applied sciences

MDPI

Article

University Student Dropout Prediction Using Pretrained Language Models

Hyun-Sik Won [1,†], Min-Ji Kim [1,†], Dohyun Kim [2], Hee-Soo Kim [1] and Kang-Min Kim [1,3,*]

[1] Department of Artificial Intelligence, The Catholic University of Korea, Bucheon 14662, Republic of Korea; abugda@catholic.ac.kr (H.-S.W.); kimmin122@catholic.ac.kr (M.-J.K.); heesu1231@catholic.ac.kr (H.-S.K.)
[2] Department of Computer Science and Engineering, Korea University, Seoul 02841, Republic of Korea; dhkim1028@korea.ac.kr
[3] Department of Data Science, The Catholic University of Korea, Bucheon 14662, Republic of Korea
* Correspondence: kangmin89@catholic.ac.kr
† These authors contributed equally to this work.

Abstract: Predicting student dropout from universities is an imperative but challenging task. Numerous data-driven approaches that utilize both student demographic information (e.g., gender, nationality, and high school graduation year) and academic information (e.g., GPA, participation in activities, and course evaluations) have shown meaningful results. Recently, pretrained language models have achieved very successful results in understanding the tasks associated with structured data as well as textual data. In this paper, we propose a novel student dropout prediction framework based on demographic and academic information, using a pretrained language model to capture the relationship between different forms of information. To this end, we first formulate both types of information in natural language form. We then recast the student dropout prediction task as a natural language inference (NLI) task. Finally, we fine-tune the pretrained language models to predict student dropout. In particular, we further enhance the model using a continuous hypothesis. The experimental results demonstrate that the proposed model is effective for the freshmen dropout prediction task. The proposed method exhibits significant improvements of as much as 9.00% in terms of F1-score compared with state-of-the-art techniques.

Keywords: university student dropout; deep learning; natural language processing; natural language inference; pretrained language model

Citation: Won, H.-S.; Kim, M.-J.; Kim, D.; Kim, H.-S.; Kim, K.-M. University Student Dropout Prediction Using Pretrained Language Models. *Appl. Sci.* **2023**, *13*, 7073. https://doi.org/10.3390/app13127073

Academic Editor: Pengjie Ren

Received: 9 May 2023
Revised: 8 June 2023
Accepted: 12 June 2023
Published: 13 June 2023

1. Introduction

According to a report from the U.S. National Center for Education Statistics, the rate of enrollment in higher-education institutions (e.g., universities) decreased to 66% in 2019 and 63% in 2020 [1]. By contrast, South Korea's enrollment rate increased to 69.4% in 2019 and 71.0% in 2020 (https://www.index.go.kr/, accessed on 15 April 2023), the highest worldwide. However, the average dropout rate among freshmen at universities in South Korea is large at 8.0% (https://www.academyinfo.go.kr/index.do, accessed on 15 April 2023). Students discontinue their studies for various reasons, including a lack of alignment with academic programs, financial constraints, and social relationships. Student dropout negatively affects not only students but also lecturers and universities, which further impacts social and economic costs [2,3]. Therefore, improving student retention is crucial for universities, necessitating the development of a system that predicts student dropout.

Several studies have been conducted to predict university and college student dropout [4–9]. Recent studies [5–7] have utilized machine learning techniques to analyze the causes of university dropout. These studies use only the demographic information of students, such as their gender, region of origin, and family income. These approaches have the advantage of predicting the dropout before the beginning of the first semester. However, students who are not predicted to drop out by models based on demographic

factors may fail to complete the course, and vice versa. For instance, some students may experience issues related to their friendships or encounter challenges when adapting to a university environment. In this case, records of academic performance and activities may play an important role in the dropout prediction task. According to previous studies [8,9], after enrolling in a university, the behaviors of students who drop out and those who successfully complete their degree programs differ. These studies empirically discovered that the predictive performance of their models increased when term exam scores and the number of course credits were included. Following these previous studies, we also utilize various pieces of academic information from students (e.g., GPA, club activities, and course evaluations), as well as demographic information.

Students' demographic and academic information throughout the semester is recorded in various formats. For instance, the GPA and student club participation status are recorded as numerical and categorical data, respectively, whereas course evaluation comments are recorded in textual format. Recently, several studies [10,11] have analyzed unstructured textual data (e.g., advisor notes) to predict student dropout. The study [10] used sentiment analysis to extract positive or negative words from the advisor's notes. The authors then utilized the extracted words as features and trained machine learning models. However, that approach considered only unstructured textual data to predict student dropout. Another study [11] developed a multimodal neural network model that combined structured data (e.g., GPA and SAT scores) and unstructured data (e.g., advising notes and forum posts). The authors showed that fusing structured and unstructured data was effective for predicting student dropout. However, this approach employed two different neural networks (i.e., long short-term memory and transformer) to analyze the structured and unstructured data. It may not capture the implicit relationships between structured and unstructured data.

Recently, in the field of natural language processing (NLP) [12], several studies [13,14] have replaced nontextual information representation with text to represent multimodal information as an integrated modality and leverage the superiority of pretrained language models (PLMs). Several researchers [15–17] began to recast various classification problems into natural language inference (NLI) tasks. The study [15] demonstrated that tabular reasoning performance could be significantly improved by recasting reasoning about tabular information as an NLI task. Another study [16] showed improved performance by training question answering and text summarization datasets by recasting them as NLI datasets.

In this paper, we propose the STUdent Dropout prediction framework (STUD), which is a novel framework that predicts which students will drop out during the semester using demographic and academic information. We first collect data from freshmen (in this study, we focused on freshmen dropout after observing that freshmen had the highest dropout rate of 45.05%) enrolled at The Catholic University of Korea from 2017 to 2021. We formulate both types of information into a natural language format using a template. In addition, we prepend the sentence (e.g., "This student will drop out.") or continuous vectors as "hypothesis" to the student information for recasting the student dropout prediction task into the NLI task. This recast NLI task is motivated by our hypothesis that the prepended hypothesis will effectively elicit the knowledge of university student dropout inherent in the language model. The contributions of this study are summarized as follows:

- We develop a novel framework that predicts university freshmen's dropout using PLMs to capture the relationships between different types of data.
- We recast the student dropout prediction task into a natural language understanding task (i.e., NLI task). To this end, we propose a novel method for modifying each type of record into a natural language form for feeding into a PLM.
- The proposed model achieves superior performance in the dropout prediction of university freshmen compared to state-of-the-art methods.

The remainder of this paper is organized as follows. The methods related to student dropout prediction and models are described in Section 2. Section 3 describes the university student dataset and introduces the proposed framework for university student dropout prediction. We present the performance evaluation results and perform in-depth analyses in Sections 4 and 5, respectively. Finally, we conclude the paper in Section 6.

2. Related Work

2.1. Research on Student Dropout in Higher Education

The amount of information on individual students and that recorded during their education have increased rapidly [18]. Several studies have been conducted to extract and utilize meaningful information from such data. Some studies [19–21] propose methods to predict dropout based on student attendance records in online environments (e.g., MOOC and Coursera). Since these studies are conducted in an online environment, there is no need to consider behavior patterns outside attendance records, such as club activities. Machine learning techniques are effective in the field of education for evaluating student performance, and several studies [4–10,22,23] compare the performance of machine learning methods on the dropout prediction task. The study [6] employed a logistic regression and artificial neural networks to predict the student dropout rate using data of high school students. In a different approach [7], the researchers focused on tree-based decision classification to predict student dropout considering students' economic circumstances. Meanwhile, the study [5] used various machine learning methods to analyze the dropout rate of freshmen engineering students and found that entering university score was a critical variable in predicting dropout. However, these studies only utilize demographic information available before admission, whereas predicting student dropout often requires academic information from universities.

Recent studies [8,24] have shown that academic information can help predict dropout probability. The study [24] applied a generalized mixed-effects random forest to predict the probability of engineering students dropping out based on student-level information and degree programs. Another study [8] used machine learning techniques to predict freshmen dropout using secondary school academic records and first-year course credits. Several researchers [11,25] predict student dropout using unstructured data. The study [25] extracted student sentiment from advisor notes written by student advisors to predict student dropout. In a recent study [11], the authors utilized temporal structured data and unstructured counseling data from all semesters to predict college student dropout using a PLM. However, because these studies train on structured and unstructured data separately, they lack the ability to capture the relationships between the data. In this study, we propose a method for formulating structured data in the format of unstructured data and concatenating them to understand the relationships between the data. To the best of our knowledge, this is the first attempt to capture the relationships between structured and unstructured data in a dropout prediction task.

2.2. Pretrained Language Models Based on Transformer

Recently, transformer-based [26] language models pretrained with large datasets [27–30] have achieved promising results in handling various NLP tasks. In particular, BERT [27], which utilizes the encoder part of the transformer, has demonstrated promising performance in various classification tasks, such as sentiment classification, question answering, and NLI. BERT adopts two pretraining objectives (i.e., masked language modeling (MLM) and next sentence prediction (NSP)) on a large-scale unlabeled text, before fine-tuning the model for a specific downstream task. MLM is an important component in the pretraining of the BERT model. It randomly masks 15% of the tokens from the input sentences and thereafter trains the model to predict the original vocabulary ID of the masked tokens based only on their context (i.e., the remaining tokens). BERT exhibits outstanding performance in most NLP tasks by fine-tuning the model with minimal task-specific architectural modifications (e.g., connecting a fully connected layer to the BERT representation). Subse-

quent studies, such as RoBERTa [30], have improved the performance by removing NSP and modifying the static MLM to a dynamic MLM. In parallel, SimCSE [31] employs a supervised method that enhances pretrained models by incorporating annotated pairs from NLI datasets into the contrastive learning framework. Transformer-based PLMs also work in multimodal tasks [32–35] and tasks built on other types of data formats (e.g., tabular data) [15,36] as well as in NLP. Inspired by these studies, we propose a framework for predicting student dropout by feeding both structured and unstructured data into a model.

3. Methodology

We developed a novel university student dropout prediction framework that uses the demographic and academic information of university freshmen. We first constructed a university student dropout dataset. Our methodology comprised two phases. First, we split the features of each student into demographic and academic information and formulated them into a natural language format. Among these, text data such as course evaluation comments were kept in the form of natural language. Next, we prepended the hypothesis to the formulated student information to recast the student dropout task as an NLI task.

3.1. Data Description

3.1.1. Data Source

We collected data from 7536 undergraduate students enrolled at The Catholic University of Korea from 2017 to 2021. Since our goal was to predict which students would drop out during their freshmen year, we only used data from the first semester for all students. The students consisted of 3073 males and 4463 females, with a total dropout rate of 10%; the rates for males and females who dropped out were 46% and 54%, respectively. As these data were official student information of a university, we did not find any outliers. We used 11 demographic and 7 academic variables. Table 1 lists the description of the collected datasets, and Table 2 lists the statistics of the five numerical data. Figure 1 shows pie charts illustrating the distributions of prominent features in categorical data.

3.1.2. Data Modality

Freshmen data can be expressed using demographic information that a student already possesses before admission (e.g., gender and age) and academic information that indicates the student's behavior during the semester (e.g., GPA and course evaluation). Table 1 describes each feature of the demographic and academic information. We divide the features into three categories based on data type: categorical, numerical, and textual. Categorical features belong to one of two or more categories, numerical features are values within a specific range, and textual features are text data. Categorical features included admission type, detailed admission type, student club, and academic activity. The admission type was determined by the university's admissions system, which consisted of evaluating students based on their Korean SAT scores, evaluating students based on their high school grades and extracurricular activities without taking an entrance exam, and evaluating foreign students. In addition, the detailed admission type was specifically applied to the second admission type (e.g., university admission interview and essay test). The student club indicated whether a student joined a club, while the academic activity indicated their involvement in university-held academic activities. In the numerical features, the parental financial income was calculated by national scholarships students received based on their income level, and the volunteer hours were the number of volunteer hours a student has performed off- and on-campus. Lastly, the course evaluation score, a numerical feature, was one of the course evaluation methods, and the course evaluation comment was written by students and was a textual feature. We formulated all the features in a natural language format that was predesigned for each feature (see Section 3.2.1). Table 3 lists the templates used in the formulation along with an example.

Table 1. Dataset description.

Group	Feature	Type	Classes	Description
Demographic information	Birth year	Categorical	24	The birth year of the student
	Gender	Categorical	2	The gender of the student
	Nationality	Categorical	12	The nationality of the student
	Graduation year	Categorical	23	The year of high school graduation
	High school	Categorical	1564	The graduating high school of the student
	High school region	Categorical	24	The region of the graduating high school
	Admission year	Categorical	5	The year the student was admitted
	Admission type	Categorical	8	The type of admission
	Detailed admission type	Categorical	14	The type of detailed admission
	Major field	Categorical	13	The major field of the student
	Parental financial income	Numerical	-	The income level divided by scholarships received
Academic information	Major GPA	Numerical	-	The average grade point of majors
	Nonmajor GPA	Numerical	-	The average grade point of nonmajors
	Volunteer hours	Numerical	-	The hours of volunteering
	Student club	Categorical	2	Whether one participates in a student club or not
	Academic activity	Categorical	2	Whether one participates in an academic activity or not
	Course evaluation score	Numerical	-	Evaluation score for courses taken
	Course evaluation comments	Textual	-	Evaluation comments for courses taken

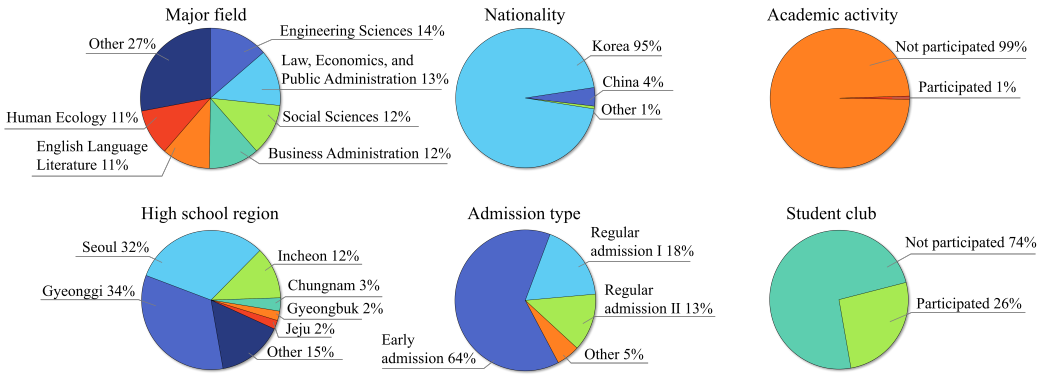

Figure 1. Distributions for the categorical data expressed as a percentage.

Table 2. Statistics of numerical data. The parental financial income is given as a percentile.

Feature	Mean	SD	Median	Min	Max
Major GPA	3.27	1.15	3.5	0.0	4.5
Nonmajor GPA	3.31	1.1	3.58	0.0	4.5
Volunteer hours	1.99	10.77	0	0	130
Parental financial income *	55.07	27.21	63	8	85
Course evaluation score	4.2	0.41	4.27	2.35	5.0

* The parental financial income is calculated only with data from students who received a national scholarship.

Table 3. An example of formulating each feature into a natural language format.

Group	Feature	Natural Language Form
Demographic information	Birth year	Born in 2002
	Gender	Male
	Nationality	South Korean nationality
	Graduation year, high school, high school region	Graduated in 2021 from Yeokgok High School, a public high school in Yeokgok-dong, Bucheon-si, Gyeonggi-do
	Admission year, admission type, detailed admission type	Admitted in 2021 with early admission —writing test admission
	Major field	Major field is natural sciences
	Parental financial income	Below 53%, similar to average students
Academic information	Major GPA	4.17 GPA in major, excellent
	Nonmajor GPA	3.5 GPA in nonmajor, normal
	Volunteer hours	No volunteer activity
	Student Club	Participated in a student club
	Academic activity	Did not participate in an academic activity
	Course evaluation score, course evaluation comments	Course evaluation : 3.93, negative, too many assignments

3.2. Modeling

3.2.1. Formulating Numerical and Categorical Variables into a Natural Language Format

We formulated the numerical and categorical variables of freshmen in a natural language format to leverage the knowledge of the PLM for the student dropout prediction task. We denoted the demographic variables of freshmen as $D = \{d_i | d_i \in \mathbb{R}, i = 1, 2, \ldots, n_d\}$, where n_d denotes the number of demographic variables. To exploit the textual understanding power of PLMs, we formulated all these variables into a natural language format and created sentences. We created a sentence using the mapping function $f_d(D) = D^*$, where D^* is a sequence consisting of the vocabulary of the PLMs. Similarly, we denoted the academic variables as $A = \{a_i | a_i \in \mathbb{R}, i = 1, 2, \ldots, n_a\}$, where n_a is the number of academic variables. We then formulated them into a sentence using $f_a(A) = A^*$. In Table 3, we used various formats depending on the specificity of each feature. First, for the categorical type, we formulated a different sentence format based on the value of the feature, excluding gender. For instance, we formulated the birth year, such as "2002", as in the sentence "Born

in 2002". In contrast, we used the gender as is, such as "Male" in the sentence. Additionally, for the categorical type indicating participation status, we formulated two types of sentences. For instance, we formulated the student club, such as "yes" which means the student participated in a student club, as "Participated in a student club", otherwise, we used "Did not participate in student club". To leverage the regional information of the pretrained model learned from the general corpus during pretraining, we included the high school location from Wikipedia if possible. We formulated the high school, such as "Yeokgok High School", with regional information as in "Yeokgok High School, a public high school in Yeokgok-dong, Bucheon-si, Gyeonggi-do". We bound features that were more naturally expressed when described together (e.g., university-admission-specific features and high-school-specific features). For instance, we formulated university-admission-specific features such as admission year "2021", admission type "early admission", and detailed admission type "writing test admission", as in the sentence "Admitted in 2021 with early admission-writing test admission".

One study [36] showed the effectiveness of using numeric values and text together. Therefore, we statistically divided the numeric features into five categories using natural language and described them using categories. For instance, we formulated the major's GPA, such as "4.17", as the sentence "4.17 GPA in major, excellent", because a score above 3.75 and below 4.25 corresponds to "excellent". We identified parental financial income using any national scholarship the student received. Because tuition varies by major, students who received the same number of scholarships may have different income levels. For accuracy, we calculated the ratio by dividing the national scholarships received by each student's major. We then formulated this ratio in the same manner as the other numerical features. In the case of the course evaluation, we formulated the course evaluation score, such as "3.93", in the same way as the sentence "Course Evaluation : 3.93, negative" and since comments were originally in textual form, we concatenated them. As a result, we formulated the course evaluation as in the sentence "Course Evaluation : 3.93, negative, Too many assignments.". If there were no comments, we formulated it as "No course evaluation comments" and if there were no course evaluations at all, we formulated it as "No course evaluations".

We then tokenized the paragraphs using the WordPiece tokenizer [37]. In that process, we did not remove stop words, in line with existing research [27]. Finally, we obtained the token sequences as follows:

$$S = \{[CLS], \hat{d}_1, \hat{d}_2, \ldots, \hat{d}_k, \hat{a}_1, \hat{a}_2, \ldots, \hat{a}_l\}. \quad (1)$$

Note that $[CLS]$ is a classification token, $\hat{d}_1, \hat{d}_2, \ldots, \hat{d}_k$ denote the tokens of the demographic token sequence and $\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_l$ denote the tokens of the academic token sequence. k and l denote the lengths of the demographic and academic token sequences, respectively. The sentence formats corresponding to each variable are listed in Table 3.

3.2.2. Recasting the Task into NLI

Following the results of previous studies [15–17], we leveraged the knowledge of our PLM by recasting the prediction of dropout probability as an NLI task. Traditionally, NLI determines whether the given "premise" and "hypothesis" logically correspond (entailment), contradict (contradiction), or are undecided (neutral). Because the student dropout prediction task is a binary classification task (drop out or not), we recast it into a binary NLI task, comprising only "entailment" and "contradiction" labels. In general, the NLI task mainly uses a hypothesis consisting of discrete words, as follows:

$$S = \{[CLS], \hat{s}_1, \hat{s}_2, \ldots, \hat{s}_m, [SEP], \hat{d}_1, \hat{d}_2 \ldots, \hat{d}_k, \hat{a}_1, \hat{a}_2, \ldots, \hat{a}_l\}, \quad (2)$$

where $[SEP]$ is a separate token, $\hat{s} \in V$ denote tokens of the discrete hypothesis, V is the vocabulary of the PLM, and m denotes the length of the hypothesis. However, as shown in a previous study, using a discrete hypothesis such as "This student will drop out" may not

be optimal. Therefore, we propose a novel approach to enhance the NLI task by attaching continuous vector representations to the hypothesis to predict whether a student will drop out. We used the pseudotoken $\hat{p}$ as the hypothesis, which was not included in the vocabulary of existing language models, as follows:

$$S = \{[CLS], \hat{p}_1, \hat{p}_2, \ldots, \hat{p}_m, [SEP], \hat{d}_1, \hat{d}_2, \ldots, \hat{d}_k, \hat{a}_1, \hat{a}_2, \ldots, \hat{a}_l\}. \quad (3)$$

Then, $\hat{p}$ was embedded through a trainable embedder consisting of a single multilayer perceptron (MLP). Training the trainable embedder with the downstream task made it possible to find a better continuous hypothesis than the discrete hypothesis that the PLM can represent. Additionally, inspired by [38], we proceeded by attaching a continuous hypothesis to the key and value of each layer to make the hypothesis have a deeper impact on the output.

3.2.3. Student Dropout Prediction

Figure 2 illustrates the architecture of the proposed model. We utilized BERT$_{base}$ [27] and SimCSE-BERT$_{base}$ [31], popular transformer-based PLMs, to extract demographic and academic representations. We input a given token sequence S into the model, which produced a sequence of contextualized token representations $\tilde{S}$ as follows:

$$\tilde{S} = \{[CLS], \tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_m, [SEP], \tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_k, \tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_l\}, \quad (4)$$

where $\tilde{S}$ is a sequence of contextualized token representations, $\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_m$ denote the representation of the continuous hypothesis and $\tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_k, \tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_l$ denote the representation of the premise. To predict the probability of a student dropout, we used the contextual representation of the $[CLS]$ token as input to the two-layer MLPs. The output of the network was passed through a sigmoid function, producing a probability score between 0 and 1, where higher values indicated a higher likelihood of dropping out. The network's parameters were trained to minimize the binary cross-entropy loss, which was calculated using the predicted probability and true label for each training instance. The binary cross-entropy loss formula is as follows:

$$L = -\frac{1}{N}\{\sum_{i=1}^{N} y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)\}, \quad (5)$$

where N is the number of training instances, y_i is the true label, and $\hat{y}_i$ is the predicted probability for the ith instance.

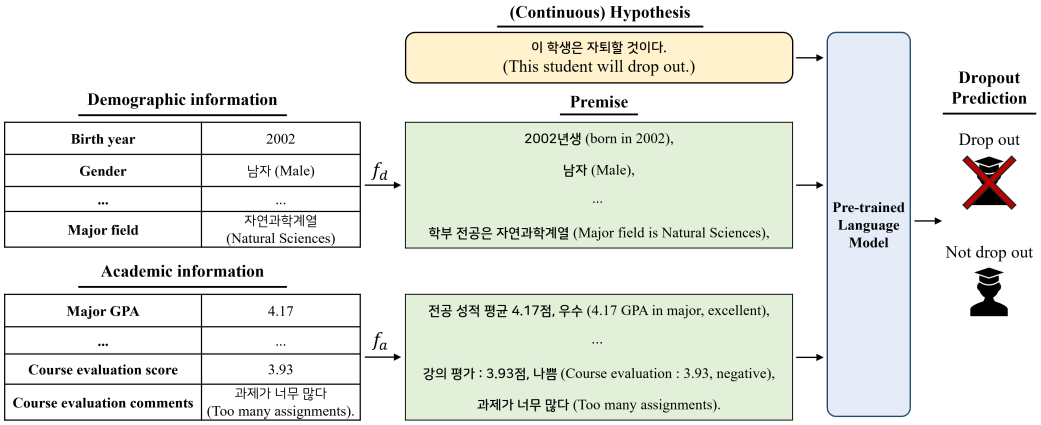

Figure 2. Illustration of the architecture for our proposed framework of STUD.

4. Experiments

4.1. Implementation Details

We used the BERT [27] and SimCSE-BERT [31] models to predict the student dropout. Both models are PLMs, and we trained them with a continuous hypothesis of length 32. We reparameterized the continuous hypothesis and used it in a reparameterization encoder comprising two MLPs with a hidden size of 512. For the hyperparameters, we set the learning rate to 5×10^{-5}, batch size to 32, and the maximum training epoch to 100. We warmed up the learning rate for the first 10% of the maximum number of training epochs and then used the AdamW optimizer [39], which employed a linear decay in the learning rate during the remainder of the training. We implemented STUD using PyTorch [40] and HuggingFace's transformers [41] libraries. Our training was conducted on a machine with three A100 GPUs with 80GB of VRAM and Intel Xeon Gold 6326.

4.2. Comparisons with Baseline Models

We evaluated the performance of seven methods. We first adopted multimodal spatiotemporal neural fusion (MSNF) [11] as a state-of-the-art method for the dropout prediction task. Other baselines included machine learning methods, such as logistic regression, support vector machine (SVM), decision tree, random forest, as well as deep learning methods, such as MLP. In our experiments, we compared the following methods:

- Logistic regression (baseline): This model is used to estimate the probability that dependent variables belong to a particular class. If the estimated probability exceeded a specific threshold (0.5), the model predicted the variables belonged to the dropout class. To counterbalance the lower proportion of student dropouts, we set the influence of the dropout class on the cost function to eight times that of the other class. The regularization parameter was set to 0.01 [5,6,22].
- SVM (baseline): This model is a binary linear classification model that determines which category the data belong to. In the feature space mapped to the dataset, the model finds the appropriate decision boundaries that divide the data into two classes. We set "rbf" as the kernel type and "scale" as a kernel coefficient. To address data imbalance, we set the influence of the dropout class on the cost function to six times that of the other class. The regularization parameter was set to one [5,8,9,22].
- Decision tree (baseline): This model shows the data patterns as predictable combinations of features. The model then proceeds with a binary classification depending on whether the features match. We used "gini" as a function to measure the quality oft he division [42] and adjusted the weight of the dropout class in the gini calculation to be six times higher than that of the other class to address data imbalance. We set the maximum depth to 5 [4,5,22].
- Random Forest (baseline): This model is an ensemble of decision trees. Instead of creating a decision tree for the entire feature set, a random forest randomly selects some features to generate several decision trees and collects the classification results from them. We use "gini" as a function to measure the quality of the division and set the weight of the dropout class in the gini calculation to be eight times higher than that of the other class. We set the maximum depth to 10 [4,5,8,9,22].
- MLP (baseline): This model is a multilayer structure that combines multiple perceptrons. We employed an MLP with a supervised learning technique using a gradient descent algorithm for training [5,6,22].
- MSNF (baseline): This model is the state-of-the-art dropout prediction method. It is a multitask model that uses long short-term memory, 1D convolutional neural networks (CNN), and BERT models to predict future dropout, next semester dropout, type of dropout, duration of dropout, and cause of dropout. Because we only used data from the first semester of the first year, we did not use temporal information. Therefore, we reimplemented the model using a 1D CNN and a BERT structure to predict future dropout [11].

- STUD: This is our proposed dropout prediction framework. We first formulated student demographic and academic information in a natural language format. We then fine-tuned PLMs (e.g., BERT [27], SimCSE-BERT [31]) by recasting the problem as an NLI task, prepending a continuous hypothesis at each layer of the model.

4.3. Experimental Results

Table 4 presents the results of our experiments, comparing the performance of the six baseline models with that of our proposed STUD model. Overall, our experiments demonstrated that STUD, which prepends continuous hypotheses to the layers, significantly outperformed the baseline methods. In particular, the F1-score of $\text{STUD}_{SimCSE-BERT}$ was 9.00% higher than that of MSNF, an existing state-of-the-art model.

Table 4. Comparison between STUD and other baselines.

Model	Dropout Prediction			
	Accuracy	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.766	0.255	**0.675**	0.370
SVM	0.847	0.339	0.533	0.414
Decision tree	0.821	0.296	0.546	0.384
Random forest	0.829	0.312	0.558	0.400
MLP	0.896	0.483	0.364	0.415
MSNF	0.877	0.405	0.442	0.422
STUD_{BERT}	0.894	0.478	0.416	0.444
$\text{STUD}_{SimCSE-BERT}$	**0.901**	**0.516**	0.416	**0.460**

This result demonstrated that our model effectively captured the relationships between structured and unstructured data when predicting dropout. In addition, $\text{STUD}_{SimCSE-BERT}$ exhibited a 10.84% higher performance in terms of the F1-score than the MLP models using only conventional numerical forms. This result showed that our attempts to formulate numeric data in textual form and train them were effective. Moreover, we observed that $\text{STUD}_{SimCSE-BERT}$ outperformed STUD_{BERT} by 3.60% in terms of the F1-score. This showed that the model additionally trained on the NLI dataset was effective for the dropout prediction task, recast as an NLI task.

Another insight that we observed from these results was that the precision was higher than the recall for the proposed method. This indicated that the model was generally passive in predicting students who drop out. In real-world applications, it may be more beneficial for administrators to use models with a high precision. By implementing a counseling program for students who are considering dropping out, universities can proactively prevent dropouts and improve student retention rates. However, if such a program is applied to students who have no intention of dropping out, it can have a negative impact on their academic experience. Balancing the precision and recall of the model can be considered in further work.

5. Analysis

5.1. Ablation Study

We conducted additional experiments to compare the performance improvements of each variable in terms of student information. To determine the effectiveness of the demographic (Dem.) and academic (Aca.) information, we removed these variables separately in distinct experiments. Moreover, we conducted a separate experiment for course evaluation comments (Cou.), which are unstructured data, to compare the performance improvement when using structured and unstructured data from academic information in detail. The results of these experiments are shown in Table 5, along with the STUD model results for comparison. When demographic information was removed, STUD_{BERT} exhibited a 54.95% performance drop in terms of the F1-score, whereas $\text{STUD}_{SimCSE-BERT}$ showed a 46.09% performance drop. When the structured academic information was removed, STUD_{BERT}

demonstrated a 42.57% performance drop based on the F1-score, and $\text{STUD}_{SimCSE-BERT}$ exhibited a 33.26% performance drop. When the unstructured course evaluation comments were removed, STUD_{BERT} showed an 8.11% performance drop based on the F1-score, and $\text{STUD}_{SimCSE-BERT}$ displayed a 9.78% performance drop.

Table 5. Ablation analysis of the STUD model.

Model	Modality			Dropout Prediction			
	Dem.	Aca.	Cou.	Accuracy	Precision	Recall	F1-Score
STUD$_{BERT}$	✗	✓	✓	0.884	0.333	0.143	0.200
	✓	✗	✓	0.892	0.424	0.182	0.255
	✓	✓	✗	0.885	0.429	0.390	0.408
	✓	✓	✓	**0.894**	**0.478**	**0.416**	**0.444**
STUD$_{SimCSE-BERT}$	✗	✓	✓	0.759	0.182	0.390	0.248
	✓	✗	✓	0.874	0.350	0.273	0.307
	✓	✓	✗	0.896	0.483	0.364	0.415
	✓	✓	✓	**0.901**	**0.516**	**0.416**	**0.460**

First, we observed the most significant drop in performance when demographic information was removed. These results indicated that demographic information was a much more influential variable than academic information for predicting student dropout. We assumed that this was because the students' academic information included only one semester, which was insufficient in terms of absolute information compared to demographic information. Second, we observed that both structured data (e.g., GPA and volunteer hours) and unstructured data (e.g., course evaluation comments) within academic information impacted a dropout prediction. This result demonstrated that course evaluation comments could not be utilized through machine learning and deep learning methods utilizing only structured data, which also influenced dropout prediction. This suggests that a combination of both structured and unstructured data is required for a more precise prediction of student dropout.

5.2. Effect of Different Types for Hypothesis

We conducted additional experiments to understand the effect of NLI task recasting on the dropout prediction task and the impact of the hypothesis on performance. We compared the model before the NLI recasting (*Classification*) with that after the NLI recasting (*NLI*). Because a previous study [43] showed that discrete tokens may be suboptimal, we ran experiments with discrete and continuous hypotheses. The results of these experiments are shown in Table 6. First, the model with the discrete hypothesis ($NLI_{Discrete}$) outperformed *Classification* by up to 5.38% in terms of the F1-score. We also observed that the performance was lower or higher depending on the discrete hypothesis. This suggested that the optimal discrete hypothesis we considered may not have been the best choice. Therefore, we conducted additional experiments using a continuous hypothesis instead of a discrete hypothesis. We found that $NLI_{Continuous}$ outperformed $NLI_{Discrete}$ by 3.32% in terms of the F1-score. Based on this result, our proposed model prepended the continuous hypothesis to the key and value of each layer to better reflect the prediction value. Our proposed model (*Ours*) outperformed $NLI_{Continuous}$ by 13.58%. These results indicated that the continuous hypothesis was more suitable than the discrete hypothesis when recasting the task to the NLI task, and attaching them to each layer was more helpful for model prediction than attaching them to the input embeddings alone.

Table 6. The comparison of dropout prediction performance for various settings.

Model	Mode	Dropout Prediction			
		Accuracy	Precision	Recall	F1-Score
$STUD_{SimCSE-BERT}$	*Classification*	0.880	0.397	0.351	0.372
	$NLI_{Discrete}$	0.881	0.409	0.377	0.392
	$NLI_{Continuous}$	0.872	0.384	**0.429**	0.405
	Ours	**0.901**	**0.516**	0.416	**0.460**

5.3. Qualitative Analysis

Figure 3 shows a visualization and comparison of the academic information between the two groups (dropout and nondropout) after admission. We found that the GPA of dropouts was 23.03% lower for majors and 24.78% lower for nonmajors compared to nondropouts, which suggests that dropouts had a lower academic motivation and achievement. We also found that dropouts participated in 84.05% fewer student clubs and 204.29% fewer volunteer hours than nondropouts, which suggested that dropouts were less motivated to participate in academic activities. Lastly, we found that dropout students' course evaluation scores were 22.38% lower than nondropout students, indicating a negative attitude towards the courses they had taken at university. As a result, we can see that dropout students were less enthusiastic about their activities at university and had a more negative attitude towards their courses than nondropout students.

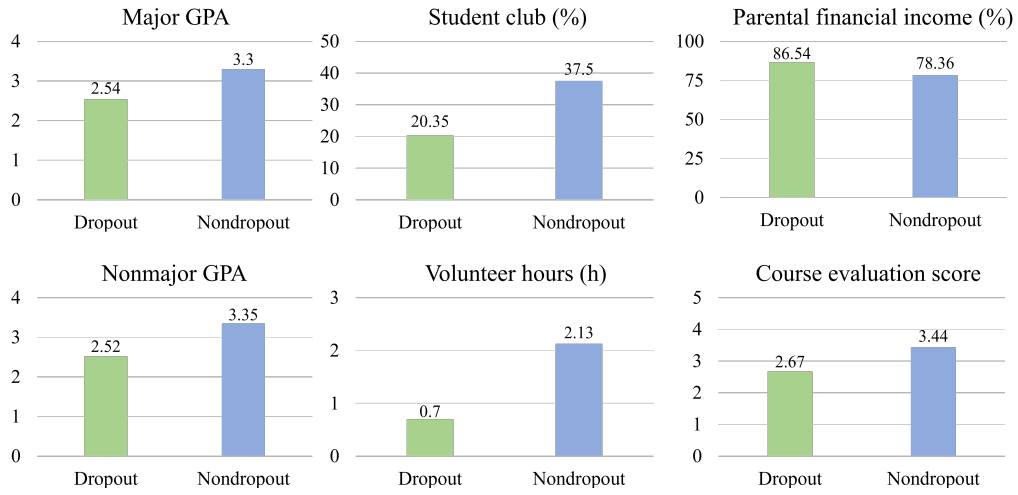

Figure 3. The difference in both numerical and categorical features between students who dropped out and those who did not drop out.

6. Conclusions

In this study, we introduced a novel framework for predicting university student dropout by combining individual demographic and academic information. We formulated each feature into a natural language format and utilized a PLM to capture the relationships between different types of data. We further recast the student dropout prediction task from a binary classification task to an NLI task to maximize the effectiveness of the language model. Experimental results demonstrated that the proposed framework STUD significantly outperformed several baseline methods. Notably, our model achieved a remarkable improvement of 9.00% on the F1-Score compared to the current state-of-the-art model. We showed that demographic and academic information was meaningful for predicting student dropout and that the recasting task was successful. However, STUD required many tokens (on average, 123 tokens) for formulating structured data into unstructured data and recasting the task as an NLI task. In future work, we aim to minimize the number of tokens for formulating and recasting the task to accommodate a wider variety of information (e.g., counseling records).

Author Contributions: Conceptualization, K.-M.K.; methodology, H.-S.W. and K.-M.K.; software, H.-S.W. and K.-M.K.; validation, H.-S.W.; formal analysis, K.-M.K. and M.-J.K.; investigation, H.-S.W. and K.-M.K.; resources, K.-M.K.; data curation, M.-J.K.; writing—original draft preparation, D.K. and H.-S.K.; writing—review and editing, H.-S.W., M.-J.K. and K.-M.K.; visualization, H.-S.W.; supervision, K.-M.K.; project administration, K.-M.K.; funding acquisition, K.-M.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Basic Research Program through the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (no. 2022R1C1C1010317)) and the Research Fund, 2021 of The Catholic University of Korea.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable. The code is available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Irwin , V.; De La Rosa, J.; Wang, K.; Hein, S.; Zhang, J.; Burr, R.; Roberts, A.; Barmer, A.; Bullock Mann, F.; Parker, S.; et al. Report on the Condition of Education 2022 (NCES 2022-144). National Center for Educ. Stat., Washington, DC, USA, NCES 2022144. 2022. Available online: https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2022144 (accessed on 8 May 2023).
2. Bound, J.; Lovenheim, M.F.; Turner, S. Why have college completion rates declined? An analysis of changing student preparation and collegiate resources. *Am. Econ. J. Appl. Econ.* **2010**, *2*, 129–157. [CrossRef] [PubMed]
3. Bowen, W.G.; Chingos, M.M.; McPherson, M.S. Crossing the finish line: Completing college at America's public universities. *Trusteeship* **2009**, *17*, 24–29.
4. Cannistrà, M.; Masci, C.; Ieva, F.; Agasisti, T.; Paganoni, A.M. Early-predicting dropout of university students: An application of innovative multilevel machine learning and statistical techniques. *Stud. High. Educ.* **2021**, *47*, 1935–1956. [CrossRef]
5. Opazo, D.; Moreno, S.; Álvarez-Miranda, E.; Pereira, J. Analysis of First-Year University Student Dropout through Machine Learning Models: A Comparison between Universities. *Mathematics* **2021**, *9*, 2599. [CrossRef]
6. Sandoval-Palis, I.; Naranjo, D.; Vidal, J.; Gilar-Corbi, R. Early dropout prediction model: A case study of university leveling course students. *Sustainability* **2020**, *12*, 9314. [CrossRef]
7. Silva, J.; Matos, L.F.A.; Mosquera, C.M.; Mercado, C.V.; González, R.B.; Llinás, N.O.; Lezama, O.B.P. Prediction of academic dropout in university students using data mining: Engineering case. *Lect. Notes Electr. Eng.* **2020**, *643*, 495–500.
8. Del Bonifro, F.; Gabbrielli, M.; Lisanti, G.; Zingaro, S.P. Student dropout prediction. In Proceedings of the International Conference on Artificial Intelligence in Education (AIED), Ifrane, Morocco, 6–10 July 2020; pp. 129–140.
9. Rodríguez-Muñiz, L.J.; Bernardo, A.B.; Esteban, M.; Díaz, I. Dropout and transfer paths: What are the risky profiles when analyzing university persistence with machine learning techniques? *PLoS ONE* **2019**, *14*, 2019. [CrossRef] [PubMed]
10. Aulck, L.; Velagapudi, N.; Blumenstock, J.; West, J. Predicting student dropout in higher education. *arXiv* **2016**, arXiv:1606.06364.
11. Alam, M.A.U. College student retention risk analysis from educational database using multi-task multi-modal neural fusion. In Proceedings of the AAAI Conference on Artificial Intelligence, Arlington, VA, USA, 22 February–1 March 2022; Volume 36, pp. 12689–12697.
12. Chen, L.C. An Improved Corpus-Based NLP Method for Facilitating Keyword Extraction: An Example of the COVID-19 Vaccine Hesitancy Corpus. *Sustainability* **2023**, 15, 3402. [CrossRef]
13. Yin, P.; Neubig, G.; Yih, W.-t.; Riedel, S. TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), Seattle, WA, USA, 5–10 July 2020; pp. 8413–8426.
14. Jun, C.; Choi, J.; Sim, M.; Kim, H.; Jang, H.; Min, K. Korean-Specific Dataset for Table Question Answering. In Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC), Marseille, France, 20–25 June 2022; pp. 6114–6120.
15. Neeraja, J.; Gupta, V.; Srikumar, V. Incorporating external knowledge to enhance tabular reasoning. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Mexico City, Mexico, 6–11 June 2021; pp. 2799–2809.
16. Mishra, A.; Patel, D.; Vijayakumar, A.; Li, X.L.; Kapanipathi, P.; Talamadupula, K. Looking Beyond Sentence-Level Natural Language Inference for Question Answering and Text Summarization. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Mexico City, Mexico, 6–11 June 2021; pp. 1322–1336.
17. Trivedi, H.; Kwon, H.; Khot, T.; Sabharwal, A.; Balasubramanian, N. Repurposing Entailment for Multi-Hop Question Answering Tasks. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN, USA, 2–7 June 2019; pp. 2948–2958.

18. Wook, M.; Yusof, Z.M.; Nazri, M.Z.A. Educational data mining acceptance among undergraduate students. *Educ. Informat. Technol.* **2016**, 22, 1195–1216. [CrossRef]
19. Dass, S.; Gary, K.; Cunningham, J. Predicting Student Dropout in Self-Paced MOOC Course Using Random Forest Model. *Information* **2021**, *12*, 476. [CrossRef]
20. Zheng, Y.; Gao, Z.; Wang, Y.; Fu, Q. MOOC Dropout Prediction Using FWTS-CNN Model Based on Fused Feature Weighting and Time Series. *IEEE Access* **2020**, *8*, 225324–225335. [CrossRef]
21. Feng, W.; Tang, J.; Liu, T.X. Understanding Dropouts in MOOCs. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Honolulu, HI, USA, 27 January–1 February 2019; pp. 517–524.
22. Kabathova, J.; Drlik, M. Towards predicting student's dropout in university courses using different machine learning techniques. *Appl. Sci.* **2021**, 11, 3130. [CrossRef]
23. Kotsiantis, S.B.; Pierrakeas, C.J.; Pintelas, P.E. Preventing student dropout in distance learning using machine learning techniques. In Proceedings of the Knowledge-Based Intelligent Information and Engineering Systems, 7th International Conference (KES), Oxford, UK, 3–5 September 2003; pp. 267–274.
24. Pellagatti, M.; Masci, C.; Ieva, F.; Paganoni, A.M. Generalized mixed-effects random forest: A flexible approach to predict university student dropout. *Stat. Anal. Data Min. ASA Data Sci. J.* **2021**, *14*, 241–257. [CrossRef]
25. Jayaraman, J.D. Predicting Student Dropout by Mining Advisor Notes. In Proceedings of the 13th International Conference on Educational Data Mining (EDM), Ifrain, Morocco, 10–13 July 2020; pp. 629–632.
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
27. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. 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 (NAACL-HLT), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
28. He, P.; Liu, X.; Gao, J.; Chen, W. Deberta: Decoding-enhanced bert with disentangled attention. In Proceedings of the 9th International Conference on Learning Representations (ICLR), Virtual Event, Austria, 3–7 May 2021.
29. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In Proceedings of the 8th International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26–30 April 2020. [CrossRef]
30. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
31. Gao, T.; Yao, X.; Chen, D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), Punta Cana, Dominican Republic, 7–11 November 2021; pp. 6894–6910.
32. Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; Liu, J. Uniter: Universal image-text representation learning. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 104–120.
33. Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 121–137.
34. Lu, J.; Batra, D.; Parikh, D.; Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; pp. 13–23.
35. Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; Dai, J. Vl-bert: Pre-training of generic visual-linguistic representations. In Proceedings of the 8th International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26–30 April 2020.
36. Hur, K.; Lee, J.; Oh, J.; Price, W.; Kim, Y.; Choi, E. Unifying Heterogeneous Electronic Health Records Systems via Text-Based Code Embedding. In Proceedings of the Conference on Health, Inference, and Learning (PMLR), Virtual, 7–8 April 2022; pp. 183–203.
37. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv* **2016**, arXiv:1609.08144.
38. Liu, X.; Ji, K.; Fu, Y.; Tam, W.; Du, Z.; Yang, Z.; Tang, J. P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), Dublin, Ireland, 22–27 May 2022; pp. 61–68.
39. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the 7th International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
40. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Chintala, S.; et al. Pytorch: An imperative style, high-performance deep learning library. In Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; pp. 8024–8035.
41. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Rush, A.M.; et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP Demos), Virtual, 16–20 November 2020; pp. 38–45.

42. Chen, C.; Liaw, A.; Breiman, L. *Using Random Forest to Learn Imbalanced Data*; University of California Berkeley: Berkeley, CA, USA, 2004; Volume 110, p. 24.
43. Liu, X.; Zheng, Y.; Du, Z.; Ding, M.; Qian, Y.; Yang, Z.; Tang, J. GPT understands, too. *arXiv* **2021**, arXiv:2103.10385.