뉴스기사 분석을 통한 사회이슈와 가격에 관한 연구 - 조류인플루엔자와 달걀가격 중심으로 -

(Analysis of the Relations between Social Issues and Prices Using Text Mining - Avian Influenza and Egg Prices -)

한무명초*, Yangsok Kim**, 이충권**
(Mu Moung Cho Han, Yangsok Kim, Choong Kwon Lee)

요 약

조류인플루엔자는 전염 속도가 매우 빠르고 양계농장을 중심으로 생산자들과 소비자들에게 심각한 영향을 끼친다. 그중에서도 2016년 말에 전국적으로 발생한 조류인플루엔자는 좁은 공간에 밀집시켜 사육하는 산란계 농장에 큰 피해를 주었다. 이에 따라 달걀과 달걀을 재료로 하는 가공식품의 가격이 급등하였고 언론은 많은 속보성 뉴스기사를 게재하였다. 본 연구는 사회이슈를 반영한 온라인 뉴스기사의 키워드 변화와 달걀가격 변동과의 상관관계를 알아보고자 하였다. 이를 위하여 2016년 11월부터 14주 동안 한국에서 발생한 조류인플루엔자 관련 온라인 뉴스기사 682건과 같은 기간의 달걀가격 변화를 분석하였다. 본 연구의 결과는 사회이슈를 반영하는 뉴스기사의 키워드와 실물가격과의 관계를 이해하는 데 기여할 것으로 기대한다.

■ 중심어 : 텍스트 마이닝 ; 온라인 뉴스기사 분석 ; 다중선형회귀 ; 조류독감

Abstract

Avian influenza (AI) is notorious for its rapid infection rate, and has a serious impact on consumers and producers alike, especially in poultry farms. The AI outbreak, which occurred nationwide at the end of 2016, devastated the livestock farming industries. As a result, the prices of eggs and egg products had skyrocketed, and the event was reported by the media with heavy emphasis. The purpose of this study was to investigate the correlation between the egg price fluctuation and the keyword changes in online news articles reflecting social issues. To this end, we analyzed 682 cases of AI-related online news articles for fourteen weeks from November 2016 in South Korea. The results of this study are expected to contribute to understanding the relationship between the actual price of eggs and the keywords from news articles related to social issues.

■ keywords : Text Mining ; Online News Article ; Multiple Linear Regression ; Avian Influenza

Ⅰ. 서 론

우리나라에서는 1996년 3월 충청북도 음성에서 처음으로 저병원성 조류인플루엔자(Low Pathogenic Avian Influenza: LPAI)가 발생한 이후로 지금까지도 계속해서 발생하고 있다. 이러한 전염성 가축 질병의 발생은 축산업계의 생산자들과 소비자들에게 큰 피해를 발생시킨다. 특히 2003년 발생한 고병원성 조류인플루엔자(Highly Pathogenic Avian Influenza : HPAI)는 가금류를 넘어 인체에 감염될 수 있어서 그 위험성이 더욱 부각되었다. 이후로도 2006년, 2008년, 2011년, 2014년, 그리고 2016년에 이르기까지 간헐적이지만 지속적으로 발생하고 있다.

정부와 학계에서는 조류인플루엔자의 원인과 확산에 관하여 다양한 연구를 수행하였다[1,2,3]. 허덕 등[4]은 조류인플루엔자

* 정회원, 계명대학교 경영정보학과
** 정회원, 계명대학교 경영정보학과

접수일자 : 2017년 08월 22일
수정일자 : 2017년 11월 22일
게재확정일 : 2018년 3월 22일
교신저자 : 이충권 e-mail : cklee@kmu.ac.kr

Smart Media Journal / Vol.7, No.1 / ISSN:2287-1322
http://dx.doi.org/10.30693/SMJ.2018.7.1.46

가 발생하면, 가금류의 단기적인 살처분으로 공급이 감소하는 것보다 양계산물의 안전성에 대한 소비자의 불안감이 높아져 소비가 크게 위축된다고 하였다.

조류인플루엔자는 2016년 발생하여 양계농장을 중심으로 매우 빠르게 퍼져 나갔다. 특히 그중에서도 좁은 공간에 밀집시켜 사육하는 산란계 농장의 피해가 육계에 비교해 크게 발생했다[5]. 이러한 산란계 농장의 큰 피해로 달걀과 달걀을 재료로 하는 가공식품의 가격이 급등하고 많은 미디어가 속보경쟁을 하듯이 대량의 뉴스기사를 게재하였다. 기존의 가축 질병에 관한 정형적 데이터만으로는 사회이슈를 반영한 종합적인 분석에 한계가 있다. 따라서 본 연구에서는 온라인 뉴스기사를 통하여 사회이슈를 반영한 분석을 시도하였다. 거기에 더해 사회 현상에 민감하게 반응하는 실물가격 데이터를 결합하여 분석하였다.

최근 텍스트 마이닝 기법의 발달은 온라인 뉴스기사와 같은 비정형 데이터를 처리하여 새로운 정보를 만들어 내는 것을 가능하게 하였다. 온라인 뉴스기사는 정치, 경제, 사회, 문화 등 이슈의 변화를 빠르게 반영한다[6,7]. 특히 온라인 뉴스기사의 경우 토픽모델링이나 트렌드 분석 등을 이용하여 이슈의 동적 변화과정을 고찰할 수도 있다[8,9]. 안성원과 조성배[10]는 비정형 데이터인 뉴스기사의 키워드와 정형 데이터인 주식가격을 결합하여 주가예측을 시도함으로써 정형 데이터와 비정형 데이터의 결합을 통한 새로운 연구의 가능성을 보여주었다.

본 연구에서는 사회이슈를 반영하는 온라인 뉴스기사의 키워드 변화와 실물가격 변동과의 상관관계를 알아보는 것을 목적으로 한다. 이를 위하여 비정형 데이터와 정형 데이터를 결합한 분석 방법을 제안한다. 첫째, 온라인 뉴스기사에서 조류인플루엔자와 관련된 키워드 빈도를 추출한다. 둘째, 추출된 키워드를 바탕으로 시계열에 따른 변화를 분석한다. 셋째, 다중선형회귀모형을 이용하여 이슈 키워드들과 달걀가격과의 상관관계를 분석한다. 이러한 연구는 간헐적 또는 주기적으로 발생하는 사회이슈를 파악하고 이를 통해 대응책을 마련할 수 있다. 즉 주기적으로 발생하는 조류 인플루엔자를 비롯하여 가축 질병으로 인한 축산물 실물가격 변동의 관계를 예측함으로써 물가 흐름을 파악하고 대책을 마련할 수 있다. 또한, 다양한 분야에서 발생하는 비정형데이터와 정형데이터의 결합을 통한 분석으로 확장할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 조류인플루엔자, 텍스트 마이닝, 다중선형회귀분석과 관련된 기존의 연구들을 소개한다. 3장에서는 본 연구를 위한 데이터 수집과 연구 결과를 설명한다. 마지막으로 4장에서는 본 연구의 결론을 기술한다.

Ⅱ. 관련 연구

1. 조류인플루엔자 관련 연구

조류인플루엔자는 닭, 오리 등과 같은 가금류와 비둘기, 철새 같은 야생조류에서 발생하는 전염병이다. 지속적인 조류인플루엔자의 발생으로 이루어지는 대대적인 살처분이 경제적 손실 및 보상금 문제를 발생시키고, 지방자치단체의 재정적 부담이 된다. 유성희 등[11]은 이러한 관점에서 방역지침의 개정 필요성과 보상 법제 전반에 관한 연구를 수행하였다. 서정순 등[2]과 전영우 등[3]은 현재까지 국내에서 발생한 조류인플루엔자를 대상으로 발생 초기 이후의 확산 범위 또는 전파지역을 예측하기 위하여 법정 가축 전염병 발생 데이터베이스를 기반으로 발생지역 간의 연관성 분석을 수행하였다. 문운경 등[12]은 공간분석과 시간분석의 한계를 극복하기 위하여 2014년 발생한 HPAI(H5N8) 바이러스 분석 시 지리정보시스템과 연계하여 시·공간을 동시에 고려한 분석으로 HPAI(H5N8) 바이러스의 전파 및 확산을 연구했다. 신정화 등[13], 오광현 등[14], 그리고 이윤정[15]은 특정 바이러스의 변화에 대하여 지역적, 시기적 조사를 통해 향후 발생하는 조류인플루엔자에 대한 유전적 정보를 제공하고자 하였다. 이처럼 조류인플루엔자 발생 시 파급효과가 광범위하고 경제적 피해가 막대하므로 이러한 피해를 줄이기 위하여 많은 연구가 수행되었다.

그러나 이러한 연구들은 정형적인 데이터를 사용하였기 때문에 조류인플루엔자로 인한 다양한 사회이슈들을 분석할 수 없었다. Gim 등[16]은 가축 질병 뉴스를 부정적인 광고효과로 가정하고 이러한 뉴스가 돼지고기 수요에 미치는 경제적 파급효과를 분석하였다. 이는 비정형 데이터인 뉴스기사와 정형데이터인 달걀가격의 상관관계를 분석하려는 본 연구의 방법과 일치한다.

2. 텍스트 마이닝과 다중선형회귀 관련 연구

온라인 뉴스와 같은 비정형 텍스트 정보를 처리하는 텍스트 마이닝은 데이터 마이닝의 한 분야로서 많은 양의 텍스트에서 의미 있는 정보를 찾는 데 사용된다. 또한, 최근에 기계학습 알고리즘을 활용한 지식모델의 생성과 분석 방법이 다양한 분야에서 제시되고 있다. 전승수[17]는 텍스트 마이닝 분석을 통하여 방대한 비정형 데이터로부터 지식 모델을 구성하는 토픽 인자와 관계 노드를 생성하고 통합하는 방법과 이를 정형화하는 알고리즘을 제시하였다.

또한, 텍스트 마이닝은 다양한 연구 분야에서 지식맵을 형성하는 데 중요한 역할을 한다. 정용복과 박의섭[18]은 암석 공학 분야에서 저널의 제목과 키워드를 수집하고 연도별 빈도 행렬의 클러스터링 및 선형회귀기법을 통하여 키워드들의 추세를 분석하였다. Cohen과 Hersh[19]는 생물의학 연구에서 전체 생물의학 지식에 대한 접근성 향상, 생물의학 문헌의 특징에 대한 더 나은 이해가 가능하도록 텍스트 마이닝 시스템을 구축하였다. 송재국 등[20], 이준석 등[21], 그리고 전성해[22]는 특허의

내용분석을 통하여 개발현황과 기술 예측을 시도하였는데, 다중 선형회귀분석을 이용한 구체적인 빅 데이터 학습에 대한 사례 연구를 수행하였다.

박희진 등[23]은 대학수학능력시험 문제 출제 시 전문가에게 의존한 기존의 방식에서 벗어나 지금까지 시행된 모의고사 및 실제 시험을 통해 축적된 자료를 바탕으로 선형회귀 및 의사결정나무 분석을 수행하여 영어영역 문제의 난이도를 예측하는 모델을 구축하고 난이도 예측에 영향을 미치는 요소를 판별하였다. 김유신 등[24]은 비정형 텍스트로 구성된 뉴스들을 수집하여 주가와 어떤 관련이 있는지 분석하였는데, 뉴스기사의 감성분석 결과 값과 주가지수 등락은 유의한 관계를 가진다고 하였다.

Ⅲ. 데이터 수집 및 연구 결과

본 연구에서는 조류인플루엔자 발생으로 인한 사회이슈의 변화와 달걀가격과의 상관관계를 분석하기 위하여 데이터 수집, 전처리, 그리고 데이터 모델링 과정을 그림 1과 같이 수행하였다.

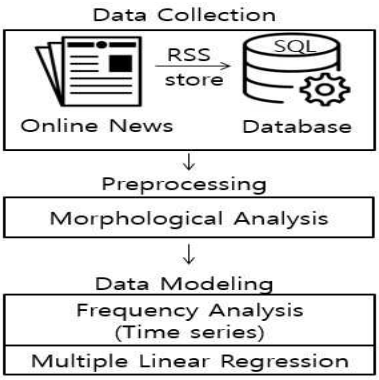

그림 1. 연구 프로세스

이를 위하여 전처리 프로그램인 한나눔 형태소 분석기와 데이터 마이닝 프로그램인 래피드마이너 7.3을 사용하였다. 래피드마이너는 텍스트 분석을 위해 Text Processing 확장 패키지를 제공한다.

래피드마이너는 데이터 과학 플랫폼 분야에서 성능이 검증된 소프트웨어로서 SAS 및 IBM과 같은 글로벌기업들과 경쟁하고 있다. Gartner Research는 2018년도 데이터 과학용 소프트웨어 보고서에서 래피드마이너가 시장을 리드하고 있는 소프트웨어 중의 하나이고, 사용 편의성뿐만 아니라 예측분석에서도 장점이 있음을 강조하였다[25]. KDnuggets는 2015년, 2016년 소프트웨어 프로그램을 사용하는 응답자를 대상으로 한 설문조사에서 가장 인기 있는 데이터 분석 소프트웨어로 래피드마이너를 선정했다[26][27]. 래피드마이너는 수백만 다운로드를 받았으며 유료고객으로 BMW, Intel, Cisco, GE 및 Samsung을 포함하여 25만 명이 넘는 사용자를 보유하고 있다. 래피드마이너는 이처럼 우수한 분석도구 이며 텍스트 분석에 있어서 비영어권의 언어 분석에도 활용할 수 있다[28].

1. 데이터 수집 및 전처리

본 연구를 수행하기 위하여 조류인플루엔자가 처음 발생한 2016년 11월 2주부터 14주 동안 RSS(Really Simple Syndication) 기반으로 제공되는 뉴스기사를 194개 웹사이트로부터 53,401건 수집하였다. 이렇게 수집된 기사 중 제목 및 요약에 '조류독감' 또는 '조류인플루엔자'라는 단어가 포함된 기사 689건을 추출하였으며 그중 중복 기사 7건을 제거하고 표 1에서 보듯이 682건의 기사를 연구에 사용하였다.

시계열에 따른 조류 인플루엔자 관련 기사의 추이를 알아보기 위하여 수집된 온라인 뉴스기사를 주(weekly) 단위로 정리하였다. 조류인플루엔자가 처음 발생한 11월 2주와 3주에는 이러한 이슈를 반영한 기사의 수가 미미하였다. 그러나 조류인플루엔자가 확산되고 달걀 수급의 불안정으로 달걀가격이 급등하면서 기사의 수도 12월 중순부터 급격히 증가하였다. 그 후 2017년 1월 말부터 조류인플루엔자 발생이 진정 국면으로 접어들면서 기사의 양이 줄어들었다.

또한, 조류인플루엔자가 발생한 기간의 달걀가격 데이터를 수집하기 위하여 축산물품질평가원에서 제공하는 축산유통종합정보센터(http://www.ekapepia.com)에 공개된 데이터를 활용하였다. 소비자가격을 제공하지 않는 공휴일에는 하루 전날의 가격을 적용하였고, 주별 평균을 구하여 분석에 사용하였다. 달걀가격은 1월 3주에 가장 높게 나타났고 그 추이는 그림 2와 같다.

표 1. 분석 데이터

주	기사 수	평균가격
11월 2주	4	5648.0
3주	6	5561.0
4주	46	5419.8
12월 1주	21	5525.0
2주	20	5702.0
3주	97	6138.6
4주	112	6882.6
5주	96	7973.4
1월 1주	65	8570.2
2주	85	9396.6
3주	43	9429.8
4주	21	9056.0
2월 1주	32	8742.0
2주	34	8154.0
합계	682	

Smart Media Journal / Vol.7, No.1 / ISSN:2287-1322
http://dx.doi.org/10.30693/SMJ.2018.7.1.48

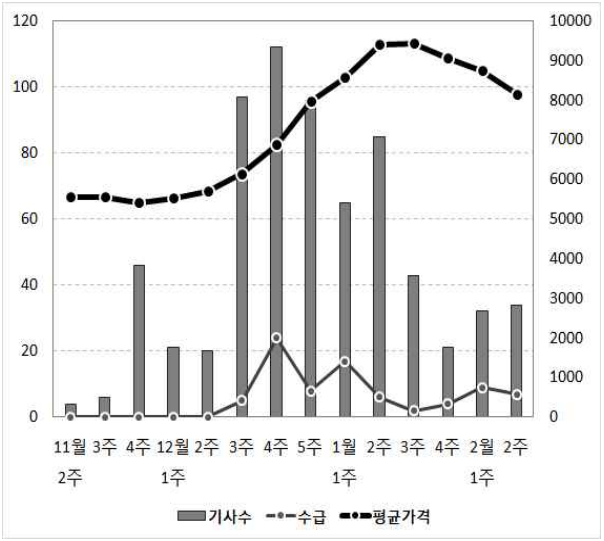

그림 2. 주별 기사 & 달걀가격 & '수급'키워드 추이

주 단위로 정리된 온라인 뉴스기사들로부터 키워드를 추출하기 위하여 한나눔 형태소 분석기를 이용해서 명사, 형용사, 부사를 추출하고 CSV 파일로 저장하였다. 이를 위하여 데이터베이스 테이블 속성에서 '제목'과 '기사내용' 속성을 사용하였다.

2. 빈도 분석

조류인플루엔자에 대한 사회이슈의 변화를 알아보기 위하여 주 단위로 온라인 뉴스기사의 단어 빈도를 용어-기사 매트릭스로 작성하였다. 첫째, 전처리된 CSV 파일을 로드한 후 래피드마이너가 제공하는 <Process Documents From Data> 오퍼레이터를 이용하여 키워드 빈도를 추출하였다. 그 결과로 '달걀', '조류', '인플루엔자', '고병원성', '확산' 등을 포함한 상위 20위 키워드를 생성하였다. 총 14주간의 주별 용어-기사 매트릭스를 작성할 상위 20위 키워드는 표 2와 같다.

표 2. 전체 기사 단어 빈도 20위

순위	키워드	빈도	순위	키워드	빈도
1	달걀	1030	11	수입	220
2	조류	825	12	축산	214
3	인플루엔자	676	13	전국	183
4	고병원성	391	14	식품	182
5	확산	342	15	농림	181
6	가격	325	16	농가	175
7	방역	307	17	인상	162
8	정부	292	18	지원	150
9	발생	242	19	가금	144
10	마리	230	20	살처분	142

둘째, 주 단위로 이슈들의 추이를 분석하기 위하여 <Process Documents From Data> 오퍼레이터를 사용하여 주별 키워드 매트릭스를 생성하였다. 주 단위 작업의 편의성을 위하여 <Multiply> 오퍼레이터와 <Filter Examples> 오퍼레이터를 사용하였다. 그 결과 14주간의 정형화된 키워드 빈도 테이블이 생성되었다. 이때 주 단위 키워드는 100위까지 생성하였으며, 전체 20위까지의 키워드 중에서 100위 이내에 포함되지 않은 키워드가 있는 경우에는 그 키워드 값을 0으로 설정하였다. 이렇게 만들어진 주 단위 키워드 빈도의 용어-기사 매트릭스는 표 3과 같다.

표 3. 주별 키워드 빈도 용어-기사 매트릭스

	1주	2주	3주	4주	5주	6주	7주	8주	9주	10주	11주	12주	13주	14주
조류	15	13	83	38	23	124	127	93	65	100	44	18	45	37
달걀	0	0	1	0	26	98	263	159	129	193	62	42	45	4
인플루엔자	8	8	55	28	21	97	112	84	59	80	39	18	34	33
고병원성	9	7	33	29	15	62	61	50	32	43	12	6	12	20
확산	2	2	35	18	24	81	74	42	19	12	5	5	5	18
가격	1	0	0	0	15	36	60	42	36	52	28	10	14	31
방역	1	2	28	9	14	71	38	37	29	20	20	15	6	17
정부	0	0	4	2	0	36	68	25	60	28	40	8	7	14
발생	1	1	13	7	6	26	39	45	27	21	12	9	0	35
마리	0	0	8	4	18	28	37	55	41	13	3	2	4	17
수입	0	0	0	0	0	4	53	9	53	64	16	17	2	2
축산	3	3	15	5	11	26	28	29	21	32	20	4	4	13
전국	0	3	13	9	20	27	29	30	10	14	6	6	0	16
식품	3	3	12	3	8	23	19	30	22	25	10	4	9	11
농림	3	3	13	3	7	21	24	26	22	27	10	4	4	14
농가	2	8	15	4	11	31	6	21	15	19	15	4	4	20
인상	0	0	0	1	7	35	33	19	15	25	10	0	1	16
지원	0	0	0	2	0	22	23	8	20	31	13	3	1	27
가금	0	0	11	0	10	36	14	24	19	7	6	5	3	9
살처분	0	4	4	3	10	21	29	18	20	12	13	1	2	5

셋째, 생성된 용어-기사 매트릭스를 시각화하였다. 키워드별 각 주의 빈도를 각 주의 전체 빈도로 나누었다. 이렇게 계산된 비율을 기준으로 14주 중 최댓값을 구하고, 각 키워드의 주 단위 비율을 최댓값으로 나누었다. 그림 3과 같이 키워드별 추이를 쉽게 인지할 수 있으며 주별 시간의 흐름에 따른 조류인플루엔자 이슈의 변화를 파악할 수 있다.

	11월2주	11월3주	11월4주	12월1주	12월2주	12월3주	12월4주	12월5주	1월1주	1월2주	1월3주	1월4주	2월1주	2월2주
조류	1.00	0.66	0.54	0.55	0.23	0.29	0.24	0.22	0.17	0.24	0.20	0.14	0.33	0.16
달걀	0.00	0.00	0.01	0.00	0.51	0.45	1.00	0.76	0.67	0.93	0.56	0.67	0.67	0.03
인플루엔자	1.00	0.76	0.68	0.76	0.39	0.42	0.40	0.38	0.29	0.36	0.33	0.27	0.47	0.27
고병원성	1.00	0.59	0.36	0.70	0.25	0.24	0.19	0.20	0.14	0.17	0.09	0.08	0.15	0.15
확산	0.51	0.39	0.88	1.00	0.91	0.72	0.54	0.39	0.19	0.11	0.09	0.15	0.14	0.30
가격	0.45	0.00	0.00	0.00	1.00	0.56	0.77	0.68	0.63	0.85	0.85	0.54	0.70	0.92
방역	0.36	0.55	1.00	0.71	0.75	0.89	0.39	0.49	0.41	0.26	0.49	0.65	0.24	0.40
정부	0.00	0.00	0.15	0.16	0.00	0.46	0.72	0.33	0.86	0.37	1.00	0.35	0.29	0.34
발생	0.44	0.33	0.56	0.66	0.39	0.39	0.49	0.71	0.46	0.33	0.35	0.47	0.00	1.00
마리	0.00	0.00	0.30	0.33	1.00	0.36	0.40	0.75	0.60	0.18	0.08	0.09	0.17	0.42
수입	0.00	0.00	0.00	0.00	0.00	0.06	0.65	0.14	0.89	1.00	0.47	0.88	0.10	0.06
축산	1.00	0.76	0.49	0.36	0.54	0.30	0.27	0.35	0.27	0.39	0.45	0.16	0.15	0.28
전국	0.00	0.77	0.43	0.66	1.00	0.31	0.28	0.37	0.13	0.17	0.14	0.24	0.00	0.35
식품	1.00	0.76	0.39	0.22	0.39	0.26	0.18	0.36	0.29	0.30	0.23	0.16	0.33	0.24
농림	1.00	0.76	0.43	0.22	0.35	0.24	0.23	0.31	0.29	0.33	0.23	0.16	0.15	0.31
농가	0.33	1.00	0.24	0.14	0.27	0.18	0.03	0.13	0.10	0.11	0.17	0.08	0.07	0.22
인상	0.00	0.00	0.00	0.18	0.86	1.00	0.78	0.57	0.48	0.75	0.56	0.00	0.09	0.87
지원	0.00	0.00	0.00	0.25	0.00	0.43	0.37	0.16	0.44	0.63	0.50	0.20	0.06	1.00
가금	0.00	0.00	0.73	0.00	1.00	0.84	0.27	0.59	0.50	0.17	0.27	0.41	0.23	0.40
살처분	0.00	1.00	0.13	0.22	0.49	0.24	0.27	0.21	0.26	0.14	0.29	0.04	0.07	0.11

그림 3. 시각화된 단어빈도 추이

‘조류’, ‘인플루엔자’, ‘고병원성’, ‘축산’ 등과 같은 키워드는 조류인플루엔자 발생 초기에 많이 나타났지만 ‘확산’, ‘방역’ 등과 같은 키워드는 조류인플루엔자 발생 3주부터 많이 나타나는 결과를 보인다. 특히 ‘달걀’ 키워드는 달걀가격의 추이를 나타내는 그림 2와 비슷하게 나타나고 있다. 또한, 정부에서 달걀 수입 발표를 함으로써 그 발표 주체인 ‘정부’라는 키워드도 ‘달걀’, ‘가격’과 밀접한 관계가 있음을 예상할 수 있고 분석결과 또한 비슷하다.

그러나 달걀가격에 영향을 크게 미치는 ‘수급’ 키워드가 전체 기사 단어 빈도 20위 안에 등장하지 않는다. 이는 조류인플루엔자 발생 6주부터 달걀가격이 급속하게 상승하면서, 처음으로 ‘수급’이라는 키워드가 5번 등장했다. 정부는 달걀가격을 안정시키기 위해 2016년 12월 19일 달걀 수급 계획을 발표했다. 그 영향으로 ‘수급’ 키워드가 조류인플루엔자 발생 7주에 가장 많은 24회 나타났다. 이후 조류인플루엔자 발생 9주를 제외하고는 10번 이하의 발생 빈도를 보인다. 따라서 ‘수급’이라는 키워드 빈도는 조류인플루엔자 발생 14주 동안의 전체 빈도를 기준으로 하는 상위 20위에는 등장하지 않았다.

3. 다중선형회귀분석

2016년 말 조류인플루엔자의 여파로 달걀 수급 부족이 물가 상승으로 이어졌다. 거기에 더해서 2017년 초 설날 명절의 영향으로 구매가 증가하면서 달걀가격은 급등하였다. ‘미국산’이라는 키워드는 달걀가격의 하락에 영향을 미친 것으로 판단된다. 달걀가격은 조류인플루엔자의 발생 초기에는 큰 변화가 나타나지 않았지만, 점차 확산되면서 달걀가격의 급등을 초래하였고 2016년 12월 19일 정부에서 외국으로부터 달걀을 수입하겠다는 발표에도 불구하고 가격의 상승세를 막지는 못하였다. 그러나 2017년 1월 13일 미국산 달걀의 수입이 시작된 이후부터는 달걀가격이 하락하는 것을 확인할 수 있었다. 이는 미국산 달걀을 수입함으로써 달걀의 공급이 원활해짐에 따라 가격이 하락하는 추세를 보인다. 따라서 조류인플루엔자와 달걀가격의 상관관계 분석을 위하여 다중선형회귀분석을 하였다.

조류인플루엔자를 반영한 데이터를 대상으로 14주 동안의 주별 키워드 추세와 달걀가격의 상관관계가 높은 키워드를 찾기 위하여 래피드마이너가 제공하는 <Linear Regression> 오퍼레이터를 이용하여 다중선형회귀분석을 실시하였다. 종속변수는 달걀가격으로 하였고, 독립변수는 주 단위 키워드 빈도 100위까지의 키워드를 사용하였다. 따라서 독립 변수에 비해 데이터 세트의 관측치가 적은 문제가 있었다. 이를 보완하기 위하여 <Bootstrapping> 오퍼레이터를 이용해서 데이터 관측치를 23배 늘렸다. 다중선형회귀분석에서 너무 많은 변수로 인한 과대모형 설명력을 제거하기 위하여 <Forward Selection> 오퍼레이터를 이용하여 모형에 대한 설명력이 높은 변수를 선택하였다. 전진선택(Forward Selection)은 회귀방정식을 적합 시키기 위하여 가장 중요한 변수들을 찾는 방법의 하나이다. 처음 한 개의 변수부터 시작해 한 번에 하나씩 변수를 모델에 추가해가며 모델에 적합한 변수를 찾는 작업을 반복하였다.

다중선형회귀 모델을 적용하여 달걀가격과 이슈 키워드의 관계를 알아본 결과는 표 4와 같다. 전진선택법에 의해 선택된 ‘지난해’, ‘물가’, ‘명절’, ‘여파’, ‘미국산’, ‘소비자’, ‘지원’, ‘처음’, ‘방역’, ‘기준’이라는 변수는 p-Value 값이 .05 이하로 모두 유의한 변수이다. 그중에서 ‘지난해’, ‘물가’, ‘명절’, ‘여파’, ‘지원’은 달걀가격에 양의 영향을 미치나, ‘미국산’, ‘소비자’, ‘처음’, ‘기준’, ‘방역’은 달걀가격에 음의 영향을 미치는 것으로 나타났다.

표 4. 다중선형회귀 결과

Attribute	Coefficient	Std.Error	Std. Coefficient	t-Stat	p-Value	Code
미국산	-35.12	0.14	-0.31	-252.52	0.00	****
소비자	-26.80	0.12	-0.13	-216.07	0.00	****
처음	-18.89	0.44	-0.05	-42.90	0.00	****
기준	-13.37	0.43	-0.04	-30.85	0.00	****
방역	-2.66	0.07	-0.03	-40.13	0.00	****
지원	14.88	0.07	0.10	199.52	0.00	****
물가	31.12	0.02	0.15	1561.11	0.00	****
여파	75.90	0.21	0.35	365.68	0.00	****
지난해	163.37	0.07	0.61	2361.53	0.00	****
명절	187.93	0.24	0.68	791.47	0.00	****
(Intercept)	5581.84	0.57	NaN	9775.34	0.00	****

Ⅳ. 결론

본 연구는 조류인플루엔자와 관련된 온라인 뉴스기사의 키워드 변화를 분석하여 달걀가격과의 관계를 알아보는 것을 목적으로 하였다. 이를 위하여 RSS에서 제공하는 온라인 뉴스기사 682건에서 주 단위로 키워드 빈도를 생성하였다. 또한, 주 단위 평균 달걀가격을 수집하고 이를 결합하여 용어-기사 매트릭스를 생성하였다.

이슈 트렌드 파악을 위하여 전체 키워드 빈도의 상위 20개를 선택하여 시각화하였다. 분석결과는 ‘달걀’, ‘가격’, ‘정부’ 키워드는 조류인플루엔자 발생 중기에 높은 빈도를 보인 것으로 나타났다. 다음으로 다중선형회귀 모델의 적용결과는 달걀가격 상승에 영향을 미치는 키워드가 ‘명절’, ‘물가’ 등인 것으로 나타났다. 반면에, 하락에 영향을 미친 키워드는 ‘미국산’, ‘소비자’ 등이라는 것을 파악할 수 있었다. 이처럼 다중선형회귀 모델의 분석결과 달걀가격에 중용한 영향을 미친 키워드는 ‘미국산’, ‘소비자’, ‘처음’, ‘기준’, ‘물가’, ‘여파’, ‘지난해’, ‘명절’로 조류인플루엔자 전체 기사 단어 빈도 20위에는 없는 키워드이다. 이는 14주 동안의 조류인플루엔자를 대표하는 높은 빈도의 키워드보다 미국산 달걀 수입과 같은 특정 사건을 대표하는 키워드가 중요함을 보여준다.

본 연구는 사회이슈 문제를 반영한 비정형 데이터인 온라인

Smart Media Journal / Vol.7, No.1 / ISSN:2287-1322
http://dx.doi.org/10.30693/SMJ.2018.7.1.50

뉴스기사와 정형 데이터인 달걀가격을 결합하여 분석하였다는 점에서 중요한 의의가 있다. 조류인플루엔자뿐만 아니라 다양한 전염성 질병으로 인한 사회이슈와 관련되어 나타나는 현상에 본 연구방법을 적용해 볼 수 있을 것으로 기대된다.

토픽모델링(Topic Modelling)은 주어진 문서에서 발견된 단어들의 분포를 분석함으로써 해당 문서가 어떤 주제를 가졌는지 파악하는 텍스트 분석 기법이다[29]. 따라서 향후 연구에서는 단순히 단어를 활용하는 것보다, 신문기사의 내용에 나타나는 주제(토픽)들을 찾아내어 상품의 가격과 어떤 관계가 있는지 연구하는 것이 필요하다.

REFERENCES

[1] 모인필, “고병원성 조류인플루엔자의 역학적 특성과 국내 발생 동향,” *한국위기관리논집*, 제3권, 제2호, 30-37쪽, 2007.

[2] 서정순, 이종욱, 박대희, 정용화, “고병원성 조류인플루엔자 전염병의 발병지역 연관성 분석,” *한국인터넷정보학회 학술발표대회 논문집*, 189-190쪽, 2015.

[3] 전영우, 이소희, 구신희, 심성삼, 박영진, “공공데이터 기반 조류인플루엔자 발생지역 예측에 관한 연구”, *한국지형공간정보학회 춘계학술대회*, 143-144쪽, 2015.

[4] 허덕, 우병준, 이형우, “고병원성 조류인플루엔자 발생이 양계산물 가격에 미치는 영향,” *농정연구속보*, 제48권, 1-13쪽, 2008.

[5] 농림축산식품부, http://www.mafra.go.kr/FMD-AI/05/01_07.jsp,(2017.03.24. 검색)

[6] L. Li, J. Hou, Z. Wang, J. Tang, P. Zhang, R. Yang, and Q. Zheng, “NewsMiner: multifaceted news analysis for event search,” *Knowledge-Based Systems*, vol. 76, pp. 17-29, 2015.

[7] 한무명초, 김양석, 이충권, “텍스트 마이닝 기법을 활용한 동남권 신공항 신문기사 분석,” *스마트미디어저널*, 제6권, 제1호, 47-53쪽, 2017.

[8] 임명수, 김남규, “비정형 텍스트 분석을 활용한 이슈의 동적 변이과정 고찰,” *지능정보연구*, 제22권, 제1호, 1-18쪽, 2016.

[9] 차준석, 김정인, 김판구, “단어 간 의미적 연관성을 고려한 어휘 체인 기반의 개선된 자동 문서요약 방법,” *스마트미디어저널*, 제6권, 제1호, 22-29쪽, 2017.

[10] 안성원, 조성배, “뉴스 텍스트 마이닝과 시계열 분석을 이용한 주가예측,” *한국컴퓨터종합학술대회 논문집*, 제37권, 제1호, 364-369쪽, 2010.

[11] 유성희, 이진홍, 김동련, “조류인플루엔자(AI) 발생으로 인한 보상제도의개선방안에 관한 연구,” *건국대학교 법학연구소 일감법학*, 제29권, 제0호, 219-246쪽, 2014.

[12] 문운경, 조성범, 배선학, “2014년 국내 발생 HPAI (고병원성 조류인플루엔자)의시, 공간 군집 분석,” *한국지리정보학회지*, 제18권, 제3호, 89-101쪽, 2015.

[13] 신정화, 왕승준, 정집설, 김용관, 엄재구, 정원화, 모인필, “야생조류의 조류인플루엔자 바이러스 분포와 이동경로 추적조사,” *한국가금학회 정기총회 및 학술발표회*, 10-12쪽, 2016.

[14] 오광현, 배연지, 이승백, 모종석, 모인필, “한국 야생조류에서 분리된 저병원성 조류인플루엔자 바이러스의 아미노산 변화에 대한 조사,” *한국가금학회 정기총회 및 학술발표회*, 43-44쪽, 2016.

[15] 이윤정, “최근 조류인플루엔자 현황과 바이러스 특성,” *대한인수공통전염병학회 학술발표초록집*, 165-184쪽, 2016.

[16] U. S. Gim, S. H. Choi, and J. H. Cho, “An Impact Analysis of FMD News on Pork Demand in Korea,” *The Korean Journal of Community Living Science*, vol. 26, no. 1, pp. 75-85, 2015.

[17] 전승수, “의미 기반의 지식모델 통합과 탐색에 관한 연구,” *인터넷정보학회논문지*, 제15권, 제6호, 99-106쪽, 2014.

[18] 정용복, 박의섭, “텍스트 마이닝을 이용한 암반공학분야 SCI 논문의 주제어 분석,” *터널과 지하공간*, 제25권, 제4호, 303-319쪽, 2015.

[19] A. M. Cohen and W. R. Hersh, “A survey of current work in biomedical text mining,” *Briefings in bioinformatics*, vol. 6, no. 1, pp. 57-71, 2005.

[20] 송재국, 류태규, 윤장혁, “특허권리성에 영향을 미치는 요인에 대한 연구: IT 관련 한국특허의 내용분석을 중심으로,” *Entrue Journal of Information Technology*, 제11권, 제3호, 57-68쪽, 2012.

[21] 이준석, 이준혁, 김갑조, 박상성, 장동식, “데이터 마이닝을 통한 기술경영 전략 수립에 관한 연구,” *한국지능시스템학회 논문지*, 제25권, 제2호, 126-132쪽, 2015.

[22] 전성해, “특허분석을 위한 빅 데이터학습,” *한국지능시스템학회 논문지*, 제23권, 제5호, 406-411쪽, 2013.

[23] 박희진, 장경애, 이윤호, 김우제, 강필성, “데이터 공학: 데이터마이닝 기법을 활용한 대학수학능력시험 영어영역 정답률 예측 및 주요 요인 분석,” *정보처리학회논문지. 소프트웨어 및 데이터 공학*, 제4권, 제11호, 509-520쪽, 2015.

[24] 김유신, 김남규, 정승렬, “뉴스와 주가: 빅데이터

감성분석을 통한 지능형 투자의사결정모형," *지능정보연구*, 제18권, 제2호, 143-156쪽, 2012.
[25] C. J. Idoine, P. Krensky, E. Brethenoux, J. Hare, S. Sicular, and S. Vashisth, "Magic Quadrant for Data Science and Machine-Learning Platforms," *Gartner Reprint*, pp. 1-38, February 2018.
[26] KDnuggets Annual Software Poll:RapidMiner and R vie for first place, KDnuggets(2013), https://www.kdnuggets.com/2013/06/kdnuggets-annual-software-poll-rapidminer-r-vie-for-first-place.html(accessed Mrch,17,2018).
[27] KDnuggets 15th Annual Software Poll:RapidMiner continues to lead., KDnuggets (2014), https://www.kdnuggets.com/2014/06/kdnuggets-annual-software-poll-rapidminer-continues-lead.html(accessed Mrch,17,2018).
[28] M. K. Saad and W. Ashour, "Arabic Morphological Tools for Text Mining," *International Conference on Electrical and Computer Systems (EECS'10)*, pp.1-6, Nov2010.
[29] Topic model(2018), https://en.wikipedia.org/wiki/Topic_model(accessed Mrch,17,2018).

저 자 소 개

한무명초(정회원)
2006년 방송통신대학교 컴퓨터과학학과 학사 졸업.
2009년 계명대학교 전산교육학과 석사 졸업.
2016년 계명대학교 경영정보학과 박사 졸업.

<주관심분야 : 데이터마이닝, 정보기술, 지식관리>

Yangsok Kim(정회원)
1995년 서울시립대하교 경제학과 학사 졸업.
2004년 University of Tasmania 컴퓨터 공학 석사 졸업
2009년 University of Tasmania 컴퓨터 공학 박사 졸업.

<주관심분야 : Big Data, Data Analytics, Knowledge-Based System>

이충권(정회원)
1995년 계명대학교 경영정보학과 학사 졸업.
1999년 Southeast Missouri State University MBA 졸업
2003년 University of Nebraska-Lincoln 박사 졸업.
2003-2006년 Georgia Southern University 조교수

<주관심분야 : Big Data, Text Mining, IT Jobs>