트위터에서의 토픽별 감정 패턴 분석

이재환, 김남윤, 임혜원, 김형주

서울대학교 컴퓨터공학부

jlee@idb.snu.ac.kr, nykim@idb.snu.ac.kr, hwlim@idb.snu.ac.kr, hjk@snu.ac.kr

Sentiment Pattern Analysis of Topics on Twitter

Jaehwan Lee, Namyoon Kim, Hyewon Lim, Hyoung-Joo Kim

Dept. of Computer Science and Engineering, Seoul National University

요 약

소셜 네트워킹 서비스는 빠른 정보의 확산과 더불어 사용자들 정보에 대한 주관성이 뚜렷하다는 특징 때문에 여러 토픽에 대한 다양한 의견을 들을 수 있는 유용한 도구로 평가 받고 있다. 특히 소셜 네트워킹 서비스를 기반으로 한 감정 분석은 특정 콘텐츠나 뉴스에 대한 사람들의 전반적인 시각을 알아보는데 유용하다. 본 논문에서는 트위터 데이터를 기반으로 주목받는 토픽에 대한 감정 분석을 수행하였다. 이를 통해 특정 토픽에 대해서 유사한 감정 패턴이 나타나며, 트윗 발생량이 아닌 감정의 변화로 이슈를 탐지할 수 있음을 발견하였다.

1. 서론

트위터는 사용자들이 관계를 맺고 관심사를 공유하는 대표적인 소셜 네트워킹 서비스이다. 사용자들이 발행하는 트윗의 양은 하루에 5억 건에 달할 정도이며 정치나 경제와 같이 굵직한 이슈부터 개인적인 잡담에 이르기까지 다양한 주제를 포함하고 있다.

사용자들은 현재 이슈가 되는 토픽들을 트위터에 공유하고 이에 대한 의견들을 주고 받는다. 이는 트윗 수의 폭발적인 증가를 불러오는데, 일례로 마이클 잭슨의 사망 소식이 전해지자 “Michael Jackson”을 포함한 트윗의 수가 22.61%나 증가했다[1]. 이처럼 트위터에서 많은 사용자들의 주목을 받는 주제를 ‘트렌드’라고 부르는데, 트렌드에 관한 트윗의 85% 이상이 헤드라인 뉴스와 일치한다[2]. 이는 온라인과 오프라인의 이슈가 트위터에 빠르게 반영되며, 트위터를 분석함으로써 실세계의 이슈를 탐지할 수 있음을 나타낸다. 또한 소셜 서비스에서 특정 이슈나 콘텐츠에 대해 사용자들이 사용하는 언어에서 나타난 감정들이 실생활에서 느끼고 있는 감정과 유사하기 때문에[3] 트위터는 여러 이슈들에 대한 사람들의 다양한 생각과 감정들을 얻을 수 있는 유용한 도구가 된다[4]-[6].

트위터에서의 감정 분석은 설문조사 등의 번거로운 과정 없이도 여러 주제에 대해 사람들의 전반적인 시각을 빠르게 알아볼 수 있게 한다. 또한 트윗 수가 급격히 증가하지는 않지만 다른 트윗들에 비해 상대적으로 감정의 증가 폭이 큰 경우와 같이, 단순히 시간 당 트윗 발생 빈도를 측정하는 것만으로는 잡아낼 수 없는 이벤트들을 탐지함으로써 새로운 트렌드의 추출이 가능해진다. 이러한 감정분석을 통해 언어지는 결과는 사용자들의 감정에 따른 추천이나 상품에 대한 마케팅 방향 수립 등에 활용할 수 있다. [7]에서는 감정 분석을 통해 카메라 브랜드에 대한 호감도를 예측했으며 75%~95%의 높은 정확도를 보였다.

긍정, 부정, 중립의 수준에서 분석했던 기존 연구[4], [6], [7]와 달리, 본 연구에서는 트위터 데이터를 대상으로, 감정 분석의 기준이 되는 NRC Emotion Lexicon (Emolex)[8]를 이용하여 주목 받는 토픽들에 대한 더 높은 수준의 감정 분석을 하였다. 이를 위해 두 가지 가설을 세웠고, 실험을 통해 특정 토픽에 대한 사람들의 감정은 시간의 흐름과 관계없이 대체로 일정함과, 감정 변화로 이벤트를 탐지할 수 있음을 발견하였다. 본 연구에서는 실험 대상을 인물에 한정하였으나 실험 방법은 토픽의 종류에 구애되지 않으므로 다른 분야의 토픽에도 적용할 수 있다.

2. NRC Emotion Lexicon

NRC Emotion Lexicon (EmoLex)는 이슈의 감정 분석에 사용되는 감정 단어 사전으로 14,200개의 단어로 구성되어 있다. Plutchik이 제시한 인간의 여덟 가지 기본 감정[9]인 기쁨(joy), 슬픔(sad), 화(anger), 두려움(fear), 신뢰(trust), 메스꺼움(disgust), 기대(anticipation), 놀람(surprise)에 긍정과 부정을 포함한 10가지 감정을 기반으로 하여, 각 단어가 어떤 감정을 내포하고 있는지 10차원 이진 벡터로 나타낸다. 본 연구에서는 0.92버전을 사용하였다.

3. 데이터

본 연구에서 사용한 데이터는 제 1회 빅데이터 분석 경진대회[1]에서 제공한 트윗 데이터로, 2013년 3월 한 달간 한국인이 발행한 트윗이다. 데이터의 크기는 총 30GB로 JSON 형태로 제공되었으며, 1,930,258명의 사용자와 145,588,301개의 트윗을 포함한다. 이

[1] http://contest.kbig.or.kr

데이터에 Jackson fast JSON Parser [2] 를 이용하여 트윗들을 날짜 별로 나누고, 한국어 이외의 언어로 작성된 트윗은 제외하였으며, 토픽 추출 및 감정 분석에 필요한 텍스트 부분 이외의 해시태그, 특수문자, URL 등을 제거하였다.

4. 감정 분석

4.1 전처리

각 트윗은 형태소 분석기를 이용하여 단어들의 집합으로 변환하였다. 영어의 경우에는 TweetMotif [3] 와 같이 트위터에 특화된 다양한 형태소 분석 도구들이 존재하지만, 이는 한국어를 다루는 데는 적합하지 않기 때문에 우리는 루씬(Lucene) [4] 에서 제공하는 WhitespaceAnalyzer를 사용하였다. 루씬의 또 다른 형태소 분석기인 CJKAnalyzer와 StandardAnalyzer의 경우에는 한국어 데이터에 대해 좋지 않은 성능을 보였으며, 특히 CJKAnalyzer의 경우에는 인터넷 용어에 대한 인덱싱이 제대로 되지 않았다. 형태소 분석을 마친 트윗들은 날짜 별로 나누어 각각 하나의 문서로 취급하였으며, 이를 이용해 문헌 빈도(Document Frequency)를 계산하였다.

4.2 감정 분석 방법

감정 분석을 위한 기준인 EmoLex가 영어 단어만을 제공하기 때문에 한국어로 된 트윗을 번역하는 과정이 필요하다. 트윗의 경우 맞춤법에 맞지 않는 경우가 많으므로 번역에서 오류가 발생하기 쉽다. 번역의 정확성을 높이기 위해 네이버에서 제공하는 맞춤법 검사기를 이용하여 트윗을 정제한 후 구글 번역기를 사용해 영어로 변환하였다.

변환된 트윗은 Java로 작성된 감정 분석기에 넣어 8가지 감정 및 긍정, 부정의 10차원으로 구성된 감정 벡터에 대응시켰다. 번역하는 과정에서 'not'이나 'never'와 같은 부정어가 들어간 경우, 긍정/부정 벡터 값을 반대로 준 벡터와 그렇지 않은 벡터의 차이가 유의미하지 않아 부정어 처리는 수행하지 않았다.

5. 실험

감정 분석에 사용될 트윗이 많을 수록 높은 정확도를 기대할 수 있다. 따라서 높은 DF값을 가진 단어를 기준으로 삼고 해당 단어에 대하여 트위터 사용자들이 갖고 있는 감정 분석을 실시하였다. 4.1절의 전처리 과정을 거쳐 높은 DF 값을 나타낸 7개 토픽(김연아, 박근혜, 박태환, 손연재, 안철수, 인피니트, 틴탑)을 선택하고, 두 가지 가설을 세워 감정 분석을 실시하였다.

가설 1 특정 토픽에 대해 사람들이 나타내는 감정은 대체로 일정할 것이다.

가설 2 감정 변화가 나타나는 순간에는 어떤 사건이 발생했을 것이다.

하나의 트윗이 대응되는 감정 벡터가 각 차원에서 0이 아닌 값을 가질 확률은 매우 낮다. 벡터의 값이 모두 작은 경우 비교하기 어려우므로 먼저 각 토픽 단어가 포함된 트윗을 날짜 별로 분류하고, 날짜별 트윗의 감정 벡터를 합하였다. 30일 간의 트윗에서 토픽당 30개씩 총 210개의 감정 벡터를 얻었으며, 같은 토픽에 해당하는 감정 벡터들의 코사인 유사도를 구한 후 7개 토픽에 대하여 평균 유사도를 계산하였더니 0.980959(최솟값 0.9584)로 나타났다.

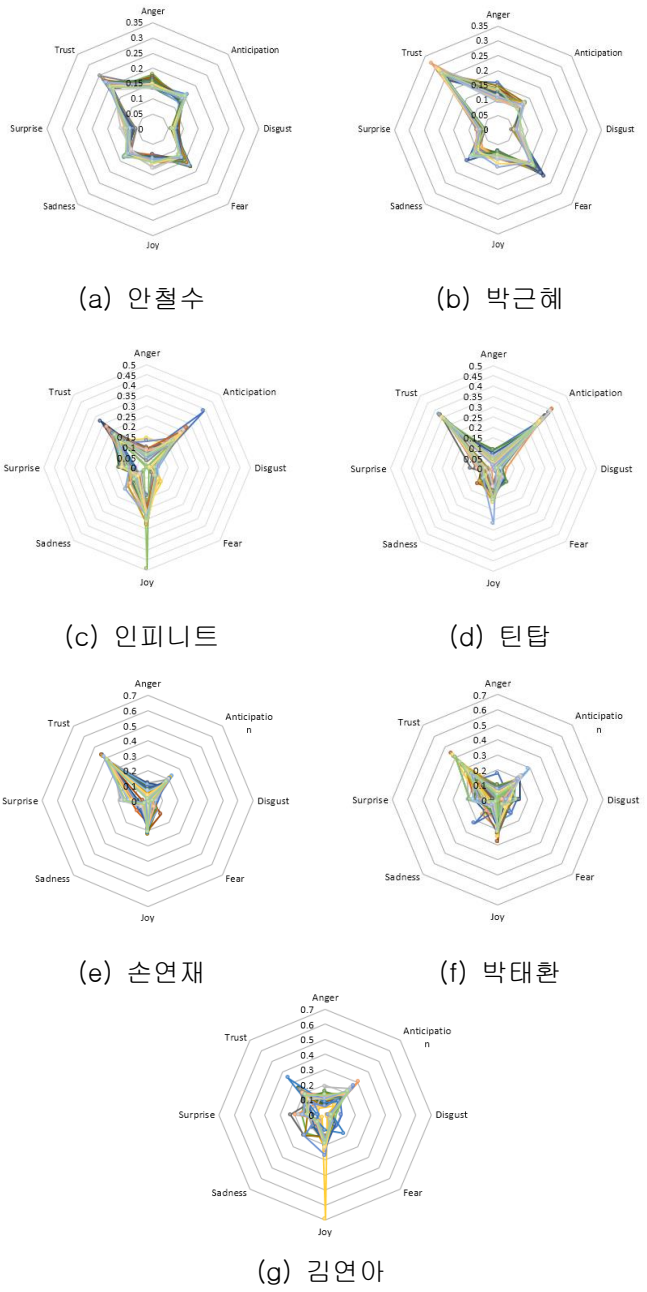
감정 분석의 결과는 그림 1과 같다. 그림 1은 각 토픽에 대한 한 달간의 감정분포를 나타낸 것으로, 특정 토픽에 대해서 사람들은 시간의 흐름과 관계없이 전반적으로 유사한 감정 패턴을 지니고 있음을 알 수 있다. 또한 계산된 값의 차이는 존재하나 정치인, 연예인 등 같은 군집에 속하는 토픽에 대해서는 비슷한 감정 분포를 보였다. 정치인의 경우에는 양극에 위치한 신뢰와 두려움이 첫 번째와 두 번째로 높게 나타났다. 이는 지지 세력과 반대 세력이 분명하게 존재하기 때문이라고 추측할 수 있다. 아이돌 가수와 운동 선수의 경우에는 신뢰, 기대, 기쁨과 같은 긍정적인 감정이 다른 감정에 비해 극단적으로 높게 나타나는 것을 알 수 있다. 표 1은 군집 간의 평균 감정 분포 유사도를 나타낸다. 흥미롭게도 가수 간의 유사도보다 가수와 운동선수 간의 유사도가 더 높게 나타났는데, 이는 아이돌 가수와 같이 확고한 팬 층을 확보하고 있는 김연아의 특성 때문으로 보인다.

표 1 군집 간 평균 유사도

	정치인	운동선수	아이돌가수
정치인	0.97329	0.882648	0.860253
운동선수		0.9294	0.917114
아이돌가수			0.914159

그림 1(g)의 감정 분석 그래프에서는 다른 감정 벡터들과 유사도가 떨어지는 한 벡터를 관찰할 수 있다. 3월 16일의 트윗인데 이 날의 감정 벡터는 다른 날의 감정 벡터들과 코사인 유사도를 계산하였을 때 평균 유사도 0.6840으로 다른 날에 비해 현저히 낮게 나타났다. 3월 16일의 트윗들을 살펴본 결과 '김연아 점프 실제로 본 후기.jpg'라는 내용의 글이 트윗에 유행처럼 퍼져나갔기 때문인데, 트윗의 발행 빈도가 가장 높았던 17일(세계 피겨 선수권 대회일)에 주목하였을 때는 찾아낼 수 없었던 사건이다. 운동선수에 관련된 트윗은 국제 대회처럼 특별한

[2] http://github.com/FasterXML/jackson
[3] http://github.com/brendano/tweetmotif
[4] http://lucene.apache.org/core/



(a) 안철수 (b) 박근혜

(c) 인피니트 (d) 틴탑

(e) 손연재 (f) 박태환

(g) 김연아

그림 1 토픽별 감정 분석

사건이 있는 날을 제외하고는 발행 빈도가 매우 낮기 때문에 사람들 사이에 이슈가 되는 내용들을 트윗 발행 빈도만으로는 찾아낼 수 없는 경우가 많은데, 이러한 경우 트위터 사용자들의 감정 변화폭을 통해 탐지할 수 있음을 알 수 있다.

6. 결론

본 논문에서는 트위터 데이터를 기반으로 토픽 별 사람들의 감정 패턴 분석을 수행하였다. 기존의 연구들과 마찬가지로 사후 분석이었다는 점과 이슈 예측을 위해 무한에 가까운 토픽들에 대한 감정 패턴을 측정해 두어야 한다는 한계가 있으나, 일정 토픽에 대해서 트위터 사용자들이 나타내는 감정 패턴은 시간의 흐름에 대해 대체로 일정한 것으로 나타났다. 또한 감정 변화를 이용해 이슈의 발생을 탐지할 수 있음을 밝혔다. 이는 특정 사건의 발생이 감정 변화를 유도한다는 [4]의 연구 결과와도 일치한다.

참 고 문 헌

[1] C. Pete, “Michael Jackson Dies: Twitter Tributes Now 30% of Tweets,” 2009. [Online]. Available: http://mashable.com/2009/06/25/michael-jackson-twitter/.

[2] H. Kwak, C. Lee, H. Park, and S. Moon, “What is Twitter, a social network or a news media?,” in *Proceedings of the 19th international conference on World wide web - WWW ’10*, 2010, p. 591.

[3] C. Orellana-rodriguez, E. Diaz-aviles, and W. Nejdl, “Mining Emotions in Short Films: User Comments or Crowdsourcing?,” in *Proceedings of the 22nd international conference on World Wide Web companion*, 2013, pp. 69–70.

[4] H. Wang, D. Can, A. Kazemzadch, F. Bar, and S. Narayanan, “A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle,” in *Proceedings of the ACL 2012 System Demonstrations. Association for Computational Linguistics*, 2012, pp. 115–120.

[5] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, 2011.

[6] M. Thelwall, K. Buckley, and G. Paltoglou, “Sentiment in Twitter Events,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 62, no. 2, pp. 406–418, 2011.

[7] T. Nasukawa and J. Yi, “Sentiment analysis: Capturing favorability using natural language processing,” in *Proceedings of the 2nd Iternational Conference on Knowledge Capture*, 2003, pp. 70–77.

[8] S. M. Mohammad and P. D. Turney, “Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon,” in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 2010, pp. 26–34.

[9] R. Plutchik, “A general psychoevolutionary theory of emotion,” *Emot. Theory, Res. Exp.*, vol. 1, no. 3, pp. 3–33, 1980.