농산물 가격 예측을 위한 뉴스 기사 감성사전 구축

Emotion dictionary construction with News articles for forecasting agricultural prices

저자 (Authors)
김미선, 트란 디엔 손, 오아란, 윤석준, 강혜정, 김경백, 김민수, 양형정
Mi-Sun Kim, Tran Dinh Son, Aran Oh, Seok-Jun Yun, Hye-Jeong Kang, Kyung-Baek Kim, Min-Su Kim, Hyung-Jeong Yang

출처 (Source)
한국정보과학회 학술발표논문집 , 2017.12, 283-285(3 pages)

발행처 (Publisher)
한국정보과학회
The Korean Institute of Information Scientists and Engineers

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07322126

APA Style
김미선, 트란 디엔 손, 오아란, 윤석준, 강혜정, 김경백, 김민수, 양형정 (2017). 농산물 가격 예측을 위한 뉴스 기사 감성사전 구축. 한국정보과학회 학술발표논문집, 283-285

농산물 가격 예측을 위한 뉴스 기사 감성사전 구축

김미선[1O], 트란 디엔 손[1], 오아란[1], 윤석준[1], 강혜정[2], 김경백[1], 김민수[3], 양형정[1**]
[1]전남대학교 전자컴퓨터공학대학원
[2]전남대학교 농업생명과학대학원
[3]전남대학교 통계학과대학원
misun_kim05@hanmail.net, trandinhson3086@gmail.com, dhdkfks9@naver.com, potoemember@gmail.com,
hjkang@chonnam.ac.kr, kyungbaekkim@gmail.com, kimms@chonnam.ac.kr, hjyang@jnu.ac.kr

Emotion dictionary construction with News articles for forecasting agricultural prices

Mi-Sun Kim[1O], Tran Dinh Son[1], Aran Oh[1], Seok-Jun Yun[1], Hye-Jeong Kang[2], Kyung-Baek Kim[1],
Min-Su Kim[3], Hyung-Jeong Yang[1*]
[1]School of Electronics and Computer Engineering, Chonnam National University
[2]School of Agricultural Life Science, Chonnam National University
[3]School of Statistics, Chonnam National University

요 약

본 논문에서는 농산물 가격 예측에 사용하기 위해 온라인 뉴스를 이용한 감성 사전을 구축하는 방법을 제안한다. 이를 위해 2015년과 2016년에 걸쳐 양파에 관련된 기사를 수집했으며, 수집한 비정형 텍스트데이터에서 문장 단위로 기사를 분할한다. 분할된 텍스트 중 분석내용과 연관 없거나, 가격 등락에 상관없이 많이 언급된 단어들을 불용어로 처리했다. 형태소 분석을 진행한 후 키워드를 추출하여 제안한 모형의 성능을 로지스틱 회귀방식과 인공신경망을 이용해 실험했다. 그 결과 로지스틱 회귀분석을 이용했을 때보다 인공신경망의 일종인 합성곱신경망(Convolutional Neural Networks, CNN)을 이용했을 때 약 10% 이상 정확도가 향상됐다.

1. 서 론

현재 농산물 가격 예측 연구로는 채소류, 과일류를 중심으로 수급전망과 시계열 분석을 통한 가격 예측[4]을 하거나 비정형 농업 기상자료를 이용한 가격 예측 연구들이 진행되었다.[1] 그러나 실제 양파 가격은 기상 변화에 따른 생산량의 변화가 심하고 유사 작목 간 대체 관계가 존재하기 때문에 예측이 어려운 실정이다.[1] 또한 양파의 가격은 수확 후 유통과정에서 결정된다. 따라서 위와 같은 연구는 현실적인 농산물 가격 예측에 적절하지 않다.

언론 보도와 소비자의 행동에 관한 연구에 따르면 기업에 관련된 부정적인 언론 기사는 어떤 미디어보다 영향이 크고 가격에 영향을 미친다. [2][3] 뉴스에 대한 오피니언 마이닝 연구로는 주가의 움직임을 예측하기 위한 많은 연구들이 진행되었다. [9][15] 이와 같은 많은 연구들은 미리 긍정 부정의 극성을 규정해 놓은 범용 감성사전을 이용하거나, 각 연구의 주제에 맞는 어휘사전을 따로 만들어 진행하였다. 그러나 이 분야의 어휘사전을 농업 관련 분석 연구에 적용시키는 것에는 한계가 있다.

본 연구에서는 온라인 기사에 나타난 어휘들의 극성을 양파가격의 등락에 미치는 영향에만 근거하여 판별하고, 이를 기반으로 양파가격 등락 관점에서의 관련 어휘들의 감성사전을 구축하고자 한다. 사람의 직관적인 판단이 아니라 양파의 가격 등락여부를 기준으로 분석과정을 제시함으로써, 유사 분야에서 주제에 맞는 감성사전을 만드는데 도움을 줄 수 있다.

2. 본 론

2.1. 예측 모형 설계

본 연구에서 제안하는 농산물 가격예측 모형은 그림1에 나타나 있다. 그림 1은 [15]에서 제안한 시스템을 양파 가격 예측 주제에 맞게 개선한 것으로, 농작물에 특화된 감성사전을 별도로 구축하여 사용한다는 점에서 가장 큰 차별성을 갖는다.

그림1의 전체 모형의 각 단계를 설명하면 다음과 같다. 첫 단계는 양파 관련 뉴스를 수집한다. 두 번째 단계에서는 뉴스의 텍스트를 형태소 분석기 모듈에 의해 분해하여 감성사전을 구성한다. 세 번째 단계에서는 분해된 어휘에 대해 감성사전을 참조하여 긍

"본 연구는 미래창조과학부 및 정보통신기술진흥센터의 대학ICT연구센터육성 지원사업의 연구결과로 수행되었음" (ITP-2017-2016-0-00314)
"본 성과물은(논문, 산업재산권, 품종보호권 등)은 농촌진흥청 연구사업(세부과제번호: PJ011823022017)의 지원에 의해 이루어진 것임"
"이 논문은 2017년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017R1A2B4011409)"
교신저자 : 양형정 e-mail : hjyang@jnu.ac.kr

정 부정의 극성을 부여한다. 완성된 감성사전을 이용하여 기계학습이론 중 합성곱신경망(convolutional Neural Networks, CNN)을 이용하여 신문기사를 이용한 양파 가격 등락 변화 예측 모형을 검증한다.

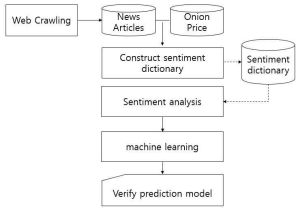

그림 1. 제안된 시스템 모형

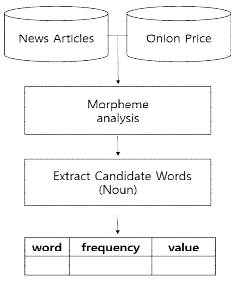

그림 2. 감성사전 구축 과정

2.2. 감성사전 구축

감성사전 구축 과정은 다음과 같다[그림2]. [15]의 알고리즘을 도메인 특성에 맞추어 개선하여 농업 뉴스에 대한 분석을 진행한다. 가격이 상승한 것을 긍정이라고 보며, 본 연구에서는 상승이라고 표현한다. 빈도수(freq)는 해당 단어가 나온 기사의 수를 합산하여 계산하고, 상승 지수(pos)는 해당 단어가 들어간 기사가 월별 양파 가격(onion month price, OMP)이 상승한 달에 속한 경우의 수를 합산하여 계산한다. 빈도수와 상승 값을 수식으로 표현하면 수식 (1)과 같다.

$$word(i,j) = \begin{cases} 1(\text{기사}j\text{에 단어}i\text{가 존재} \\ 0(\text{그 외의 경우}) \end{cases}$$

$$freq(i) = \sum_{j=1}^{n} word(i,j)$$

$$OMP(j) = \begin{cases} 1(\text{기사}j\text{가 상승달에 속할 경우}) \\ 0(\text{그 외의 경우}) \end{cases}$$

$$pos(i) = \sum_{j=1}^{n} \{word(i,j) \times OMP(j)\} \quad (1)$$

마지막으로 추출한 어휘들의 상승 지수를 계산하여 감성사전을 완성한다. 상승 지수는 상승 값을 빈도수로 나누어 나타내며, 식으로 표현하면 수식 (2)와 같다.

$$P(i) = \frac{\sum_{j=1}^{n} \{word(i,j) \times OMP(j)\}}{\sum_{j=1}^{n} word(i,j)} \quad (2)$$

2.3. CNN 기반 기사 문장 분류 모델

[17]을 참고하여 CNN을 텍스트 처리에 적용하였다. CNN은 문장의 지역 정보를 보존함으로써 단어의 등장순서를 학습에 반영하는 아키텍처이다. 본 연구에서는 신문기사와 위에서 만들어진 감성사전을 이용하여 각 기사가 상승인지 하락인지 CNN을 이용해서 분류한다.

그림 3은 CNN 기반으로 한 모델을 나타낸다. n 개의 단어로 이뤄진 기사를 각 단어별로 k 차원의 행벡터로 임베딩한다. 아키텍처의 입력 값은 파라미터를 설정할 때 단어벡터 값도 초기 값을 랜덤으로 주고 학습 과정에서 업데이트를 하는 방식을 사용했다. 만들어진 입력 값은 문장의 길이만큼 해당하는 개수만큼 나열되고, 그 단어벡터들은 (문장길이×사용자가 지정한 단어벡터 차원 수)로 lookup 테이블을 만들게 된다. CNN은 배치(batch) 단위로 레이어(layer)에 들어가기 때문에 배치를 반영한 입력 값의 차원수를 설정해야 한다.

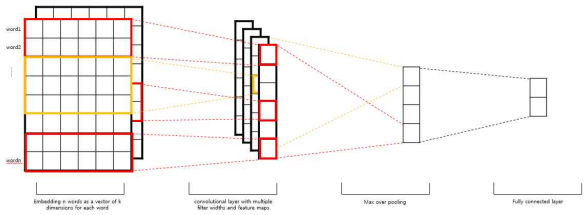

그림 3. CNN 기반 기사 문장 분류 모델

3. 실험내용 및 결과

3.1 데이터수집

본 연구에서는 2015년 1월 1일부터 2016년 12월 31일까지 작성된 양파와 관련된 기사를 Python을 이용해 크롤링하여 농업신문 630건, 일반 언론 12,132건의 기사를 수집하여 실험에 활용하였다.

또한 제시하는 방안을 검증하기 위해 훈련용으로 70%, 검증용으로 30%로 데이터 집합을 랜덤으로 나누어 사용하였다.

양파의 가격은 감성사전을 만드는데 긍정과 부정의 기준이 된다. 농산물유통정보(KAMIS)에서 해당 년도의 양파 가격을 수집하고 소비자물가지수 지표를 통해 가격을 조정해서 사용한다. 실제 양파 20kg당 가격을 해당 년 월의 소비자물가 지수로 나눠서 조정한 금액을 기준으로 사용한다. 상승 달과 하락 달을 구분하기 위해 각 년도의 조정 평균가를 구하고 그보다 높으면 상승 달, 낮으면 하락 달로 구분한다. 표1는 연도별로 조정된 평균가를 기준으로 상승 달과 하락 달을 구분한 것이다.

표1. 연도별 상승 월과 하락 월

year	상승 달	하락 달
2015	7,8,9,10,11,12	1,2,3,4,5,6
2016	1,2,3,9,10,11	4,5,6,7,8,12

3.2 실험결과

최종적으로 일별 뉴스기사의 빈도수와 상승지수, 소비자물가를 적용한 양파 가격 항목을 추가해 양파 가격 예측 모형을 위한 데이터로 사용한다. 본 연구에서는 인공신경망의 일종인 CNN을 이용하였고, 로지스틱 회귀분석과 비교하였다.

본 실험에서는 총 65,563개의 단어를 이용하였다. 예측모형간의 성능을 비교하기에는 데이터의 양이 충분하지 않다고 판단하여 교차검증을 하였다.[16] 분류 기법은 인공신경망의 일종인 CNN을 이용했다. 성능측정은 5중 교차 검증(5-fold cross-validation) 방식을 사용하였으며 결과는 표2과 같다.

CNN은 batch(배치)단위로 데이터가 들어가는데 배치를 반영한

입력 값의 파라미터는 표2과 같다. embedding_size는 단어 벡터의 차원 수로써 필터의 너비를 의미한다. 채널 수는 1로 고정하였다. batch_size는 배치의 크기를, epochs_size는 훈련할 때 각 자료가 몇 번이나 학습에 사용되었는지 결정한다.

표2. 예측 모형 성능 평가 결과

logistic regression	Convolutional Neural Network(CNN)			
result	result	embedding_ size	batch_ size	epochs_ size
	0.78	128	64	10
	0.81	**256**	**32**	**10**
0.63	0.68	256	32	100
	0.59	256	64	10
	0.52	512	128	200

로지스틱 회귀분석을 이용했을 때는 63%의 정확도를 보였고, 인공신경망의 일종인 CNN을 이용했을 때 파라미터를 [embedding_size=256, batch_size=32, epochs_size =10] 으로 설정했을 때 80% 내외의 정확도를 보였다. 이 결과는 Kim(2013)과 Chun,(2013)이 주가예측을 위해 뉴스 콘텐츠와 오피니언마이닝을 이용한 것[10][11]보다 향상된 것을 확인할 수 있다.

4. 결 론

본 논문에서는 농산물 양파의 가격의 등락을 예측하기 위해 뉴스 기사와 소비자물가를 적용한 양파 가격을 바탕으로 생성된 감성사전을 이용하여 새로운 기사데이터가 들어왔을 때 가격 등락에 어떤 영향을 미치는지를 예측하는 모델을 제시한다.

본 연구를 통해 농산물 가격 예측을 하는 것에 기상데이터가 쓰이는 것 이외에 비정형데이터로 언론기사를 사용할 수 있는 한 가지 방안을 제시하였다. 향후 연구에서는 뉴스 기사와 예측 정확도와의 상관관계에 대한 연구를 진행하여 예측 모델의 한계점을 보환할 수 있는 대책을 강구할 필요가 있을 것이다.

참 고 문 헌

[1] Nam, Kuk-Hyun, and Young-Chan Choe. "A Study on Onion Wholesale Price Forecasting Model." *Journal of Agricultural Extension & Community Development* 22.4, pp. 423-434, 2015.

[2] Dean, Dwane Hal. "Consumer reaction to negative publicity: Effects of corporate reputation, response, and responsibility for a crisis event." *The Journal of Business Communication,* 1973.

[3] Fiske, Susan T. "Attention and weight in person perception: The impact of negative and extreme behavior." *Journal of personality and Social Psychology* No. 38-6, pp. 889, 1980.

[4] 장수희, et al. "비정형 농업기상자료를 활용한 배추 도매가격 예측 모형 연구." *한국데이터정보과학회지* No. 28-3, pp. 617-624, 2017.

[5] 이철성, et al. "한글 마이크로블로그 텍스트의 감정 분류 및 분석." *정보과학회논문지: 데이타베이스* No. 40-3, pp. 159-167, 2013.

[6] 김동영, 박제원, and 최재현. "SNS 와 뉴스기사의 감성분석과 기계학습을 이용한 주가예측 모형 비교 연구." *한국 IT 서비스학회지* No. 13 pp. 211-233, 2014.

[7] Shim, Kwang-Seob, and Jae-Hyung Yang. "High speed Korean morphological analysis based on adjacency condition check." *Journal of KIISE: Software and Applications* No. 31-1, pp. 89-99. 2004.

[8] 김민정, and 김철주. "텍스트마이닝을 활용한 숭례문 관련 기사의 트렌드 분석." *한국콘텐츠학회논문지* No. 17-3, pp. 474-485, 2017.

[9] 정지선, 김동성, and 김종우. "온라인상의 뉴스 감성분석을 활용한 개별 주가 예측에 관한 연구." *한국지능정보시스템학회 학술대회 논문집, pp.* 45-58, 2015.

[10] Kim, Kwang-Yong, Gyeong-Rak Lee, and Seong-Weon Lee. "A Comparative Analysis of Artificial Intelligence System and Ohlson model for IPO firm's Stock Price Evaluation." *Journal of Digital Convergence* No. 11.5, pp. 145-158, 2013.

[11] 천세원, 김유신, and 정승렬. "뉴스 콘텐츠의 오피니언 마이닝을 통한 매체별 주가상승 예측정확도 비교 연구." *한국지능정보시스템학회 학술대회논문집,* pp. 133-137, 2013.

[12] 함효민, et al. "로지스틱 판별 분석을 통한 한국어 구어체 문장 감정 분류 정확도 분석." *한국정보과학회 학술발표논문집*, pp. 1870-1872, 2017.

[13] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics*, 2002.

[14] Dave, Kushal, Steve Lawrence, and David M. Pennock. "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews." *Proceedings of the 12th international conference on World Wide Web. ACM*, 2003.

[15] 유은지, et al. "주가지수 상승 예측을 위한 주제지향 감성사전 구축 방안." *한국지능정보시스템학회 학술대회논문집*, pp. 42-49, 2012.

[16] Frank, Eibe, et al. "Weka-a machine learning workbench for data mining." *Data mining and knowledge discovery handbook. Springer US*, pp. 1269-1277, 2009.

[17] Kim, Yoon. "Convolutional neural networks for sentence classification." *arXiv preprint arXiv* , No. 1408.5882, 2014.

[18] 이상훈, et al. "텍스트 마이닝을 활용한 영화흥행 예측 연구." *한국데이터정보과학회지*, No. 26.6, pp. 1259-1269, 2015.

[19] 윤홍준, 김한준, and 장재영. "오피니언 마이닝 기술을 이용한 효율적 상품평 검색 기법." *정보과학회논문지: 컴퓨팅의 실제 및 레터*