네이버와 다음 뉴스의 경쟁 업체에 대한 뉴스 차이 양상 탐구

김준형[1], 이원혁[2], 황현수[3]

요약

본 연구는 pyhton을 활용한 텍스트 마이닝 및 토픽 분석에 관한 연구로 특정 단어에 대한 개별 포털의 보도 양상을 탐구하는데 연구촛점이 있다. 포털에서 뉴스를 제작하는게 아니라 언론사별 뉴스를 제작 후 포털에 보도하는 거라 비슷할 것이라고 예상되는 바, 이를 실증적으로 파악하고 차이가 있다면 어떤 차이가 있는지 확인해보려고 한다.. 네이버와 다음에서 네이버주가와 카카오주가를 검색하여 나오는 뉴스들을 이용하였다. 이를 바탕으로 '네이버주가'와 '카카오주가'를 키워드로 검색한 네이버와 다음 포털에서 최신부터 약400~600개의 기사들을 수집하였다. 기사 수집에는 Python으로 작성된 웹크롤링을 사용하였다. 수집된 결과에 대하여 각 키워드 별 포털을 KoNLP의 Okt를 활용하여 형태소 분석을 수행하였다. 이를 바탕으로 기사들에 사용된 단어의 빈도 분석을 실시하였고 그리고 감성분석도 실시하였다. 이를 통해 단어 사용에서 드러나는 포털간의 차이점을 확인하고 감성라벨링을 붙여 모델링도 진행하였다. 또한 더욱 심층적이고 논리적인 보도 양상 유추를 위하여 LDA 토픽 분석을 수행하였고 그 결과를 제시하였다.

주요용어 : 텍스트마이닝, 뉴스기사분석, 출현단어분석, 한국어형태소분석, KoNLP, LDA토픽모델링, 감성분석, 워드클라우드

1. 서론

인터넷은 '정보의 바다' 지만 한 사람이 볼 수 있는 정보는 한정돼 있다. 디지털 시대, 어떤 정보를 PC나 스마트폰의 상단에 띄우느냐 하는 '정보의 배열' 이 막강한 권력인 이유다. 이 권력이 알고리즘에 위임되고 있다. 그러나 각 포털에서 뉴스 기사를 취사 선택한다는 논란은 계속 지속해 오고 있다. 포털 사이트는 하루 평균 3만여 개의 기사가 송고되는데 이 가운데 인공지능을 갖춘 알고리즘이 부적절한 기사를 걸러 3000여 개를 선정한다. 카카오는 지난 2015년부터 포털 다음과 모바일 메신저 카카오톡 내 뉴스를 AI 알고리즘 '루빅스(現 카카오i)'를 통해 편집하고 있다. 네이버의 경우 지난 2017년 4월부터 AI 알고리즘인 '에어스(AiRS)'를 도입해 이용자별 맞춤형 뉴스를 제공하고 있다. AI 알고리즘이 뉴스 편집의 공정성과 신뢰성 시비를 불식시키기 위한 효율적인 대안으로 떠올랐기 때문이다. 포털 뉴스의 '손 편집' 을 없앤 결정적 계기는 지난 2018년 전 민주당원의 포털 뉴스 댓글조작 사건인 일명 '드루킹 사건' 이다. 당시 뉴스 편집 공정성 논란이 거세지자 한성숙 네이버 대표는 "네이버 편집자가 더 이상 기사를 배열하지 않겠다" 고 공식화했다. 이후 네이버는 초기 화면에서 뉴스를 없애고 검색창과 최소한의 정보만 남기는 방식으로 전환했다. 하지만 알고리즘이 정치적 상업적 편향성, 조작과 오류 가능성에서 자유로울 수 없다는 지적이 나온다. 알고리즘도 사람이 만드는 것이라는 점과, 인공지능은 스스로 반복 학습을 하면서 진화하는데 처음 설계했을 때보다 편향성이 커졌을 수도 있다는 우려가 크다. 자신만의 관심사나 관점에 갇히는 '필터 버블' 우려도 제기된다. 알고리즘의 지배 시대에 '알고리즘에 대한 감시' 가 새로운 과제로 떠오르고 있다. 이러한 점에서 우리나라 주식에서 코로나19 이후 제일 화제가 되는 주식이 카카오와 네이버이다. 각 포털이 서로 경쟁업체로서 뉴스 기사를 취사 선택하지 않고 있다면, 각 포털에서 각 포털과 관련된 주식에 관해 게시된 기사들이 차이가 없을 것이다. 이를 확인하기 위해 네이버뉴스와 다음뉴스에서 게시되는 기사 중에 '네이버 주식', '카카오주식'과 관련된 기사들을 텍스트 마이닝 기법을 통해 비교 분석했다.

텍스트 마이닝이란 비정형 데이터에 대한 마이닝 과정이다. 마이닝이란 데이터로부터 통계적인 의

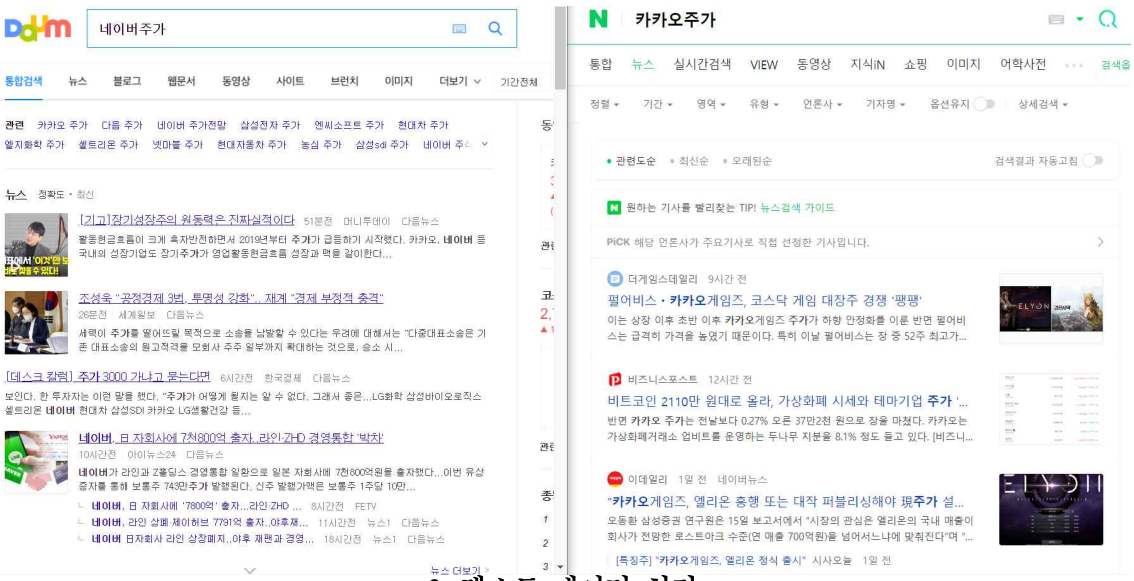
[1] 30019 세종특별자치시 세종로 2511, 고려대학교 세종캠퍼스 빅데이터전공
[2] 30019 세종특별자치시 세종로 2511, 고려대학교 세종캠퍼스 빅데이터전공
[3] 30019 세종특별자치시 세종로 2511, 고려대학교 세종캠퍼스 빅데이터전공

미가 있는 개념이나 특성을 추출하고 이것들 간의 패턴이나 추세 등의 고품질의 정보를 끌어내는 과정이다. 데이터는 형태에 따라 고정된 구조 형태로 구성된 데이터를 정형데이터로, 정해진 구조가 없을 때는 비정형 데이터로 구분한다. 정형 데이터를 이용한 마이닝을 데이터 마이닝, 비정형 데이터를 이용한 마이닝을 텍스트 마이닝이라고 한다.

텍스트 마이닝은 일반적으로 텍스트 범주화, 텍스트 클러스터링, 클러스터의 특징과 그것들의 관계를 기반으로 개념이나 특성 추출을 하고 개념과 특성 간의 관계 예측 등의 과정을 수행하게 된다. 본 연구는 비정형 데이터인 네이버와 다음에서의 뉴스기사를 이용하여 분석을 진행한다. 코로나19 이후 가장 화제 되고 있는 주식인 네이버주식과 카카오주식을 각 포털사이트에서 검색하여 나온 뉴스들을 기반으로 분석을 한다. 네이버와 다음이 경쟁사로서, 자신들과 관련된 주식기사를 더 많이 게시하고 있는지 빈도분석, 감성분석을 통하여 각 포털이 자신의 포털과 관련된 기사들을 더 긍정적으로 평가하고 있는지, 부정적으로 평가하고 있는지를 분석한다. 또한 각 키워드인 '네이버 주가' 와 '카카오 주가'를 바탕으로 워드클라우드를 추출하고 토픽 모델링을 진행하여 결과값이 차이가 있는지를 비교해 본다.

2. 데이터

본 연구에서는 11월 20일을 기준으로 네이버 뉴스에서 '네이버 주가', '카카오 주가'를 검색, 다음 뉴스에서 '네이버 주가', '카카오 주가'를 검색하여 나오는 최근 500개의 뉴스를 파이썬을 이용하여 크롤링하였다



3. 텍스트 데이터 처리

먼저, 네이버 포털에서 '카카오주가', '네이버주가'라는 키워드를 통하여 뉴스를 검색하고 다음 포털에서 '카카오주가', '네이버주가'라는 키워드를 통하여 뉴스를 검색했다. 이 과정에서 Python 3.6.1 버전에서 작성한 뉴스 수집기(크롤러)를 통하여서 검색된 뉴스를 날짜와 최신순으로 뉴스를 수집했다. 뉴스 기사 수집에는 python 3.6.1버전에서 작성한 웹크롤링 코드를 통해 네이버와 다음 포털에서 키워드 뉴스부분 검색 기능을 사용하였다. '네이버주가'와 '카카오주가'를 검색 키워드로 지정하고 페이지 수를 입력한 뒤 포털별로 검색한다. 검색결과는 Figure 2와 같은 화면으로 등장한다.

Figure 1. 다음 뉴스포털 검색 결과

본 검색 결과에서 나오는 기사들의 제목과 기사들을 python내에서 csv파일로 저장합니다. Python으로 작성한 뉴스 수집기(크롤러)에서는 포털 별 뉴스 검색 결과에서 나타나는 링크들에 있는 뉴스들을 수집하여 1~50페이지 정도의 뉴스데이터를 불러온다. 불러온 본문과 제목이 담긴 기사들 4개의 csv파일로 저장한다.

수집된 포털과 검색어 기준으로 기사들이 저장되어 있는 csv파일을 생성한다. csv파일은 수집된 기사 본문과 제목이 나열되어 있는 형태로 되어있다. 이 안에는 검색키워드와 크게 관련이 없는 내용이 담긴 기사들과 각종 특수문자들이 포함되어 있다. 그리하여 분석에 들어가기전 먼저 형태소분석에 python의 한국어 형태소 분석기(KoNLP)를 사용하였고 모듈은 그 속에 Okt를 사용하였다. 기사들의 명사를 모두 추출하여 특수문자들을 제거한 후 '네이버주가'와 '카카오주가'와 관련이 없는 주식 광고 뉴스에 들어가있는 다른 주식들을 불용어 처리사전에 넣었다. 그렇게 생성된 명사들을 저장한 파일을 이용하여 본 연구를 진행하였다.

4. 워드클라우드

네이버 뉴스 포털과 다음 뉴스 포털에서 각각 '카카오주가','네이버주가'라는 키워드를 통하여 다음(카카오주가), 다음(네이버주가), 네이버(카카오주가), 네이버(네이버주가) 총 4개의 데이터를 수집하였다. 파이썬에서 작성한 뉴스수집기(크롤러)를 통하여 검색어와 기간을 입력하여 최근 400~500개정도의 뉴스데이터를 각각 수집하였다. 4개의 분류된 데이터로 각각 워드클라우드를 그리고 비교했다. 순서대로 다음에서 네이버주가, 카카오주가를 검색한 데이터 네이버에서 네이버주가, 카카오주가를 검색한 데이터를 워드클라우드한 이미지이다.

워드클라우드를 통해 같은 검색어를 입력했지만 다른 양상을 보이는 것을 확인할 수 있다. 네이버에서는 네이버주가를 검색하면 네이버관련, 카카오주가를 입력하면 카카오관련 주가 기사가 많지만 다음에서는 네이버주가를 검색하면 네이버자체보다는 주가의 전반적인 기사가 많이 게시되는 것을 확인할 수 있다.

4. 감성 분석

기사본문에서는 긍부정 단어가 너무 섞여 있어 기사제목으로 감정분석을 시도하였다. 감성분석은 지도 기계학습기반 감성분석으로 LSTM 신경망을 이용하여 모델을 구축하였다. 지도 기계학습 기반 감성분석이란 인간이 만든 감성사전으로 기계학습을 시켜 모델을 만들고 새로운 들어왔을 때 긍정 부정 평가 하는 것이 목표이다. LSTM (Long Shot Term Memory)구조의 순환 신경망은 3개의 게이트를 사용하여, 시간 단위로 입력 노드를 통해 들어오는 데이터를 입력, 저장, 출력할 수 있도록 제어함으로써 일반 DNN 대비 장기기억(과거의 정보를 참조하는)에 대해 연결 성능치가 높은 신경망이다. 감정분석 모델은 인간 코더가 판단한 긍정 및 부정 라벨링으로 데이터를 생성하고 데이터 중 일부 문서가 훈련 데이터가 되어 기계학습 모델을 생성한 후 생성된 모델을 이용해 새로운 테스터 데이터가 들어왔을 때 긍정 및 부정 여부를 판별하는 방식이다.

먼저 감성사전은 주가에 대한 기사를 보면서 긍정단어와 부정단어를 구별하여 생성하였다.

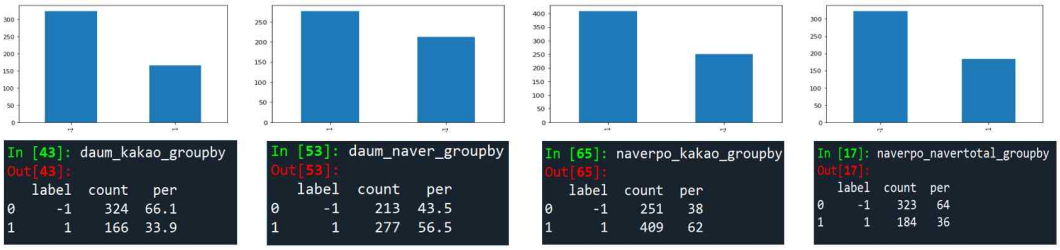
긍정	부정
↑	↓
급등	폭락
폭등	하락
상승	암울
웃음	.롤러코스터
최고점	사지마
화답	주춤
업	뿔난
뛴	백지화
...	...

만든 감성사전으로 각 데이터에 긍정 부정을 판별하여 부정 '-1', 긍정'+1'인 label변수를 추가하였다.

index	title	label
6	[경제 브리핑] 기관 '팔자'에 카카오게임즈 주가 7% 뚝	-1
7	[특징주] 카카오게임즈, 기관 보호매수 풀린 첫날 주가 7.9%↓	-1
9	카카오게임즈, 430만주 의무보유 해제 주가 5만 원대 무너지나	-1
11	카카오게임즈 임직원 첫 스톡옵션 행사... '주가 희석' 우려	-1
12	카카오게임즈 주가 급락에 상장 앞두고 답답한 빅히트	-1
.	...	-1

index	title	label
39	카카오게임즈 후광효과...카카오.넷마블 주가↑	1
40	카카오게임즈 '따따상. 기대하는 투자자	1
41	'거침없는' 카카오, 주가 40만원 첫 돌파	1
42	카카오 지붕 뚫고 최고가...주가 40만원 첫 돌파(종합)	1
43	장외주가 6만원 돌파... '카카오게임즈 대박, 엘리온에 달렸다'	1
...	...	1

'다음(카카오주가, 네이버주가)'와 네이버(카카오주가)에대한 뉴스기사제목과 라벨을 훈련데이터로 LSTM신경망을 이용해 모델을 만든 후 데스트데이터는 네이버포털에서 검색한 '네이버주가'에 대한 뉴스기사로 테스트를 하였다. 테스트결과 optimizer를 rmsprop로 설정했을 때 90.07, adam 설정시 88.30의 정확도가 나왔다. 만들어진 모델을 가지고 긍정 부정을 판별한 후 감정분석을 시도 했다. 감정분석은 다음과 네이버에서 각각 카카오주가와 네이버주가를 검색한 데이터에 긍정 부정 라벨을 붙인 데이터의 빈도를 통해 분석을 시도하였다.



```
In [43]: daum_kakao_groupby
Out[43]:
   label  count   per
0     -1    324  66.1
1      1    166  33.9
```

```
In [53]: daum_naver_groupby
Out[53]:
   label  count   per
0     -1    213  43.5
1      1    277  56.5
```

```
In [65]: naverpo_kakao_groupby
Out[65]:
   label  count  per
0     -1    251   38
1      1    409   62
```

```
In [17]: naverpo_navertotal_groupby
Out[17]:
   label  count  per
0     -1    323   64
1      1    184   36
```

분석결과 다음포털에서는 카카오주가에 대해 부정하는 기사가 많았고 네이버주가에 대해 긍정하는 기사가 많았다. 네이버 포털에서는 네이버주가에 대해 부정하는 기사가 많았고 카카오주가에 대해서는 긍정하는 기사가 더 많았다. 분석 전 각 포탈에서 자사와 관련된 검색어의 기사를 더 긍정적으로 게시 할 것으로 예상했지만 결과가 반대로 나왔다. 하지만 같은 언론사들이 기사를 내지만 게시하는 포털은 분명히 기사를 다르게 게시한다는 것을 알 수 있다.

5. 토픽 분석

토픽 분석이란 문서 내에서 은닉된 주제들을 찾아내기 위해 개발된 통계 추론 모델이다. 이는 문서라고 상정되는 단어의 결합으로 이루어진 의미를 가지는 결합체가 특정 주제들을 가지고 있을 것이라고 가정하는 것이다. 이와 같은 토픽 분석 중에서도 가장 대표적으로 사용되는 것이 LDA(latent dirichlet allocation) 알고리즘이다. 이 방법은 문서 내의 단어와 같은 관찰 가능한 변수들로 은닉된 주제와 같은 관찰되지 않는 변수들을 추론한다. 이를 통하여 각 문서들의 주제 비율, 단어들이 각 주제에 포함될 확률들을 알아낼 수 있다.

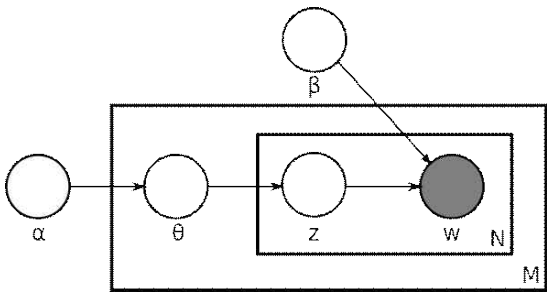


Figure 3. Graphics of LDA topic modeling

LDA 토픽 분석에서 단어는 이산 데이터의 기본 단위이며 각각의 단어는 단어집(vocabulary)의 인덱스 항목이며 $\{1,....,V\}$로 표기할 수 있다. 단어 벡터 w는 V-벡터로 표기하며 $w^v = 1$, $w^u = 0$, $u \neq v$를 만족한다. 또한 문서(document)는 N개의 단어들의 연속으로 $W = (w_1, w_2, ..., w_N)$으로 나타낸다. 코퍼스는 M개의 문서 집합으로 D=$\{W_1, W_2, ..., W_M\}$으로 표기한다.

이 때, LDA는 코퍼스 D에 있는 각각의 W에 대하여 여러 생성 과정(generative process)을 가정한다. 첫 째로 $N \sim Poisson(\xi)$을 선택한다. 둘 째, $\theta \sim Dir(\alpha)$를 선택한다. 셋 째로 문서 내의 단어 $w_n \in W$에 대해서 $z_n \sim Mltinomial(\theta)$를 선택하고 z_n이 주어졌을 때, w_n은 $p(w_n|z_n,\beta)$로부터 선택한다.

α는 각 문서가 어떠한 주제 비율로 구성될지를 나타내는 θ값을 결정하는 k차원 디리클레 분포의 매개변수이다. θ는 Dirichlet 분포를 따르는 k차원 벡터이며, 따라서 α값에 따라 θ가 분포하게 될 Dirichlet 분포의 형태가 결정된다. 이 때 θ^i는 문서가 i번째 주제에 속할 확률 분포를 나타낸다. z_n은 k차원 벡터이며 z_n^i은 단어 w_n이 i번째 주제에 속할 확률 분포를 나타낸다. β는 $k \times V$ 크기의 행렬 매개변수로, β_{ij}는 i번째 주제가 단어집 j번째 단어를 생성할 확률을 나타낸다.

이에 LDA 토픽 모형은 각 문서에 대해 k개의 주제에 대한 가중치 θ가 존재할 때, 문서 내의 각

단어 w_n은 k개의 주제에 대한 가중치 z_n을 가지는 데, 이 때 z_n은 θ에 의한 다항 분포로 선택된다. 마지막으로 실제 단어 w_n이 z_n에 기반하여 선택되는 것이다. 잠재 변수 α, β가 주어졌을 때 $\theta, z = z_1, \cdots, z_N, w$에 대한 결합 분포는 다음과 같다.

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta)$$

여기서 $p(z_n|\theta)$가 $z_n^i = 1$과 같은 유일한 i에 대한 단순한 θ^i라면, z과 θ의 적분 합을 통해 다음과 같은 문서의 주변 분포를 구할 수 있다.

$$p(w|\alpha, \beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum p(z_n|\theta)p(w_n|z_n, \beta)\right)d\theta$$

마지막으로 각 각의 단일 문서에 대한 주변 확률을 모두 곱하여 다음과 같은 코퍼스의 확률을 구할 수 있다(Blei at al., 2003).

$$p(D|\alpha, \beta) = \prod_{d=1}^{N}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_n} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn}, \beta)\right)d\theta_d$$

본 논문에서는 통계분석 프로그램인 Python을 활용하여 수집된 포털 별 뉴스 기사들에 대하여 토픽 분석을 진행하였다.

Table . Result of LDA topic modeling

Category	Keyword
Naver에서 네이버 주가 검색	카카오, 가치, 증가, 카카오게임즈, 카카오 뱅크, 상장, 올해, 증가
Naver에서 카카오주가 검색	투자리포트, 코스피, 하락, 삼성전자, 상승
Daum에서 네이버주가 검색	카카오, 카카오게임즈,상승, 다음카카오, 기대감, 목표, 주가
Daum에서 카카오 주가 검색	네이버, 목표, 라인, 미래, 하락, 목표주가, 급등

각 포털에서 네이버주가와 카카오주가를 검색한 데이터 분석 결과에서 포털별 키워드가 다르게 나타나고 있다. 같은 검색어를 검색하는 것인데도 포털별로 차이가 있고 나오는 많이 나오는 키워드가 달랐다.

6. 결론

본 연구는 경쟁사인 네이버와 카카오에서 '네이버 주가'와 '카카오 주가'를 검색하여 나온 각 500개의 뉴스 데이터를 이용하여 탐구를 진행하였다. 알고리즘의 지배 시대에 '알고리즘에 대한 감시'가 새로운 과제로 떠오르고 있고, 각 포털에서 뉴스기사를 취사 선택한다는 논란이 지속되기에 이 데이터로 각 포털 간의 게시된 기사들의 차이를 비교 분석해보았다. 뉴스라는 것은 언론사가 올리는 것이고 AI 시스템이 도입되었기 때문에 각 포털 간의 차이가 없을 것이라 예상했다. 하지만 같은 검색어로 워드클라우드를 추출해 보았을 때 나온 단어들이 각 포털에서 서로 차이를 보였다. 또한 토픽모델링의 결과로 나온 키워드를 비교해보아도 서로 다른 단어들이 나왔다. 감성분석에서도 긍부정 기사

개수를 비교해 보았을 때 포털별로 서로 반대의 결과가 나타났다. 이렇듯 각 포털 간에 같은 검색어를 검색을 하여도 서로 다른 결과값이 나오고 분명히 차이가 있다는 것을 알 수 있었다. AI 시스템이 뉴스기사를 편집하는 지금에도 포털 간에 차이가 존재하는 것을 보아 포털 뉴스 편집에 관한 논란은 앞으로도 계속 지속될 것이라고 판단된다. AI 알고리즘은 인간의 개입, 해석, 조작 등으로부터 자유로울 수 없다는 점을 분명히 인식하고, 알고리즘에 대한 사회적 비판과 감시가 필요할 것이다. 또한 알고리즘의 결과에 문제가 생기면 알고리즘 뒤에 숨지 말고 반드시 그 책임을 그 기업이 져야만 할 것이다.

References

http://www.munhwa.com/news/view.html?no=2020091501031607000001

http://www.hani.co.kr/arti/economy/it/820804.html

서대호. (2019). "잡아라!텍스트마이닝with 파이썬"

송민. (2017). "텍스트마이닝 Text Mining"