한국경제학보 제29권 제2호
The Korean Journal of Economics Vol. 29, No. 2 (Autumn 2022)

Machine Learning과 Google Trends Data를 이용한 유가 예측 및 분석*

김 선 미 · **조 두 연*****

요 약

유가의 등락은 에너지 수입국인 우리나라에 있어 매우 중요한 사안 중 하나이며, 유가의 변동이 한국경제에 미치는 부정적인 영향을 줄이고, 안정적인 운영을 위해 유가를 예측하는 것은 중요하다. 본 연구를 통해 거시경제변수 외에도 웹 검색어 기반의 Google Trends Data를 이용하여 유가 등락에 영향을 주는 요인을 분석하고자 한다. WTI 유가 예측 모형에서 석유 수요 및 공급 관련 설명변수를 활용한 모형의 예측력과 유가 하락기 동안 빈도가 높은 단어의 검색어 추세 변화량을 추가한 모형의 예측력을 비교하였을 때, 검색어 추세를 추가한 모형의 예측력이 개선되는지를 분석하였다. 본 연구에서는 2004년 1월부터 2020년 12월까지의 데이터를 이용하여 WTI 유가 예측에 영향을 주는 석유 공급 및 수요 변수 이외에 Google Trends 검색어 추세를 추가함으로써 예측력을 높일 수 있음을 보였다. WTI 유가 예측 모형의 예측력을 비교하기 위해 Adaptive LASSO, Ridge Regression, Random Forest 모형 이외에 최근 가격 예측모형에서 많이 활용되고 있는 LSTM 알고리즘을 적용한 결과, 정형데이터만 이용한 모형의 예측력에 비하여 Google Trends Data를 함께 이용한 모형의 예측력이 개선된다는 점을 보였다.

주제분류 : B030104, B030109
핵심 주제어 : Google Trends Data, LSTM 모형, 유가 예측, Web scraping

* 본 논문의 개선을 위해 유익한 논평을 해 주신 두 분의 익명의 심사자들에게 감사드린다.
** 제1저자, 성균관대학교 경제대학 퀀트응용경제학과 석사, e-mail: ssmkim912@gmail.com
*** 교신저자, 성균관대학교 경제대학 경제학과/퀀트응용경제학과 부교수, e-mail: dooyeoncho@g.skku.edu

Ⅰ. 서 론

유가의 등락은 에너지 수입국인 우리나라 이외에도 세계 경제에 많은 영향을 주는 것으로 알려져 있다. 2008년 미국 서브프라임 모기지, 2020년 초에 시작된 코로나19 등의 영향으로 국제 유가는 하락하였다가 최근 다시 상승하였다. 유가 하락이 지속되면 에너지 수입국의 경우, 에너지 수입액 감소로 인하여 경상수지 흑자폭이 증가한다. 반면에 에너지 수출국의 경우 교역조건이 악화되어 경상수지 적자폭이 증가하며, 실물경기 둔화 등으로 경제기초여건이 크게 악화된다. 유가의 등락은 에너지 수입국과 수출국의 경제에 모두 영향을 주기 때문에, 유가 예측에 대한 지속적인 연구가 이루어져 왔다.

그러나, 유가 등락에 영향을 미치는 요인 중에는 OPEC국가 간의 갈등, 중동지역의 분쟁, 코로나19와 같은 예측하기 어려운 외부요인들의 영향이 크게 작용하게 된다. 과거에는 시계열 데이터인 유가를 예측하기 위해 ARIMA 혹은 GARCH 모형 등 시계열 분석 위주로 연구가 이루어졌지만, 최근 들어 기계학습(Machine Learning)기법 기반으로 다양한 외부요인들을 설명할 수 있는 변수를 모형의 Feature로 활용하고 있다. 예를 들어, 송경재・양회민(2009)은 WTI 유가 예측 모형으로 ARIMA 모형을 적용하여 1984~2004년까지 총 84분기의 WTI 유가 가격을 이용하여 유가의 추세를 예측하였다. 예측 정합성 결과, 3년(12분기) 동안 9분기의 추세를 정확히 예측하였고, 예측값과 실제값의 오차는 1.83~18.07%로 평균적으로 10%의 예측 오차가 발생하였다. 하지만, 시계열 분석의 특성상 단일변수에 의한 연구로서 유가에 영향을 미치는 변수에 대한 영향도는 측정이 불가능하다는 한계점이 존재한다.

박강희・신현정(2011)은 시계열 데이터인 유가를 벡터형태로 전환한 후, Semi-Supervised Learning(SSL)을 적용하여 다양한 요인들을 설명변수로 사용하여 유가를 예측하였다. 유가는 특성상 다른 경제 지표 간의 상호작용이 유가 등락에 영향을 받지만, ANN, SVM과 같은 방법론은 입력변수 간의 상호작용을 분석하기는 어려움이 존재한다. 1992년 1월부터 2008년 7월까지의 WTI 유가를 월별 데이터로 변환한 후 사용하였으며, 설명변수로는 세계 석유 총 생산량, OPEC 석유 생산량, 사우디아라비아

석유 생산량 등 공급 관련 변수와 세계 석유 총 수요량, OECD 석유 수요량, Non-OECD 석유 수요량 등 수요 관련 변수 등을 사용하였다. SSL은 입력 값으로 벡터 타입의 변수를 사용해야 하기 때문에, 시계열 형태의 유가 데이터를 벡터 타입으로 변환하여 사용한다. 벡터타입으로 변환할 때, 파라미터 값의 개수에 따라 고차원 문제를 유발하거나 모델의 과적합이 발생할 수 있기 때문에 추정에 유의해야 한다.

또한, 한동우(2019)는 Google Trends Data를 활용하여 유가의 변화량과 검색어 추세와의 유의성을 확인하여 유가를 예측하는데 키워드 추세가 어떠한 영향을 주는지에 대해 분석하였다. 2014년 8월 1일부터 2014년 11월 28일까지를 유가 하락기로 2016년 1월 4일부터 2016년 4월 29일까지를 유가 상승기로 분류한 뒤, 두바이 유가 변화량과 유가와 밀접한 관계가 있는 수요와 공급, 지정학적 요인, 금융, 기타 4개의 키워드 카테고리를 선정하여 총 12개의 키워드 검색 추세간의 관계를 분석하였다. 또한, 검색어 별 검색 추세와 두바이 유가 변화량과의 관계를 상관관계분석을 통해 파악하고, 단순회귀분석을 통하여 검색추세와 두바이 유가 변화량과의 유의성을 확인하였다. 유가 상승기와 유가 하락기의 변화량과 12개의 키워드 검색 추세 간 단순회귀분석 결과, 모형의 설명력이 다소 낮았으며 총 24개의 모형 중 5개의 모형만 유의하였다. 따라서, 본 연구에서는 키워드 검색 추세가 유가를 예측하는데 어떠한 영향을 주는지를 알아보고자 하며, 유가를 예측할 때, 거시변수 이외에 유가 하락기 동안 높은 빈도를 보이는 단어를 선정하여 검색 추세를 WTI 유가 예측모형의 설명변수로 활용하고자 한다.

기존 연구에서는 유가 예측 모형에 ARIMA 모형과 단순회귀모형을 적용하여 종속변수인 유가 이외에 영향을 주는 다양한 요인들을 설명변수로 활용하지 못하기 때문에 유가에 영향을 미치는 변수에 대한 영향도 측정이 불가능하다는 한계점이 있었다. 이에 따라, 본 연구에서는 이를 보완하고, 유가 예측 모형의 예측력을 향상시키기 위하여 WTI 유가에 영향을 줄 수 있는 석유 공급 및 수요 데이터 이외에도 유가 하락기의 신문기사에서 추출한 단어의 Google Trends Data를 설명변수로 이용하여 유가를 예측하고 분석하였다.

국제 에너지 시장은 2000년대 신흥국의 수요증가와 경기회복으로 인하여 가격이 폭등하였으나, 중국의 성장 둔화 및 대체 에너지 개발, 2020년

초에 시작된 코로나19로 인해 유가가 하락하는 추세를 보였다. 유가의 등락은 에너지 수입국인 우리나라에 있어 매우 중요한 사안 중 하나이며, 유가의 변동이 한국경제에 미치는 부정적인 영향을 줄이고, 안정적으로 운영될 수 있도록 유가를 예측하는 것이 필요한 실정이다. 따라서, 본 연구를 통해 거시변수 외에도 웹 검색어 기반의 Google Trends Data를 이용하여 유가 등락에 영향을 주는 요인을 도출하고자 한다. 즉, WTI 유가 예측 모형에서 석유 공급, 수요 관련 설명변수를 활용한 모형의 예측력과 유가 하락기 동안 빈도가 높은 단어의 검색어 추세 변화량을 추가한 모형의 예측력을 비교하였을 때, 검색어 추세를 추가한 모형의 예측력이 개선되는지를 분석하였다.

본 논문의 구성은 다음과 같다. 제Ⅱ장에서는 실증모형과 데이터를 제시하고, 제Ⅲ장에서는 Adaptive LASSO, Ridge Regression, Random Forest, LSTM 등 다양한 예측모형을 이용하여 추정한 결과를 제시한 뒤 예측 결과를 분석한다. 마지막으로, 제Ⅳ장에서는 결론과 본 연구를 통하여 도출한 시사점을 제시한다.

Ⅱ. 실증모형 및 데이터

1. 연구방법 및 실증모형

(1) Adaptive LASSO 모형

LASSO Regression 모형은 최소제곱추정법과 달리 약간의 bias를 허용함으로써 분산을 감소시키고, 예측 정확도를 높일 수 있으며, 변수 선택의 목적까지 달성할 수 있다. 하지만, bias가 존재함으로써 변수 선택 일치성이 충족되지 않을 수 있는 단점이 존재한다. Adaptive LASSO는 LASSO의 추정량의 bias를 줄이기 위한 방법으로 식은 다음과 같이 주어진다.

$$Q(\beta|X,y,w) = \frac{1}{2}\|y - X\beta\|^2 + \lambda\sum_j w_j|\beta_j|,\ where\ w_j = |\widetilde{\beta}_j|^{-1}$$

회귀계수의 초기추정치의 역수인 가중치 벡터 ω의 크기가 0에 가까울 때는 penalty를 많이 주고, 0이 아닌 회귀계수에 대해서는 penalty를 적게 준다. Adaptive LASSO 모형의 장점은 ① LASSO보다 더 적은 수의 변수로 더 작은 MSE가 산출되어 예측력이 개선되며, ② Training 자료가 충분하고, 비교적 Noise가 적을 때 정확하게 변수를 선택한다는 점이다.

(2) Ridge Regression 모형

Ridge Regression 모형은 최소제곱법과 매우 유사하지만, 각 계수의 곱을 더한 값을 식에 포함하여 계수의 크기를 최소화할 수 있도록 설계된 모형이다. 아래 식과 같이 조절 모수인 λ가 회귀추정치와 관련된 두 항의 영향을 조절해주는 역할을 하게 된다.

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=RSS+\lambda\sum_{j=1}^{p}\beta_j^2$$

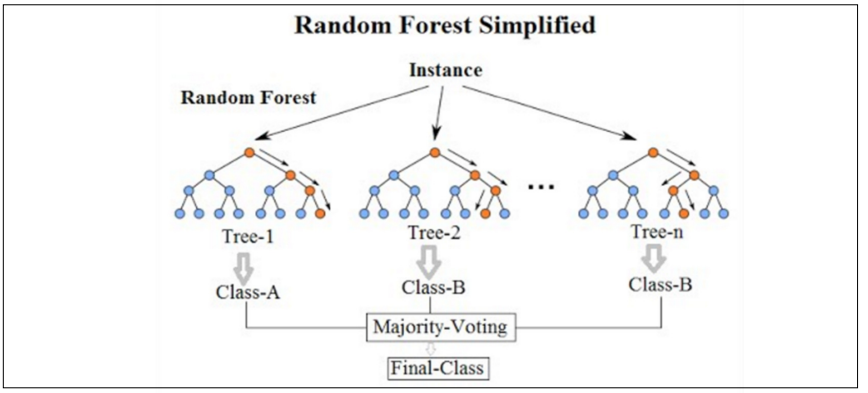
즉, $\lambda=0$ 일 때, penalty항은 효과가 없으므로, Ridge Regression 모형은 최소제곱추정치를 생성하게 되며, $\lambda=\infty$ 일 때, shrinkage penalty의 영향이 커지고, Ridge Regression 계수 추정치는 0에 가까워진다. Ridge Regression 모형의 장점은 오차를 최소화하는 함수에 penalty를 줌으로써 보다 smooth하게 계수를 선택할 수 있다는 점이다.

(3) Random Forest 모형

Random Forest 모형은 Decision Tree의 분류보다 정확도를 개선시키기 위해, 여러 개의 Tree를 생성하여 각각 Tree의 예측을 총 조합하여 결론을 내리는 구조를 가지고 있다. 아래 〈그림 1〉은 Random Forest 모형의 구조를 보여주고 있다.

내부적으로 랜덤 샘플링을 수행하여 각 Tree마다 훈련시키는 데이터가 다르게 적용되며, Tree의 개수가 많을수록 정확도는 개선되지만, 기울기가 급변하는 지점이 존재하게 된다.

〈그림 1〉 Random Forest 구조(Random Forest structure)

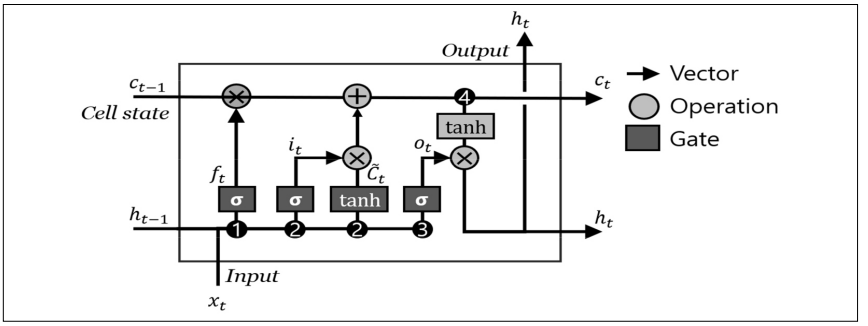


(4) Long-Short Term Memory(LSTM) 모형

기존의 Recurrent Neural Networks(RNN) 알고리즘에서 발생하는 문제인 기울기 소실 문제(Gradient Vanishing Problem)를 제어하기 위한 방법으로 제시되었다. 기울기 소실 문제란, 은닉층이 많은 다층 퍼셉트론에서, 은닉층을 많이 거칠수록 전달되는 오차가 크게 줄어들어 학습이 되지 않는 현상을 의미하며, LSTM 알고리즘은 〈그림 2〉와 같이 4개의 상호작용 계층과 셀 스테이트(Cell State)라 불리는 컨베이어 벨트를 사용하며, 각 라인은 벡터로 구성된다. 아래 〈그림 2〉는 LSTM 모형의 데이터 처리 알고리즘을 보여주고 있다.

LSTM 모형의 장점은 RNN 알고리즘보다 장기간의 정보를 기억함으로써, 더 정교한 예측을 수행할 수 있다는 점이라고 할 수 있다.

〈그림 2〉 LSTM의 데이터 처리 알고리즘(Data processing algorithm for LSTM)



2. 데이터

본 연구에서는 2004년 1월부터 2020년 12월까지의 WTI 유가의 일별 시계열 자료를 종속변수로 사용하였다. 국제원유시장에서 생산지에 따라 WTI유, Brent유, 두바이유로 분류되며, 이 중 WTI는 서부 텍사스 중질유로, 현물 및 선물로 거래되며 북미 전역에 영향을 주는 지표이며, 미국 내 뉴욕상업거래소(NYMEX)에서 1983년부터 원유선물을 도입하였다.

유가와 연관되어 있는 변수를 찾기 위해 유가와 관련된 여러 참고문헌을 참조하여 주요 나라의 석유 총 생산량과 미국 내 석탄 생산량/소비량/수출량/수입량 등 총 10개의 데이터를 수집하였다. 아래 〈표 1〉은 WTI 유가 예측을 위해 수집한 변수 리스트, 주기 및 출처를 보여준다.

〈표 1〉 WTI 유가 예측을 수행하기 위해 수집한 변수 리스트(List of variables and data sources used in analysis to forecast WTI crude oil price)

NO	데이터명	주기	출처
1	WTI 유가	일별	Opinet
2	중국 석유 총 생산량	월별	EIA
3	인도 석유 총 생산량	월별	EIA
4	러시아 석유 총 생산량	월별	EIA
5	사우디아라비아 석유 총 생산량	월별	EIA
6	미국 석탄 생산량	월별	EIA
7	미국 석탄 소비량	월별	EIA
8	미국 석탄 수출량	월별	EIA
9	미국 석탄 수입량	월별	EIA
10	원/달러 환율	일별	NAVER 금융
11	Google Trends Data - 'oil'	월별	Google
12	Google Trends Data - 'crude'	월별	Google

주: EIA는 Energy Information Administration 을 의미함.
Note: EIA stands for Energy Information Administration.

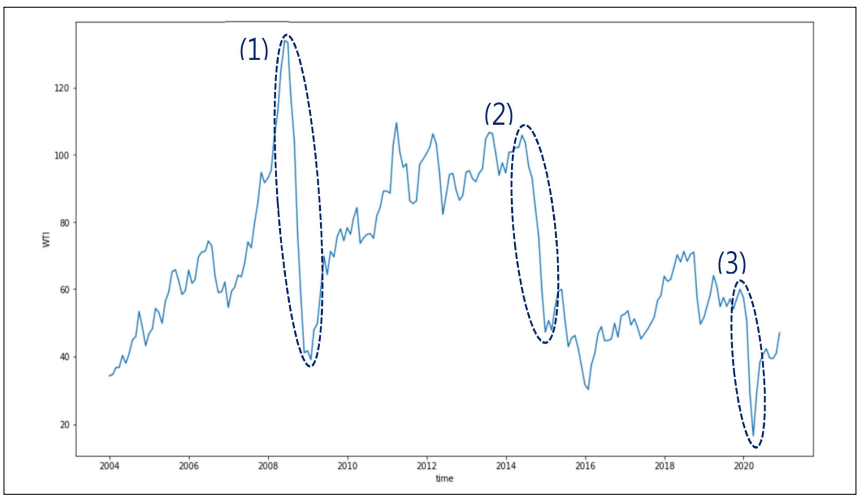
종속변수인 WTI 유가의 데이터 주기는 일별이지만, 설명변수로 활용할 변수들의 주기가 월별이기 때문에, 데이터 주기가 일별인 데이터를 월별 평균값으로 계산하여 월별 데이터로 변환하여 사용하였다. 또한, 유가 관련 검색 키워드 추세를 지수화한 Google Trends Data를 변수로 활용하였다. Google Trends Data는 주어진 기간 동안 Google의 검색 엔진에 입력하

는 검색어에 대한 전체적인 볼륨 및 검색어 추세를 볼 수 있는 서비스로 다음과 같은 특징이 있다.

① 원하는 검색어, 검색 기간 설정
② 검색 주기에 해당하는 검색 추이
③ 지역별 검색량 비교
④ 검색어와 관련 있는 주제 및 검색어 확인

본 연구에서는 Google Trends Data를 이용하여 경제변수 이외의 키워드 검색 추세가 유가를 예측하는 데 도움이 되는지를 분석하였다. 〈그림 3〉은 최근 WTI 유가 추이를 나타내고 있으며, 2004년 이후 WTI 유가의 가격 하락이 60%이상인 경우가 세 차례 발생하였음을 알 수 있다.

〈그림 3〉 최근 WTI 유가 추이(Recent trend for WTI crude oil price)



첫번째 하락기는 2008년 7월부터 12월까지에 해당하는 시기로, 유가가 약 68.9% 하락하였으며, 이 시기에는 미국의 금융위기 여파로 세계 경기가 악화되어 석유 수요가 감소하였다. 두번째 하락기는 2014년 6월부터 2016년 1월까지에 해당하는 시기로, 유가가 70.0% 하락하였으며, 중동 산유국을 중심으로 한 공급 과잉이 원인이었다. 또한, 글로벌 경제 성장 둔화로 주요국의 원유 수요가 이전 수준에서 정체되었던 시기였다. 세번째 하

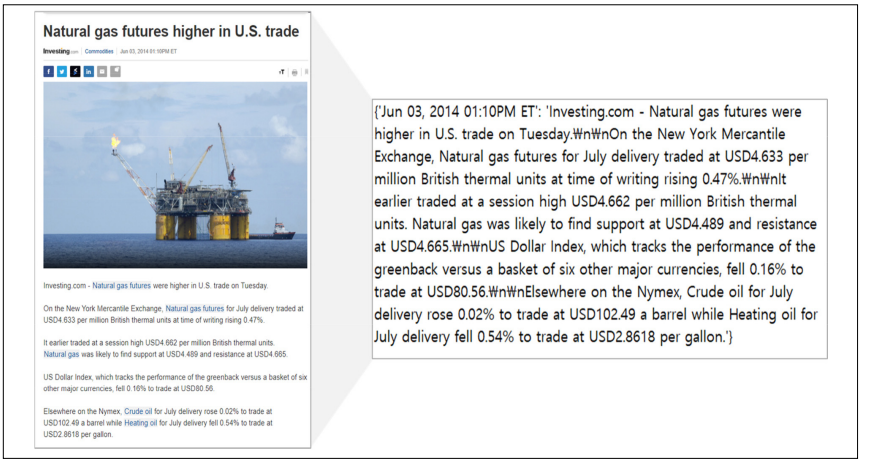
락기는 2020년 1월부터 4월까지에 해당하는 기간으로, 유가가 84.0% 하락하였으며, 코로나 19로 인하여 에너지 소비를 수반하는 경제/사회 활동을 위축시켜 국제 유가 하락 및 에너지 수요 감소를 초래하였다. 〈표 2〉는 유가가 60% 이상 하락하였던 시기에 대한 WTI 유가 변동률을 보여주고 있다.

〈표 2〉 WTI 유가 변동률(Changes in WTI crude oil prices)

	2008.7~2008.12		2014.6~2016.1		2020.1~2020.4	
	유가($)	변동률(%)	유가($)	변동률(%)	유가($)	변동률(%)
WTI	133.4→41.4	-68.9	105.2→31.5	-70.0	62.7→31.5	-84.0

본 연구에서는 2014년 6월부터 2016년 1월까지의 WTI 관련 신문기사를 수집한 후, 유가 하락 기간 동안 가장 빈번하게 나온 단어를 추출하여 Google Trends Data의 검색어 추이와 WTI 유가와의 영향도를 파악하고자 한다. 〈그림 4〉에 제시된 바와 같이 Web scraping을 이용하여 WTI 유가 관련 신문기사를 총 5,320개 수집하였으며, 유가 관련 사이트인 Investing.com에 WTI 관련 신문기사만 별도로 모아 놓은 커뮤니티에 기재된 신문기사를 대상으로 수집하였다.[1)]

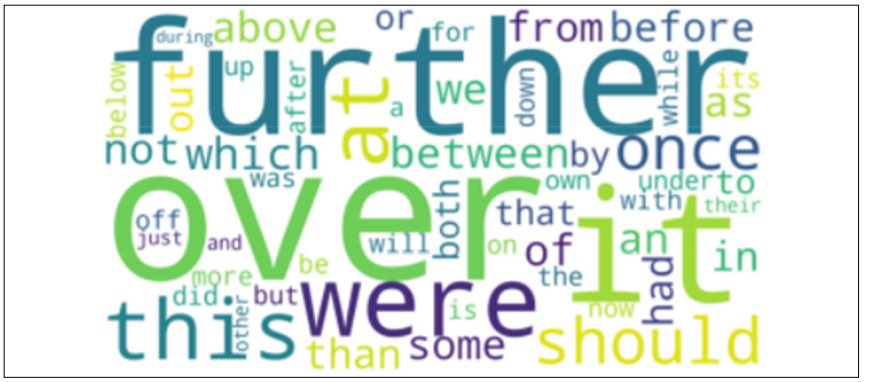
〈그림 4〉 Web scraping 결과 예시(Example of web scraping)



1) www.investing.com/commodities/crude-oil.

수집된 기사 중 핵심 단어를 추출하기 위해 텍스트 분석을 통해 토큰(token) 단위로 분리한 후, Natural Language Toolkit(NLTK)이 정의한 영어 불용어(Stop word) 리스트를 사용하여 불용어를 제거하였다. 〈그림 5〉는 불용어를 제거한 예시를 보여주고 있으며, 〈그림 6〉은 불용어가 제거된 단어들 중 명사만 추출하여 제시하고 있다.

〈그림 5〉 불용어 제거 예시(Example of stop words removal)



〈그림 6〉 명사 추출 예시(Example of noun phrase extraction)

아래 〈표 3〉은 추출된 명사 중 상위 3개의 최빈 단어와 빈도를 보여주고 있으며, 'oil'이 20,549회로 가장 높은 빈도수를 보여주고 있으며, 그 다음으로 'crude'가 9,586회로 두 번째로 높은 빈도수를 보여주고 있음을 알 수 있다.

〈표 3〉 최빈 단어 추출(Top3)(High frequency words extraction (Top3))

순위	단어	빈도
1	oil	20,549
2	crude	9,586
3	futures	8,624

유가 하락기의 신문기사를 수집하여 추출한 단어들 가운데 빈도가 가장 높은 두 단어('oil'과 'crude')를 구글의 검색 키워드 추세를 지수화한 Google Trend Data로 검색하여 유가를 예측하기 위한 설명변수로 사용하여 유가의 예측력을 개선할 수 있는지 분석하였다.

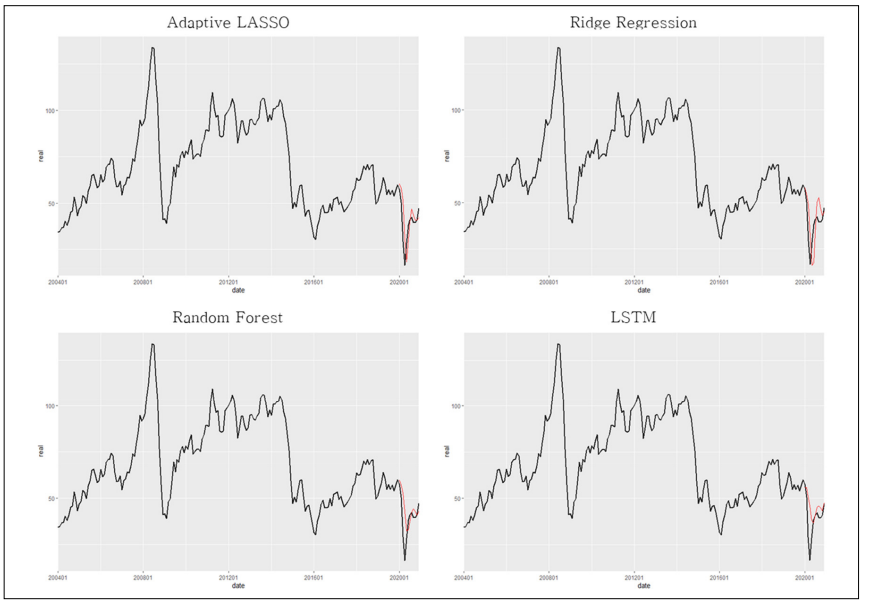
Ⅲ. 연구결과

1.정형데이터만 사용한 유가 예측 결과

유가 예측 모형의 예측력을 정교하게 비교하고자 총 4개의 예측 모형을 적용하여 모형의 Root Mean Squared Error(RMSE)를 비교하였다. 본 연구에 사용한 데이터는 2004년 1월부터 2019년 12월까지의 월별 데이터를 학습데이터(Training Set)로 사용하였고, 2020년 1월부터 2020년 12월까지의 데이터를 테스트데이터(Test Set)로 사용하였다. 연구 가설을 검증하기 위해 정형데이터를 활용하였을 때와 유가 하락기 동안 수집된 신문기사 중 가장 빈도가 높은 단어의 검색어 추세를 변수로 추가하였을 때를 각각 Adaptive LASSO 모형, Ridge Regression, Random Forest 모형과 최근 가격 예측 모형에서 많이 활용되고 있는 LSTM 알고리즘을 적용하여 모형의 예측력 비교하였다.[2] 석유 공급 및 수요 관련 정형데이터를 활용한 유가모형 예측 결과는 다음 〈그림 7〉과 같다.

2) 정형데이터는 〈표 1〉에 제시된 바와 같이, 종속변수로 1) WTI 유가를 이용하였으며, 설명변수로 2) 중국 석유 총 생산량 - 10) 원/달러 환율에 해당하는 데이터를 이용하여 모형을 추정하였다.

〈그림 7〉 정형데이터만 사용하여 추정한 예측 모형 결과(Estimation results from forecasting models using structured data only)



추정 결과, Adaptive LASSO 모형의 RMSE는 9.160이며, Ridge Regression 모형의 RMSE는 9.753로 나타났다. 또한, Random Forest 모형의 RMSE는 10.796, LSTM 모형의 RMSE는 11.339로 추정되었다. 추정 결과를 정리하자면, 석유 공급 및 수요 관련 정형데이터를 활용한 유가모형 중 Adaptive LASSO 모형의 예측력이 가장 좋은 것으로 나타났으며, 그 다음으로 Ridge Regression 모형, Random Forest 모형, 마지막으로 LSTM 모형의 순으로 예측력이 좋은 것으로 나타났다.

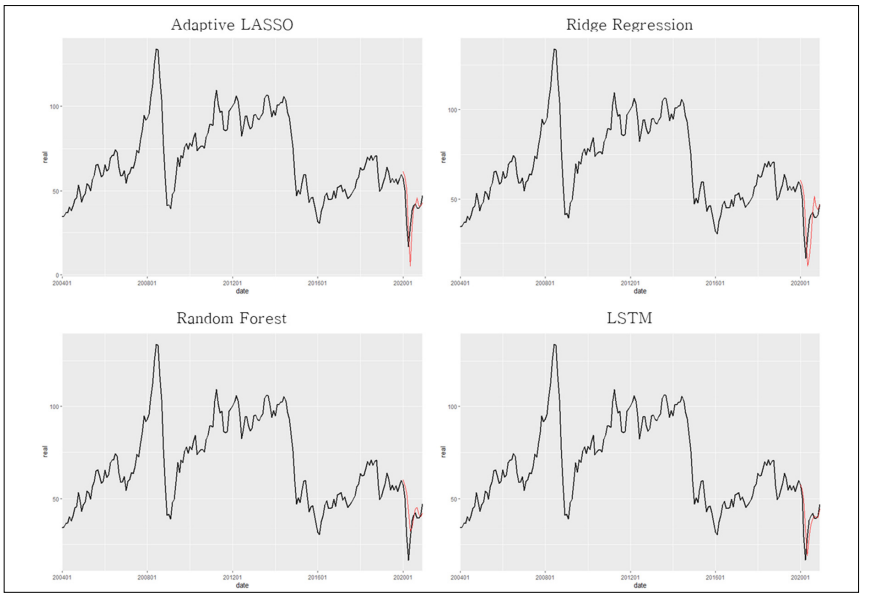
다음 장에서는 석유 공급 및 수요 관련 정형데이터 뿐만 아니라 Google Trends Data를 사용하여 유가모형을 예측하고 정형데이터만을 사용했을 경우의 결과를 비교하여 분석하였다.

2. 정형데이터와 Google Trends Data를 함께 사용한 유가 예측 결과

유가 하락기 동안 신문기사 중 빈도가 높은 단어인 'oil'의 검색어 추세(Google Trend Data)를 추가한 유가모형 예측 결과는 〈그림 8〉과 같이

나타났다.

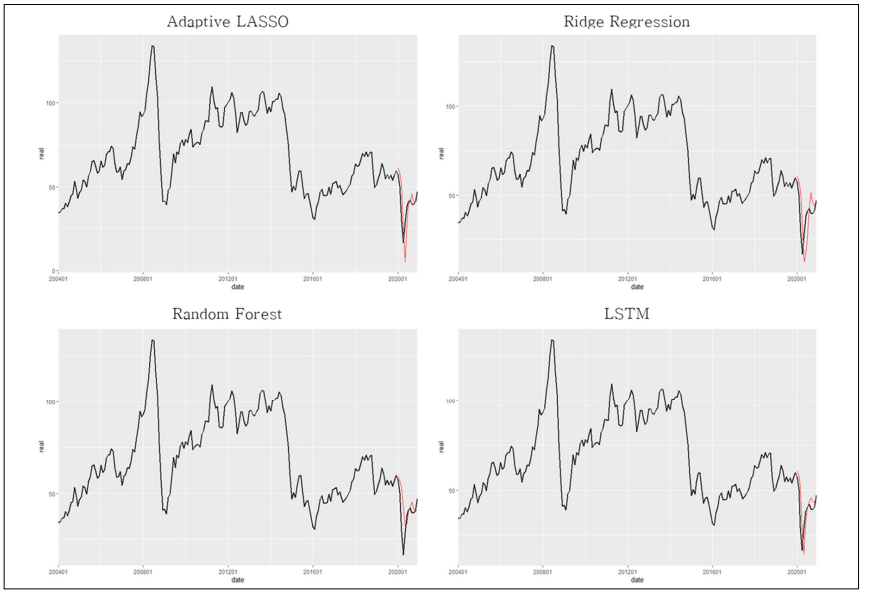
〈그림 8〉 정형데이터와 비정형데이터인 'oil'의 검색어 추세를 함께 사용하여 추정한 예측 모형 결과(Estimation results from forecasting models using both structured data and unstructured data ('oil'))



모형 추정 결과, Adaptive LASSO 모형의 RMSE는 9.056, Ridge Regression 모형의 RMSE는 11.320로 나타났다. 또한, Random Forest 모형의 RMSE는 10.722로, LSTM 모형의 RMSE는 8.828로 추정되었다. 추정 결과를 정리하자면, 석유 공급 및 수요 관련 정형데이터 및 비정형데이터인 'oil' 의 검색어 추세를 활용한 유가모형 중 LSTM 모형의 예측력이 가장 좋은 것으로 나타났으며, 그 다음으로 Adaptive LASSO 모형, Random Forest 모형, 마지막으로 Ridge Regression 모형의 순으로 예측력이 좋은 것으로 나타났다.

아래 〈그림 9〉에 유가 하락기 동안 신문기사 중 빈도가 높은 두 단어인 'oil'와 'crude'의 검색어 추세(Google Trends Data)를 추가한 유가모형 예측 결과가 제시되어 있다.

〈그림 9〉 정형데이터와 비정형데이터인 'oil', 'crude'를 함께 사용하여 추정한 예측 모형 결과(Estimation results from forecasting models using both structured data and unstructured data ('oil' and 'crude'))



추정 결과, Adaptive LASSO 모형의 RMSE는 10.564, Ridge Regression 모형의 RMSE는 12.539로 나타났다. 또한, Random Forest 모형의 RMSE는 10.780로, LSTM 모형의 RMSE는 9.420으로 추정되었다. 추정 결과를 정리하자면, 석유 공급 및 수요 관련 정형데이터 및 비정형데이터인 'oil'과 'crude'의 검색어 추세를 활용한 유가모형 중 LSTM 모형의 예측력이 가장 좋은 것으로 나타났으며, 그 다음으로 Adaptive LASSO 모형, Random Forest 모형, 마지막으로 Ridge Regression 모형의 순으로 예측력이 좋은 것으로 나타났다.

유가 예측 모형을 정형데이터를 사용한 모형과 유가 하락기 동안 신문기사 중 빈도가 높은 단어 2개를 선정하여 검색어 추세를 변수로 추가한 모형의 예측력을 비교한 결과는 〈표 4〉에 제시되어있다.

석유 공급 및 수요 관련 정형데이터를 활용한 모형의 RMSE는 Adaptive LASSO, Ridge Regression, Random Forest, LSTM 순으로 낮게 추정되었다. 따라서, Adaptive LASSO 모형의 RMSE가 가장 낮기 때문에 4개의 모형 중 예측력이 가장 좋은 모형이라고 할 수 있다. 정형데이터와

유가 하락기 동안 신문기사 중 가장 빈도가 높게 추출된 단어 'oil'의 Google Trends Data를 추가한 모형의 RMSE는 LSTM, Adaptive LASSO, Random Forest, Ridge Regression 순으로 낮게 추정되었다. 따라서, LSTM 모형의 RMSE가 가장 낮기 때문에 4개의 예측 모형 중 예측력이 가장 좋은 모형이라고 할 수 있다.

〈표 4〉 유가 예측 모형 결과(RMSE) 비교(Comparison among the results from forecasting models (RMSE))

	Adaptive LASSO	Ridge Regression	Random Forest	LSTM
i) 정형데이터	**9.160**	9.753	10.796	11.339
ii) 정형+비정형 ('oil') 데이터	9.056	11.320	10.722	**8.828**
iii) 정형+비정형 ('oil', 'crude') 데이터	10.564	12.539	10.780	**9.420**

주: i) 정형데이터, ii) 정형+비정형('oil') 데이터, iii) 정형+비정형('oil', 'crude') 데이터를 사용하여 각각 추정한 4가지 모형 중 가장 낮은 RMSE값을 가지는 모형은 Bold체로 표시되어 있다.

Note: The lowest value of RMSE among four forecasting models is in bold using i) Structured data, ii) Structured + Unstructured ('oil') data, and iii) Structured + Unstructured ('oil', 'crude') data.

정형데이터와 유가 하락기 동안 신문기사 중 빈도가 가장 높게 추출된 단어 'oil'과 두번째로 빈도가 높은 단어 'crude'의 Google Trends Data를 추가한 모형의 RMSE는 LSTM, Adaptive LASSO, Random Forest, Ridge Regression 순으로 낮게 추정되었다. 따라서, LSTM 모형의 RMSE가 가장 낮기 때문에 4개의 모형 중 예측력이 가장 좋은 모형이라고 할 수 있다.

LSTM 알고리즘 특성상 장기간의 데이터를 기억함으로써 Training 데이터가 길면 길수록 다른 예측 모형에 비해 예측력이 좋게 나타남을 알 수 있다. 또한, 정형데이터를 사용한 모형보다 제일 빈도가 높은 단어 'oil'의 검색어 추세를 추가한 모형의 예측 결과, 4개의 모형 중 Ridge Regression 모형을 제외한 3개 모형의 RMSE가 낮게 나타났다. 'oil' 뿐 아니라, 두번째로 빈도가 높은 단어 'crude'를 추가한 모형의 예측결과, Adaptive LASSO, LSTM 모형의 RMSE가 낮게 나타났다.

따라서, 본 연구에서는 WTI 유가를 예측할 때, 석유 공급 및 수요 관련

정형데이터만 사용한 모형의 예측력보다 유가 하락기 동안 신문기사 중 추출된 빈도가 높은 단어들의 검색어 추세를 설명변수로 추가한 모형의 예측력이 개선된다는 사실을 밝혀냈다.

Ⅳ. 결론 및 시사점

유가 예측은 우리나라와 같은 에너지 수입국에 있어 매우 중요한 주제이자 필수불가결한 부분이기 때문에 과거부터 오랜 시간동안 다양한 거시지표 변수와 원자재의 공급 및 수요 데이터를 활용하여 유가를 예측하려는 연구들이 많았다. 최근에는 미국의 서브프라임 모기지, 코로나19와 같은 문제 외에도 유가의 등락에 영향을 주는 원인이 다양해지고 있는 실정이다. 따라서, 본 연구에서는 유가 하락기 동안 수집된 신문기사 중 빈도가 높은 단어들을 추출하여 Google Trends Data를 통해 검색어 추이를 설명변수로 추가하여 유가 예측모형의 예측력을 향상시킬 수 있음을 보였다.

구체적으로, 본 연구는 2004년 1월부터 2020년 12월까지의 데이터를 이용하여 WTI 유가 예측에 영향을 주는 석유 공급 및 수요 변수 이외의 검색어 추세를 추가함으로써 예측력을 높일 수 있음을 보였다. WTI 유가 예측 모형의 예측력을 비교하기 위해 Adaptive LASSO, Ridge Regression, Random Forest 이외에 최근 가격 예측모형에서 많이 활용되고 있는 LSTM 알고리즘을 적용하여 각 모형의 RMSE를 산출한 결과, 정형데이터만 이용한 모형의 예측력보다 비정형데이터인 Google Trends Data를 함께 이용한 모형의 예측력이 개선된다는 점을 보였다.

본 연구를 통하여 정형데이터 외에 비정형데이터를 활용하여 유가 예측모형의 예측력을 높일 수 있다는 것을 검증하였으며, 이는 앞으로 GPI와 같이 데이터 빈도가 낮은 경제지표를 예측할 때, Google Trends Data와 같은 검색어 추세 데이터를 활용할 수 있는 가능성을 보여주었다. 국제 유가와 석유 공급 및 수요의 영향도를 측정할 수 있으며, 유가 관련 기사들로부터 추출된 핵심단어가 유가에 영향을 주는 외부요인을 뒷받침할 수 있는 변수로 활용될 수 있다는 점을 보였다.

과거에도 중장기 유가를 예측하기 위한 다양한 연구들이 진행되었으며,

그만큼 중장기 유가 예측은 우리나라와 같은 에너지 수입국에 있어 매우 필요하고 중요한 연구분야 중 하나라고 할 수 있다. 본 연구에서는 다양한 예측 모형을 이용하여 추정된 중장기 유가 예측을 통하여 개별적 시장 참여자는 원유 관련 파생상품에 투자할 수 있는 정보로도 활용할 수 있고, 원유 관련 민간 사업자는 안정적인 원유 구매계획 수립 시 활용 가능하다는 점에서 중요한 시사점을 제공한다.

투고 일자: 2022. 11. 21. 심사 및 수정 일자: 2022. 12. 9. 게재 확정 일자: 2022. 12. 9.

◈ 참고문헌 ◈

박강희 · Tianya Hou · 신현정 (2011), "기계학습기법에 기반한 국제 유가 예측 모델,"『대한산업공학회지』, 제37권, 64-73.

Park, Kanghee, Tianya Hou, and Hyunjung Shin (2011), "Oil Price Forecasting Based on Machine Learning Techniques," *Journal of the Korean Institute of Industrial Engineers*, 37, 64-73.

박동욱 (2018), "국제 유가가 국내 주식시장 및 업종별 주가에 미치는 영향력 변화," 박사학위논문, 부경대학교.

Park, Dongwook (2018), "The Change of Effect of Oil Price on the Korean Stock Market and Industrial Sector Indices," Department of Economics, The Graduate School, Pukyong National University.

서유정 · 조철근 (2019), "국제유가 변동성 예측과 유가 확률분포 추정,"『에너지경제연구원 기본연구보고서』, 19-07.

Cho, Cheol-keun, and Yujung Suh (2019), "Predicting the Volatility of WTI Futures Prices and the Probability Distribution of Oil Price Changes," *Korea Energy Economics Institute (KEEI) Report*, 19-07.

송경재 · 양회민 (2005), "시계열 분석에 의한 국제유가 예측; Nymex-WTI 선물가격 중심으로,"『통계연구』, 제10권, 4-4.

Song, Kyong-Jae, and Hoi-Min Yang (2005), "A Study on the Nymex WTI Prices Forecasting Using Time Series Analysis," *Journal of the Korean Official Statistics*, 10, 4-4.

이철용 (2011), "베이지안 모형을 이용한 중장기 국제유가 전망 연구,"『POSRI 경영경제연구』, 제11권, 58-86.

Lee, Chul Yong (2019), "Long-term Crude Oil Price Forecast Using the Bayesian Model," *POSRI Business and Economic Review*, 11, 58-86.

한동우 (2019), "인터넷 검색 자료를 활용한 유가 분석," 석사학위논문, 한양대학교.

Han, Dongwoo (2019), "An Analysis of Crude Oil Prices Using Internet Search Data," Department of Energy and Mineral Resources Engineering, The Graduate School, Hanyang University.

Forecasting Crude Oil Prices with Google Trends Data Based on Machine Learning Methods*

Seonmi Kim** · Dooyeon Cho***

Abstract

Forecasting crude oil prices is an important issue, especially for Korea which is the importer of crude oil, since fluctuations in crude oil prices may have a negative effect on the economy. This study investigates some factors that may cause fluctuations in crude oil prices with macro variables as well as Google Trends Data. By employing data on oil demand and supply mainly used in forecasting models for WTI crude oil prices and trends on keywords highly searched during a period of a decline in oil prices, it analyzes whether it can improve forecasting power. We find that including Google Trends Data, besides data on oil demand and supply, can improve predictive ability over the sample period January 2004 to December 2020. To compare predictability in various models, we employ Adaptive LASSO, Ridge Regression, Random Forest, and LSTM. The results suggest that the LSTM model outperforms other models when both structured data and Google Trends Data are jointly used.

KRF Classification : B030104, B030109
Key Words : Google Trends Data, LSTM model, forecasting oil prices, web scraping

---

* The authors are grateful to two anonymous referees for their helpful comments that improved this manuscript.

** First Author, Master in Economics, Department of Quantitative Applied Economics, Sungkyunkwan University, Seoul 03063, Republic of Korea, e-mail: ssmkim912@gmail.com

*** Corresponding Author, Associate Professor, Department of Economics and Department of Quantitative Applied Economics, Sungkyunkwan University, Seoul 03063, Republic of Korea, e-mail: dooyeoncho@g.skku.edu