IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 April 2024, accepted 16 June 2024, date of publication 21 June 2024, date of current version 28 June 2024.
Digital Object Identifier 10.1109/ACCESS.2024.3417180

RESEARCH ARTICLE

Novel Curriculum Learning Strategy Using Class-Based TF-IDF for Enhancing Personality Detection in Text

NAAE KWON, YUENKYUNG YOO, AND BYUNGHAN LEE, (Member, IEEE)
Department of Electronic Engineering, Seoul National University of Science and Technology, Seoul 01811, South Korea
Corresponding author: Byunghan Lee (bhlee@seoultech.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT under Grant NRF-2022R1C1C1006511, in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) under Grant RS-2022-II220043, and in part by the Information Technology Research Center (ITRC) Support Program under Grant IITP-2024-RS-2022-00156295.

ABSTRACT Personality detection plays a pivotal role in social interactions, machine learning (ML), and natural language processing (NLP). Its goal is to discern an individual's traits from their behavior and expressions. The prevalence of text-based communication has sparked interest in inferring personality from written content. However, challenges persist in accurately interpreting traits like the Big-Five or Myers-Briggs Type Indicator. These challenges stem from the reliance on self-reported surveys for labeling, which introduces uncertainties as individual assessments may not consistently align with their actual personality. In this paper, we propose novel curriculum strategies that employ class-based term frequency-inverse document frequency (c-TF-IDF) to enhance personality detection performance. By leveraging a curriculum approach that mirrors human learning progression, starting from simpler tasks and moving toward more complex ones, these strategies aim to train models on progressively challenging scenarios. Our experimental results demonstrate that these proposed curriculum-based strategies improve the accuracy of personality detection compared to previously suggested methods. This study contributes to advance understanding of text-based cues for personality inference. It has the potential to enrich various fields, including human-computer interaction, personalized recommendations, and targeted marketing.

INDEX TERMS Personality detection, curriculum strategy, c-TF-IDF, language models, big-five personality.

I. INTRODUCTION

Personality refers to unique patterns displayed in an individual's behavior, cognitive process, and emotional expression. Personality detection aims to analyze these patterns to understand and often predict people's behaviors and inclinations. Understanding and grasping personality traits in social interactions facilitate effective communication in human interactions. Recently, there has been a growing trend in sharing opinions and emotions in textual form on social media platforms [1], [2]. Analyzing such text data to comprehend an individual's trait supports personalized recommendations [3], [4], targeted marketing [5], and facilitates human-computer interaction [6] in the fields of machine learning (ML) and natural language processing (NLP) [7].

The task of personality detection encounters various challenges, with dataset construction standing out as a primary obstacle. Generating data pairs that label a user's speech or written text with their personality traits requires significant time and resources. Furthermore, the inherent difficulty of laypeople in assessing personality exacerbates the challenge. In such circumstances, expert interpretation becomes indispensable for deciphering an individual's accurate personality traits [8]. Moreover, the labeled personality traits often rely

The associate editor coordinating the review of this manuscript and approving it for publication was Seifedine Kadry.

on self-reported survey outcomes, such as NEO-PI-3 [9] or Myers-Briggs Type Indicators (MBTI) [10]. Datasets based on these self-assessment methods can pose challenges for automatic personality detection due to respondent's subjective interpretations.

To address these challenges, previous approaches have used traditional machine learning [11], [12], word embeddings like Word2Vec [13], FastText [14], and psychological features [15] techniques. With the advancement of language models (LM), diverse approaches incorporating LM have been proposed [15], [16]. The growing feasibility of integrating contextual information has led to an increased adoption of pre-trained LM, such as BERT [17]. Building upon established detection methods utilizing LM, we propose a method that incorporates a curriculum-based approach. Curriculum learning [18] mirrors the gradual learning process of humans, which proceeds from simple to complex tasks. The process entails training models on easy data first and progressively introducing more challenging data. The key element lies in defining an appropriate score that determines the level of difficulty of the data. Commonly employed methods in the text domain of NLP encompass factors such as sentence length [19], word frequency [20], conjunction [21], and parse tree depth [22]. Although these approaches are easy-to-use, there are limited curriculum methods tailored for personality detection.

Hence, we propose two curriculum strategies, namely weighted averaged-based and cumulative sum-based methods. These approaches are constructed using a class based-term frequency-inverse document frequency (c-TF-IDF) [23] adjustment to the traditional TF-IDF [24] method at the class level. While the typical TF-IDF method is utilized to find relevant information across multiple documents, the c-TF-IDF approach is tailored to be effective for data where classes exist. We integrate c-TF-IDF to extract informative words within the data, utilize them in the acquisition of the curriculum score and differentiate the difficulty assessment process by employing average and sum methods for comparative experiments. According to the experimental results compared with various curriculum strategies reveal that our proposed approach surpasses existing methods. The experiments also provide insights into determining a suitable method for utilizing c-TF-IDF scores.

Our contributions can be summarized as follows:

- We utilized c-TF-IDF on text-based personality datasets to extract the importance of informative words. This approach operates at the class level, effectively discerning words associated with each trait.
- Through the implementation of a prioritization function for individual words and a curriculum score algorithm to gauge data difficulty, we devised an effective curriculum learning strategy.
- Our experimental results demonstrate that the proposed strategy enhances personality detection performance compared to previous methods.

II. RELATED WORKS

A. PERSONALITY DETECTION

Research on detecting and predicting an individual's personality is widespread across various fields because it provides insights into user behavior, preferences, and decision making [25]. Employing various methodologies, automatic personality detection has garnered attention for its perceived superiority compared to human assessments [26]. This study focuses on text-based personality data, including user speeches, essays, and social media posts annotated with self-reported personality traits.

Conventional machine learning methods, including the support vector machine (SVM) [27], random forest (RF) [28], and K-nearest neighbors (KNN) [29], have been employed for personality recognition. Wang et al. [11] improved classification performance, addressing challenges such as a small sample size and severe sample distribution imbalance. They utilized particle swarm optimization (PSO) and synthetic minority oversampling technique tomek link (SMOTETomek) methods, combining feature optimization and data resampling. In their study, SVM and KNN served as classification models. Moraes et al. [8] used personality detection to understand candidates' ideas and ideals during the job hiring process. The decision tree [30], Naive Bayes [31], SVM, and KNN were employed. They utilized TF-IDF vectorization to calculate word importance within the data, incorporating these weights into word embeddings.

With the emergence of high-performing pre-trained language models like BERT, researches have conducted studies that leverage them for personality detection. Kazameini et al. [32] employed contextual word embeddings from BERT with bagged SVM to achieve outstanding detection performance. Mehta et al. [15] leveraged both traditional psychological features and BERT embeddings for personality detection. Their experimental results revealed that utilizing a language model was more effective than psychological features. This emphasizes the importance of leveraging contextual information. Kazemeini et al. [16] identified the drawbacks of current personality detection methods by leveraging embeddings from big-five inventory (BFI) statements [33], a self-reported questionnaire defining personality traits. A considerable embedding distance between BFI statements and the state-of-the-art model [15] was observed. Attributing this to the loss of semantic information in sentences, the study adopted sentence-BERT [34] to derive sentence embeddings. Although they did not observe a significant improvement in performance, their study enhanced the interpretability of the process.

In recent research, Zhou et al. [35] highlighted the constraints associated with relying on a limited set of text features for classification. The study redirected its attention to the prevalent use of emojis across various social media platforms. In contrast to using traditional embeddings derived from sentences or words for classification, this study integrated emoji embeddings into data. It underscored the significance of extracting valuable features. Building upon studies investigating the relationship between personality traits and emotions, Li et al. [1] introduced a novel multi-task framework for enhanced performance. Lin et al. [36] critiqued the predominant emphasis on language models in existing detection methods, noting their tendency to overlook sentiment information. Their study also pointed out inconsistencies in the current selection process of psycholinguistic features. To address these issues, the study proposed a method based on high-dimensional psycholinguistic features and introduced an improved distributed gray wolf optimizer feature selection (IDGWOFS) process as an additional feature selection approach.

B. CURRICULUM LEARNING

Curriculum learning [18], inspired by the human learning process, involves training a model on simple data first and progressing to complex data. It has shown notable performance improvements in terms of the learning convergence speed and overall performance of the model [18]. Various methods can be used to define simple datasets during the learning process [37], [38]. In this paper, we aim to explore several methods that are well-utilized in the text domain.

The analysis of sentence length is a widely adopted method in NLP that is influenced by the concept of humans learning from shorter to longer texts [19], [21], [39], [40]. Spitkovsky et al. [39] implemented two curriculum learning approaches, namely "Baby steps" and "Less is more". The "Baby steps" approach adopts an incremental learning strategy, starting with easy data and gradually advancing to more difficult data. This method involves a step-by-step increase in the amount of data with each learning iteration. In line with these observations, the "Less is more" approach proposed by Bengio et al. [18] suggests that simpler samples can occasionally convey richer information. Kocmi and Bojar [21] employed a straightforward method, dividing data based on sentence length and setting thresholds such as 8 tokens or 16 tokens to create bins. The learning process commenced with shorter length bins and progressed accordingly.

Another widely used method is the word rarity approach, which is rooted in the concept of learning from a small and easy-to-understand set of words when studying or reading books [20], [21], [40], [41]. Various strategies for applying this learning approach have been developed. Kocmi and Bojar [21] ranked words based on their frequency of occurrence within the data in descending order. For instance, 5,000 of the most frequently occurring words were assigned to rank one. Then, sentences containing words in rank one were placed in the first bin for initial learning, and subsequent bins were determined based on the subsequent ranks. Platanios et al. [40] calculated the frequency of all unique words composing the data, sorted them in descending order, computed the difficulty score for each word in a sentence, and averaging them. This difficulty score was then utilized in the final data sorting process.

Based on the sentence structure, conjunctions can be used to determine the difficulty score [21], [22]. The difficulty score is commonly defined by the number of conjunctions. Kocmi and Bojar [21] determined the difficulty score using coordinating conjunctions, exploiting the idea that although sentences with many conjunctions may be easy to interpret, the hierarchical structures of the sentences can make them challenging for models to learn. Another approach involves using prepositional phrases. Tsvetkov et al. [22] exploited the idea that an increase in the number of conjunctions results in higher sentence complexity. Such sentences are more challenging for a model to comprehend at once.

Beyond the aforementioned methods widely used in the field, numerous curriculum methods have been proposed in the NLP text domain. In the domain closely related to personality, namely emotion recognition, Yang et al. [42] suggested conversation-level curriculum (CC) and utterance-level curriculum (UC) methods to enhance learning performance in emotion recognition tasks. CC defines the difficulty score based on the degree of emotion shift within a conversation and is tailored to tasks involving emotion recognition within dialogues. It is based on the notion that frequent changes in emotions during a conversation pose a challenge in terms of data reliability. UC does not differentiate between similar emotions (e.g., excitement and happiness) during the early stages of learning. It gradually distinguishes them through learning to grasp more nuanced emotions as it evolves. In intent detection field, Gong et al. [43] pointed out the risk of overfitting to easy samples during training and proposed a density-based method. It measures the eigenvector's density for each sample, utilizes a density based clustering algorithm to determine the difficulty level of each data point, and adjusts the training from simple to complex samples over the learning process.

III. PROPOSED METHODOLOGY

We propose two curriculum learning strategies based on the c-TF-IDF scores: a *weighted average-based* approach and a *cumulative sum-based* approach. Figure 1 illustrates the organizational procedure for our curriculum learning strategy. B-1 shows the weighted average-based approach, and B-2 shows the cumulative count-based approach. Both methodologies follow a sequence involving the sequence adjustment with curriculum strategy subsequent to the extraction of informative words. However, there are differences in the priority function assigning priority values to input words and in the calculation function of the curriculum score for each data point. These variances are elaborated upon in the following section.

A. INFORMATIVE WORDS EXTRACTION

In Figure 1, step A shows the process of extracting informative words. Let $\mathcal{D}_i$ be the dataset annotated for a trait i (such as OPN, I/E), and the value (0 or 1) of that trait be the class c of $\mathcal{D}_i$ (A.1). We calculate the c-TF-IDF score to extract informative words for each class c of $\mathcal{D}_i$'s (A.2).

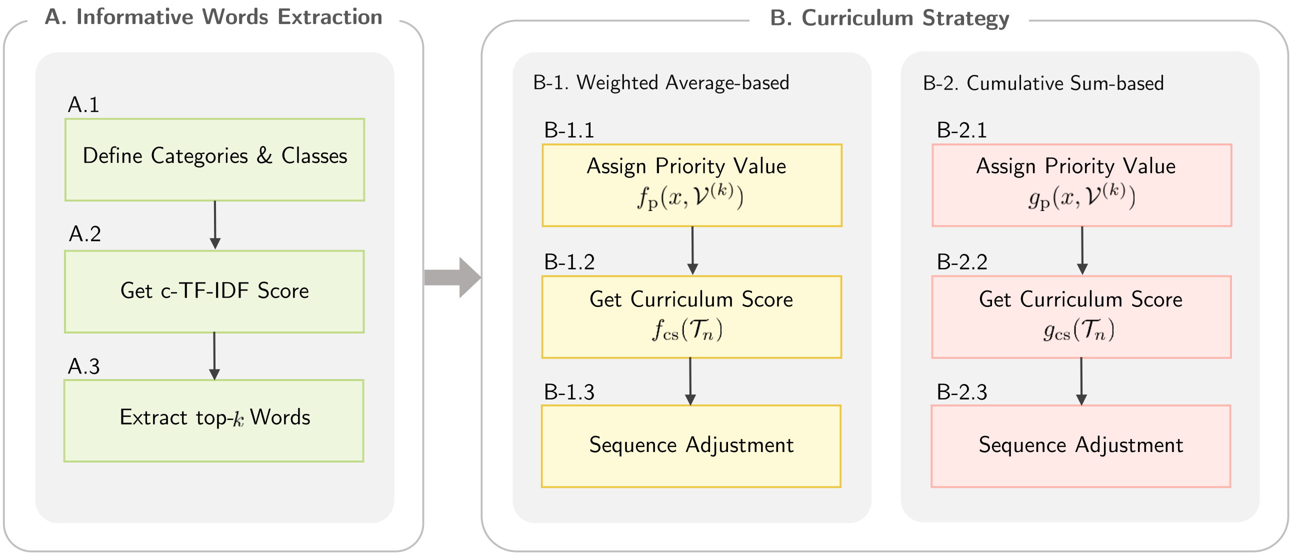


FIGURE 1. Overview of our organization procedure for curriculum learning. Step A involves the extraction of informative top-k percent words for all unique words through c-TF-IDF. Subsequently, c-TF-IDF is employed. There are two optional methodologies for achieving the curriculum strategy: the weighted average-based approach and cumulative sum-based approach. Following sequence adjustment, an ordered dataset reflecting the difficulty of the data is obtained based on the curriculum score.

Unlike the conventional document-level TF-IDF approach, the c-TF-IDF is adjusted to operate at the class (topic or category) level, utilizing the priority of specific words within a class. This adjusted expression of TF-IDF, known as c-TF-IDF, can explain that one class document differs from the other. The c-TF-IDF score s for a word x in a class c can be defined as follows:

$$s_{x,c} = f_{x,c} \times \log\left(\frac{A}{f_x}\right) \tag{1}$$

where $f_{x,c}$ represents the class-based term frequency of the word x in a specific class c, f_x denotes the frequency of the word x across all classes, and A is the average number of words per class. Then, we extract the top-k percent of significant words based on calculated c-TF-IDF scores (A.3). The set of words comprising k percent of all the unique words extracted from class c of $\mathcal{D}_i$ is defined as vocabulary set $\mathcal{V}_{i,c}^{(k)}$. We leverage the c-TF-IDF score to capture the distinctiveness of words between specific classes and their relevance to characterize personality traits.

B. CURRICULUM STRATEGY

1) WEIGHTED AVERAGE-BASED STRATEGY

The vocabulary set consisting of the top-k percent words extracted in step A is denoted as $\mathcal{V}^{(k)}$. Then, each of these words is assigned a priority value according to a priority function (B-1.1). For input word x, the priority function $f_{\mathrm{p}}(x, \mathcal{V}^{(k)})$ can be defined as follows:

$$f_{\mathrm{p}}(x, \mathcal{V}^{(k)}) = 1 + \frac{1 - \mathrm{rank}_{\mathcal{V}^{(k)}}(x)}{k}, \tag{2}$$

where $\mathrm{rank}_{\mathcal{V}^{(k)}}(x)$ returns the rank of the word x with reference to the vocabulary set $\mathcal{V}^{(k)}$. Through this function, we can determine the weight of word x.

Let $\mathcal{V}_{i,c}^{(k)}$ be the top-k percent vocabulary set for class c of a trait i, and $\mathcal{T}_n$ be the set of tokens comprising a n-th data point within $\mathcal{D}_i$. The curriculum score for a data point $\mathcal{T}_n$ can be obtained using weighted values from Eq. (2) by the curriculum score function f_{cs} as follows (B-1.2):

$$\mathcal{W}_n = \mathcal{V}_{i,c}^{(k)} \cap \mathcal{T}_n, \qquad \mathcal{W}_n = \{w_1, w_2, \ldots, w_m\} \tag{3}$$

$$f_{\mathrm{cs}}(\mathcal{T}_n) = \frac{\sum_{j=1}^{m} f_{\mathrm{p}}(w_j, \mathcal{V}_{i,c}^{(k)})}{m}. \tag{4}$$

Here, $\mathcal{W}_n$ is the intersection of $\mathcal{V}_{i,c}^{(k)}$ and $\mathcal{T}_n$, and m represents the number of elements in $\mathcal{W}_n$. The resulting score aligns with the average of word weights included in $\mathcal{T}_n$. Consequently, we conduct sequence adjustment following the order of the highest curriculum score (B-1.3).

2) CUMULATIVE SUM-BASED STRATEGY

In contrast to the previous section, the allocation of priority values in this section differs by employing the c-TF-IDF score s rather than the weight of input word x. When considering the set of c-TF-IDF scores for the top-k percent words extracted in the preceding step A as $\mathcal{S}^{(k)}$, the priority function $g_{\mathrm{p}}(x, \mathcal{S}^{(k)})$ for assigning the priority value of the input word x can be defined as follows (B-2.1):

$$g_{\mathrm{p}}(x, \mathcal{S}^{(k)}) = \frac{s_x - \min(\mathcal{S}^{(k)})}{\max(\mathcal{S}^{(k)}) - \min(\mathcal{S}^{(k)})}, \tag{5}$$

where s_x denotes the c-TF-IDF score associated with the input word x. The resulting value is normalized to fall within a range between 0 and 1, ensuring consistency in the priority allocation process.

Let $\mathcal{S}_{i,c}^{(k)}$ be the set of top-k percent c-TF-IDF scores for class c of a trait i, and $\mathcal{T}_n$ be the set of tokens comprising

TABLE 1. Comparison of personality detection accuracy (%) between the existing curriculum and proposed methods. The values highlighted in bold face indicate the highest performance for each category within the divided sections. The text underlined in bold face signifies the highest average value across each dataset. The curriculum strategies proposed in this paper are denoted by W for the weighted average-based strategy and C for the cumulative sum-based approach. The numbers are associated with the parameter *k*.

	Essays						Kaggle MBTI				
	OPN	CON	EXT	AGR	NEU	Average	I/E	N/S	F/T	P/J	Average
RAN	61.58	56.48	55.75	57.61	58.43	57.970	77.24	86.30	69.07	61.26	73.468
CO [21]	66.31	63.47	62.57	63.96	62.41	63.744	77.89	86.49	71.52	63.55	74.863
LEN [40]	62.36	57.09	58.15	58.96	56.12	58.536	77.29	86.30	67.98	61.89	73.365
WR [40]	65.66	**63.55**	**64.04**	64.52	**62.74**	**64.102**	77.91	**86.51**	71.55	**63.96**	**74.983**
PS [22]	**67.36**	61.85	63.22	**64.69**	62.09	63.842	**78.33**	86.49	**71.66**	63.43	74.978
W-10	66.80	62.42	65.34	64.12	62.74	64.284	78.08	86.35	71.46	64.13	75.005
W-25	66.64	63.89	62.66	**65.51**	**64.61**	64.662	**78.42**	86.35	71.16	**64.45**	75.095
W-50	**66.97**	**64.94**	**66.39**	63.31	63.80	**65.082**	78.14	86.40	70.56	63.78	74.720
W-75	66.07	60.31	64.86	64.78	63.72	63.948	78.07	86.40	71.75	64.42	**75.160**
W-100	65.86	62.69	65.13	64.40	62.85	64.186	78.19	**86.47**	**71.88**	63.69	75.058
C-10	67.29	**64.69**	**64.37**	64.61	**63.96**	**64.984**	77.99	86.35	71.62	**64.59**	**75.138**
C-25	**67.45**	61.77	62.99	**65.18**	62.50	63.978	78.19	86.47	71.41	63.18	74.813
C-50	67.05	63.72	64.21	64.53	63.80	64.662	78.08	86.33	**71.88**	63.55	74.960
C-75	66.40	64.53	**64.37**	64.53	63.47	64.660	**78.31**	86.38	71.42	63.69	74.950
C-100	66.56	64.53	**64.37**	64.53	63.47	64.692	77.89	**86.65**	71.12	64.18	74.960

TABLE 2. Standard deviation of c-TF-IDF scores for category-specific extracted vocabulary sets based on parameter *k* in the Essays dataset.

	OPN		CON		EXT		AGR		NEU		Average
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
k=10	5.50E-04	5.98E-04	5.76E-04	5.68E-04	5.79E-04	5.64E-04	5.84E-04	5.58E-04	5.84E-04	5.60E-04	5.72E-04
k=25	3.67E-04	3.99E-04	3.85E-04	3.79E-04	3.86E-04	3.77E-04	3.90E-04	3.73E-04	3.89E-04	3.74E-04	3.82E-04
k=50	2.66E-04	2.88E-04	2.78E-04	2.74E-04	2.79E-04	2.72E-04	2.82E-04	2.70E-04	2.81E-04	2.71E-04	2.76E-04
k=75	2.19E-04	2.37E-04	2.29E-04	2.25E-04	2.30E-04	2.24E-04	2.32E-04	2.22E-04	2.31E-04	2.23E-04	2.27E-04
k=100	1.91E-04	2.06E-04	1.99E-04	1.96E-04	2.00E-04	1.95E-04	2.02E-04	1.93E-04	2.01E-04	1.94E-04	1.98E-04

an n-th data point within $\mathcal{D}_i$. The curriculum score for a data point $\mathcal{T}_n$ can be calculated using Eq. (5). The curriculum score function g_{cs} is given as follows (B-2.2):

$$g_{\text{cs}}(\mathcal{T}_n) = \sum_{j=1}^{m} g_{\text{p}}(w_j, \mathcal{S}_{i,c}^{(k)}). \tag{6}$$

Here, the resulting score aligns with the cumulative sum of priority values included in $\mathcal{T}_n$. Similar to B-1, we conduct sequence adjustment following the order of the highest curriculum score (B-2.3).

The curriculum learning strategies introduced in this section diverge in their emphasis: the weighted average-based approach assesses the collective significance within data points, while the cumulative sum-based approach emphasizes the importance of words contained in those data points rather than their overall significance. These strategies indicate that a higher curriculum score corresponds to a robust representation of category-specific words within the data point. As a result, these data points can be categorized as '*easier*' based on their sorted order, reflecting the increased presence of informative terms related to the targeted category.

IV. EXPERIMENTS AND DISCUSSION

A. DATASET

In this study, we utilize publicly available datasets to evaluate personality detection performance. The specific number of data points corresponding to each dataset is shown in Table 4:

Essays [44]: This dataset comprises 2,467 essays authored by anonymous students, written as a stream of consciousness. The essays are labeled with binary labels ("y" and "n") derived from self-reported surveys conducted on the authors, following the Big-Five personality traits model (also known as the Five Factor Model - FFM) [45]. It consists of openness to experience (OPN), conscientiousness (CON), extraversion (EXT), agreeableness (AGR), and neuroticism (NEU).

Kaggle MBTI[1]: This dataset comprises 8,675 unique values collected from the PersonalityCafe forum, a dataset based on MBTI. It includes users' self-reported MBTI types and the most recent 50 posts authored by users on the website. The MBTI type is defined by four scales. Introversion or extroversion (I/E), intuition or sensing (N/S), feeling or thinking (F/T), and perceiving or judging (P/J).

[1] https://www.kaggle.com/datasnaek/mbti-type

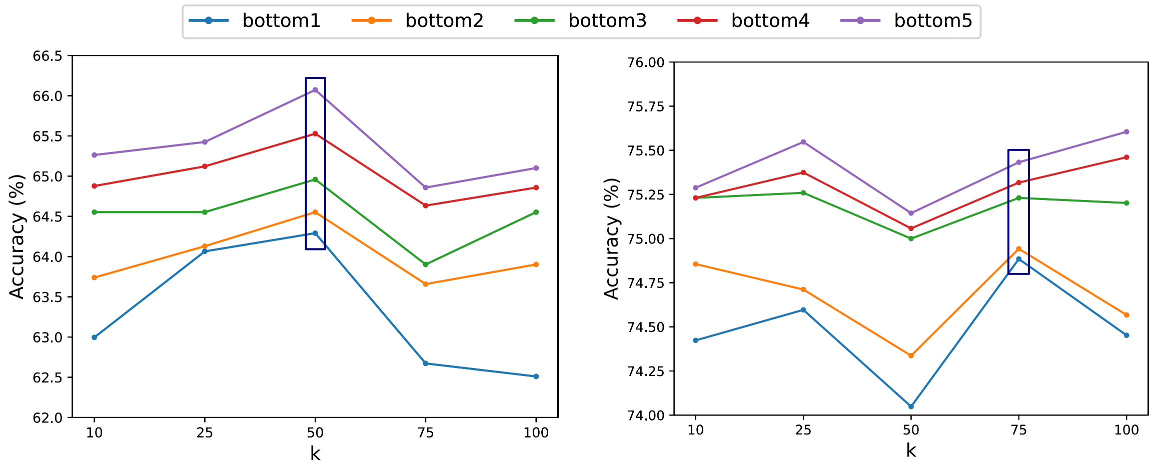


FIGURE 2. Performance differences in personality detection accuracy for the bottom five folds based on the weighted average-based approach across parameter *k* values. (a) represents the performance variations in the Essays dataset, while (b) signifies the performance variations in the Kaggle MBTI dataset.

TABLE 3. Standard deviation of the c-TF-IDF scores for category-specific extracted vocabulary sets based on parameter *k* in the Kaggle MBTI dataset.

	I/E		S/N		T/F		P/J		Average
	Left	Right	Left	Right	Left	Right	Left	Right	
k=10	2.38E-04	2.45E-04	2.47E-04	2.39E-04	2.30E-04	2.50E-04	2.40E-04	2.39E-04	2.41E-04
k=25	1.56E-04	1.60E-04	1.61E-04	1.56E-04	1.51E-04	1.63E-04	1.57E-04	1.56E-04	1.58E-04
k=50	1.12E-04	1.15E-04	1.16E-04	1.12E-04	1.08E-04	1.17E-04	1.13E-04	1.12E-04	1.13E-04
k=75	9.16E-05	9.41E-05	9.49E-05	9.17E-05	8.86E-05	9.57E-05	9.23E-05	9.20E-05	9.26E-05
k=100	7.95E-05	8.17E-05	8.24E-05	7.96E-05	7.69E-05	8.31E-05	8.01E-05	7.98E-05	8.04E-05

TABLE 4. Number of data per label in each dataset. For the Kaggle MBTI dataset, label "left" and "right" represent I and E in I/E, respectively, and the other personality traits are the same.

Dataset	Essays		Dataset	Kaggle MBTI	
Label	Yes (y)	No (n)	Label	Left	Right
OPN	1,271	1,196	I/E	6,676	1,999
CON	1,254	1,213	S/N	1,197	7,478
EXT	1,275	1,192	T/F	3,981	4,694
AGR	1,309	1,158	P/J	5,241	3,434
NEU	1,234	1,233	-	-	-
Total	2,467		Total	8,675	

B. EXPERIMENTAL DETAILS

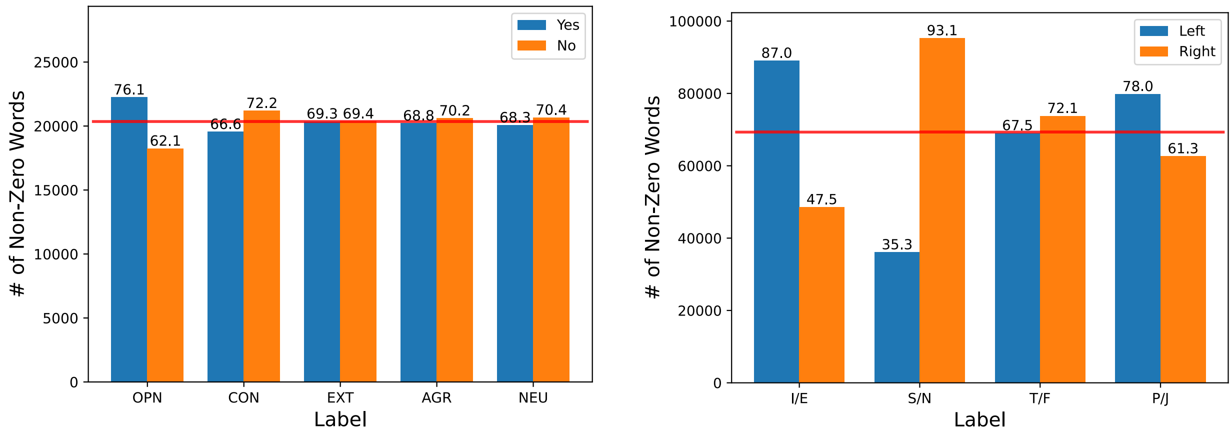
To validate the effectiveness of the two proposed curriculum methods, we established the following comparisons.

- **Random (RAN)**: This represents the most basic approach, where the learning order is randomized by shuffling.
- **Coordinate Conjunction (CO)** [21]: This method considers the count of coordinating conjunctions (e.g., and, or, but) within a sentence. The part of speech (POS) tagger function from the NLTK[2] library is used in this process.
- **Sentence length (LEN)** [40]: A widely used curriculum method in NLP, it involves tokenizing each sentence and using the number of tokens to determine sentence length, which is then employed as the curriculum score.
- **Word rarity (WR)** [40]: This method utilizes the frequency of words in the dataset. First, frequencies are computed for all unique words, and then, for each sentence, the average frequency value for all composing words is used as the curriculum score. The CountVectorizer from sklearn[3] is employed for frequency computation.
- **Prepositional/Subordinating Conjunction (PS)** [22]: Similar to the coordinate method, PS uses POS tags to determine the count of relevant words (e.g., on, with, before) for learning.

Similar to the configuration of [15] for personality detection, we employed a multi-layer perceptron (MLP) with 50 hidden units and the rectified linear unit (ReLU) activation function [46]. Preprocessing steps were employed to eliminate data noise by removing hyperlinks, hashtags, smileys, emojis, and spaces. We employed the BERT-base pre-trained language model for word embedding and set a maximum of 512 tokens from the beginning of each

[2] https://www.nltk.org/

[3] https://scikit-learn.org/stable/



(a) Essays

(b) Kaggle MBTI

FIGURE 3. Proportion of non-zero c-TF-IDF scored words among class-specific unique words in the datasets. (a) represents the Essays dataset, (b) represents the Kaggle MBTI dataset, and the solid red line indicates the average value across all classes.

TABLE 5. Complementary experiment results for searching parameter k. The average value obtained by considering the number of non-zero c-TF-IDF words per class was utilized as the parameter k, indicating W-avg.

		WR	W-50	W-avg
Essays	OPN	65.66	**66.97**	66.64
	CON	63.55	**64.94**	63.23
	EXT	64.04	**66.39**	64.85
	AGR	64.52	63.31	**64.94**
	NEU	62.74	63.80	**64.29**
	Average	64.102	**65.082**	64.790
		WR	W-75	W-avg
Kaggle MBTI	I/E	77.91	78.07	**78.11**
	S/N	86.51	86.40	**86.53**
	T/F	71.55	**71.75**	71.70
	P/J	63.96	**64.42**	64.25
	Average	74.983	**75.160**	75.148

sentence. We used the optimizer Adam [47], coupled with a categorical cross-entropy loss function. To ensure fairness in the comparison process, we introduced our own random seed set to zero. In addition, to prevent data bias, we employed stratified 10-fold cross-validation. Mitigating the impact of dataset idiosyncrasies observed in certain folds on model evaluation, we considered average values based on the bottom five folds, rather than the entire folds. In the experiment of proposed methods, parameter k – utilized for extracting the top-k percent vocabulary set – was configured at 10, 25, 50, 75, and 100.

C. RESULTS

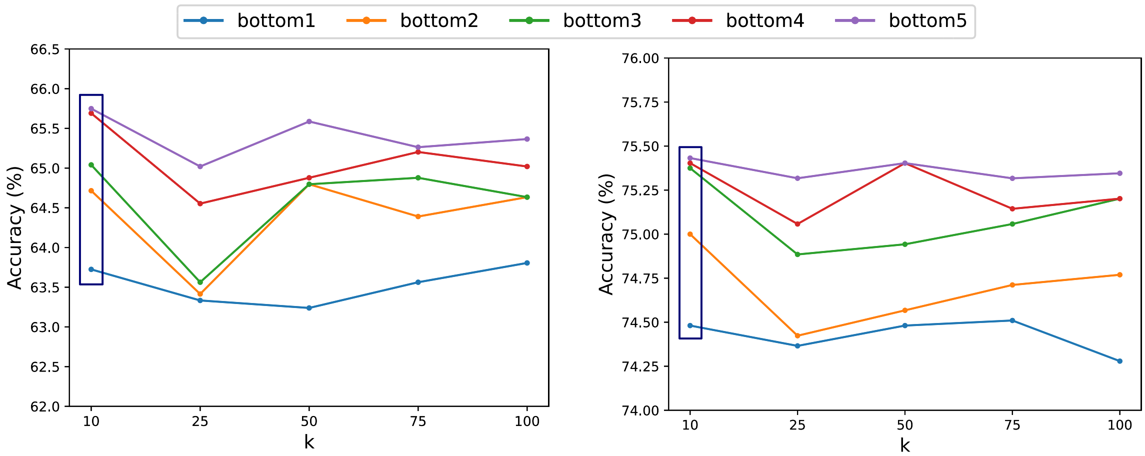
Table 1 lists the personality detection performance for each dataset. Our proposed weighted average-based approach denoted as W-k demonstrated the highest performance out of all the other methods. In this section, we provide an analysis of the results obtained from each proposed methodology.

1) WEIGHTED AVERAGE-BASED APPROACH

In Table 1, the Essays dataset achieved the highest performance at 50 with 65.082%, while the Kaggle MBTI dataset peaked at 75 with 75.160% compared to conventional curriculum methods. This was an improvement of 7.112%p and 1.692%p over the RAN method. When compared to the WR method, which achieved the highest performance in the comparative experiments, the improvements were 0.98%p and 0.18%p. When compared to the existing CO, LEN, and PS methods, there was an average improvement of about 3.04%p and 0.76%p.

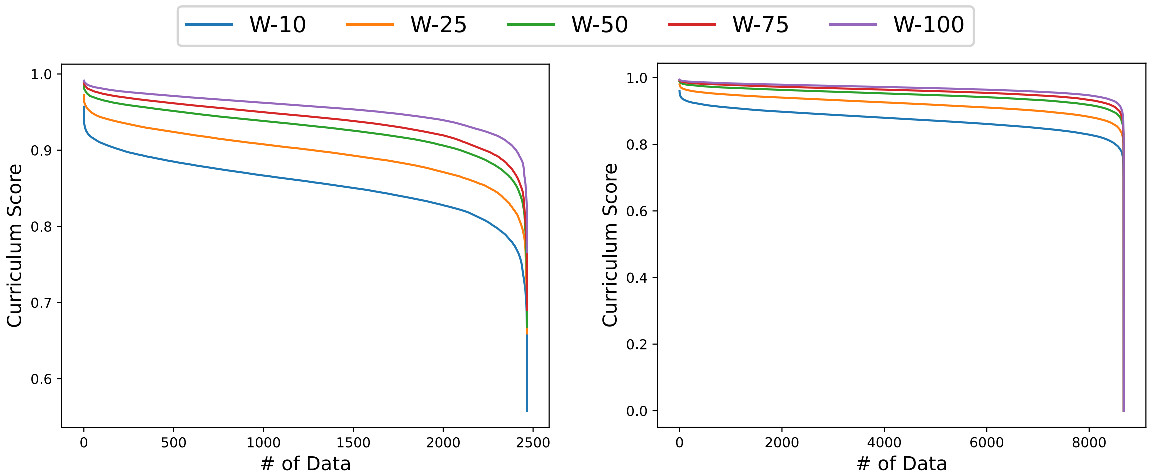
However, the performance improvement in the Kaggle MBTI dataset was marginal compared to that of the Essays dataset. We regarded this as a result stemming from the inherent characteristics of the datasets. The Kaggle MBTI dataset, shown in Table 4, was an imbalanced dataset skewed towards one side in the I/E and S/N traits. In imbalanced situations, models commonly tend to exhibit higher accuracy for the majority class. This trend was reflected in the performance results of Table 1, where I/E and S/N traits demonstrated higher accuracy on average compared to others. Furthermore, the Kaggle MBTI dataset contained fewer meaningful words than the Essays dataset. Table 2 and Table 3 represent the standard deviation of the c-TF-IDF scores for each parameter k. A higher standard deviation implies significant differences in the c-TF-IDF scores among extracted words, indicating meaningful differences. The standard deviation for the Essays dataset was on average 2.431 times higher for all classes at each k compared to the Kaggle MBTI dataset. This indicated that the words from the Kaggle MBTI dataset had less impact on each trait that of other dataset. In summary, these observations explained why the performance improvement in the Kaggle MBTI dataset was not substantial.

We hypothesized that variations in traits observed within specific datasets would impact the determination of



(a) Essays (b) Kaggle MBTI

FIGURE 4. Performance differences in personality detection accuracy for the bottom five folds based on the cumulative sum-based approach across parameter *k* values. (a) represents the performance variations in the Essays dataset, while (b) signifies the performance variations in the Kaggle MBTI dataset.



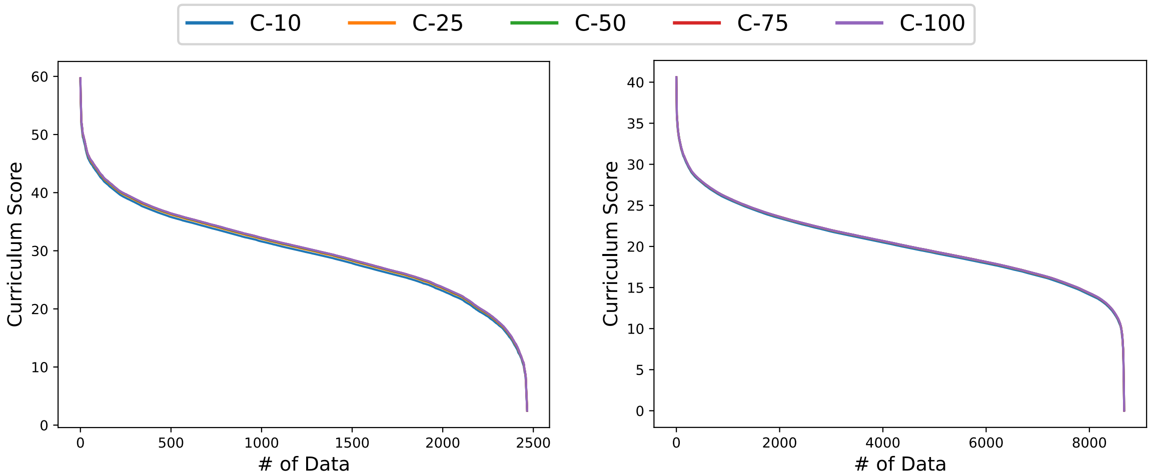
(a) Essays (b) Kaggle MBTI

FIGURE 5. Curriculum score of the data points based on parameter *k* in the weighted average-based approach. (a) represents the Essays dataset and (b) represents the Kaggle MBTI dataset.

parameter k. Figure 2 reveals that k performed best at 50 for the Essays dataset and at 75 for the Kaggle MBTI dataset. Paying attention to the c-TF-IDF scores and performance variations based on the parameters, we examined the c-TF-IDF scores and observed performance variations based on the parameters. Our findings revealed that certain words in the dataset were assigned a score of zero by the c-TF-IDF function. This occurs when words have a very low frequency in a specific class, indicating their lack of significance (Figure 3). To further explore this, we conducted additional experiments based on the count of non-zero c-TF-IDF scores per class. Based on the count of non-zero c-TF-IDF scores per class, we considered the mean values from Figure 3. This led us to set k to 69 for the Essays dataset and 68 for the Kaggle MBTI dataset for further analysis (Table 5). The results indicated that W-avg demonstrated performance improvement in both datasets. When compared to the highest performances in the proposed method, the difference ranged from a minimum of 0.012%p to a maximum of 0.292%p. By understanding the characteristics of the dataset and considering the distribution of c-TF-IDF scores, the selection of an appropriate parameter k may result in a better performance than that of traditional methods.

2) CUMULATIVE SUM-BASED APPROACH

In Table 1, the C approach achieved the highest performance at 10 for both the Essays dataset with 64.984% and the Kaggle MBTI dataset with 75.138%. Compared to those of the RAN method, these values represent an improvement of 7.014%p and 1.67%p. When compared to those of the WR method,



(a) Essays (b) Kaggle MBTI

FIGURE 6. Curriculum score of data points based on parameter k in the cumulative sum-based approach. (a) represents the Essays dataset and (b) represents the Kaggle MBTI dataset.

the improvements were 0.882%p and 0.155%p. Furthermore, compared to the existing CO, LEN, and PS methods, there was an average improvement of approximately 2.94%p and 0.74%p. In Figure 4, the C approach showed minimal variation with the smallest k achieving the highest performance. This observation was explained by examining the curriculum score of the data points (Figure 5 and Figure 6). In the W approach, variations in the results were observed in both datasets according to changes in the value of k, and there was a discernible trend. However, for the C approach, significant variations were not detected. This was attributed to the cumulative nature of the C approach: even words with low curriculum scores were included in the scoring calculation, they did not exert a significant influence. Although this method demonstrated more effective performance compared to conventional methods, it is expected to be more effective for datasets with a larger distribution of scores.

V. CONCLUSION

We presented curriculum strategies for determining data difficulty using class-specific informative words to improve personality detection performance. The experimental results indicated that our proposed method outperformed existing approaches. Given the challenge of detecting intricate individual traits in sparse personality datasets, prioritizing data based on class-specific frequencies is crucial for enhancing accuracy in personality detection.

ACKNOWLEDGMENT

(Naae Kwon and Yuenkyung Yoo contributed equally to this work.)

REFERENCES

[1] Y. Li, A. Kazemeini, Y. Mehta, and E. Cambria, "Multitask learning for emotion and personality traits detection," *Neurocomputing*, vol. 493, pp. 340–350, Jul. 2022.

[2] A. F. Ab. Nasir, E. S. Nee, C. S. Choong, A. S. A. Ghani, A. P. P. A. Majeed, A. Adam, and M. Furqan, "Text-based emotion prediction system using machine learning approach," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 769, no. 1, Feb. 2020, Art. no. 012022.

[3] S. C. Matz, M. Kosinski, G. Nave, and D. J. Stillwell, "Psychological targeting as an effective approach to digital mass persuasion," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 48, pp. 12714–12719, Nov. 2017.

[4] Y. Guan, Q. Wei, and G. Chen, "Deep learning based personalized recommendation with multi-view information integration," *Decis. Support Syst.*, vol. 118, pp. 58–69, Mar. 2019.

[5] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, and M. Ispir, "Wide & deep learning for recommender systems," in *Proc. 1st Workshop Deep Learn. Recommender Syst.*, 2016, pp. 7–10.

[6] Z. Lv, F. Poiesi, Q. Dong, J. Lloret, and H. Song, "Deep learning for intelligent human–computer interaction," *Appl. Sci.*, vol. 12, no. 22, p. 11457, 2022.

[7] X. Zhao, Z. Tang, and S. Zhang, "Deep personality trait recognition: A survey," *Frontiers Psychol.*, vol. 13, May 2022, Art. no. 839619.

[8] R. Moraes, L. L. Pinto, M. Pilankar, and P. Rane, "Personality assessment using social media for hiring candidates," in *Proc. 3rd Int. Conf. Commun. Syst., Comput. IT Appl. (CSCITA)*, Apr. 2020, pp. 192–197.

[9] R. R. McCrae, P. T. Costa Jr., and T. A. Martin, "The NEO–PI–3: A more readable revised NEO personality inventory," *J. Personality Assessment*, vol. 84, no. 3, pp. 261–270, Jun. 2005.

[10] I. B. Myers, *The Myers-Briggs Type Indicator: Manual*. Washington, DC, USA: Consulting Psychologists Press, 1962.

[11] Z. Wang, C. Wu, K. Zheng, X. Niu, and X. Wang, "SMOTETomek-based resampling for personality recognition," *IEEE Access*, vol. 7, pp. 129678–129689, 2019.

[12] D. E. Cahyani and A. F. Faishal, "Classification of big five personality behavior tendencies based on study field with Twitter analysis using support vector machine," in *Proc. 7th Int. Conf. Inf. Technol., Comput., Electr. Eng. (ICITACEE)*, Sep. 2020, pp. 140–145.

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. Int. Conf. Learn. Represent.*, 2013.

[14] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017.

[15] Y. Mehta, S. Fatehi, A. Kazameini, C. Stachl, E. Cambria, and S. Eetemadi, "Bottom-up and top-down: Predicting personality with psycholinguistic and language model features," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2020, pp. 1184–1189.

[16] A. Kazemeini, S. S. Roy, R. E. Mercer, and E. Cambria, "Interpretable representation learning for personality detection," in *Proc. Int. Conf. Data Mining Workshops (ICDMW)*, Dec. 2021, pp. 158–165.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.* Minneapolis, MI, USA: Association for Computational Linguistics, vol. 1, Jun. 2019, pp. 4171–4186.

[18] Y. Bengio, J. Louradour, and R. Collobert, "Curriculum learning," in *Proc. Int. Conf. Mach. Learn.*, Aug. 2009, pp. 41–48.

[19] D. Campos, "Curriculum learning for language modeling," 2021, *arXiv:2108.02170*.

[20] B. Jafarpour, D. Sepehr, and N. Pogrebnyakov, "Active curriculum learning," in *Proc. 1st Workshop Interact. Learn. Natural Lang. Process.*, 2021, pp. 40–45.

[21] T. Kocmi and O. Bojar, "Curriculum learning and minibatch bucketing in neural machine translation," in *Proc. RANLP - Recent Adv. Natural Lang. Process. Meet Deep Learn.*, Nov. 2017, pp. 379–386.

[22] Y. Tsvetkov, M. Faruqui, W. Ling, B. MacWhinney, and C. Dyer, "Learning the curriculum with Bayesian optimization for task-specific word representation learning," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2016, pp. 130–139.

[23] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," 2022, *arXiv:2203.05794*.

[24] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, Jan. 1988.

[25] M. M. Leary, M. D. Reilly, and F. W. Brown, "A study of personality preferences and emotional intelligence," *Leadership Org. Develop. J.*, vol. 30, no. 5, pp. 421–434, Jul. 2009.

[26] W. Youyou, M. Kosinski, and D. Stillwell, "Computer-based personality judgments are more accurate than those made by humans," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 4, pp. 1036–1040, Jan. 2015.

[27] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[28] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.

[29] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.

[30] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986.

[31] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York, NY, USA: Wiley, 1973.

[32] A. Kazemeini, S. Fatehi, Y. Mehta, S. Eetemadi, and E. Cambria, "Personality trait detection using bagged SVM over BERT word embedding ensembles," 2020, *arXiv:2010.01309*.

[33] O. P. John, E. M. Donahue, and R. L. Kentle, "Big five inventory," *J. Personality Social Psychol.*, 1991.

[34] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.

[35] L. Zhou, Z. Zhang, L. Zhao, and P. Yang, "Attention-based BiLSTM models for personality recognition from user-generated content," *Inf. Sci.*, vol. 596, pp. 460–471, Jun. 2022.

[36] H. Lin, C. Wang, and Q. Hao, "A novel personality detection method based on high-dimensional psycholinguistic features and improved distributed gray wolf optimizer for feature selection," *Inf. Process. Manage.*, vol. 60, no. 2, Mar. 2023, Art. no. 103217.

[37] X. Wang, Y. Chen, and W. Zhu, "A survey on curriculum learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4555–4576, Sep. 2022.

[38] Formatted P. Soviany, R. T. Ionescu, P. Rota, and N. Sebe, "Curriculum learning: A survey," *Int. J. Comput. Vis.*, vol. 130, no. 6, pp. 1526–1565, 2022.

[39] V. I. Spitkovsky, H. Alshawi, and D. Jurafsky, "From baby steps to leapfrog: How 'less is more' in unsupervised dependency parsing," in *Proc. Human Lang. Technol., North Amer. Chapter Assoc. Comput. Linguistics*, 2010, pp. 751–759.

[40] E. A. Platanios, O. Stretcu, G. Neubig, B. Poczos, and T. Mitchell, "Competence-based curriculum learning for neural machine translation," in *Proc. Conf. North*, 2019, pp. 1162–1172.

[41] C. Liu, S. He, K. Liu, and J. Zhao, "Curriculum learning for natural answer generation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 4223–4229.

[42] L. Yang, Y. Shen, Y. Mao, and L. Cai, "Hybrid curriculum learning for emotion recognition in conversation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 11595–11603.

[43] Y. Gong, C. Liu, J. Yuan, F. Yang, X. Cai, G. Wan, J. Chen, R. Niu, and H. Wang, "Density-based dynamic curriculum learning for intent detection," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2021, pp. 3034–3037.

[44] J. W. Pennebaker and L. A. King, "Linguistic styles: Language use as an individual difference," *J. Personality Social Psychol.*, vol. 77, no. 6, pp. 1296–1312, 1999.

[45] C. J. McCrae, R. Robert, and T. Paul, *The Five-Factor Theory of Personality*. New York, NY, USA: Guilford Press, 2008.

[46] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*

NAAE KWON received the B.S. and M.S. degrees in electronic engineering from Seoul National University of Science and Technology, South Korea, in 2022 and 2024, respectively. She is currently an Assistant Researcher with Seoul National University of Science and Technology. Her research interests include deep learning, natural language processing, and large language model.

YUENKYUNG YOO received the B.S. degree in electronic engineering from Seoul National University of Science and Technology, South Korea, in 2023, where she is currently pursuing the M.S. degree. Her research interests include deep learning and natural language processing.

BYUNGHAN LEE (Member, IEEE) received the B.S. degree in electrical engineering from Korea University, South Korea, in 2011, and the Ph.D. degree in electrical and computer engineering from Seoul National University, South Korea, in 2018. He is currently an Assistant Professor with the Department of Electronic Engineering, Seoul National University of Science and Technology. His research interests include machine learning, artificial intelligence, and their biomedical applications.