이학석사 학위논문

NLP 모델을 이용한 KOSPI 키워드집합 확장 및 키워드 검색량을 활용한 KOSPI 예측

2023년 8월

서울대학교 대학원

통계학과

고 우 진

NLP 모델을 이용한 KOSPI 키워드집합 확장 및 키워드 검색량을 활용한 KOSPI 예측

지도 교수 오 희 석

이 논문을 이학석사 학위논문으로 제출함
2023년 4월

서울대학교 대학원
통계학과
고 우 진

고우진의 이학석사 학위논문을 인준함
2023년 7월

위 원 장 김 용 대 (인)

부위원장 오 희 석 (인)

위 원 박 건 웅 (인)

국문 초록

NLP 모델을 이용한 KOSPI 키워드집합 확장 및 키워드 검색량을 활용한 KOSPI 예측

본 논문은 두개의 연구 주제로 구성되어 있다.

먼저 NLP 모델을 이용한 KOSPI 관련 키워드 집합 확장을 연구한다. 처음 설정한 KOSPI 관련 159개의 original seed keyword로부터 시작해 text와 candidate keyword를 수집하고 KLUE-RoBERTa/large, KPF-BERT, KB-ALBERT의 NLP 모델을 활용하여 벡터화를 진행한다. 임베딩 벡터들간의 코사인 유사도를 계산하여 similarity/importance score를 구하고 1차 스크리닝에 활용해 828개의 generated keyword를 생성한다. 또한 generated keyword의 검색량과 KOSPI의 상관관계를 기준으로 2차 스크리닝을 진행해 KOSPI와 관련성 있는 키워드 집합을 얻는다.

다음으로 확장된 KOSPI 관련 키워드의 검색량을 활용하여 KOSPI 예측을 연구한다. 예측모형으로는 LSTM을 활용하며 학습데이터의 기간 구조를 1~7년으로 다양하게 설정해 실험을 진행하였고 각 기간에 맞는 최적의 하이퍼파라미터 조합을 찾으려 했다. 실험결과 seed keyword 대비 새롭게 확장된 KOSPI 키워드의 검색량을 사용하였을 때 KOSPI 예측 성능이 더 우수함을 확인하였고 이를 통해 키워드 집합 확장 태스크가 잘 수행되었음을 알 수 있었다. 또한 기간별 7개 모델을 앙상블하여 예측하고자 하는 날과 가까운 시점에 큰 가중치를 준 앙상블모델이 향상된 예측 성능을 보이는 것을 확인할 수 있었다.

주요어 : KOSPI, 자연어처리, 키워드추출, 주가예측, 검색량, 딥러닝
학 번 : 2020-27834

목 차

표 목차

그림 목차

제 1 장 서 론

검색 엔진에 특정 키워드를 검색하는 행위는 인터넷 사용자들이 관심을 표출하는 행위이다. 따라서 인터넷 검색량은 대중들의 관심 정도를 나타내는 지표로써 해석이 가능하다. 실제로 광범위한 분야의 연구에서 인터넷 검색량이 활용되고 있다. Ettredge et al.(2005)는 실업과 관련된 인터넷 검색량과 실업률의 연관 관계를 연구하였다. Cooper et al.(2005)은 암에 대한 인터넷 검색량과 실제 암 발생빈도의 관계를 조사하였다. Polgreen et al.(2008)과 Ginsberg et al.(2009)는 독감과 연관된 검색 결과와 실제 독감 발생 건수와의 관계를 조사하였다. Preis et al.(2013)은 주식투자자가 경제 상황에 대해 갖는 관심에 대한 민감도가 경제 관련 인터넷 검색량과 관계가 있음을 밝혔다.

KOSPI는 한국거래소의 유가증권시장에 상장된 회사들의 시가총액의 기준시점과 비교시점을 비교하여 나타낸 지표로써, 대한민국 경제를 대표하는 가장 큰 경제지표 중 하나이다. 행동경제학적 관점에서 볼 때 주식시장은 투자자의 심리에 의해 비이성적으로 작동한다는 증거가 Baker & Wurgler(2006), Stambaugh, Yu, & Yuan(2012) 등의 연구에 의해 제시되었다. 따라서 KOSPI 관련 연구 및 투자에 특정 키워드 검색량을 활용한다면, 대중의 관심 및 심리를 반영한다는 점에서 의의가 있을 것으로 기대한다.

하지만 연구에서 활용할 검색량의 대상이 되는 KOSPI 관련 키워드 선정방법에 대한 고민이 필요하다. 기존에 주로 사용되는 KOPSI 관련 키워드들은 오래된 문헌이나 연구경험적 직관에 대부분 의존한다. 이처럼 한정적인 KOSPI 관련 키워드 집합을 사용한다면 효과적이고 생각치 못한 키워드를 놓칠 수 있다. 그렇다고 모든 단어를 고려 대상으로 두기에는 시간적, 비용적 리소스가 많이 소요된다는 문제점이 존재한다.

따라서 본 연구에서는 NLP 키워드 추출 태스크를 통해 효과적으로 KOSPI 관련 키워드의 집합을 확장하고자 한다. 이후 확장된 KOSPI 관련 키워드 집합 안에서 KOSPI와 상관관계가 높은 키워드를 선정하고, 해당 키워드들의 검색량을 사용하여 KOSPI를 예측하는 LSTM 모델을 학습하고자 한다.

본 논문의 구성은 다음과 같다. 제 2장에서는 KOSPI 관련 키워드 집합을 확장하는데 사용할 NLP 모델과, KOSPI 예측에 사용할 LSTM 모델을 간략히 소개한다. 제 3장에서는 전체 프로세스를 순서도로 정리하여 소개한다. 제 4장에서는 본 연구에서 사용할 데이터를 소개한다. 제 5장에서는 분석 과정과 실험 결과를 설명한다. 제 6장에서는 결과를 정리하고 본 연구의 의의와 한계점을 제시한다.

제 2 장 모 델

제 2.1 절 BERT

2018년 구글에서 발표한 BERT(Bidirectional Encoder Representations for Transformers) 모델은 트랜스포머 구조를 사용한 사전 학습 언어모델이다. 이전까지의 NLP 모델은 특정 태스크를 수행하기 위해서는 모델을 처음부터 학습시켜야 하고, 태스크에 적합한 많은 양의 레이블링된 데이터를 필요로 했다. 하지만 현실에 존재하는 레이블링된 데이터의 수는 제한적이며, 데이터 레이블링 작업은 어렵고 오래 걸리기 때문에 사전 학습 언어모델의 필요성이 대두되었다.

BERT는 위키피디아(약 25억개의 단어)와 BooksCorpus(약 8억개의 단어)와 같은 방대한 양의 레이블링 되지 않은 텍스트 데이터로 사전 학습을 진행하여 언어지식을 갖춤으로써 이와 같은 문제를 해결한다. 레이블링 되지 않은 텍스트 데이터로부터 레이블링 된 텍스트 데이터를 자동으로 생성하여 지도학습에 사용하기 때문에 준지도학습(Semi-supervised learning)이라고도 불린다.

BERT가 사전 학습단계에서 수행하는 작업에는 MLM(Masked Language Modeling)과 NSP(Next Sentence Prediction)가 있다. MLM은 주어진 텍스트의 토큰의 일부분을 랜덤으로 마스킹하고 마스킹된 위치의 토큰을 예측하는 작업이다. MLM 작업을 통해 모델은 단방향이 아닌 양방향에 해당하는 주변의 문맥 정보를 고려할 수 있게 된다. NSP는 두 문장 A, B가 주어질 때, B가 실제로 A 다음에 나오는 문장이 맞는지 이진 분류하는 작업이다. NSP 작업을 통해 모델은 여러 문장 간의 관계를 파악할 수 있게 된다. 이렇게 학습된 BERT 모델은 다양한 NLP 태스크에 활용된다. 주로 태스크에 적합한 학습데이터를 활용하여 추가학습을 진행해 태스크에 특화된 모델을 만드는 Fine-tuning 과정을 거치게 된다. Fine-tuning시에는 사전 학습된 모델을 사용하기 때문에 상대적으로 적은 양의 학습데이터와 학습시간이 소요된다. 사전 학습된 언어모델을 이용해 Fine-tuning 과정을 통해 추가 학습된 NLP모델은 그렇지 않은 모델에 비해 해당 태스크에 대해 더 좋은 성능과 강건성을 보인다고 알려져 있다.

제 2.2 절 RoBERTa

RoBERTa(Robustly optimized BERT Pretraining Approach)는 학습데이터 및 학습방식을 수정하여 기존 BERT의 성능을 향상시킨 모델이다. BERT와 비교해 주요 차이점은 다음과 같다.

우선 학습데이터의 양을 증가시켰다. BERT의 사전 학습시에 사용한 위키피디아, BookCorpus 외에도 CC-News, OpenWebText, Stories를 학습데이터로 사용하였다. 이는 BERT 대비 약 10배 수준이다. 배치 사이즈 역시 256에서 8000으로 증가시켜 더 큰 배치사이즈로 학습을 진행했다.

사전 학습 과정에서도 차이점이 존재한다. MLM 작업에서 정적 마스킹 대신 동적 마스킹 방법을 사용하였다. 기존의 BERT 모델에서는 한 문장에 대해 적용한 마스킹을 여러 epoch 동안 동일한 input으로 사용했다. 이에 반해 RoBERTa에서는 한 문장을 여러 개로 복사해 마스크의 위치를 동적으로 결정한다. 따라서 다양하게 마스킹 처리된 데이터의 학습을 가능하게 한다. 또한 NSP의 효용성에 의문을 제기하며 NSP 작업을 제거하였다. NSP 태스크를 수행하지 않고 학습하는 것이 BERT의 성능향상으로 이어진다는 사실을 실험을 통해 논문에서 밝혔다.

BERT의 경우 WordPiece 토크나이저를 사용한 반면 RoBERTa에서는 BBPE 토크나이저를 사용한다. BERT의 사전크기는 30,000토큰이고 RoBERTa의 경우는 50,000토큰이다. 또한 하나 이상의 문서를 이용하여 최대 토큰 길이에 가깝게 구성하는 FULL-SENTENCES와 같은 개선된 학습 방법을 제시하였다.

제 2.3 절 ALBERT

사전 학습 언어모델의 모델 사이즈 증가는 종종 Downstream 태스크의 성능 향상으로 이어진다. 하지만 메모리 및 학습시간의 한계로 모델의 크기를 계속해서 증가시키기 어렵다. 이러한 문제를 해결하기 위해 ALBERT가 등장하였다.

ALBERT는 학습데이터로 BERT와 동일하게 위키피디아, BookCorpus 데이터를 사용한다. ALBERT는 RoBERTa와 마찬가지로 NSP의 문제점을 제기한다. NSP의 경우 MLM 대비 난이도가 높지 않아 유용하지 않으며, 문장의 일관성 뿐만 아니라 주제에 대한 예측을 포함하는 태스크임을 지적한다. 따라서 기존 NSP작업 대신 SOP(Sentence Order Prediction)를 제안한다. SOP는 연속되는 두 문장(Positive)과 문장 순서를 앞뒤로 바꾼 문장(Negative)을 이용하여 문장의 순서가 옳은지를 예측하는 방식이다. 또한 BERT에서는 히든 차원과 임베딩 차원이 동일한데 반해, ALBERT에서는 임베딩 차원을 따로 분리하여 더 작은 차원으로 적용하였고, 각 Layer의 파라미터를 공유하는 방식을 통하여 모델의 파라미터의 수를 줄이고 학습시간을 단축시켜 더 큰 모델의 학습이 가능하게 했다.

제 2.4 절 LSTM

RNN(Recurrent Neural Network)은 히든 노드가 방향을 가진 엣지로 연결돼 순환구조를 이루는 인공신경망의 한 종류로써 시계열 데이터를 다루는데 효과적이다. 하지만 RNN은 관련정보와 그 정보를 사용하는 지점 사이의 거리가 멀어지는 경우에 학습능력이 현저히 저하되는 장기의존성(Long-Term Dependency) 문제를 가지고 있다.

LSTM(Long Short-Term Memory)은 이러한 문제를 해결하기 위해 나온 모델로써 1997년 Hochreiter & Schmidhuber에 의해 소개되었다. LSTM은 RNN과 다르게 cell state와 3개의 gate(input gate, forget gate, output gate)를 가지고 있다. Cell state는 LSTM 구조의 핵심요소로써 사소한 linear interaction만을 적용시키며 과거의 정보를 변형없이 전달하는 역할을 한다. Forget gate는 과거의 정보를 버릴지 말지를 결정한다. Input gate는 새로운 정보를 cell state에 저장/추가할 것인가하는 업데이트 정보를 결정하고, output state는 어떤 정보를 output으로 내보낼지 결정한다.

이처럼 LSTM은 3개의 gate를 활용해 cell state를 보호/제어하며 RNN의 장기의존성 문제를 효과적으로 해결하여 길고 복잡한 데이터도 손실 없이 학습할 수 있다.

제 3 장 방 법

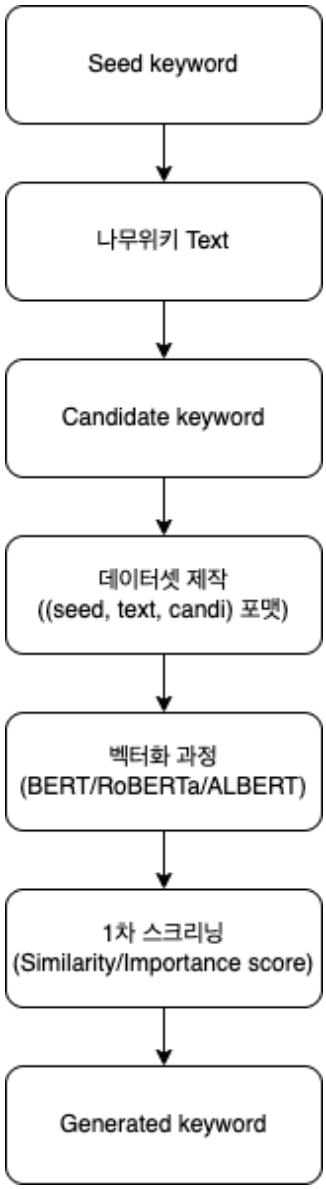

그림 1. KOSPI 관련 키워드 집합 확장 프로세스 순서도

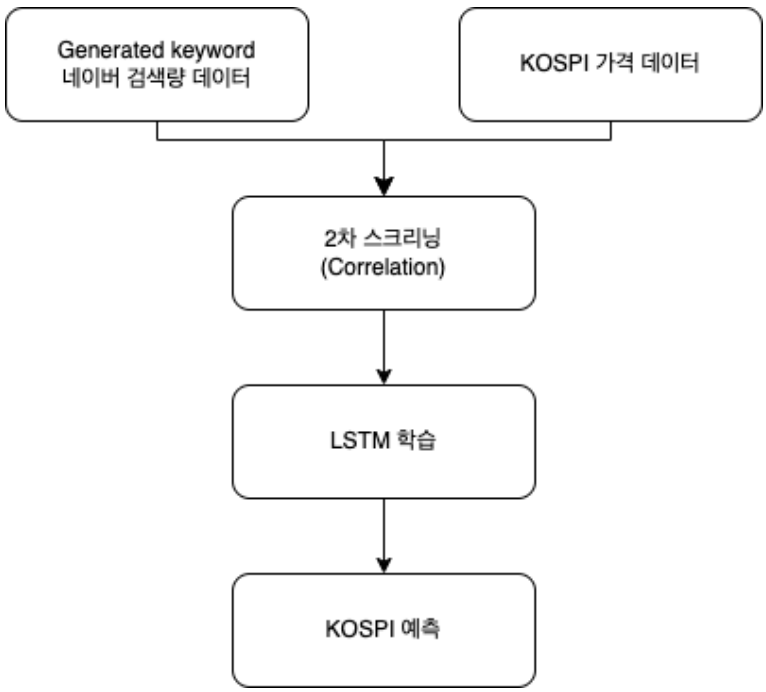

그림 2. KOSPI 예측 프로세스 순서도

제 4 장 데 이 터

제 4.1 절 나무위키 Text 데이터셋

KOSPI 관련 seed keyword로부터 KOSPI 관련 키워드 집합을 확장하기 위해 본 연구에서는 나무위키 문서 데이터셋을 활용하고자 하였다. 나무위키는 크롤링을 통한 접근을 제한하며 DB(dump) 데이터를 제공한다. 하지만 2023년 2월 기준 dump 파일 역시 일시적으로 제공이 중지되어 있다. 이에 Huggingface에 업로드 되어있는 heegyu/namuwiki-extracted 데이터셋을 사용하였다. 해당 데이터셋은 title, text, contributors, namespace의 총 4개의 열과 571,308개의 행으로 이루어져 있으며 2.19GB의 크기를 가진다. 이 중 title과 text를 사용하였다. Title과 text는 각각 나무위키 문서의 제목과 내용을 나타낸다.

제 4.2 절 야후 파이낸스 KOSPI 데이터

Yahoo finance 홈페이지의 주식 가격 데이터에서 KOSPI (코드:^KS11)에 대한 정보를 수집한다. Ran Aroussi의 파이썬 라이브러리 yfinance를 이용하여 데이터를 불러왔다. 파이썬 데이터프레임 형식의 결과가 반환되며 날짜, 시가, 고가, 저가, 종가, 거래량 등의 정보가 제공된다. 조회 기간 및 간격을 인자로 주어 조정할 수 있다. 본 연구에서는 일별 KOSPI 종가를 활용하였다.

제 4.3 절 네이버 검색량 데이터

네이버 Developers 데이터랩 API를 이용하여 키워드에 대한 검색량을 수집하였다. 현재 API를 통해 2016년 1월 1일부터의 검색량을 수집할 수 있다. 설정한 키워드에 대해 기간을 일간, 주간, 월간 단위로 조회할 수 있다. 설정한 기간 중 검색 횟수가 가장 높은 시점을 100으로 두고 나머지는 상대적인 값으로 하여 데이터를 제공한다.

제 5 장 실 험

제 5.1 절 KOSPI 관련 키워드 집합 확장

제 5.1.1 절 Seed keyword, text, candidate keyword 수집

우선 KOSPI 관련 키워드 집합 확장의 기초가 되는 original seed keyword를 수집한다. 통계청에서 제공하는 기간별(19~22년) 경제 분야 주요 키워드, 국내 주식시장 분석 보고서 및 관련 논문으로부터 총 159개의 KOSPI 관련 original seed keyword를 선정하였다(표 1). 선정된 모든 original seed keyword가 나무위키에서 검색되지는 않았다. 나무위키는 같은 대상을 나타내는 동의어에 대해 해당 문서로 연결해주는 리다이렉트 검색 알고리즘을 사용한다. Original seed keyword를 나무위키에 검색하여, 나무위키에 검색되는 것들과 새롭게 리다이렉트된 문서의 제목을 수합하여 seed keyword로 선정하였다. seed keyword는 총 101개이다(표 2).

표 1. Original seed keyword

Original seed keyword (159)
ESG(비재무적성과), GDP, S&P 500, banned stocks, brokerage stocks, low priced stocks, old stock holders, securities networks, simulated stock trading, stock code, stock market quotations, 가상자산, 개념주, 거품, 경제, 경제데이터, 경제성장률, 계좌개설, 고용, 공급망, 공매도, 구리, 국내총생산, 국민연금, 금리, 금리인상, 금융, 금융뉴스, 금융시장, 금융협업, 급락, 기관, 기업공개, 기준금리, 나스닥, 내부자, 뉴욕증시, 닛케이, 다우, 다크 호스, 달러, 대출, 대형마트, 마스크, 막대도표, 메타버스, 면세점, 모빌리티, 무역분쟁, 무역수지, 무역협상, 미국, 미국대선, 미세먼지, 바이오, 배당, 배터리, 백신, 보너스, 부, 부동산, 분양가, 분양가상한제, 불 마켓, 브렉시트, 블록체인, 비트코인, 빅데이터, 사회적거리두기, 삼성전자, 상승, 상장지수펀드, 선물, 선물거래, 세금, 소득, 소매, 소비자물가지수, 소상공인, 수익률, 수출, 수출규제, 수출액, 스타트업, 시가총액, 시장, 시장상황, 신주, 암호화폐, 양적완화, 업무협약, 에너지, 예금금리, 오늘의 시장, 오늘의 주식시장, 외국인, 외국인순매도, 외국인순매수, 원달러환율, 원자재, 위험관리, 유가, 유동성, 유량주, 은행 융자, 인공지능, 인플레이션, 인플레이션율, 일일한도, 일자리, 임대료, 자기자본(보통주), 자본, 자산관리, 자율주행, 재건축, 재정보증, 재택근무, 재판매, 전기차, 전셋값, 정보공개, 정책, 주가, 주가지수, 주가지수선물, 주력, 주식, 주식거래, 주식계좌개설, 주식소개, 주식시장, 주식자보금, 주식추천, 중국, 중국 주식, 중소기업, 증권, 증권시장, 증권투자기금, 차익거래, 청약, 최저임금, 치료제, 친환경, 코로나 19, 코스닥, 코스피, 클라우드, 탄소중립, 통화, 투기자, 투자와 자산관리, 투자자, 파산, 팬데믹, 펀드회사, 핀테크, 환율

표 2. Seed keyword

Seed keyword (101)
S&P 500, 거품경제, 경제, 경제성장률, 계좌개설, 고용, 공매도, 구리, 국내총생산, 국민연금, 금융, 금융기관, 기업공개, 기준금리, 내부자거래, 뉴욕증권거래소, 닛케이 225, 다우 존스 산업평균지수, 다크 호스, 달러, 대출, 대통령 선거, 마스크, 메타버스, 면세점, 무역수지, 무역전쟁, 미국, 미세먼지, 바이오, 배당, 배터리, 백신, 범유행전염병, 보너스, 보증, 보통주, 부동산, 분양가상한제, 브렉시트, 블록체인, 비트코인, 빅 데이터 프로세싱, 사회적 거리두기, 삼성전자, 상장지수펀드, 선물(금융), 세금, 소득, 소비자물가지수, 소상공인, 수출, 스타트업, 시가총액, 시장(경제), 암호화폐, 약, 양적완화, 우량주, 원자재, 유가, 유가증권, 유동성, 이자, 인공지능, 인플레이션, 임대료, 자본, 자본금, 자율주행, 재건축, 재산, 재택근무, 재테크, 전기자동차, 전세, 정책, 주가 지수, 주식, 주식시장, 중국, 중소기업, 직장, 차익거래, 청약, 최저임금제, 친환경, 코로나바이러스감염증-19, 코스닥, 코스피, 클라우드 컴퓨팅, 탄소 중립, 탈것, 투자자, 파산, 펀드, 핀테크, 한도대출, 할인점, 화폐, 환율

다음으로 그림 3과 같이 seed keyword로 검색된 나무위키 문서를 seed keyword에 대응하는 text로 수집하였다. 이후 과정에서 text는 NLP 모델의 input으로 사용되는데 그 크기를 일정하게 맞춰주기 위해서, seed keyword로 검색된 나무위키 문서 전문을 그대로 사용하는 대신 적절히 쪼개서 사용하였다. 개행 문자(₩n)를 기준으로 나무위키 문서를 문단(소주제)별로 분리한 후, 길이(토큰 기준)가 500이 넘지 않도록 합쳐서 text로 사용하였다. 이때 토크나이저로는 RoBERTa 토크나이저를 사용하였다.

"貸出돈이나 물건 등을 빌려 주는 일. 그 반대로 빌려오는 것은 차입(借入)이라고 한다. 은행은 이걸 토대로 돈을 번다. 지급준비제도 참고. 역으로 빚내서 돈을 번다. 레버리지 참고.보통 시중에 있는 은행들은 대출이자로 먹고사는 것이다. 예금과 대출의 금리차를 '예대마진'이라고 한다.금융시장을 인체에 비유하면 대출 상품들은 적혈구와 같은 역할을 한다고 볼 수 있다. 잘 알려진 대항해시대 또한 스페인, 포르투갈 등지의 왕가로부터의 대출을 통해 시작되었다. 개인은 그냥 대출이라고 한다. 혹은 융자라고도 부르는데, 융자는 빌리는 입장에서 지불해야 하는 채무만을 뜻하는 용어로 대출에 비해 의미가 좁다. 또한 시대적으로도 신세대는 잘 쓰지 않아 사어가 될 가능성이 있다."

그림 3. Seed keyword에 대응하는 나무위키 문서 text 수집

이후 단계에서는 파이썬 라이브러리 KoNLPy를 사용하여 앞서 수집한 text로부터 candidate keyword를 추출한다. KoNLPy는 한국어 정보처리를 위한 파이썬 패키지로 kkma, mecab 등의 형태소 분석기를 제공한다. kkma는 띄어쓰기 오류에 강한 형태소 분석기로 다양한 종류의 품사 태그를 지원한다. 하지만 정제되지 않은 언어에 대해서는 분석 정확도가 상대적으로 낮다는 단점이 존재한다. mecab은 일본어 형태소 분석기를 한국어에 맞게 수정한 분석기이다. 분석속도가 빠르고

어휘사전에 쉽게 새로운 단어를 추가할 수 있지만 미등록어 및 동음이의어 처리에 관한 한계점이 존재한다. 본 연구에서는 kkma와 mecab 분석기를 사용해 text로부터 명사품사의 단어를 추출하였다. kkma 형태소 분석기를 사용한 결과 31,655개의 단어가, mecab 형태소 분석기를 사용한 결과 16,040개의 단어가 각각 추출되었다. 두 분석기의 교집합에 있는 13,318개의 단어를 최초의 candidate keyword로 선정하였다.

최종적으로 (seed keyword, text, candidate keyword) 형식의 데이터를 구축하였다. 전체 데이터의 행의 개수는 127,351개 이며, 해당 데이터는 이후의 벡터화 및 스크리닝 작업에 사용된다. 데이터셋의 일부를 표 3에서 확인할 수 있다.

표 3. (Seed keyword, Text, Candidate keyword) 형식의 데이터셋 일부

Index	Seed key	Text	Candidate key
0	소상공인	소상공인마당 - 소상공인시장진흥공단에서 운영하는 사이트"소상공인"이란 소기업(小企業) 중 다음 각 호의 요건을 모두 갖춘 자를 말한다(소상공인 보호 및 지원에 관한 법률 제 2 조).상시 근로자 수가 10 명 미만일 것업종별 상시 근로자 수 등이 대통령령으로 정하는 기준에 해당할 것구체적인 기준은 중소기업의 기준과 마찬가지로 좀 복잡하게 되어 있으나, 기본적으로 상시 근로자 수 기준은 아래와 같이 되어 있다(같은 법 시행령 제 2 조 제 1 항).	마당
1	소상공인	소상공인마당 - 소상공인시장진흥공단에서 운영하는 사이트"소상공인"이란 소기업(小企業) 중 다음 각 호의 요건을 모두 갖춘 자를 말한다(소상공인 보호 및 지원에 관한 법률 제2조).상시 근로자 수가 10명 미만일 것업종별 상시 근로자 수 등이 대통령령으로 정하는 기준에 해당할 것구체적인 기준은 중소기업의 기준과 마찬가지로 좀 복잡하게 되어 있으나, 기본적으로 상시 근로자 수 기준은 아래와 같이 되어 있다(같은 법 시행령 제2조 제1항).	소상
2	소상공인	소상공인마당 - 소상공인시장진흥공단에서 운영하는 사이트"소상공인"이란 소기업(小企業) 중 다음 각 호의 요건을 모두 갖춘 자를 말한다(소상공인 보호 및 지원에 관한 법률 제2조).상시 근로자 수가 10명 미만일 것업종별 상시 근로자 수 등이 대통령령으로 정하는 기준에 해당할 것구체적인 기준은 중소기업의 기준과 마찬가지로 좀 복잡하게 되어 있으나, 기본적으로 상시 근로자 수 기준은 아래와 같이 되어 있다(같은 법 시행령 제2조 제1항).	공인
…	…	…	…
127348	경제성장률	경제성장률이 평균적으로 낮은 것은 대침체, 서브프라임 모기지 사태 시기의 수치를 합산하여서 그렇다. 자료의 수치는 아래의 통계를 합산하여 7 로 나눈 결과물이다.G20/경제성장률대한민국/경제성장률필리핀/경제성장률중국/경제성장률두산백과-경제성장률국력국가별 경제기업 관련 정보솔로우-스완 모형	정보
127349	경제성장률	경제성장률이 평균적으로 낮은 것은 대침체, 서브프라임 모기지 사태 시기의 수치를 합산하여서 그렇다. 자료의 수치는 아래의 통계를 합산하여 7로 나눈 결과물이다.G20/경제성장률대한민국/경제성장률필리핀/경제성장률중국/경제성장률두산백과-경제성장률국력국가별 경제기업 관련 정보솔로우-스완 모형	스완
127350	경제성장률	경제성장률이 평균적으로 낮은 것은 대침체, 서브프라임 모기지 사태 시기의 수치를 합산하여서 그렇다. 자료의 수치는 아래의 통계를 합산하여 7로 나눈 결과물이다.G20/경제성장률대한민국/경제성장률필리핀/경제성장률중국/경제성장률두산백과-경제성장률국력국가별 경제기업 관련 정보솔로우-스완 모형	모형

제 5.1.2 절 벡터화과정

이후 제 5.1.3 절에서 similarity score 및 importance score를 활용하여 candidate keyword를 1차 스크리닝 한다. 해당 score들은 seed keyword, text, candidate keyword를 각각 벡터로 표현한 후 이들 간의 코사인 유사도로 계산된다. 따라서 seed keyword, text, candidate keyword를 각각 벡터화하는 과정이 필요하다. KLUE-RoBERTa/large, KPF-BERT, KB-ALBERT의 총 3가지 사전 학습 언어모델을 사용하여 벡터화과정을 진행한다.

KLUE는 한국어 언어모델의 성능을 평가하기 위한 목적으로 만든 한국어 자연어이해 벤치마크 데이터셋이다. KLUE-RoBERTa/large는 한국어 데이터셋(MODU, CC-100-Kor, NAMUWIKI, NEWSCRAWL, PETITION)을 사용하여 RoBERTa 아키텍처를 기반으로 사전 학습한 거대언어모델이다. 파라미터의 수가 337M개에 달하며, KLUE 리더보드의 다양한 NLP 태스크에서 전반적으로 높은 성능을 보인다.

KPF-BERT는 한국언론진흥재단이 보유한 2000~2021년 빅카인즈 기사 8천만 건 중 1차 정제를 통해 추려낸 약 4천만 건을 BERT 아키텍처를 기반으로 사전 학습한 언어모델이다. BERT를 활용한 기존 한국어 모델들은 위키백과나 웹 문서 등을 주로 학습하는데 반해, 뉴스기사 코퍼스를 사용하여 기계독해에 강점을 보이고 뉴스기사에 특화된 모델이다.

KB-ALBERT는 KB 국민은행에서 제공하는 경제/금융 도메인에 특화된 ALBERT 아키텍처 기반의 한국어 언어모델이다. 금융 도메인과 관련된 약 15GB의 문서를 추가적으로 학습하였지만, 해당 도메인으로만 치우치지 않도록 약 25GB의 일반 도메인의 텍스트도 사전 학습에 사용하였다. 금융 도메인의 전문 용어들이 어휘사전에서 버려지지 않도록 일반적인 경우보다 큰 약 50,000개 정도 크기의 어휘사전을 사용하였다. 또한 음절단위 모델을 사용하여 조금 더 세분화된 형태소 분석 결과를 제공한다.

벡터화과정은 다음과 같다.

Step1. Seed keyword, text, candidate keyword를 토큰화한 후 각각 사전 학습 언어모델을 통과시킨다.

Step2. 특정 layer(last hidden state, embedding layer)의 출력벡터들을 pooling(max pooling, mean pooling)하여 벡터화한다.

다양한 경우를 고려하기 위해 Model, Layer, Pooling 방식에 따라 총 8가지의 경우를 계산하여 사용하였다(표 4).

표 4. Model, Layer, Pooling에 따른 8가지 벡터화 과정

Case	Model	Layer	Pooling
1	KLUE/RoBERTa-large	Last Hidden State	Max Pooling
2	KLUE/RoBERTa-large	Last Hidden State	Mean Pooling
3	KLUE/RoBERTa-large	Embedding Layer	Max Pooling
4	KLUE/RoBERTa-large	Embedding Layer	Mean Pooling
5	KPF/BERT	Last Hidden State	Max Pooling
6	KPF/BERT	Last Hidden State	Mean Pooling
7	KB/ALBERT	Last Hidden State	Max Pooling
8	KB/ALBERT	Last Hidden State	Mean Pooling

제 5.1.3 절 1차 스크리닝(Similarity score, Importance score)

최초 13,318개의 candidate keyword에 대한 1차 스크리닝의 기준으로 similarity score 및 importance score 라는 2가지 metric을 사용하였다.

Similarity score는 벡터화된 candidate keyword와 seed keyword의 코사인 유사도로 계산한다. 해당 스코어는 candidate keyword가 seed keyword와 얼마나 유사한지 나타내는 지표이고, seed keyword와 동떨어져 관련성이 적은 candidate keyword를 제외하는 도구로써 사용된다.

Importance score는 벡터화된 candidate keyword와 text의 코사인 유사도로 계산한다. 해당 스코어는 candidate keyword가 text에서 어느 정도로 중요한 단어인지 나타내는 지표이다. 수집한 text가 KOSPI와 관련된 글이라고 가정했을 때, 해당 값이 높을수록 candidate keyword가 KOSPI와 관계가 있고 text의 주요내용을 포함한다고 판단할 수 있다.

그림 4는 언어모델(KLUE-RoBERTa/large, KPF-BERT, KB-ALBERT), layer(last hidden state, embedding layer), pooling(max, mean) 방식에 따른 총 8가지 벡터화 경우에 대한 similarity score와 importance score의 분포를 나타낸 그림이다. 그림 4를 살펴보면 언어모델, layer, pooling 방식에 따라 similarity score와 importance score의 분포가 상이한 것을 확인할 수 있다.

또한 동일한 코사인 유사도 값이라도 벡터화한 방법에 따라 내포하는 의미가 다르기 때문에 직접적인 비교가 어렵다. 따라서 다양한 결과를 고려하기 위해 언어모델, layer, pooling 방식에 따른 총 8가지 벡터화 경우에 대해서, similarity score와 importance score가 모두 상위 5% 에 해당하는 candidate keyword를 각각 필터링하고

합집합하여 1차 스크리닝 결과로 사용한다.

1차 스크리닝 결과와 original seed keyword를 합집합하여 새로운 KOSPI 관련 키워드라는 의미에서 generated keyword로 이름을 붙여 사용한다. Generated keyword의 개수는 총 828개이고 표 5와 같다.

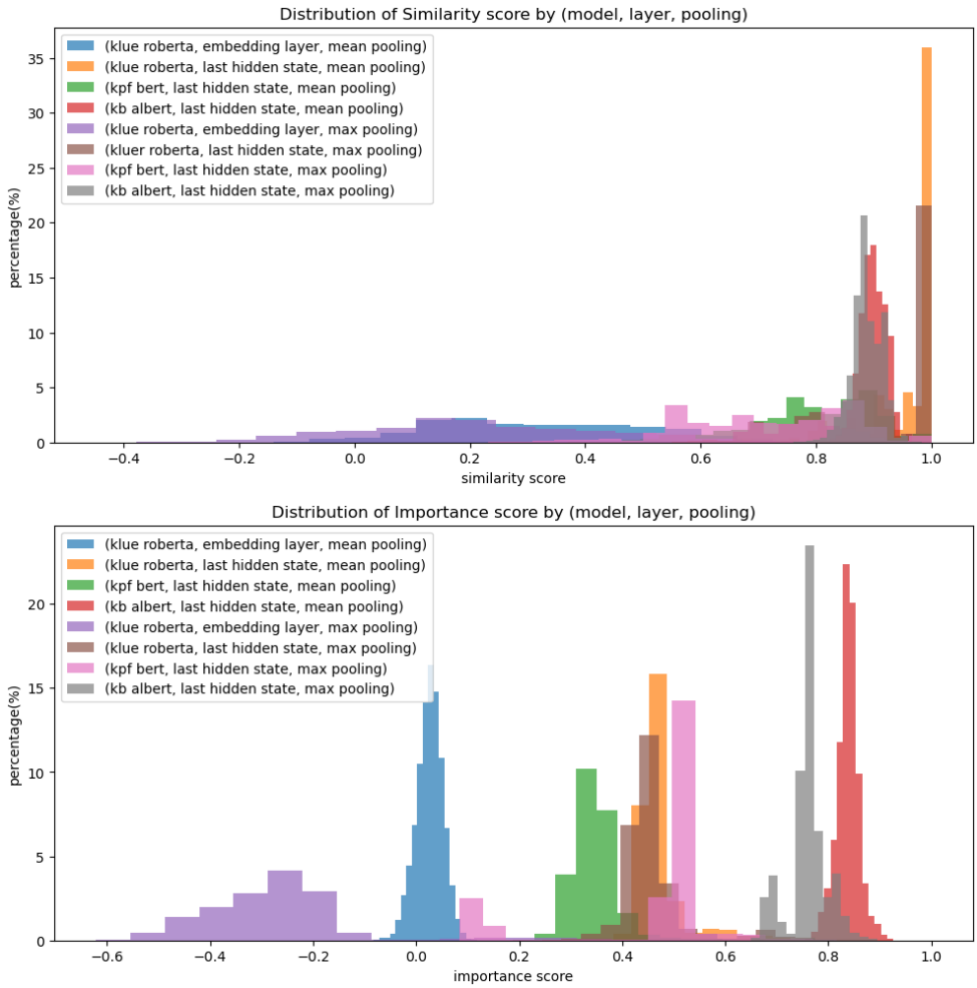

그림 4. 8가지 벡터화 과정에 따른 similarity score 및 importance score 분포

표 5. Generated keyword

Generated keyword (828)
ESG(비재무적성과), GDP, S&P 500, banned stocks, brokerage stocks, low priced stocks, old stock holders, securities networks, simulated stock trading, stock code, stock market quotations, 가격, 가능, 가상, 가상자산, 가시거리, 가전제품, 가정주부, 가치, 간섭, 감속기, 강대국, 개념, 개념주, 개발, 갭, 거, 거기, 거래, 거래세, 거주, 거품, 건축물, 검, 검사, 결국, 결산, 결제, 경국제민, 경세제민, 경영자, 경우, 경제, 경제데이터, 경제성장, 경제성장률, 경제학, 계산서, 계승, 계약서, 계약자, 계정, 계좌개설, 고도성장, 고속철도, 고용, 골자, 공공, 공공건물, 공급망, 공매도, 공무원, 공산주의, 공화국, 과잉, 과학자, 관리, 관세법, 관세청, 교대근무, 교육세, 구리, 구매, 구성, 구자학, 구조, 구축, 국가, 국내총생산, 국무총리, 국민연금, 국민은행, 국민주택기금, 국회의원, 귀신, 그것, 그라운드, 그레이트, 근로계약서, 글로벌, 금리, 금리인상, 금융, 금융뉴스, 금융시장, 금융업, 금융업자, 금융협업, 급락, 기관, 기관지, 기관지염, 기능, 기댓값, 기사, 기술원, 기술자, 기습, 기업, 기업가, 기업공개, 기존, 기준금리, 기프트카드, 기하급수, 껍질, 꿈, 나노미터, 나스닥, 날, 남아프리카, 내국세, 내부자, 내역, 네트워크, 년, 노르웨이, 눈병, 뉴욕증시, 닛케이, 다국적, 다세대, 다우, 다운로드, 다음, 다이어트, 다크 호스, 다크호스, 단어, 단위, 단정, 달러, 답변, 당국자, 당대, 당시, 당연지사, 당원, 대상, 대상자, 대전, 대체, 대출, 대통령, 대학, 대한민국, 대한전선, 대형마트, 데, 데이터베이스, 도착지, 독가스, 돈, 돈세탁, 동기전동기, 동안, 동작, 동화면세점, 뒷받침, 듀플렉스, 드라이브, 드래프트, 등, 등록증, 디노미네이션, 디스토피아, 디스플레이, 디플레이션, 디플레이터, 땅, 때문, 라디에이터, 라스파이레스, 라이선스, 라이트닝, 랩, 량, 레벨, 레이어스, 레플리카, 롯데카드, 리하르트, 릴리스, 마르크, 마스크, 마운틴뷰, 마음가짐, 마이너, 마이너스, 마인, 마찬가지, 막대도표, 망라, 매입, 매한가지, 머리카락, 메모리, 메소포타미아, 메시지, 메타버스, 면세, 면세점, 면세품, 면허, 면허증, 명실상부, 모니터링, 모빌리티, 모집, 모터스, 목동, 목표, 무기, 무언가, 무역분쟁, 무역수지, 무역의존도, 무역협상, 문제, 물가, 물자, 물적분할, 미국, 미국대선, 미국식, 미국인, 미만, 미분양, 미사일, 미성년, 미성년자, 미세, 미세먼지, 미시, 미제, 미주리, 미지수, 미터법, 미합중국, 밀착, 바이러스, 바이오, 바이오매스, 반복, 반영, 반지하, 발달, 발원지, 발음, 발전시설, 밧데리, 방법, 방식, 방어기제, 방촌, 배당, 배당금, 배치, 배터리, 백색가전, 백신, 백화점, 버전, 버지니아, 베, 베네룩스, 베스, 벤처, 벤치마킹, 보너스, 보유, 보증서, 보증인, 보증채무, 보통학교, 보험업법, 보호법, 복용, 본래, 본원통화, 부, 부가세, 부녀회, 부대, 부동산, 부동산업, 부분, 부양, 부양가족, 분석, 분양가 , 분양가상한제, 불 마켓, 불꽃놀이, 브라우저, 브레이크, 브렉시트, 브뤼셀, 브리튼, 블록, 블록체인, 블루칩, 비거주자, 비교, 비용, 비즈니스, 비지니스, 비트코인, 비판, 빅데이터, 사건, 사기, 사람, 사례, 사용, 사용료, 사용자, 사우디아라비아, 사유, 사이, 사찰, 사태, 사회적거리두기, 산소, 산업은행, 삼성그룹, 삼성메디슨, 삼성물산, 삼성시계, 삼성전기, 삼성전자, 삼성증권, 삼성타운, 상담, 상승, 상장지수펀드, 상태, 상품, 생각, 생산, 생존, 서비스, 서비스업, 선거인단, 선물, 선물거래, 선물환, 선진국, 선하증권, 설립, 설법, 설탕, 섭취, 성행, 세관, 세금, 세기, 세르게이, 세월, 센터, 소기업, 소득, 소득세, 소리, 소매, 소매업, 소매점, 소문자, 소비자물가지수, 소상공인, 소액결제, 소프트웨어, 손해, 송금, 수, 수동변속기, 수수료, 수원시, 수의대, 수익, 수익률, 수입, 수준, 수출, 수출규제, 수출액, 슈퍼마켓, 스마트폰, 스크립트, 스타트업, 스탠퍼드, 스탠포드, 스토리지, 스톡옵션, 스튜디오, 스트라이크, 스트레스, 스트리밍, 스티글리츠, 승승장구, 시, 시가총액, 시계열, 시공사, 시기, 시리즈, 시뮬레이션, 시스템, 시장, 시장상황, 시중, 시중금리, 신세계, 신세계백화점, 신주, 신주인수권, 실물, 실패, 싱가포르, 아르바이트, 아시아, 아이디어, 아이템, 아케이드, 아키텍처, 아파치, 아폴로, 아프리카, 아황산가스, 악화, 안구, 안드로이드, 알고리즘, 알레르기, 알루미늄, 암호, 암호학자, 암호화폐, 애서, 약리학, 약육강식, 양산, 양적완화, 어른, 어불성설, 어원, 어처구니, 언어, 업데이트, 업무협약, 에너지, 에이전트, 에피소드, 엑슨모빌, 엔진, 여파, 역할, 연대보증, 연합뉴스, 영향, 예금금리, 예금보험공사, 예불, 오늘의 시장, 오늘의 주식시장, 오사카, 오아시스, 오프라인, 온도, 온라인, 외국, 외국어, 외국인, 외국인순매도, 외국인순매수, 요구, 우리나라, 우후죽순, 운반, 워싱턴, 원달러환율, 원자재, 원천, 월, 위, 위원장, 위키백과, 위험관리, 유가, 유가상승, 유동성, 유량주, 유럽, 유발, 유일, 유흥업소, 은행, 은행 융자, 의도, 의료, 의미, 의사소통, 이, 이것, 이남신, 이념, 이득, 이미지, 이사국, 이코노미, 이탈리아, 이하, 이해, 이후, 익스플로러, 인, 인공지능, 인공호흡기, 인디펜던트, 인사말, 인삼, 인센티브, 인수, 인식, 인플레이션, 인플레이션율, 인플루엔자, 일, 일렉트릭, 일물일가, 일반인, 일본, 일일한도, 일자리, 일종, 임대료, 잉여금, 자, 자금, 자기자본(보통주), 자동, 자동변속기, 자동차, 자리매김, 자본, 자본주의, 자산, 자산관리, 자산운용, 자연환경, 자유자재, 자유주의, 자율주행, 장본인, 재개발, 재건축, 재단, 재정보증, 재택근무, 재판매, 재화, 저출산, 저희, 전기, 전기난로, 전기차, 전동차, 전망, 전매특허, 전문가, 전성시대, 전세가, 전세권, 전세금, 전세차, 전셋값, 전역, 전월세, 전자석, 전자파, 전자화폐, 전차선, 전형, 전화, 정규직, 정기예금, 정도, 정반대, 정보, 정보공개, 정부, 정신, 정책, 정책금융, 정책학, 정치, 정치학, 제권판결, 제일기획, 제품, 조, 조건, 조달, 조리, 조언, 조정, 조직, 존재, 종사자, 좌절, 주가, 주가지수, 주가지수선물, 주공, 주도, 주력, 주말, 주변, 주식, 주식거래, 주식계좌개설, 주식소개, 주식시장, 주식자보금, 주식추천, 주식회사, 주택, 줄다리기, 중, 중공업, 중국, 중국 주식, 중국어, 중독, 중소기업, 중앙은행, 중앙일보, 중진국, 중화민국, 증권, 증권가, 증권거래소, 증권시장, 증권투자기금, 증시, 지경, 지방, 지역, 지원, 지원금, 지주회사, 지지대, 지지부진, 진공청소기, 질병관리본부, 집단, 집적회로, 쪽, 차감, 차관회의, 차례, 차이점, 차익거래, 차지, 차후, 창업, 채, 채권자, 처리, 천연가스, 천연자원, 천정부지, 천하무적, 첨단, 청년, 청년유니온, 청년층, 청소년, 청약, 청약서, 체크, 최고, 최빈국, 최저임금, 최저한도, 최하, 추이, 축출, 취급, 치료제, 친환경, 침체, 카메라, 캐피털, 캔터베리, 캘리포니아, 커뮤니티, 컨설턴트, 컴퓨터, 코로나 19, 코스닥, 코스트코, 코스피, 콘스탄티노플, 콧김, 쿠로시오, 크래프트, 클라우드, 클라이언트, 키프로스, 탄소, 탄소중립, 터키, 테크놀로지, 텍사스, 텔레비전, 통화, 통화승수, 투기자, 투자, 투자와 자산관리, 투자자, 트랜스미션, 트랜잭션, 트레이딩, 트레일러, 트롤리버스, 트린, 특정, 티메로살, 파산, 파산법, 파산자, 파이낸셜, 파이널, 파키스탄, 파트너십, 판매자, 패러다임, 팬데믹, 펀드회사, 페이스북, 페인트칠, 편도, 편서풍, 편승, 편의점, 편제, 평가절하, 포름알데히드, 포트폴리오, 포함, 푸른색, 풀프레임, 프라이부르크, 프랑스, 프랜차이즈, 프랭크, 프레스콧, 프로, 프로그래밍, 프로그램, 프로세서, 프로세스, 프로세싱, 프로젝터, 프로젝트, 프로토콜, 프리랜서, 프리미어, 프리미엄, 프리우스, 플라스틱, 플래티넘, 플레이어, 피해자, 핀테크, 필두, 필라델피아, 하나, 하드웨어, 하락, 하락세, 하우스, 하이브리드, 하향, 하향조절, 학생, 한국, 한국식, 한국어, 한국은행, 한국증권금융, 한류, 한정판, 할아버지, 합장, 항목, 해결, 해킹, 행복감, 허위, 헬리콥터, 현상, 현존, 협동조합, 호텔신라, 호흡기계, 홈페이지, 확률, 환율, 환율제, 황산화물, 후, 후지필름, 후진국, 흑역사, 희소성

제 5.1.4 절 2차 스크리닝(Correlation)

고려하고 싶은 특정기간 동안의 generated keyword의 네이버 검색량과 KOSPI의 피어슨 상관관계를 계산한다. 이때 keyword 검색량이 KOSPI에 반영되는 시간 차를 고려하여 1~10의 lag를 주어 10개의 상관관계를 모두 계산하고, 10개의 상관관계 값의 절대값 중 최대값을 사용한다. 해당 값이 0.5 이상인 generated keyword만을 선택하여 2차 스크리닝 결과로 사용한다.

1, 2차 스크리닝을 통해 원하는 기간 동안의 KOSPI의 흐름과 가장 유사한 검색량 변화 양상을 가지는 KOSPI 관련 keyword를 얻을 수 있다. 표 6, 7은 실제로 특정기간을 설정해 keyword를 스크리닝한 결과이다. KOSPI가 대한민국 경제를 대표하는 가장 큰 경제지표 중 하나인 만큼 스크리닝된 keyword들은 해당 기간의 국내외 이슈와 크게 관련성이 있어 보인다. 이를 추후 보완·발전시킨다면 트렌드 분석등의 분야에서도 해당 방법론을 활용할 수 있을 것으로 생각한다.

표 6은 (2020.01.01 ~ 2020.03.01) 기간을 기준으로 했을 때의 2차 스크리닝된 상위 10개의 keyword이다. 해당 기간은 코로나-19가 발발한 시기로 발원지, 마스크, 백신, 정부 등 추출된 키워드가 해당 기간의 이슈와 크게 관련됨을 확인할 수 있다.

표 6. (2020.01.01 ~ 2020.03.01) 기간 동안의 검색량과 KOSPI의 상관관계 기준 상위 10개의 키워드

Keyword	Max_lag	Max_abs_corr
마르크	3	0.843637
수원시	1	0.811066
발원지	1	0.810843
마스크	2	0.805925
페이스북	1	0.797280
무역의존도	1	0.797144
백신	3	0.796805
정부	3	0.789430
대통령	5	0.780595
연합뉴스	3	0.780333

표 7은 (2022.01.01 ~ 2022.12.31) 기간을 기준으로 했을 때의 2차 스크리닝된 상위 10개의 keyword이다. 메타버스, 페이스북 등의 키워드는 해당기간의 메타버스이슈를, 사회적거리두기, 코로나19, 질병관리본부 등의 키워드는 해당기간의 코로나-19이슈를 반영함을 알 수 있다.

표 7. (2022.01.01 ~ 2022.12.31) 기간 동안의 검색량과 KOSPI의 상관관계 기준 상위 10개의 키워드

Keyword	Max_lag	Max_abs_corr
싱가포르	1	0.867734
페이스북	7	0.850718
인수	7	0.780638
투자	1	0.744987
메타버스	7	0.738880
사회적거리두기	2	0.738270
코로나19	10	0.734242
시리즈	1	0.731982
동화면세점	6	0.725220
질병관리본부	7	0.717162

제 5.2 절 KOSPI 예측

제 5.2.1 절 데이터셋 제작

본 연구에서는 KOSPI 예측 모형으로 LSTM을 사용하였고 LSTM 모델의 학습 및 검증에 사용되는 데이터셋을 Sliding Window 방식을 적용하여 제작하였다. Window size는 KOSPI 예측에 앞서 참고하는 데이터의 크기를 의미한다. 7일, 14일, 30일의 3가지 경우로 다양하게 실험하였다. 모델의 input은 window size일 동안의 KOSPI 종가 데이터 및 상관관계 기준 상위 top-k개 키워드의 검색량이다. 모델의 target은 다음날의 KOSPI 종가이다. 이때 모델의 KOSPI 종가 데이터와 키워드 검색량의 단위가 다르기 때문에 input과 target 모두 min-max scaling을 통해 -1 ~ 1 범위로 맞춰서 사용한다.

최근 몇 년치 데이터를 학습에 사용하는 것이 적절한가에 대한 고민이 있었다. 긴 기간의 데이터는 많은 양의 학습데이터를 확보할 수 있다는 장점이 있지만, 과거의 정보가 최근의 추세와 다를 수 있다는 단점이 있다. 이에 반해 짧은 기간의 데이터는 최근 추세위주의 학습을 할 수 있다는 장점이 있지만, 학습데이터의 양이 적다는 단점이 있다. 이에 학습데이터의 기간을 1 ~ 7년으로 달리하여 다양하게 실험을 진행하였다. 실험에서 사용한 KOSPI 종가 및 검색량 데이터는 표 8과 같다. 각 경우의 80%를 Train data로, 나머지 20%를 Validation data로 사용하였다.

표 8. 기간에 따른 7개의 학습데이터

Dataset	Period
1 year Dataset	1 year (2022.01.01 ~ 2023.01.26)
2 year Dataset	2 year (2021.01.01 ~ 2023.01.26)
3 year Dataset	3 year (2020.01.01 ~ 2023.01.26)
4 year Dataset	4 year (2019.01.01 ~ 2023.01.26)
5 year Dataset	5 year (2018.01.01 ~ 2023.01.26)
6 year Dataset	6 year (2017.01.01 ~ 2023.01.26)
7 year Dataset	7 year (2016.02.01 ~ 2023.01.26)

제 5.2.2 절 모델 학습 및 하이퍼파라미터 튜닝

다양한 하이퍼파라미터를 사용하여 데이터셋 크기(1~7년)에 따른 Validation loss(MSE) 기준의 최적 하이퍼파라미터 조합을 찾기 위한

실험을 진행하였다. 하이퍼파라미터 튜닝시에 고려한 하이퍼파라미터는 표 9와 같다.

표 9. KOSPI 예측 LSTM 모형의 하이퍼파라미터

Hyperparameter	Meaning	Values
look_back	다음날 KOSPI를 예측하기 위해 참고할 이전 데이터의 일수(=window size)	7, 14, 30
top_k	KOSPI 예측시에 사용할 상관계수 기준 상위 키워드의 개수	0, 1, 3, 5
batch_size	Batch size	16, 32
lr	Learning rate	0.1, 0.01
num_layers	LSTM 모델의 layer의 수	1, 2
h_dim	LSTM 모델의 layer의 차원크기	16, 32
epochs	Epochs	50

Grid search 방식으로 모든 하이퍼파라미터 조합을 고려해서 각 크기의(1~7년) 데이터셋 마다 196번의 실험, 총 1372번의 실험을 진행했다. Validation loss(MSE) 기준, 학습 데이터셋의 크기에 따른 최적의 하이퍼파라미터 조합은 표 10과 같다.

표 10. Validation loss(MSE) 기준, 학습 데이터셋 크기에 따른 최적의 하이퍼파라미터 조합

Dataset	look_back	top_k	batch_size	lr	num_layer	h_dim	epochs	Val loss
1 year	7	0	32	0.1	1	32	49	0.003055
2 year	7	3	32	0.01	2	32	32	0.002234
3 year	30	0	32	0.1	1	32	38	0.0008077
4 year	14	3	32	0.01	2	32	46	0.0008746
5 year	30	5	32	0.01	2	16	36	0.0009195
6 year	14	1	32	0.01	1	16	40	0.0009771
7 year	14	1	32	0.01	2	32	35	0.0009475

이후 제 5.2.4 절에서 각 기간의 데이터마다 최적 하이퍼파라미터의 조합으로 KOSPI 예측 LSTM 모델을 학습하였고, 앙상블 방법을 활용해 7개 모델의 예측값의 평균을 최종 예측값으로 사용하였다.

이 같은 방법을 사용한 이유는 (1) 각 기간별로 최적화된 모델을 학습하고, (2) 예측하고자 하는 날과 가까운 시점의 데이터에는 큰 가중치를 부여, 먼 시점의 데이터에는 상대적으로 작은 가중치를

부여하기 위해서이다.

제 5.2.3 절 결과 1 – Seed keyword vs generated keyword

앞서 제 5.1 절(KOSPI 관련 키워드 집합 확장)에서 적합한 키워드가 추출되었는지 확인해보기 위한 실험을 진행하였다. KOSPI 예측 LSTM 모델의 input으로 KOSPI 종가 데이터 및 상관관계 기준 상위 top-5개의 키워드 검색량을 사용해 다음날의 KOSPI 종가를 예측하였다. 이때 top-5개의 키워드의 후보군을 달리하여 실험을 진행했다. Case1은 seed keyword에서의 top-5개의 키워드를 사용한 경우이고, Case2는 generated keyword에서의 top-5개의 키워드를 사용한 경우이다.

top-k=5로 고정하였기 때문에 top-k를 제외한 다른 모든 하이퍼파라미터에 대해 다시 하이퍼파라미터 튜닝을 진행해 최적 세팅에서의 MSE를 비교하였다. 실험결과는 표 11에 정리되어 있다.

표 11. 학습데이터 크기에 따른 case1과 case2 에서의 MSE

Dataset size	MSE(Case 1)	MSE(Case 2)
1 year	**0.003129**	0.003218
2 year	0.002874	**0.002789**
3 year	0.0008992	**0.0008311**
4 year	0.0009015	**0.0008863**
5 year	0.0009279	**0.0009195**
6 year	0.001003	**0.0009868**
7 year	**0.000962**	0.0009678

7개의 각 연도별 경우 중 5개(2, 3, 4, 5, 6년)에서 case1 보다 case2가 더 좋은 성능을 기록하였다. Case1이 더 좋은 성능을 기록한 1년의 경우는 학습데이터의 양이 적고, 7년의 경우는 case1과 case2의 MSE 차이가 다른 년도에 비해 가장 작다는 점을 고려할 때 case2의 성능이 case1의 성능보다 좋다고 판단할 수 있다.

해당 실험결과를 통해 generated keyword가 seed keyword 대비 KOSPI 예측에 적절함을 알 수 있고, 앞선 제 5.1 절(KOSPI 관련 키워드 집합 확장)에서 새로운 KOSPI 관련 키워드 집합 확장이 잘 되었다는 정량적 평가자료로 활용할 수 있다.

제 5.2.4 절 결과 2 – Ensemble

기존 개별 모델들과 기간별 7개 모델을 앙상블한 Ensemble 모델의 성능을 비교하기 위한 실험을 진행하였다. (2022.12.08 ~ 2023.01.20)에 해당하는 30일 동안의 KOSPI를, 2022.12.08 기준 최근 1, 2, 3, 4, 5, 6, 7년 동안의 데이터를 학습데이터로 사용해 모델을 학습하여 예측하였다. 모델 학습시에 사용한 하이퍼파라미터의 값은 표 10 에서 구한 각 기간별 데이터에서의 최적 하이퍼파라미터 조합이다. 이후 7개 모델의 예측 결과를 앙상블하여 새롭게 예측하였다. 최종 8개 모델의 예측 결과는 그림 5~12 및 표 12에 정리되어 있다.

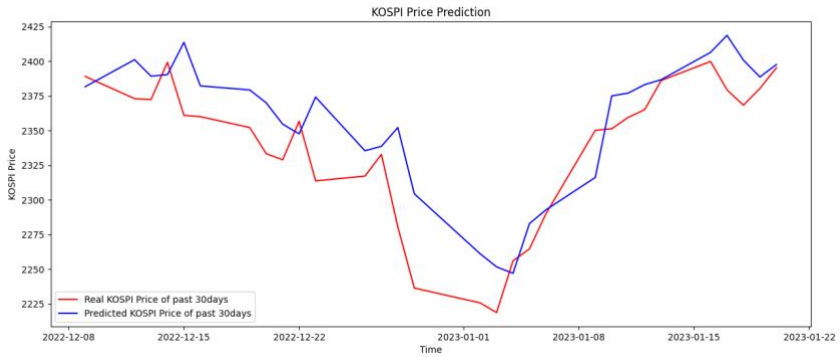

그림 5. 2022.12.08 기준 최근 1년의 데이터를 학습데이터로 사용했을 때의 (2022.12.08 ~ 2023.01.20)에 해당하는 KOSPI 예측값과 실제값

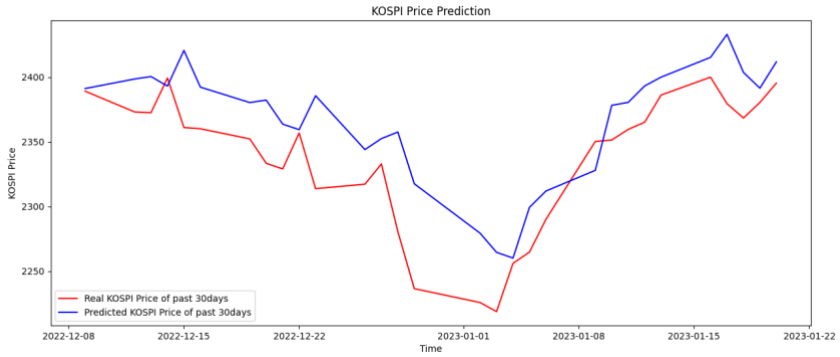

그림 6. 2022.12.08 기준 최근 2년의 데이터를 학습데이터로 사용했을 때의 (2022.12.08 ~ 2023.01.20)에 해당하는 KOSPI 예측값과 실제값

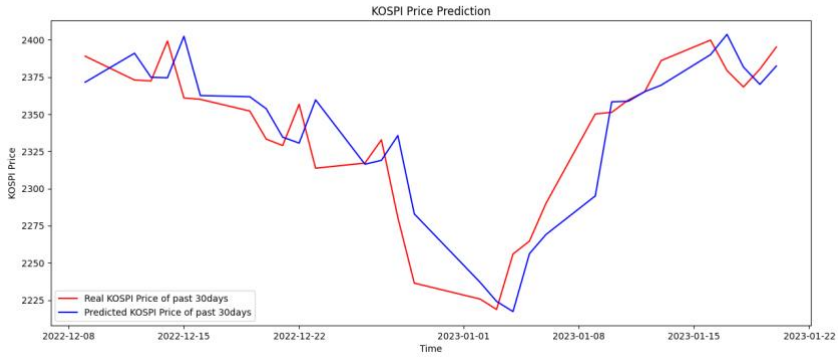

그림 7. 2022.12.08 기준 최근 3년의 데이터를 학습데이터로 사용했을 때의 (2022.12.08 ~ 2023.01.20)에 해당하는 KOSPI 예측값과 실제값

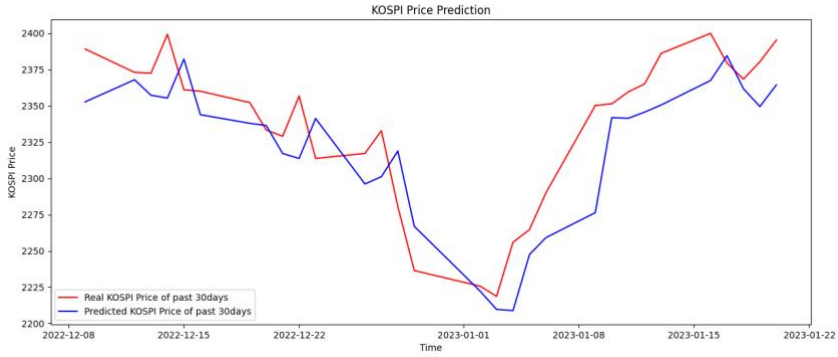

그림 8. 2022.12.08 기준 최근 4년의 데이터를 학습데이터로 사용했을 때의 (2022.12.08 ~ 2023.01.20)에 해당하는 KOSPI 예측값과 실제값

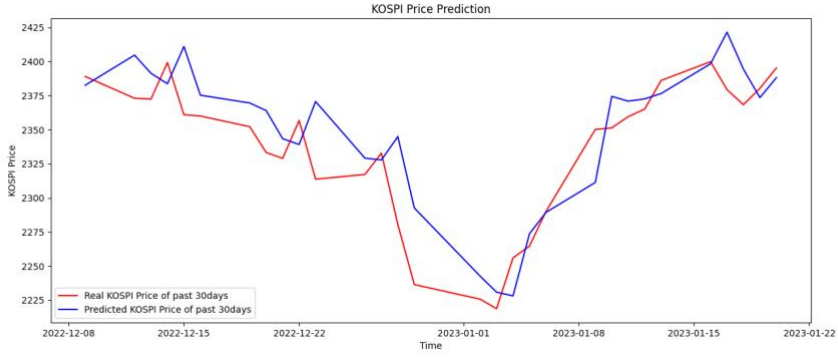

그림 9. 2022.12.08 기준 최근 5년의 데이터를 학습데이터로 사용했을 때의 (2022.12.08 ~ 2023.01.20)에 해당하는 KOSPI 예측값과 실제값

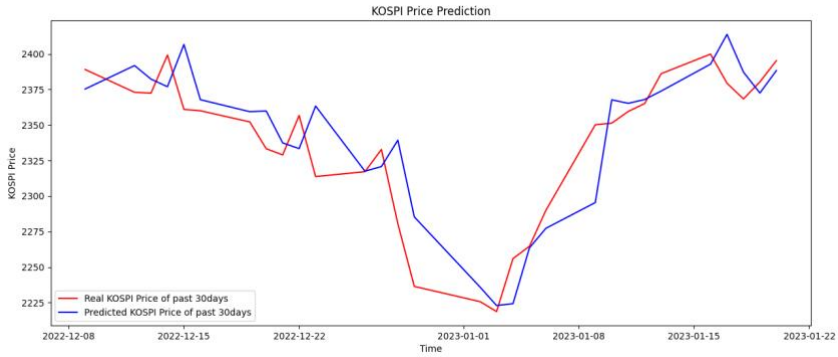

그림 10. 2022.12.08 기준 최근 6년의 데이터를 학습데이터로 사용했을 때의 (2022.12.08 ~ 2023.01.20)에 해당하는 KOSPI 예측값과 실제값

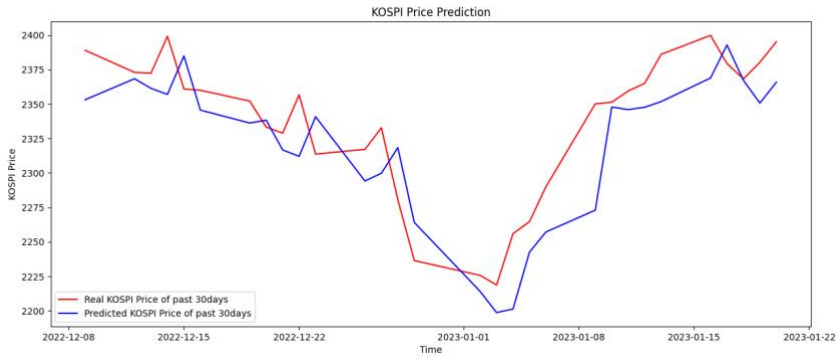

그림 11. 2022.12.08 기준 최근 7년의 데이터를 학습데이터로 사용했을 때의 (2022.12.08 ~ 2023.01.20)에 해당하는 KOSPI 예측값과 실제값

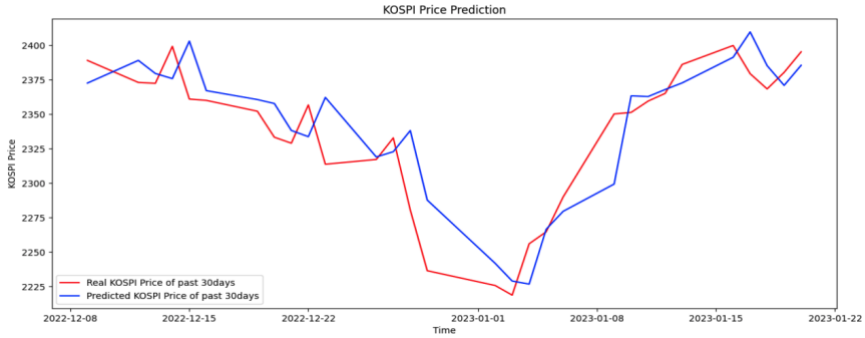

그림 12. 7개 모델을 앙상블 했을 때의 (2022.12.08 ~ 2023.01.20)에 해당하는 KOSPI 예측값과 실제값

표 12. 학습데이터의 크기를 1~7년으로 다르게 했을 때와 이들을 앙상블 했을 때의 KOSPI 예측 LSTM 모형의 성능(MSE)

Dataset size	MSE
1 year	963.8987
2 year	1441.0094
3 year	612.8418
4 year	838.3274
5 year	767.7130
6 year	652.3897
7 year	886.2186
Ensemble	**612.4413**

각 년도마다 해당기간 동안의 최대, 최소값이 달라지기 때문에 동일한 값이라도 min-max scaling 이후의 결과가 달라진다. 따라서 각 년도 및 Ensemble 모델의 예측결과를 원활하게 비교하기 위해, 원래의 KOSPI 단위로 scaling back 한 이후의 수치를 사용하여 MSE를 계산하였다. 따라서 이전에 비해 MSE값의 단위가 커졌다.

7개 모델을 앙상블 했을 때의 MSE값은 612.4413으로 7개의 개별 모델 대비 향상된 예측 성능을 기록하였다.

앙상블 방법을 통해 예측하고자 하는 날과 가까운 시점의 데이터에는 큰 가중치를 주고, 먼 시점의 데이터에는 상대적으로 작은 가중치를 주는 셈이 되었는데, 해당 방법이 KOPSI 예측에 효과적임을 확인할 수 있다.

제 6 장 결 론

본 논문은 KLUE/BERT, KPF/RoBERTa, KB/ALBERT의 사전 학습 언어모델을 활용하여 KOSPI 관련 키워드집합을 확장했고, 그 결과 828개의 Generated keyword를 생성하였다. 이후 확장된 키워드 집합에서 해당 검색량과 KOSPI 가격의 상관관계가 큰 키워드들을 뽑아서 그 검색량을 KOSPI 가격 데이터와 함께 KOSPI 예측 LSTM 모델의 학습데이터로 사용하였다. 또하 다양한 하이퍼파라미터 조합 실험과 앙상블 방법론을 활용하여 KOSPI 예측 LSTM 모델 최적화를 이루었다.

본 논문의 다음과 같은 의의를 가진다. 첫번째는 사전 학습 언어모델 및 NLP 키워드 추출 태스크를 이용하여 KOSPI 관련 키워드의 집합을 확장시킨 점이다. 이는 문헌이나 직관에 주로 의존했던 과거와 달리 문맥 및 의미적인 부분을 고려한 방법이고 생각치 못한 키워드를 찾을 수 있으며, 프로세스를 자동화했다는 점에서 의의가 있다. 두번째는 키워드의 검색량을 KOSPI 예측에 활용했다는 점이다. 인터넷 검색량은 대중들의 관심의 정도를 나타내는 지표로써 해석이 가능한 만큼 행동경제학적 관점에서 대중의 심리를 KOSPI 예측에 반영했다는 의의가 있다. 마지막으로는 새로운 분야에서의 활용 가능성이다. KOSPI는 대한민국 경제를 대표하는 가장 큰 경제지표 중 하나로 국내외의 시장 상황이 반영되어 있다. 2차 스크리닝 단계에서 기간을 특정하였을 때 해당 기간에 이슈가 되거나 중요했던 키워드들이 추출되는 것을 확인해 볼 수 있다. 이를 보완·발전시킨다면 기간별 경제 트렌드 분석 등의 분야에서도 활용할 수 있을 것으로 기대한다.

한계점으로는 다음이 있다. 첫번째로 키워드 추출을 태스크로 하는 Fine-tuning을 통해서 KOSPI 관련 키워드 집합 확장을 시도해보려 했지만 레이블링된 데이터셋의 부재로 수행하지 못한점이다. 본 논문에서는 대신해서 seed key, text, candidate key의 임베딩 벡터들간의 코사인 유사도를 기준으로 키워드 추출 및 스크리닝 작업을 수행하였다. 추후에 적합한 데이터를 찾거나 직접 데이터셋을 제작하여 비교 실험을 해보려고 한다. 또한 코사인 유사도 외에 다양한 거리 metric을 활용해 볼 수 있을 것이다. 두번째는 키워드의 검색량과 KOSPI의 관련성을 서열화할 때 그 기준으로 상관관계를 사용했다는 점이다. 대중의 심리가 가격에 영향을 미친다는 행동경제학적인 관점에서 보았을 때 단순 상관관계보다 Granger causality 분석 등을 통한 인과관계를 고려해 키워드 추출을 해볼 수도 있을 것이다. 마지막으로 KOSPI 예측 LSTM 모델 학습시에 입력변수로 키워드 검색량과 과거 KOSPI 데이터만을 사용했다는 점이다. 실험의 편의 및 NLP 키워드 추출 태스크를 통해 새롭게 추출된 키워드의 성능을 확인하기 위해 위의 세팅으로 실험을 진행했다. 거시경제지표를 비롯한 다양한 변수를 후보로 두어 학습에 사용한다면 더 향상된 KOSPI 예측 성능을 얻을 수 있을 것이다.

참고 문헌

[1] 김동규, 이동욱, 박장원, 오성우, 권성준, 이인용, & 최동원. (2022). KB-BERT: 금융 특화 한국어 사전학습 언어모델과 그 응용. *지능정보연구*, *28*(2), 191-206.

[2] Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The journal of Finance*, *61*(4), 1645-1680.

[3] Cooper, C. P., Mallon, K. P., Leadbetter, S., Pollack, L. A., & Peipins, L. A. (2005). Cancer Internet search activity on a maj or search engine, United States 2001-2003. *Journal of medica l Internet research*, *7*(3), e36.

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Be rt: Pre-training of deep bidirectional transformers for languag e understanding. *arXiv preprint arXiv:1810.04805*.

[5] Ettredge, M., Gerdes, J., & Karuga, G. (2005). Using web-bas ed search data to predict macroeconomic statistics. *Communic ations of the ACM*, *48*(11), 87-92.

[6] Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolin ski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012-101 4.

[7] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term me mory. *Neural computation*, *9*(8), 1735-1780.ISO 690

[8] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & So ricut, R. (2019). Albert: A lite bert for self-supervised learnin g of language representations. *arXiv preprint arXiv:1909.11942* .ISO 690

[9] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

[10] Park, S., Moon, J., Kim, S., Cho, W. I., Han, J., Park, J., ... & Cho, K. (2021). Klue: Korean language understanding evaluation. *arXiv*

preprint arXiv:2105.09680.ISO 690

[11] Polgreen, P. M., Chen, Y., Pennock, D. M., Nelson, F. D., & Weinstein, R. A. (2008). Using internet searches for influenza surveillance. *Clinical infectious diseases*, *47*(11), 1443-1448.

[12] Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying t rading behavior in financial markets using Google Trends. *Sci entific reports*, *3*(1), 1684.

[13] Tang, X., Lei, N., Dong, M., & Ma, D. (2022). Stock Price Pr ediction Based on Natural Language Processing. *Complexity*, *2 022*.ISO 690

Abstract

KOSPI Keyword Set Expansion using NLP Model & KOSPI Prediction using Keyword Search Index

Woojin Ko
The Department of Statistics
The Graduate School
Seoul National University

This paper consists of two study topics.

First, we study the expansion of KOSPI keyword set using NLP model. Starting from KOSPI related 159 original seed keywords, we collect text and candidate keywords. Then using NLP models KLUE-RoBERTa/large, KPF-BERT and KB-ALBERT, we vectorize seed keywords, text and candidate keywords. Calculating cosine similarity between embedding vectors, we get similarity/importance score and by utilizing these scores at 1st screening we extract 828 generated keywords. Also, by proceeding 2nd screening on the basis of correlation between search index of generated keywords and KOSPI, we get keyword set highly related with KOSPI.

Next, we study KOSPI prediction using search index of KOSPI related keywords. As a prediction model we use LSTM. We conduct experiments in various length setting of training data from 1 year to 7 year and try to find the best combination of hyperparameter for each length of data. As a result of experiment, we could find that KOSPI prediction performance of using search index of newly generated keywords is better than using search index of seed keywords, which means keyword set expansion task was well carried out. In addition, an ensemble model which combined 7 different models for each length of training data shows much improved prediction performance. It implies giving a large weight to the data which is close to the day we want to predict is useful.

Keywords : KOSPI, NLP, Keyword Extraction, Stock Prediction, Search Index, Deep Learning
Student Number : 2020-27834