검색어 트렌드와 뉴스기사 수를 이용한 주간 경제심리보조지수 개발

박채린 (소프트웨어학부), 이혜승 (소프트웨어학부)

December 11, 2020

요약. 정부는 현 경제 상황에 맞는 합당한 정책을 펼치고자 한다. 현 경제 상황은 소비심리가 과하게 위축되고 불황이 이어지면서, 정책 시행 전후의 단기적인 경제 상황의 이해가 필수적이다. 따라서 본 연구는 단기적으로 소비심리를 파악할 수 있는 주간 경제보조심리지수를 개발한다. 주간 경제보조심리지수에 반영되는 자료는 '침체', '금융위기', '불황', '폭락', '외환위기'의 키워드를 포함하는 뉴스 기사 수, 구글, 네이버 포털의 '경제' 키워드 검색량이다. 7개의 자료들을 예측변수로 대표적인 월간 지표인 소비자심리지수를 반응변수로 활용하여, 소비자심리지수 예측 문제를 해결한다. 소비자심리지수는 월간 지표이고, 본 연구에서 개발하는 지표는 주간 지표이므로, 적절한 데이터 전처리가 필요하다. 따라서 해당 달의 모든 주차의 변수값들을 평균내어 월간 데이터로 변환한다. 이후, 적합한 모형에 주간 자료들을 인입하여 주간 경제심리보조지수를 생성할 수 있다.

주요 용어: 소비자심리, 주간 경제심리보조지수, 예측 모형

1 서론

2008년의 경제 불황부터 2020년 한 해 동안 지속된 코로나19 바이러스까지, 굵직한 이슈들로 인해 국가 경제는 계속해서 무너지고 있다. 이러한 상황에서, 정부는 경제 정책 시행에 대한 빠른 피드백을 필요로 한다. 본래 정부는 '경제 안정과 성장'의 두 마리 토끼를 목표로 하는 경제 주체로서, 두 요인의 균형을 맞추기 위해 당시 경제 상황에 따른 합당한 정책을 펼친다. 따라서, 정책 시행 전과 현 경제 상황에 대한 이해가 필수적이다.

그 이해를 돕기 위해, 본 연구는 부정적인 경제 상황에서의 단기적인 흐름 및 현황을 빠르게 파악할 수 있는 주 단위 경제지표를 개발한다. 본 지표는 경제 활동의 주축을 이루는 가계(개인)의 소비에 영향을 미치는 소비자 심리에 대한 세부 지표로써 기존의 월 단위 지표인 소비자심리지수의 보조 지수로 활용하고자 한다. 이어서, 본 연구의 필요성 및 효과 3가지를 제시하고자 한다.

첫째, 기존의 경제지표보다 빠른 소비자의 피드백을 통해 부정적인 경제 상황에 보다 탄력적으로 대응이 가능하다는 장점이 있다. 본 연구는 주 단위 경제지표로서, 매주 주간 지표를 생성하기 때문에 경기 변동 및 단기적인 흐름에 발빠르게 대처할 수 있다. 현재 참조하고 있는 기존의 경제지표는 월말에 각 값이 발표되어, 급박한 상황에서 사용이 불가능하다는 맹점을 지니기 때문이다. 이에, 정부의 효과적인 단기 경제정책 시행이 가능해질 수 있으며, 기업(비즈니스)의 관점에서도 소비자 니즈 및 시장에 신속히 대응하는 '마켓 센싱' 역량 강화의 효과도 노릴 수 있다고 판단된다.

둘째, 본 연구(주간 경제보조심리지수)는 대표적인 기존의 지표 대비 속보성의 효과를 지니고 있다. 본 연구에서 개발한 모형에 수집한 주간 자료를 인입하면, 즉시 지수를 얻을 수 있기 때문이다. 기존의 경제지표인 소비자심리지수가 설문조사를 통해 생성되는 자료의 성격 상 지니는 단점을 보완하고자 한다. 소비자심리지수는 경제 주체들을 대상으로 경기 인식 및 전망에 관하여 조사하며, 현재생활형편, 생활형편전망, 가계수입전망, 소비지출전망, 현재경기판단, 향후 경기전망의 6개 개별 지수를 통해 계산된다. 이에, 설문조사 결과 취합부터 데이터 가공 및 분석을 거쳐 지수를 생성하는 과정까지 상당한 시간이 소요된다.

셋째, 표본 설계의 오류에서 벗어나 보다 정확한 지표를 얻고자 한다. 본 연구는 3개 포털의 검색량을 변수로 활용한다. 기존의 소비자심리지수는 층화다단추출법과 확률계통추출법을 이용하여 대상 가구를 선정한 후, 설문조사 결과를 바탕으로 생성된 지수이다. 이 과정에서 표본 설계로 인한 오류가 발생할 우려가 있다. 또한, 무성의한 대답 및 신뢰성 여부, 개인의 심리를 충분히 반영하지 못한 결과가 발생할 수 있다. 따라서, 포털의 검색량을 이용하는 본 연구는 설문조사의 맹점에서 탈피할 수 있다. 표본 설계의 비용과 오류에서 비교적 자유로우며, 집단의 대표성을 어느 정도 가정할 수 있고, 개인의 내면 심리를 파악할 수 있기 때문이다. 결과적으로 소비자심리지수의 문제점을 보완할 수 있는 주간 경제심리보조지수로 활용될 수 있는 여지가 충분하다 여겨진다.

본 연구(주 단위 경제지표)는 경기침체에 관한 뉴스 기사 수와 포털에서 개인들이 '경제' 키워드에 대해 검색한 수량을 활용한다. 이는 R-word index에서 아이디어를 착안하였다. R-word index란, 미국의 경기 침체 당시 이코노미지에서 처음 등장한 지수다. 뉴스 기사에 'recession'이라는 단어가 많이 등장하면, 민간소비가 줄어 경기 침체가 찾아온다는 전제 하에 경기 침체를 예측하는 지표다. 그리고 이 지수는 실제로 미국의 1981년과 1990년 경기 침체의 시작을 예측하는 데 성공했다. 본 연구에서는 키워드 수를 모두 카운팅하지 않고, 키워드를 포함하는 뉴스 기사 수를 이용하여 주간 지표를 생성한다. 단기적인 경제 흐름을 파악할 수 있으며, 경기 판단에 도움을 줄 것으로 기대한다.

해당 연구에 활용되는 데이터는 크게 3가지이다. '침체', '금융위기', '불황', '폭락', '외환위기'의 키워드를 포함하는 뉴스 기사 수, 구글, 네이버, 카카오 포털의 '경제' 키워드에 대한 검색량 비율, 그리고 소비자심리지수이다. 경제 상황에 대한 부정적인 단어들이 포함된 기사 수는 빅카인즈를 통해 수집하며, 포털 키워드 트렌드는 구글, 다음, 네이버에서의 '경제' 키워드의 검색량을 이용한다. 경기동행지수와 소비자심리지수는 한국은행에서 제공하는 경제통계 open API 서비스를 활용하여 수집한다.

경제 지표와 경기 심리, 전망, 태도 등과 경제 활동의 상관 관계에 대하여 많은 연구가 선행되었다. 선행 연구로는 포털 검색 통계를 이용한 경제 분석, SNS 게시글을 이용한 경제 분석이 있다. [1]'SNS 데이터를 활용한 소비자성향 분석'에서는 SNS에서 발생하는 경제 상황 메시지의 감성지수와 소비자심리지수를 비교 분석한다. 본 연구는 SNS가 아닌 검색 통계를 선택한다. 키워드 중심의 비정형인 소셜 네트워크 데이터의 텍스트 분석보다 검색 통계는 상대적으로 오류가 적고 의도성이 낮아, 비교적 객관적이기 때문이다.

[2]'빅데이터를 이용한 경기판단지표 개발: 네이버 검색 경기지수 작성과 유용성 검토' 연구는 네이버 내의 호황, 불황 검색 데이터를 바탕으로 네이버 검색 경기지수를 작성하였다. 6년 전에 이루어진 이 연구는 당시 우리나라 포털 점유율에서 네이버가 독보적인 비중을 차지하고 있어, 네이버 검색 통계만 이용하였다. 그러나, 현재는 네이버, 구글, 다음이 삼각편대를 이루고 있는 구조이다. 여전히 네이버가 점유율 1위를 차지하고 있으나, 구글과 다음의 비중이 이전에 비해 많이 높아졌기 때문이다. 현재의 포털 점유율 구조를 반영하여 검색 통계 대상으로 포털 3개를 고려한다.

본 연구의 목적은 소비자심리지수의 추세를 잘 따르는 주간 경제심리보조지수 개발에 있다. 구성은 다음과 같다. 2장에서는 경기 지표와 검색 통계에 대해 소개한다. 3장에서는 자료를 어떻게 수집하여 구성하고 어떠한 테이블 형식을

[1] SNS 데이터를 활용한 소비자성향 분석, 황영자, 2016, 통계개발원 연구보고서

[2] 빅데이터를 이용한 경기판단지표 개발: 네이버 검색 경기지수 작성과 유용성 검토, 이긍희, 황상필, 2014, 경제분석, 한국은행 경제연구원 제20권 제4호

가졌는지는 소개한다. 4장에서는 예측변수 선택과 전처리 및 데이터 분할 방법, 사용한 지도학습 모델들을 제시한다. 5장에서는 최종 모델을 선택하는 T검정 과정과 최종 선택 모형인 Lasso 선형 회귀 모형으로 생성한 주간 지수를 소개한다. 6장에서는 연구 결과를 요약하고 한계 및 향후 과제를 정리한다.

2 배경

- 경기종합지수(Composite Economic Indexes)

  경기종합지수는 경기변동의 국면, 전환점과 속도, 진폭을 측정하기 위해 고안된 경기 지표이다. 우리나라의 경기종합지수는 선행종합지수, 동행종합지수, 후행종합지수 3개가 사용되며, 여러가지의 자료를 가공, 종합하여 계산된다.

- 소비자심리지수(Composite Consumer Sentiment Index, CCSI)

  한국은행이 매월 전국 2,200가구를 대상으로 우편조사와 전화인터뷰를 통해 소비자동향지수를 실시한다. 소비자심리지수는 소비자동향조사를 바탕으로 현재생활형편, 생활형편전망, 가계수입전망, 소비지출전망, 현재경기판단, 향후경기전망에 대한 6개의 개별지수를 표준화하여 합성한 지수이다. 소비자심리지수가 100보다 높을 경우, 경제상황에 대한 소비자의 주관적인 기대심리가 낙관적임을, 100보다 낮을 경우에는 비관적임을 나타낸다.

- 경기동행지수 Component of Coincident Index, CCI)

  경기동행지수란 현재의 경기 상태를 나타내는 지표로서 국민경제 전체의 경기변동과 거의 동일한 방향으로 움직이는 7개 지표로 구성된다. 7개의 지표는 고용 부문의 비농립어업취업자수, 생산 부문의 광공업생산지수, 서비스업생산지수, 소비 부문의 소매판매액지수, 내수출하지수, 투자 부문의 건설기성액, 대외 부분의 수입액으로 구성된다. 7개 지표가 한달간 어느 정도 움직이는지를 토대로 경기동행지수가 만들어진다.

- 동행지수 순환변동치(Cyclical Component of Coincident Index)

  경기동행지수에서 추세변동을 제거한 순환요인으로, 현재의 경기상황을 파악하는 보조 지표이다. 순환변동치로 경기 변화를 더 뚜렷하게 판단할 수 있다. 동행지수 순환변동치가 100보다 높을 경우 경기가 호황이고, 100 미만이면 경기가 불황이라고 판단된다.

- 검색 통계

  검색 통계는 사용자가 검색 포털에 검색어를 입력하여 주요 정보를 찾을 때 검색 질의어별 검색수를 요약하여 정리한 것이다. 본 연구에서 이용하는 3개의 포털(네이버, 구글, 다음)에서는 검색 통계를 트렌드라는 사이트를 통해 제공하고 있다. 검색어 트렌드는 사람들이 검색한 검색 수를 단순히 합한 것이 아닌 상대값으로 표현된다. 해당 포털에서 검색어가 검색된 횟수를 합산하여 조회기간 내 최다 검색량을 100으로 설정하여 상대적인 변화를 나타낸다.

3 자료

본 연구에서는 데이터셋을 크게 세 가지를 활용한다. 경제 지표인 소비자심리지수와 경기동행지수 순환변동치, 특정 키워드를 포함하는 뉴스 기사 수 정보, 그리고 포털에의 '경제' 키워드 검색비율 정보다.

3.1 자료 소개

id	year	month	ccsi
1	2016	1	99.5
2	2016	2	101.1
3	2016	3	98.6
4	2016	4	98.2

Table 1: 소비자심리지수(CCSI)

id	year	month	cci
1	2016	1	101.9
2	2016	2	102
3	2016	3	102.2
4	2016	4	102.4

Table 2: 동행지수 순환변동치

첫 번째로, 경제 지표인 소비자심리지수과 경기동행지수는 [3]한국은행에서 제공하는 경제통계 open API 서비스를 활용하여 수집한다. 소비자심리지수는 주간 경제심리보조지수를 생성을 위한 예측 문제를 해결할 때, 반응변수로 활용한다. 경기동행지수 순환변동치는 포털에의 '경제' 키워드 검색량과 실제 경제 상황과의 상관 관계가 있는 지를 파악할 때 활용한다.

id	year	month	day	google	naver	kakao
1	2016	1	4	53	19.65163	0
2	2016	1	11	58	25.16599	0
3	2016	1	18	74	27.38639	0
4	2016	1	25	59	24.72622	0

Table 3: 구글 트렌드, 네이버 데이터랩, 카카오 트렌드의 '경제' 키워드 검색 비율

두 번째로, 포털의 '경제' 키워드 검색 비율 정보는 [4]구글 트렌드, [5]네이버 데이터랩, [6]카카오 트렌드에서 받아온다. 구글과 카카오는 API 제공을 하지 않으며, 웹 스크래핑도 불가하여 각각 csv파일과 엑셀 파일로 다운받는다. 네이버는 데이터랩을 통해 naver API를 활용하여 데이터를 받아온다. 포털 검색 비율 정는 예측문제에서의 예측변수들로 활용한다.

id	year	month	day	keyword1	keyword2	keyword3	keyword4	keyword5
1	2016	1	4	825	419	437	986	168
2	2016	1	11	770	348	351	626	182
3	2016	1	18	918	422	441	445	147
4	2016	1	25	1067	425	567	383	148

Table 4: 경제 상황 키워드를 포함하는 뉴스 기사 수

마지막으로, 특정 키워드를 포함하는 뉴스 기사 수 정보는 한국언론진흥재단에서 운영하는 뉴스 분석 서비스인 빅카인즈에서 json 파일로 데이터를 수집한다. 포털에 키워드를 검색해서, 주간 뉴스 기사 수를 스크래핑을 시도했으나, 결과의 재현성을 고려했을 때, 스크래핑보다는 빅카인즈의 키워드 트렌드 서비스를 활용하는 것이 더 안정적이다. 뉴스 키워드 수는 부정적인 경제상황 키워드를 참고하여 침체, 금융위기, 불황, 저성장, 외환위기, 금리인하, 디플레이션, 불경기, 폭락 등을 고려한다. 선택 기준은 일주일에 100건 이상 등장하고, '침체' 단어를 기준으로 상관계수가 높은 키워드이다. 최종적으로 뉴스 키워드는 침체, 금융위기, 불황, 폭락, 외환위기로 하여, 뉴스 기사 수 정보를 받는다.

[Table 4]의 keyword1는 '침체', keyword2는 '금융위기', keyword3는 '불황', keyword4는 '폭락', keyword5는 '외환위기' 키워드를 포함한 뉴스기사 수이다.

[3]한국은행 경제통계시스템(ECOS): http://ecos.bok.or.kr/
[4]구글 트렌드: https//trends.google.co.kr/trends/?geo=KR
[5]네이버 데이터랩:https://datalab.naver.com/
[6]카카오: https://datatrend.kakao.com/

3.2 자료 전처리 과정

예측문제의 반응변수인 소비자심리지수는 월말에 제공되는 지수로 한달 단위이다. 2016년부터 2020년 10월까지 총 58개월의 소비자심리지수 값을 갖는다. 그러나 본 연구에서는 주간 지표를 얻기 위해, 변수들을 주간으로 수집했고, 총 252개의 값들을 갖고 있다. 반응변수와 예측변수의 지표 단위를 맞추기 위해 전처리 과정이 필요하여 크게 두 가지를 고려한다. 첫 번째로는 주간 데이터들을 소비자심리지수와 동일하게 월간 데이터로 변환, 두 번째로는 월간 데이터들을 주간지수에 맞게 주간 데이터로 변환하는 방법이다. 데이터셋에 해당 전처리 기법들을 적용한 후에 테이블 간 병합을 진행할 수 있기 때문에, 전처리 과정을 진행한 후의 테이블로 ER diagram을 그린다.

id	year	month	keyword1	keyword2	keyword3	keyword4	keyword5	google	naver	ccsi
1	2016	1	895.00	403.50	449.00	610.00	161.25	61.00	25.834265	99.5
2	2016	2	811.80	396.60	359.20	341.60	124.40	60.00	24.908160	97.2
3	2016	3	667.25	267.00	293.00	98.50	107.50	67.75	28.296107	99.5
4	2016	4	718.25	258.25	367.00	142.25	116.75	69.00	25.843962	101.1

Table 5: 평균 전처리를 적용한 데이터셋

첫 번째 방법에서는 월의 마지막 주차만 활용해서 월간지표로 활용하는 것을 고려했으나, 월 마지막 주의 레코드가 월 단위의 소비자 심리를 충분히 반영하지 않는다. 따라서, 월 단위로 변수마다 매 주차의 평균과 중앙값을 구해 월간 데이터로 변환해준다. 상단의 [Table 5]과 부록의 [Table 14]은 각각 평균, 중앙값 전처리를 진행한 후, 부록의 [Figure 3]와 같이 병합된 형태이다.

id	year	month	day	keyword1	keyword2	keyword3	keyword4	keyword5	google	naver	ccsi
1	2016	1	4	825	419	437	986	168	53.00	26.05846	93.575
2	2016	1	11	770	348	351	626	182	58.00	25.16599	95.550
3	2016	1	18	918	422	441	445	147	74.00	27.38639	97.525
4	2016	1	25	1067	425	567	383	148	59.00	24.72622	99.500

Table 6: 내삽을 적용한 데이터셋

두 번째 방법에서는, 내삽을 적용하여, 월간 소비자심리지수 값들이 주어졌을 때, label이 없는 주차의 레코드들에 대해 등간격으로 매 주차의 소비자심리지수를 추정하여 주간 지표로 변환한다. [Table 6]는 내삽 전처리를 진행한 후, 부록의 [Figure 4]와 같이 병합된 형태이다.

4 방법론

소비자 심리 지수를 예측하고 주간 지표를 생성하기 위해 반응변수는 소비자심리지수(CCSI), 예측변수는 포털 트렌드의 '경제' 키워드 검색량과 5개의 키워드에 대한 뉴스 기사 수로 설정하여 지도학습을 실시한다. 반응 변수는 월 단위, 예측변수는 주 단위이므로 이를 맞춰주기 위해 변수값들의 평균값, 중앙값, 내삽 적용으로 3가지 방법을 고려한다. 분할 방법으로는 랜덤 분할, 시간을 고려한 분할로 2가지를 고려한다. 지도학습 모델은 다중 선형 회귀 모형, 랜덤 포레스트 모형, 일반화가법모형, 라쏘 회귀 모형으로 4가지이며 각 파라미터는 다양하게 시도한다.

4.1 변수 상관관계 탐색

먼저 예측변수로 고려한 것은 포털 트렌드 검색량과 뉴스 키워드 수이다. 포털은 구글, 네이버, 카카오를 고려

하고, 어떤 포털을 예측변수로 선택할 것인지 판단하기 위해 경기가 좋지 않을 때 포털의 '경제' 검색량이 높은지 확인한다. 각 포털과 경기동행지수(CCI)와의 단순회귀분석 및 상관성을 확인하였다.

- 포털: 구글

id	coef	std err	t	$P > \|t\|$
intercept	0.8123	0.024	33.742	0.000
google	-0.4099	0.078	-5.265	0.000

Table 7: 구글 트렌드의 '경제' 키워드 검색 비율과 경기동행지수 CCI와의 단순 회 귀분석

단순 회귀 분석 결과 [Table 7]을 보면 구글의 '경제'키워드 검색량 변수의 계수는 -0.41로 두 변수는 음의 관계를 보이며, 유의수준 0.05에서 p-value는 0.05보다 작으므로 통계적으로 유의미하다고 할 수 있다. 피어슨 상관계수는 -0.401755로 강한 음의 상관 관계를 가진다.

- 포털: 카카오

id	coef	std err	t	$P > \|t\|$
intercept	0.7699	0.076	10.161	0.000
kakao	-0.3473	0.258	-1.344	0.189

Table 8: 카카오 트렌드의 '경제' 키워드 검색 비율과 경기동행지수 CCI와의 단순회귀분석

단순 회귀 분석 결과 [Table 8]을 보면 계수는 -0.3473으로 두 변수는 음의 관계를 보이며, 유의수준 0.05에서 p-value는 0.189로 유의하지 않다. 피어슨 상관계수 또한 -0.23828이므로 카카오와 CCI는 선형관계가 약함을 알 수 있다.

- 포털: 네이버

id	coef	std err	t	$P > \|t\|$
intercept	0.9331	0.046	20.075	0.000
naver	-0.7144	0.158	-4.532	0.000

Table 9: 네이버 트렌드의 '경제' 키워드 검색 비율과 경기동행지수 CCI와의 단순회귀분석

단순 회귀 분석 결과 [Table 9]을 보면 계수는 -0.7144로 두 변수는 음의 관계를 보이며, 유의수준 0.05에서 p-value는 0.05보다 작으므로 통계적으로 유의미하다고 할 수 있다. 피어슨 상관계수는 -0.52491로 강한 음의 상관 관계를 가진다.

포털 트렌드와 경기동행지수 순환변동치와의 단순회귀분석 및 상관성의 결과 카카오 트렌드 변수는 CCI와의 상관계수의 절댓값이 작다. 따라서, 경제 상황이 좋지 않을 때, 사람들이 포털에 '경제'를 많이 검색할 것이라는 가정이 유의하지 않다는 것을 확인했다. 반대로 네이버와 구글 트렌드 변수는 경기동행지수 순환변동치와의 상관계수 절대값이 크고 경제상황이 좋지 않을 때, 사람들이 포털에 '경제'를 많이 검색할 것이라는 가정이 유의하다는 것을 확인했다. 따라서, 예측 변수로 카카오를 제외한 구글, 네이버 트렌드 검색량를 활용한다. 따라서, 최종 예측 변수들은 구글, 네이버 트렌드의 검색량과 침체, 금융위기, 불황, 폭락, 외환위기 키워드의 뉴스 기사 수로 총 7가지다.

4.2 전처리 및 데이터 분할

변수들을 활용하여 실제 모형을 적합할 때에는 스케일링을 거친다. 사용된 스케일러는 StandardScaler로 평균을 제거하고 데이터를 단위 분산으로 조정한다. 즉, 각 예측변수들의 평균을 0, 분산을 1로 변경하여 모든 변수들이 같은 스케일을 갖게 된다.

예측 모델을 학습시키고 테스트하기 위하여 데이터들을 트레이닝셋과 테스트셋으로 분할한다. 먼저 데이터를 트레이닝셋과 테스트셋으로 분할한 후, 모델 적합 시 10-fold Cross Validation 과정 내에서 트레이닝 셋을 모형 적합할 자료와 모형 평가할 자료로 분할하여 적합 및 평가를 진행한다. 트레이닝 셋과 테스트 셋으로 나누는 방법으로는 2가지를 고려한다. 첫 번째로 데이터를 랜덤으로 섞은 후 트레이닝 셋 : 테스트 셋 = 8 : 2로 분할한다. 두 번째로는 2016년 1월 - 2019년 10월을 트레이닝 셋으로, 2019년 11월 - 2020년 10월을 테스트 셋으로 분할한다.

4.3 예측 모형

소비자심리지수 예측을 위해 다음과 같이 4가지 모형족을 활용한다. 각 모형마다 적합한 파라미터들을 찾기 위해 여러 파라미터를 대입해보고 제일 작은 test RSME를 갖는 파라미터를 선택한다.

- 다중 선형 회귀
  예측변수 7개를 이용하여 소비자심리지수를 예측한다. 추정한 회귀 계수로 각각의 예측변수가 소비자심리지수에 얼마나 영향을 미치는지 알 수 있다.
- 랜덤 포레스트
  랜덤 포레스트의 주요 파라미터는 n_estimators(생성할 트리의 개수), max_features(최대 선택할 변수의 수)이다. 본 연구에서 생성한 트리의 개수는 $2^3, 2^4, 2^5, 2^6$, 최대 선택 변수의 개수는 1,2,3,4,5로 총 16개의 파라미터를 고려한다.
- 일반화 가법
  예측변수의 개수는 7개로 7개의 spline term을 가진다. 파라미터 λ는 랜덤으로 100*7개를 생성하여 모델을 적합한다.
- 라쏘 회귀
  라쏘 회귀에서 이용한 파라미터는 α=[10, 1, 0.1, 0.01, 0.001]을 이용하여 적합한다.

데이터 전처리 방법 3가지와 데이터 분할 방법 2가지로 총 6가지의 데이터를 각각 4가지 모델에 대해 10-fold cv를 적용한다. 이 과정에서 최적의 파라미터를 찾고 T검정을 통해 최적의 모델을 1개씩 선택한다. 이에 관한 자세한 내용은 5.2에서 설명한다.

4.4 예측 문제 해결 과정

3가지의 전처리 방법과 2가지의 데이터 분할 방법을 고려하여, 총 6가지 방법으로 전처리한 데이터로 소비자심리지수 예측모형들을 생성한다. 각 방법에서 10-fold Cross Validation(10겹 교차검증)을 통해 적합한 다중 선형 회귀 모형, 라쏘 회귀 모형, 랜덤 포레스트 회귀 모형, 일반화 가법 모형을 비교한다.

4가지의 모형의 성능 차이가 없다는 귀무가설을 기각하고자, T검정을 진행하여 성능의 차이가 있는 지, 어떤 모형이 가장 좋은 성능을 보이는지 분석한다.

가설은 다음과 같다.

$$H_0\text{: model A의 RMSE} = \text{model B의 RMSE}$$
$$\text{즉, 두 모델의 성능에 차이가 없다.}$$
$$H_1\text{: 두 모델의 성능에 차이가 있다.}$$

유의수준 0.05에서 양측 검정을 진행하고, [7] $|T \text{ 통계량}| > t_{0.025,9} = 2.262$ 이면 귀무 가설을 기각하여 두 모델 A와 B의 성능에 차이가 있다고 판단한다. 그러나 이때, T검정에 활용하는 10개의 자료들이 10-fold cv로 얻은 RMSE값들이기 때문에 서로 의존적이다. 따라서 T검정을 하는 데에 어느 정도의 위험성이 따른다. 독립적이지 않은 자료들로 인하여 실제적인 신뢰구간이 더 길어지며, p-value가 더 커지는 경향을 보인다. 따라서 오차범위 안에서는 거의 차이가 없다는 전제 하에, T검정 결과를 활용하고자 한다.

4.5 비교 지표

데이터 전처리 및 분할에 따른 각 데이터셋마다의 최적의 모델 6개의 성능을 비교하기 위해 test RMSE와 test MAE를 이용한다.

- RMSE(평균제곱오차)

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y_i})^2}$$

  예측값과 실제값을 뺀 후 제곱한 값들의 평균에 루트를 취한 것이다. MSE는 오류의 제곱 평균이므로, 실제 반응변수의 스케일과 다르고 실제 오류 평균보다 더 커질 수 있다. 따라서, MSE에 루트를 취한 RMSE를 이용한다.

- MAE(평균절대오차)

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y_i}|$$

  실제값과 예측값의 차이의 절댓값을 평균한 것이다. 소비자심리지수에 특이값이 있는 본 연구의 데이터 셋에 잘 부합하는 지표로 보인다.

[7] $T* = mean/(std/n)$

5 실험 결과

5.1 연구 환경

소비자심리지수를 예측하는 모형을 생성한 후, 주간 변수 값들을 대입하여 얻은 값을 주간 경제심리보조지수로 활용한다.

먼저 예측 문제에서는 3가지의 전처리 방법과 2가지의 데이터 분할 방법을 고려하여, 총 6가지 방법으로 전처리한 데이터로 소비자심리지수 예측 모형들을 생성했다. 각 방법에서 10-fold Cross Validation을 통해 적합한 다중 선형 회귀 모형, 라쏘 회귀 모형, 랜덤 포레스트 회귀 모형, 일반화 가법 모형을 비교하여 본 연구의 목적에 잘 부합하며, 성능이 좋은 모형을 최종으로 선택했다. 이후, 최종 모델에 주간 변수들을 인입하여 얻은 값을 주간 경제심리보조지수로 활용한다.

5.2 소비자심리지수 예측

5.2.1 검정 과정

데이터들을 2019년 10월까지 데이터를 훈련 자료로, 2019년 11월 이후 데이터를 테스트 자료로 활용하고, 전처리 방법으로 주차 데이터들의 평균 내는 방법을 활용했을 때의 T검정 과정이다.
부록 [Table 15]는 해당 데이터셋에서 각 모형들의 10-fold CV를 적용하여 얻은 RMSE값들이다. 10개의 RMSE값을 이용하여 T검정을 진행한다.

	multiple linear regression – random forest
1-fold RMSE	1.7356
2-fold RMSE	-0.7198
3-fold RMSE	4.6099
4-fold RMSE	0.1795
5-fold RMSE	0.6840
6-fold RMSE	0.1468
7-fold RMSE	1.2165
8-fold RMSE	0.8454
9-fold RMSE	0.9781
10-fold RMSE	0.6093
RMSE average	1.0279
RMSE std	1.3507

Table 10: 다중 선형 회귀 모형과 랜덤 포레스트 모형의 RMSE 차이

[Table 10]는 다중 선형 회귀 모형과 랜덤 포레스트 모형의 RMSE의 차이를 나타내는 표다. 표의 값들을 이용하여, 두 모형을 비교하는 T검정을 진행한다.

0.05의 유의수준에서 $t_{0.025,9} = 2.262$이고, 해당 검정의 T통계량은 2.4065로, $t_{0.025,9} = 2.262$보다크므로 귀무 가설이 기각된다. 따라서, 다중 선형 회귀 모형과 랜덤 포레스트 모형의 성능에는 차이가 있으며, 이때 평균 RMSE 가 양수이기 때문에, 랜덤 포레스트 모형의 RMSE가 더 작은 값을 가지는 것을 알 수 있다. 즉, T검정 결과, 랜덤 포레스트 모형이 다중 선형 회귀 모형보다 성능이 좋다고 판단된다. 모든 전처리 및 분할 방법을 적용한 데이터셋에서 각 6번의 T검정을 통해, 최적의 모형들을 선택한다.

5.2.2 모형 비교 및 선택

	Shuffle = True	Shuffle = False
Mean	LR Test RMSE: 4.006 Test MAE: 3.122	Lasso($\alpha = 0.1$) Test RMSE: 8.438 Test MAE: 2.568
Median	LR Test RMSE: 6.203 Test MAE: 3.664	Lasso($\alpha = 0.1$) Test RMSE: 9.336 Test MAE: 2.605
Interpolation	RF(n_estimators=32, max_features=4) Test RMSE: 3.490 Test MAE: 2.483	RF (max_features=3, n_estimators=64) Test RMSE: 13.889 Test MAE: 3.381

Table 11: 전처리 및 분할 방법에 따른 최적 모형

전처리 및 분할 방법에서의 최적 모형들은 [Table 11]와 같다. RMSE와 MAE 오류 측도만으로 비교하면, 내삽을 적용한 후 임의로 섞은 데이터 셋으로 훈련시킨 랜덤 포레스트 모형이 가장 좋은 예측력을 보여준다.

그러나 2020년과 같이 예상치 못하게 경제 상황이 급격하게 나빠진 경우에, 경제심리보조지수로서 주간지수가 유용할 것으로 판단된다. 이러한 주간지수의 목적을 고려하였을 때, 2016년부터 2019년의 데이터로 모델을 훈련시키고, 2020년 데이터로 test를 진행하는 방식이 예상하지 못한 경제 상황을 예측하고자 하는 목적에 부합한다. 또한 세 가지 전처리 방법 중, 가장 예측력이 좋았던 평균 전처리 방법의 최적 모델인 α가 0.1인 라쏘 회귀 모형을 최종으로 선택한다.

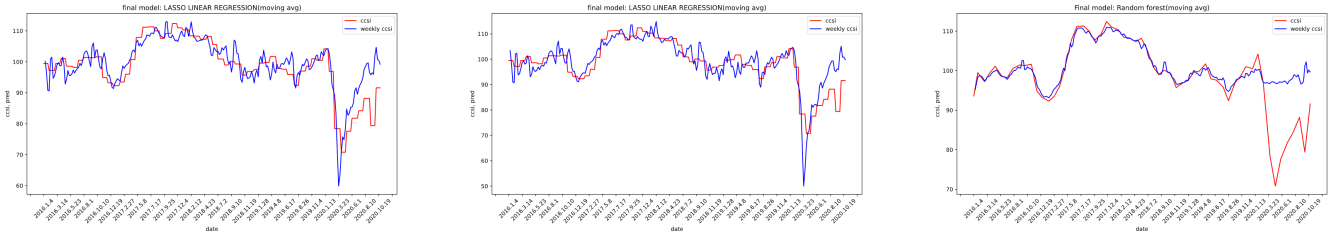

Figure 1: 평균, 중앙값, 내삽 전처리 방법

Shuffle = false에 대한 방법 3개 모형들의 시각화 결과는 [Figure 1]으로 확인할 수 있다. 세 전처리 방법 중, 내삽 적용 방법은 test set에 대해 거의 제대로 예측하지 못한다. Interpolation 방법을 제외하고, 평균과 중앙값은 성능에의 차이가 뚜렷하지 않기 때문에 test RMSE가 약 1정도 작은 평균 전처리 방법의 모형을 최종으로 선택한다.

5.2.3 최종 예측 모형 해석

함수 공간에의 제약을 가하여, 손실함수 최소화를 방해하는 L1 Norm으로 인하여 모델의 분산이 줄어들고, test 오류를 감소시키고 해석력이 좋기 때문에 성능이 비슷했던 다중 선형 회귀 모형, 랜덤 포레스트 모형이 아닌 라쏘 회귀 모형을 선택했다. 라쏘 회귀 모형의 식은 아래 식과 같다.

$$(\hat{\beta_{\lambda_0}}, \hat{\beta_\lambda}) = \underset{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j) + \lambda \sum_{j=1}^{p} |\beta_j|$$

라쏘 회귀 모형은 규제와 변수선택을 동시적으로 하여, y 추정에 필요한 변수들만 활용한다. 10-fold Cross Validation을 진행하여, 가장 최적의 λ인 0.1을 활용한다. 예측 모형에서의 예측변수는 X_1: 키워드1(침체),

X_2: 키워드2(금융위기), X_3: 키워드3(불황), X_4: 키워드4(폭락), X_5: 키워드5(외환위기), X_6: 구글 검색어 트렌드, X_7: 네이버 검색어 트렌드가 사용되며, 반응 변수 Y는 CCSI(소비자심리지수)이다. 2016 - 2019년의 트레이닝 셋이 적합 시킨 라쏘 회귀 모형은 아래 식과 같다.

$$Y = 101.48–2.25X_1–0.316X_2–2.568X_3 + 0.907X_4–0.395X_5–1.954X_7$$

'침체' 키워드를 포함한 뉴스기사 수, '금융위기' 키워드를 포함한 뉴스기사 수, '불황' 키워드를 포함한 뉴스기사 수, '외환위기' 키워드를 포함한 뉴스기사 수, 네이버의 '경제' 키워드 검색량은 음의 계수를 가지며, '폭락' 키워드를 포함한 뉴스기사 수는 양의 계수, 그리고 구글의 '경제'키워드 검색량은 변수 선택에서 제거되어 0의 계수를 갖는다.

따라서, 적합된 모형에서는 '침체', '금융위기', '불황', '외환위기' 키워드의 뉴스 기사 수와 네이버 검색량이 많은 경우 소비자심리지수는가 낮았다. 즉, 부정적인 경제 상황을 다루는 뉴스 기사 수가 많을 수록, 사람들이 경제에 관심을 갖고 포털에 많이 검색해볼 수록, 소비 심리가 위축된다 해석해볼 수 있다.

변수	CCSI와의 상관계수
keyword1	-0.844307
keyword2	-0.666754
keyword3	-0.655859
keyword4	-0.259032
keyword5	-0.422807
google	-0.359626
naver	-0.367600
CCSI	1.000000

Table 12: 변수와 CCSI 상관계수

라쏘 회귀 모형의 경우, 반응변수와의 조건부 상관계수가 낮은 것은 중요하지 않은 변수로 판단되어 계수를 0으로 추정한다. [Table 12]를 보면, 구글의 '경제' 키워드 검색량은 소비자심리지수와의 상관계수가 -0.359626로 작고, 네이버 변수와의 상관계수도 0.675322로 두 변수간의 다중공선성 존재한다. 또한 규제항이 없는 다중 선형 회귀 모형을 적합했을 때, 구글 변수의 계수 절댓값이 가장 작기 때문에 반응변수에 영향을 가장 덜 미치는 것으로 보인다. ($Y = 101.480–1.9933X_1–0.5722X_2–2.9120X_3 + 1.2553X_4–0.3958X_5 + 0.3118X_6–2.4645X_7$) 이러한 이유로 변수 선택에서 제외되었다. 그러나 동일 검색어에 대하여 구글 검색의 중요성이 네이버 검색의 중요성보다 떨어진다는 의미는 아니며, 라쏘 회귀 모형의 수리적 특성에 의하여 한 변수만 선택되었다.

또한 폭락 키워드에 대한 뉴스 기사 수 변수는 양의 계수를 갖는다. 그러나 단일 변수만 살펴보았을 때는 다른 변수들과 같이 소비자심리지수와 음의 관계를 갖는다. 경제 상황을 생각해보았을 때, 폭락은 비교적 단기적으로 발생하는 키워드이고, 침체, 금융위기, 불황, 외환위기는 비교적 거시적인 경제 상황을 가리키는 표현들이다. 따라서 폭락은 다른 키워드들과 다른 방향으로 움직일 가능성이 있다. 그러나 y값을 구성하는 다른 변수들이 모두 음의 계수를 갖기 때문에, 폭락 키워드로만 y값이 결정되지 않을 것이며, 폭락 키워드가 y값의 보정 역할을 한다고 판단된다.

5.3 주간 경제심리보조지수 생성

	date	weeklyCLI
1	2016.1.4	100.114651
2	2016.1.11	100.218829
3	2016.1.18	96.924625
4	2016.1.25	90.786863
5	2016.2.1	90.612911
...	...	...
248	2020.9.28	101.857171
249	2020.10.5	104.707534
250	2020.10.12	100.781811
251	2020.10.19	100.454511
252	2020.10.26	99.235899

Table 13: 주간 경제심리보조지수

모형 적합과 테스트는 월간 데이터에 관련해서 진행했지만, 본 연구의 목표는 주간 경제심리보조지수의 생성이므로, 월간 데이터로의 전처리 이전의 전체 데이터를 모델에 인입한다. 이 때, 최종모형은 라쏘 회귀 모형이므로, 모형 적합시 활용했던 정규화 스케일링을 적용한 형태의 데이터를 인입한다. 결과적으로 [Table 13]와 같이, 해당 주의 경제심리보조지수를 생성한다.

5.4 주간 경제심리보조지수의 유효성 입증

본 연구에서 개발한 주간 경제심리보조지수가 과연 선행경제지표로써 유효한지에 대해 두 가지 방법으로 그 성능을 입증하고자 한다. 첫 번째로는 소비자심리지수의 추세를 잘 따르는지 확인한다. 두 번째로는 test set에 대해 추정한 값과 실제 소비자심리지수값과의 피어슨 상관계수가 충분히 큰 값을 갖는지 확인한다.

- CCSI의 추세를 잘 따르는지 확인
  입증된 경기선행지수인 소비자심리지수의 추세를 따르는 것을 보임으로써 개발한 주간 경제보조심리지수가 경기에 선행되는 지수의 역할을 하고 있음을 확인할 수 있다. [Figure 1]의 왼쪽 그림을 보면, 개발한 주간 경제심리보조지수(파란색 선)는 소비자심리지수(빨간색 선)의 추세를 잘 따르고 있다.
  그러나 2020년의 주간 데이터들에 대해서는 추세는 잘 따르지만, 원래의 소비자심리지수보다 과대 추정되었다. 과대 추정된 부분의 데이터를 살펴보고, 추가적인 분석은 5.5에서 설명한다.

- 테스트셋의 추정값과 실제값의 상관계수가 충분히 큰 지 확인
  테스트 셋의 소비자심리지수 추정값과 소비자심리지수의 상관계수는 0.776이다. 추정값과 실제값이 강한 양의 상관관계를 보이기 때문에, 2020년과 같이 경제 상황이 급격히 안좋아진 시기에도, 충분히 소비자의 심리를 나타내는 지표로 활용이 가능하다.

5.5 추가 분석

2020년에 대한 과대 추정의 원인 파악하고자 한다.

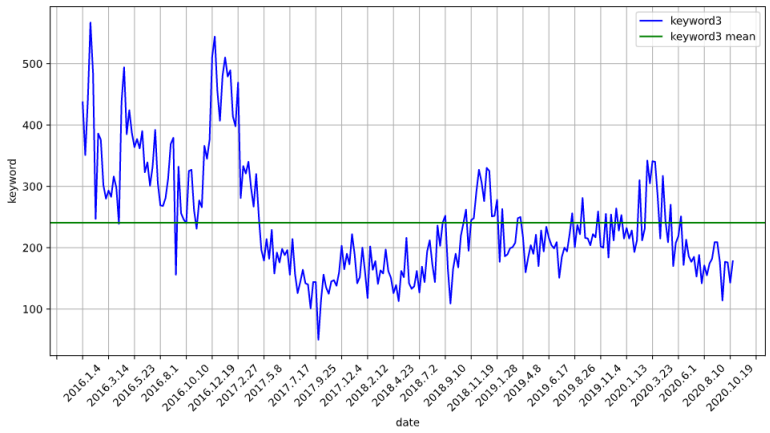

Figure 2: 불황(keyword3)을 포함하는 뉴스 기사 수

예측 모형이 2020년에 대해서 값을 높게 추정하고 있다. [Figure 2]를 보면 y값이 높게 추정되는 2020년에 대해서는 2016년에 비해 상대적으로 불황 키워드에 대한 뉴스 기사 수가 낮다. 2016년에는 대외적으로는 중국의 사드 보복 사건, 유가 하락, 수출 불황 극심, 대내적으로는 박근혜, 최순실 게이트 사건으로 인해 대내외적 경제 상황이 좋지 않았다. 이로 인해 2016년 뉴스 기사에 불황 키워드가 많이 사용되어 상대적으로 2020년에는 불황 키워드에 대한 뉴스 키워드가 낮음을 파악할 수 있다. 또한 불황 키워드 변수의 계수 절댓값이 가장 크기 때문에 Y에 영향을 가장 크게 미친다. 즉, 2020년의 낮은 불황 키워드의 뉴스 기사 수의 영향으로 반응 변수가 크게 추정된다.

소비자심리지수가 최저를 찍는 2020년 3월 이후로는 검색량이 평소의 패턴을 찾았으나, 여전히 좋지 않은 경제 상황으로 인하여 소비자들의 심리는 위축되어 있다. 따라서 소비자들의 심리를 더 반영할 수 있는 추가 키워드의 분석이 필요할 것으로 판단된다.

6 요약 및 결론

6.1 연구 요약

변수로 크게 세 가지 포털 트렌드 검색량, 뉴스 키워드 수를 주간 데이터를 수집했으나 예측하고자 하는 값은 소비자심리지수로 월간 지표이기 때문에 월의 모든 주차 값들에 평균, 중앙값 전처리를 적용했다. 그리고 월별 소비자심리지수에 내삽을 적용하여 반응 변수를 주간 지표로 생성했다. 최종으로 선택한 전처리 방법으로는 주간 변수값들을 평균내는 것으로 이용했고, 데이터 분할 방법으로는 시간 순으로 2016년 1월 2019년 10월을 트레이닝 셋으로 2019년 11월 이후를 테스트 셋으로 활용했다. 모형은 총 4가지 다중 선형 회귀, 랜덤 포레스트, 일반화 가법 모형, 라쏘 회귀를 고려했으며, 최종적으로 성능이 가장 좋았던 α가 0.1인 라쏘 회귀 모형으로 예측 문제를 해결했다. 최종 모형에 주간 자료를 넣어 주간 경제심리보조지수를 개발했다. 개발한 주간 경제심리보조지수가 소비자심리지수의 추세 및 값을 잘 따르고 있기 때문에, 경기선행지표 역할을 충분히 수행한다고 판단된다. 따라서 2020년 같이 펜데믹으로 인한 불안정한 경제 상황이나 새로운 경제 정책에 관한 빠른 피드백 및 탄력적인 대응이 필요할 때, 주간 경제심리보조지수를 활용하는 것이 효과적일 것으로 보인다.

6.2 한계 및 향후 과제

- 다양한 키워드 반영

  경제 키워드들이 아닌 다양한 factor의 키워드들에 대한 뉴스 기사수를 카운팅 하여, 새로운 인사이트를 얻고 모형의 예측력을 더 높일 수 있을 것으로 기대된다. 예로, 중국의 리커창 총리는 전력 사용량 40%, 은행 대출량 35%, 철도 화물 운송량 25%를 반영하여 경제 상황을 나타내는 리커창 지수를 개발하였다. 해당 지수는 직접적인 경제 키워드들은 아니지만, 경제 상황을 반영하는 직관적 요소들로 구성되어, 경기를 잘 반영하여 새로운 인사이트를 제공하기도 했다.

- 포털 데이터의 추가 수집

  뉴스 기사 수 데이터는 1990-01-01부터 수집이 가능하지만, 포털의 주간 검색 비율 데이터는 2016-01-01부터 수집이 가능하다. 따라서 데이터 수집기간 중, 2020년과 같이 소비자심리지수가 7, 80대의 값을 갖는 경우가 없었다. 2008년의 경제 대공황때의 포털 데이터 수집이 가능했다면, 낮은 소비자심리지수에 대한 충분한 훈련이 되어, 모형의 예측력이 더 좋아졌을 것으로 판단된다.

- 포털 자료간 평균 전처리

  구글 트렌드와 네이버 데이터랩의 자료 간의 상관 계수는 0.675322이다. 양의 상관관계를 가지는 두 자료의 다중공선성을 제거하기 위해, 두 자료의 평균을 하나의 예측 변수로 활용해볼 수 있다.

7 부록

- 코드
  주간 경제심리보조지수 구현은 해당 github repository에서 확인 및 재현이 가능하다.

- 테이블

id	year	month	keyword1	keyword2	keyword3	keyword4	keyword5	google	naver	ccsi
1	2016	1	871.5	420.5	439.00	535.50	158.00	58.50	25.612225	99.5
2	2016	2	889.0	364.0	376.00	289.00	123.00	60.00	23.764760	97.2
3	2016	3	663.00	260.50	288.00	106.00	106.00	68.50	27.666635	99.5
3	2016	4	663.00	250.00	367.50	116.00	124.50	68.50	25.756655	101.1

Table 14: 중앙값 전처리를 적용한 데이터셋

	multiple linear regression(A)	Random forest(B)	GAMs for regression(C)	Lasso regression(D)
1-fold RMSE	5.3051	3.5694	9.9997	4.4412
2-fold RMSE	1.3125	2.0323	6.5000	2.5307
3-fold RMSE	5.6360	1.0261	1.7354	2.8046
4-fold RMSE	2.5682	2.3887	6.1808	2.2995
5-fold RMSE	3.6963	3.0122	7.2112	2.7202
6-fold RMSE	2.7389	2.5921	9.8191	2.5736
7-fold RMSE	3.1159	1.9052	7.5407	1.5171
8-fold RMSE	2.2055	1.3601	11.9615	1.9567
9-fold RMSE	3.6255	2.6474	6.0791	2.7369
10-fold RMSE	4.0446	3.4353	5.0167	5.0838
RMSE average	3.4248	2.3969	7.2044	2.8664
RMSE std	1.2717	0.7913	2.7368	1.0299

Table 15: 각 모형 10-fold CV 적용하여 얻은 RMSE

- ER 다이어그램

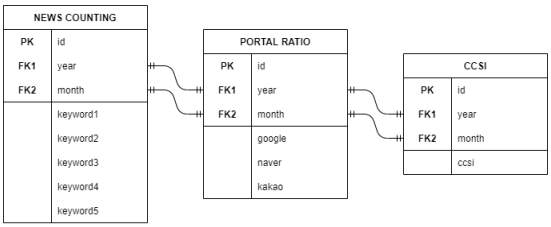

Figure 3: 평균, 중앙값 전처리를 적용한 테이블 ERD

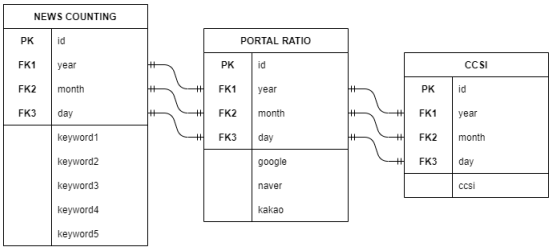

Figure 4: 내삽을 적용한 테이블 ERD