대학원 도시빅데이터융합학과

Pairing 프로그램 연구 활동 결과 보고서

<table>
<tr><td rowspan="6">Pairing
프로그램
연구 활동
개요</td><td colspan="2">활동학기</td><td colspan="2">2021학년도
제2학기</td><td>교과목명
(교과번호)</td><td colspan="2">비정형데이터 분석
(88017)</td></tr>
<tr><td rowspan="5">참여연구원</td><td>연번</td><td>학과</td><td>학위과정</td><td>학번</td><td colspan="2">성명</td></tr>
<tr><td>1</td><td>전자전기컴퓨터공학과</td><td>석사과정</td><td>G202149025</td><td colspan="2">정주은</td></tr>
<tr><td>2</td><td>공간정보학과</td><td>석사과정</td><td>G202193010</td><td colspan="2">배수현</td></tr>
<tr><td>3</td><td>도시빅데이터융합학과</td><td>석사과정</td><td>G202146005</td><td colspan="2">송태현</td></tr>
<tr><td>4</td><td>도시빅데이터융합학과</td><td>박사과정</td><td>G20204602</td><td colspan="2">양진욱</td></tr>
<tr><td rowspan="2">연구과제
개요</td><td colspan="2">연구과제명</td><td colspan="5">코로나 트위터 감정 분석(covid19 tweet sentiment analysis)</td></tr>
<tr><td colspan="2">연구목적 및
필요성</td><td colspan="5">딥러닝 기반 코로나19 관련 트위터 감정분석 성능 비교

인공지능이라는 큰 틀 안에 머신러닝(기계학습)이 있고, 그 안에 딥러닝이 있다. 그리고 NLP라고 불리는 자연어 처리는 머신러닝과 딥러닝을 교집합으로 가지고 있는 영역이 있다. 본 연구에서는 다양한 자연어 처리 기술을 분석하고 연구 동향을 파악하고자 한다. 이에 따라, 대표적인 자연어 처리 기술들을 선정하고 이를 통해 코로나 관련 트위터 메세지 기반의 감정 분석을 진행하였다.</td></tr>
</table>

	국내외 연구동향	텍스트로부터 의미를 이해하고 유용한 정보를 추출하는 자연어처리 기술은 딥러닝 기술의 발전에 따라 최근 몇 년간 비약적인 발전을 이루었다(Young, 2018). ● 과거에는 SVM이나 로지스틱 회귀 모델같은 shallow models에 기반한 머신러닝이 사용되었다. 이는 특징을 사람이 직접 추출하여야 하며 이에 의존적인 단점이 존재한다(Young, 2018). ● 최근에는 딥러닝을 이용한 자동화된 특징 추출이 가능하다. 이는 워드 임베딩(Milokov et al., 2010, 2013a)과 딥러닝 기법(Socher et al., 2013)의 성공에 따른 것이다(Young, 2018). ● 최근 진행된 연구에서는 관계 기반 질의응답(relation-based question-answering) 과제에서 별도의 지식 그래프(knowledge graph)를 이용해 사실 관계(entity relations)를 표현(Ye, 2019; Lv, 2020)하고자 한다. 이벤트 간의 관계(event relations)를 표현한 추론적 지식 그래프(inferential knowledge graph)를 GPT과 결합해 추론 능력을 높이는 연구를 진행하였다. ● 모델 학습 및 추론에 드는 시간과 자원을 획기적으로 줄여 언제 어디서나 언어이해 기능을 이용할 수 있도록 하는 경량화 연구인 MobileBERT(Sun, 2020)나, SentenceBERT(Reimers, 2019)가 있다.
	주요연구내용	자연어 처리 분야에서 대표적으로 사용되는 머신러닝과 딥러닝기법을 사용하여 감정분석을 수행하고 각 성능을 비교분석하였다. 머신러닝 분석방법으로는 서포트벡터머신(LinearSVC), 랜덤포레스트, Multinomial NB_나이브베이즈를 사용하였으며, 딥러닝 분석방법으로 LSTM, GRU, BERT를 활용하였다.
	연구의 추진전략 및 방법	본 연구에서는 전세계적인 소셜 미디어 '트위터'에서 수집한 데이터를 가지고 코로나19 관련 감정분석을 수행하고자 한다. 트위터가 가지는 구조적 특징을 유지하면서도 효율적인 데이터 학습을 위해 불용어 제거, 정규표현식, tf-idf를 이용한 텍스트 전처리를 진행하였다. 텍스트를 컴퓨터가 이해할 수 있는 언어로 변환하기 위해서 통계적 기반의 임베딩 기법과 문장 수준의 임베딩 기법을 적용하였다. 감정분석을 위한 자연어 처리를 위해서는 머신러닝의 서포트 벡터 머신, 랜덤 포레스트, 나이브베이즈 기법과 딥러닝의 LSTM, GRU, BERT를 통해 추론하고 그 결과를 비교분석하였다.
	기대효과	자연어 처리는 인공지능 초기부터 관심도가 높았으며 다른 분야가 알파고 이후 급격히 증가한 것에 비하면 꾸준히 성장해 온것으로 보인다. 언어와 기술에 대한 이해를 동시에 필요로 하기에 특히 SNS 텍스트 데이터에 대해 기업들이 관심을 가지고 있다. 우리가 진행한 sentiment analysis는 기업에 대한 평가를 빠르게 파악하여 고객의 요구 사항을 처리할 수 있게 해준다. 플레시먼힐러드의 소셜 부대표인 스미스는 NLP 기술이 구체적인 감정을 파악할 수 있을 만큼 정교 해졌다고 밝혔다. 앞으로도 시장에 오래 머물수록 성능은 향상될 것으로 보인다.

연구활동 결과보고 (요약)	머신러닝으로 학습하여 추론된 트위터 메서지 감정분석을 Confusion Matrix로 확인한 결과는 다음과 같다. LinearSVC (서포트벡터머신)의 정확도(acc)는 79%이다. 랜덤포레스트의 경우 추론 결과가 약 47%로 기대 이하의 매우 낮은 성능을 보였다. 머신러닝으로 추론한 감정분석의결과를 분류별로 분석하였다. 'Neutral'과 'Negative' 카테고리의 경우 f1-score 값이 0.80, 0.83으로 분류된 반면, 'Positive'는 0.68로 다른 감정 카테고리에 비해 분류 성능이 상당히 떨어지는 것으로 나타났다. 이에 대해서는 머신러닝 모델에 대한 분석 뿐만 아니라 학습 데이터 분석이 추가적으로 진행되어야 할 것이다. 딥러닝 모델로 학습하여 추론된 모델별 결과는 다음과 같다. GRU는 LSTM보다 학습 속도가 빠르다고 알려져있지만 반드시 LSTM 대신 사용하는 것이 좋다고 하지 않는다(Denny Britz et al, 2017). 본 실험에서도 GRU의 학습 속도가 빨랐으나, 정확도는 LSTM보다 떨어지는 것으로 나타났다. LSTM의 정확도는 87%이고, GRU의 정확도는 80%로 도출되었다. 연구 결과, 머신러닝보다 딥러닝 모델을 사용한 감정분석의 결과가 훨씬 높음을 알 수 있었다. 머신러닝은 미리 확인된 컨텍스트를 입력하거나 사람이 개입하여 의미를 밝히고 관계를 정의하여야 하지만, 딥러닝에서는 단어 또는 구문의 의미와 관계를 원문을 통해서 바로 학습할 수 있기 때문이다. 그러나 본 연구에서 활용한 딥러닝 모델인 LSTM과 GRU는 hidden state 벡터에 모든 단어의 의미를 담아야 하기 때문에 모든 정보를 담기 어렵다. 이 문제를 해결하기 위해 Attention 기반 BERT 모델 실험을 추가로 진행한다면 성능을 높일 수 있을 것이라 기대한다.
향후 활용 계획	자연어 처리 분야에서 RNN 계열의 모델들이 갖고 있던 문제들을 Transformer 아케텍처를 기반으로 한 최신 고성능 모델로 해결한다. 대표적으로 BERT와 GPT이다. 2020년 구글은 GPT-3를 발표하여 기존 GPT-1의 1000배, GPT-2의 100배 보완했다. BERT와 GPT의 차이는 Attention 참조방향이고 GPT는 문장 생성에, BERT는 문장의 의미를 추출하는데 강점을 지닌 것으로 알려져 있다. 우리의 목표는 딥러닝 기반으로 한 모델 성능 비교이므로 최근 트랜드인 GPT-3를 활용할 계획이다.

제출일자	2021/12/22	
작성자	정주은	(서명/인)
	배수현	(서명/인)
	송태현	(서명/인)
	양진욱	양진욱(서명/인)

[붙임] 연구 활동 결과 보고(자유 양식으로 별도로 연구결과보고서 등 첨부)

첨부 파일 참고