[2020-2 개별연구 보고서]

오피니언마이닝를 활용한 국내 코로나 19 확진자 수 증가 감지 방안 연구

과목명: 개별연구

지도교수: 윤병운 교수님

학과: 산업시스템공학

학번: 2018112455

이름: 이 솔

가. 연구의 필요성과 중요성

○ 코로나 19 상황 장기, 상시화 전망

- 세계 코로나 19 누적 확진자 수는 62,697,693 명이며 사망자 수는 1,460,841 명(2020.11.29 기준)으로 지속적으로 증가 중
- 대한민국 코로나 19 누적 확진자 수는 33,824 명이며 사망자 수는 523 명(2020.11.29 기준)
- 사회·경제적 파급효과가 상당하여 이에 따른 불확실성이 장기간 지속 우려

○ '사회적 거리두기' 단계 조정을 통한 지역사회 감염 차단 노력

- 코로나 19 유행의 심각성과 방역조치의 강도에 따라 1~5 단계로 구분하여 시행
- 거리두기 단계 격상 기준은 권역별 중증환자 병상 여력 및 주간 유행 양상을 중심으로 설정하며 중환자실 병상 여력으로 감당 가능한 주평균 일일 확진자 수를 핵심 지표로 활용하되, 감염 재생산 지수 등 다양한 보조 지표 고려
- 코로나 19 확진자 수 파악 및 예측의 중요도 증가

○ 코로나 19 이후 SNS 사용량 급증

- 코로나 19 대유행 시점마다 인스타그램, 트위터 채널에서의 정보량 급격하게 증가
- 코로나 19 확산 방지를 위한 자가격리, 원격근무, 사회적 거리두기 시행 등으로 집에 있는 시간은 늘고 대면 접촉은 줄어들면서 나타난 결과로 보임
- 코로나 19 관련 키워드 언급 및 해시태그 사용량 증가

○ 소셜 데이터를 수집하고 분석하여 코로나 19 확진자 수의 증가 감지하는 방법론 필요

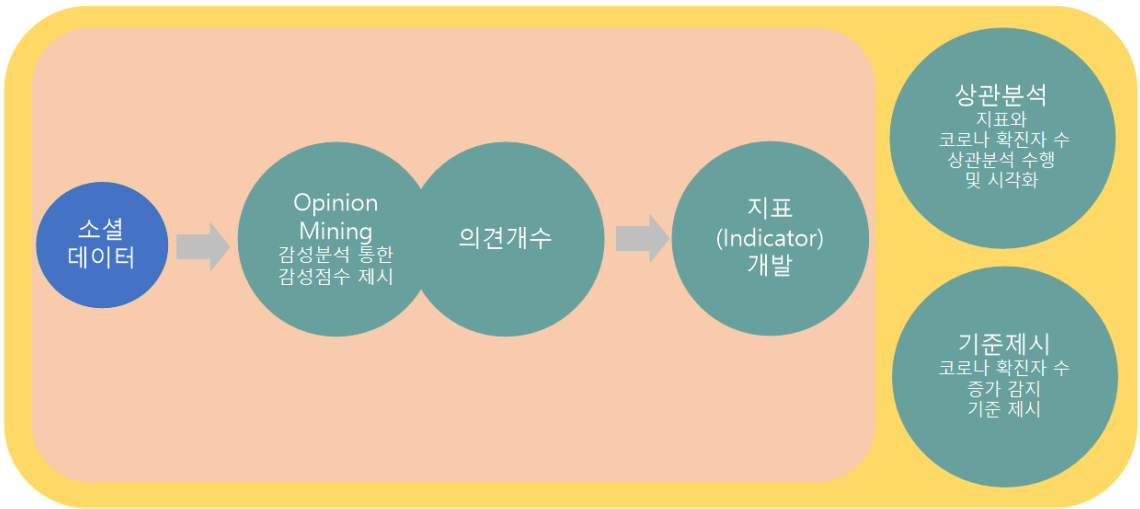
- 소셜 빅데이터의 분석은 방대한 양의 데이터를 활용하여 다양한 참여자의 생각과 의견을 확인할 수 있기 때문에 사회적 문제의 파악과 예측에 활용되어 옴
- 소셜 데이터를 이용한다면 코로나 19 확진자 수에 관한 유의미한 정보를 추출할 수 있을 것으로 보임

나. 연구 목표 및 내용

(1) 최종 목표

○ 본 연구는 소셜 데이터를 통해 코로나 19 확진자 수 증가를 감지하기 위한 기준을 제시하기 위해 소셜 데이터 감성분석을 통한 지표 생성과 지표와 코로나 19 확진자 수와의 상관분석 수행

- 1 단계에는 코로나 19 와 관련된 소셜 데이터(Social data)를 웹 크롤링(Web crawling)을 통해 수집하여 데이터베이스를 구축한 후, 의견 추출(opinion extraction)을 통한 오피니언 마이닝(Opinion mining)과 의견 개수 기반의 지표(Indicator) 개발
- 2 단계에는 지표와 코로나 19 확진자 수와의 상관분석을 수행하여 상관관계 파악, 최종적으로 코로나 19 확진자 수 증가를 감지하기 위한 기준 제시



[그림 1] 연구 개념도

▪ 1 단계에는 코로나 19 와 관련된 소셜 데이터베이스로부터 데이터를 수집하고 감성분석을 통해 감성점수를 파악하고 이에 의견 개수를 반영, 코로나 19 에 관해 개인의 생각과 상황을 포함하고 있는 지표를 개발할 것임

▪ 소셜 데이터와 같이 방대한 데이터를 웹 크롤링을 통해 수집하고, 감성분석(sentiment analysis)에 활용할 수 있도록 데이터베이스를 구축

▪ 구축한 데이터베이스를 기반으로 word2vec 을 이용하여 전처리한 후 감성분석(sentiment analysis)을 통해 감성점수를 도출

▪ 감성분석을 통해 도출된 감성점수에 의견개수를 포함해 새로운 지표 개발

▪ 2 단계에는 코로나 19 확진자 수 증가를 감지하기 위한 기준을 제시하기 위해 지표와 코로나 19 확진자 수와의 상관분석을 수행하여 상관관계를 파악

(2) 연구 범위 및 내용

○ 코로나 19 관련 상황 및 의견 데이터 수집 및 데이터베이스 구축

- SNS 의견 데이터 수집
- 유의미한 정보 추출을 통해 상황 및 의견 데이터베이스 구축

○ 코로나 19 에 관해 개인의 생각과 상황을 포함하고 있는 지표 개발

- 감성 분석을 통한 각 의견데이터의 감성점수 도출
- 감성점수와 의견개수를 반영 지표 개발

○ 지표와 코로나 19 확진자 수와의 상관관계 분석 및 시각화

- 지표와 코로나 19 확진자 수와의 상관관계 분석
- 코로나 확진자 수의 증가를 감지할 수 있는 기준 제시

다. 연구수행내용 및 연구결과

(1) 연구프로세스

○ 데이터 수집

▪ 코로나 19 확진자 수 증가 감지 기준 마련을 위해서는 코로나 19 발생에 관하여 사전적 의견 데이터 수집이 필요하다 판단되어, 본 연구에서는 코로나 19 사전적 의견 데이터를 비대면모임, 대면모임, 규제, 방역, 코로나 총 5 개의 키워드로 구분하여 수집

▪ 즉각적이고 자유로운 의견 표출에 대표적인 SNS 인 '트위터'에서 데이터 수집을 진행하였으며 각 키워드 별 상세 키워드는 키워드와 연관성이 높으며 사용 횟수가 많은 것을 기준으로 선정

▪ 수집 기간은 국내 첫 코로나 19 감염자 발생일 약 2 주 전인 2019.11.22~2020.12.02 로 설정

▪ 각 키워드 별 상세 키워드를 포함한 트윗을 대상으로 작성시간(작성 일자) 및 트윗 내용을 텍스트 형태로 추출

▪ 특성이 다른 다섯 키워드 데이터를 모두 분석함으로써 코로나 19 에 대한 국민의 의견을 반영할 수 있음

▪ 추후 이루어질 상관분석에서 사용될 국내 코로나 19 확진자 발생 데이터는 일자 및 확진자 수를 텍스트 형태로 추출(출처: 공공 데이터 포털, 보건복지부_코로나 19 감염_현황)

데이터베이스 유형	키워드	상세 키워드	수집 가능한 데이터
소셜 데이터 (트위터)	비 대면모임	온라인모임 온라인파티 온라인회식 온라인집회 홈트 집밥 홈카페	작성시간, 트윗 내용
	대면모임	모임 파티 회식 집회 헬스장 맛집 카페	
	규제	사회적거리두기 사회적거리두기실패	
	방역	방역 마스크 손소독제	
	코로나	코로나	

[표 1] 코로나 19 관련 사전 의견 데이터베이스

○ 각 키워드별 지표(Indicator) 개발

▪ (감성사전 정의) 감성사전이란 긍정적이거나 부정적인 감정을 나타내는 단어들을 모아놓은 사전을 의미하며, 본 과제에서는 수집한 전체 코로나 19 의견데이터에서 긍·부정적으로 쓰이는 단어들과 표현들을 이용하여 감성사전을 정의함. 특히 이 단계에서 감성값(Sentimental value)이 은 단어들을 먼저 시드 키워드(seed keyword)로 정의함

▪ (Word embedding) 전처리한 데이터를 word2vec 알고리듬을 이용하여 각 단어들을 임베딩하여, 개별 단어의 고유한 특징을 나타낼 수 있도록 단어별 벡터 값을 도출함

▪ (Propagation 을 통한 단어 별 감성 값 계산) 앞서 도출된 시드 키워드와 graph-based semi supervised learning 을 이용하여, 시드 키워드와 벡터 거리가 가까운 단어들로 시드 키워드의 감성점수를 거리에 비례하여 할당함. 시드 키워드와의 거리는 word2vec 을 통해 도출된 단어별 벡터 값 간 차이를 통해 계산되며, 시드 키워드와 가까운 위치에 있는 단어들은 이와 유사한 감성 값 및 특징을 갖는다고 볼 수 있음

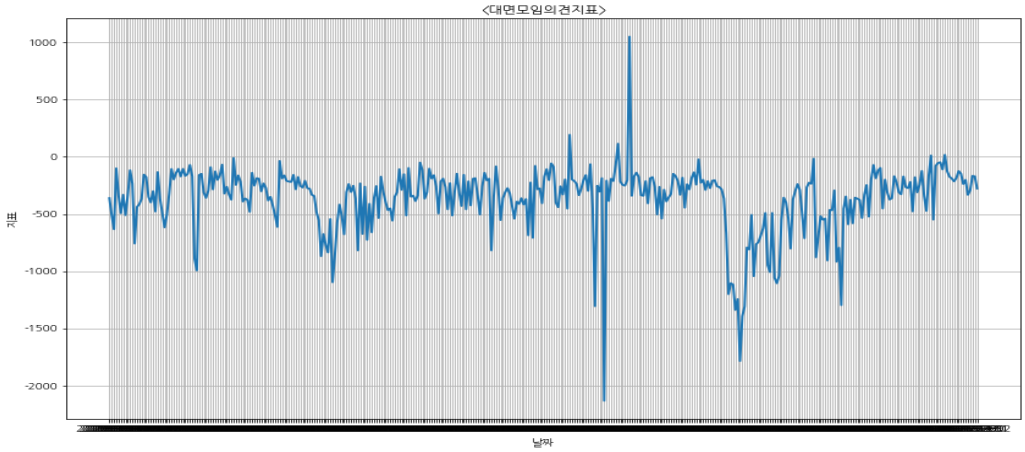
▪ (키워드 별 지표 도출) 일별로 하루에 발생한 의견 개수와 각 의견 감성점수들의 평균을 곱하는 방식으로 키워드(대면모임, 비 대면모임, 방역, 규제, 코로나)별 지표 도출

○ 각 키워드별 지표(Indicator) 추이 일 기준 시각화

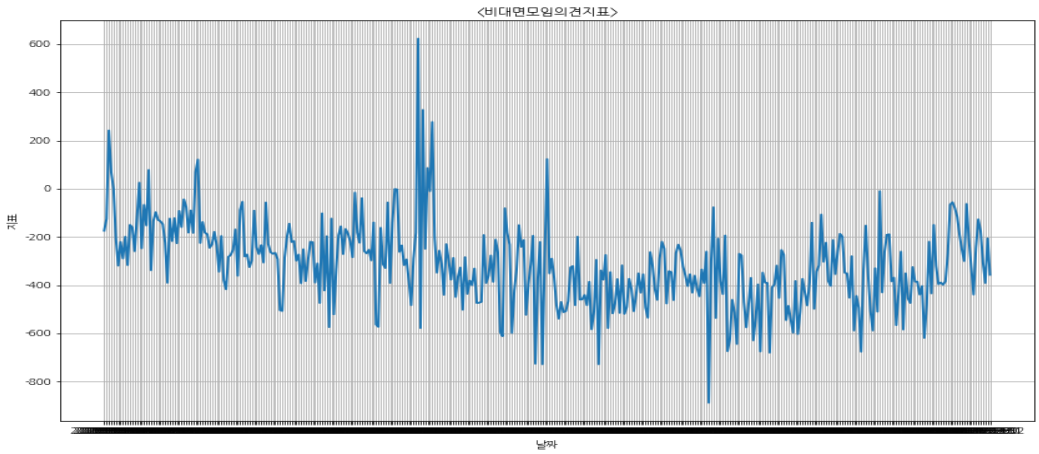
▪ 각 키워드별 지표(Indicator)은 총 5 개로 대면모임의견지표, 비대면모임의견지표, 방역의견지표, 규제의견지표, 코로나의견지표가 있으며 이를 일 기준으로 그래프를 생성하여 시각화 함

▪ x 축은 기간으로 2019.11.22 ~ 2020.12.02 에 해당하며 y 축은 각 키워드별 지표(Indicator)

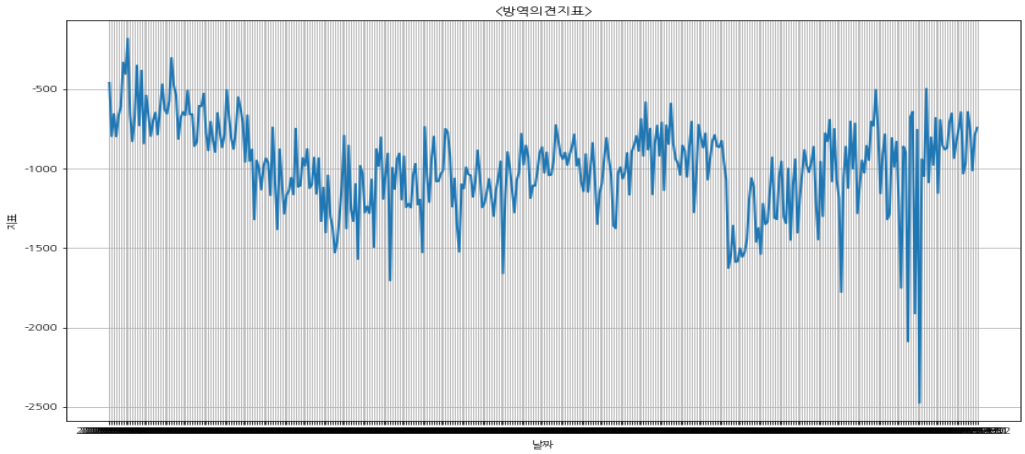
▪ 값이 증가할수록 해당 키워드에 대한 사람들의 의견이 긍정적으로 변하는 것이고 감소할수록 해당 키워드에 대한 사람들의 의견이 부정적으로 생성되는 것을 의미



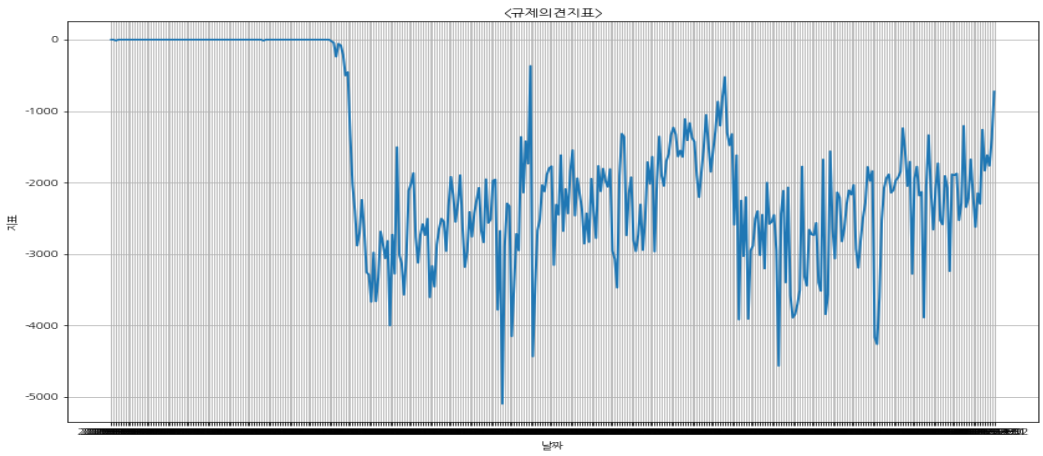
[그림 2] 대면모임의견지표의 추이 그래프(일 기준)



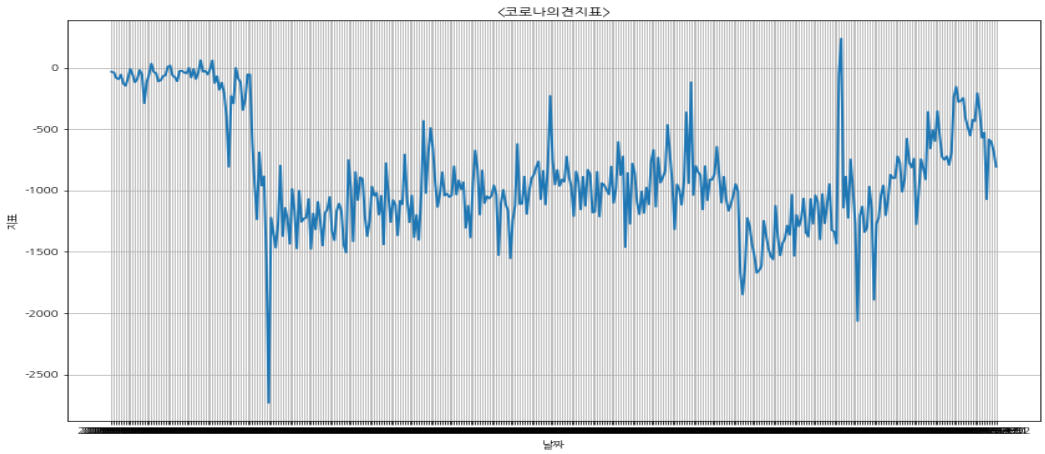
[그림 3] 비대면모임의견지표의 추이 그래프(일 기준)



[그림 4] 방역의견지표의 추이 그래프(일 기준)



[그림 5] 규제의견지표의 추이 그래프(일 기준)



[그림 6] 코로나의견지표의 추이 그래프(일 기준)

○ 각 키워드 별 지표(Indicator)과 코로나 19 확진자 수의 상관관계 분석

▪ 코로나 19 관련 사전의견데이터를 나타내는 지표로 키워드(대면모임, 비 대면모임, 방역, 규제, 코로나)별 지표 사용, 추가적으로 일별 코로나 19 확진자 수 데이터 사용

일별	대면모임 의견지표	비대면모임 의견지표	방역 의견지표	규제 의견지표	코로나 의견지표	코로나 19 확진자 수
2019-11-21	-361	-174	-461	0	-32	0
2019-11-22	-509	-125	-797	0	-37	0
…	…	…	…	…	…	…
2020-12-02	-278	-206	-743	-727	-806	511

[표 2] 일별 의견지표 및 코로나 19 확진자 수

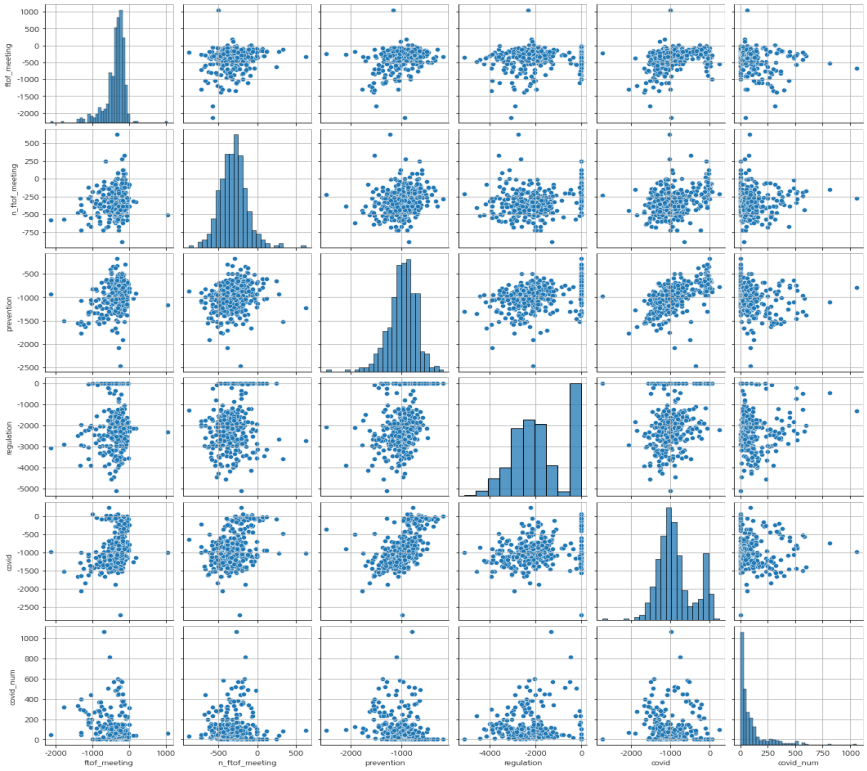
▪ 이용 데이터는 위에서 수집한 데이터들을 전처리 함(Null data 를 포함하는 행은 제거)

▪ 최종적으로 본 연구에서 이용할 변수와 별수 별 설명은 [표 3] 참고

변수명	설명
date	일별
ftof_meeting	대면모임 의견지표
n_ftof_meeting	비 대면모임 의견지표
Prevention	방역 의견지표
Regulation	규제 의견지표
covid	코로나 의견지표
covid_num	코로나 19 확진자 수(국내)

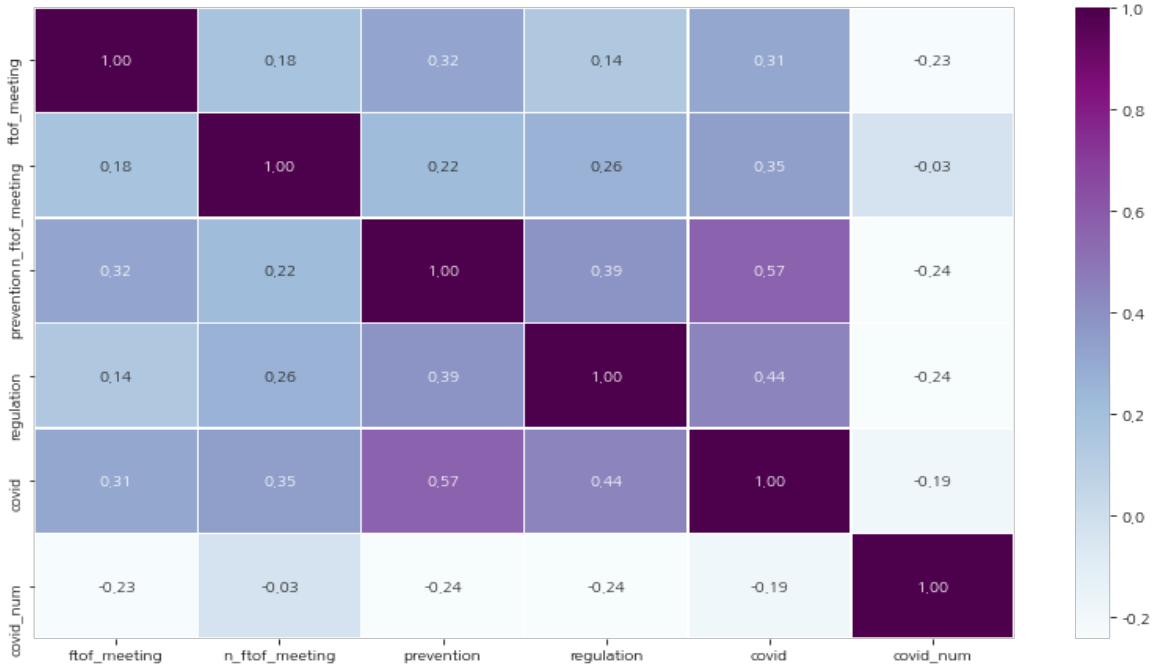
[표 3] 변수 설명

▪ 변수 별 상관관계를 산점도와 히스토그램을 통해 시각화하면 [그림 7]와 같음



[그림 7] 변수 별 상관관계 시각화

▪ 변수 별 상관행렬을 heatmap 으로 표현하였으며 그 결과는 [그림 8]와 같음



[그림 8] 변수 간 상관행렬

▪ covid_num(코로나 19 확진자 수)와 타 변수들의 상관계수들에 대해 통계적 검증을 실시하였고 유의수준 95% 하에서 n_ftof_meeting 을 제외한 나머지의 상관관계가 통계적으로 유의미함

변수	P 값
ftof_meeting	4.29559901204629e-06
n_ftof_meeting	0.5344099263542007
prevention	1.7409157927938073e-06
regulation	2.054637448077559e-06
covid	0.00015465727733229943

[표 4] 변수 별 상관계수의 p 값

○ 코로나 19 확진자 수 증가를 감지할 수 있는 기준 제시

▪ [그림 7] 피어슨 상관 계수(Pearson correlation coefficient)에 따르면 covid_num(코로나 19 확진자 수)와는 n_ftof_meeting(비대면모임의견지표)를 제외하고 나머지지표와 약한 음의 선형관계를 보임

▪ covid_num(코로나 19 확진자 수)와 가장 유의미한 상관관계를 보이는 변수는 ftof_meeting, prevention, regulation 이며 이 변수(지표)의 증가율과 covid_num(코로나 19 확진자 수)의 증가의 관계를 분석한 결과 각 지표의 기준일 기점 이전 3 일간의 지표 증가율의 평균이 0 미만일때 기준일 1 일 이후 covid_num 이 증가할 확률이 각각 72%, 100%, 64%로 다소 유의미한 수치를 보임

▪ 코로나 확진자 수 증가를 감지하기 위해 ftof_meeting, prevention, regulation 의 변수의 기준일 기점 이전 3 일간의 지표 증가율의 평균값을 참고하여 평균값이 0 미만이라면 기준일의 코로나 확진자 수가 증가할 수 있다고 할 수 있음

(2) 연구결과

본 연구에서는 코로나 19 확진자 수 증가를 감지할 수 있는 기준 탐색을 위해 가용한 데이터셋을 검토하고, 사전적 의견 데이터를 수집함. '대면모임', '비대면모임', '방역', '규제', '코로나' 총 5 개의 유형으로 나누어 수집한 사전적 의견 데이터를 바탕으로 감성사전을 재정의하고 재정의한 감성사전을 바탕으로 감성분석을 통해 각 의견 데이터에 해당하는 감성지수를 도출함. 감성지수와 의견 데이터 개수를 바탕으로 각 유형을 대표하는 5 개의 지표(Indicator)을 생성함. 생성한 지표(Indicator)들과 전국 코로나 19 확진자 수와의 상관분석을 수행하여 이들의 상관관계를 파악하고 가장 유의미한 상관관계를 보이는 지표 3 가지 '대면모임의견지표', '방역의견지표', '규제의견지표'에 대해 코로나 19 확진자 수 증가를 감지할 수 있는 기준을 제시하기위해 각 지표와 일별 코로나 19 확진자 수와의 관계를 재분석함. 그 결과 '코로나 19 확진자 수 증가를 감지하기 위해 '대면모임의견지표', '방역의견지표', '규제의견지표'에 대해서는 기준일 기점 이전 3 일간의 지표 증가율의 평균값을 참고하여 평균값이 0 미만이라면 기준일의 코로나 확진자 수가 증가할 수 있다'라는 결론을 도출함.

하지만 결론 도출에 있어 이에 대한 정확도를 명확히 계산하지 못했다는 한계점이 존재. 도출된 결론을 바탕으로 새로운 test data 를 선정하여 정확도를 측정하고 개선하는 단계가 추가된다면 보다 유의미한 코로나 19 확진자 수를 감지할 수 있는 기준이 마련될 것으로 기대됨.