**연 구 용 역 보 고 서**

**페이크 뉴스 자동 검출 응용 서비스의 인공지능 표준 시스템 분석**

**수탁기관: 동의대학교 산학협력단**

한 국 전 자 통 신 연 구 원

**연 구 용 역 보 고 서**

**페이크 뉴스 자동검출 응용 서비스의 인공지능 표준 시스템 분석**

**수탁기관: 동의대학교 산학협력단**

한 국 전 자 통 신 연 구 원

**제 출 문**

**한국전자통신연구원장 귀하**

본 보고서를 “페이크 뉴스 자동검출 응용 서비스의 인공지능 표준 시스템 분석”의 최종 연구보고서로 제출합니다.

2020년 11월 07일

수 탁 기 관 : 동의대학교 산학협력단

수 탁 기 관 장 : 이 임건 (인)

연 구 책 임 자 : 김 근형

참 여 연 구 원 : 한 윤진

박 예진

이 승찬

**요 약 문**

1. **제 목**

페이크 뉴스 자동검출 응용 서비스의 인공지능 표준 시스템 분석

1. **연구의 목적 및 중요성**

**1) 연구의 목적**

**2) 연구의 필요성**

1. **연구의 내용 및 범위**
2. **연구 결과**
3. **활용에 대한 건의**



**목차**

**그림 목차**

**표 목차**

# 제 1 장 개요

## 제 1 절 인공지능 서비스 배경 및 현황

1. **서비스 활용 사례**
2. **시장/제품 분석**

## 제 2 절 신뢰성 기술

1. **정의**
2. **중요성, 필요성, 파급효과**
3. **국내외 현황**

# 제 2 장 인공지능 신뢰성 분석

## 제 1 절 인공지능 서비스 현황

(문제점, 해결 방안)

## 제 2 절 표준화 현황

1. **국외 표준화 현황**
2. **국내 표준화 현황**

# 제 3 장 인공지능 신뢰성 프레임워크

1. **서비스 현황 사례**

(참조구조 및 기능, 공통 인터페이스)

1. **인공지능 신뢰성 요구 사항**

(참조구조 및 기능, 공통 인터페이스)

# 제 4 장 설명 가능한 인공지능 기반 서비스

## 제 1 절 설명 가능한 리뷰의 감성 탐지 서비스

1. **단어 임베딩**

일반적으로 컴퓨터는 자연어처리에서 단어의 의미와 상관없이 각각을 개별적인 기호로 취급한다. 컴퓨터가 자연어를 처리하도록 하기 위해선 자연어를 계산 가능한 형식으로 변형시켜야 한다. 문장에서 총 단어의 수만큼을 벡터의 차원으로 두고, 단어마다 해당하는 차원에 1로 표시하여 단어 벡터로 생성하는 것을 one-hot encoding이라고 한다. 각각의 단어 벡터들을 모두 더한 것을 해당 문장의 벡터라고 할 수 있는데, 단어의 의미를 담지 않은 단어 벡터들의 합이므로 문장의 의미 또한 표현하는데 어려움이 있다.

[http://doc.mindscale.kr/km/unstructured/11.html, 한국어 임베딩(책)]

이를 해결하기 위해 단어의 의미를 좀 더 세밀한 차원에 벡터로 표현하려 한 것이 단어 임베딩(word embedding)이다. 단어 임베딩은 one-hot encoding과 다르게 각각의 단어 벡터에 의미를 포함시키므로 비슷한 의미의 단어가 있는 경우 벡터 상에서 가까운 곳에 위치하게 된다. 그러므로 단어끼리 의미 비교를 위해 임베딩된 단어 벡터간에 유사도 계산이 가능하다. 유사도가 높으면 의미가 비슷하다는 것을 알 수 있다. 단어 임베딩의 종류에는 LSA, Word2Vec, GloVe, FastText 등이 있는데 이와 같은 단어 임베딩은 단어마다 벡터 값이 고정되어 있기 때문에 단어가 사용된 문맥은 고려하지 않는다. 즉 같은 단어이지만 다른 맥락에서 사용되었을 때도 같은 벡터 값으로 출력되기 때문에 차이를 알 수 없다.

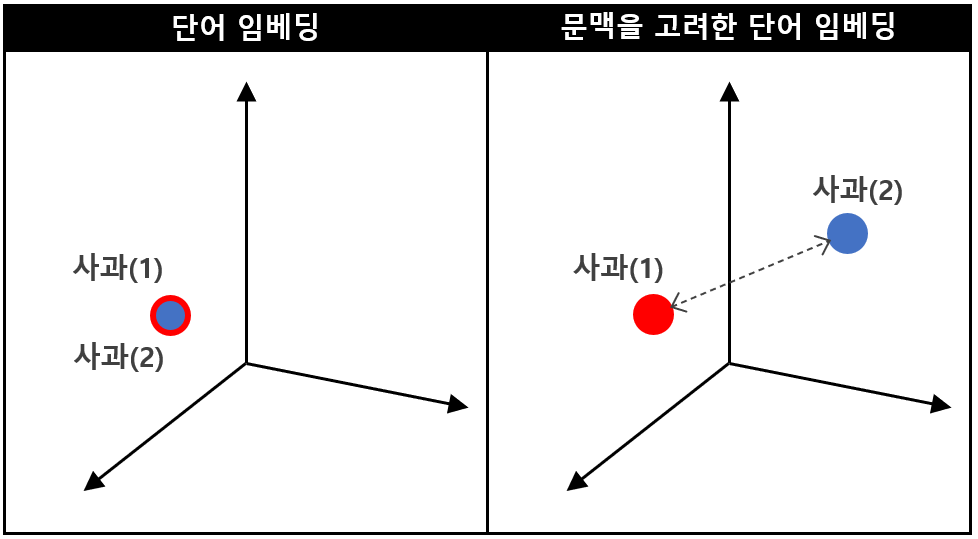


그림 1. 임베딩 방법 비교

단어가 사용된 문장의 맥락을 파악하기 위해 등장한 것이 문맥을 고려한 단어 임베딩(contextualized word embedding)이다. 문맥을 고려한 임베딩에선 개별 단어뿐만 아니라 문장 전체의 문맥적 의미를 함축하기 때문에 임베딩된 벡터를 인공지능을 통해 학습시켰을 때 기존의 단어 임베딩보다 학습 효과가 좋은 것으로 알려져있다. 그림 1은 단어 임베딩과 문맥을 반영한 단어 임베딩의 차이를 그림으로 나타낸 것이다. 단어 임베딩은 같은 문장에서 사용된 동일한 단어가 사용된 의미에 상관없이 같은 벡터를 출력하지만, 문맥을 고려한 임베딩의 경우 파악한 문맥에 맞게 각각의 벡터를 출력하게 된다. 문맥을 고려한 단어 임베딩 종류로는 BERT(Bidirectional Encoder Representtions from Transformer), ELMo(Embeddings from Language Models), GPT(Generative Pre-Training)등이 있다.

본 연구에서는 BERT가 사용되었다. BERT의 출력으로 얻을 수 있는 단어 벡터들은 문맥을 파악하여 생성된 벡터이기 때문에 단어의 의미뿐만 아니라 문장의 의미 또한 유지시킨다. 단어 벡터는 12개의 encoder layer를 거쳐 출력되는데, 마지막 레이어의 출력보다 마지막 4개 레이어 출력의 연결이 벡터의 표현을 더 잘 나타낼 수 있다. BERT는 출력 뒤에 원하는 분류 모델을 추가시켜 분류 학습을 하게 된다. 본 연구에선 설명가능한 인공지능을 위해 문장 벡터에 어떤 단어 벡터가 영향을 미쳤는지 해당 라벨로 분류되었는지 분석하였다.

1. **BERT(Bidirectional Encoder Representtions from Transformer)**

BERT는 인공지능 기반 자연어처리 분야에 등장한 구글의 강력한 언어모델이다. 기본적으로 Wikipedia 및 Book data와 같은 대용량 unlabeled data로 모델을 사전 학습(pre-training)하여 문맥을 파악할 수 있도록 구축된 모델이다. 사전학습된 모델 BERT에 분류하고자 하는 task에 맞게 labeled data로 추가 학습하여 분류 성능을 높인다. BERT는 사전학습에는 Masked Language Model(MLM)과 Next Sentence Prediction(NSP) 두 가지 방법이 사용된다. MLM은 랜덤한 위치의 토큰을 마스킹 시킨 뒤, 나머지 토큰들을 기반으로 마스킹된 토큰의 단어를 예측하고, NSP는 두 개의 문장을 입력하여 두 문장이 연속된 문장인지 예측하며 학습 성능을 높인다.

BERT Encoder는 MLM과 NSP를 위해 Encoder-Decoder 아키텍처로 이루어진 Transformer의 Transformer Encoder 구조를 기반으로 한다. Transformer는 RNN(Recurrent Neural Network)과 마찬가지로 시퀀스 데이터를 처리할 수 있도록 설계된 모델이다. 하지만 RNN과 다르게 Transformer는 시퀀스 데이터를 순차적으로 처리하지 않는다. 순차적으로 처리되는 RNN은 문장이 길어지는 경우 기울기 소실 문제(Vanishing gradient problem)로 인해 초기 토큰에 대한 정보가 사라질 수 있다는 단점이 있는 반면, Transformer는 병렬처리하여 초기 토큰의 정보를 보존할 수 있다. 또한 데이터 병렬처리 방법은 학습 시간을 단축시킨다.

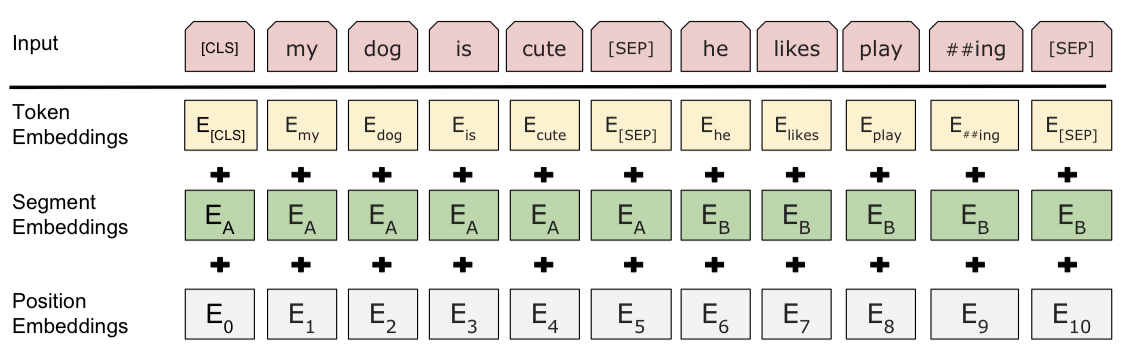


그림 2. BERT의 입력데이터 임베딩

BERT의 입력으로는 단어 토큰화된 토큰들이 사용된다. 그림 1의 Input에서 [CLS] 토큰은 BERT Encoder의 전체 layers를 거치고 난 후 token sequence의 결합된 의미를 가지게 되어 분류에 사용되는 토큰이며, [SEP] 토큰은 문장이 두 개 들어갈 때 문장을 구분하는 것에 사용되는 토큰이다. BERT Encoder에 입력하기 위해선 Embedding 과정을 거쳐야 하는데, Token embeddings과 position embeddings이 사용되는 Transformer Encoder과는 다르게, BERT Encoder에서는 토큰이 어느 문장에 속하는지 알 수 있도록 segment embeddings까지 더하여 입력 데이터를 생성한다.

BERT Encoder는 Multi-Head Attention이 사용되는데 이는 임베딩된 입력 데이터의 차원을 Head 수만큼 나누어 각각 Attention을 계산하고 추후에 합치는 과정이다. Attention 은 모든 토큰 간의 상관관계에 따라 가중치를 계산하는 메커니즘이다. 주어진 토큰 벡터 Query에 대해서 다른 토큰 Key와의 관계를 계산하여 가중치를 구하게 되는데 이를 Value에 적용시켜 출력값을 생성한다. 모든 토큰에 대한 attention 계산은 다음과 같이 최적화된 행렬 계산으로 표현될 수 있다.

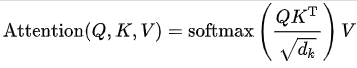


그림 3. Attention 계산 수식

설정한 layers 수만큼 BERT Encoder가 반복된 후 [CLS] 토큰은 token sequence의 결합된 의미를 가지게 되는데, 여기에 분류 모델을 추가되면 문장 분류 학습이 가능하게 된다. BERT는 다양한 task에 사용될 수 있으며, 입력 데이터와 분류 모델을 어떻게 Fine-tuning 하느냐에 따라 다른 유형으로 변형 가능하다.

표 1. BERT 클래스 종류

|  |  |
| --- | --- |
| **클래스** | **요약** |
| BertModel | 모든 모델의 틀  (사전 학습 완료) |
| BertForPreTraining | MLM과 NSP가 추가된 모델  (사전 학습 완료) |
| BertForMaskedLM | MLM가 추가된 모델  (사전 학습 완료) |
| BertForNextSentencePrediction | NSP가 추가된 모델  (사전 학습 완료) |
| BertForSequenceClassification | 문장 분류 모델이 추가된 모델  (사전 학습되었으나 분류 모델은 추가 훈련되어야 함) |
| BertForTokenClassification | 토큰 분류 모델이 추가된 모델  (사전 학습되었으나 분류 모델은 훈련되어야 함) |
| BertForQuestionAnswering | 질문에 맞는 토큰 분류 모델이 추가된 모델  (사전 학습되었으나 분류 모델은 훈련되어야 함) |

표 1은 BERT를 다양한 유형으로 사용할 수 있게 해주는 클래스 종류다. 모든 모델의 틀이 되는 BertModel을 기반으로 다양한 사전학습에 사용되는 MLM과 NSP 모델이 각각 추가된 클래스와 두 가지 모델 모두 추가된 클래스가 있다. 해당 클래스들은 모두 사전 학습되었으며 추가 학습시킬 필요가 없고, 이와 다르게 사전 학습이 되어있으나 추가 학습이 필요한 클래스들도 있다. BertForSequenceClassification 클래스는 문장을 label에 맞게 분류해주는 모델이 추가되어있기에, 분류 모델을 원하는 분류 유형에 맞는 데이터셋으로 학습시킬 필요가 있다. BertForTokenClassification 클래스 또한 문장의 토큰 각각을 label에 맞게 분류해주므로 추가 학습이 필요하다. 마지막으로 BertForQuestionAnswering 클래스는 Q&A를 위해 문장에서 정답을 찾아낼 수 있도록 추가 학습시켜야한다.

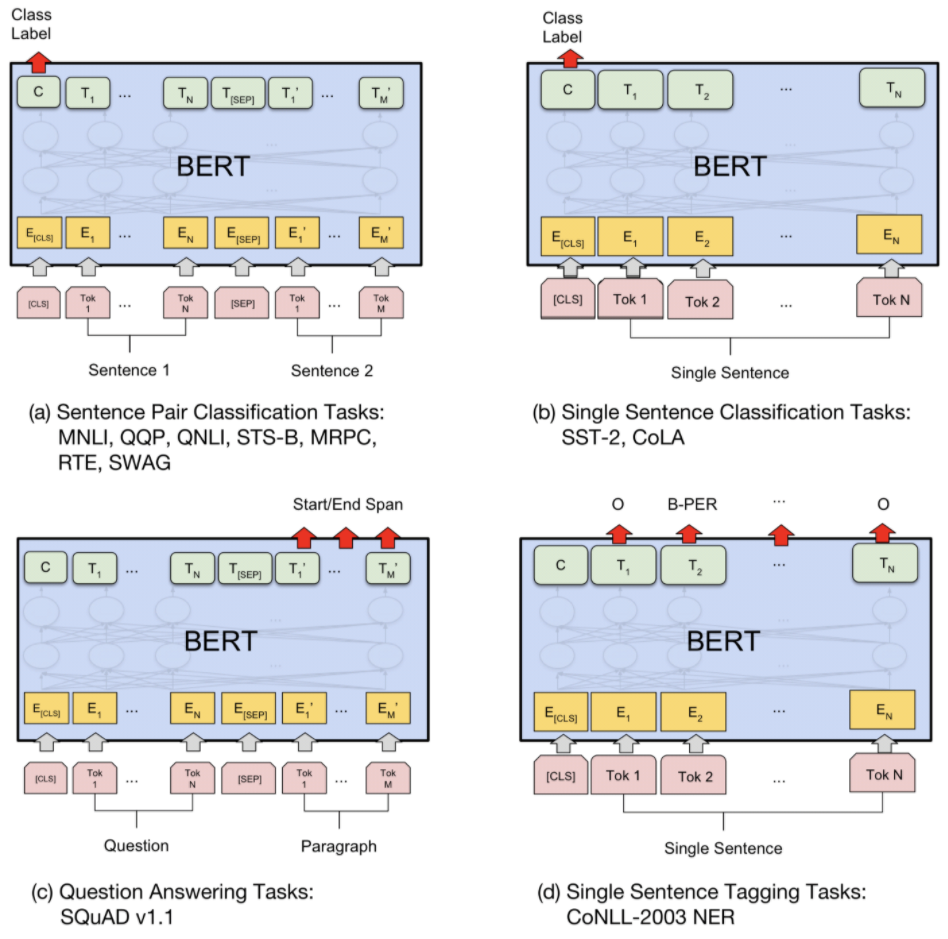


그림 4. BERT의 사용 유형

그림 3의 (a)와 (b)와 같은 경우는 BertForSequenceClassification 클래스를 생성하여 사용할 수 있는 유형으로, 분류하고 싶은 개수에 따라 분류 모델을 조정하고, 문장의 의미를 가지는 [CLS] 토큰의 벡터를 사용하여 class를 분류한다. (c)와 같은 Question Answering(QA) task 경우엔 BertForQuestionAnswering 클래스를 생성하여 사용할 수 있는 유형으로, Question와 Paragraph를 [SEP] 토큰으로 분류하고, Question에 정답을 Paragraph에서 찾아낸다. (d)와 같은 경우는 BertForTokenClassification 클래스를 생성하여 사용할 수 있는 유형으로, Named Entity Recognition(NER)이나 형태소 분석처럼 각 토큰이 어떤 class를 가지는지 각각을 분류할 수 있다.

BERT는 위의 내용처럼 downstream task 학습을 통해 사전학습된 parameter들을 변경하며 진행되는 Fine-tuning 접근법뿐만 아니라 Feature-based 접근법으로도 사용 가능하다. Feature-based 접근법은 BERT를 통해 진행된 Feature extraction을 특정 분류를 위한 추가 모델에 입력하여 학습시키는 형태이다. Feature-based 접근법은 다른 모델인 ELMo(Embeddings from Language Model)에서도 사용되는 접근법으로, 사용하는 이유는 Fine-tuning 접근법을 사용했을 때 사용되는 pooler는 분류 모델에 들어가기 전에 [CLS] 토큰을 처리하는 과정인데, pooler의 출력 값은 분류에는 도움을 주지만 문장의 의미를 표현하기엔 좋은 값이 아니기 때문이다.

* (This output is usually not a good summary of the semantic content of the input, you’re often better with averaging or pooling the sequence of hidden-states for the whole input sequence.) [출처]

BERT 논문에 따르면 Feature-based 접근법에는 다양한 방법들이 존재하는데 이 방법들은 성능에도 영향을 미친다. 여러 방법들을 간단하게 설명하자면 첫 번째로BERT Encoder의 입력으로 사용되는 토큰을 사용하는 방법과 BERT Encoder의 출력인 토큰을 사용하는 방법이 있다. BERT Encoder의 layer 중 마지막에서 두 번째 layer의 토큰을 사용할 수도 있으며 모든 layer를 합해서도 사용 가능하다. 가장 많이 사용되는 방법 두 가지로는 BERT Encoder의 마지막 layer 4개를 합하는 방법과 layer 4개를 Concat하는 방법이 있는데 Concat하는 방법은 Fine-tuning 접근법의 성능과의 차이가 크게 나지 않는다. 이를 보아, BERT는 Feature-based에서도 효과적이라고 볼 수 있다.

1. **BERT 세부 기능 분석**

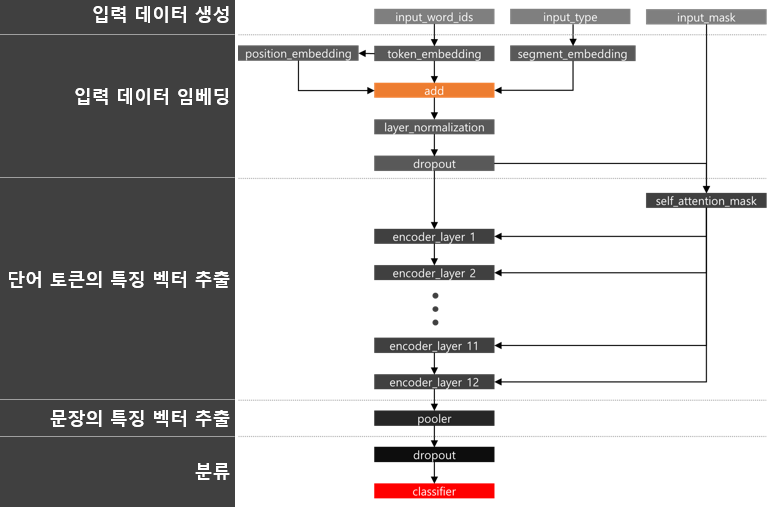


그림 5. BERT의 블록 다이어그램

그림 4는 BertForSequenceClassification 클래스를 생성하였을 때 전체 모델의 구조를 블록 다이어그램으로 표현한 것이다. 마지막 단계에 분류 모델을 추가하여 문장 분류가 이루어진다. BERT의 입력으로는 세 가지 데이터가 준비되어야 한다. 첫 번째로 입력 문장을 토큰화 한 뒤, 토큰을 index 값으로 변환시킨 input\_word\_ids 데이터가 필요하다. 두 번째로는 여러 문장이 입력으로 들어올 때 토큰이 어느 문장에 소속되어 있는지 알 수 있도록 input\_type 데이터가 있어야 하며, 마지막으로 토큰과 토큰이 아닌 패딩 위치를 알 수 있도록 해주는 input\_mask 데이터가 필요하다.

입력 받은 데이터들은 임베딩 과정을 거치게 되는데 input\_word\_ids 데이터로 token\_embedding 벡터와 토큰의 순서를 알 수 있는 position\_embedding 벡터를 얻게 된다. input\_type 데이터로 소속 문장을 알 수 있도록 해주는 segment\_embedding 벡터를 얻게 되며, input\_mask는 다음 단계인 encoder 단계에서 토큰과 패딩을 구분할 수 있도록 활용된다. 임베딩 과정을 통해 얻은 세 가지 벡터를 합하여 임베딩된 입력 데이터를 얻을 수 있다.

임베딩된 입력 데이터는 토큰의 특징 벡터를 추출할 수 있는 encoder로 입력된다. 이 과정에서 데이터는 Multi-Head Attention을 통해 각 단어 사이의 관계성에 맞게 가중치가 계산된다. 설정된 여러 layer를 거치게 되면서 입력된 문장의 문맥을 파악하게 되고 그에 맞는 특징 벡터를 출력한다.

encoder 단계에서 출력된 토큰의 특징 벡터들 중에서 첫 번째 토큰인 [CLS] 토큰은 다음 단계인 pooler의 입력으로 사용된다. pooler에서는 분류 모델이 분류를 더 효율적으로 진행할 수 있도록 문장 벡터인 [CLS] 토큰을 처리하는 과정이다. 분류 모델은 pooler의 결과인 문장 벡터를 입력으로 받아 결과에 맞는 label을 학습하게 된다.

1. **실험**

표 2. 리뷰 감정 탐지 서비스 실험 환경

|  |  |
| --- | --- |
| **프로그램** | Colab |
| **런타임** | GPU |
| **모델** | BertForSequenceClassification |
| **토크나이저** | BertTokenizer |
| **데이터셋** | Naver sentiment movie corpus |
| **데이터 수** | 학습 데이터: 150,000 |
| 테스트 데이터: 50,000 |
| **라벨** | 긍정: 100,000 |
| 부정: 100,000 |
| **테스트 정확도** | 87% |

실험을 위해 BERT 모델에 긍정 및 부정으로 이루어진 영화 리뷰 데이터를 학습시켰으며, 각 라벨로 분류될 때 어떤 단어의 영향이 컸는지 테스트했다. 임베딩된 벡터는 사칙 연산을 통해 단어와 문장 사이의 의미적 관계를 도출할 수 있다. 이를 기반으로 우리는 문장 벡터와 단어 벡터들 사이의 유사도를 측정하였을 때, 유사도가 높은 벡터가 문장에 영향을 크게 미쳤다고 판단했다. 표 3은 해당 테스트를 위한 실험 환경이다.

실험에서는 Transformers가 제공하는 클래스 중 BertForSequenceClassification를 사용했다. BertForSequenceClassification는 데이터 분류를 위해 단일 Linear layer가 추가된 BERT 모델이다. 클래스 생성 시 사전 학습된 BERT 모델과 추가된 분류 레이어가 입력 데이터에 대하여 특정 분류 작업을 학습한다. 사전 학습된 BERT 모델 중에서도 다양한 언어가 지원되는 bert-base-multilingual-cased 모델이 사용되었는데, 해당 모델은 Wikipedia 지분율이 높은 상위 104개의 언어로 학습되었기에 영어 이외의 언어로도 사용 가능하다. 사용 가능한 언어 목록에 한국어도 포함되어있어 해당 모델을 선택하였다.



그림6. BertTokenizer의 한국어 토큰화

생성된 모델에 데이터를 입력하기 전에 그림 5와 같이 문장을 토큰으로 분리하는 토큰화 과정이 필요하다. 토큰화에 사용될 토크나이저도 모델과 마찬가지로 Transformers의 사전 학습된 토크나이저인 BertTokenizer가 사용되었다. BertTokenizer는 WordPiece를 기반의 토크나이저로, WordPiece란 서브워드 토큰화(Subword Tokenization) 기반의 토크나이즈 알고리즘이다. 이는 자주 등장하여 이미 학습되어 있는 단어는 두고, 자주 등장하지 않는 단어를 더 작은 단위로 분리하여 의미 있는 단어로 토큰화하는 방법이다. 예를 들어 'ab'라는 단어가 있을 때 'ab'가 등장할 확률보다 'a'와 'b'로 등장할 확률이 높다면 더 작게 분리하는 것이다. 영어에 맞는 형태소 분석이 사용되지 않고, 분리된 토큰 사이에 '##' 기호로 연결을 표시하여 한국어도 처리가 가능하다. BertTokenizer를 사용하여 토큰화된 토큰들은 index 값으로 변환하여 모델에 입력으로 들어가게 된다.

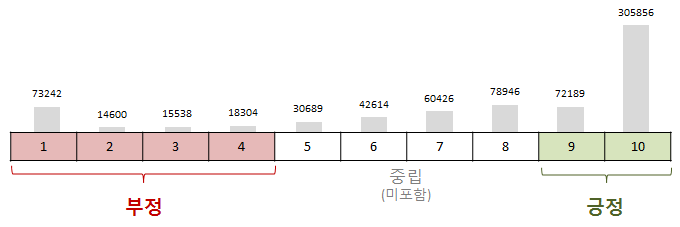


그림 7. 네이버 영화 리뷰 데이터 수집

실험 데이터로는 한국어 데이터 텍스트 분류 테스트에 많이 사용되는 네이버 영화 리뷰 데이터 nsmc(Naver sentiment movie corpus)가 사용되었다. 학습 데이터 150,000개와 테스트 데이터 50,000개를 합쳐 총 200,000개로 구성된 데이터로써 문장이 긍정일 때 label이 1, 부정일 때 label이 0으로 표시되어 있다. 네이버 영화의 리뷰 중 영화당 100개의 리뷰를 수집하였으며, 그림과 같이 1~10점까지의 평점 중에서 중립적인 평점인 5~8점 리뷰는 제외하고 1~4점은 부정으로, 9~10점은 긍정으로 포함시켰다. 긍정, 부정 label는 동일한 비율로 각 100,000개 수집되었다.

모델과 토크나이저, 데이터셋이 준비되었으면 첫 번째로 데이터셋을 BertTokenizer을 사용하여 토큰화하는 것이 필요하다. 토큰화된 토큰들은 숫자로 된 index로 변환하여 input\_word\_ids 데이터를 생성하였다. BERT 모델은 입력 시퀀스의 길이 제한이 있기 때문에 최대 시퀀스 길이에 맞춰 길이를 토큰을 잘라내거나 모자란 부분을 0으로 패딩 처리하여야 한다. 이때 문장에서 토큰이 존재하는 부분과 의미없는 부분을 알려주는 input\_mask 데이터를 생성하였다. 학습 시 input\_mask 데이터가 입력되면 BERT 모델에서 패딩 부분은 학습하지 않아 속도를 향상시킬 수 있다. 전처리된 학습 데이터를 분리하여 10%는 검증 데이터로 활용되었다.

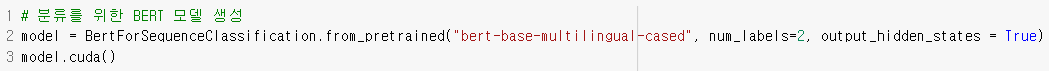


그림 8. 모델 생성

모델을 생성할 땐 그림 8과 같이 output\_hidden\_states를 True로 설정해야 BERT Encoder의 계산 값을 출력할 수 있다. optimizer로는 AdamW가 사용되었으며 epochs를 4로 설정하여 반복하였다. 생성된 모델에 전처리된 학습 데이터를 사용하여 학습시킨 후, 테스트 데이터도 위와 같은 전처리를 거쳐 모델을 테스트한 결과로 87%의 정확도를 얻을 수 있었다.

우리는 BERT Encoder를 거쳐 출력된 토큰 벡터와 문장 벡터를 보고 어떤 토큰이 문장에 큰 영향을 주었길래 분류 모델이 해당 label로 분류를 했는지 분석하고자 하였다. 첫 번째로 알 수 있었던 결과는 그림 8과 같이 BERT Encoder의 마지막 layer에서 출력된 토큰 벡터들을 히스토그램으로 그려보면 유사한 패턴을 가진다는 것이다.

뿐만 아니라 여러 문장으로 테스트했을 때도 같은 label로 분류되는 문장들 사이에 유사한 패턴이 존재한다는 것을 알 수 있었다. 그림 9의 첫 번째 히스토그램은 label이 부정으로 분류된 문장들의 문장 벡터들을 나타낸 것이며, 두 번째 히스토그램은 label이 긍정으로 분류된 문장들의 문장 벡터들을 나타낸 것이다. 각 10개(더 추가)의 문장으로 테스트하였으며, 문장들의 백터 값을 비교했을 때 유사도가 크게 높진 않았지만 각 분류 label에 따라 다른 패턴을 가진다는 것을 알 수 있었다. 각 label에 맞는 패턴이 있다는 근거를 기반으로 label 패턴에 가깝게 생성된 문장 벡터와 유사도가 높은 토큰 벡터가 label 분류에 영향을 미쳤다고 판단했다.

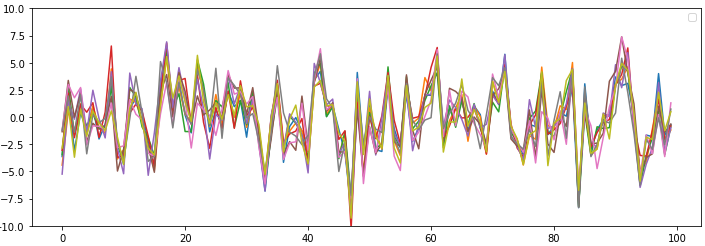


그림 9. 토큰 벡터 히스토그램



그림 11. 코사인 유사도 계산 코드

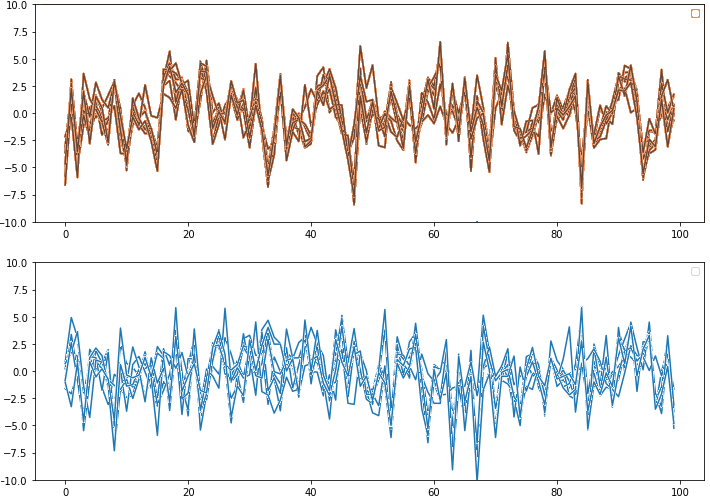


그림 10. 부정 및 긍정 문장 벡터 히스토그램

코사인 유사도를 사용하여 문장 벡터와 토큰 벡터 두 벡터 간의 유사도를 계산했다. BERT의 출력 벡터는 768차원을 가지게 된다. 두 벡터가 완전히 동일한 경우에는 유사도 결과는 1의 값을 가지며 두 벡터가 관련성이 없을수록 0에 가까워진다. 문장 벡터는 토큰 벡터의 합으로 이루어진 벡터이기 때문에 768차원을 구성하고 있는 정보들은 같다. 문장 벡터와 토큰 벡터들의 코사인 유사도를 계산하여 문장의 특징 벡터와 가장 유사한 토큰, 즉 단어는 어떤 것들이 있는지 분석했다.

그림은 코사인 유사도 테스트한 여러 문장 중 하나인 ‘재미없다 지루하고. 같은 음식 영화인데도 바베트의 만찬하고 넘 차이남....’ 이라는 문장의 결과를 히스토그램으로 나타낸 것이다. 각 토큰들과 코사인 유사도 계산을 한 [CLS] 토큰은 문장 벡터를 의미하며, 숫자 값은 유사도 결과를 나타낸 것이다. 해당 문장 벡터는 분류 모델을 거쳤을 때 부정으로 분류된 문장이다. 문장 벡터의 패턴은 그림 10의 첫 번째 히스토그램인 부정 패턴과 유사하게 출력된다. 부정 패턴에 가까운 문장 벡터와 각 토큰들의 코사인 유사도를 계산했을 때 결과를 보면, 토큰들 중에서도 유사도가 높은 부분이 존재한다는 것을 알 수 있다. 히스토그램을 네 등분하여 비교했을 때 첫 번째 부분과 나머지의 결과 값이 차이가 나는 것을 볼 수 있다. 첫 번째 부분의 토큰으로는 ‘재’, ‘##미’, ‘##없’, ‘##다’, ‘지’, ‘##루’, ‘##하고’, ‘.’ 등으로 부정의 의미를 가지는 단어들로 구성되어 있다. 나머지 부분의 내용을 보면 감정을 나타내기보단 타영화와 비교하는 내용으로써, 긍정인지 부정인지를 알 수 있는 감정적인 단어가 사용되지 않았음을 알 수 있다.

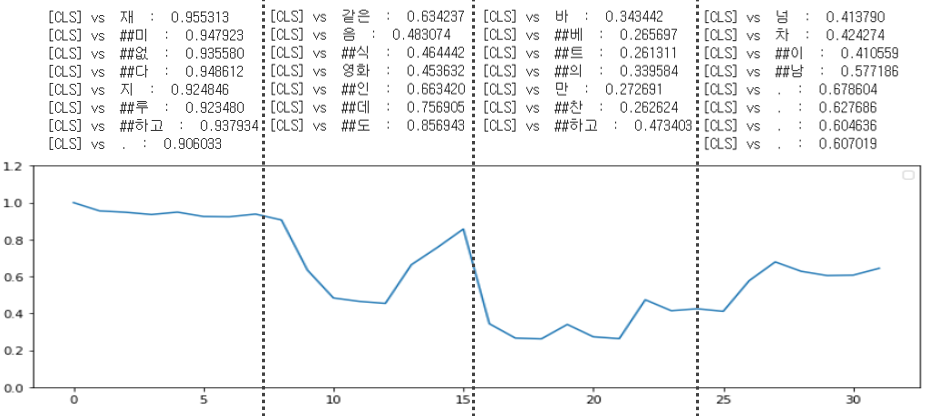


그림 12. 문장 벡터와 토큰 벡터의 코사인 유사도

우리는 50개의 문장으로 위와 같은 테스트를 반복하여 문장 벡터와 토큰 벡터의 코사인 유사도를 수동으로 확인했다. 히스토그램으로 강조되는 부분의 토큰이 감정을 나타내는 단어가 맞는지 확인하였으며, 긍정 및 부정으로 분류될 때 각 label의 문장에서 어떤 단어들이 강조되고 있는지 정리하였다. 50개의 문장을 테스트하였을 때 분류에 맞는 단어가 강조되고 있다고 판단되었다.

더 많은 문장에서 강조되는 단어가 어떤 것들이 있는지 분석하기 위해 for문을 사용하여 코사인 유사도가 높은 부분의 키워드를 출력하였다. 1000개의 문장에서 강조되는 단어를 출력하였다. BERT의 토큰화는 형태소 단위로 토큰화되지 않기 때문에 감정을 나타내는 명사를 정확하게 찾아낼 수는 없었으나 긍정으로 분류된 문장에서는 ‘좋다’, ‘깔끔해서’, ‘짱이다’, ‘최고’, ‘감동’ 등의 긍정적인 부분이 코사인 유사도가 높게 계산되었으며, 부정으로 분류된 문장에서는 ‘지루하다’, ‘재미없다’, ‘아깝다’, ‘실망’, ‘유치하다’ 등의 부정적인 부분이 코사인 유사도가 높게 계산되었다. 다음 표는 1000개의 문장에서 강조된 단어들의 통계를 나타낸 것이다. (통계내서 표 추가)

[https://drive.google.com/drive/u/1/folders/1\_VUVpnYyLYGB5fmjCA95E461HHkuicy0]

## 제 2 절 설명 가능한 가짜 뉴스 탐지 서비스

1. **가짜뉴스**

지난 2016년 미국 대통령 선거에서 가짜뉴스(FakeNews) 문제가 많은 영향을 미치면서 전 세계적으로 사회적 이슈가 되었다. 가짜뉴스가 사회적 이슈로 대두된 만큼 해당 개념에 대한 정의가 명화하지 않은 상황이다. 가짜 선거뉴스 현황과 각국의 대응 사례에 따르면 각국의 가짜뉴스 정의와 범위가 다른 것으로 알려져있다. 독일의 경우 소셜 네트워크 서비스를 통해 정치 목적, 경제적 이득을 위해 사실이 아닌 글을 유통하는 것을 가짜뉴스의 개념으로 정의하며, 게시물이 법률에 위반하는지의 여부를 판단기준으로 한다. 미국의 정치적 차원에서 가짜뉴스란 일반적으로 정치 선동의 일종으로 정치적, 경제적 이득을 취하려는 목적의 잘못된 정보를 포함하는 모든 형태의 기사나 영상을 의미한다. 가짜뉴스는 미국 대통령 선거와 같은 정치뿐만 아니라, 의료, 엔터테인먼트, 예술, 문화 등 다양한 주제의 뉴스에서도 찾아볼 수 있다. 일반적으로 뉴스가 사실에 근거하지 않거나, 사실의 일부를 왜곡하거나, 논쟁적인 이슈에 대한 편향적인 주장을 가지고 있는지의 여부를 판단기준으로 한다. 영국에서의 가짜뉴스는 대중의 관심을 확보하고 그것으로 광고수익을 창출하기 위해 조작된 뉴스로 정의할 수 있다. 실제뉴스가 편향적인 주장을 가지더라도 가짜뉴스로 구분하지 않으며, 선동 및 거짓을 포함하고 있는지의 여부를 판단기준으로 한다. 이렇듯 각국마다 가짜뉴스의 다양한 개념을 정의하고 있지만 해당 정의도 명확하게 정리된 것이 아니며, 한국도 각국과 마찬가지로 가짜뉴스의 개념 정의가 확실하지 않고 가짜뉴스의 확산에 대한 대응 방안이 체계적으로 마련되지 않은 상황이다. [가짜 선거뉴스 현황과 각국의 대응 사례]

가짜뉴스가 사회적 이슈로 대두되면서 다양한 연구들과 함께 가짜뉴스 탐지 방법이 제시되고 있다. 가짜뉴스 탐지를 위한 기술적 방법으로는 인공지능 기반 가짜뉴스 탐지 방법과 시맨틱 기반 가짜뉴스 탐지 방법 등이 있다. 인공지능 기반 가짜뉴스 탐지 방법은 가짜뉴스에 사용된 언어와 구문을 기계에 학습시켜 판단할 뉴스의 가짜뉴스 확률을 추정하는 방법이다. 시맨틱 기반 가짜뉴스 탐지 방법은 뉴스에 사용된 단어, 어절, 문장, 맥락 등을 분석하여 내용의 사실성을 검증하는 방법이다. 문장의 형태소와 구문, 의미 등을 추출하여 지식 베이스를 구축하게 되고 구축된 지식 베이스에 따라 뉴스의 사실 여부를 판단한다. 가짜뉴스 탐지 방법은 다양한 방법을 혼합하여 발전해나가고 있으며, 인공지능 기반 가짜뉴스 탐지 방법과 다른 방법의 하이브리드 방법이 확산될 것이라 예상된다. [페이크 뉴스 탐지 기술 동향과 시사점]

표 3. 가짜뉴스 탐지 논문 조사

|  |  |  |  |
| --- | --- | --- | --- |
| **논문** | **가짜뉴스 정의** | **데이터셋** | **네트워크** |
| A Deep Ensemble Framework for Fake News Detection and Classification | 주제와 맞지 않는 내용이 본문에 존재하며, 주장에 근거가 없는 뉴스 | LIAR | CNN,  Bi-LSTM |
| CSI: A Hybrid Deep Model for Fake News Detection | 주제와 맞지 않는 내용이 본문에 존재하며, 신뢰되지 않는 출처로부터 유포되는 뉴스 | Twitter,  Weibo | LSTM |
| DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning | 다른 기사와 상반되는 주장이 사람들이 오해할 수 있는 선동적인 어휘로 작성된 뉴스 | Snopes,  PolitiFact,  NewsTrust,  SEMEVAL-2017 TASK 8 | GloVe,  Bi-LSTM |
| Deep Diffusive Neural Network based Fake News Detection from Heterogeneous Social Networks | 주제와 맞지 않는 내용이 본문에 존재하며, 신뢰되지 않는 출처로부터 유포되는 뉴스 | PolitiFact | HFLU,  GDU |
| Detecting Incongruity between News Headline and Body Text via a Deep Hierarchical Encoder | 기사의 제목과 상관없는 내용이 본문에 작성되어 독자가 원하지 않은 정보를 제공하는 뉴스 | 수집한  한국어 기사를 가공하여 생성한 데이터 | GRU |
| EMET: EMBEDDINGS FROM MULTILINGUAL-ENCODER TRANSFORMER FOR FAKE NEWS DETECTION | 소셜 미디어 네트워크를 통해 유포되는 게시물 중 다른 뉴스의 일반적인 주장들과 상반되는 정보를 포함한 게시물 | Twitter | CNN |
| FNDNet – A deep convolutional neural network for fake news detection | 주제와 맞지 않는 문장이 있고, 근거가 없는 주장을 보이는 내용 | Kaggle | GloVe,  CNN |
| Fake News Detection Using A Deep Neural Network | 가짜뉴스에서 많이 사용되는 선동적인 어휘로 작성된 뉴스 | Kaggle | CNN |
| A Deep Learning Approach for Automatic Detection of Fake News | 문장에서 주제 및 다른 문장들과 관계가 없고, 근거가 없는 선동적인 주장을 보이는 뉴스 | FakeNews AMT,  Celebrity news | ELMo,  GRU,  LSTM |
| Deep Two-path Semi-supervised Learning for Fake News Detection | 제한된 단어들을 사용하며, 문법적인 실수 등 문맥상 공통적인 특징들을 가짐 | PHEME | CNN |
| exBAKE: Automatic Fake News Detection Model Based on Bidirectional Encoder Representations from Transformers (BERT) | 기사의 제목과 상관없는 내용이 본문에 작성되어 독자가 원하지 않은 정보를 제공하는 뉴스 | FNC-1 | BERT |

본 연구에서는 인공지능 기반 가짜뉴스 탐지 방법으로 가짜뉴스를 탐지하기 위하여, 먼저 인공지능 기반의 다양한 탐지 기술에서 정의하고 있는 가짜뉴스 개념을 정리하였다. 가짜뉴스 정의에 따라 어떤 인공지능 구조로 구축하였는지 알아보기 위해, 딥러닝 네트워크를 사용하여 탐지한 논문을 중점으로 조사하였다. Arjun Roy [A Deep Ensemble Framework for Fake News Detection and Classification]의 논문에 따르면 주제와 맞지 않는 내용이 본문에 존재하며, 주장에 근거가 없는 뉴스를 가짜뉴스로 정의한다고 하였다. 주제를 찾기 위한 CNN(Convolutional Neural Network)과 문맥을 탐색하기 위한 Bi-LSTM(Bidirectional Long Short-Term Memory Network)이 사용되었다. Natali Ruchansky []의 논문도 마찬가지로 주제와 맞지 않는 내용이 본문에 존재하는 뉴스를 가짜뉴스로 정의했으며, 추가로 신뢰되지 않는 출처로부터 유포되는 뉴스까지 포함했다. 문맥 파악을 위한 LSTM이 사용되었다. Stephane Schwarz [] 의 논문에서는 소셜 미디어 네트워크를 통해 유포되는 게시물 중 다른 뉴스의 일반적인 주장들과 상반되는 정보를 포함한 게시물을 가짜뉴스로 정의했다. 소셜 미디어 네트워크의 게시물과 답변 게시물, 게시물과 관련된 뉴스에서 CNN으로 주장을 추출하여 비교하는 방법이 사용되었다. Rohit Kumar Kaliyar[]의 논문에서는 주제와 맞지 않는 내용이 본문에 존재하며, 주장에 근거가 없는 뉴스를 가짜뉴스로 정의했다. 사전 학습된 언어 모델 GloVe(Global Vectors for Word Representation)를 사용하여 뉴스의 단어 임베딩을 거치고 CNN을 사용하여 주장을 추출하는 방법이 사용되었다. 마지막으로 Heejung Jwa[]의 논문에서는 기사의 제목과 상관없는 내용이 본문에 작성되어 독자가 원하지 않은 정보를 제공하는 뉴스를 가짜뉴스로 정의하였다. 사전 학습된 언어 모델 BERT를 사용하여 단어 임베딩을 거치고 분류 레이어를 추가시켜 가짜뉴스를 판단하는 방법이 사용되었다. 그 외 딥러닝 네트워크 기반의 다양한 논문을 조사하였으며 결과는 표 4와 같다.

표 4. 가짜뉴스 탐지 모델 테스트 결과

|  |  |
| --- | --- |
| **모델** | **정확도** |
| LSTM[] | 89.92% |
| LSTM + CNN[] | 94.13% |
| BERT[] | 99.13% |

다양한 논문 조사 결과 인공지능 기반 가짜뉴스 탐지 방법에는 CNN, LSTM, 그리고 GloVe, ELMo, BERT와 같은 사전 학습된 언어 모델들이 가장 많이 사용된다는 것을 알 수 있었다. 본 연구에서는 가짜뉴스 탐지를 위하여 사용할 모델을 선택하기 위하여 여러 모델에 동일한 영어 데이터셋을 학습시켜 테스트 정확도를 확인하고자 하였다. 데이터셋은 Kaggle의 Fake News Dataset[]의 Train.csv 파일을 사용하였으며, 해당 데이터셋은 진짜뉴스 10,387개, 가짜뉴스 10,413개로 이루어져있다. 데이터셋의 90%는 학습 데이터로, 10%는 테스트 데이터로 사용하였다. 테스트에 사용된 모델은 LSTM 모델, LSTM과 CNN을 사용한 하이브리드 모델, 그리고 사전 학습된 언어 모델들 중 성능이 좋은 것으로 알려져 있는BERT 모델이다. 이 실험을 하기 위한 가짜뉴스 정의는 가짜뉴스에서 많이 사용되는 어휘로 작성된 뉴스이다. 세 가지 오픈소스 모델은 모두 진짜뉴스와 가짜뉴스 각각의 본문에서 많이 사용되는 어휘를 학습시켜 가짜뉴스를 탐지할 수 있도록 하는 모델이다. 세 가지 모델에 동일한 데이터를 학습시키고 테스트한 결과는 표와 같다. 본 연구는 가장 높은 정확도를 보였던 인공지능과 시맨틱의 하이브리드 모델 BERT를 기반으로 한국어 가짜뉴스를 탐지하고 결과를 설명할 수 있는 모델을 제안하고자 한다.

1. **영어 가짜뉴스 탐지**

본 연구에서는 영어로 된 가짜뉴스 데이터셋을 학습시키고 테스트하기 위한 모델로 감성 탐지 서비스와 마찬가지로 BertForSequenceClassification 클래스를 사용하였다. 영어 데이터이기 때문에 클래스 안의 BERT 모델은 어느 유형을 선택해도 사전 학습되어있지만, 영어 데이터 테스트 후 실행할 한국어 데이터 테스트를 대비하여 다양한 언어가 지원되는 bert-base-multilingual-cased 모델을 사용하였다. BERT는 사전 학습된 언어 모델로 엄청난 정보를 포함하고 있지만, 다방면으로 자세한 정보를 가지지 않았기에 분류에 맞는 추가 학습이 필요하다. 데이터셋으로는 앞서 설명한 Kaggle의 Fake News Dataset의 Train.csv 파일을 사용하였다. 같은 출처의 Test.csv 파일을 사용하지 않은 이유는 모델의 정확도를 계산하기 위해선 데이터마다 label이 필요했지만 속성으로 label을 가지고 있지 않은 Test.csv 파일은 적합하지 않다고 판단했기 때문이다.

BERT를 이용한 영어 가짜뉴스 탐지 모델은 표 5와 같이 99.13%의 정확도를 보였다. 그러나 BERT는 지식 베이스의 모델이기 때문에 학습 데이터와 테스트 데이터의 출처가 달라졌을 때는 정확도가 떨어짐을 알 수 있었다. 같은 모델로 Kaggle Fake News dataset 데이터로 학습시키고 Kaggle Fake and real news Dataset[] 데이터로 테스트하였을 때 50.46%의 정확도를 보였다. 같은 Kaggle의 데이터이며 두 가지 모두 정치적 뉴스로 이루어진 데이터이지만, Fake News dataset의 경우 경쟁을 위하여 미국 테네시 대학교의 UTK Machine Learning Club에서 제공한 데이터로 약 5개의 Kaggle 데이터에서 수집했다고 하였으나 정확한 출처가 나와있지 않다. Fake and real news Dataset의 경우 진짜뉴스는 신뢰가능한 뉴스 제공 사이트 Reuters.com을 통해, 가짜뉴스는 팩트체크 사이트 Politifact에서 수집하였으며 주로 2016년에서 2017년 사이의 뉴스로 이루어져있다.

해당 실험을 통해 테스트 데이터가 학습된 데이터와 같은 정치 관련 데이터라고 해도 수집 시기나 정치 분야 등에 따라 학습된 지식과 다르다면 정확한 분류가 어렵다는 것을 알 수 있었다. 학습 데이터는 어떤 유형의 테스트 데이터가 들어오더라도 정확하게 분류될 수 있도록 더 넓은 시기와 다양한 분야 데이터로 이루어져야 한다. 또한 데이터를 수집하는 출처를 하나의 사이트로 제한하지 않고 여러 사이트로 넓혀 작성된 기사의 어휘가 편향되지 않아야 한다. 우리는 해당 결과를 통해 지식 베이스 모델인 BERT는 데이터셋의 영향을 크게 받는다는 것을 알 수 있었다.

1. **한국어 가짜뉴스 탐지**

한국어 가짜뉴스 탐지를 위해 첫 번째로 서울대학교 언론정보연구소가 운영하는 SNU 팩트체크에서 뉴스 데이터를 수집했다. SNU 팩트체크에서 다루는 데이터는 주로 정치인이나 공직자 후보의 발언, 뉴스나 소셜 미디어 등을 통해 대중에게 알려지는 정보들이다. 그러므로 SNU 팩트체크 사이트에는 일반적인 온라인 기사뿐만 아니라 기사, 뉴스, 블로그, 카페 등의 다양한 형태의 데이터를 검증한다. 한 뉴스의 사실 판단을 한 결과에서 얻을 수 있는 데이터는 사실 판단을 하고자 한 내용의 요약 정리, 내용의 출처, 검증 내용, 검증의 출처 등이다.

우리는 지식 베이스로 label을 분류하는 BERT를 테스트하기 위하여 기사, 블로그 등 출처와 상관없이 요약 정리 데이터를 수집하였다. 요약 정리 데이터는 10-30개의 단어로 이루어져있으며 판단하려는 내용을 간단하게 설명한 데이터다. SNU 팩트체크에서의 판단 결과는 '사실', '대체로 사실', '절반의 사실', '대체로 사실 아님', '전혀 사실 아님', '판단 유보' 등 6가지로 분류되는데 진짜뉴스와 가짜뉴스의 정확한 판단을 위해 '사실' 데이터를 진짜뉴스로, '전혀 사실 아님' 데이터를 가짜뉴스로 판단하기로 하였다. 그러나 SNU 팩트체크에서 다루고 있는 데이터는 검증이 필요한 '전혀 사실 아님'에 가까운 데이터가 많았기 때문에 '사실'로 판단되는 데이터의 수가 현저히 적었다. 모델 학습을 위해선 많은 데이터가 필요하기에 우리는 '대체로 사실' 데이터도 진짜뉴스에 포함시켜 수집하였다. 그러나 SNU 팩트체크에서 제공하고 있는 데이터의 수 자체가 많지 않기에 정치, 경제, 코로나 분야에서 진짜뉴스 100개, 가짜뉴스 100개의 데이터를 수집할 수 있었다.

높은 성능의 모델로 학습시키기에 많은 데이터는 아니지만 BERT가 한국어 뉴스도 잘 분류할 수 있는지 테스트하기엔 문제되지 않았다. 총 200개의 데이터 중 90%는 학습 데이터로, 나머지 10%를 테스트 데이터로 사용하였다. 데이터가 적은만큼 Epoch를 여러 번 반복시켰다. 학습 정확도가 98.41%가 되었을 때 학습된 모델로 테스트한 결과 80.95%의 정확도를 얻었다. 데이터가 적더라도 학습된 데이터와 테스트 데이터의 시기와 분야가 비슷하여 지식이 겹친다면 어느정도 분류가 가능하다는 것을 알 수 있었다.

우리가 최종적으로 하고자 한 실험은 텍스트로 이루어진 온라인 기사의 사실 검증이다. SNU 팩트체크 사이트에서 요약 정리 데이터가 아닌 판단하고자 하는 내용의 출처를 통해 온라인 기사를 수집하였다. 영어 가짜뉴스 데이터셋의 메타데이터들을 기반으로 다양한 모델에서 다양한 방법으로 유용하게 사용될 수 있도록 한국어 가짜뉴스 데이터셋을 만들고자 하였다. 다음 표는 한국어 가짜뉴스 탐지 서비스를 위한 실험 환경이다.

표 5. 한국어 가짜뉴스 탐지 실험 환경

|  |  |
| --- | --- |
| **개발 환경** | Colab |
| **런타임 환경** | GPU |
| **모델** | BertForSequenceClassification |
| **토크나이저** | BertTokenizer |
| **데이터셋** |  |
| **데이터 수** | 학습 데이터: |
| 테스트 데이터: |
| **라벨** | 진짜뉴스: |
| 가짜뉴스: |
| **테스트 정확도** |  |

# 제 5장 결론

**참고문헌**

[1]