Journal of the Korean Data & Information Science Society
2015, **26**(5), 1167–1173

http://dx.doi.org/10.7465/jkdi.2015.26.5.1167
한국데이터정보과학회지

텍스트마이닝을 이용한 사회 이슈 찬반 분류에 관한 연구[†]

강선아[1] · 김유신[2] · 최상현[3]

[123]충북대학교 경영정보학과, BK21+BSO사업팀

접수 2015년 8월 5일, 수정 2015년 9월 21일, 게재확정 2015년 9월 22일

요 약

정보통신기술의 발전은 SNS, 블로그, 게시판 등 자신의 생각이나 의견을 표출할 수 있는 장소의 다양성을 제공하였고 이는 빅데이터 성장을 가능케 하였다. 특히 매순간마다 엄청난 수의 사용자가 이용가능하고 다양한 이슈에 대한 의견을 작성할 수 있는 SNS의 특징으로 인해 많은 사람들이 트위터 등에 사회적 이슈에 대한 자신의 의견을 드러낸다. 따라서 본 연구에서는 트위터에서 작성되는 사회 이슈에 대한 의견을 수집하여 사회이슈를 주제로 하는 감성사전을 구축하고 구축된 감성사전을 통해 감성 분석을 실시하고자 한다. 사용된 데이터는 '비키니', '나꼼수'를 포함하는 트윗 글이다. 사회이슈에 특화된 주제지향 감성사전을 구축하고 구축된 감성사전을 통해 긍부정 의견을 분석한 결과 Precision은 61%로 나타났으며 F1-score는 74%의 성능을 보여주었다. 본 연구는 정치적 색을 띄고 있는 특정 사회 이슈에 대한 트윗 작성자의 의견이 긍정인지 부정인지 자동으로 분류할 수 있도록 하는 사전 구축의 하나의 기준을 제시할 것이라 기대한다.

주요용어: 감성 분석, 감성 사전, 사회이슈, 오피니언 마이닝, 텍스트마이닝.

1. 머리말

정보통신기술의 발전은 SNS, 블로그, 게시판 등 자신의 생각이나 의견을 표출할 수 있는 장소의 다양성을 제공하였다. 이는 빅데이터 성장을 가능케 하였고, 관심을 고조시켰다 (Choi 등, 2013). 특히 소셜 네트워크 서비스 (socail network service; SNS)는 매순간마다 엄청난 수의 사용자가 이용 가능하여 제품, 서비스 등에 대한 리뷰를 남길 수 있어 긍정/부정에 대한 오피니언의 변화가 끊임없이 발생한다 (Park 등, 2011). 즉 개인의 의견을 빠르게 파악하여 목적에 따라서 분석하고 이를 활용할 수 있는 오피니언 마이닝에 대한 연구가 매우 중요해지고 있다 (Kim과 Cho, 2013).

다양한 소셜 네트워크 서비스 중에 트위터는 개인의 사적인 내용을 포함하여 정치, 경제, 사회, 문화, 스포츠 등 다양한 분야에 대한 개인의 의견과 가치관을 표출할 수 있는 장소이다. 따라서 사고, 사건등 사회적 이슈가 발생할 경우 트위터를 통해 상황을 중계하기도 하고 트윗 작성자의 관심사, 가치관을 드러내기도 한다 (Hur와 Choi 2012). 이러한 트위터의 특성으로 많은 사람들이 다양한 주제에 대하여 자신의 의견을 작성하여 여론을 형성 하는 등 여론의 장으로서 자리매김하고 있다 (Shamma 등, 2009). 최근에는 트위터의 트윗 글을 분석하여 선거의 결과를 예측하는 연구들이 국내외로 진행되어 왔다. 이

[†] 본 연구는 미래창조과학부 및 정보통신산업진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음 (NIPA-2014-H0301-14-1022).
[1] (362-763) 충북 청주시 서원구 충대로 1, 충북대학교 경영정보학과, 석사과정.
[2] (362-763) 충북 청주시 서원구 충대로 1, 충북대학교 경영정보학과, BK 21+ 연구교수.
[3] 교신저자: (362-763) 충북 청주시 서원구 충대로 1, 충북대학교 경영정보학과, 교수.
E-mail: chois@chungbuk.ac.kr

는 정치적 특색을 지닌 트윗 작성자들이 정치적 상황에 영향을 미치고 있음을 보여준다. 이처럼 소셜미디어는 영향력 있는 몇몇 작성자에 의해 여론이 형성되므로 정치적 사건이나 사회이슈에 대한 여론을 조성하는 데 중요한 역할을 한다 (Bae 등, 2013).

본 논문에서는 정치적 색을 띄고 있는 특정 사회 이슈에 대한 트윗 작성자의 의견이 긍정인지 부정인지 자동으로 분류할 수 있는 사전을 구축하고자 한다. 논문에서 다룬 사회이슈는 비키니 시위 사건이다. 이 시위는 딴지일보에서 제작한 팟캐스트인 '나는 꼼수다'에서 주로 논의가 되었던 허위사실 유포 혐의로 구속 수감된 정봉주 전의원의 석방을 요구하는 시위이다. 연구는 다음과 같이 시행된다. 먼저 데이터를 수집한 후, 수집된 데이터로부터 자연어 처리 프로세싱을 통해 긍부정 어휘를 추출하여 감성 사전을 구축한다. 또한 구축된 감성 사전을 토대로 트윗 의견의 긍/부정을 분류하여 사전의 유효성을 검증한다. 따라서 본 연구 내용을 활용하면 사회 이슈에 대한 대중의 의견이 긍정인지 부정인지 자동으로 분류가 될 것이라고 기대한다.

논문은 다음과 같이 구성된다. 제 2절에서는 관련연구를 기술하고 제 3절에서는 사회 이슈에 대한 감성 사전 구축을 위한 일련의 순서를 소개와 더불어 분석 데이터에 대한 내용을 소개한다. 그리고 제 4절에서는 감성 사전 구축에 대한 연구 결과에 대해 보여주며 마지막 제 5절에서 본 연구의 시사점 및 한계, 향후 연구과제로 결론을 맺는다.

2. 관련연구

오피니언 마이닝은 텍스트 데이터에서 긍정, 부정의 의견을 판단하고 활용하는 목적으로 사용된다 (Jang 등, 2015). 오피니언 마이닝은 정치, 문화, 경영 등 다양한 분야에서 활용되므로 분야마다 긍정과 부정을 나타내는 감성 어휘가 상이할 수 있다. 그러므로 도메인 지향적 어휘를 추출하는 것이 중요하다.

Owsley 등 (2006)은 글의 감성 분류를 위해 주제에 특화된 사전을 사용하는 것의 중요성을 설명하였다.

주제에 특화된 오피니언 마이닝의 중요성이 증명되면서 주식, 영화, 관광 등 다양한 분야에서 연구가 진행되고 있다. Aue와 Gamon (2005)는 서로 다른 4개의 도메인인 책, 영화, 제품, 지식에 관한 웹 설문조사 데이터를 수집하여 새로운 도메인을 위한 감성 분류를 위한 방법을 제시하였다.

Yu 등 (2013)은 주가 도메인에 특화된 주제지향 감성사전을 구축하기 위해 뉴스에 나타난 어휘들의 극성을 주가 지수 등락에 미치는 영향으로 판별하여 어휘 감성사전을 구축하여 주가 도메인에 부합한 감성사전 구축의 하나의 기준을 제시하였다.

또한 Yoon과 Kim (2011)은 네이버에서 제공한 영화 박쥐에 대한 10,000건의 영화 평점 정보를 토대로 새로운 영화평에 대한 평점 자동 예측 방식을 연구하여 영화 도메인의 단어사전을 자동으로 구축하는 방법을 제안하였다. Cho 등 (2015)은 SNS 상에서 발생하는 충북 관광 관련 글을 수집하여 이슈분석, 연관분석, 감성분석을 통해 충북 관광을 위한 홍보 전략과 관광 진흥 정책을 수립하는 방안을 제시하였다. Kang 등 (2015)은 사회네트워크 분석과 텍스트마이닝을 이용하여 공격, 패스 등의 패턴을 찾아내고 배구경기력과 관련된 키워드를 추출하여 경기력을 평가하였다.

3. 연구 방법론

본 연구는 정치적 색을 띈 사회 이슈라는 특정 도메인 안에서 긍정과 부정 어휘를 추출하여 감성 사전을 구축하고 구축된 감성 사전으로 문장의 긍/부정을 정확하게 분류하고자 한다.

연구 프레임 워크는 Figure 3.1로 설명될 수 있다. 모형의 첫번째 단계에서는 분석 대상이 되는 트윗 글을 수집한다. 이를 위해 '나꼼수', '비키니'단어를 포함하는 트윗 글을 수집하였다. 두 번째 단계에서

는 수집된 트윗 글에 대해 긍정과 부정을 태깅하고 긍정과 부정 트윗 글 각각 세종국어사전을 이용하여 형태소를 분리하였다. 세 번째 단계에서는 감성사전을 구축하기 위해 형태소가 분리된 긍정과 부정 어휘의 빈도를 도출하여 어휘의 긍부정을 판단하였다. 구축된 감성 사전의 정확도를 파악하기 위해 긍정 트윗 글에서 무작위로 1000개를 부정 트윗 글에서 무작위로 1,000개를 뽑아 감성 사전으로 문장의 긍부정 여부를 판별하여 정확도와 F1-score로 사전의 예측 정확성을 증명하였다.

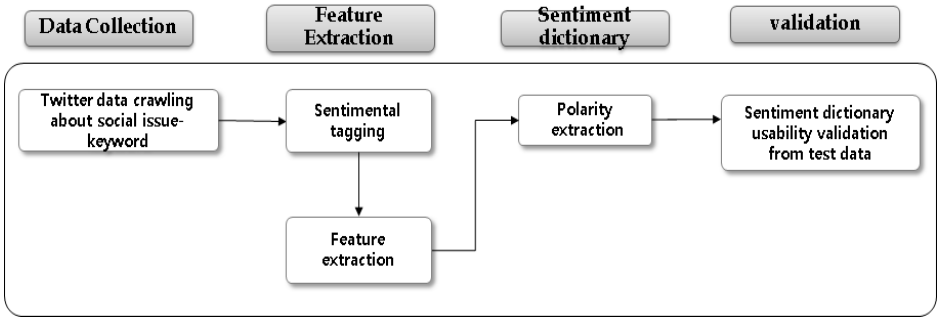

Figure 3.1 Framework of sentiment dictionary construction

4. 감성사전을 이용한 긍부정 판별

4.1. 실험 데이터

본 연구에서는 '나꼼수', '비키니'를 포함하는 트윗 글을 수집하였다. 비키니 시위 사건은 제 1절에서 언급하였듯이 허위사실 유포 혐의로 구속 수감된 정봉주 전 의원 석방을 요구하는 시위이다. 시위에 대한 트윗 글은 최대 140자까지 작성될 수 있으며 당시 논란이 되고 있던 비키니 시위는 의사 표현을 위한 수단이라 주장하며 이를 옹호하는 글에는 긍정, 해당 시위를 비판하는 글에는 부정을 태깅하였다.

데이터 수집 기간은 논쟁이 진행 중이던 2012년 1월 26일부터 2012년 2월 7일까지 13일이며 총 수집된 데이터는 4,363건이다. 각 트윗 글에 긍정과 부정, 중립을 태깅한 후 중립인 글과 중복으로 수집된 글을 제거하여 2,797건으로 축소시켰다. 2,797개의 트윗 글 중에 1,225개의 트윗은 부정, 1,572개의 트윗은 긍정으로 비율은 각각 44%, 56%이다.

4.2. 감성 어휘 추출

의견의 긍,부정을 판별하기 위해서 감성 어휘는 매우 중요한 역할을 하며 감성 사전에 감성 어휘가 어떻게 구성되는가에 따라 결과가 좌지우지 된다.

본 연구에서 감성 어휘를 결정하기 위해 긍정과 부정이 태깅된 글을 분리하여 따로 형태소 분석을 진행하였다. 형태소 분석 후 어휘의 빈도를 계산하여 최다 빈도 1,500개의 어휘를 추출하였다. 추출 후 다음의 규칙으로 어휘를 정제하였다.

1. 어휘 출현 빈도가 2이하인 어휘 제거
2. 작성자 아이디 등 의미 파악이 어려운 영어, 한자 제거
3. 접속사, ㅋ, ㅠ 단어 제거
4. 동의어 취합
5. 긍정어휘와 부정어휘에서 어휘가 중복될 경우 더 높은 빈도의 감성으로 판단

위의 규칙을 적용하여 긍정어휘 923개, 부정어휘 689개로 추출되었다.

Table 4.1 Positive/negative term

Positive Term	Negative Term
apology	criticism
demand	unpleasant
expression	attack
courage	explanation
appreciation	falsehood
state	rage
supporter	bad
defender	doubt
pleasure	inconvenience
relax	circulation

Table 4.1은 긍정과 부정 어휘를 비교한 표로서 특정 사건에 적용된 어휘이며 '사과', '용기', '통쾌'와 같이 해당 시위를 옹호하는 표현들이 긍정 어휘로 출현하였고, 해당 시위를 비난하는 '비판', '불쾌', '거짓' 등이 부정 어휘에 출현하였다.

4.3. 감성 사전을 이용하여 문장의 긍부정 판별 및 평가

본 절에서는 4.2절에서 설명된 방법으로 구축된 감성사전을 이용하여 문장의 긍/부정을 판별하고자 한다. 먼저 긍정과 부정 트윗 글 각각 1,000개씩 무작위로 추출하여 세종 국어사전으로 형태소를 분리하였다. 2,000개의 트윗 어휘와 감성사전의 어휘를 비교하여 같은 어휘가 존재하면 긍정은 +1을 부정을 -1의 값을 부여하였다.

다음 단계에서는 식(4.1)과 같이 긍정 용어의 출현된 수와 부정 용어의 출현된 수의 합을 구하여 만약 긍정 용어의 출현된 수의 합이 부정 용어의 출현된 합보다 크다면 해당 트윗 글을 긍정으로 판별하였고, 긍정 용어의 출현된 수의 합보다 부정 용어의 출현된 합이 크거나 0이면 부정으로 판별하였다.

$$\text{IF Sum(appr\_Pos\_term)} > \text{Sum(appr\_Neg\_term) THEN Word(i)senti} = 1; \quad (4.1)$$
$$\text{ELSE Sum(appr\_Pos\_term)} \leq \text{(appr\_Neg\_term) THEN Word(i)senti} = 0; \text{ END IF;}$$

"나꼼수 비키니 발언에 여성삼국에서 성명서를 준비하는 모양인데 개인적으론 나꼼수팀의 대응은 현명하지 못하다고 본다. 현재와 같은 태도와 상황이 계속된다면 동력..."라는 트윗 글을 식(4.1)의 방식으로 예시를 들어보면 Table 4.2와 같다.

먼저 세종국어사전에 의해 분류된 형태소 17개 중 감성사전에 존재하는 12개의 어휘를 분석 대상으로 놓고 12개의 어휘가 긍정 사전에 존재하는가 부정 사전에 존재하는가를 파악한다. 긍정 사전에 존재하는 어휘이면 +1을 부정 사전에 존재하는 어휘이면 +1을 부여한다. 부여된 값의 결과를 계산하여 값이 0을 기준으로 크면 긍정으로 작거나 같으면 부정으로 판단한다. Table 4.2의 경우 각 어휘에 +1과 -1를 부여한 결과이며 총 합이 -6이므로 부정이라 분류하게 된다.

Table 4.2 Example of positive/negative analysis

morpheme analysis	comment	woman	statement	preparation	shape	maneuver	wisdom
dictionary	negative	negative	positive	negative	negative	negative	negative
	-1	-1	+1	-1	-1	-1	-1

morpheme analysis	present	attitude	situation	continuously	mobilize	result
dictionary	negative	negative	negative	positive	positive	negative
	-1	-1	-1	+1	+1	-6

분류 성능을 평가하기 위해 precision과 F1- score를 이용하기로 하였다.

5. 연구결과

본 절에서는 3절과 4절에서 제시한 연구 방법을 토대로 연구 결과를 도출하여 작성하였다. '나꼼수', '비키니' 키워드인 트윗 글에 대한 긍/부정 의견을 판별하기 위해 먼저 긍정어휘 1,037개와 부정어휘 905개로 사회 이슈에 대한 긍/부정 사전을 구축하였다. 구축된 감성 사전으로 긍/부정을 판별한 결과 Figure 5.1과 같이 Precision은 57% 로 나타났으며 F1-score는 67%의 성능을 보여주었다.

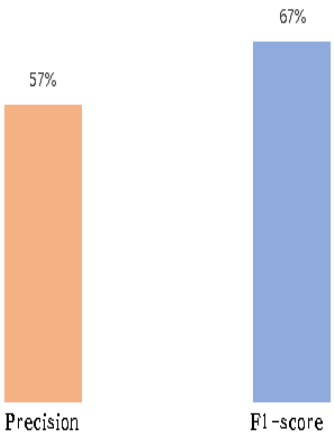

Figure 5.1 Precision and F1-score

사회 이슈 감성 사전의 성능이 높지 않은 이유는 2,797건의 트윗 글로 사전을 구축하여 사전의 수가 적었기 때문이다. 또한 해당 사건의 트윗 의견이 '나는 꼼수다'라는 팟 캐스트를 비판하거나 '비키니 사건'을 비판하는 글이 대부분이어서 긍정 사전과 부정 사전의 어휘가 명확하게 구분이 되지 않았다.

6. 결론

본 논문에서는 사회 이슈에 대한 트위터를 수집하여 감성 사전을 구축한 후 감성 분석을 수행하여 도메인에 특화된 감성 사전 구축을 목표로 하였다. 분석 수행과정은 수집 데이터에 대해 자연어 처리를 하고, 감성 사전을 구축하여 1037개의 긍정 어휘와 905개의 부정 어휘를 도출하였다. 구축된 감성사전으로 2000개의 트윗 글을 샘플링하여 감성 분석을 실시한 결과 61%의 정확도와 74%의 F1-score에 성능을 보였다.

본 연구의 한계점은 사회 이슈를 '비키니 사건'으로 한정지어 3,000건의 트윗 글로 감성 사전을 구축하여 감성분석의 예측 정확도가 높지 않다는 점이다. 따라서 향후 연구에서는 예측 정확도를 높이기 위해 사회 이슈의 사건을 늘리고 감성 사전을 확장하여 사회 이슈 도메인 감성사전을 보완하여야 한다.

References

Kim, Y. D. and Cho, K. H. (2013). Big data and statistics. *Journal of the Korean Data & Information Science Society*, **24**, 959-974.

Park, K. M., Park, H. G., Kim, H. G. and Ko, H. D. (2011). The opinion mining study in SNS. *Communications of the Korean Institute of Information Scientists and Engineer*, **29**, 54-60.

Hur, S. H. and Choi K. S. (2012). A study on characteristics and types of tweet in twitter. *Hanminjok Emunhak*, **61**, 455-494.

Bae, J. H., Son, J. E. and Song, M. (2013). Analysis of twitter for 2012 south korea presidential election by text mining techniques. *Journal of Intelligent and Information System*, **19**, 141-156.

Jang, G. E., Park, S. H. and Kim, W. J. (2015). Automatic construction of a negative/positive corpus and emotional classification using the internet emotional sign. *Journal of KIISE*, **42**, 512-521.

Yu, E. J., Kim Y. S., Kim, N. G. and Jeong, S. R. (2013). Predicting the direction of the stock index by using a domain-specific sentiment dictionary. *Journal of Intelligent and Information System*, **19**, 95-110.

Cho, W. S., Cho, A., Kwon, K. E. and Yoo, K. H. (2015). Implementation of smart chungbuk tourism based on SNS data analysis. *Journal of the Korean Data & Information Science Society* , **26**, 409-418.

Kang, B. U., Huh, M. K. and Choi, S. B. (2015). Performance analysis of volleyball games using the social network and text mining techniques. *Journal of the Korean Data & Information Science Society*, **26**, 619-630.

Yoon, D. M. and Kim, K. J. (2011). *Prediction of rating score from short comments on movies using word-rating correlation analysis*, The HCI Society of Korea, Korea.

Aue, A. and Gamon, M. (2005). Customizing sentiment classifiers to new domains: A case study. In *Proceedings of Recent Advances in Natural Language Processing*, Borovets, Bulgaria.

Owsley, S., Sood, S., and Hammond, K. J. (2006). Domain Specific Affective Classification of Documents. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, USA.

Shamma, D. A., Kennedy, L. and Churchill, E. F. (2009). Tweet the debates: Understanding community annotation of uncollected source. *Advancing Computing as a Science & Profession(ACM)*, USA.

Choi, H. S., Park, H. W. and Park, C. Y. (2013). Support vector machines for big data analysis. *Journal of the Korean Data & Information Science Society*, **24**, 989-998.

Journal of the Korean Data & Information Science Society
2015, **26**(5), 1167–1173

http://dx.doi.org/10.7465/jkdi.2015.26.5.1167
한국데이터정보과학회지

Study on the social issue sentiment classification using text mining[†]

Sun-A Kang[1] · Yoo Sin Kim[2] · Sang Hyun Choi[3]

[123]Department of MIS, Chungbuk National University

Received 5 August 2015, revised 21 September 2015, accepted 22 September 2015

Abstract

The development of information and communication technology like SNS, blogs, and bulletin boards, was provided a variety of places where you can express your thoughts and comments and allowing Big Data to grow, many people reveal the opinion of the social issues in SNS such as Twitter. In this study, we would like to pre-built sentimental dictionary about social issues and conduct a sentimental analysis with structured dictionary, to gather opinions on social issues that are created on twitter. The data that I used is "bikini", "nakkomsu" including tweet. As the result of analysis, precision is 61% and F1- score is 74%. This study expect to suggest the standard of dictionary construction allowing you to classify positive/negative opinion on specific social issues.

Keywords: Opinion mining, sentimental analysis, sentimental dictionary, social issue, text mining.

[†] This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2014-H0301-14-1022) supervised by the NIPA (National IT Industry Promotion Agency).

[1] Master student, Department of MIS, Chungbuk National University, Cheongju 362-763, Korea.
[2] Research professor, Department of MIS, Chungbuk National University, Cheongju 362-763, Korea.
[3] Corresponding author: Professor, Department of MIS, Chungbuk National University, Cheongju 362-763, Korea. E-mail: choi@chungbuk.ac.kr