Detect AI-Generated Korean Text via KoBERT by Building Custom Dataset

한양대학교 2024년 2학기 AI+X:딥러닝 과목 프로젝트

기계공학부 2018013309 김승희
기계공학부 2018014002 유용준

1. Introduction

1.1 한국어 AI 생성 텍스트 탐지 기술의 중요성

인공지능의 발전은 글쓰기, 콘텐츠 생성, 질문 응답 등 다양한 분야에서 영향력을 넓혀가고 있다. 특히, 대형 언어 모델(Large Language Model, LLM)은 텍스트 생성에 있어 뛰어난 성능을 보여주며 주목받고 있다. 그러나 이러한 기술의 무분별한 사용은 표절, 허위 정보 생성, 가짜 뉴스, 스팸 확산과 같은 심각한 문제를 초래한다. 이에 따라 AI가 생성한 텍스트를 신뢰성 있게 감지할 수 있는 기술의 필요성이 점차 대두되고 있다.

해외에서는 이와 같은 AI 생성 텍스트 감지 기술의 중요성이 부각되면서 관련 기업의 설립 및 기술 경진 대회, 연구가 활발히 이루어지고 있다. 하지만 한국어와 관련해서는 아직 관련 기업이나 데이터셋이 전무하며, 연구 또한 초기 단계에 머물러 있는 상황이다.

이러한 배경에서, 한국어로 작성된 AI 생성 텍스트를 감지하기 위한 학습 데이터셋을 직접 구축하고, 이를 활용하여 한국어 AI 생성 텍스트 판별 모델 개발을 진행하고자 한다.

2. Related Work

2.1 AI-Generated Text Detection 서비스 기업

- **ZeroGPT**
  ZeroGPT 웹사이트
  ZeroGPT는 GPT-3, LLaMA, Google Bard와 같은 AI 모델이 생성한 콘텐츠를 감지하는 도구이다. 한국어도 지원하지만, 학습 데이터가 주로 영어 중심이라 한국어에 대한 정확도가 매우 낮다.

- **GPTZero**
  GPTZero 웹사이트
  GPTZero 역시 모든 언어를 지원한다고 하나, 한국어에 대한 정확도는 낮다는 피드백이 많다. 특히, 한글 콘텐츠가 AI 생성이라고 잘못 판단하는 경우(거짓 양성)가 자주 발생한다.

- **OpenAI - AI Text Classifier**
  OpenAI AI Classifier
  OpenAI는 AI 텍스트 분류기의 정확도가 낮다고 경고하고 있으며, 현재는 사용 불가 상태이다. 이 분류기는 주로 영어 텍스트에 사용되는 것이 좋으며, 9%의 거짓 양성(FP) 오류를 가질 수 있어 AI 생성 텍스트와 인간 텍스트 간의 차이를 정확하게 구별하는 것이 매우 어려운 일임을 알 수 있다.

2.2 AI-Generated Text Detection 논문

- **Can AI-Generated Text be Reliably Detected?**
  논문 링크

이 논문은 대형 언어 모델(LLMs)로 생성된 텍스트의 탐지가 패러프레이징(재구성)에 취약하며, 탐지 성능이 랜덤 분류기 수준으로 떨어질 수 있다는 것을 입증한다. 특히 AI 생성 텍스트의 탐지 한계는 특정 글쓰기 스타일이나 고의적 조작까지도 가능하게 만든다고 설명한다. 또한, 워터마킹이 적용된 모델도 공격에 취약하다는 점을 강조하고 있다.

- **Safeguarding Authenticity in Text with BERT-Powered Detection of AI-Generated Content**
  논문 링크
  이 논문은 BERT 모델을 활용하여 AI 생성 텍스트를 식별하는 방법을 제시한다. BERT의 문맥적 임베딩을 분석하여 AI와 인간 작성 텍스트를 구별할 수 있는 성능을 입증한다.

2.3 AI-Generated Text Detection 기술 경진대회

- **Kaggle: LLM - Detect AI Generated Text**
  대회 링크
  이 대회는 언어 모델로 생성된 텍스트를 탐지하는 경진 대회로, 영어 데이터셋을 활용해 진행되었다. 그러나 한국어와 같은 비영어권 언어에 대한 고려가 부족하여, 해당 데이터셋 및 대회 내용을 한국어 텍스트 탐지에 직접적으로 적용하기 어렵다.

이와 같이 기존의 AI 탐지 기술 및 연구는 주로 영어를 중심으로 이루어져 있으며, 한국어 데이터셋의 부재와 한국어 탐지 모델 부재라는 명확한 한계점이 존재한다. 이는 **한국어 AI 생성 텍스트 탐지를 위한 데이터셋 구축 및 모델 개발**이라는 본 연구 프로젝트의 필요성을 더욱 부각시킨다.

3. Custom Dataset Construction

- 한국어 AI Generated Text 분류 데이터셋이 존재하지 않기 때문에 직접 구축한다.
- 한국어로 이뤄진 Base Source 자연어 데이터셋을 구하고, 해당 데이터를 GPT-4o-mini 등의 LLM으로 재구성해서 AI Generated Text / Human Written Text 이진 분류 데이터셋을 구축한다.
- 3.1~3.3에 해당하는 코드는 `1_dataset_preprocess.ipynb`에서 확인할 수 있다.

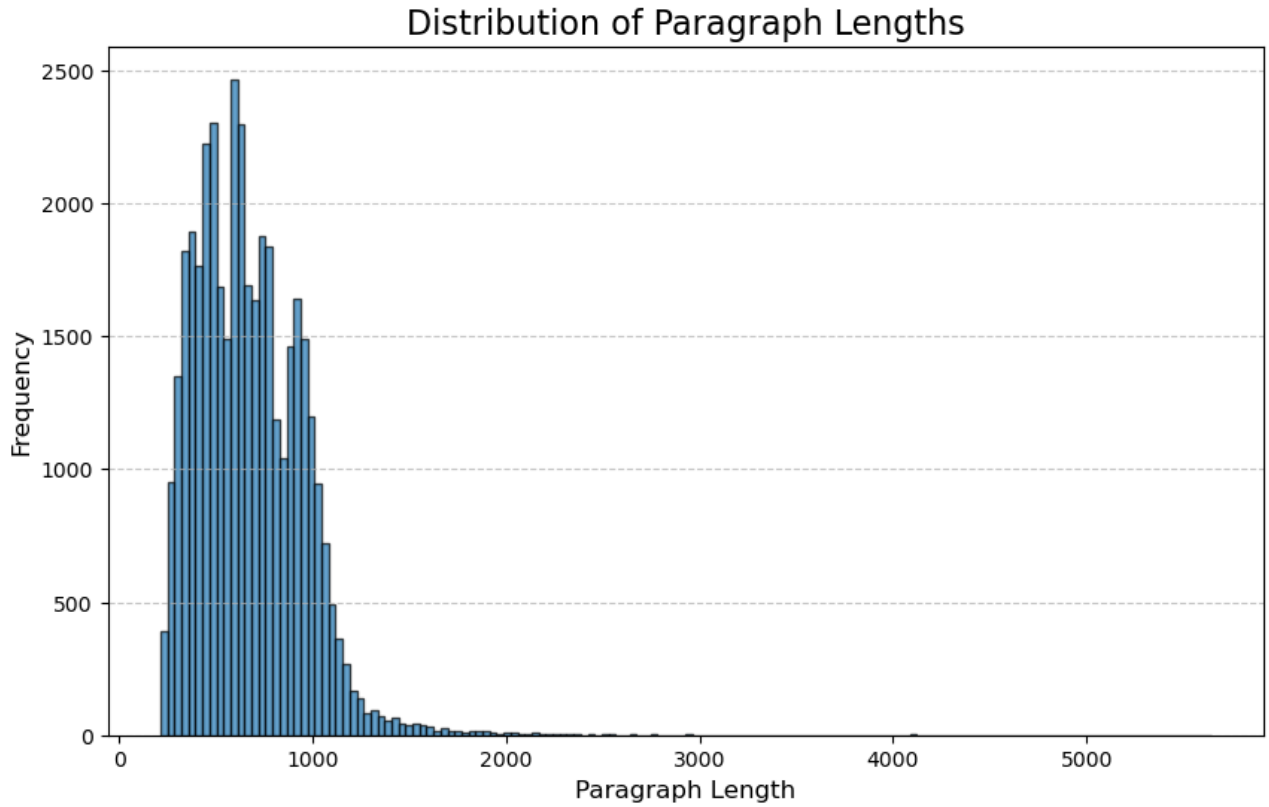
3.1 Base Source Data

- AI-Hub의 에세이 글 평가 데이터 를 Base Source Data로 활용
- 해당 데이터셋은 초등, 중등, 고등학생 등 다양한 학년군의 에세이 및 에세이 평가 점수로 구성된 데이터로, 에세이는 전문가의 자문을 통해 구성된 50가지 주제로 구성되어 있다.
- 'essay_type'으로 글짓기, 대안제시, 설명글, 주장, 찬성반대 등이 존재한다.

3.2 Base Source Dataset Preprocess

1. JSON to Pandas Dataframe: Raw JSON 파일에서 필요한 정보만 뽑아서 통합한 후, Pandas dataframe으로 변환하는 전처리 단계
2. 데이터셋 결측치 제거 전처리

3. paragraph_txt의 길이 계산한 후 분포 확인, 너무 짧거나 긴 경우 제거하는 전처리



4. Detect Placeholder: #@이름#, #@소속# 등의 태그를 추출하고, 해당 태그가 있는 경우에는 해당 데이터를 제거해주는 전처리
5. HTML 태그 제거 전처리
6. 'essay_prompt' 전처리: \n 및 공백 전처리

3.3 Split Train & Valid & Test Dataset

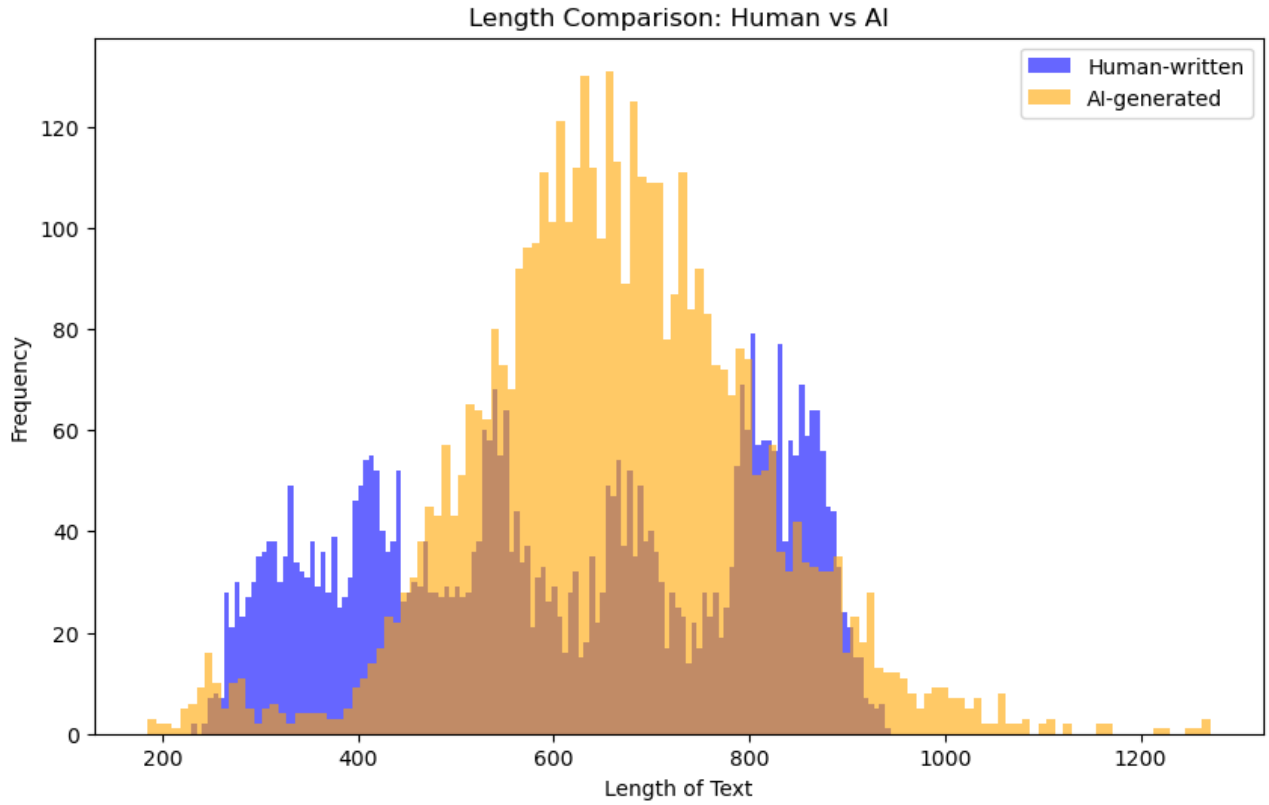
- 글짓기, 대안제시, 설명글, 주장, 찬성반대 5개의 에세이 타입이 존재하고, 초등, 중등, 고등학생으로 나뉘기 때문에 균등한 분포를 위해서 Train & Valid & Test 데이터셋에 대해 동일한 양을 설정한다.
- 이 데이터셋은 추후 Human-Written Text가 되고 Label이 0을 부여하며, 추후 AI-Generated Text를 생성할 때의 Input이 된다.
- Train Dataset: 4500개 (초,중,고 별로 300개씩)
  글짓기(900), 대안제시(900), 설명글(900), 주장(900), 찬성반대(900)
- Validation Dataset: 450개 (초,중,고 별로 30개씩)
  글짓기(90), 대안제시(90), 설명글(90), 주장(90), 찬성반대(90)
- Test Dataset: 450개 (초,중,고 별로 30개씩)
  글짓기(90), 대안제시(90), 설명글(90), 주장(90), 찬성반대(90)

3.4 Generate AI-Generated Text

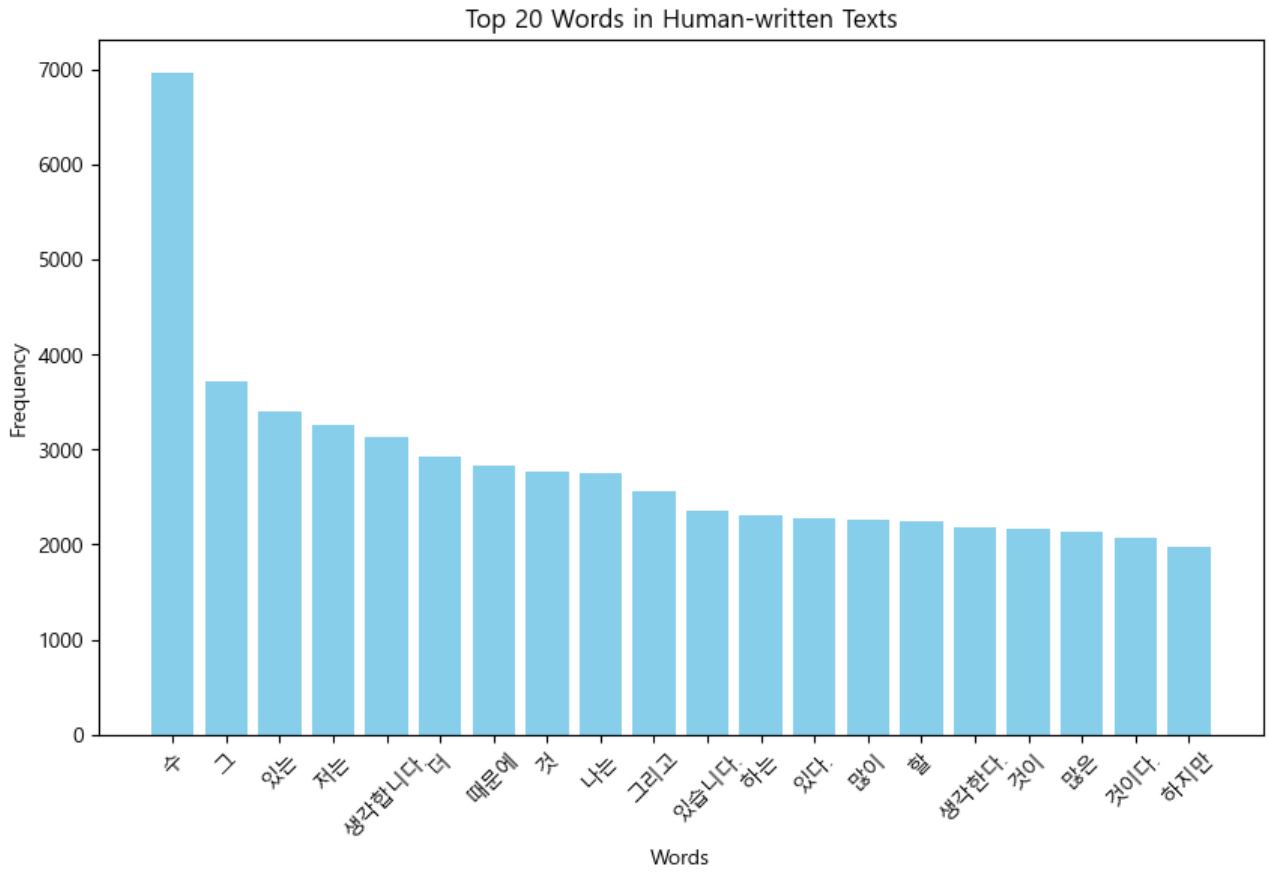
- OpenAI의 GPT-4o-mini 모델의 API를 활용하여 AI-Generated Text를 생성하고, Label 1을 부여한다.
- 3.4~3.5에 해당하는 코드는 `2_build_custom_dataset_AI_Generated_Text.ipynb`에서 확인할 수 있다.

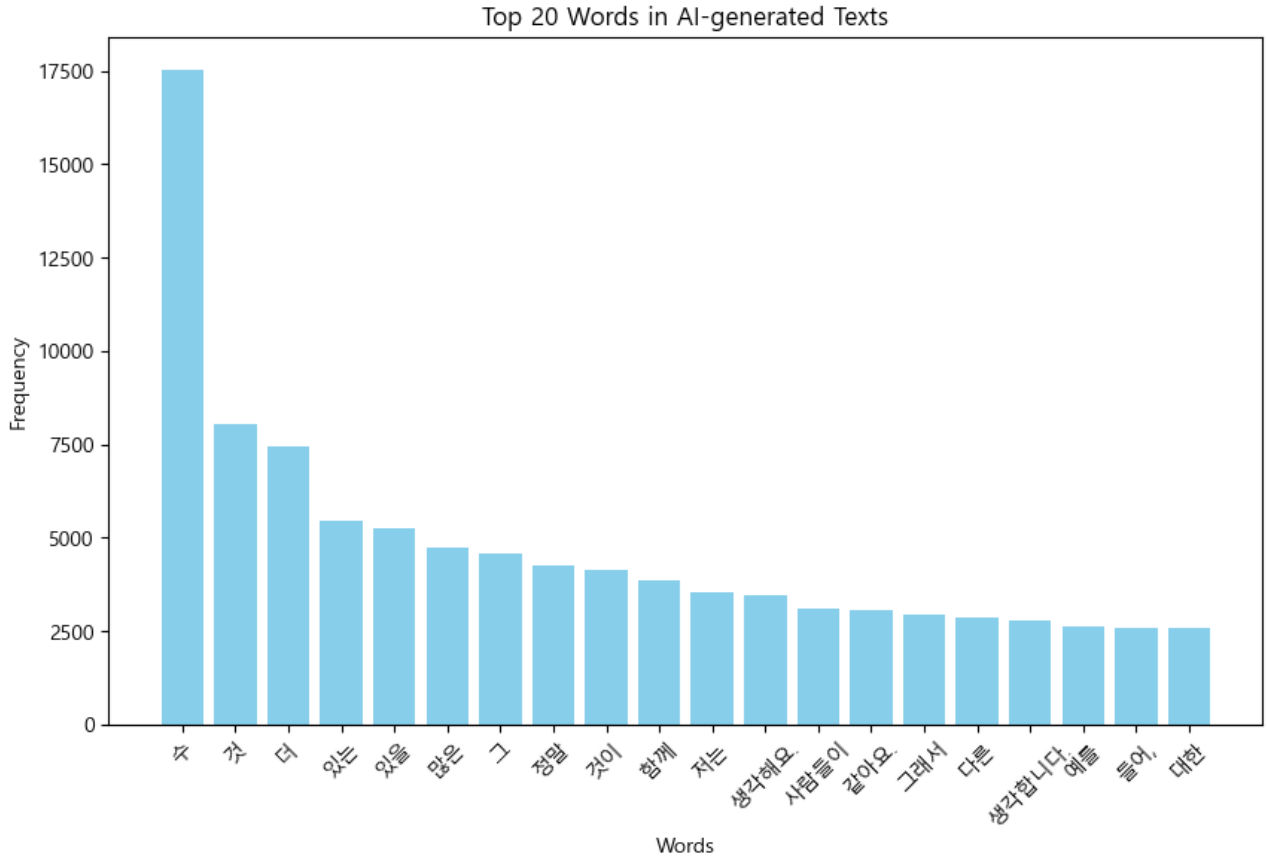
3.5 Human-Written Text vs AI-Generated Text Analysis

1. Length 비교

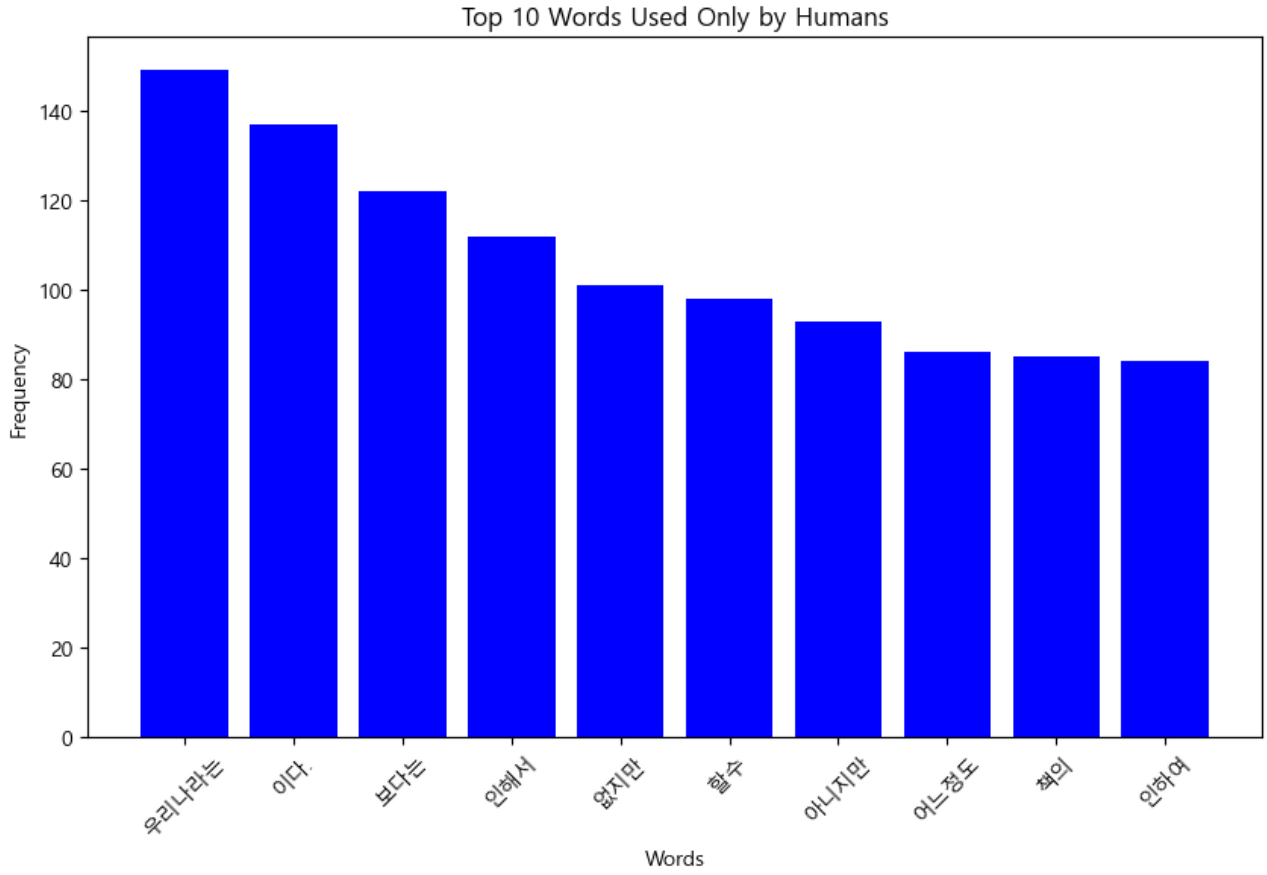


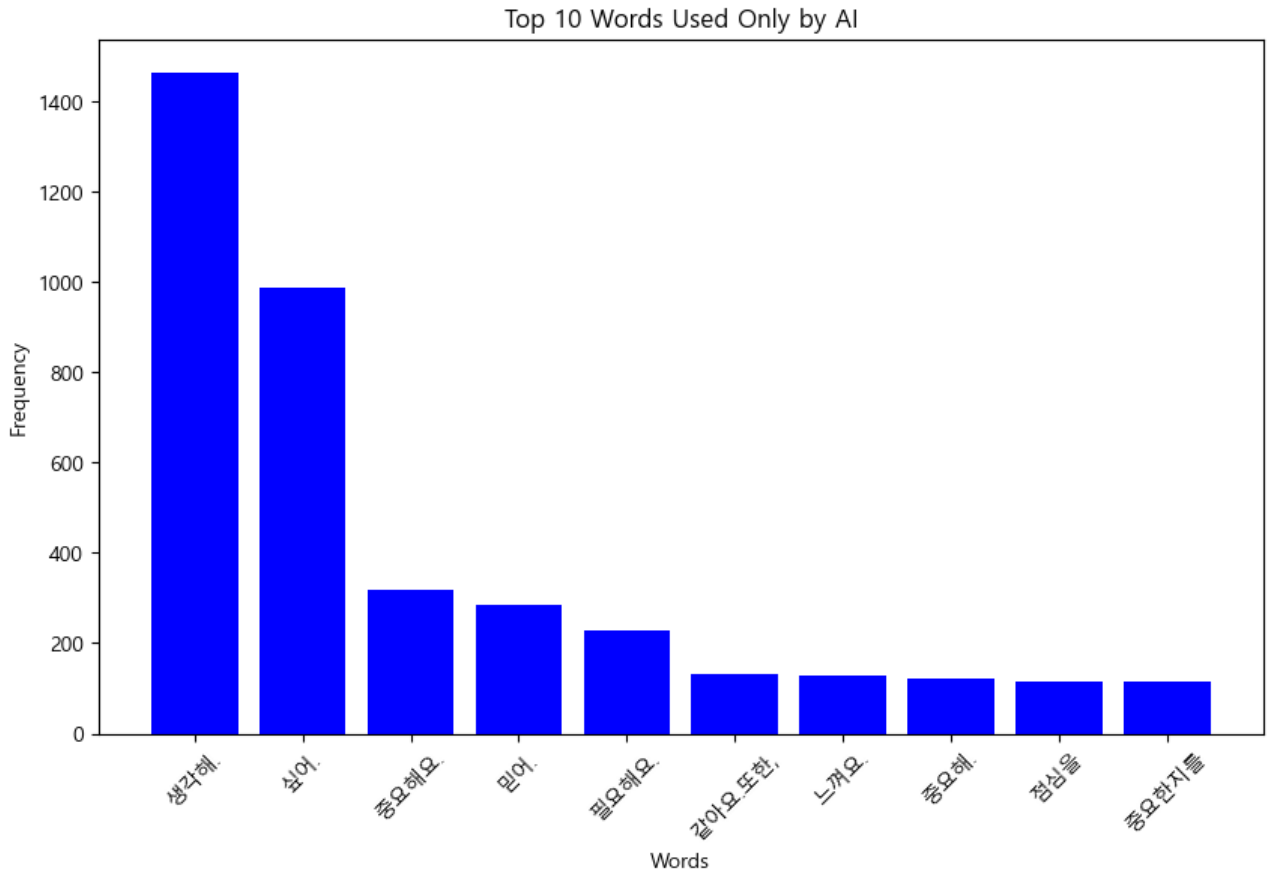
2. 사용 빈도가 높은 단어 Top 20 비교 (공백 기준으로 단어 분리)





3. 사람만 쓰는 단어 VS AI만 쓰는 단어 (공백 기준으로 단어 분리)





4. Methods

학습 방법 - 구축한 이진 분류 데이터셋으로 파인튜닝
모델 선정 - KoBERT

5. Experiments Analysis

실험 결과 분석 내용 작성

6. Conclusion

결론 내용 작성