

## Development of an AutoML Web Platform for Text Classification Automation

Ha-Yoon Song<sup>†</sup> · Jeon-Seong Kang<sup>††</sup> · Beom-Joon Park<sup>††</sup> · Junyoung Kim<sup>†††</sup> ·  
Kwang-Woo Jeon<sup>†††</sup> · Junwon Yoon<sup>††</sup> · Hyun-Joon Chung<sup>††††</sup>

## ABSTRACT

The rapid advancement of artificial intelligence and machine learning technologies is driving innovation across various industries, with natural language processing offering substantial opportunities for the analysis and processing of text data. The development of effective text classification models requires several complex stages, including data exploration, preprocessing, feature extraction, model selection, hyperparameter optimization, and performance evaluation, all of which demand significant time and domain expertise. Automated machine learning (AutoML) aims to automate these processes, thus allowing practitioners without specialized knowledge to develop high-performance models efficiently. However, current AutoML frameworks are primarily designed for structured data, which presents challenges for unstructured text data, as manual intervention is often required for preprocessing and feature extraction. To address these limitations, this study proposes a web-based AutoML platform that automates text preprocessing, word embedding, model training, and evaluation. The proposed platform substantially enhances the efficiency of text classification workflows by enabling users to upload text data, automatically generate the optimal ML model, and visually present performance metrics. Experimental results across multiple text classification datasets indicate that the proposed platform achieves high levels of accuracy and precision, with particularly notable performance when utilizing a Stacked Ensemble approach. This study highlights the potential for non-experts to effectively analyze and leverage text data through automated text classification and outlines future directions to further enhance performance by integrating Large language models.

Keywords : Text Classification, Automated Machine Learning (AutoML), Web Platform, Natural Language Processing, H2O

## 텍스트 분류 자동화를 위한 AutoML 웹 플랫폼 개발

송 하 윤<sup>†</sup> · 강 전 성<sup>††</sup> · 박 범 준<sup>††</sup> · 김 준 영<sup>†††</sup> · 전 광 우<sup>†††</sup> · 윤 준 원<sup>††</sup> · 정 현 준<sup>††††</sup>

## 요 약

인공지능과 머신러닝 기술의 급격한 발전은 다양한 산업 분야에 혁신을 일으키고 있으며, 특히 자연어 처리(NLP) 기술은 텍스트 데이터 분석 및 처리에 새로운 가능성을 제공하고 있다. 텍스트 분류 모델을 효과적으로 개발하려면 데이터 탐색, 전처리, 특징 추출, 모델 선택, 하이퍼파라미터 튜닝, 성능 평가 등의 복잡한 단계를 거쳐야 하며, 이는 많은 시간과 전문 지식을 요구한다. 자동화된 머신러닝(AutoML)은 이러한 과정을 자동화하여 비전문가도 고성능 모델을 쉽게 생성할 수 있도록 돕는다. 그러나 기존 AutoML 도구는 주로 정형 데이터에 특화되어 있어, 비정형 텍스트 데이터 처리에는 전처리와 특징 추출 과정에서 수작업이 필요하다. 본 연구에서는 이러한 한계를 해결하기 위해 텍스트 전처리, 단어 임베딩, 모델 학습 및 평가 과정을 자동화하는 웹 기반 AutoML 플랫폼을 개발하였다. 이 플랫폼은 사용자가 텍스트 데이터를 업로드하면 최적의 머신러닝 모델을 자동으로 생성하고 성능을 시각적으로 제공함으로써 텍스트 분류 작업의 효율성을 크게 향상시킨다. 다양한 텍스트 분류 데이터셋을 활용한 실험 결과, 제안된 플랫폼은 높은 정확도와 정밀도를 보였으며, 특히 Stacked Ensemble 모델 사용 시 우수한 성능을 나타냈다. 본 연구는 텍스트 분류 자동화를 통해 비전문가도 손쉽게 텍스트 데이터를 분석하고 활용할 수 있는 가능성을 제시하며, 향후 대규모 언어 모델(LLM)을 적용하여 성능을 더욱 향상시킬 계획이다.

키워드 : 텍스트 분류, 자동화된 머신러닝, 웹 플랫폼, 자연어처리, H2O

※ 본 연구는 방위산업기술지원센터의 지원으로 수행되었음(UC200019D).

† 정 회 원 : 한국로봇융합연구원 인공지능로봇연구본부 주임연구원

†† 비 회 원 : 한국로봇융합연구원 인공지능로봇연구본부 주임연구원

††† 비 회 원 : 한국로봇융합연구원 인공지능로봇연구본부 선임연구원

†††† 정 회 원 : 한국로봇융합연구원 인공지능로봇연구본부 책임연구원

Manuscript Received : August 1, 2024

Accepted : August 26, 2024

\* Corresponding Author : Hyun-Joon Chung(hjchung@kiro.re.kr)

## 1. 서 론

최근 몇 년간 인공지능과 머신러닝(Machine Learning, ML) 기술의 급격한 발전은 다양한 산업에 혁신적인 변화를 가져왔다. 특히 자연어 처리(Natural Language Processing, NLP) 기술의 발전은 텍스트 데이터를 분석하고 처리하는 데 있어 새

로운 가능성을 열어주었다. 텍스트 분류는 NLP의 핵심 과제 중 하나로, 이메일 스팸 감지[1], 감정 분석[2], 문서 번역[3] 등 다양한 응용 분야에서 중요한 역할을 하고 있다. 그러나 효과적인 텍스트 분류 모델을 개발하는 과정은 데이터 탐색, 전처리, 특징 추출, 모델 선택, 하이퍼파라미터 튜닝, 성능 평가 등 여러 복잡한 단계를 포함하며, 이를 위해서는 전문 지식과 많은 시간이 필요하다[4].

자동화된 머신러닝(Automated Machine Learning, AutoML)은 이러한 복잡한 과정을 자동화하여 머신러닝 모델 개발의 효율성을 크게 향상시키는 기술로 주목받고 있다[5]. AutoML은 모델 선택, 하이퍼파라미터 최적화, 모델 평가 등의 과정을 자동으로 수행하여, 머신러닝에 대한 지식이 부족한 사용자도 유의미한 모델을 손쉽게 구축할 수 있도록 돕는다[6]. H2O AutoML은 H2O.ai에서 제공하는 강력한 AutoML 프레임워크로, 사용이 간편하고 R, Python, Java 등 다양한 프로그래밍 언어를 지원하여 데이터 과학자 및 엔지니어가 효율적으로 작업할 수 있도록 한다[7]. 그러나 H2O AutoML은 정형 데이터에 주로 특화되어 있어, 비정형 데이터인 텍스트 분류 작업을 직접적으로 지원하지 않는 한계가 있다. 이로 인해 사용자는 텍스트 데이터를 수치화하기 위해 단어 임베딩(Word Embedding)과 같은 기법을 활용한 특징 추출 과정을 수동으로 처리해야 한다[8].

본 연구에서는 이러한 한계를 극복하기 위해 텍스트 데이터의 전처리, 단어 임베딩을 통한 특징 추출, 모델 학습 및 평가 과정을 하나의 자동화된 파이프라인으로 통합한 H2O 기반 AutoML 웹 플랫폼을 개발하였다. Fig. 1은 텍스트 분류를 위한 기존의 AutoML 워크플로우와 본 연구에서 제안하는

AutoML 워크플로우의 구조적 차이를 나타낸다. 기존 AutoML 시스템은 텍스트 데이터를 처리할 때, 데이터 전처리와 특징 추출을 수동으로 수행해야 했다(Fig. 1a). 이러한 작업이 완료된 후에야 비로소 모델 학습, 하이퍼파라미터 조정, 성능 평가가 자동으로 진행된다. 그러나 본 연구에서 제안하는 AutoML 플랫폼은 텍스트 데이터의 전처리, 특징 추출, 모델 학습, 그리고 성능 평가까지 전 과정을 자동화하는 파이프라인을 제공한다(Fig. 1b). 이 플랫폼은 텍스트 데이터의 특성에 적합한 전처리 및 특징 추출 기법을 자동으로 적용하여, 연구자와 개발자가 텍스트 분류 모델을 보다 효율적으로 개발하고 최적화할 수 있도록 지원한다.

본 논문은 다음과 같은 구조로 구성된다. 2장에서는 관련 연구를 검토하고, 3장에서는 제안하는 AutoML 웹 플랫폼의 설계 및 구현 방식을 설명한다. 4장에서는 실험 환경과 결과를 제시하며, 5장에서는 결론 및 향후 연구 방향을 논의한다.

## 2. 관련 연구

### 2.1 텍스트 분류

텍스트 분류는 자연어 처리(NLP)에서 핵심적인 연구 분야로, 텍스트 데이터를 특정 카테고리 또는 클래스로 자동으로 분류하는 작업을 의미한다[9]. 텍스트 분류는 주로 데이터 수집 및 전처리, 특징 추출, 모델 선택 및 학습의 세 단계로 구성된다[10]. 먼저, 데이터 수집 및 전처리 단계에서는 텍스트 데이터를 다양한 출처에서 수집한 후, 불필요한 요소를 제거하고 필요한 부분을 정제한다. 이 과정에서는 텍스트 정규화, 불용어 제거, 형태소 분석 등의 기법이 사용되며, 이러한 전처리 과정을 통해 모델 성능을 향상시킬 수 있는 정제된 데이터를 도출한다. 다음으로, 특징 추출 단계에서는 텍스트 데이터를 수치화하여 머신러닝 모델에 입력할 수 있도록 변환하는 작업을 수행한다. 주로 사용되는 기법으로는 단어 임베딩(Word Embedding)이 있으며, 대표적인 방법으로는 Word2Vec[11], GloVe[12], 그리고 BERT[13] 등이 있다. 이들 임베딩 기법은 텍스트의 의미를 보존한 채 수치화하여 머신러닝 모델이 텍스트의 의미를 효과적으로 이해할 수 있도록 돕는다. 마지막으로, 모델 선택 및 학습 단계에서는 분류 작업에 적합한 머신러닝 알고리즘을 선택한 후 학습을 통해 최적의 분류 모델을 생성한다. 이 과정에서 자주 사용되는 알고리즘으로는 Naive Bayes, SVM(Support Vector Machine), 의사결정 나무(Decision Tree), 심층 신경망(Deep Neural Network) 등이 있다[14].

최근 대규모 언어 모델(Large Language Models, LLM)의 발전은 텍스트 분류 작업의 패러다임을 크게 변화시켰다. 특히, BERT[13], BART[15], GPT-3[16]와 같은 모델들은 다양한 텍스트 분류 작업에서 우수한 성능을 보여주고 있으며, 높은

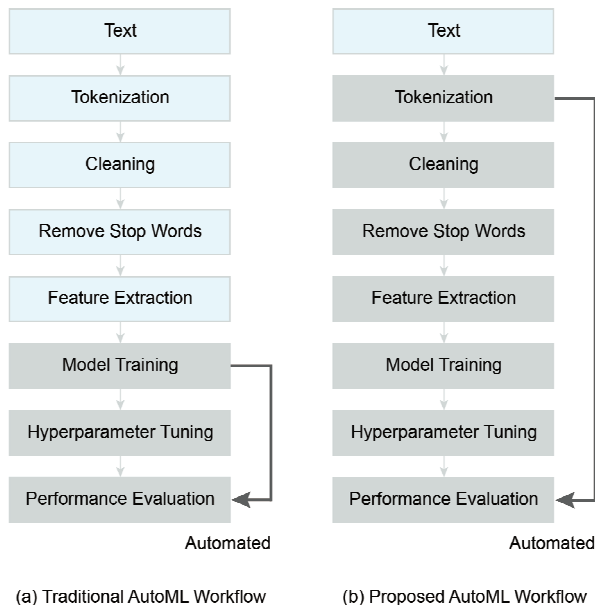


Fig. 1. Comparison of Traditional and Proposed AutoML Workflows for Text Classification

정확도를 달성하고 있다[17]. 이들 모델은 대규모 데이터셋에서 사전 학습(pre-training)을 거친 후, 특정 작업에 맞게 미세 조정(fine-tuning)되어 활용된다. 그러나 LLM 기반 방법에는 몇 가지 한계가 존재한다. 첫째, LLM 기반 모델은 전통적인 머신러닝 모델에 비해 훨씬 더 많은 계산 자원과 비용을 요구한다. LLM은 수십억 개 이상의 매개변수를 학습해야 하므로, 대규모 GPU 클러스터와 같은 고성능 컴퓨팅 자원이 필요하며, 이로 인해 운영 및 유지 비용이 크게 증가한다. 반면, Naive Bayes나 SVM과 같은 전통적인 머신러닝 모델은 상대적으로 적은 자원으로도 효율적인 학습이 가능하다. 둘째, 학습 시간에서도 큰 차이가 있다. 전통적인 머신러닝 모델은 학습 시간이 짧아 신속하게 모델을 구축할 수 있는 반면, LLM 기반 모델은 매우 긴 학습 시간이 소요된다. 이러한 제약은 실시간 응용 프로그램이나 자원이 제한된 환경에서 LLM 기반 모델의 활용을 어렵게 만드는 주요 요인으로 작용할 수 있다.

### 2.2 AutoML 시스템

AutoML은 머신러닝 모델 개발의 복잡한 과정을 자동화하여, 사용자가 별도의 전문 지식 없이도 고성능 모델을 생성할 수 있도록 돕는 기술이다[18]. AutoML 시스템은 데이터 전처리 자동화, 모델 선택, 하이퍼파라미터 최적화, 모델 평가 및 앙상블 등의 주요 구성 요소로 이루어져 있다. 데이터 전처리 자동화는 데이터 정제, 누락 값 처리, 특징 선택 및 생성 등의 작업을 자동으로 수행한다. 모델 선택 단계에서는 다양한 머신러닝 알고리즘 중에서 최적의 모델을 자동으로 선택한다. 하이퍼파라미터 최적화는 모델의 성능을 극대화하기 위해 하이퍼파라미터를 자동으로 조정하며, 모델 평가 및 앙상블 단계에서는 여러 모델의 성능을 평가하고, 필요에 따라 앙상블 기법을 사용하여 성능을 향상시킨다.

H2O.ai의 H2O AutoML은 이러한 AutoML 도구 중 하나로, 데이터 과학자와 엔지니어를 위한 오픈 소스 머신러닝 플랫폼을 제공한다. H2O AutoML은 사용자 친화적인 웹 기반 인터페이스를 제공하며, R과 Python 등 여러 프로그래밍 언어를 지원하여 사용자 편의성을 높였다. 또한, 회귀, 분류, 군집화, 딥러닝 등 다양한 머신러닝 알고리즘을 지원하며, 데이터 전처리, 모델 선택, 하이퍼파라미터 튜닝, 모델 평가 등의 과정을 자동으로 수행한다. 그러나 H2O AutoML은 텍스트 데이터를 직접 처리하는 기능이 제한적이므로, 텍스트 데이터를 효과적으로 처리하기 위해서는 추가적인 단어 임베딩 과정이 필요하다는 한계가 있다.

본 연구는 텍스트 분류 작업의 자동화를 위한 AutoML 웹 플랫폼 개발을 목표로 한다. 특히, H2O AutoML의 한계를 보완하여 텍스트 데이터에 특화된 자동화된 전처리 및 특징 추출 과정을 통합함으로써 사용자 편의성을 높이고, 작업 효율성을 극대화하고자 한다.



Fig. 2. AutoML Platform System Architecture

## 3. 시스템 설계 및 구현

### 3.1 시스템 개요

본 연구에서는 텍스트 분류 작업을 자동화하기 위해 H2O 기반의 AutoML 웹 플랫폼을 개발하였다. 이 플랫폼은 사용자가 텍스트 데이터를 업로드하면 최적의 머신러닝 모델을 자동으로 생성하여 텍스트 분류 작업을 수행할 수 있도록 설계되었다. 주요 구성 요소는 실행 환경 관리, 실행 알고리즘 관리, 학습 플랫폼, 사용자 인터페이스로 구성된다. Fig. 2는 플랫폼 시스템 구성도를 나타낸다.

### 3.2 실행 환경 관리

데이터 수집 및 처리는 안정적이고 확장 가능한 구조로 설계되었으며, Apache Kafka를 통해 실시간 데이터 스트리밍과 안정적인 데이터 수집 큐를 구현하였다. 대용량 데이터는 Hadoop Distributed File System(HDFS)를 통해 효율적으로 저장 및 관리하였으며, 분석을 위한 대규모 데이터 처리는 HBase와 같은 컬럼 베이스 저장소를 활용하여 성능을 극대화하였다. 모델 학습 및 예측 결과는 RESTful API를 기반으로 웹 서비스 형태로 제공하였으며, 자동화된 API 관리 기능을 통해 클라이언트와의 상호작용이 원활하고 안정적으로 이루어지도록 하였다. 특히, Docker를 사용하여 서비스를 컨테이너화하고, Kubernetes로 컨테이너의 배포와 관리를 자동화함으로써 플랫폼의 확장성과 유연성을 극대화하였다. 또한, 인프라 관리에는 Infrastructure as Code(IaC) 기법을 도입하여 하드웨어, 가상 리소스, 플랫폼, 컨테이너 시스템, 서비스 등을 코드로 관리하였다. 이를 통해 인프라의 프로비저닝 및 관리를 자동화하여 시스템의 일관성과 안정성을 보장할 수 있었다.

### 3.3 실행 알고리즘 관리

플랫폼은 파라미터, 시각화 설정, 소스 코드 등을 포함하여 다양한 알고리즘을 효율적으로 관리할 수 있다. 파라미터 관리 기능은 학습 및 검증 데이터 비율, 학습 반복 횟수, 랜덤 시드 고정 여부, 배치 사이즈 등 기본적인 실행 설정을 관리한다. 시각화 설정 관리 기능은 성능 지표를 나타내는 그래프 및 차트를 설정하는 데 사용된다. 마지막으로, 소스 코드 관리 기능은 알고리즘의 버전 관리, 재현성 보장, 코드 공유 등을 지원하며, 필요 시 업데이트 및 수정이 가능하다. 이를 통해 사용자는 다양한 알고리즘을 쉽게 적용하고 관리할 수 있으며, 일관된 결과를 도출할 수 있다.

### 3.4 학습 플랫폼

학습 플랫폼은 데이터 전처리, 모델 학습, 결과 분석의 세 가지 주요 단계로 구성된다.

#### 1) 데이터 전처리

비정형 텍스트는 구조화되어 있지 않고 일정한 규격이 없기 때문에, 분석에 적합한 형태로 데이터를 정제하는 과정이 필요하다. 본 연구의 데이터 전처리 단계에서는 텍스트 데이터의 토큰화(tokenization), 정제(cleaning), 정규화(normalization), 불용어 제거 작업을 수행한다. 먼저 텍스트 데이터 세트인 코퍼스(corpus)를 의미 있는 단위로 나누는 토큰화를 진행한다. 그 다음 정제 및 정규화 과정에서는 불필요한 기호, 숫자, 공백 등을 제거하고, 모든 텍스트를 소문자로 변환한다. 마지막으로 불용어 제거 과정에서는 분석에 불필요한 일반적인 단어들을 제거한다.

#### 2) 모델 학습

임베딩 모델 학습 단계에서는 처리된 텍스트 데이터를 벡터 형태로 변환하여 머신러닝 모델 학습에 사용할 수 있도록 한다. 이를 위해 Word2Vec과 같은 대표적인 방법을 사용한다. Word2Vec은 단어를 벡터 공간에 임베딩하여 단어 간의 유의미한 유사성을 반영할 수 있다[9]. 학습 방식에는 CBOW(Continuous Bag of Words)와 Skip-Gram이 있다[11]. CBOW는 문장의 문맥을 통해 찾고자 하는 단어를 예측하는 알고리즘이며, Skip-Gram은 특정 단어를 이용하여 그 주변에 나올 단어들을 예측하는 모델이다. Skip-Gram이 CBOW보다 성능이 뛰어나다고 알려져 있어[19], 본 플랫폼에서는 Skip-Gram 방식을 지원한다.

머신러닝 모델 학습 단계에서는 H2O AutoML을 사용하여 다양한 머신러닝 알고리즘을 적용한 모델들을 학습시킨다. H2O AutoML은 XGBoost[20], GBM(Gradient Boosting Machine)[21], GLM(Generalized Linear Model with regularization)[22], Neural Network[23], DRF(Distributed Random Forest)[24] 등의 알고리즘을 통해 여러 모델을 자동으로 생

성한다. 또한, 모델 학습 후에는 기존 모델을 사용하여 추가적인 Stacked Ensemble을 학습시킨다. Stacked Ensemble에는 'All Models'와 'Best of Family' 두 가지 유형이 있다. 'All Models'는 모든 기본 모델을 포함하며, 'Best of Family'는 각 알고리즘 패밀리에서 가장 성능이 좋은 모델만을 포함한다. 'Best of Family' 앙상블은 6개 이하의 기본 모델로 구성되므로, 예측 속도가 빠르고 복잡성이 낮아 대규모 앙상블에 비해 성능 저하가 크지 않으면서도 높은 효율성을 제공한다.

#### 3) 결과 분석

결과 분석 단계에서는 학습된 모델의 성능을 평가하고, 결과를 시각화하여 사용자에게 제공한다. H2O AutoML은 결과 분석의 일환으로 리더보드(leaderboard)와 교차 검증 결과의 세부 요약(cross-validation metrics summary)을 제공한다. 리더보드는 생성된 여러 모델의 성능을 종합적으로 비교할 수 있는 기능을 제공하며, 클래스별 평균 오류율(mean per class error), 로그 손실(log loss), RMSE(Root Mean Squared Error) 등의 지표를 기준으로 모델의 순위를 제공한다.

교차 검증 결과의 세부 요약은 각 모델의 성능을 더욱 상세하게 평가할 수 있도록 다양한 지표를 제공한다. 여기에는 각 폴드(fold) 및 폴드 평균에 따른 정확도(accuracy), 정밀도(precision), 재현율(recall), F1-score 등이 포함된다. 또한, 혼동 행렬(confusion matrix)을 통해 모델의 분류 결과를 시각화함으로써 사용자가 모델의 예측 성능을 직관적으로 이해할 수 있도록 돕는다. 혼동 행렬은 실제 클래스와 예측된 클래스 간의 관계를 시각적으로 나타내어, 어떤 클래스에서 주로 오류를 범하는지 쉽게 파악할 수 있게 한다.

이러한 결과 분석을 통해 사용자는 모델의 성능을 종합적으로 평가하고, 최적의 모델을 선택할 수 있다. 모델의 성능을 향상시키기 위해 필요할 경우 모델을 재학습하거나 하이퍼파라미터를 조정할 수 있다. 이를 통해 플랫폼 사용자는 텍스트 분류 모델을 구축하고, 이를 다양한 응용 분야에서 효과적으로 활용할 수 있다.

### 3.5 사용자 인터페이스

사용자 인터페이스는 사용자가 텍스트 분류 작업을 쉽고 효율적으로 수행할 수 있도록 도와주는 다양한 기능을 제공하며, 직관적이고 시각적인 정보를 통해 시스템의 상태를 쉽게 파악할 수 있게 설계되었다. Fig. 3은 본 연구에서 개발한 H2O 기반 AutoML 웹 플랫폼의 대시보드 화면을 나타낸다. 대시보드에서는 학습 플랫폼, AI 모델 저장소, 추론 플랫폼 등 주요 기능에 대한 접근이 용이하며, 데이터셋과 모델 현황, 알고리즘의 상태를 한눈에 확인할 수 있다. 사용자는 각종 지표와 모델 상태를 시각적으로 확인하여 작업 진행 상황을 쉽게 파악할 수 있다.

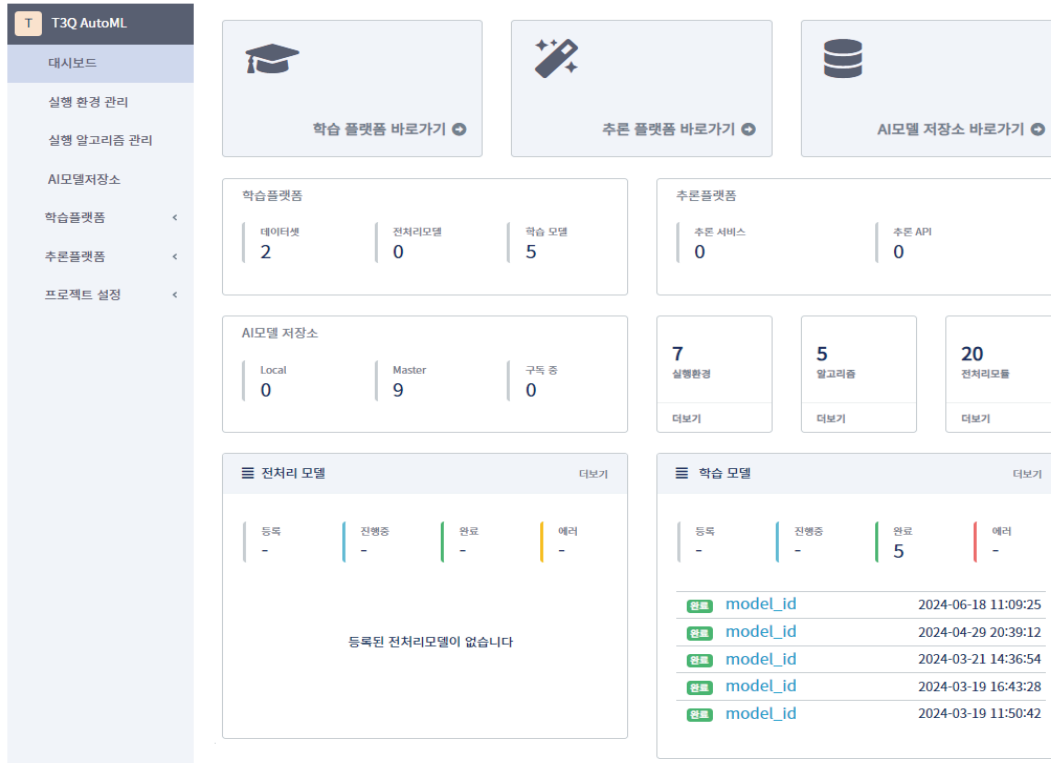


Fig. 3. Dashboard Screen of the Developed AutoML Web Platform

## 4. 실험 및 결과

### 4.1 데이터셋

본 연구에서는 AutoML 웹 플랫폼의 성능을 평가하기 위해 다양한 텍스트 분류 실험을 수행하였다. 실험 대상은 스팸 문자 메시지 분류와 구인 광고 분석의 두 가지 주요 작업이었다. 각 데이터셋은 학습 데이터와 테스트 데이터로 나뉘었으며, 비율은 각각 80%와 20%로 설정되었다.

#### 1) 스팸 문자 메시지 데이터셋

이 데이터는 데이터 분석 경진대회 플랫폼인 Kaggle에서 공개된 스팸 문자 메시지 데이터셋으로, 총 5,157개의 메시지를 포함하고 있다. 메시지는 '스팸(spam)'과 '스팸 아님(ham)'으로 분류되며, 이진 분류 문제를 다룬다.

#### 2) Craigslist 구인 광고 데이터셋

이 데이터셋은 Craigslist 웹사이트에서 수집된 구인 광고 텍스트로 구성되어 있으며, 각 광고는 직업 카테고리 라벨을 포함하고 있다. 총 13,730개의 광고로 이루어져 있으며, 6개의 직업 카테고리로 분류된다. 카테고리는 회계(accounting), 관리(administrative), 고객 서비스(customerservice), 교육(education), 식음료(foodbeverage), 노동(labor)이다.

### 4.2 실험 환경 및 설정

실험은 Ubuntu 22.04.3 LTS 운영체제, Intel(R) Xeon(R) CPU 2.00GHz, 251GB RAM 환경에서 진행하였다. 실험에 사용한 Python 버전은 3.8.1이다. 모델 학습에 사용된 하이퍼파라미터는 epoch 10, vector 100으로 설정하였다. 추가적으로, 모델의 일반화 능력을 평가하기 위해 K-fold 교차 검증 방법을 적용하였다.

### 4.3 실험 결과

Table 1은 스팸 문자 메시지 데이터셋과 Craigslist 구인 광고 데이터셋에 대한 리터보드를 나타낸다. 각 모델의 클래스별 평균 오류율, 로그 손실, RMSE 값이 포함되어 있으며, 이를 지표로 모델의 성능을 비교하였다.

#### 1) 스팸 문자 메시지 데이터셋

로그 손실을 기준으로 가장 성능이 우수한 모델은 Stacked Ensemble('Best of Family')이었다. 이 모델은 로그 손실 0.06319, 클래스별 평균 오류율 0.04353, RMSE 0.12292로 높은 성능을 보였다. 그 뒤를 이어 Stacked Ensemble('All Models')과 GBM\_1 모델이 높은 성능을 나타내었다. Table 2는 스팸 문자 메시지 데이터셋에 대한 성능 평가 지표를 제시한다. 가장 좋은 성능을 보인 Stacked Ensemble('Best of Family') 모델은 정확도 98.36%, 정밀도 96.71%, 재현율 98.11%, F1-Score 97.39%의 성능을 보였다.



Table 1. Leaderboard Performance Metrics for Each Dataset

Dataset	Models	Log loss	Mean per class error	RMSE
Spam Text Message	StackedEnsemble BestOfFamily	0.06319	0.04353	0.12292
	StackedEnsemble AllModels	0.06325	0.04398	0.12291
	GBM_1	0.06475	0.04583	0.12149
	XGBoost_2	0.06907	0.05219	0.12885
	XGBoost_1	0.07277	0.04664	0.13228
	GLM_1	0.08027	0.04833	0.14166
	DRF_1	0.11807	0.06200	0.14556
Craigslis Job	StackedEnsemble AllModels	0.53123	0.19189	0.41125
	StackedEnsemble BestOfFamily	0.53231	0.19113	0.41156
	GBM_1	0.54145	0.19524	0.40895
	XGBoost_1	0.56914	0.20066	0.41244
	XGBoost_2	0.57820	0.20225	0.41156
	GLM_1	0.65400	0.21705	0.45132

Table 2. Performance Evaluation Results on the Spam Text Message Dataset

Metric	Value (%)
Accuracy	98.36
Precision	96.71
Recall	98.11
F1-score	97.39

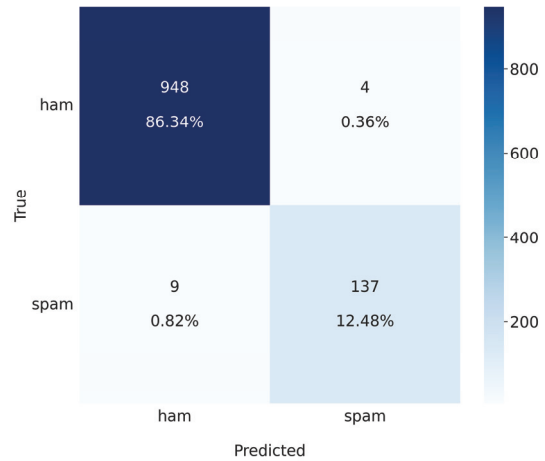
Table 3. Performance Evaluation Results on the Craigslis Job Dataset

Metric	Value (%)
Accuracy	80.88
Precision	81.36
Recall	81.44
F1-score	81.37

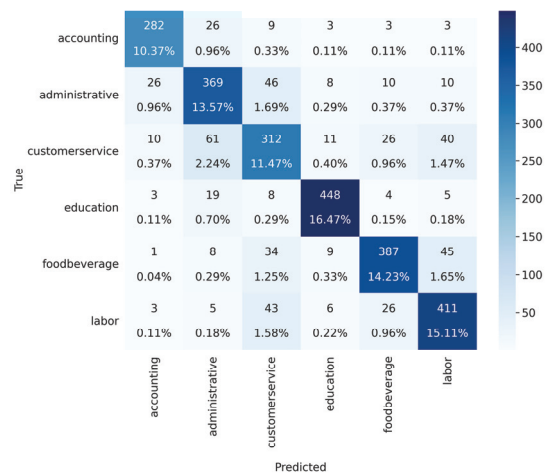
2) Craigslist 구인 광고 데이터셋

로그 손실을 기준으로 Stacked Ensemble('All Models') 모델이 가장 우수한 성능을 보였다. 이 모델은 로그 손실 0.53123, 클래스별 평균 오류율 0.19189, RMSE 0.41125를 보였다. Stacked Ensemble('Best of Family') 모델이 그 뒤를 이었으며, 다른 모델들에 비해 상대적으로 낮은 로그 손실, 클래스별 평균 오류율, RMSE 값을 나타냈다. Table 3은 Craigslist 구인 광고 데이터셋에 대한 성능 평가 지표표를 보여준다. 가장 좋은 성능을 보인 Stacked Ensemble('All Models') 모델은 정확도 80.88%, 정밀도 81.36%, 재현율 81.44%, F1-Score 81.37%의 성능을 보였다.

Fig. 4는 두 데이터셋의 혼동 행렬을 시각화한 것이다. Fig. 4a는 스팸 문자 메시지 데이터셋의 혼동 행렬을, Fig. 4b는 Craigslist 구인 광고 데이터셋의 혼동 행렬을 보여준다. 두 데이터셋 모두에서 클래스별로 균형 잡힌 성능을 보여 모델이 분류 작업을 효과적으로 수행했음을 확인하였다.



(a) Spam Text Message Dataset



(b) Craigslist Job Dataset

Fig. 4. Confusion Matrix for Each Dataset

실험 결과, 제안된 H2O 기반 AutoML 웹 플랫폼은 다양한 텍스트 분류 작업에서 높은 성능을 보였다. 각 모델의 성능은 데이터셋에 따라 다소 차이가 있었으나, 전반적으로 높은 정확도와 정밀도를 보였다. 특히, Stacked Ensemble을 사용한 모델은 대부분의 경우 단일 모델보다 더 우수한 성능을 보였다.

5. 결 론

본 연구에서는 텍스트 분류 자동화를 위한 H2O 기반 AutoML 웹 플랫폼을 개발하였으며, 이를 통해 텍스트 데이터 처리의 복잡한 단계를 통합적으로 자동화하는 방법을 제시하였다. 제안된 플랫폼은 기존 AutoML 시스템의 한계를 보완하여, 특히 비정형 텍스트 데이터를 효과적으로 처리할 수 있는 기능을 제공한다. 이 플랫폼은 텍스트 데이터의 전처리, 특징 추출, 모델 학습, 그리고 평가까지의 모든 과정을 자동화하여, 사용자가 복잡한 작업 없이도 효율적이고 고성능의 텍스트 분류 모델을 구축할 수 있도록 지원한다. 다양한 데이터셋을 활

용한 실험 결과, 본 플랫폼은 높은 정확도와 효율성을 보여주었으며, 특히 Stacked Ensemble 모델을 적용한 경우 우수한 성능을 나타냈다.

본 연구의 주요 기여는 다음과 같다. 첫째, 텍스트 분류 과정의 복잡한 단계를 자동화하여, 비전문가도 쉽게 고성능의 텍스트 분류 모델을 구축할 수 있도록 하였다. 둘째, H2O AutoML과 단어 임베딩 기법을 결합하여 텍스트 데이터를 보다 효과적으로 처리할 수 있는 방법을 제안하였다. 셋째, 데이터 업로드부터 모델 평가에 이르는 전 과정을 시각적으로 관리할 수 있는 사용자 친화적인 인터페이스를 제공하여 사용자 편의성을 크게 향상시켰다. 이 플랫폼은 데이터 전처리, 단어 임베딩, 머신러닝 모델 학습 및 평가의 모든 단계를 자동화하여 사용자가 텍스트 분류 작업을 보다 쉽고 효율적으로 수행할 수 있도록 돕는다. 본 연구를 통해 다양한 응용 분야에서 텍스트 데이터를 효과적으로 처리할 수 있는 가능성을 확인하였다.

향후 연구에서는 대규모 언어 모델(LLM)을 활용하여 텍스트 분류 성능을 더욱 향상시키는 방안을 모색할 예정이다. LLM의 사전 학습된 언어 처리 능력을 플랫폼에 적용함으로써, 더 복잡한 텍스트 데이터 처리 및 분류 문제를 해결할 수 있을 것으로 기대된다. 이러한 발전을 통해 다양한 분야에서 텍스트 분류의 성과와 활용 가능성을 더욱 증대시킬 수 있을 것이다.

### References

- [1] M. Zareapoor and K. R. Seeja, "Feature extraction or feature selection for text classification: A case study on phishing email detection," *International Journal of Information Engineering and Electronic Business*, Vol.7, No.2, pp.60, 2015.
- [2] O. Bruna, H. Avetisyan, and J. Holub, "Emotion models for textual emotion classification," *Journal of Physics: Conference Series*, Vol.772, No.1, 2016.
- [3] T. Moh, T.-S. Teng, and Z. Zhang, "Cross-lingual text classification with model translation and document translation," in *Proceedings of the 50th Annual Southeast Regional Conference*, 2012.
- [4] R. Desai et al., "TextBrew: Automated Model Selection and Hyperparameter Optimization for Text Classification," *International Journal of Advanced Computer Science and Applications*, Vol.13, No.9, 2022.
- [5] X. He, K. Zhao, and X. Chu, "AutoML: A survey of the state-of-the-art," *Knowledge-Based Systems*, Vol.212, pp. 106622, 2021.
- [6] J. Han, K. S. Park, and K. M. Lee, "An automated machine learning platform for non-experts," in *Proceedings of the International Conference on Research in Adaptive and Convergent Systems*, 2020.
- [7] E. LeDell and S. Poirier, "H2o automl: Scalable automatic machine learning," in *Proceedings of the AutoML Workshop at ICML*, Vol.2020, San Diego, CA, USA: ICML, 2020.
- [8] S. Brändle et al., "Evaluation of representation models for text classification with AutoML tools," in *Proceedings of the Future Technologies Conference (FTC) 2021*, Vol.2, Springer International Publishing, 2022.
- [9] K. Kowsari et al., "Text classification algorithms: A survey," *Information*, Vol.10, No.4, pp.150, Apr. 2019.
- [10] M. M. Mirończuk and J. Protasiewicz, "A recent overview of the state-of-the-art elements of text classification," *Expert Systems with Applications*, Vol.106, pp.36-54, 2018.
- [11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
- [12] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. 2014 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2014.
- [13] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [14] R. Li, "A review of machine learning algorithms for text classification," *Cyber Security*, Vol.226, 2022.
- [15] M. Lewis, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.
- [16] T. B. Brown, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.
- [17] Y. Chae and T. Davidson, "Large language models for text classification: From zero-shot learning to fine-tuning," Open Science Foundation, 2023.
- [18] F. Stoica and L. F. Stoica, "AutoML Insights: Gaining Confidence to Operationalize Predictive Models," 2024.
- [19] O. Levy, Y. Goldberg, and I. Dagan, "Improving distributional similarity with lessons learned from word embeddings," *Transactions of the Association for Computational Linguistics*, Vol.3, pp.211-225, 2015.
- [20] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.
- [21] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, Vol.29, No.5, pp.1189-1232, 2001.
- [22] N. E. Breslow, "Generalized linear models: checking assumptions and strengthening conclusions," *Statistica Applicata*, Vol.8, No.1, pp.23-41, 1996.
- [23] A. Candel et al., "Deep learning with H2O," H2O.ai Inc., pp.1-21, 2016.
- [24] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, Vol.63, pp.3-42, 2006.



**송 하 윤**

<https://orcid.org/0009-0007-5935-9797>  
e-mail : hayoon@kiro.re.kr  
2020년 가톨릭대학교 심리학과(학사)  
2023년 고려대학교 뇌공학과(석사)  
2023년~현 재 한국로봇융합연구원  
인공지능로봇연구본부 주임연구원

관심분야: Artificial Intelligence, Deep Learning,  
Human-Robot Interaction



**전 광 우**

<https://orcid.org/0000-0002-4889-6083>  
e-mail : jeonkw@kiro.re.kr  
2011년 한밭대학교 기계설계공학과(학사)  
2013년 한밭대학교 기계설계공학과(석사)  
2024년 경북대학교 기계공학과(박사수료)  
2012년~2014년 한국로봇융합연구원  
주임연구원

2014년~2015년 한국화이바 특수사업본부 설계기술센터  
선임연구원

2015년~현 재 한국로봇융합연구원 인공지능로봇연구본부  
선임연구원

관심분야: Exo-Skeleton Robot, Human-Robot Interaction,  
Design Optimization



**강 전 성**

<https://orcid.org/0000-0002-7061-5794>  
e-mail : kjs2605@kiro.re.kr  
2016년 수원대학교 전기전자공학부(학사)  
2018년 동국대학교 전자전기공학부(석사)  
2018년~2021년 전문연구요원(AI Dept.)  
2021년~현 재 한국로봇융합연구원  
인공지능로봇연구본부 주임연구원

관심분야: Computer Vision, AutoML, Generative AI, Deep  
Learning



**윤 준 원**

<https://orcid.org/0009-0000-1737-5181>  
e-mail : jyoon@kiro.re.kr  
2020년 고려대학교 기계공학과(학사)  
2022년 고려대학교 기계공학과(석사)  
2023년~현 재 한국로봇융합연구원  
인공지능로봇연구본부  
주임연구원

관심분야: Artificial Intelligence, Deep Learning,  
Generative Model



**박 범 준**

<https://orcid.org/0000-0001-6905-7740>  
e-mail : parkbj26@kiro.re.kr  
2017년 인하대학교 전기공학과(학사)  
2019년 연세대학교 전기전자공학과(석사)  
2019년~2021년 한국과학기술연구원  
로봇미디어연구소 연구원

2022년~현 재 한국로봇융합연구원 인공지능로봇연구본부  
주임연구원

관심분야: Deep Reinforcement Learning, Swarm  
Intelligence, Space Robotics, Motion Planning



**정 현 준**

<https://orcid.org/0000-0001-5533-9587>  
e-mail : hjchung@kiro.re.kr  
2005년 University of Iowa  
기계공학과(석사)  
2009년 University of Iowa  
기계공학과(박사)

2015년~2021년 한국로봇융합연구원 선임연구원

2021년~현 재 한국로봇융합연구원 인공지능로봇연구본부  
책임연구원, 본부장

관심분야: Dynamics and Control, Optimization  
Algorithms, Computational Decision Making,  
Robotics, Modeling and Simulation



**김 준 영**

<https://orcid.org/0000-0002-9888-4822>  
e-mail : junyoung.kim@kiro.re.kr  
2015년 광운대학교 제어측정공학과(석사)  
2021년 대구경북과학기술원 로봇공학전공  
(박사)  
2021년~2023년 대구경북과학기술원  
로봇및기계전자공학연구소  
(PostDoc)

2023년~현 재 한국로봇융합연구원 인공지능로봇연구본부  
선임연구원

관심분야: Wearable Robot, Rehabilitation Engineering,  
Nonlinear Control, Actuator