언론사 프레임 분석을 위한 벡터기반의 단어 표현: 코로나 19 를 중심으로

이다인*, 김유섭**
*한림대학교 융합소프트웨어학과
**한림대학교 소프트웨어융합대학
Dainee96@gmail.com, Yskim01@hallym.ac.kr

Vector-based word representation for media frame analysis: focused on covid-19

Da-In Lee*, Yu-Seop Kim**
*Convergence Software, Hallym University
**College of Software, Hallym University

요 약

본 논문에서는 언론사 프레임 분석을 위해 2020 년 2 월 1 일부터 7 개월간 코로나 19 를 언급한 기사 데이터를 수집하여 단어 임베딩을 수행하고, 언론사별 중복단어 행렬로 K-Means Clustering 을 수행하여 군집별로 모인 언론사들의 분포를 살펴본다. 또한, 언론사별 중복되지 않는 유일단어들의 긍정, 부정, 정치적, 경제적 등의 특성에 따라 프레임을 분석하여 파악한다. 이를 통해, 특정 기간동안 코로나 19 관련 기사에서 나타나는 언론사별 프레임을 비교 및 분석하고자 한다.

1. 서론

2019 년 12 월 31 일 중국 후베이성 우한에서 처음 발생된 코로나 19 는 현재까지도 전 세계에 기하급수적으로 감염되어 인류가 고통받고 있다. 국내에서는 2020 년 1 월 20 일 첫 확진자가 발생하였고, 그 이후로 국내 확진자 수는 점진적으로 증가세를 보이고 있다. 이와 같이 코로나 19 는 사회, 경제적으로 극심한 변화를 일으켰으며, 현재 전세계적으로 가장 큰 영향력을 행사하고 있다. 코로나 19 에 대해 정확한 정보를 전달해야 하는 각 언론사들의 역할은 감염의 확산과 예방, 치료와 같은 국민의 건강을 위해 매우 중요하다. 그러므로 언론사 프레임 분석을 통해 각 매체별로 중요하게 생각하는 단어의 우선순위를 부여하고 코로나 19 이슈에 대한 여론의 형성을 분석하고자 한다.

프레임 분석이란 동일한 사건에 대해 그 사건의 해석이나 의사결정이 달라지는 인식의 왜곡현상을 말하며, 구조화 효과 라고도 한다. 예를 들어, 긍정적인 프레임에 속한 해석은 불확실한 이득보다는 확실한 이득을 선호하고, 부정적인 인식의 프레임에 속한 해석은 확실한 손실보다는 불확실한 손실을 선호한다. 언론사 프레임을 분석함으로써, 코로나 19 에 대한 다양한 해석을 분석하고 각각의 해석에 대한 해결책을 찾아보는 것 또한 가능하다.

본 논문에서는 2020 년 2 월 1 일 ~ 8 월 30 일 까지 보도된 11 개의 중앙지 (경향신문, 국민일보, 내일신문, 동아일보, 문화일보, 서울신문, 세계일보, 조선일보, 중앙일보, 한겨레, 한국일보)와 8 개의 경제지 (매일경제, 머니투데이, 서울경제, 아시아경제, 아주경제, 파이낸셜뉴스, 한국경제, 헤럴드경제), 5 개의 방송사 (KBS, MBS, OBS, SBS, YTN) 에서 코로나 19 를 언급한 기사 데이터를 수집한다. 그 후, 전체 기사들을 각 언론사별로 구분하여 단어 임베딩을 수행하고 코로나 19 와 유사한 벡터표현을 가진 단어 상위 100 개씩 추출한다. 또한, 추출된 단어들의 중복단어 행렬에 대해 차원을 축소한 후 K-Means Clustering 을 수행하고 그 분포를 비교한다. 최종적으로, 정렬된 각 언론사별 유일한 단어들을 파악하여 프레임을 분석한다.

2. 관련 연구

뉴스 프레임 분석에 대해 연구한 논문들 중 [1]은 코로나 19 와 같이 바이러스인 에이즈 (AIDS)를 중심으로 분석하고, 이 이슈에 대한 각 언론사별 주제 프레임과 뉴스 정보원과의 관계를 파악한다. [2]는 특정 생명과학 뉴스에 대한 국내와 미국의 언론사별 주요 프레임을 비교함으로써, 뉴스 프레임의 형식적 특성을 파악한다. [3]은 Goff man 프레이밍 이론을 적용하

여 질적 뉴스 프레임 분석의 방법론을 제시한다. 사건에 있어서 사람들이 인식하고 식별하는 것을 해석의 스키마에 레이블을 지정하여, 뉴스 프레임을 정의한 후 분석하는 방법을 설명한다.

뉴스 프레임과 관련이 없지만 본 논문과 비슷한 분석방식을 사용한 연구들도 있다. 그 중 [4]은 과학 출판물에 대한 단어 추출 및 생성을 위해 단어 임베딩 기반의 접근 방식을 제안했으며, [5]는 단어의 벡터 표현을 문맥에서 학습하는 방식인 CBOW 모델에 가중치를 적용하여 의료 관리 트윗을 분류했다. 이와 같이 단어 임베딩을 사용하면 문맥으로부터 단어벡터를 생성하여 그 활용성은 무궁무진하다.

3. 언론사 프레임 분석

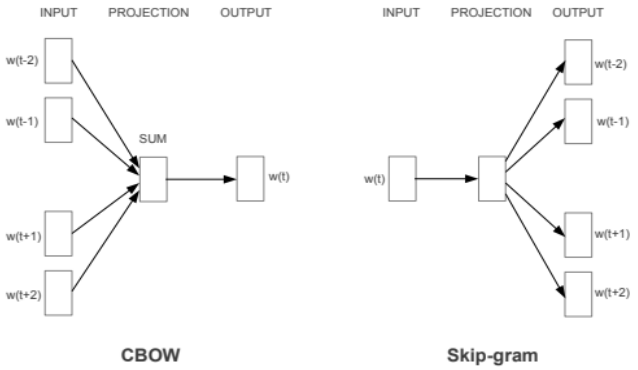
3.1 데이터수집 및 단어 임베딩

코로나 19 를 언급한 2020 년 2 월 1 일 ~ 8 월 30 일까지의 뉴스기사 데이터를 한국언론재단의 KINDS 뉴스 검색을 통해 중앙지 (경향신문, 국민일보, 내일신문, 동아일보, 문화일보, 서울신문, 세계일보, 조선일보, 중앙일보, 한겨레, 한국일보) 80,201 개, 경제지 (매일경제, 머니투데이, 서울경제, 아시아경제, 아주경제, 파이낸셜뉴스, 한국경제, 헤럴드경제) 78,417 개, 방송사 (KBS, MBS, OBS, SBS, YTN) 20000 개로 총 178,614 개를 수집한다. 해당 데이터에서 각 기사마다 순차적으로 구분되어 있는 단어들을 중심으로 활용한다.

단어 임베딩을 위해 genism 패키지의 Word2Vec 모델을 통해 학습하여 각 단어의 벡터를 생성한다. 주변 단어를 보고 중심 단어를 예측하는 CBOW (Continuous Bag-of-Words) 방식을 사용하고 생성된 벡터의 크기 (size)는 300 으로 제한한다. 또한, 학습 시 고려할 앞뒤 단어 폭 (window size)은 5, 최소 단어빈도 수 (min_cou nt)는 5, 동시에 처리할 작업 수 (workers)는 4 로 설정하여 학습한다. 이를 통하여 각 단어들은 문맥 정보가 포함된 300 차원의 벡터로 변환되고, 두 벡터 간의 비교의 척도인 코사인 유사도가 높은 순서대로 내림차순 정렬을 한다. 의미가 없는 단어들은 수작업으로 제거하고 그 결과로 코로나 19 와 코사인 유사도가 높은 단어들을 100 개씩 추출한다.

3.2 Word2Vec

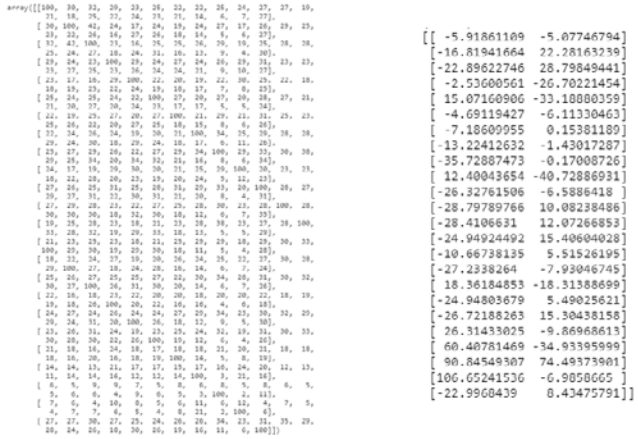
Word2Vec [6]은 2013 년에 제안된 모델로, NNLM 신경망 기반 언어 모델을 효율적 학습이 가능하도록 만든 모델이다. 언어모델 (Language Model)이란 이전에 나온 단어를 바탕으로 다음 단어를 예측하는 모델을 말하며, Word2Vec 모델은 CBOW 와 Skip-Gram 두 가지 방식이 있다. (그림 1)과 같이 전자는 주변단어를 통해 중심단어를 예측하고, 후자는 중심단어로 주변단어를 예측한다. 일반적으로, Skip-Gram 방식이 CBOW 방식보다 성능이 더 좋은 편이지만, 본 논문에서는 CBOW 방식으로 벡터기반의 단어표현을 생성하였을 때 더 좋은 성능을 보인다.



(그림 1) Word2Vec 모델의 CBOW, Skip-Gram 방식

3.3 중복단어 행렬 및 차원축소

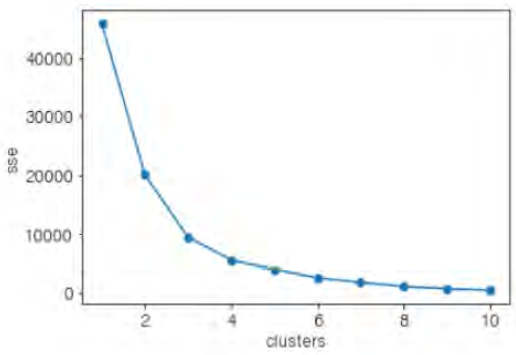
각 언론사 별로 코로나 19 와 코사인 유사도에 따라 내림차순으로 나열된 100 개의 단어들을 비교하기 위해 중복단어 행렬 (24x24)을 구한다. 일반적인 모든 데이터 행렬은 주성분 분석 (Principal Component Analysis, PCA)로 단순화하는 것이 가능하고, 동일한 데이터에 대한 예측을 하기위해 사용된다 [7]. 다음 (그림 2)와 같이 행렬에 주성분 분석을 수행하여 기존 데이터를 유지한 채, 24 차원을 2 차원으로 축소 (24x2) 하여 수행한다. 이처럼 언론사별 중복단어 행렬 벡터를 통해 유사한 위치정보를 파악한다.



(그림 2) 중복단어 행렬 (좌)과 차원축소 행렬 (우)

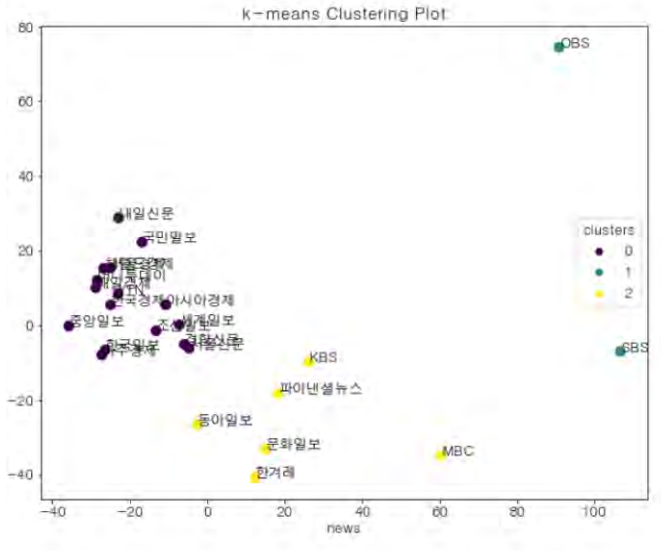
3.4 K-Means Clustering 수행

K-Means Clustering 이란, 클러스터 내 오차제곱합 (SSE)의 값이 최소가 되도록 클러스터의 중심을 결정하는 방법이다. 차원 축소된 행렬을 통해 K-Means Clustering 을 수행하기 앞서 최적의 클러스터 개수를 찾기 위한 Elbow 기법을 활용한다. 이 기법은 클러스터 개수에 따른 SSE 의 값이 있을 때, 급격하게 줄어들다가 완만해지는 구간이 생기는 팔꿈치 부분과 같은 지점이 최적의 클러스터 개수이다. (그림 3)과 같이, 클러스터 개수가 3 개일 때 최적의 개수이며, 이를 적용하여 K-Means Clustering 을 수행한다.



(그림 3) 최적의 클러스터 개수(Elbow 기법)

Elbow 기법을 통해 정해진 클러스터 수로 K-Means Clustering 을 한 결과는 다음 (그림 4)과 같은 군집을 이룬다. 먼저, Cluster0 는 헤럴드경제, 아시아경제, 한국경제 아주경제, 서울경제, 세계일보, YTN, 중앙일보, 매일경제, 머니투데이, 내일신문, 국민일보, 조선일보, 서울신문, 한국일보, 경향신문으로 데이터 포인트 사이의 거리가 상당히 좁다. Cluster1 은 SBS, OBS 로 다른 언론사들의 군집과 먼 거리에 위치하고 두 데이터 포인트의 거리도 멀다. Cluster2 는 파이낸셜뉴스, 문화일보, 한겨레, 동아일보, MBC, KBS 이며 데이터 포인트들 사이의 거리가 Cluster0 보다 멀고 Cluster1 보다 가깝게 위치한다. 그러므로, 비교적 군집을 잘 이룬 클러스터는 Cluster0 이며 Cluser1 은 매우 동떨어져 있다. 이와 같은 그래프는 각 언론사별로 비슷한 프레임을 가진 군집을 파악하는데 용이하다.



(그림 4) K-Means Clustering 그래프

4. 프레임 분석결과

다음 <표 1>는 다른 언론사와 중복되지 않은 상위 유일단어로 프레임을 분석한다. 유일단어는 해당 언론사에서만 나온 단어를 의미하는 것이 아닌, 다른 언론사들에 비해 더 중요하게 여기는 단어이다. 이를 통해, 각 언론사별 사회적, 경제적, 국가적, 국제적인 희망, 위기, 대응과 같은 특성을 찾아 비교 분석한다. 동일한 Cluster 내에 속한 언론사들은 기본적으로 중복단어가 많다는 전제에 있으므로 유일단어가 다르다고 해서 프레임이 크게 다르지 않다.

<표 1> 언론사별 상위 유일단어

언론사(cluster) '프레임 분석'	중복단어를 제거한 후의 유일단어 top10
경향신문(0) '사회적 희망 프레임'	결핍, 병원장, 안정화, 중앙임상, 자연적, 확산방지, 악조건, 온정, 파동, 비결, 메디시티
국민일보(0) '경제적 희망 프레임'	잔불, 방도, 식량안보, 무역업계, 방어선, 미흡, 대변화, 점입가경, 여행수요, 복원력
동아일보(2) '사회적 대응 프레임'	평온, 광운학원, 동시다발, 감염력, 문화생활, 대처법, 연구기관들, 재조명, 이색적, 주택시장
문화일보(2) '경제적 프레임'	헌혈운동, 파급, 인류, 고통, 위기의식, 소비심리, 산업계, 저개발국, 경제전문가, 건설업계
세계일보(0) '국가적 위기 프레임'	초당적, 반부패, 단기직, 전파경로, 비상경영체제, 한타바이러스, 오찬간담회, 절실, 국가기관, 대재난
조선일보(0) '경제적 위기 프레임'	경기침체, 일본프로야구, 변종바이러스, 재발령, 집단, 홈쿡, 의원발, 중소상공인들, 연쇄적, 요동
중앙일보(0) '국제적 위기 프레임'	외식업계, 세계시민, 해빙, 당면, 비상체제, 일본사회, 비상등, 마비, 패닉상태, 초기대응
한겨레(2) '사회적 위기 프레임'	위기, 방역체계, 예측, 참상, 방송가, 의사단체, 부도, 보건학, 치료약, 희망적
매일경제(0) '경제적 위기 프레임'	줄도산, 진퇴양난, 염려, 집중호우, 비극적, 중대고비, 실적악화, 큐코노미, 사회활동, 콜센터발
서울경제(0) '경제적 대응 프레임'	극복방안, 수출업계, 명륜진사갈비, 산업전략, 워킹맘들, 면세점들, 공연예술계, 소식지, 방제, 대응전략
아주경제(0) '경제적 대응 프레임'	추가대책, 혈관, 금리정책, 의학적, 수보회의, 유통가, 여행시장, 공세적, 앱노멀, 민생안정
파이낸셜뉴스(2) '경제적 위기 프레임'	사멸, 사회변화, 민생경제, 경제질서, 부정, 돌발변수, 조직력, 부산국제광고제, 거시경제, 내수침체
한국경제(0) '경제적 대응 프레임'	지역기업들, 근원적, 역전할머니맥주, 채비, 지역경제계, 대구광역시, 컬처웍스, 임시방편, 극우세력, 미온적
헤럴드경제(0) '국제적 경제 프레임'	경기둔화, 삼성바이오로직스, 폐원, 점막, 전인미답, 골프계, 영업활동, 국가적, 국제질서, 방역모범국
KBS(2)	유럽축구연맹, 조직위, 단백질, 피

‘국제적 프레임’	서지, 중앙아시아, 광어, 악전고투, 센터발, 점거, 개도국
MBC(2) ‘국제적 프레임’	대륙, 연구진, 스웨덴, 절정, 반려동물, 주요국, 나이롱, 월북, 경계심, 충격
OBS(1) ‘사회적 희망 프레임’	안정, 동참, 규제, 의료, 제시, 발언, 요구, 위축, 건강, 업계
SBS(1) ‘국제적 위기 프레임’	비상사태, 이탈리아, 혈장, 주석, 플로리다주, 모더나, 워싱턴포스트, 외국, 부활절, 보도
YTN(0) ‘경제적 위기 프레임’	노쇼, 우라늄, 내수시장, 업무환경, 진폭, 모범국, 전전긍긍, 황폐화, 분기점, 실직

우선, Cluster0 의 언론사들을 살펴보면, 경향신문은 ‘안정화’, ’확산방지’, ’악조건’과 같은 단어를 통해 악조건에서도 안정화를 찾는 희망적인 프레임을 갖고, 국민일보는 ‘무역업계’, ‘여행수요’와 같은 경제적인 면과 ‘방도’, ‘복원력’과 같은 긍정적 희망을 바라보는 프레임을 갖는다. 세계일보는 ‘한타바이러스’ 발생에 대한 언급과 ‘비상경영체제’, ‘국가기관’, ‘대재난’과 같은 단어로 국가적 위기 프레임을 갖는다. 조선일보는 ‘경기침체’, ‘중소상공인들’과 같은 단어를 보면 경제적 위기 프레임을 갖고, 중앙일보는 ‘세계시민’, ‘일본사회’, ‘비상등’과 같이 국제적 위기를 바라본다. 한국일보는 ‘지역경기’, ‘중기중앙회’, ‘임대인들’로 중립적인 경제적 프레임을 갖고, 매일경제는 ‘큐코노미’라는 격리경제라는 신조어를 언급하며 ‘진퇴양난’, ‘실적악화’로 경제적 위기 프레임을 갖는다. 서울경제는 ‘수출업계’, ‘면세점들’을 보아 경제프레임과 ‘극복방안’, ‘대응전략’과 같은 단어를 통해 긍정적 대응 프레임이 보인다. 아주경제는 ‘금리정책’, ‘추가대책’, ‘민생안정’을 보아 경제적 대응 프레임을 갖으며, 한국경제도 ‘지역기업들’, ‘지역경제계’와 ‘채비’, ‘임시방편’을 통해 경제적 대응 프레임을 갖는다. 헤럴드경제는 ‘경기둔화’, ‘국제질서’, ‘방역모범국’을 통해 국제적 경제 프레임을 나타내며, YTN 은 ‘내수시장’, ‘황폐화’, ‘실직’과 같은 단어를 통해 경제적 위기 프레임을 갖는다.

다음으로 Cluster1 의 언론사들 중 OBS 는 ‘안정’, ‘동참’, ‘규제’로 사회적 희망 프레임을 갖고, SBS 는 ‘비상사태’, ‘이탈리아’, ‘플로리다주’, ‘워싱턴포스트’와 같이 국제적 위기를 바라본다. 같은 Cluster 에 속하지만 데이터 포인트 간의 거리가 있는 만큼 서로 다른 프레임을 갖는다. 마지막으로, Cluster2 의 언론사들 중 동아일보는 ‘평온’, ‘대처법’, ‘문화생활’과 같은 단어를 보아 사회적 대응 프레임을 갖고, 문화일보는 ‘소비심리’, ‘산업계’, ‘경제전문가’로 중립적인 경제적 프레임이라 판단된다. 한겨레는 ‘위기’, ‘방역체계’, ‘참상’을 통해 사회적 위기 프레임을 갖고, 파이낸셜 뉴스는 ‘경제질서’, ‘부정’, ‘내수침체’로 경제적 위기 프레임을 갖는다. KBS 는 ‘유럽축구연맹’, ‘중앙아시아’, ‘개도국’과 같은 단어를 보면 외국에 대한 국제적 프레임을 가지며, MBC 도 마찬가지로 ‘대륙’, ‘스웨덴’, ‘월북’을 통해 국제적 프레임으로 판단된다. 이와 같이 Cluster 별 언론사 프레임을 분석해 보았는데, 동일한 Cluster 내에서도 데이터 포인트 사이의 밀접한 정도가 다르기 때문에 구분되는 프레임을 가진다.

5. 결론 및 향후 연구계획

본 논문에서는 언론사 프레임 분석을 위해 2020 년 2 월 1 일부터 7 개월간 코로나 19 를 언급한 기사 데이터를 각 언론사별로 구분하여 수집한다. 기사들에 대해 단어 임베딩 모델인 Word2Vec 을 적용하여 벡터기반의 단어표현을 생성하고 코로나 19 와 유사한 벡터인 100 개의 단어들을 내림차순으로 추출한다. 또한, 언론사별 중복된 단어행렬을 구하여 주성분 분석으로 차원 축소한 후 K-Means Clustering 을 수행한다. 최종적으로 군집별로 모인 언론사들의 분포를 살펴보고, 각 언론사마다 중복되지 않는 유일한 단어들의 사회적, 경제적, 국가적, 국제적인 희망, 위기, 대응과 같은 특성을 찾아 프레임 분석을 한다. 이 연구를 통하여 각 언론사들의 군집 분포를 확인하고 프레임들을 파악하는 계기가 되었지만, 유일 단어들을 보고 판단하는 과정이 다소 주관적이라는 것이 한계점이다. 향후에는 각 단어들을 분야별로 분류하여 라벨링하고 언론사 프레임을 좀더 구체적으로 분석하고자 한다.

감사의 글

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2019R1A2C2006010)

참고문헌

[1] 정의철. "에이즈 뉴스 프레이밍 분석: 비판적 헬스 저널리즘 관점을 중심으로." 한국언론학보 52.4 (2008): 223-249.

[2] 김수정; 조은희. 생명과학에 대한 한국과 미국의 뉴스 프레임 비교연구. 한국언론학보, 2005, 49.6: 109-139.

[3] Linstrom, Margaret, and Willemien Marais. "Qualitative news frame analysis: A methodology." *Communitas* 17 (2012): 21-38.

[4] WANG, Rui; LIU, Wei; MCDONALD, Chris. Using word embeddings to enhance keyword identification for scientific publications. In: *Australasian Database Conference*. Springer, Cham, 2015. p. 257-268.

[5] Kuang, Sicong, and Brian D. Davison. "Learning word embeddings with chi-square weights for healthcare tweet classification." Applied Sciences 7.8 (2017): 846.

[6] Mikolov, T., Chen, K., Corrado, G., & Dean, J." Efficient Estimation of Word Representations in Vector Space *arXiv preprint arXiv: 1301.3781*, 2013.

[7] Wold, Svante, Kim Esbensen, and Paul Geladi. "Principal component analysis." *Chemometrics and intelligent laboratory systems* 2.1-3 (1987): 37-52.