC) value of XGBoost from 95.1% to 95.5%.

Determining the misclassification cost is crucial for the performance of cost-sensitive models. Although the minority class proportion is a common way to calculate misclassification costs, alternative functions have been examined, including power, exponential, and logarithmic misclassification cost functions. For its greater performance in terms of the F1-score and calculation cost, the exponential weight has been proposed [31]. This approach can potentially surpass the performance of models using oversampling methods, making it a promising solution for addressing data imbalance in MBTI classification.

Considering the benefits of the ALBERT model and cost-sensitive learning for imbalanced data classification, the researchers propose the hybrid model of ALBERT and CNN with cost-sensitive learning to address the problem in MBTI classification. The research contributes in two aspects. First, the proposal of the ALBERT model for feature extraction to improve the feature representation of the data can lead to the improvement of the classification results. Second, the research provides insights of implementing cost-sensitive learning to eliminate the effect of imbalanced data that appears in the dataset for MBTI classification.

II. Research Method

A. Myers-Briggs Type Indicator (MBTI) Personality Model

The MBTI is a personality theory developed by Isabel Myers and Katharine C. Briggs, grounded in Jung's personality theory [3]. The MBTI theory categorizes human personality into four dimensions, namely the Introversion/Extraversion (I/E) dimension, the Intuition/Sensing (N/S) dimension, the Thinking/Feeling (T/F) dimension, and the Judgement/Perception (J/P) dimension. The comprehensive description of each dimension is as follows [3, 32]:

1) Introversion/Extraversion (I/E) measures how individuals interact with their environment and whether they are outwardly (E) or inwardly (I) oriented.
2) Intuition/Sensing (N/S) measures how individuals process information through direct experience (S) or instinct and imagination (N).
3) Thinking/Feeling (T/F) measures how individuals make decisions through logic (T) or moral principles (F).

TABLE I
A SAMPLE OF THE PERSONALITYCAFE DATASET.

Type	Posts
INTP	'Good one https://www.youtube.com/watch?v=fHiGbolFFGw |||–Of course, to which I say I know; that's my blessing and my curse. ||| Does being absolutely positive that you and your best friend could be an amazing couple count? If so, than yes. Or it's more I could be madly in love in case I reconciled my feelings (which at... ||| ...'

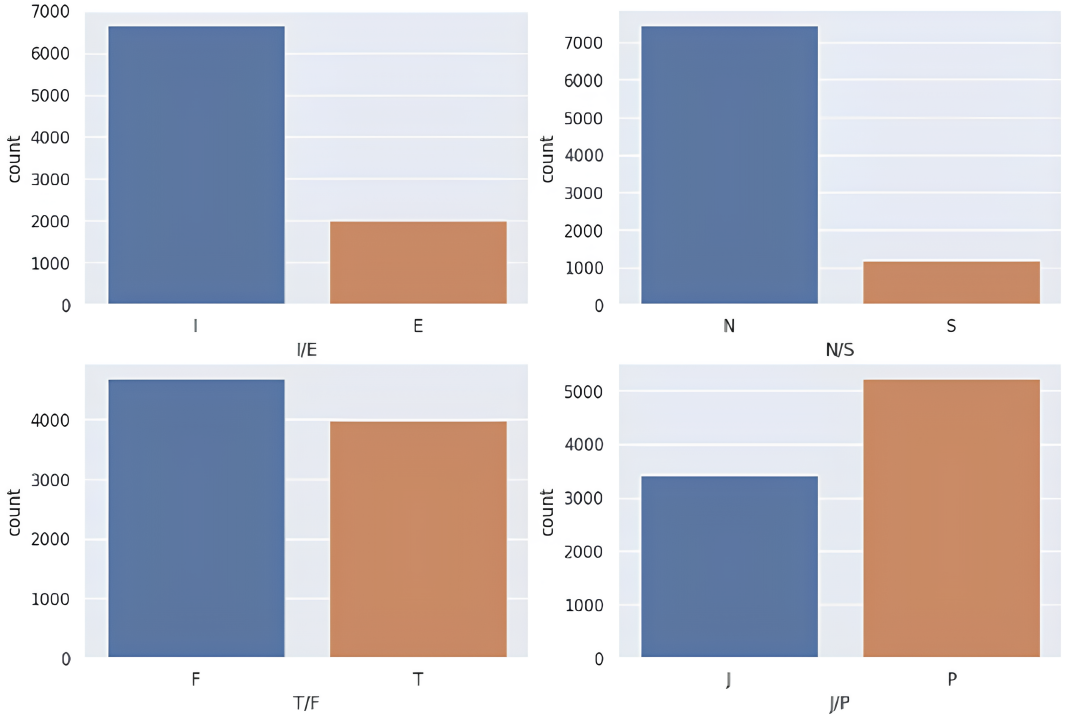

Fig. 1. Frequency distribution of each class in each dimension.

4) Judgment/Perception (J/P) measures how individuals organize their surroundings. It assesses whether they tend to create and refer to a plan (J) or they tend to be flexible and accept change (P).

In MBTI theory, personality types are determined by combining the four dimensions mentioned. Therefore, each individual falls into one of possible 16 types: INFJ, ENFJ, INFP, ENFP, INTJ, ENTJ, INTP, ENTP, ISFJ, ESFJ, ISFP, ESFP, ISTJ, ESTJ, ISTP, or ESTP. Each letter in the type signifies the dimension that an individual actively employs. For example, an individual with an INFJ personality type typically employs introversion, intuition, emotion, and judgment.

B. Dataset

The proposed model is trained using the PersonalityCafe dataset [24]. The PersonalityCafe dataset is a public dataset accessed via the Kaggle website. The dataset contains text data collected by web scraping from the PersonalityCafe website, an online forum that discusses topics about personality theory, including MBTI. The dataset comprises 8,675 users with an MBTI type assigned to each user and the respective user's last 50 posts. Each user is labeled by one of 16 MBTI types. Each user has at most 50 posts, with every post separated by a three-line symbol (|||). Table I displays a sample in the dataset.

The distribution of the count of each dimension is shown in Fig. 1. Some dimensions are underrepresented in the dataset, while others dominate the class distribution. The dataset exhibits an imbalanced class distribution in the I/E and N/S dimensions, with the E and S classes being underrepresented. Meanwhile, for the T/F dimension and J/P dimension, the class frequencies do not differ greatly. Hence, both dimensions

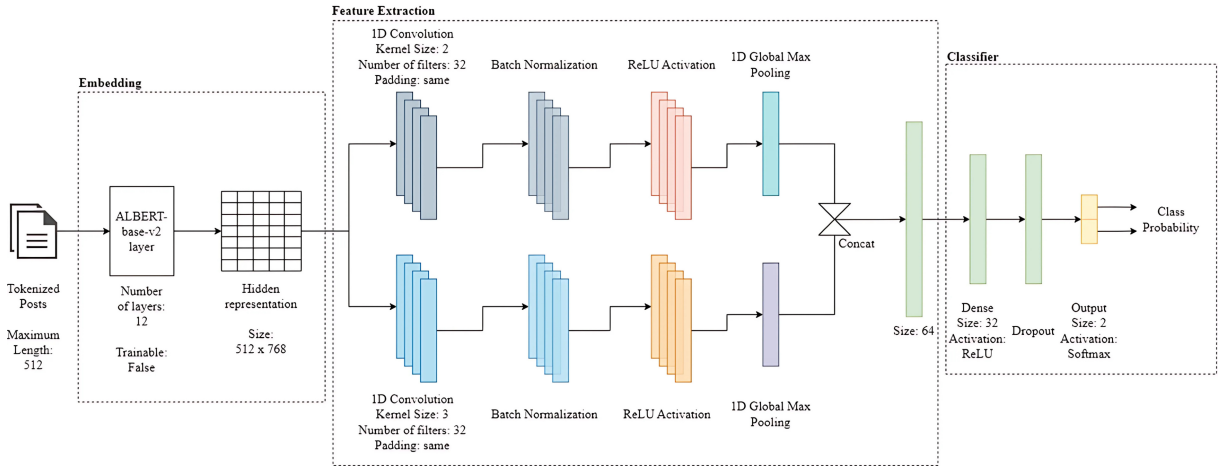

Fig. 2. A Lite Bidirectional Encoder Representations from Transformers (ALBERT) + Convolutional Neural Network (CNN) model.

do not possess an imbalance characteristic.

C. Data Preprocessing

The preprocessing stage of a dataset is the first stage of the experiment. The initial step involves text cleaning by omitting posts exceeding 25 words following the Linguistic Inquiry and Word Count (LIWC) guidelines for personality classification [33] and performing standard text cleaning procedures, including eliminating contractions, stopwords, and unwanted characters like periods, commas, quotation marks, underscores, and URLs. Tokenization follows text cleaning, dividing sentences into tokens. This experiment utilizes the SentencePiece tokenization [34]. The ALBERT limitation then truncates the longer tokenized text, including only the first 512 tokens. Besides the preprocessing steps on the post data, the MBTI type data is also preprocessed by splitting it into four dimensions.

After all data in the dataset are cleaned and tokenized, the dataset is split into three non-overlapping sets: training set, validation set, and testing set. The training set, which comprises 80% of the dataset, is used to train the model. The trained model is validated by the validation set, which has 10% of all data in the dataset. Lastly, the resulting model is evaluated to assess its performance using the testing set, which has 10% of all data in the dataset.

D. Proposed Model

The research proposes a hybrid model combining the pre-trained language model ALBERT with a One-Dimensional CNN (1D-CNN) illustrated in Fig. 2. The pre-trained ALBERT model is an improvement of the BERT model, where the ALBERT model has fewer parameters but still has roughly the same performance as BERT [22]. The ALBERT model has a similar architecture to BERT, with a base model of 12 layers and a 768-dimensional hidden layer. The ALBERT model utilizes a lower embedding space of 128 dimensions before projecting into the hidden space and employs parameter sharing across all layers to minimize the number of parameters. These processes enable ALBERT-base to possess merely 12 million parameters, in contrast to BERT-base, which contains 108 million parameters [22]. The proposed model in the research utilizes the ALBERT-based model, generating word vectors with a dimensionality of 768. The research utilizes the pre-trained ALBERT-based model provided by the HuggingFace library.

The word vectors that the ALBERT model obtains are projected to the convolution layers. The research implements multiple convolution layers with different kernel sizes to extract information in different n-grams of words. The kernel window will slide over the word sequence to detect n-gram features in different positions. Each convolution layer uses the same number of filters and padding mechanism to ensure that all convolution layers produce the output with the same dimension. After the word vectors have been filtered, they are sent to a batch normalization layer and a Rectified Linear Unit (ReLU) layer. Finally, the global max-pooling layer is used to reduce the number of dimensions. The researchers choose the global max-pooling layer to extract the most prominent information from the filtered features.

Table II shows the hyperparameters that are used in the model. The convolutional layers utilize 32 filters with a stride of 1 to efficiently extract features from the input data. The model incorporates a dense layer

TABLE II
MODEL HYPERPARAMETERS.

Hyperparameter	Value
Number of filters	32
Stride	1
Dense output	32
Dropout rate	0.5
Learning rate	1e-3
Optimizer	Adam
Batch size	32

TABLE III
SEARCH SPACES FOR HYPERPARAMETERS (β AND γ).

Weight Function	β	γ
Power	{0.85, 0.9, 0.95}	{1/3, 1/2}
Logarithm	{0.9, 0.95}	-
Exponential	{0.75, 0.9}	-

with 32 output units in conjunction with a dropout rate of 0.5, a strategy employed to mitigate overfitting by randomly disabling half of the neurons during the training process. The model is optimized using the Adam optimizer with a learning rate of 0.001 (1e-3), offering a favorable trade-off between speed and convergence stability. Furthermore, a batch size of 32 is employed, signifying that the model updates its weights following every 32 training samples.

The convolutional network outputs are concatenated to merge the extracted information from each convolution to obtain the post's predicted MBTI dimension. The merged features are the input of the dense layer. The dense layer extracts more information contained in the merged features, with the ReLU activation function implemented as a nonlinear function. The dropout layer is connected between the output and dense layers during the training phase to lessen the chance of overfitting. To do this, a random selection of input values is made to be deactivated with a probability known as the dropout rate. The output layer has two nodes with the softmax activation function, which gives the probability of each class. As an MBTI type has four dimensions, there are four different models, each producing a prediction for one dimension.

The ALBERT and CNN models are combined with the transfer learning paradigm. The CNN model replaces the top layer of the ALBERT model, and ALBERT parameters are frozen during the training phase. The model is trained using the early-stopping procedure to avoid overfitting. After the training phase, the model is evaluated, and its performance is compared to that of the other models. It notes that the models for each dimension are trained using the same hyperparameters and training procedures.

E. Model Training

Cost-sensitive learning is implemented to train the model. The cost-sensitive learning aims to minimize the model's total cost (or simply the loss) by implementing the unequal cost of misclassification between classes. A higher cost is applied when the model classifies a minority class instance as the majority [26]. The misclassification cost is represented by a weight that influences the loss value, where the loss of classifying the majority class is multiplied by a factor that is less than one so that the loss of classifying the minority class would be higher.

Suppose that in a binary classification problem, an instance x_i that belongs to class y_i is classified as the member of class $\hat{y}_i$. Equation (1) provides the total loss that must be minimized during training. It shows that m is the size of the training set and $q(\lambda)$ is the weight. The λ represents the proportion of the majority class within the training set. There are several formulas of $q(\lambda)$ implemented in the research. The simplest formula used is the Imbalance Ratio (IR), which is formulated by Eq. (2).

$$J = \frac{1}{m}\sum_{i=1}^{n} -q(\lambda) y_i \log \hat{y}_i - (1-y_i)\log(1-\hat{y}_i), \tag{1}$$

$$q(\lambda) = \frac{\lambda}{1-\lambda}. \tag{2}$$

The research also implements the misclassification cost functions introduced by previous research [31], namely the power, logarithm, and exponential functions. These functions are formulated by Eqs. (3)–(5), respectively. The introduction of those functions is necessary because the imbalance ratio is occasionally ineffective in enhancing the model's performance when dealing with severely imbalanced data [31]. The β and γ are hyperparameters. Since β and γ are hyperparameters, those values are tuned to obtain the appropriate minority class' misclassification cost value for each dimension. The search spaces for β and γ are given in Table III. For the power cost function, both β and γ are explored, with β taking values from the set {0.85, 0.9, 0.95} and γ from the set {1/3, 1/2}, providing flexibility in cost adjustments. The logarithmic and exponential cost functions are associated exclusively with the parameter β, which is varied over the set {0.9, 0.95} and {0.75, 0.9}, respectively. Applying distinct sets of values guarantees that the resulting misclassification cost values do not become excessively extreme, enabling them to represent the actual class distribution more accurately. This approach fosters bal-

anced learning, particularly in instances of imbalanced datasets.

$$q(\lambda) = \beta \left(\frac{\lambda}{1-\lambda} \right)^{\gamma}, 0 < \beta, \gamma \leq 1, \tag{3}$$

$$q(\lambda) = \beta \log \left(\frac{\lambda}{1-\lambda} \right), 0 < \beta \leq 1, \tag{4}$$

$$q(\lambda) = \beta \cdot 10^{2\lambda - 1}, 0 < \beta \leq 1. \tag{5}$$

F. Model Evaluation

This experiment has two versions of the trained proposed model: the standard (non-cost-sensitive) model and the cost-sensitive model. The performance of the standard model is compared with other models used in previous studies, while the efficacy of the cost-sensitive model is evaluated to determine the optimal weight for each class through a comparison with the standard model's performance. The ideal cost-sensitive model is subsequently compared to other data balancing techniques discussed in previous studies, including ROS, Random Undersampling (RUS), and SMOTE. The research uses the deep learning models from the previous works as a baseline:

1) CNN: The research chooses the 1D-CNN as the baseline, with the FastText embedding as the text representation. The model is used to set the hyperparameters [17]
2) Bidirectional Long Short-Term Memory (BiLSTM): The research chooses the BiLSTM model as the baseline with the FastText embedding, similar to the model implemented [17]
3) ALBERT-base: As the research proposes utilizing the ALBERT model for feature extraction, the proposed model is compared with the ALBERT-base model [22]. The model consists of a block of ALBERT architecture connected directly with the output layer (softmax). The layers in the ALBERT model are frozen during the training phase.
4) BERT-base: The architecture of this model is similar to the ALBERT-base, with the ALBERT block substituted by the BERT-base model [21].
5) BERT+CNN: This model is a hybrid similar to the proposed model, except it utilizes BERT instead of ALBERT as the feature extractor.
6) ALBERT+BiLSTM: This hybrid model is similar to the proposed model, except the model uses BiLSTM as the classifier instead of 1D-CNN.

Then, the F1-score is selected as the evaluation metric for the comparison between the baseline models and the proposed model, as the imbalance characteristic of the dataset will skew the prediction toward the majority class. The F1-score metric is the harmonic mean between precision and recall, where both metrics consider the number of true positive [35]. The precision, recall, and F1-score formula are given by Eqs. (6)–(8), respectively. TP, FP, and FN denote the number of true positives, number of false positives, and number of false negatives, respectively.

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{6}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{7}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} + \text{Recall}}{\text{Precision} \times \text{Recall}}. \tag{8}$$

III. Results and Discussion

With the Google Cloud TPU serving as the accelerator, the research experiment is conducted within the Google Colaboratory environment. Huggingface's transformer v4.35.2 and Tensorflow v2.12.0 are the main packages used for model implementation in Python 3.10. The utilization of Google Cloud TPU has been demonstrated to markedly accelerate the training process by facilitating high-performance parallel computation. This benefit is particularly salient when dealing with large-scale deep-learning models. Then, Google Colaboratory offers a flexible and accessible experimental platform, while Huggingface Transformers facilitates the seamless integration of pre-trained models and state-of-the-art architectures. TensorFlow functions as the fundamental deep learning framework, providing substantial tools for model construction, training, and evaluation.

A. Results

Each model is trained without implementing cost-sensitive learning or oversampling. Due to the random initialization of the model (except for the pre-trained model), the experiment is repeated 10 times. The F1-score is calculated for each run, and the average is measured across runs to summarize the model performance. The model's performances are reported in Table IV.

The combination of the ALBERT and 1D-CNN models yields better results than the 1D-CNN model alone, with overall F1-scores of 77.67% for ALBERT + CNN and 70.89% for 1D-CNN, respectively in Table IV. The proposed model also reaches a higher F1-score for each MBTI dimension than the ordinary 1D-CNN. This outcome shows that using the ALBERT model instead of the FastText word embedding produces a higher-up word hidden representation. However, because the ALBERT model is simpler, it performs worse than BERT. Nonetheless, the F1-scores of both models varied by less than 1%.

TABLE IV
COMPARISON OF AVERAGE F1-SCORE BETWEEN PROPOSED MODEL AND BASELINES. THE OVERALL PERFORMANCE IS THE AVERAGE OF THE F1-SCORE FOR EACH DIMENSION.

Model	I/E	N/S	T/F	J/P	Overall
CNN	69.87%	58.32%	82.77%	72.59%	70.89%
BiLSTM	62.54%	54.72%	59.97%	62.51%	59.93%
ALBERT-base	53.52%	49.35%	73.22%	56.39%	58.12%
BERT-base	53.97%	48.75%	76.83%	58.51%	59.51%
BERT+CNN	**75.69%**	**75.77%**	**83.93%**	78.26%	**78.41%**
ALBERT+BiLSTM	73.89%	73.99%	83.02%	75.96%	76.71%
ALBERT+CNN (proposed)	75.60%	73.11%	83.63%	**78.35%**	77.67%

Note: Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (BiLSTM), A Lite Bidirectional Encoder Representations fromTransformers (ALBERT), Bidirectional Encoder Representations from Transformers (BERT), Bidirectional Long Short-Term Memory (BiLSTM), Introversion/Extraversion (I/E), Intuition/ Sensing (N/S), Thinking/Feeling (T/F), and Judgement/Perception (J/P).

TABLE V
COMPARISON OF AVERAGE F1-SCORE USAGE OF DIFFERENT MISCLASSIFICATION COST VALUES ON THE PROPOSED MODEL.

CS Strategy	I/E	N/S	T/F	J/P	Overall
No CS	75.60%	73.11%	83.63%	78.35%	77.67%
Imbalance ratio	77.01%	78.55%	83.63%	78.51%	79.42%
Balanced	77.67%	78.51%	83.37%	78.39%	79.48%
Power	78.04%	80.13%	**84.21%**	**79.06%**	80.36%
Logarithm	77.36%	79.44%	79.05%	77.66%	78.38%
Exponential	**78.60%**	**80.15%**	84.20%	78.58%	**80.38%**

Note: Cost-Sensitive (CS), Introversion/Extraversion (I/E), Intuition/Sensing (N/S), Thinking/Feeling (T/F), and Judgement/Perception (J/P).

Each MBTI dimension is predicted more accurately using the 1D-CNN model than when the BiLSTM model is implemented. The BiLSTM model achieves an overall F1-score of 59.93%, whereas the proposed model reaches 77.67%. The combination of ALBERT and BiLSTM likewise performs worse than the proposed model, with an overall F1-score of 76.71%. The 1D-CNN model works better in predicting the MBTI dimension than a recurrent model like BiLSTM [36]. It is because the 1D-CNN model is better at detecting local features in the text than the BiLSTM model. This argument stems from the correlation between an individual's personality and the qualities of words frequently used, emphasizing word frequency and length. Consequently, the interdependence among words is of lesser significance than the local attributes to MBTI prediction.

The research implements cost-sensitive learning to overcome the proposed model's insensitivity in identifying the minority class across each dimension. The procedure for training the cost-sensitive model is similar to the non-cost-sensitive model. The experiment is repeated 10 times, and the average F1-score of each run is calculated to obtain the result. The results of each cost-sensitive learning strategy are provided in Table V. The results of the non-cost-sensitive model are equal to those of the proposed model in Table IV. Only the best β and γ results are reported for each misclassification cost function.

In Table V, the cost-sensitive learning approach successfully improves the performance of the proposed model. Using five different misclassification cost settings, the cost-sensitive model yields the highest overall F1-score of 80.38%, representing an approximate 3% increase over the standard model. Compared to all cost-sensitive strategies, the exponential misclassification cost has the best performance among those cost strategies in detecting the I/E and N/S dimensions, with the F1-score of 78.60% and 80.15%, respectively. Meanwhile, the best results are obtained using power misclassification cost for the T/F and J/P dimensions, with an overall F1-score of 84.21% and 79.06%, respectively.

Table VI shows the results of the cost-sensitive learning and data-approach imbalance handling methods. Similar to the previous section, each imbalance handling strategy is repeated 10 times, and the average F1-score of each run is calculated to obtain the result. It should be noted that the results of no imbalance handling are equal to those of the proposed model in Table IV. The optimal outcome for each dimension in Table V has been included in this table for the cost-sensitive results. The overall column is the average of the data from all dimensions.

The cost-sensitive learning results are more favorable compared to other data-balancing strategies. The methods of ROS, RUS, and SMOTE have a lower

TABLE VI
COMPARISON OF AVERAGE F1-SCORE BETWEEN IMBALANCE HANDLING STRATEGIES ON THE PROPOSED MODEL.

Imbalance Handling Strategy	I/E	N/S	T/F	J/P	Overall
No Data Balancing	75.60%	73.11%	83.63%	78.35%	77.67%
ROS	77.40%	78.36%	83.91%	78.08%	79.44%
RUS	75.15%	73.42%	83.22%	77.95%	77.43%
SMOTE	77.03%	78.39%	**84.23%**	78.71%	79.59%
Cost-Sensitive	**78.60%**	**80.15%**	84.21%	**79.06%**	**80.50%**

Note: Random Oversampling (ROS), Synthetic Minority Oversampling Technique (SMOTE), Random Undersampling (RUS), Introversion/Extraversion (I/E), Intuition/Sensing (N/S), Thinking/Feeling (T/F), and Judgement/Perception (J/P).

TABLE VII
PERFORMANCE OF THE PROPOSED MODEL AND MODELS IN PREVIOUS RESEARCH.

Model	Metric	I/E	N/S	T/F	J/P	Overall
SVM (No CS) [15]	Accuracy	78.00%	86.00%	73.00%	66.00%	75.75%
	F1-score	88% (I), 1% (E)	92% (N), $<$ 1% (S)	75% (F), 69% (T)	42% (J), 76% (P)	55.38%
XGBoost(SMOTE) [12]	Accuracy	77.30%	85.00%	75.60%	67.30%	76.30%
	F1-score	**87.00%**	**92.00%**	72.00%	55.00%	76.50%
ALBERT + CNN (Cost-Sensitive)	F1-score	78.60%	80.15%	**84.21%**	**79.06%**	**80.50%**

Note: the term in parentheses next to the model name denotes the imbalance handling strategy used in the experiment. The overall column contains the average of metric values across dimensions. The table has Cost-Sensitive (CS), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Synthetic MinorityOversampling Technique (SMOTE), A Lite Bidirectional Encoder Representations from Transformers (ALBERT)+Convolutional Neural Network (CNN), Introversion/Extraversion (I/E), Intuition/Sensing (N/S), Thinking/Feeling (T/F), and Judgement/Perception (J/P).

overall F1-score than the cost-sensitive learning. They have an overall F1-score of 79.44%, 77.43%, 79.59%, and 80.50% for ROS, RUS, SMOTE, and cost-sensitive learning, respectively. Furthermore, the cost-sensitive learning approach excels in predicting I/E, N/S, and J/P. Conversely, the SMOTE method achieves the highest score for the T/F dimension, with an F1-score of 84.23%, which is slightly above the cost-sensitive learning score of 84.21%.

B. Discussion

Table VII shows the performance of the model in previous studies with the proposed model in the research. The proposed model demonstrates greater performance in detecting the J/P dimension when compared to the model used in previous studies regarding the model's capacity to detect all MBTI dimensions. Moreover, whereas alternative models exhibit heightened sensitivity and bias towards specific dimensions, such as N/S or I/E, the new model demonstrates a balanced efficacy in identifying each dimension. The proposed model identifies an individual's MBTI type with greater accuracy than the previously suggested model in earlier studies.

The architecture of the hybrid model, which integrates ALBERT for contextual feature extraction and 1D-CNN for local feature detection, is a key factor in its superior performance. ALBERT's parameter sharing mechanism and reduced embedding dimensionality enable efficient representation learning without sacrificing contextual depth. When coupled with 1D-CNN, which captures recurrent local patterns such as word n-grams associated with personality traits, the model becomes capable of handling both global and local semantic structures in the data. This complementary relationship enhances generalization across different MBTI traits. Importantly, the model maintains relatively consistent predictive accuracy across dimensions, including the more challenging J/P, as evidenced by its F1-score of 79.06% in Table VII, an improvement over previous models that have exhibited strong biases toward certain dimensions.

In addition to the architectural innovation, the research highlights cost-sensitive learning as a superior strategy for dealing with class imbalance compared to traditional techniques such as ROS, RUS, and SMOTE. Unlike these data-level methods that alter the training distribution, potentially introducing redundancy or noise, cost-sensitive learning modifies the training objective by assigning greater penalties to misclassifications involving minority classes, thus preserving the original linguistic characteristics of the dataset. As shown in Table VII, the cost-sensitive model achieves the highest overall F1-score of 80.50%. In contrast, previous studies using SVM and XGBoost combined with SMOTE show a strong bias toward certain MBTI dimensions, such as a high F1-score on N/S but poor performance on J/P, leading to less balanced results. The proposed approach's ability to dynamically emphasize minority classes through calibrated loss weighting, particularly with the exponential cost function, demonstrates a more reliable and equitable

solution for unbalanced text classification, supporting its effectiveness in personality type prediction.

IV. Conclusion

The research proposes a hybrid model using ALBERT and 1D-CNN for predicting the MBTI type of an individual based on social media post data, particularly text data. The ability to capture the contextual information in the data with a lighter model that ALBERT has and the 1D-CNN model that is powerful in extracting local features in sequence data makes the proposed model improve in detecting each MBTI dimension. To address data imbalance, the researchers use a cost-sensitive method in several misclassification cost strategies to enhance the proposed model's efficacy in identifying minority classes. The proposed model achieves an overall F1-score of 77.67%, and the cost-sensitive learning successfully improves the model with an overall F1-score of 80.50%. These results outperform other models, demonstrating that the proposed model better predicts MBTI types using textual data.

In the future research, it will be possible to include the part-of-speech and emotional features of the words in the input. The research highlights the limitation of cost-sensitive learning, emphasizing the necessity of optimizing misclassification costs through hyperparameter search or metaheuristic optimization methods in future research to achieve the most suitable misclassification cost for optimal model performance. Furthermore, additional models and a synthesis of models will be incorporated in the next studies. Consequently, the objective of attaining an improved F1-score for each dimension will be achieved.

Acknowledgement

The authors would like to acknowledge the support from Bina Nusantara University for funding the publication fee of the research. The research is not supported by any external grants.

Author Contribution

Conceived and designed the analysis, R. C. P. and D. S.; Collected the data, R. C. P.; Contributed data or analysis tools, R. C. P.; Performed the analysis, R. C. P.; Wrote the paper, R. C. P.; and Supervised the writing process, D. S.

Data Availability

The data that support the findings of the research are openly available in Kaggle at https://www.kaggle.com/datasets/datasnaek/mbti-type, reference number [24].

References

[1] S. Kemp, "Digital 2023 April global statshot report," 2023. [Online]. Available: https://datareportal.com/reports/digital-2023-april-global-statshot

[2] G. Ryan, P. Katarina, and D. Suhartono, "MBTI personality prediction using machine learning and SMOTE for balancing data based on statement sentences," *Information*, vol. 14, no. 4, pp. 1–15, 2023.

[3] I. B. Myers, M. H. McCaulley, N. L. Quenk, and A. L. Hammer, *MBTI manual: A guide to the development and use of the Myers-Briggs Type Indicator*. Consulting Psychologists Press, 1998.

[4] A. Furnham, "The big five facets and the MBTI: The relationship between the 30 NEO-PI (R) Facets and the four Myers-Briggs Type Indicator (MBTI) scores," *Psychology*, vol. 13, no. 10, pp. 1504–1516, 2022.

[5] K. El-Demerdash, R. A. El-Khoribi, M. A. I. Shoman, and S. Abdou, "Deep learning based fusion strategies for personality prediction," *Egyptian Informatics Journal*, vol. 23, no. 1, pp. 47–53, 2022.

[6] E. Utami, A. D. Hartanto, S. Adi, I. Oyong, and S. Raharjo, "Profiling analysis of DISC personality traits based on Twitter posts in Bahasa Indonesia," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 2, pp. 264–269, 2022.

[7] M. Song, H. J. Choi, and S. S. Hyun, "MBTI personality types of Korean cabin crew in Middle Eastern Airlines, and their associations with cross-cultural adjustment competency, occupational competency, coping competency, mental health, and turnover intention," *International Journal of Environmental Research and Public Health*, vol. 18, no. 7, pp. 1–20, 2021.

[8] M. H. Akhtar, A. Ashfaq, A. M. Khalid, and M. Baig, "Stress levels among pre-clinical medical students and their coping strategies," *Journal of University Medical & Dental College*, vol. 14, no. 1, pp. 524–528, 2023.

[9] E. Azoulay, F. Pochard, J. Reignier, L. Argaud, F. Bruneel, P. Courbon, A. Cariou, K. Klouche, V. Labbé, F. Barbier *et al.*, "Symptoms of mental health disorders in critical care physicians facing the second COVID-19 wave: A cross-sectional study," *Chest*, vol. 160, no. 3, pp. 944–955, 2021.

[10] M. H. Amirhosseini and H. Kazemian, "Machine learning approach to personality type prediction based on the Myers–Briggs Type Indicator®," *Multimodal Technologies and Interaction*, vol. 4,

no. 1, pp. 1–15, 2020.

[11] M. Gjurković and J. Šnajder, "Reddit: A gold mine for personality prediction," in *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*. Louisiana, USA: Association for Computational Linguistics, June 2018, pp. 87–97.

[12] K. N. P. Kumar and M. L. Gavrilova, "Personality traits classification on Twitter," in *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. Taipei, Taiwan: IEEE, Sep. 18–21, 2019, pp. 1–8.

[13] T. Pradhan, R. Bhansali, D. Chandnani, and A. Pangaonkar, "Analysis of personality traits using natural language processing and deep learning," in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*. Coimbatore, India: IEEE, July 15–17, 2020, pp. 457–461.

[14] H. Zhang, "MBTI personality prediction based on BERT classification," *Highlights in Science, Engineering and Technology*, vol. 34, pp. 368–374, 2023.

[15] M. T. Zumma, J. A. Munia, D. Halder, and M. S. Rahman, "Personality prediction from Twitter dataset using machine learning," in *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. Kharagpur, India: IEEE, Oct. 3–5, 2022, pp. 1–5.

[16] N. Čerkez and V. Vareškić, "Machine learning approaches to personality classification on imbalanced MBTI datasets," in *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)*. Opatija, Croatia: IEEE, Sep. 27–Oct. 1, 2021, pp. 1259–1264.

[17] N. Cerkez, B. Vrdoljak, and S. Skansi, "A method for MBTI classification based on impact of class components," *IEEE Access*, vol. 9, pp. 146 550–146 567, 2021.

[18] R. K. Hernandez and I. Scott, "Predicting Myers-Briggs type indicator with text," in *31st Conference on neural information processing systems (NIPS 2017)*, Long Beach, California, Dec. 4–9, 2017.

[19] V. G. Dos Santos and I. Paraboni, "Myers-Briggs personality classification from social media text using pre-trained language models," *JUCS: Journal of Universal Computer Science*, vol. 28, no. 4, pp. 378–395, 2022.

[20] Y. Wang, J. Zheng, Q. Li, C. Wang, H. Zhang, and J. Gong, "XLNet-Caps: Personality classification from textual posts," *Electronics*, vol. 10, no. 11, pp. 1–16, 2021.

[21] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "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, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, Jun. 2019, pp. 4171–4186.

[22] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," 2019. [Online]. Available: https://arxiv.org/abs/1909.11942

[23] N. A. Schaubhut and R. C. Thompson, "Indonesia (Indonesian) technical brief for the MBTI® global step I™ and step II™ assessments," 2020. [Online]. Available: https://www.themyersbriggs.com/-/media/Myers-Briggs/Files/Manual-Supplements/MBTI-Global-Manual-Tech-Brief_IDN.pdf

[24] M. Jolly, "(MBTI) Myers-Briggs personality type dataset," 2017. [Online]. Available: https://www.kaggle.com/datasets/datasnaek/mbti-type

[25] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," *ACM Computing Surveys (CSUR)*, vol. 52, no. 4, pp. 1–36, 2019.

[26] B. Krawczyk, M. Woźniak, and G. Schaefer, "Cost-sensitive decision tree ensembles for effective imbalanced classification," *Applied Soft Computing*, vol. 14, pp. 554–562, 2014.

[27] H. T. Madabushi, E. Kochkina, and M. Castelle, "Cost-sensitive BERT for generalisable sentence classification with imbalanced data," 2020. [Online]. Available: https://arxiv.org/abs/2003.11563

[28] D. Mercier, S. T. R. Rizvi, V. Rajashekar, A. Dengel, and S. Ahmed, "ImpactCite: An XLNet-based method for citation impact analysis," 2020. [Online]. Available: https://arxiv.org/abs/2005.06611

[29] V. Ravi, "Attention cost-sensitive deep learning-based approach for skin cancer detection and classification," *Cancers*, vol. 14, no. 23, pp. 1–26, 2022.

[30] B. Ghoshal and A. Tucker, "On cost-sensitive calibrated uncertainty in deep learning: An application on COVID-19 detection," in *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*. Aveiro, Portugal: IEEE, June 7–9, 2021, pp. 503–509.

[31] K. Li, B. Wang, Y. Tian, and Z. Qi, "Fast and accurate road crack detection based on adaptive cost-sensitive loss function," *IEEE Transactions on Cybernetics*, vol. 53, no. 2, pp. 1051–1062, 2021.

[32] K. Y. Kim, Y. B. Yang, M. R. Kim, J. S. Park, and J. Kim, "MBTI personality type prediction model using WZT analysis based on the CNN ensemble and GAN," *Human-Centric Computing and Information Sciences*, vol. 13, 2023.

[33] A. Koutsoumpis, J. K. Oostrom, D. Holtrop, W. Van Breda, S. Ghassemi, and R. E. De Vries, "The kernel of truth in text-based personality assessment: A meta-analysis of the relations between the Big Five and the Linguistic Inquiry and Word Count (LIWC)," *Psychological Bulletin*, vol. 148, no. 11-12, pp. 843–868, 2022.

[34] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2018, pp. 66–71.

[35] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, Inc., 2022.

[36] H. Ahmad, M. U. Asghar, M. Z. Asghar, A. Khan, and A. H. Mosavi, "A hybrid deep learning technique for personality trait classification from text," *IEEE Access*, vol. 9, pp. 146 214–146 232, 2021.