트윗 키워드 네트워크를 이용한 구제역의 감성분석

채희찬*, 이종욱**, 최윤아**, 박대희**, 정용화**
*고려대학교 컴퓨터정보학과
**고려대학교 컴퓨터융합소프트웨어학과
e-mail:chay219@korea.ac.kr

Sentiment Analysis of Foot-and-mouth Disease using Tweet Keyword Network

Heechan Chae*, Jonguk Lee**, Yoona Choi**, Daihee Park**, Yongwha Chung**
*Dept of Computer and Information Science, Korea University
**Dept of Computer and Convergence Software, Korea University

요 약

구제역으로 인하여 국내 축산업계 및 관련 산업분야는 매년 막대한 피해를 입고 있다. 구제역과 관련한 다양한 학술적 연구들이 현재 진행되고는 있으나, 구제역의 발병에 따른 사회적 파급효과에 관한 공학적 분석 연구는 매우 제한적이다. 본 연구에서는 구제역에 관한 일반 시민들의 감성적 반응을 텍스트 마이닝 방법론을 사용하여 분석하는 체계적인 방법론을 제안한다. 제안하는 시스템은 먼저, 트위터에 게시된 트윗 중 구제역과 관련된 데이터를 수집한 후, 감성사전을 기반으로 극성탐지 과정을 거친다. 둘째, 토픽 모델링의 대표적인 기법 중 하나인 LDA를 활용하여 트윗으로 부터 키워드들을 추출하고, 추출된 키워드들로부터 극성별 동시출현 키워드 네트워크를 구성한다. 셋째, 키워드 네트워크을 통해 각 구간별 구제역의 사회적 파급효과를 분석한다. 사례 분석으로써, 2010년 7월부터 2011년 12월까지 국내에서 발생한 구제역에 관한 일반 시민들의 감성적 변화를 분석하였다.

1. 서론

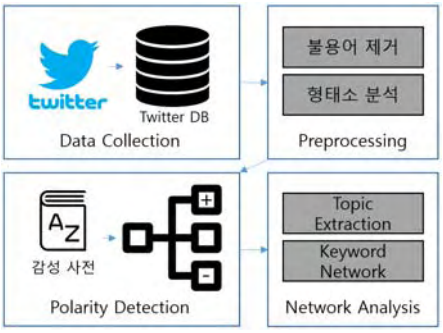
매년 발병하는 구제역은 축산업계뿐만 아니라 일반 소비자들 및 사회 전반에 큰 피해를 야기한다. 특히, 지난 2010년 발생한 구제역은 전국적으로 확산되어 총 피해액이 3조원이 넘는 막대한 피해가 발생하였다[1].

최근 정부의 주도하에 구제역을 비롯한 가축 질병에 관한 공공 데이터의 수집과 동시에 가축전염병의 확산방지를 위한 학술적 연구들이 활발하게 진행 중이다. 예를 들어, 가축질병 공공데이터를 이용하여 OLAP(On-Line Analytical Processing)분석을 수행한 연구[2], 네트워크 분석을 통해 조류독감의 전파를 예측하는 연구[3], 그리고 온라인 뉴스 데이터를 활용한 네트워크 분석[4] 등의 연구가 보고되었다. 그러나 이러한 공공 데이터와 같은 정형적 데이터나 온라인 뉴스 데이터만으로는 구제역으로 인한 피해 농가 및 시민들이 느끼는 진솔한 감성적 문제를 파악하기에는 한계가 있다. 반면, SNS(Social Network Service)와 같은 온라인 매체는 정형화된 공공 데이터베이스에서 다루지 않는 다양한 사회적 이슈들이 신속하게 전파될 뿐만 아니라 해당 문제에 관한 일반 시민들의 진솔한 감성 정보를 포함하고 있다[5].

본 논문에서는 대표적 SNS인 트위터(Twitter)를 대상으로 토픽 모델링 및 네트워크 분석을 활용하여 구제역으로 인한 일반 시민들의 개인적 감성을 살펴보고자 한다. 먼저, 트위터에 게시된 트윗(Tweet)중에서 구제역과 관련된 데이터를 수집하고, 전처리 과정과 극성(긍정 또는 부정)탐지 과정을 거친다. 이후, 토픽 모델링의 대표적인 기법 중 하나인 LDA(Latent Dirichlet Allocation)를 활용하여 키워드들을 추출하고, 동시출현 키워드 네트워크를 구성한다. 마지막으로 구성된 네트워크를 통해 극성별, 구제역의 위험구간별 사회적 파급효과를 분석한다. 사례 분석으로써, 2010년 7월부터 2011년 12월까지 국내에서 발생한 구제역으로 인한 일반 시민들의 감성적 변화를 분석한다.

2. 키워드 네트워크를 이용한 구제역의 감성분석

본 논문에서 제안하는 키워드 네트워크 기반의 감성 분석 시스템은 크게 데이터 수집 모듈, 전처리 모듈, 극성탐지 모듈, 네트워크 분석 모듈로 구성되며, 시스템 구조는 그림 1과 같다.



(그림 1) 키워드 네트워크 기반의 구제역 감성분석 시스템

2.1 데이터 수집 모듈

데이터 수집 모듈에서는 크롤러(crawler)를 이용하여 트위터에서 '구제역'이 포함된 트윗만을 수집한다. 데이터 수집 시, 시간 변화에 따른 분석 등 다양한 분석을 위하여 트윗의 게재 시간 및 내용을 함께 수집한다.

2.2 전처리 모듈

전처리 모듈에서는 정제된 결과 도출을 위해 트윗의 불용어 제거, 형태소 분석을 수행한다. 불용어 제거 과정에서는 선택된 트윗에 포함된 불용어와 @트윗, #태그, URL, 광고 등을 제거한다. 또한, 트윗을 분석 가능한 단위로 만들기 위하여 형태소 분석을 수행한다.

2.3 극성탐지 모듈

해당 사건에 대한 긍정 또는 부정적인 의견 여부를 판단하기 위하여, 극성탐지 모듈에서는 감성사전을 이용하여 해당 트윗의 극성(긍정, 부정, 중립)을 판단한다. 감성사전이란 긍정적이거나 부정적인 감정을 나타내는 단어들을 모아놓은 사전이다. 극성 판단은 트윗이 포함하고 있는 긍정 단어 수, 부정 단어 수에 의해 결정되며, 트윗이 포함하고 있는 단어들을 감성사전과 비교한 후 극성 탐지 수식에 대입하여 결과 값이 음수이면 부정, 양수이면 긍정, 0이면 중립으로 판단한다. 극성탐지 수식은 식 1과 같다[6]. 본 연구에서 사용한 감성사전은 'KOSAC'에서 제공하는 사전을 기반으로 구제역과 관련된 명확한 긍·부정 표현 단어들을 추가하여 사용하였다.

$$polarity = \frac{p-n}{p+n},\ p = \text{긍정}, n = \text{부정 단어 수} \qquad \text{(식 1)}$$

2.4 네트워크 분석 모듈

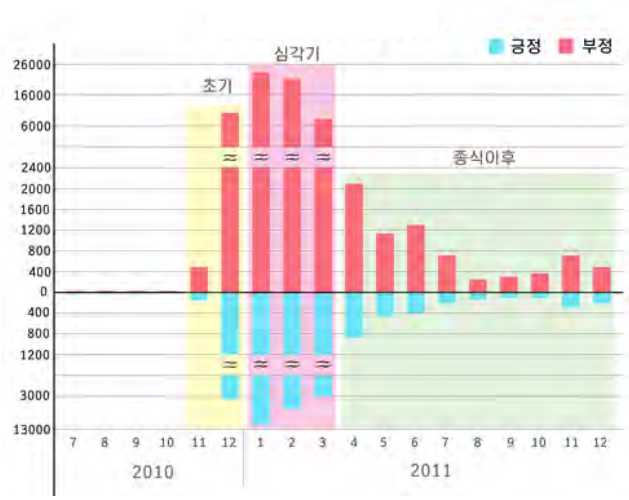
네트워크 분석 모듈에서는 극성별 키워드들의 관계를 통한 파급효과 분석을 위해 먼저, 극성탐지 모듈에서 결정된 '긍정', '부정'에 해당하는 트윗을 대상으로 각각 토픽 모델링 방법 중 하나인 LDA를 사용하여 키워드들을 추출한다. 다음으로 추출된 키워드를 활용하여 동시출현 키워드 네트워크를 구성하고, 구제역으로 인한 시기별, 극성별 파급효과들을 분석한다. 동시출현 키워드 네트워크는 키워드가 해당 트윗에 출현한 횟수를 나타내는 문서-키워드 행렬(document-keyword matrix)을 키워드간의 인접 행렬(adjacency matrix)로 변환하고, 이를 네트워크로 구성한다. 동시출현 키워드 네트워크의 노드(node)의 크기는 키워드의 출현 빈도를 나타내고, 엣지(edge)의 두께는 키워드간의 동시출현 빈도를 의미한다. 따라서 두 키워드의 동시출현 빈도가 높을수록 두 키워드의 연관성이 높음을 의미한다.

3. 실험 및 분석 결과

3.1 실험 설계

본 실험에서는 트위터에 '구제역' 키워드를 포함하는 게시글을 수집하였으며, 수집 기간은 국내에서 구제역으로 가장 큰 피해가 발생한 시기의 전·후 기간인 2010년 7월부터 2011년 12월까지로 설정하였다. 형태소 분석은 트위터에서 제공하는 한국어 형태소 분석기를 사용하였고, 불용어 제거 및 토픽 모델링은 통계프로그램 R의 KoNLP 패키지와 topicmodels 패키지를 사용하였다. 또한, 네트워크 시각화 및 분석 패키지인 igraph 패키지를 사용하여 '구제역' 키워드를 중심으로 성형(star) 구조의 동시출현 키워드 네트워크를 구성하였다. 그림 2와 같이 정부의 구제역 위기경보단계를 기준으로 구제역 발생 시기를 세 구간('발생 초기', '심각기', '종식 이후')으로 구분하고, 각 구간에서의 긍·부정에 해당하는 파급효과를 키워드 네트워크를 통해 분석하도록 설계하였다.

트윗 데이터 수집 결과, 해당 기간에 중복된 트윗들을 제거하고 약 13만 건의 트윗 데이터가 수집되었으며, 중립성을 띠는 데이터를 제외한 9만 여건의 데이터의 극성이 탐지되었다(그림 2 참조).

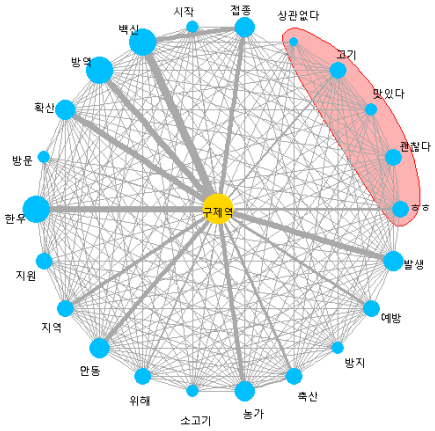


(그림 2) '구제역' 키워드를 포함하는 트윗의 극성 추이

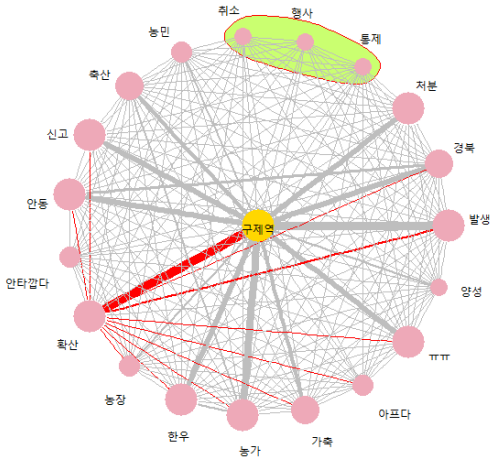
3.2 실험 결과 분석

3.2.1 구제역 발생 초기구간

구제역 발생 초기 구간의 키워드 네트워크 분석결과는 그림 3과 같다. 먼저, 그림 3 (a)에 의하면 초기 긍정 네트워크에서는 '백신', '접종', '방역' 등과 같은 구제역 발생과 확산에 관련하여 방역 및 백신 접종에 대한 긍정적인 인식들이 주를 이루는 것을 확인 할 수 있다. 또한, 고기와 관련해서 '괜찮다', '상관없다' 와 같은 키워드의 등장은 구제역 발생 초기 구간임을 감안했을 때, 사람들이 구제역에 대한 심각성을 미처 깨닫지 못한 것으로 파악된다. 반면, 초기 부정 네트워크는 그림 3 (b)와 같으며, 발생 초기 구제역의 '확산'에 대한 부정적 키워드들이 강하게 등장한다. 또한, '행사', '취소' 등과 같은 키워드들의 등장은 구제역 발생으로 갑작스럽게 행사가 취소된 것과 관련한 시민들의 실망감을 보여준다.

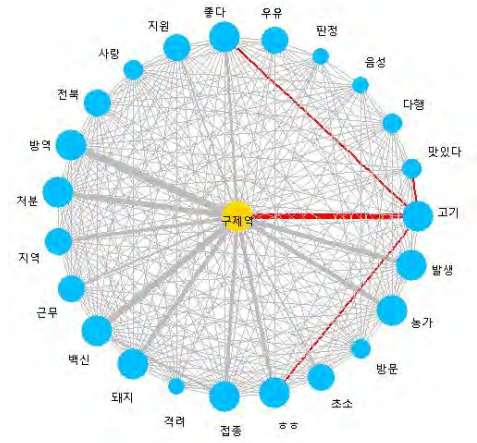


(a) 긍정 키워드 네트워크



(b) 부정 키워드 네트워크

(그림 3) 구제역 발생 초기의 키워드 네트워크



(a) 긍정 키워드 네트워크

(b) 부정 키워드 네트워크

(그림 4) 구제역 발생 심각기의 키워드 네트워크

3.2.2 구제역 심각기 구간

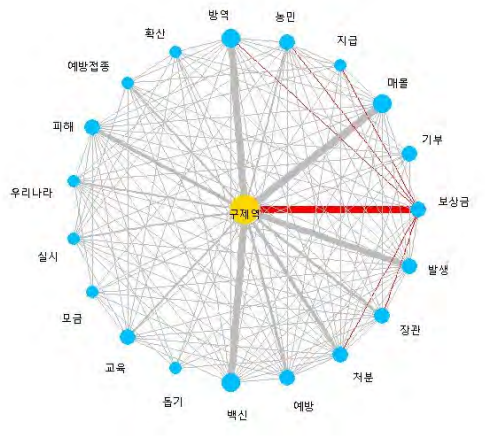
구제역 발생 심각기 구간의 키워드 네트워크 분석결과는 그림 4와 같다. 먼저, 심각기 긍정 네트워크는 그림 4 (a)와 같으며, 심각기에서는 초기와 마찬가지로 '백신', '접종', '방역', '지원' 등과 같은 방역 및 확산 방지에 관한 키워드들이 주로 나타난다. 한편, 고기에 관한 키워드는 '맛있다'와 같이 고기의 맛에 대한 긍정적인 반응은 초기와 비교해 비슷하지만, '괜찮다', '상관없다'와 같은 키워드들은 사라지고 부정 네트워크(그림 4 (b))에서 '가격', '비싸다'와 같은 고기의 가격에 대한 부정적 키워드들의 등장했음을 확인할 수 있다. 심각기 부정 네트워크에서는 '구제역'과 '정부'라는 키워드가 아주 강하게 연결되어 있고, '정부'와 연결되어 있는 '대응', '대처', '실패', '문제'라는 키워드들로부터, 구제역이 심각해짐에 따라 시민들이 정부의 구제역에 관한 대응 및 대처에 강한 감성적 불만과 불신을 느끼고 있음을 유추할 수 있다.

3.2.3 구제역 종식 이후 구간

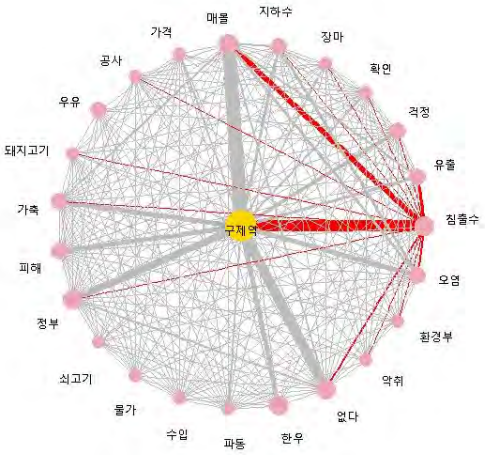
구제역 종식 이후 구간의 키워드 네트워크 분석결과는 그림 5와 같다. 먼저, 종식 이후 긍정 네트워크는 그림 5 (a)와 같으며, '보상금'이라는 키워드가 강하게 나타나고 있음을 확인할 수 있다. 구제역 이후 정부에서 피해 농민들에 대한 보상금 지급 정책과 관련하여 긍정적으로 생각하는 것을 '보상금', '농민', '지급', '장관'과 같은 키워드가 함께 등장하는 것을 통해 알 수 있다. 반면, 종식 이후 부정 네트워크는 그림 5 (b)와 같다. '침출수', '매몰'과 같은 키워드들이 '구제역'과 강한 연결을 보이고 있다. 또한 '장마', '유출', '오염' 등의 키워드가 함께 등장하는 것으로 보아, 장마철 가축의 매몰로 인한 오염으로 발생하는 침출수의 피해 문제가 구제역 종식 이후 사람들에게 가장 큰 우려의 대상이 되는 것을 확인할 수 있다.

4. 결론

본 논문에서는 다양한 사회적 이슈들에 즉각적으로 반응하는 대표적 SNS인 트위터를 대상으로 텍스트 마이닝 방법론을 활용하여 구제역이 사람들에게 실질적으로 어떤 감정적 영향을 미치는 지를 분석하였다. 구제역 발생 시기를 '발생 초기', '심각기', '종식 이후'로 구분하고, 각 구간에서의 극성에 따른 파급효과들을 동시출현 키워드 네트워크를 사용하여 분석하였다. 분석결과를 요약하면, 전 구



(a) 긍정 키워드 네트워크



(b) 부정 키워드 네트워크

(그림 5) 구제역 종식 이후의 키워드 네트워크

간에 걸쳐 구제역 예방 및 대처에 대한 키워드들이 많이 등장하면서 구제역의 피해로부터 벗어나고자 하는 사람들의 긍정적인 의견을 엿볼 수 있었다. 또한, 구제역 확산 및 정부 대응에 대한 강한 불만과 가축 매몰로 인한 침출수의 문제가 사람들에게 가장 큰 문제이자 부정적 측면으로 다가옴을 확인했다.

본 연구는 구제역의 발생 초기부터 심각기를 거치면서 종식 이후까지 일반 시민들이 느끼는 진솔한 감성적 변화를 트윗을 통하여 공학적으로 추적하는 시도라고 할 수 있다. 추후 보다 다양한 건설적 차원의 후속 연구들이 기대된다.

5. 감사의 글

본 연구는 2015년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2015R1D1A3A01018731).

참고문헌

[1] 박선일, 배선학, “구제역의 시,공간 군집 분석”, 한국지역지리학회지, Vol. 18, No. 4, pp. 464-472, 2012.

[2] 경민주, 염재홍, “Open Source SOLAP기반의 가축전염병 예찰 및 방역 의사결정 지원시스템 구현”, 한국측량학회지, Vol. 30, No. 3, pp. 287-294, 2012.

[3] Hyungjin Lee, Kyo Suh, Namsu Jung, Inbok Lee, Ilhwan Seo, Ounkyug Moon, and Jeongjae Lee, “Prediction of the Spread of Highly Pathogenic Avian Influenza Using a Multifactor Network: Part 2-Comprehensive Network Analysis with Direct/Indirect Infection Route,” Biosystems Engineering, Vol. 118, pp. 115-127, 2014.

[4] 노병준, 서정순, 이종욱, 박대희, 정용화, “온라인 뉴스를 활용한 키워드 네트워크 기반의 구제역 파급효과 분석”, 한국정보기술학회논문지, Vol. 14, No. 9, pp. 143-152, 2016.

[5] 우현지, 김영훈, “토픽 모델링을 이용한 트위터 데이터의 공간 분포 패턴 분석”, Vol. 23, No. 2, pp. 376-387, 2017.

[6] Bautin Mikhail, Lohit Vijayarenu, and Steven Skiena, “International Sentiment Analysis for News and Blogs,” ICWSM, pp. 19-26, 2008.