

## Deep Learning-Based Classification of Special Education Logs Using DistilKoBERT

Won-Cheol Park\*

\*Ph. D, Dept. of Computer Engineering, Kongju National University, Cheonan, Korea

### [Abstract]

This study proposes a DistilKoBERT-based model that automatically classifies disability types by reflecting the learning characteristics of special education students. The dataset consists of 3,425 sentences extracted from special education support logs and is categorized into five disability types: Intellectual Disability, Emotional and Behavioral Disorders, Communication Disorders, Autism Spectrum Disorder, and Physical Disability. Text mining and natural language processing techniques were applied for data preprocessing. Additionally, the performance of the DistilKoBERT model was compared with TF-IDF + SVM and XGBoost. The evaluation results demonstrated that the DistilKoBERT model achieved the highest accuracy (0.89) and F1-score (0.87). This confirms that the proposed automatic classification system outperforms existing methods. This study highlights the potential of automated diagnostic systems in the field of special education. Future research aims to enhance the system's practicality by expanding the dataset and optimizing the model.

▶ **Key words:** DistilKoBERT, Special Education Support Log, Automatic Classification, Natural Language Process (NLP), Text Mining, Disability Types

### [요 약]

본 연구는 특수교육 대상 학생의 학습 특성을 반영하여 장애 유형을 자동으로 분류하는 DistilKoBERT 모델을 구축하였다. 데이터셋은 특수교육 지원 일지에서 추출한 3,425개 문장으로 구성되었으며, 5가지 장애 유형(지적 장애, 정서 행동 장애, 의사소통 장애, 자폐 스펙트럼 장애, 지체 장애)으로 분류된다. 본 연구에서는 텍스트 마이닝과 자연어 처리 기법을 적용하여 데이터 전처리를 수행하였다. 또한, DistilKoBERT 모델의 성능을 TF-IDF + SVM, XGBoost와 비교 분석하였다. 성능 평가 결과, DistilKoBERT 모델이 가장 높은 정확도(Accuracy 0.89)와 F1 점수(0.87)를 기록하였다. 이를 통해 제안된 자동 분류 시스템이 기존 방법보다 뛰어난 성능을 보임을 검증하였다. 본 연구는 특수교육 분야에서 자동화된 진단 시스템의 가능성을 제시한다. 향후 연구에서는 데이터 확장 및 모델 최적화를 통해 시스템의 실용성을 더욱 향상시키고자 한다.

▶ **주제어:** DistilKoBERT, 특수교육 지원 일지, 자동 분류, 자연어 처리(NLP), 텍스트 마이닝, 장애 유형

• First Author: Won-Cheol Park, Corresponding Author: Won-Cheol Park  
\*Won-Cheol Park (pwcfrog@kongju.ac.kr), Dept. of Computer Engineering, Kongju National University  
• Received: 2025. 04. 15, Revised: 2025. 05. 15, Accepted: 2025. 06. 04.

## I. Introduction

교육의 효과는 교사의 역량과 헌신에 따라 달라지며, 이는 특수교육에서도 예외가 아니다. 특수학교에서는 학생 개개인의 요구를 반영하여 교사가 1:1 맞춤형 교육을 제공하지만, 효과적인 교육을 위해서는 교사가 학생의 학습 특성과 개별적인 교육 데이터를 충분히 이해하고 활용할 수 있어야 한다. 그러나 실제 교육 현장에서는 다양한 장애 유형과 학습 특성을 고려한 체계적인 교육과정과 환경이 충분히 갖추어져 있지 않아, 개별 맞춤형 교육이 원활하게 이루어지기 어려운 실정이다[1]. 또한 특수교사 양성기관에서 교육실습과 교육과정들에 관한 많은 연구들이 이루어지고 있으나 학생들에게 적합한 맞춤형 교육을 진행하는 데 있어 아직 불완전한 상태이다[2]. 특수교육에서 학생의 학습 특성과 장애 유형을 정확하게 진단하는 것은 맞춤형 교육 제공의 핵심적인 요소이다. 특히, 지적 장애(Intellectual Disability), 정서 행동 장애(Emotion Behavior), 의사소통 장애(Communication), 자폐 스펙트럼 장애(Autism), 지체 장애(Physical Disability) 등 다양한 장애 유형에 따라 적절한 교육적 접근 방식이 달라질 수 있으므로, 신속하고 정밀한 진단 시스템이 필수적이다[3]. 기존의 특수교육 진단 과정은 주로 전문가의 관찰과 수동적 분석에 의존해 왔다[4]. 그러나 이러한 방법은 많은 시간과 비용이 소요될 뿐만 아니라, 평가자의 주관적인 해석이 개입될 가능성이 높아 진단의 일관성과 신뢰성을 확보하는 데 한계가 있다. 이에 따라, 최근에는 텍스트 마이닝 및 딥러닝 기반 자연어 처리(NLP) 기법을 활용하여 자동화된 진단 시스템을 개발하는 연구가 활발히 진행되고 있으며, 이는 특수교육 분야에서 더 객관적이고 효율적인 진단을 가능하게 하는 새로운 해결책으로 주목받고 있다[5-7].

본 연구의 목적은 DistilKoBERT 모델을 기반으로 한 자동 진단 시스템을 구축하여, 특수교육 지원 일지를 분석하고 장애 유형을 자동으로 분류하는 것이다. 이를 통해 특수교육 대상 학생들의 학습 특성을 더욱 체계적으로 파악하고, 개별 맞춤형 교육 지원을 위한 객관적인 진단 시스템을 제공하고자 한다. 또한, 본 연구는 DistilKoBERT 기반의 장애 유형 분류 시스템을 구축하고, 기존 머신러닝 모델(TF-IDF + SVM, XGBoost)과 비교하여 그 성능을 검증하는 것을 목표로 한다. 이러한 연구 결과는 특수교육 분야에서 딥러닝 기반 자연어 처리 기법을 적용하는 새로운 가능성을 제시하며, 더욱 정밀한 맞춤형 교육 지원 시스템 개발을 위한 기초 자료로 활용될 수 있을 것이다.

본 논문은 다음과 같은 구성으로 이루어진다. 2장에서는 텍스트 마이닝 및 자연어 처리 기반의 특수교육 데이터 분석과 관련된 선행연구를 고찰하고, 3장에서는 제안 시스템의 구조 및 DistilKoBERT 모델 구현과 학습 과정을 설명한다. 4장에서는 실험 설계 및 성능 평가 결과를 제시하고, 5장에서는 연구의 시사점 및 활용 가능성을 논의하며, 6장에서 결론과 향후 연구 방향을 제시한다.

## II. Preliminaries

### 1. Related works

#### 1.1 Text Mining and Special Education Data

텍스트 마이닝은 비정형 텍스트 데이터를 분석하여 유의미한 패턴과 정보를 추출하는 기법으로, 교육 데이터 분석뿐만 아니라 의료, 법률, 감성 분석 등 다양한 분야에서 활용되고 있다[8-10]. 특히, 교육 분야에서의 텍스트 마이닝 연구는 학생들의 학업 성취도 분석, 교육 콘텐츠 추천, 감성 분석 등을 중심으로 이루어져 왔다[11-13]. 그러나 특수교육 데이터 분석에 텍스트 마이닝 기법을 적용한 연구는 상대적으로 부족하며, 기존 연구는 주로 정형 데이터(예: 평가 점수, 출석률 등)를 활용한 분석에 집중되어 있다. 특수교육 지원 일지는 개별 학생의 학습 및 행동 패턴을 기록한 문서로, 학생별 맞춤형 지원을 제공하기 위한 중요한 자료이다[14]. 하지만 이러한 데이터는 대부분 비정형 텍스트 형식으로 존재하며, 학습자의 학습 및 행동 특성을 자동으로 분석하고 분류하는 것이 어렵다는 한계를 가진다. 이에 따라, 특수교육 지원을 위한 텍스트 마이닝 기반 연구가 요구되고 있으며, 최근 자연어처리(NLP, Natural Language Processing) 기법을 활용한 연구가 증가하는 추세이다[15][16]. 최근 연구들은 딥러닝과 NLP 모델을 활용하여 특수교육 데이터 분석의 정확도를 높이는 방안을 탐색하고 있다. 특히, 감성 분석(Sentiment Analysis)을 통해 특수교육 대상자의 감정 변화를 분석하거나, 토픽 모델링(Topic Modeling) 기법을 활용하여 주요 관심사를 도출하는 연구가 진행되고 있다[17][18]. 또한, 머신러닝 기반 분류 모델을 적용하여 장애 유형별 특성을 자동으로 분류하는 연구도 이루어지고 있으며, 이러한 접근 방식은 교육자의 업무 부담을 줄이고, 학생 개개인의 필요에 맞춘 맞춤형 지원을 가능하게 한다는 점에서 높은 실용성을 가진다.

## 1.2 DistilKoBERT and NLP-based Models

자연어처리(NLP) 기술이 발전함에 따라, 딥러닝 기반의 분류 모델들이 다양한 분야에서 활용되고 있다. 특히, Transformer 모델의 등장 이후 NLP 연구는 급격한 발전을 이루었으며, BERT(Bidirectional Encoder Representations from Transformers)는 자연어 이해 및 문맥 학습 능력에서 뛰어난 성능을 보이며 대표적인 NLP 모델로 자리 잡았다. 하지만 BERT는 높은 연산 비용과 방대한 학습 데이터 요구로 인해 실제 현장에서 활용하는 데 한계가 있다[19]. 이를 보완하기 위해 등장한 모델이 DistilBERT이며, 본 연구에서 활용하는 DistilKoBERT는 DistilBERT의 한국어 버전이다. 최근에는 다양한 자연어 처리 기반 모델이 특수교육 데이터를 분석하기 위해 활용되고 있으며, 특히 Transformer 기반의 모델이 문맥 이해와 정밀 분류에 강점을 보이고 있다. 다음의 Table 1은 전통적 방식부터 최신 경량화 모델까지 주요 텍스트 마이닝 기법의 특성과 활용 분야를 비교한 것이다.

Table 1. Comparison of Text Mining Techniques

<b>[1] TF-IDF ( Traditional Approach )</b>
Frequency-based, lacks contextual understanding
Applications : Document Classification, Information retrieval
<b>[2] Word2Vec ( Embedding Techniques )</b>
Captures semantic similarity between words
Applications : Sentiment analysis, recommendation systems
<b>[3] BERT ( Pre-trained Language Model )</b>
Bidirectional contextual learning with high accuracy
Applications : Question answering systems, text summarization
<b>[4] DistilKoBERT ( Lightweight Model )</b>
~40% smaller and 60% faster than BERT, retains ~97% performance
Applications : Small-scale text classification, real-time analysis

이러한 비교를 통해 확인할 수 있듯, 기존의 TF-IDF, Word2Vec 등의 기법은 구조적 한계나 문맥적 정보의 부족으로 인해 특수교육 지원 일지와 같은 비정형 텍스트 데이터의 분류에는 다소 제약이 있다. 이에 반해 BERT 및 DistilKoBERT와 같은 사전 학습 기반의 모델은 문맥 이해도가 높고, 특히 DistilKoBERT는 경량화 구조를 바탕으로 실시간 처리와 소규모 데이터에도 적합한 성능을 제공한다. DistilBERT는 BERT의 구조를 단순화하여 모델의 크기를 40% 축소하고, 학습 속도를 60% 향상시키면서도 원

본 BERT 대비 97%의 성능을 유지하는 특징을 갖고 있다 [20]. 이러한 경량화된 구조는 상대적으로 적은 데이터에서도 우수한 성능을 발휘할 수 있으며, 특수교육 지원과 같은 실시간 분석이 필요한 환경에서 더욱 적합한 모델로 평가된다. 특수교육 데이터 분석에서 기존에 사용되었던 머신러닝 기반 분류 모델과 비교했을 때, DistilKoBERT는 문맥을 보다 효과적으로 반영할 수 있다는 강점을 가진다 [21]. DistilKoBERT는 Transformer 기반의 문맥 학습 구조를 활용하여 단어의 순서 및 문맥적 의미까지 반영할 수 있어, 특수교육 지원 일지와 같은 비정형 데이터 분석에 더욱 적합하다. 최근 연구들에서도 Transformer 기반의 모델이 머신러닝 기반의 모델보다 텍스트 분류에서 우수한 성능을 보이며, 특히 다중 클래스 분류(Multi-class Classification)에서 높은 정확도를 기록하고 있다[22].

본 연구에서는 DistilKoBERT를 기반으로 특수교육 지원 일지 데이터를 분석하여 장애 유형을 자동으로 분류하는 시스템을 구축하고, 기존 머신러닝 모델과 성능을 비교하여 최적의 접근 방안을 도출하고자 한다. 기존 연구에서는 감성 분석, 질의응답 시스템, 문서 요약, 질병 분류 등 다양한 분야에서 자연어 처리(NLP) 모델이 활용되어 왔으나, 특수교육 지원 데이터처럼 도메인 특화된 비정형 데이터를 대상에 대한 딥러닝 기반 연구는 여전히 부족한 실정이다. 특히, 다중 장애 유형을 정밀하게 분류하는 연구는 제한적으로 이루어져 왔다. 이러한 한계를 보완하기 위해 본 연구에서는 DistilKoBERT 모델을 적용한 특수교육 데이터 자동 분류 시스템을 제안하고, 기존 머신러닝 모델과 비교 분석하여 가장 효과적인 분류 방법을 도출하는 것을 목표로 한다. 이를 통해 특수교육 데이터 분석의 실용성을 높이고, 개별 학습자의 특성을 반영한 맞춤형 교육 지원을 가능하게 하여 현장 적용성 및 실용성 측면에서의 새로운 연구 방향을 제시하고자 한다.

## III. The Proposed System

### 1. System Architecture and Process

본 연구에서 제안하는 시스템은 학생별 특수교육 지원 일지데이터를 기반으로 하여 장애 유형을 자동으로 분류하는 DistilKoBERT 모델을 중심으로 설계되었다. 전체 시스템은 데이터 수집, 데이터 전처리, 모델 학습, 성능 평가, 결과 분석 및 적용 가능성 평가의 5단계로 구성되며, 각 단계는 상호 연계되어 효율적인 분류 작업을 수행한다. 다음의 Figure 1은 제안 시스템의 전체적인 구성도를 보

여주며, 시스템은 입력 데이터 수집부터 결과 출력 및 활용까지의 전체 흐름을 포괄한다.

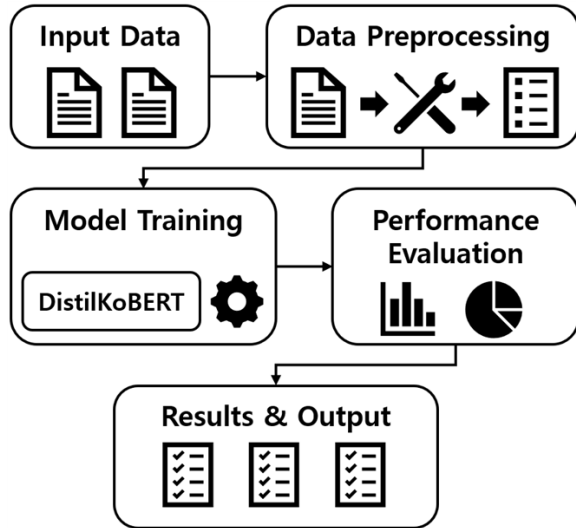


Fig. 1. Architecture of the proposed system

첫 번째 단계는 입력 데이터 단계로, 특수교육 대상 학생들의 특수교육 지원 일지 데이터를 수집한다. 이 데이터는 지적 장애, 정서 행동 장애, 의사소통 장애, 자폐 스펙트럼 장애, 지체 장애의 다섯 가지 주요 장애 유형별로 구성된다.

두 번째 단계는 데이터 전처리 과정으로, 모델 학습의 정확도를 향상시키기 위해 텍스트 정제, 토큰화, 라벨 인코딩 등의 전처리 과정이 수행되었다. 텍스트 정제(Text Cleaning) 단계에서는 불용어(stopwords) 및 특수문자 제거를 통해 모델이 의미 있는 단어를 집중할 수 있도록 하였다. 토큰화(Tokenization) 단계에서는 Hugging Face의 DistilBERT TokenizerFast를 사용하여 문장을 숫자형 토큰 시퀀스로 변환하였으며, 128 토큰의 최대 시퀀스 길이를 설정하여 메모리 사용량 절감과 정보 손실 최소화를 동시에 달성하였다. 이후, 라벨 인코딩(Label Encoding)을 수행하여 각 장애 유형을 정수형 레이블로 변환하였다. 데이터셋은 학습 데이터(80%), 검증 데이터(10%), 테스트 데이터(10%)로 분할되었으며, 이를 통해 학습 과정의 객관성을 확보하고 과적합(overfitting)을 방지하였다.

세 번째 단계는 모델 학습 단계이다. DistilKoBERT는 BERT 모델의 경량화 버전으로, 40% 더 작은 크기와 60% 더 빠른 학습 속도를 제공하면서도 높은 성능을 유지할 수 있다. 본 연구에서는 Epoch 5회, Batch Size 16, Learning Rate  $3e-5$ , Optimizer AdamW, Early Stopping Patience 값 2 등의 하이퍼파라미터(Hyperpara-

eters)를 적용하여 학습을 진행하였다. 또한, Cross Entropy Loss를 적용하고, 에포크별 검증 수행 및 조기 종료(Early Stopping) 전략을 통해 학습 과정에서 과적합 방지와 성능 최적화하기 위한 작업을 수행하였다.

네 번째 단계인 성능 평가 단계에서는 모델의 성능을 검증하기 위해 정확도(Accuracy), F1 점수(F1 Score), 정밀도(Precision), 재현율(Recall)과 같은 다양한 지표를 활용하였다. 특히, 혼동 행렬(Confusion Matrix) 분석을 수행하여 각 장애 유형별 분류 정확도와 오분류 비율을 분석하였다. 이를 통해 모델의 분류 성능을 세부적으로 평가하고, 장애 유형 간의 분류 패턴을 파악하였다.

마지막 단계는 결과 분석 및 적용 가능성 평가 단계로, 최종적으로 분류된 장애 유형을 출력한다. 제안된 DistilKoBERT 기반 시스템이 실제 교육 현장에서 어떻게 활용될 수 있는지에 대한 적용 가능성을 평가하고, 실제 특수교육 환경에 적용 가능한 실용적 모델로서의 가치를 입증할 수 있는 자료가 될 수 있다.

## 2. Dataset Construction and Preprocessing

### 2.1 Dataset Construction

본 연구에서는 특수교육대상자의 학생별 특수교육 지원 일지로부터 데이터셋을 구축하고, 이를 제안 시스템의 학습 및 성능 평가에 활용하였다. 데이터는 특수교육에서 주요하게 다루어지는 다섯 가지 장애 유형(지적 장애, 정서 행동 장애, 의사소통 장애, 자폐 스펙트럼 장애, 지체 장애)으로 분류되었으며, 총 3,425개의 실제 진단 문장으로 구성되어 있다. 데이터셋은 각 장애 유형별로 고르게 분포되도록 구성하여, 특정 유형에 대한 학습 편향을 방지하고 모델 평가의 신뢰성을 확보하였다. 수집된 문장들은 교육 현장에서 발생하는 다양한 사례를 반영하고 있으며, 모델이 각 장애 유형의 언어적 특성을 정밀하게 학습할 수 있도록 설계되었다. 장애 유형별로 살펴보면, 지적 장애는 지적 발달 지연과 학습 곤란을 주요 특징으로 하며, 인지 지연과 낮은 이해력이 나타나는 850개의 문장이 포함되었다. 정서 행동 장애는 정서적 불안정과 행동 문제를 중심으로 구성되며, 불안, 공격성, 사회적 문제 등의 특성을 담은 704개의 문장으로 이루어져 있다. 의사소통 장애는 언어적 및 비언어적 의사소통에 어려움을 겪는 아동을 대상으로 하며, 말하기 어려움과 이해력 부족이 포함된 700개의 문장으로 구성되었다. 자폐 스펙트럼 장애는 사회적 상호작용의 결함과 반복적 행동을 주요 특징으로 하며, 해당 특성이 반영된 625개의 문장이 수록되어 있다. 마지막으로, 지체 장애는 신체적 이동 능력과 운동 기능의 어려움

을 중심으로 하며, 이동성 문제 및 협응 능력 부족이 드러난 546개의 문장이 포함되었다. 이와 같은 장애 유형별 데이터 분류는 특수교육 대상자의 특성을 명확히 이해하고, 향후 유형별 맞춤형 지원 전략 수립에 중요한 기반을 제공한다. 다음의 Table 2은 각 카테고리별 주요 특성을 요약한 것이다.

Table 2. Dataset Statistics and Key Characteristics

Category	Details
<b>Intellectual_Disability</b> Sentences : 850	Intellectual development delays, learning difficulties
Key Characteristics : Cognitive delays, low comprehension	
<b>Emotion_Behavior</b> Sentences : 704	Emotional instability, behavioral issues
Key Characteristics : Anxiety, aggression, social issues	
<b>Communication</b> Sentences : 700	Verbal & non-verbal communication impairments
Key Characteristics : Speech difficulties, comprehension issues	
<b>Autism</b> Sentences : 625	Autism spectrum disorder characteristics
Key Characteristics : Social deficits, repetitive behaviors	
<b>Physical_Disability</b> Sentences : 546	Physical mobility & motor function challenges
Key Characteristics : Mobility issues, coordination deficits	

## 2.2 Data Preprocessing

제안 시스템의 학습 성능과 효율성을 극대화하기 위해 체계적인 데이터 전처리 과정을 수행하였다. 전처리 과정은 텍스트 정제(Text Cleaning), 토큰화(Tokenization), 라벨 인코딩(Label Encoding), 데이터 분할(Data Split)의 네 단계로 구성된다.

첫 번째 단계인 텍스트 정제에서는 모델 학습의 정확성을 높이기 위해 불필요한 텍스트 요소를 제거하였다. 문맥 이해에 기여하지 않는 불용어(stopwords)를 제거하여 모델이 의미 있는 핵심 단어에 집중할 수 있도록 하였으며, 학습 과정에서 혼란을 유발할 수 있는 특수문자 및 불필요한 기호를 제거하여 데이터의 일관성을 확보하였다.

다음으로 수행된 토큰화 과정에서는 텍스트 데이터를 모델이 이해할 수 있는 숫자 형태의 토큰 시퀀스로 변환하였다. 본 연구에서는 Hugging Face의 DistilBert TokenizerFast를 사용하여 문장 단위 토큰화를 수행하였다. 해당 토큰라이저는 DistilKoBERT 모델과의 높은 호환성과 빠른 처리 속도를 제공하여 대규모 텍스트 데이터 처리에 적합하다. 또한, 최대 시퀀스 길이를 128로 설정하여 메모리

사용량을 줄이는 동시에 정보 손실 없이 효율적인 학습이 가능하도록 하였다. 데이터 분석 결과, 본 연구의 데이터셋 문장은 대부분 128 토큰 이내에 포함되어 추가적인 정보 손실 없이 모델 학습이 이루어졌다.

라벨 인코딩 과정에서는 각 장애 유형을 모델 학습에 적합한 정수형 레이블로 변환하여 분류 작업을 위한 목표 변수(Target Variable)를 생성하였다. 구체적으로, 지적 장애는 0, 정서 행동 장애는 1, 의사소통 장애는 2, 자폐 스펙트럼 장애는 3, 지체 장애는 4로 인코딩하였다. 이러한 정수형 레이블 변환을 통해 머신러닝 모델이 각 장애 유형을 목표 변수로 인식할 수 있도록 하였으며, Cross-Entropy Loss와 같은 분류 손실 함수를 적용할 수 있도록 하였다.

마지막으로, 데이터 분할을 통해 모델의 성능 평가의 객관성 확보와 과적합 방지를 달성하였다. 전체 데이터셋을 학습 데이터(80%), 검증 데이터(10%), 테스트 데이터(10%)로 분할 하였으며, 학습 데이터는 모델이 패턴과 특성을 학습하는 데 사용되었다. 검증 데이터는 하이퍼파라미터 튜닝과 모델의 일반화 성능을 평가하는 데 사용되었으며, 테스트 데이터는 최종적으로 학습된 모델의 성능을 객관적으로 평가하는 데 활용되었다. 이와 같은 일련의 전처리 과정은 모델 학습의 효율성과 정확성을 높이는 데 기여하였다. 각 단계는 모델의 일반화 성능과 실제 데이터 환경에서의 적용 가능성을 극대화하는 데 중점을 두었다. 불필요한 요소 제거, 최적의 시퀀스 길이 설정, 정확한 레이블 인코딩, 체계적인 데이터 분할을 통해 DistilKoBERT 모델의 성능을 극대화할 수 있는 학습 환경을 구축하였다. 결과적으로, 본 연구의 데이터 전처리 과정은 모델이 각 장애 유형의 고유한 언어적 특성을 효과적으로 학습할 수 있도록 지원하였으며, 실제 교육 환경에서도 적용 가능한 실용적인 모델 개발의 기반을 마련하였다.

## 3. Model Training with DistilKoBERT

### 3.1 Overview of DistilKoBERT Model

본 연구에서는 특수교육 대상자의 특수교육 지원 일지 데이터를 자동으로 분류하기 위한 모델로 DistilKoBERT를 채택하였다. DistilKoBERT는 BERT 모델의 경량화 버전으로, 상대적으로 적은 양의 데이터로도 빠른 학습과 높은 분류 성능을 유지할 수 있어 본 연구의 데이터 환경에 적합하다. 특히, 본 연구에서 사용한 데이터는 한국어 기반의 비정형 텍스트로 구성되어 있으며, 다국어(multilingual) 처리를 지원하는 'distilbert base multilingual cased' 모델은 이러한 데이터의 분석에 유

리하다. DistilKoBERT는 BERT의 기본 구조인 Transformer Encoder를 기반으로 하지만, 지식 증류(distillation)를 통해 모델의 크기와 연산량을 줄이면서도 성능 저하를 최소화하였다. 이러한 특성은 학습 시간 단축과 자원 절감의 장점을 제공하며, 실무 환경에서의 적용 가능성을 높인다. 본 연구에서는 이러한 모델 특성을 바탕으로 특수교육 지원 일지 데이터에 대한 분류 모델을 구현하고 학습하였다.

### 3.2 DistilKoBERT Model Implementation and Training

본 연구에서는 DistilKoBERT 모델의 효율적인 학습을 위해 사전 실험과 기존 연구 결과를 바탕으로 도출된 하이퍼파라미터 설정을 다음의 Table 3과 같이 적용하였다. 이러한 설정은 모델의 안정적인 수렴과 일반화 성능 향상에 중점을 두고 설계되었다.

Table 3. DistilKoBERT Model Hyperparameter Setup

Hyperparameter	Value
Epochs	5
Batch Size	16
Learning Rate	3e-5
Optimizer	AdamW
Weight Decay	0.01
Evaluation Strategy	epoch
Early Stopping	Patience-2
Max Sequence Length	128

학습 반복 횟수(Epoch)는 과적합을 방지하면서 수렴을 유도할 수 있는 수준인 5회로 설정하였다. 배치 크기(Batch Size)는 메모리 자원과 학습 안정성의 균형을 고려해 16으로 설정하여 학습 안정성과 메모리 효율성을 동시에 확보하였다. 학습률(Learning Rate)은 3e-5로 설정하여 과적합(Overfitting)을 방지하고, 점진적인 성능 향상이 이루어지도록 하였으며, 이는 BERT 계열 모델 학습 시 안정적인 수렴을 유도하는 대표적인 값이다. 최적화 함수로는 일반적으로 텍스트 분류에서 우수한 성능을 보이는 AdamW를 사용하였고, 과도한 가중치 업데이트를 방지하기 위해 weight decay 값은 0.01로 설정하여 모델의 일반화 성능을 높였다. 학습 과정에서는 에포크 단위 평가 전략(Evaluation Strategy)을 적용하여 학습 도중 성능을 실시간으로 모니터링하였으며, Early Stopping 기법을 통해 과적합을 방지하였다. Patience 값은 2로 설정하여 성능 향상이 2회 연속 발생하지 않을 경우 학습을 종료하도록 하였다. 이는 소규모 데이터셋에서의 과적합 가능성을 최소화하고 학습 시간을 절약하기 위한 설정이다. 문장 길

이(Max Sequence Length)는 메모리 효율성과 정보 손실 방식을 동시에 고려하여 128 토큰으로 제한하여 연산 자원을 절약하고 학습 효율을 높였다. 입력 데이터는 Hugging Face의 DistilBertTokenizerFast를 활용해 문장 단위로 토큰화되었으며, 이 과정에서 [CLS], [SEP] 등의 특수 토큰이 자동으로 삽입되었다. 모든 입력 문장은 Padding 및 Truncation 처리를 통해 128 토큰으로 길이를 통일함으로써, 메모리 사용량을 줄이고 병렬 처리 효율성을 확보하였다.

학습 과정은 Hugging Face의 Trainer API를 활용하여 수행되었으며, 에포크마다 학습 손실(Loss)과 검증 정확도(Validation Accuracy)를 모니터링하였다. 분류 작업에는 Cross-Entropy Loss를 손실 함수로 사용하였고, 앞서 설정된 AdamW 옵티마이저와 함께 적용하여 학습 성능을 최적화하였다. 이와 같은 일련의 설정은 사전 실험 및 관련 선행연구들을 바탕으로 조정되었으며, 제한된 데이터 환경에서도 과적합을 방지하고, 학습 효율성과 모델의 일반화 성능을 확보할 수 있도록 설계되었다.

본 모델 구현 및 학습 전략은 실제 특수교육 데이터 분석 환경에서의 적용 가능성과 현실적인 조건을 반영한 실용적인 접근으로 설계되었으며, 이는 본 연구의 목적을 달성하는 데 있어 핵심 기술로 기능하였다.

## IV. Experimental Results and Analysis

### 1. Experimental Setup

본 연구에서는 제안된 DistilKoBERT 기반 특수교육대상자의 학생별 특수교육 지원 일지 데이터 분석 시스템의 성능을 검증하기 위해, 기존 모델(TF-IDF + SVM, XGBoost)과의 성능 비교 및 분석을 수행하였다. 본 실험의 목적은 제안된 DistilKoBERT 모델이 특수교육 지원 일지 데이터 분류 작업에서 기존 모델 대비 얼마나 우수한 성능을 발휘하는지 평가하는 것이다. 실험에 사용된 데이터셋은 총 3,425개의 특수교육 진단 보고서 문장으로 구성되었다. 각 문장은 지적 장애, 정서 행동 장애, 의사소통 장애, 자폐 스펙트럼 장애, 지체 장애의 다섯 가지 주요 장애 유형 중 하나로 분류되었다. 데이터셋의 구성은 각 장애 유형별로 균등하게 분포되어 있어, 학습 과정에서 특정 카테고리로의 편향을 최소화하고, 평가 과정에서의 공정성을 확보할 수 있도록 설계되었다. 제안 시스템의 학습과 검증을 위한 데이터 분할은 80%를 학습용(training set),

10%를 검증용(validation set), 10%를 테스트용(test set)으로 설정하였다. 이러한 데이터 분할 전략은 모델 학습 과정에서 과적합(overfitting)을 방지하고, 테스트 단계에서 모델의 일반화 성능을 객관적으로 평가할 수 있도록 하였다. 성능 평가는 다음의 네 가지 주요 지표를 사용하여 수행되었다. 먼저, 정확도(Accuracy)는 전체 데이터 중에서 올바르게 분류된 비율을 나타내며, 분류 모델의 전반적인 성능을 평가하는 지표로 사용된다. 이어서, F1 점수(F1-Score)는 정밀도(Precision)와 재현율(Recall)의 조화 평균으로, 특히 데이터셋에 불균형이 존재할 경우 성능 평가에서 중요한 역할을 한다. 정밀도(Precision)는 모델이 특정 장애 유형으로 예측한 사례 중에서 실제로 올바르게 예측된 비율을 측정하여 모델의 정확성을 평가한다. 반면, 재현율(Recall)은 실제 특정 장애 유형에 속하는 데이터 중에서 모델이 얼마나 정확하게 해당 유형으로 분류했는지를 나타내며, 모델의 민감도를 측정하는 지표로 활용된다. 모델 학습과 실험은 다음의 하드웨어 및 소프트웨어 환경에서 수행되었다. 프로그래밍 언어로는 Python 3.12(64bit)를 사용하였으며, 모델 구현과 학습을 위해 PyTorch와 Hugging Face의 Transformers 라이브러리를 활용하였다. 특히, Hugging Face의 Transformers 라이브러리를 사용함으로써 모델 구현과 학습 프로세스를 표준화하고, DistilKoBERT 모델의 빠른 학습 속도와 높은 성능을 실험적으로 검증할 수 있었다. 이러한 실험 설정을 통해, 본 연구는 제안된 시스템의 성능과 실용성을 기존 모델들과 비교 분석하여 특수교육 데이터 분석 분야에서의 적용 가능성을 평가하고자 한다.

## 2. Comparative Performance Analysis

본 연구에서는 제안 모델인 DistilKoBERT와 기존 모델(TF-IDF + SVM, TF-IDF + XGBoost)의 성능을 정확도, F1 점수, 정밀도, 재현율의 네 가지 주요 지표를 통해 비교 및 분석하였다. 다음의 Table 4는 각 모델의 성능을 요약한 것이다.

Table 4. Model Performance Comparison

Model	Distil KoBERT	TF-IDF + SVM	TF-IDF + XGBoost
Accuracy	0.89	0.84	0.77
F1 Score	0.87	0.83	0.74
Precision	0.88	0.85	0.76
Recall	0.89	0.76	0.77

DistilKoBERT 모델은 모든 성능 지표에서 가장 뛰어난 결과를 기록하였다. 정확도(0.89)와 F1 점수(0.87)는 물론 정밀도(0.88)와 재현율(0.89)에서도 높은 수치를 달성하였다. 이러한 결과는 DistilKoBERT가 텍스트의 의미적 맥락과 언어적 특징을 효과적으로 학습했음을 보여준다. 특히, F1 점수에서 정밀도와 재현율 간의 균형을 이루었다는 점은 모델의 분류 안정성과 일반화 성능을 입증하는 중요한 결과이다.

반면, TF-IDF + SVM 모델은 정확도 0.84, F1 점수 0.83으로 상대적으로 안정적인 성능을 보였지만, 언어적 맥락 이해에서는 DistilKoBERT에 미치지 못했다. TF-IDF는 단어 빈도 기반의 단순한 특징을 추출하므로, 문맥적 정보의 반영이 부족하여 텍스트의 의미적 관계를 학습하는 데 한계를 보였다. 또한, TF-IDF + XGBoost 모델은 정확도 0.77, F1 점수 0.74로 가장 낮은 성능을 기록하였다. 이는 XGBoost가 수치적 관계 분석에는 강점을 가지지만, 텍스트 데이터의 복잡한 언어적 특성을 효과적으로 처리하는 데는 한계가 있음을 보여준다. 결과적으로, TF-IDF 기반 모델들은 단순한 피처 기반 접근 방식의 한계로 인해 텍스트의 문맥적 특성을 효과적으로 학습하지 못하는 성능 저하를 보였다.

DistilKoBERT 모델의 성능 우수성은 특히 특수교육대상자의 학생별 특수교육 지원 일지와 같이 복잡한 언어적 특성을 지닌 데이터에서 두드러졌다. Transformer 기반의 언어 모델은 단어 간 관계와 문맥적 의미를 반영할 수 있어, 텍스트의 의미적 정확성을 높이며 카테고리별 분류 성능을 향상시킬 수 있었다. DistilKoBERT 모델은 정확도, F1 점수, 정밀도, 재현율 모든 지표에서 기존 모델들을 능가하는 성능을 보여주었으며, 특수교육 데이터 분석에 최적화된 모델임을 입증하였다.

## 3. Confusion Matrix Analysis

본 절에서는 제안 시스템(DistilKoBERT)의 혼동 행렬(confusion matrix)을 통해 카테고리별 분류 성능을 분석하고, 모델의 분류 정확도와 신뢰성을 평가한다. 혼동 행렬은 각 카테고리별로 정확히 분류된 데이터와 오분류된 데이터의 분포를 시각적으로 제공하여 모델의 강점과 약점을 식별하는 데 도움을 준다. 다음의 Figure 2는 혼동 행렬 분석의 결과를 요약한 것이다. 각 카테고리별 실제 클래스(Actual Class)와 모델이 예측한 클래스(Predicted Class)를 교차 비교하여 모델의 분류 성능을 세부적으로 평가하였다.

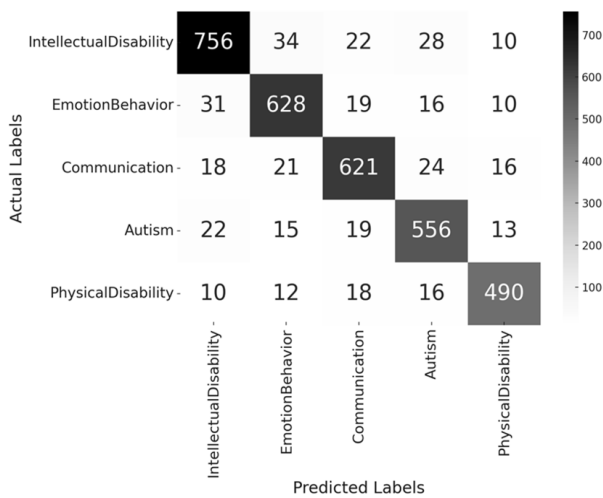


Fig. 2. Confusion Matrix for DistilKoBERT Classification

Figure 2의 대각선 상에 위치한 값들은 정확하게 예측된 샘플 수(True Positives)를 나타내며, 비대각선 값들은 잘못 분류된 샘플(False Positives 및 False Negatives)을 나타낸다. 각 행의 마지막 열인 Total (Actual Samples)은 해당 카테고리에 속하는 전체 샘플 수를, 마지막 행인 Total (Predicted Samples)은 각 카테고리가 예측된 총 샘플 수를 나타낸다. 이와 같은 구성은 모델의 분류 성능을 세부적으로 평가하고, 카테고리 간 오분류 경향 분석을 통해 모델 개선 방향을 탐색하는 데 중요한 정보를 제공한다. 혼동 행렬 분석 결과, DistilKoBERT 모델은 전체적으로 89%의 높은 분류 정확도를 기록하였다. Intellectual Disability(756/850), Emotion Behavior(628/704), Communication(621/700), Autism(556/625), Physical Disability(490/546)에서 모두 85% 이상의 높은 분류 정확도를 달성하였다. 특히 Communication과 Physical Disability 카테고리에서는 각각 88.7% 및 89.7%의 높은 정확도를 기록하여, DistilKoBERT 모델이 언어적 특성과 행동적 특성을 구분하는 능력이 뛰어남을 보여주었다. 이러한 결과는 모델이 장애 유형별 고유한 언어적 특성을 효과적으로 학습했음을 나타내며, 실제 교육 환경에서의 적용 가능성을 높인다. 그러나 일부 카테고리 간의 오분류도 발생하였다. Intellectual Disability와 Autism 간에는 각각 28건 및 22건의 오분류가 발생했는데, 이는 두 장애 유형 간의 언어적 표현 유사성으로 인한 것으로 해석된다. 예를 들어, "추상적 개념 이해의 어려움"과 같은 문장은 두 카테고리 모두에서 나타날 수 있는 특징으로, 모델이 해당 문장을 특정 카테고리로 분류하는 데 혼동을 겪었을 가능성이 있다. 또한 Emotion Behavior와 Communication 간에도 각각 19건 및 21건의 오분류가 관찰되었다. 이는

정서적 반응과 의사소통 방식에서 나타나는 언어적 중첩으로 인해 발생한 것으로 판단된다. Physical Disability의 경우에도 Autism 및 Communication과의 소규모 오분류가 발생하였으며, 이는 행동적 특성과 의사소통의 언어적 표현이 일부 겹치기 때문으로 해석된다. 이러한 분석 결과는 DistilKoBERT 모델이 대부분의 카테고리에서 높은 분류 정확도를 달성하였음을 보여주며, 오분류 사례는 주로 언어적 맥락이 유사한 카테고리 간에서 발생했음을 시사한다. 모델의 분류 성능을 향상시키기 위해서는 이러한 유사한 카테고리 간의 오분류를 줄이기 위한 추가적인 특성 엔지니어링이나 데이터 증강 기법의 적용이 필요할 것으로 보인다. 예를 들어, 문맥 기반 임베딩 기법을 활용하여 문장 내에서 의미적으로 중요한 단어와 문구를 강조하거나, 각 카테고리의 특성을 더욱 명확히 반영할 수 있는 추가적인 데이터 확보를 통해 모델의 분류 능력을 향상시킬 수 있다. 성능 평가 요약 결과, DistilKoBERT 모델은 정확도 0.89, F1 점수 0.87, 정밀도 0.88, 재현율 0.89의 우수한 성능을 보였다. 이러한 성능 지표는 모델이 분류 작업에서 정밀도와 재현율 간의 균형을 효과적으로 유지하고 있음을 의미한다. 특히 Communication과 Physical Disability 카테고리에서의 높은 분류 정확도는 모델의 강점을 보여주는 지표이다. 추가적으로 수행된 오분류 사례 분석을 통해 모델의 개선 방향을 도출할 수 있었으며, 이는 향후 실제 교육 현장에서의 실용적 적용 가능성을 높이는 데 기여할 것으로 기대된다. 결론적으로, 본 연구에서 수행한 혼동 행렬 분석과 성능 평가 결과는 DistilKoBERT 모델이 특수교육대상자의 학생별 특수교육 지원 일지 분류 작업에서 뛰어난 성능을 발휘할 수 있음을 보여준다. 향후 오분류 사례의 원인 분석과 데이터 보강, 고도화된 전처리 과정의 적용을 통해 모델의 분류 성능을 더욱 향상시킬 수 있으며, 이는 실제 교육 현장에서의 효과적인 장애 유형 분류 및 지원 전략 수립에 기여할 수 있을 것으로 판단된다.

## V. Conclusions

본 연구에서는 특수교육대상자의 학생별 특수교육 지원 일지 데이터를 활용하여 학생의 장애 유형을 자동으로 분류하는 시스템을 제안하였다. 제안된 시스템은 경량화된 딥러닝 모델인 DistilKoBERT를 기반으로 하여, 상대적으로 적은 양의 데이터에서도 높은 성능을 발휘할 수 있도록 설계되었다. 이를 통해 특수교육 현장에서 개별 학생의 특

성을 신속하고 정확하게 파악할 수 있는 자동 분류 시스템의 실용성을 입증하고자 하였다. 연구 과정에서는 총 3,425개의 특수교육 지원 일지 데이터를 수집하여 지적 장애, 정서 행동 장애, 의사소통 장애, 자폐 스펙트럼 장애, 지체 장애의 다섯 가지 유형으로 분류하였다. 데이터 전처리 과정에서는 텍스트 정제, 토큰화, 라벨 인코딩 및 데이터 분할을 통해 모델 학습의 효율성을 높였다. 학습 과정에서는 DistilKoBERT 모델을 사용하여 학습 반복 횟수, 학습률, 배치 크기 등 최적의 하이퍼파라미터를 적용하여 학습 효율성을 극대화하였다. 실험 결과, DistilKoBERT 모델은 정확도 0.89, F1 점수 0.87, 정밀도 0.88, 재현율 0.89의 성능을 달성하였다. 이는 비교 모델인 TF-IDF + SVM의 정확도(0.84)와 TF-IDF + XGBoost의 정확도(0.77)보다 우수한 성능을 보여, 본 연구에서 제안한 시스템의 효용성을 강하게 입증하였다. 또한, 혼동 행렬 분석을 통해 장애 유형별 분류 정확도와 오분류 사례를 심층적으로 분석하였으며, 모델의 일반화 성능과 분류 신뢰성을 검증하였다. 이를 통해 DistilKoBERT 기반 시스템이 실제 교육 현장에서 맞춤형 교육 지원 방안을 제공하는 실용적인 도구로 활용될 수 있음을 확인하였다. 그러나 본 연구에는 몇 가지 제한점이 존재한다. 첫째, 데이터셋이 특정 교육 기관에서 수집된 사례로 한정되어 있어 다양한 교육 현장에서의 일반화 가능성에 대한 추가적인 검증이 필요하다. 둘째, 분류된 각 장애 유형에 따른 개별화된 교육 지원 전략에 대한 자동 추천 기능이 포함되지 않아, 추후 해당 기능의 개발이 필요하다. 셋째, 실시간 분석 및 적용성을 위한 시스템 성능 최적화와 배포 전략에 대한 후속 연구가 필요하다. 향후 연구에서는 다양한 교육 현장에서 수집된 데이터셋을 통합하여 모델의 범용성을 향상시킬 예정이다. 또한, 학생 맞춤형 교육 전략 자동 추천 시스템을 구현하여 분류 결과에 기반한 교육적 피드백을 제공할 수 있도록 연구를 확장할 계획이다. 마지막으로, 실시간 분석 및 모바일 환경 적용을 통해 교사와 교육 관계자가 현장에서 즉각적으로 활용할 수 있는 시스템으로 발전시키는 것을 목표로 한다. 이를 기반으로 향후에는 본 시스템을 교사용 교육 지원 도구로 개발할 수 있다. 예를 들어, 교사는 특수교육 지원 일지를 시스템에 입력하기만 하면, 자동으로 해당 학생의 장애 유형을 분류하고, 유형별로 최적화된 교육 전략 또는 지원 방안을 추천받을 수 있다. 해당 시스템은 웹 기반 혹은 모바일 애플리케이션 형태로 구현되어, 교사들이 실시간으로 학급 상황에 맞는 피드백과 개입 전략을 확인할 수 있도록 지원할 수 있다. 또한, 누적된 분석 결과를 통해 학생의 장기적인 발달 추세를 시각화하여 학부모 상담 및 지원 계획 수립에도 활용될

수 있으며, 교사 간 협업이나 행정 보고에도 유용한 자료로 기능할 수 있다. 이러한 활용 시나리오는 단순한 분류 모델을 넘어, 특수교육 현장에서 실질적인 의사결정을 지원하는 통합 도구로의 발전 가능성을 보여준다.

본 연구에서 제안한 DistilKoBERT 기반 특수교육 데이터 분석 시스템은 특수교육 현장에서 학습자의 개별적 특성을 고려한 맞춤형 교육 지원을 효과적으로 제공할 수 있는 실용적인 도구임을 입증하였다. 높은 분류 정확도와 효율적인 학습 구조는 시스템의 실용성과 적용 가능성을 강화하였으며, 이를 통해 향후 특수교육 환경에서 보다 정교한 맞춤형 교육 전략 수립에 기여할 것으로 기대된다.

## REFERENCES

- [1] J.G. Kim, "Analysis of Research Trends Related to Individualized Education in Special Education Using Keyword Network Analysis," *SECI*, Vol. 17, No. 4, pp. 97-115, December 2024. DOI: 10.24005/seci.2024.17.4.97
- [2] E.J. Lee, S.H. Kim, S.Y. Seo, N. Gantumur, and S.H. Lee, "A Study on the Experiences and Perceptions of Special and General Education Teachers, and Administrators regarding the Co-Homeroom Teacher System in Inclusive Preschools," *Journal of Special Education Studies*, Vol. 41, No. 1, pp. 203-237, February 2025. DOI: 10.31863/JSE.2025.02.41.1.203
- [3] B. Lee, "Analysis on the Special Education Course in the Department of Secondary Special Education," *Korean Journal of Special Education*, Vol. 53, No. 2, pp. 159-178, June 2018. DOI: 10.15861/kjse.2018.53.2.159
- [4] D.H. Son, "Analysis of Topics on Inclusive Education for Children with Disabilities Using Text Mining Techniques," *The Journal of Humanities and Social Sciences* 21, Vol. 14, No. 3, pp. 1793-1808, June 2023. DOI: 10.22143/HSS21.14.3.124
- [5] M.J. Kwon, J.K. Park, S.H. Bae, D.B. Gu, K.Y. Kim, and H.S. Kang, "Analysis of the Current Status of Art Therapy Using Text Mining Techniques," *Journal of Practical Engineering Education*, Vol. 16, No. 6, pp. 833-842, December 2024. DOI: 10.14702/JPEE.2024.833
- [6] M.R. Jeong, "A Study on the Trends of Special Education-Related Media Coverage Using Text Mining in Korea: 2008-2022," *PERSPECTIVES IN KOREAN STUDIES*, Vol. 28, pp. 313-364, August 2024. DOI: 10.33835/cghs.28.9
- [7] H.S. Jung, "Analysis of Research Trends in the Basic Curriculum of Special Education Using Text Mining," *SECI*, Vol. 17, No. 3, pp. 24-48, September 2024. DOI: 10.24005/seci.2024.17.3.24
- [8] G.D. Park, "Text Mining to Characterise Silver Town Services: Focusing on Social Awareness," *Journal of the Korea Contents*

- Association, Vol. 25, No. 1, pp. 633-643, January 2025. DOI: 10.5392/JKCA.2025.25.01.633
- [9] Y.H. Kim, "Application of Text Mining for Legal Information System: Focusing on Defamation Precedent," *Journal of the Korean Society for Library and Information Science*, Vol. 54, No. 1, pp. 387-409, February 2020. DOI: 10.4275/KSLIS.2020.54.1.387
- [10] S.C. Kwon, J.E. Kim, and B.C. Jang, "Comparative Study of User Reactions in OTT Service Platforms Using Text Mining," *JICS*, Vol. 25, No. 3, pp. 43-54, June 2024. DOI: 10.7472/jksii.2024.25.3.43
- [11] G.J. Yoo, and K. Lee, "Seeking Directions for Parental Education Programs through Sentiment Analysis Based on Text Mining: Centered on Play," *EME*, Vol. 22, No. 4, pp. 201-222, December 2021. DOI: 10.22154/JCLE.22.4.9
- [12] G.H. Lee, "A Study on the Perception Analysis of Future Education through Big Data," *KALCI*, Vol. 22, No. 24, pp. 25-37, December 2022. DOI: 10.22251/jlcci.2022.22.24.25
- [13] Y.E. Kim, and H.I. Cho, "Analysis of Hope Related Research Trends of Adolescents Using Text Mining," *Korean Society for Holistic Convergence Education*, Vol. 27, No. 4, pp. 113-138, October 2023. DOI: 10.35184/kshe.2023.27.4.113
- [14] E.S. Koh, and J. Park, "Perceptions of Inclusive Classroom Teachers, Special Education Teachers, Parents, and Special Education Paraprofessionals of the Necessary Traits and Job Competencies Related to Inclusive Education Support that are Required of Special Education Paraprofessionals," *The Journal of Special Education: Theory and Practice*, Vol. 25, No. 3, pp. 105-129, September 2024. DOI: 10.19049/JSPED.2024.25.3.05
- [15] H.M. Lee, and Y.M. Kim, "Development and effectiveness of NLP group counseling program for PTSD symptom reduction and PTG enhancement in PTSD risk group elementary school students," *The Korean Journal of Elementary Counseling*, Vol. 18, No. 2, pp. 181-210, April 2019. DOI: 10.28972/kjec.2019.18.2.181
- [16] H.S. Jung, "Analysis of Research Trends in the Basic Curriculum of Special Education Using Text Mining," *Korean Special Education Association for Curriculum and Instruction*, Vol. 17, No. 3, pp. 24-48, September 2024. DOI: 10.24005/seci.2024.17.3.24
- [17] H.Y. Kwak, S.H. Kim, and B.I. Lee, "An analysis on research trends related to families of young children with special needs using text mining analysis: With a special reference to the Korean professional journal articles (2003 - 2021)," *Journal of Special Education*, Vol. 38, No. 3, pp. 119-139, August. 2022. DOI: 10.31863/JSE.2022.08.38.3.119
- [18] Y.Y. Park, S.Y. Kim, Y.M. Yoo, Y.I. Choi, and Y. Ju, "Roles of Occupational Therapists Working in Special Education Support Centers: A Mixed Method Design Using Text Mining and Qualitative Research," *Korean Journal of Occupational Therapy*, Vol. 32, No. 4, pp. 23-37, November 2024. DOI: 10.14519/kjot.2024.32.4.02
- [19] J.B. Kim, Y.G. Kim, and J.M. Park, "Collaborative Filtered Enhanced Recommendation System Using BERT," *The Journal of the Institute of Internet, Broadcasting and Communication*, Vol. 24, No. 5, pp. 61-67, October 2024. DOI: 10.7236/JIIBC.2024.24.5.61
- [20] S. Sun, Y. Cheng, Z. Gan, and J. Liu, "Patient Knowledge Distillation for BERT Model Compression," *arXiv preprint arXiv:1908.09355*, pp. 4323-4332, August 2019. DOI: 10.48550/arXiv.1908.09355
- [21] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, pp. 1-5, March 2020. DOI: 10.48550/arXiv.1910.01108
- [22] T.Y. Kim, S.U. Baek, M.H. Lim, B. Yun, D. Paek, K.E. Zoh, K. Youn, Y.K. Lee, Y. Kim, J.W. Kim, E. Choi, M.Y. Kang, Y.H. Cho, K.E. Lee, J. Sim, J. Oh, H. Park, J. Lee, J.U. Won, Y.M. Lee, and J.H. Yoon, "Occupation classification model based on DistilKoBERT: using the 5th and 6th Korean Working Condition Surveys," *Annals of Occupational and Environmental Medicine*, Vol. 36, No. e19, pp. 1-13, August 2024. DOI: 10.35371/aoem.2024.36.e19

## Authors



Won-Cheol Park received the B. S., M. S. and Ph. D. degrees in Computer Engineering from Kongju National University in 2011, 2013, and 2018, respectively. He is currently teaching in the Department of Computer

Engineering, Kongju National University. He is interested in cloud computing, big data, image processing and machine learning.