Journal of KIIT. Vol. 22, No. 1, pp. 61-69, Jan. 31, 2024. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2024.22.1.61

감정분석 모델성능 향상을 위한 GPT기반 데이터 증강 방법

김철민*, 정현준**

GPT-based Data Augmentation Method for Improving Emotion Analysis Model Performance

Cheolmin Kim*, Hyunjun Jung**

※ 이 연구는 정부(과학기술정보통신부)의 재원으로 한국 연구재단의 지원을 받아 수행되고 있습니다.
(No. NRF-2022R1G1A1008493)

요 약

머신러닝에 중요한 데이터 셋의 구성과 품질을 올리는 기술인 데이터 증강은 적은 양의 데이터를 바탕으로 다양한 알고리즘을 통해 데이터의 양을 늘리는 기술이다. 본 연구에서는 대규모 언어 모델인 GPT(Generative Pre-trained Transformer)를 활용한 데이터 증강으로 감정 분석 모델의 성능을 향상시키는 방법을 제안하고 평가한다. 데이터 셋의 클래스 불균형 문제를 해결하기 위해 가중치를 적용한 로직을 사용하였고, 생성된 데이터의 품질 및 다양성에 대한 한계를 극복하기 위해 프롬프트 엔지니어링을 적용했다. 실험결과, 제안한 방법은 데이터의 품질을 유지하면서 다양성을 높이고, 클래스 불균형 문제를 효과적으로 해결할 수 있어 KoBERT 모델을 이용해 GPT를 활용한 데이터 증강이 모델의 성능을 향상시킬 수 있음을 보였다.

Abstract

Data augmentation, a technology that increases the composition and quality of datasets important for machine learning, is a technology that increases the amount of data through various algorithms based on a small amount of data. In this study, we propose and evaluate ways to improve the performance of emotion analysis models by augmenting data using a large language model, Generative Pre-trained Transformer(GPT). We used weighted logic to solve the class imbalance problem of datasets and applied prompt engineering to overcome limitations on the quality and diversity of generated data. As a result of the experiment, it was shown that the proposed method can increase diversity while maintaining the quality of the data and effectively solve the class imbalance problem, so that data augmentation using GPT can improve the performance of the model using the KoBERT model.

Keywords
GPT, data augmentation, large-scale language model, prompt engineering

* 군산대학교 소프트웨어학부 학사과정
- ORCID: https://orcid.org/0009-0004-7579-8771
** 군산대학교 소프트웨어학부 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-6717-1395

· Received: Oct. 31, 2023, Revised: Dec. 26, 2023, Accepted: Dec. 29, 2023
· Corresponding Author: Hyunjun Jung
Dept. of Software at Kunsan National University, 558, Daehak-ro, Kunsan-si, Jeollabuk-do, Republic of Korea
Tel.: +82-63-469-8917, Email: junghj85@kunsan.ac.kr

Ⅰ. 서 론

머신러닝은 의료, 자동차, 통신, 금융과 같은 다양한 산업에서 중요한 역할을 하고 있다. 머신러닝에서 사용되는 심층 신경망은 여러 은닉 층을 포함해 모델의 입력과 출력 사이의 관계를 학습하는 표현적인 모델이다. 이러한 기술의 핵심은 학습 알고리즘과 그 알고리즘이 학습하는 데이터 셋에 있다[1]. 인공지능 모델의 성능은 학습 데이터 셋에 의존성이 높으며, 데이터의 품질, 균형, 다양성이 모델의 예측 정확도와 일반화 능력에 결정적인 영향을 미친다.

하지만 이러한 중요성에도 불구하고, 실제 데이터 셋은 자주 불균형한 특성을 보이곤 하는데, 이러한 클래스 불균형은 모델 학습에 매우 부정적인 영향을 미칠 수 있으며, 데이터가 부족하거나 소수 클래스의 데이터가 충분히 학습되지 않을 경우, 모델의 일반화 능력이 떨어지게 되는 과소적합(Underfitting) 현상과 과적합(Overfitting)현상이 발생한다[2].

이런 문제에 대한 다양한 해결 전략 중 볼츠만 머신(RBM, Restricted Boltzmann Machine)을 통해 신경망 각 계층을 사전학습(Pre-training) 하는 방법[3]과 일부 노드를 무작위로 학습시키는 드롭아웃(Dropout)[4]처럼 다양한 방법들이 시도되고 있다[5][6]. 그 중 데이터 증강은 데이터 셋이 부족한 상황에서 기존 데이터를 변형하거나 새로운 데이터를 생성하여 학습 데이터 셋을 확장하는 기술을 말한다.

이 기법은 모델이 보다 다양한 데이터 패턴을 학습하도록 도와주고, 과적합과 과소적합의 위험을 줄여줄 수 있다. 그러나 데이터 증강 방법에는 여러 문제점이 있다. 증강된 데이터의 품질이 원래 데이터 셋과 비교해 일관성이 떨어질 수 있고, 계산 비용이 증가하며, 시간이 오래 걸릴 수 있고 특정 분야의 전문 지식이 필요할 수 있다는 문제점이 존재한다[7][8].

이러한 문제를 극복하고자 본 논문에서는 생성형 AI(Artificial Intelligence) 모델인 GPT(Generative Pre-trained Transformer)를 사용한 새로운 데이터 증강 방법을 제안한다. GPT는 대규모의 텍스트 데이터를 학습하여 복잡한 문맥을 이해할 수 있는 고성능의 자연어 처리 모델이다[9]. 이 모델을 활용하여 각 클래스의 데이터를 자연스럽게 증강할 수 있다면, 기존 데이터 증강 방법의 여러 문제점을 극복하고, 모델의 성능을 높일 수 있을 것으로 예상된다. GPT의 이러한 장점을 적극 활용하여, 데이터 증강을 통한 클래스 불균형 문제 해결에 대한 새로운 접근법을 제시한다.

GPT 모델의 프롬프트를 조작하여 원하는 결과를 얻기 위해서는 프롬프트 엔지니어링 기법을 개발할 필요가 있다. 프롬프트 엔지니어링은 모델의 입력을 조작하여 원하는 결과를 얻기 위한 방법론이며, GPT 기반에서 프롬프트를 목적에 맞게 활용하여 불균형 클래스에서 데이터를 생성하고자 한다.

감정 분석 모델을 개발할 때는 종종 특정 감정에 대한 데이터가 부족하거나 불균형적인 경우가 많다. 이로 인해 모델이 특정 감정을 인식하는 능력이 제한되거나 편향될 수 있다. 데이터 증강은 이러한 불균형을 해소하고, 모델이 다양한 감정 상황에 더 잘 대응할 수 있도록 돕는다. 특히 GPT와 같은 고급 언어 모델을 사용한 데이터 증강은 실제와 유사한 다양한 감정 표현을 생성할 수 있으며, 이를 통해 감정 분석 모델은 더욱 정확하고 균형 잡힌 성능을 보일 수 있다.

감정 분석에 맞는 프롬프트 엔지니어링을 통해, GPT 기반의 데이터 증강이 실제로 모델 성능 향상에 얼마나 기여하는지에 대한 실험적 검증을 실시한다. 나아가, 다양한 자연어 처리 분야에 어떻게 적용될 수 있는지에 대한 가능성도 확인할 수 있을 것이다.

머신러닝 기술의 중요성과 심층 신경망의 구조, 그리고 학습 데이터 셋의 중요성을 강조하며, 데이터의 품질과 균형, 다양성이 모델 성능에 미치는 영향, 그리고 클래스 불균형 문제와 그 해결을 위한 다양한 기존 방법들, 특히 데이터 증강의 필요성과 문제점을 설명하였고, 본 연구에서 GPT를 이용한 새로운 데이터 증강 방법을 제안하고, 프롬프트 엔지니어링을 통한 불균형 클래스 데이터 생성의 가능성을 탐색한다. 이를 통해 감정 분석 모델 개발에 있어서 GPT 기반 데이터 증강의 효과를 실험적으로 검증하고자 하며, 데이터 증강이 머신러닝과 인공지능 분야에서 어떻게 활용될 수 있는지에 대한 새로운 통찰을 제공하고자 한다.

II. 관련 연구

머신러닝에서 널리 사용되는 데이터 증강은 모델의 성능을 향상시키는 널리 알려진 기술인만큼, 다양한 관련 연구가 이뤄지고 있다. 기존의 데이터 증강은 원본 학습 데이터를 변형하여 새로운 학습 샘플을 생성하며, 모델의 일반화 능력을 향상시키는 방법을 적용해 왔다. 대표적인 데이터 증강 연구인 EDA:Easy Data Augmentation 연구[10]를 살펴보면 동의어 교체, 단어 무작위 삽입, 삭제 기법으로 텍스트를 약간씩 변형하여 새로운 학습 샘플을 생성한다. 또한 적은 양의 데이터에 적용할 수 있는 계층별 데이터 증강 알고리즘에 관한 연구[11]에서는 기존 데이터의 개수가 부족한 클래스의 샘플을 보간하여 새로운 소수 클래스의 데이터를 생성하는 SMOTE 기법이 있었으며[12], 그걸 발전시켜 클래스의 샘플을 3개의 그룹(분류 모델 성능 향상, 분류기 성능 감소, 분류가 어려운 데이터 포인트)로 나눈 MSMOTE 기법을 볼 수 있었다[13].

하지만 이런 기법들은 기존의 데이터 포인터의 다양성을 기반으로 새로운 샘플을 생성하기에 새로운 특성이나 패턴을 만들어 낼 수 없는 문제점이 있다. 또한 맥락을 완전히 이해하지 못해 새로운 데이터가 원본 데이터의 의미를 왜곡하거나 잘못된 정보를 포함해 생성할 수 있으며, 원본의 복잡성과 다양성을 완전히 반영하지 못할 수 있다.

OpenAI의 GPT 시리즈는 대규모 언어 모델이다. GPT-3는 이 시리즈의 주요 모델 중 하나로, 놀라운 언어 이해 및 생성 능력을 갖추고 있으며, 이러한 기술을 바탕으로 다양한 특화 모델이 개발되어왔다. Codex는 프로그래밍 코드 생성에 특화되어 있으며, InstructGPT는 사용자의 지시에 더 정확하게 응답하기 위해 개발되었고 ChatGPT는 대화형 챗봇 및 대화 기반 응용 프로그램에 적합하며, GPT-4는 GPT-3보다 더 향상된 성능을 가진 발전된 모델로 개발되었다. 이러한 모델들은 자연어 처리 분야에서 다양한 응용 가능성을 보여준다[14].

따라서 GPT는 Large Language Model로 대규모 데이터 셋에서 훈련된 언어 모델이자 수십억 개의 파라미터를 가짐으로 다양한 자연어처리 작업에 높은 성능을 보이니, 부족한 데이터 클래스에 대해 새로운 특성이나 패턴을 가진 샘플을 만들어 내 효과적으로 감정 분류 작업의 성능을 향상시키는 방안이 될 수 있으며, 목적에 맞는 프롬프트 엔지니어링을 통해 원본 데이터의 복잡성을 유지하면서도 다양성을 보강하여 보다 품질 좋은 데이터가 증강된 데이터 셋을 확보할 수 있다.

III. GPT기반 감정 데이터 증강 방법

본 장에서는 본 논문이 제안하는 GPT를 활용한 불균형한 감정분석 데이터 셋을 보완하기 위한 증강 전략에 대해 상세히 설명한다.

3.1 클래스 불균형 해결 전략

데이터 셋의 클래스 불균형은 머신러닝 모델의 성능에 큰 영향을 미치는 문제 중 하나이다. 이 문제를 해결하기 위해 본 연구에서는 동적 확률 가중치 로직을 적용하였다. 이 로직은 각 클래스에 대한 데이터 수가 적은 클래스에 높은 확률을 부여하여, 증강 과정에서 그 클래스의 데이터가 더 많이 생성되도록 한다. 확률 가중치는 각 라벨 i에 속하는 데이터의 개수 N(i)를 전체 데이터의 개수 N를 활용하여 동적으로 적용한다. 여기서 N은 $\sum_{i=1}^{K} N(i)$로 계산한다. 이렇게 확률 가중치 로직을 적용하게 된다면, 데이터가 생성될 때마다 확률 가중치가 실시간으로 조정되어 부족한 클래스의 데이터를 더 많이 생성할 수 있다.

$$Dynamic\,\mathrm{P}robability\,Weighting(DPW): P_i = \frac{1}{R_i},$$
$$Weight\,Normalization(WN): P_i = \frac{P_i}{\sum_{j=1}^{K} P_j} \quad (1)$$

동적 확률 가중치 로직(Dynamic probability weighting logic)은 식 (1)로 표현할 수 있다. 클래스별 데이터량의 상대적인 비율을 나타내는 R을 식으로 표현하면 $R_i = \frac{N}{N_i}$로 표현이 가능하며, 여기서 N은 전체 데이터의 수, N(i)는 각 클래스 I에 속하는 데이터의 수이다.

K는 전체 클래스 수를 의미한다. 가중치 R(i)이 크다는 것은 클래스 i의 데이터가 전체 데이터 셋에서 차지하는 비율이 낮다는 것을 의미하기에 R(i)이 큰 클래스에 가중치를 더 높게 부여함으로 가중치 정규화를 통해 모든 클래스에 대한 가중치 합이 1이 되도록 정규화 하여 모든 클래스에 대해 균일하게 데이터를 생성할 수 있도록 한다.

이러한 동적 확률 가중치 로직은 데이터 량의 변화에 따라 동적으로 확률이 조정되어 부족한 클래스에 대한 불균형 문제를 해결하는데 유연하게 적용될 수 있다. 사용할 데이터 셋은 6가지 감정 클래스로 구분되어 있고, 로직을 적용한 결과, 그림 1과 같이 데이터량이 다른 각 클래스마다 의도한대로 데이터가 생성되어 균일한 데이터 셋이 되었음을 확인할 수 있었다.

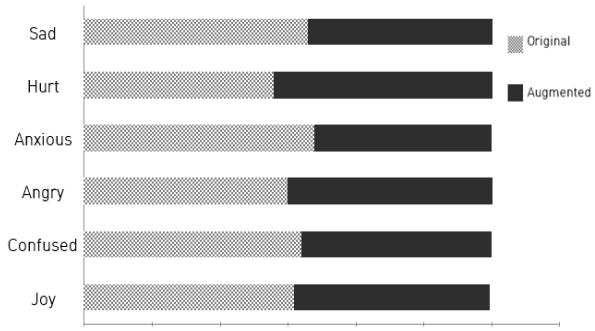

그림 1. 클래스별 데이터 비율
Fig. 1. Percentage of data by class

3.2 프롬프트 엔지니어링을 통한 감정 데이터 생성

가중치 로직으로 증강시킬 데이터 클래스가 결정되었다면, GPT 모델은 데이터 증강을 위해 어떤 클래스의 감정 데이터를 생성해야 하는지 결정한다. 이어서 GPT의 자연스러운 텍스트 생성 능력을 이용해 품질이 뛰어나고 넓은 다양성을 가진 데이터를 생성하기 위해서는 GPT에게 세밀하고 적절한 프롬프트를 설계하여 지시해야 한다[15]-[17]. 본 논문에서는 감정 데이터 생성을 위해 크게 Generative Prompt와 Conditional Prompt전략을 사용하여 그림 2와 같이 프롬프트 엔지니어링을 적용했다.

Generative Prompt는 모델의 창의성을 높이기 위해 사용하는 초기문맥(Initial context)전략으로, 자연어 모델이 문장 또는 문단을 생성하기 시작할 때 주어지는 초기 입력을 의미한다. 기존 데이터의 참조 없이 새로운 주제를 가진 문장을 생성하려면, GPT에 구체적이고 창의적인 프롬프트를 제공하는 것이 필요하다. 이를 통해 Generative Prompt를 진행할 때, GPT는 기존 데이터의 내용을 참조하지만 다른 컨텍스트와 주제를 가진 문장을 생성하도록 지시받게 되며, 이는 데이터 셋의 다양성을 향상시키는 데 중요한 역할을 한다. 이러한 접근 방식은 모델이 기존 데이터셋과는 다른 새로운 상황과 시나리오를 학습하게 하여, 데이터의 다양성과 일반화 능력을 증진시킬 수 있다.

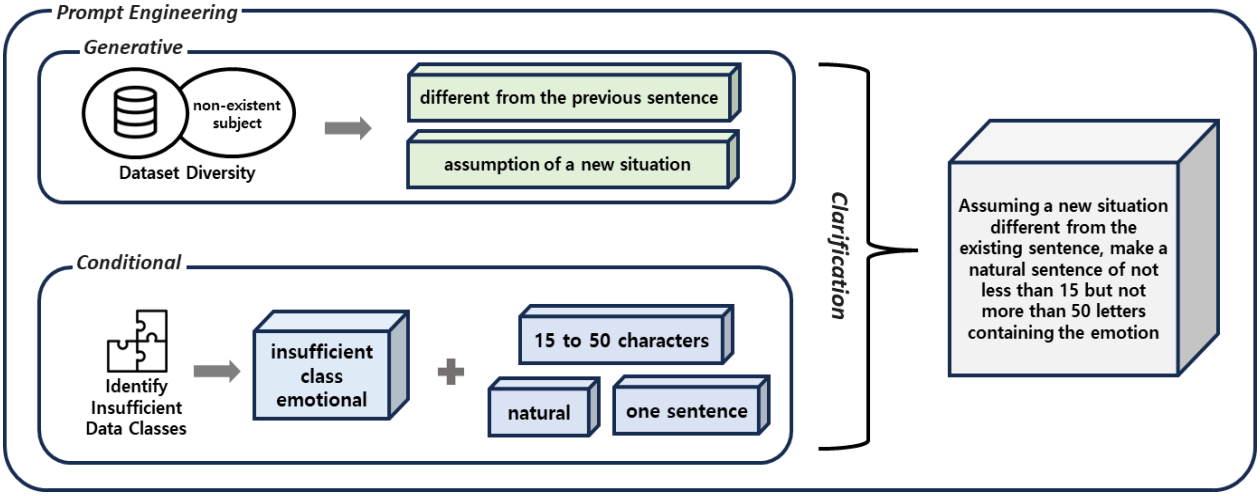

그림 2. 프롬프트 엔지니어링
Fig. 2. Prompt engineering

Conditional Prompt는 모델이 특정 조건을 만족하도록 하는 조건 토큰(Conditioning tokens)방법으로, 생성 모델에 특정 조건을 부여하는 데 사용되는 특수한 입력 토큰을 의미한다. 여기서는 GPT 모델이 가중치 로직에서 결정된 데이터 클래스의 감정을 받아서 문장이 생성될 때, 해당 감정을 반영하도록 하는 것을 목표로 한다. 또한 글자 수가 너무 적거나 많지 않게 위해 기존 데이터와 큰 차이가 나지 않도록 생성 글자 수를 조절하도록 하였고, 자연스러운 문장이 생성되도록 하는 조건을 추가하여 목적에 맞는 문장이 생성될 수 있도록 지시하였다.

이 연구에서는 GPT 모델이 데이터 증강 명령을 보다 효과적으로 이해하고 처리할 수 있도록, Prompt Clarification과 쿼리 재구성(Query reformulation) 기법을 적용하였다. 이는 그림 2의 Clarification이 적용된 예시 문장처럼 필요한 프롬프트를 재구성하여 모델이 보다 정확하게 파악하고 관련된 데이터를 효과적으로 생성할 수 있도록 한다. 이러한 접근은 GPT 모델이 프롬프트를 더 명확하게 이해하고, 감정 데이터의 다양성과 품질을 높이는 데 기여한다.

이렇게 프롬프트 엔지니어링을 통해 원본 데이터의 복잡성을 유지하면서 기존 데이터가 가지고 있지 않은 다양성을 보강하여 품질 좋은 데이터가 증강된 데이터 셋으로 모델이 전보다 개선된 성능을 보이게 된다.

IV. 실험 및 평가

4.1 실험 환경

실험에 사용할 데이터는 AI Hub의 감성대화 말뭉치 데이터 셋을 사용했다[18]. 이 데이터 셋은 6가지 감정 대분류 코퍼스를 가지고 있는 다중분류 학습을 위한 데이터 셋이다. 한국어 감정 데이터 셋을 사용함에 따라, BERT를 활용해 만들어졌고 한국어에 특화되어있는 KoBERT 모델을 사용해 실험을 진행했다. 데이터 증강에 따른 모델의 성능 평가를 위해 원본 데이터만 학습한 모델, 그리고 그림 3의 증강과정을 통해 만들어진 데이터 셋을 바탕으로 증강 데이터의 품질과 다양성을 확인하기 위해 증강 데이터만 학습한 모델, 증강 데이터가 실제로 원본데이터만 사용했을 때보다 다양성이 증가하여 성능 향상을 보임을 확인하기 위해 원본과 증강 데이터를 50%씩 추출한 데이터를 학습한 모델, 그리고 원본 데이터 셋에 생성된 데이터가 합쳐져 증강된 데이터를 학습한 모델을 비교하며 실험을 진행했다.

4.2 실험 결과 및 비교

앞서 설명한 데이터 증강 방법으로 GPT 모델이 데이터를 생성한 결과, 표 1이 보여주는 것처럼 데이터가 생성됨을 확인하였다. 생성된 데이터가 실제로 모델의 성능 향상에 기여하는지 평가하기 위해, 먼저 데이터 불균형을 해소하였는지 확인하였다.

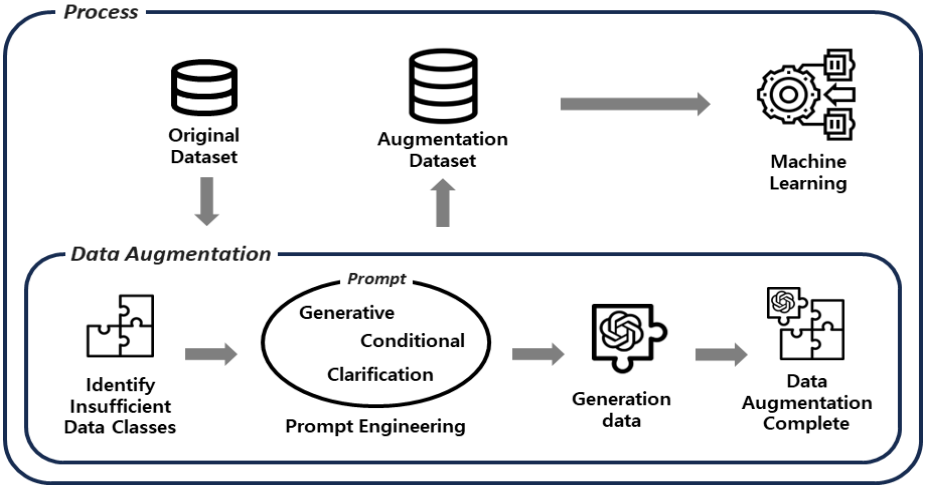

그림 3. 증강 과정
Fig. 3. Augmentation process

표 1. 데이터 생성 결과
Table 1. Data create results

Sentence	Emotion
언제까지 너에게 상처 받아야 하는지 나도 이제 지쳤어.	상처
내 꿈은 배우가 되는 건데, 현실이 그렇게 허락하지 않는 것 같아.	슬픔
지금은 알바생이지만, 월급 첫 날 얼마나 행복한지 모르겠더라!	기쁨
메시지 한 통 보내는 게 그렇게 어렵니? 화가 나서 손이 떨릴 지경이야.	분노
시험 전날이라 불안해서 그런지 잠이 안 오네, 어서 책이라도 읽어야겠어.	불안
기한 내 준비해야 하는 보고서가 삭제되어서 다시 작성해야 해.	당황

그림 1에서처럼 의도한대로 데이터를 생성할 때 더 적은 수의 데이터를 가진 클래스를 대상으로 샘플이 생성한 것을 확인할 수 있었다. 이어서 데이터 품질을 검증하기 위해 문장의 자연스러움과 복잡성을 측정하는 가독성 점수(Readability scores)를 측정하였다.

Flesch Reading Ease와 Gunning Fog Index는 텍스트의 가독성과 복잡성을 측정하는 지표다. FRE는 206.835-1.015(총 단어 수/총 문장 수) - 84.6(총 음절 수/총 단어 수)의 공식을 가지며 값이 높을수록 텍스트가 읽기 쉽다는 것을 의미한다. GFI는 0.4((총 단어 수/총 문장 수)+100(세 음절 이상 단어 수/총 단어 수))의 공식을 가지며 값이 낮을수록 읽기 쉽다는 것을 의미한다. 표 2의 결과로 원본 데이터와 생성 데이터의 값이 큰 차이를 보이지 않고 비교군으로 부자연스러운 데이터가 GFI에서 큰 차이를 보이고 있으므로 생성된 데이터의 품질이 부족하지 않음을 확인하였다.

표 2. 데이터 품질 검증
Table 2. Data quality verification

	Original data	Generated data	Unnatural data
Flesch reading ease	114.1865	112.1063	117.6266
Gunning fog index	3.1727	3.9905	1.8116

이어서 생성된 데이터의 다양성을 검증하기 위해 문장의 유사성 측정으로 비슷한 토픽을 다루고 있는지 Cosine Similarity with TF-IDF Vectors와 LDA Topic Modeling을 측정하였다. 코사인 유사도는 문서 내 단어의 벡터간의 유사성을 측정하여 주제적 유사성을 수치적으로 평가하는 방법이며, LDA는 문서 집합에서 주제를 찾아내는 토픽 모델링 방법으로 데이터셋 내에서 주제들의 분포를 파악해 주제를 식별하여 기존 데이터 셋과 얼마나 주제적으로 유사한지를 평가한다. 원본과 다른 상황을 생성한 데이터의 다양성을 확인하기 위해, 원본의 상황을 참조해 생성한 데이터를 비교 군으로 두고 측정했다. 표 3의 결과로 평균 코사인 유사도와 동일 도미넌트 토픽에서 원본과 다른 상황을 생성한 데이터가 낮은 값을 보이고 있으므로 생성된 데이터의 다양성이 확장되었음을 확인하였다[19].

표 3. 데이터 다양성 검증
Table 3. Data diversity verification

	Different situation	Reference situation
Cosine similarity with TF-IDF vectors	0.06636	0.4055
Topic modeling	0.2575	0.3892

추가적으로 PCA, t-SNE 그래프를 그림 4와 같이 생성하여 확인한 결과, 원본과 생성 데이터가 서로 유사한 형태를 보임으로 데이터의 품질이 큰 차이가 없음을 알 수 있었으며, 생성 데이터가 원본 데이터의 영역을 커버함을 보여줌으로써 데이터의 다양성이 향상될 수 있음을 재확인 할 수 있었다.

이렇게 클래스 불균형 문제점을 해소하고 생성된 데이터의 품질과 다양성을 확인하여 GPT기반 생성 데이터가 유효한 데이터임을 확인하고 KoBERT 모델에 4가지 데이터 셋을 학습한 결과, 학습 과정에 있어 데이터 증강이 모델에 보다 다양한 특징을 잘 학습하게 하였고, 그림 5와 같이 기존 데이터 모델 대비 10% 이상의 학습 성능 개선을 보였다.

학습된 모델을 테스트 해본 결과, 표 4와 그림 6에서 볼 수 있듯이 증강된 데이터 셋으로 학습한 모델이 원본 데이터만 학습한 모델보다 평균 18%의 성능 향상을 보였다.

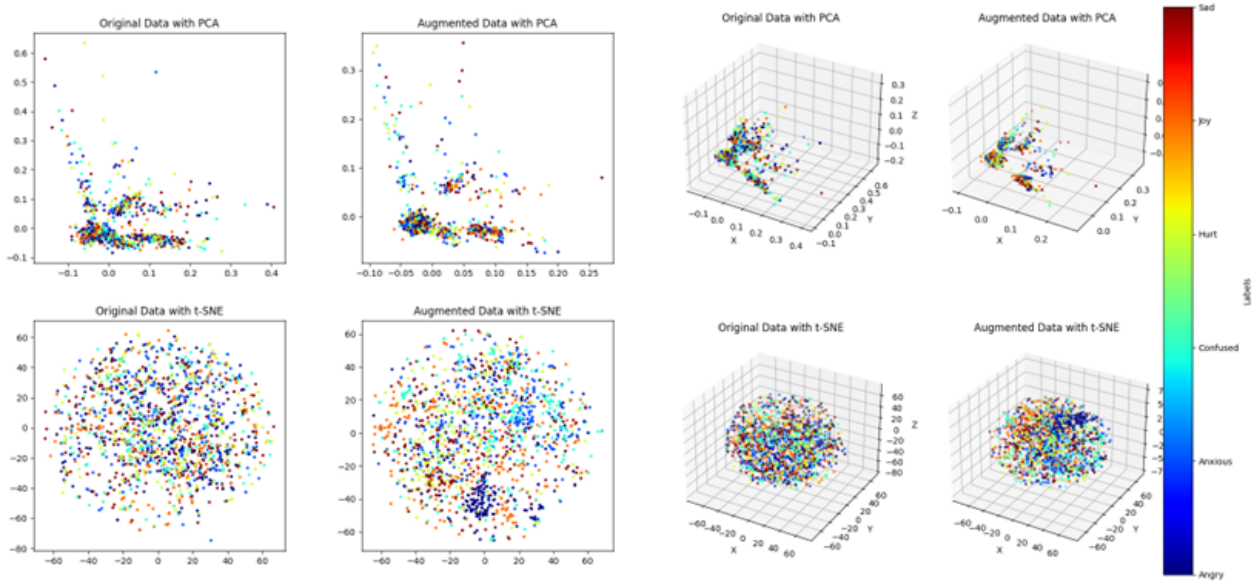

그림 4. PCA, t-SNE 그래프 그림의 선명도를 높여주세요, 텍스트 선명도도 높여주세요
Fig. 4. PCA, t-SNE graph

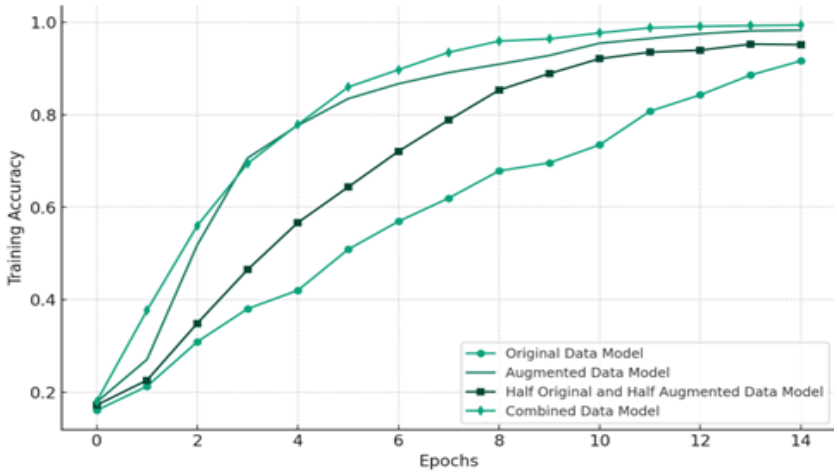

그림 5. 훈련 정확도 그래프
Fig. 5. Training accuracy graph

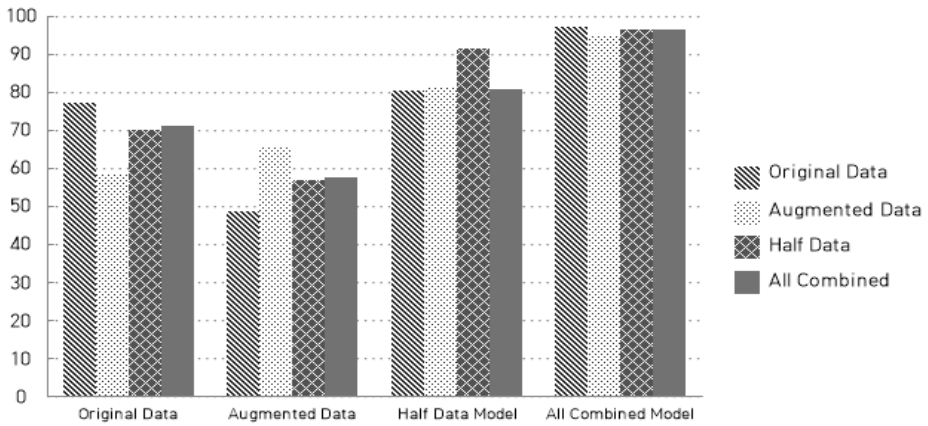

그림 6. 테스트 정확도 그래프
Fig. 6. Test accuracy graph

표 4. 테스트 정확도 검증
Table 4. Test accuracy verification

Model \ Data	Original data	Augmented data	Half data	Combined data
Original model	77.07	58.4	69.84	70.84
Augmented model	48.35	65.11	56.89	57.27
Half model	80.33	81.27	91.40	80.64
Combined model	97.12	94.8	96.28	96.35

기존 데이터로만 학습한 모델이 77%정도의 정확도를 보였는데, 생성된 데이터의 성능 향상을 확인하기 위해 원본 50%와 생성 50% 데이터를 학습한 모델이 보다 높은 정확도를 기록함으로 생성된 데이터가 성공적으로 모델 성능 향상을 끌어냈음을 보여주고 있으며, 최종적으로 원본 데이터에 생성 데이터가 증강된 모델에서는 95%정도의 높은 정확도를 보여주었다.

Ⅴ. 결론 및 향후 과제

이 논문에서는 감정 분석 데이터 셋의 클래스 불균형과 데이터의 품질 및 다양성 문제를 해결하기 위한 새로운 접근법을 제시하였다.

GPT를 활용한 데이터 증강 방법을 중심으로, 프롬프트 엔지니어링 기법을 이용하여 데이터의 품질과 다양성을 동시에 높이는 방법을 탐구하였다. 실험을 통해 본 연구의 방법이 클래스 불균형 문제를 효과적으로 해결하고, 데이터의 품질을 유지하면서도 다양성을 높일 수 있음을 입증하였다.

KoBERT 모델을 이용한 성능 평가에서는 기존 데이터셋 대비 약 10%의 학습 성능 향상과 18%의 테스트 성능 향상을 보였다. 이러한 결과는 GPT와 프롬프트 엔지니어링을 이용한 데이터 증강이 머신러닝 모델의 성능 향상에 기여할 수 있음을 보여준다.

또한, 본 연구에서는 다양한 프롬프트 엔지니어링 기법과 확률 가중치 적용 로직을 조합하여 데이터 셋을 구성하였다. 이를 통해 생성된 데이터 셋은 감정 분석뿐만 아니라, 다양한 자연어 처리 문제에도 적용 가능하며, 이로 인해 해당 분야의 연구에 중요한 기여를 할 것으로 기대된다.

향후 연구 방향으로는 다른 생성 모델과의 성능 비교, 실제 문제되는 상황에서 적용이 가능한 부분을 탐구하고 어떻게 적용할 수 있는지, 다양한 자연어 처리 문제에 본 연구의 방법론을 적용하여 효과를 검증하는 것이 중요하다.

이러한 방법과 결과를 통해 본 연구는 머신러닝과 자연어 처리 분야에서 데이터의 품질과 다양성, 그리고 클래스 불균형 문제를 효과적으로 해결할 수 있는 새로운 방법론을 제시하였다.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning", Nature, Vol. 521, pp. 436-444, May 2015. https://doi.org/10.1038/nature14539.

[2] R. Caruana, S. Lawrence, and L. Giles, "Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping", Proc. of the 13th International Conference on Neural Information Processing Systems, Cambridge, United States, pp. 381-387, Jan. 2000.

[3] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets", Neural computation, Vol. 18, No. 7, Jul. 2006. https://doi.org/10.1162/neco.2006.18.7.1527.

[4] N. Srivastava, et al., "Dropout: A Simple Way to Prevent Neural Networks from Overfitting", Journal of Machine Learning Research, Vol. 15, No. 1, pp. 1929-1958, Jan. 2014.

[5] J. Salamon and J. P. Bello, "Deep convolutional neuralnetworks and dataaugmentation for environmental soundclassification", IEEE Signal Processing Letters, Vol. 24, No. 3, Mar. 2017. https://doi.org/10.1109/LSP.2017.2657381.

[6] G. Parascandolo, H. Huttunen, and T. Virtanen. "Recurrent neuralnetworks for polyphonic sound event detection inreal life recordings", 2016 IEEE InternationalConference on Acoustics,Speech and Signal Processing (ICASSP), Shanghai, China, Mar. 2016. https://doi.org/10.1109/ICASSP.2016.7472917.

[7] A. A. Khan, O. Chaudhari, and R. Chandra, "A review of ensemble learning and data augmentation models for class imbalanced problems: combination, implementation and

evaluation", Expert Systems with Applications, Vol. 244, Jun. 2024. hhttps://doi.org/10.1016/j.eswa.2023.122778.

[8] C. Shorten, T. M. Khoshgoftaar, and B. Furht, "Text Data Augmentation for Deep Learning", Journal of Big Data, Vol. 8, No. 101, Jul. 2021. https://doi.org/10.1186/s40537-021-00492-0.

[9] J. Ye et al., "A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models", arXiv:2303.10420, Mar. 2023. https://doi.org/10.48550/arXiv.2303.10420.

[10] J. Wei and K. Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks", arXiv: 1901.11196, Jan. 2019. https://doi.org/10.48550/arXiv.1901.11196.

[11] H. Cho and J. Moon, "A layered-wise data augmenting algorithm for small sampling data", Journal of Internet Computing and Services, Vol. 20, No. 6, pp. 65-72, Dec. 2019. https://doi.org/10.7472/jksii.2019.20.6.65.

[12] N. V. Chawla, et al., "Smote: Synthetic minority oversampling technique", Journal of Artificial Intelligence Research, Vol. 16, Jun. 2002. https://doi.org/10.1613/jair.953.

[13] S. Hu, Y. Liang, L. Ma, and Y. He, "MSMOTE: improving classification performance when training data is imbalanced", 2009 second international workshop on computer science and engineering, Qingdao, China, Vol. 2, Oct. 2009. https://doi.org/10.1109/WCSE.2009.756.

[14] W.-M. Lee and B.-W. On, "Generating Emotional Sentences Through Sentiment and Emotion Word Masking-based BERT and GPT Pipeline Method", The Journal of Korean Institute of Information Technology, Vol. 19, No. 9, pp. 29-40, Sep. 2021. https://doi.org/10.14801/jkiit.2021.19.9.29.

[15] D. Baidoo-Anu and L. O. Ansah, "Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning", SSRN, pp. 1-22, Jan. 2023. http://dx.doi.org/10.2139/ssrn.4337484.

[16] T. Sorensen et al., "An Information-theoretic Approach to Prompt Engineering Without Ground Truth Labels", arXiv:2203.11364, Mar. 2022. https://doi.org/10.48550/arXiv.2203.11364.

[17] S.-H. Lee and K.-S. Song, "Prompt engineering to improve the performance of teaching and learning materials Recommendation of Generative Artificial Intelligence", Journal of the Korea Society of Computer and Information, Vol. 28, No. 8, pp. 195-204, Aug. 2023. https://doi.org/10.9708/jksci.2023.28.08.195.

[18] AI Hub, "Emotional conversation corpus", https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=86 [accessed: Dec. 2022.]

[19] J. Wang and Y. Dong, "Measurement of text similarity: a survey", Information, Vol. 11, No. 9, pp. 421, Aug. 2020. https://doi.org/10.3390/info11090421.

저자소개

김 철 민 (Cheolmin Kim)

2017년 3월 ~ 현재 : 군산대학교 소프트웨어학과 학사과정
관심분야 : 인공지능, 프롬프트 엔지니어링

정 현 준 (Hyunjun Jung)

2008년 : 삼육대학교 컴퓨터과학과(학사)
2010년 : 숭실대학교 컴퓨터학과(공학석사)
2010년 : 고려대학교 컴퓨터·전파통신공학과(공학박사)
2017년 8월 ~ 2020년 8월 : 광주과학기술원 블록체인인터넷경제연구센터 연구원
2021년 ~ 현재 : 군산대학교 소프트웨어학과 교수
관심분야 : 블록체인, 데이터 사이언스, 센서 네트워크, 사물인터넷, 머신러닝