SNS 데이터를 이용한 사회 불안의 시공간 기반 시각화

김재민*, 이주홍*, 최용석†
*한양대학교 컴퓨터·소프트웨어학과
† 한양대학교 공과대학 컴퓨터공학부
e-mail:skymygo@hanyang.ac.kr

Spatio-temporal Visualization of Social Anxiety Using SNS Data

Jae-Min Kim*, Joo-Hong Lee*, Yong-Suk Choi† *
*Dept of Computer and Software, Hanyang University
† Division of Computer Science and Engineering. Hanyag University

요 약

본 논문에서는 SNS에서 수집한 데이터를 이용하여 사회 불안의 시공간 분포를 시각화 하는 기법을 소개한다. Open API인 twitter4j를 이용하여 트위터로부터 시공간 정보를 포함한 데이터를 수집한 뒤, 이 트윗의 작성자가 불안한지 아닌지 표시한 훈련 데이터를 준비한다. 이 훈련 데이터와 한글 형태소 분석기 Open API인 KOMORAN을 이용해 사전을 구축하고, 불안 분류기를 개발한다. 트위터로부터 수집한 시공간 정보를 포함한 데이터를 분류기로 분류하여, 지도에 표시해줌으로써 사회 불안을 시각화 한다. 사회 과학자들이 이를 이용하여 불안을 체계적으로 연구함으로써 불안으로부터 생기는 다양한 사회 문제들을 해결할 수 있다.

1. 서론

최근 한국 사회는 정치적 양극화, 사교육 경쟁, 청년 실업, 자살, 노후 빈곤, 저출산, 증오 범죄 등 다양한 사회적 문제 현상을 겪고 있다. 이는 갈수록 심화되는 경쟁과 반목으로 인한 것이며, 사회적 연대의 해체와 신뢰 저하가 우려되는 상황이다. 이러한 문제들은 상호 긴밀하게 연관되어 있으며 이들을 포괄하는 근본 원인으로 사람들이 느끼는 '불안'에 주목해야 한다.[1-5]

기존의 사회 과학 연구는 사회 구성원 일부에 대한 설문조사를 통해 수행되는 경우가 많다. 이러한 방법은 불안같은 감정과 정서에 대한 체계적 연구를 하기에 개인이 자신의 실제 감정을 정확히 이해할 수 없고, 분석에 시간이 필요하다는 제약이 존재한다.[6]

본 논문에서는 기존의 방법과 달리 소셜 네트워크로부터 데이터를 수집한다. 사람들은 설문조사 보다 소셜 네트워크에서 더 솔직하기 때문에 기존의 방법보다 효과적이고 신뢰성이 있는 방법으로 수집이 가능하다. 추가적으로 데이터를 실시간으로 수집하고, 분류하므로 시간적 제약도 존재하지 않는다.

본 논문에서 제안하는 불안 분석 시각화는 위에서 나열한 많은 사회 문제 해결에 도움을 줄 수 있다. 뿐만 아니라 사회 과학자들이 해당 자료를 활용하여 불안의 변화 추세와 정치, 경제, 사회, 문화, 등의 다양한 현상들과의 관련성을 탐구하고 지역별 상황과 상대적 격차에 대한 세부적 진단 및 정책 마련에 활용 가능하다.

본 논문은 2장에서 관련 연구들을 소개한다. 3장에서는 본 논문에서 사용한 알고리즘 및 사전 구축 방법에 대해서 소개한다. 4장에서는 분류기의 성능을 평가하고, 5장에서 결론 및 추후 연구에 대해서 논의한다.

2. 관련연구

2.1 베이즈 정리

베이즈 정리는 한 사건이 일어났을 때, 다른 사건이 일어날 확률을 설명하는 정리다.[7] 예를 들어, 한 사람의 나이가 30일 확률을 P(B), 한 사람이 암에 걸렸을 확률을 P(A)라고 할 때, 한 사람의 나이가 30살이면, 그 사람이

*본 연구는 산업통상자원부의 재원으로 기술혁신사업의 지원을 받아 수행한 연구 과제 (No. 10060086, 개인 서비스용 로봇을 위한 지능-지식 집약 ·개방 ·진화형 로봇지능 소프트웨어 프레임워크 기술 개발)입니다
**2017년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초 연구사업임 (No.NRF-2015R1D1A1A01060950).
† 교신저자(Corresponding Author) : 한양대학교 공과대학 컴퓨터공학부 교수 최용석(cys@hanyang.ac.kr)

암에 걸렸을 조건부 확률 P(A|B)를 알 수 있다.

베이즈 정리에서는 수식 (1)로 조건부 확률을 정의한다.

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)} \quad \cdots\text{수식 (1)}$$

2.2 나이브 베이즈 분류

나이브 베이즈 분류는 확률들 사이의 독립을 가정하는 베이즈 정리를 적용한 확률 분류이다.[8] 나이브 베이즈 분류기는 지도 학습 모델에서 효율적으로 학습되며, 분류에 필요한 파라미터를 추정하기 위한 트레이닝 데이터의 양이 적어도 되고, 간단한 디자인과 단순한 가정에도 불구하고 많은 복잡한 실제 상황에서 잘 작동한다.

2.3 라플라스 평활화

본 논문에서 사용하는 나이브 베이즈 분류의 경우 사전에 없는 단어가 등장할 경우 확률이 0이 되므로 평활화 기법을 사용해야 한다. 이를 해결하기 위해 사전에 없던 단어가 등장할 경우 해당 단어와 사전에 등록된 모든 단어에 빈도 수 1을 더하고, 계산을 진행하는 라플라스 평활화를 사용한다.

2.4 사회 불안

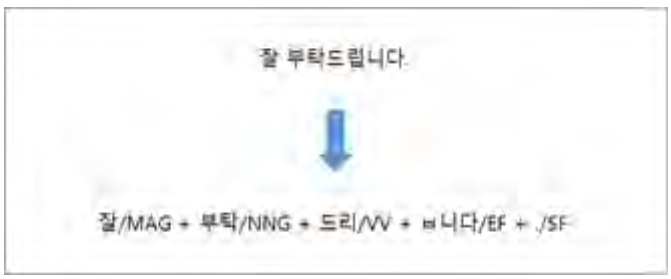
현대 사회에서 많은 사람들이 느끼는 불안은 나쁜 심리적 효과를 유발하고, 심하게는 심리 장애로도 연결된다. 양극화, 경쟁 과열, 자살 등의 사회 문제는 불안을 해소함으로써 해결할 수 있다.

3. SNS데이터의 불안 분류와 시각화

3.1 데이터 수집 및 사전 구축

Open API인 twitter4j[9]를 이용하여 트위터로부터 시공간 정보가 포함된 트윗을 수집한다. 이 과정에서 수집된 데이터의 품질을 높이기 위해 외국 관광객의 트윗으로 추정되는 한글의 비중이 낮은 트윗과 광고로 추정되는 #(해시태그)가 3개 이상인 트윗을 제거한다. 이렇게 수집된 트윗들을 불안과 비불안으로 분류한다.

분류기 개발을 위한 불안 단어 사전을 구축하기 위해 Open API인 KOMORAN[10]을 이용해 수집된 트윗들을 형태소 단위로 나눈다. 본 논문에서는 형태소의 품사 중 NNG(명사), VV(동사), VA(형용사), MM(관형사), MAG(일반 부사)만 사용한다.



(그림 1) KOMORAN을 이용한 형태소 분석의 예시

(그림 1)은 KOMORAN을 이용하여 '잘 부탁드립니다.'를 분석한 예시이다. '잘/MAG', '부탁/NNG', '드리/VV', 'ㅂ니다/EF', './SF'로 분석된다. 이 중 '부탁/NNG'와 '드리/VV'만을 사전 구축에 활용한다.

3.2 분류기 시스템 설계

위에서 구축된 사전을 이용하여 나이브 베이즈 분류기를 개발한다. 본 논문에서 사용한 분류기는 임의의 트윗이 들어왔을 때, 최대 우도 추정을 이용하여 해당 트윗이 불안일 확률과 비불안일 확률을 계산해 더 확률이 높은 쪽으로 분류한다. 사전이 <표 1>, <표 2>와 같을 때, 단어1, 단어3, 단어2로 구성된 트윗이 비불안일 확률은 수식 (2), 불안일 확률은 수식 (3)이다. 비불안일 확률이 더 높으므로, 해당 트윗은 비불안으로 분류된다.

<표 1> 비불안 사전 예시

단어/품사	횟수	분류
$단어_1/NNG$	400	비불안
$단어_2/VV$	300	비불안
$단어_3/MAG$	200	비불안
$단어_4/VA$	100	비불안
총	1000	

<표 2>불안 사전 예시

단어/품사	횟수	분류
$단어_1/NNG$	30	불안
$단어_2/VV$	30	불안
$단어_3/MAG$	10	불안
$단어_4/VA$	30	불안
총	100	

$$P(단어_1, 단어_3, 단어_2|비불안) = \frac{400}{1000} \times \frac{200}{1000} \times \frac{300}{1000} = 0.024 \quad \cdots\text{수식 (2)}$$

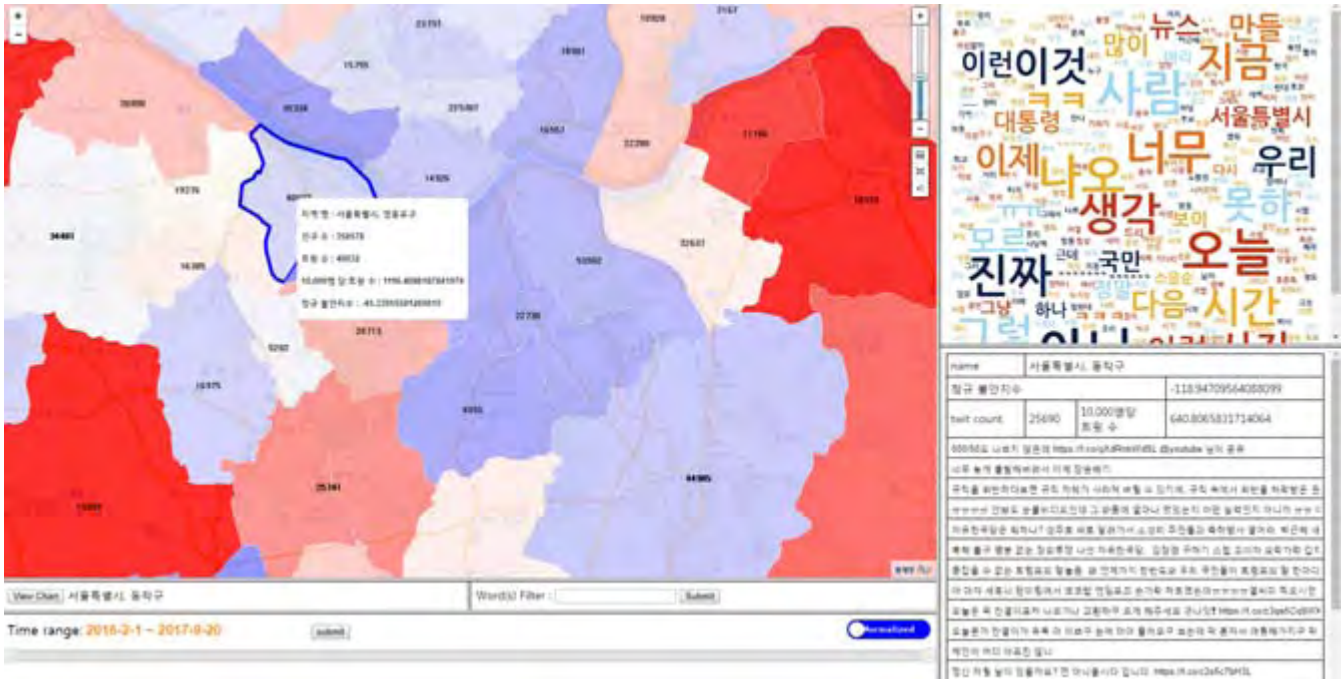
$$P(단어_1, 단어_3, 단어_2|불안) = \frac{30}{100} \times \frac{10}{100} \times \frac{30}{100} = 0.009 \quad \cdots\text{수식 (3)}$$

3.3 시각화

트위터로부터 시공간 정보를 포함한 데이터를 수집하여, 불안 분류기로 해당 트윗이 불안한지 분류하여 지도에 표시해준다. 지도는 통계청의 Open API를 사용한다. 시각화 시스템은 본 연구팀의 홈페이지에서 사용해볼 수 있다.*

지도는 축적에 따라 한국, 대도시(특별시·도), 소도시(시·군·구)로 표시된다. (그림 2)의 지도는 소도시를 표시하고

* http://166.104.140.75:62000/Emotional_Analyzation/EA/ main.html



(그림 2)2016년 2월부터 2017년 9월 까지의 불안 분포

있다. 하단의 타임 테이블을 조정하여 지도에 표시되는 데이터들의 시간을 조정할 수 있다. (그림 2)의 지도는 2106년 2월부터 2017년 9월까지의 데이터를 표시하고 있다. 타임 테이블 위에 있는 토글 버튼을 누르면 지도에 불안 지수가 표시되는 방법이 바뀐다. 불안 지수가 표시되는 방법은 두 가지로 절대 지수와 정규 지수로 표시된다. 절대 지수는 해당 도시의 트윗 1,000개 중 불안 트윗이 몇 개인지를 나타낸다. 정규 지수는 해당 도시의 트윗 중 불안 트윗의 비율이 다른 도시와 비교해 어느 정도 수준인가를 나타낸다. (그림 2)의 지도는 정규 지수로 표시되고 있다. 지도의 도시에 마우스 커서를 올리면 해당 도시의 정보가 툴팁으로 제공된다. 지역 명, 인구 수, 트윗 수, 10,000명당 트윗 수, 불안 지수가 표시된다. (그림 2)의 지도는 서울 특별시, 영등포구 위에 마우스 커서를 올려놓은 상태이다. 지도의 도시를 클릭하면, 우측에 도시의 상세 정보가 표시된다. 우측 상단에는 해당 도시의 트윗들로 생성한 워드 클라우드가 표시되고, 우측 하단에는 해당 도시의 상세 정보가 표시된다. 상세 정보에는 툴팁에서 표시되던 정보와 해당 도시의 트윗들이 출력된다. (그림 2)에서는 서울특별시, 동작구의 정보를 출력하고 있다.

4. 성능

4.1 최대 사후 확률과 최대 우도 추정

본 논문에서 트레이닝에 사용한 트윗은 75,051개이다. 이 75,051개의 트윗 중 불안 트윗은 7,366개이고, 비불안 트윗은 67,685개이다. 이는 서로의 등장 확률 차이가 매우 큰 균형적이지 못한 상황이다.

$$P(\text{불안}) = \frac{7366}{75051} \approx 0.098$$
$$P(\text{비불안}) = \frac{67685}{75051} \approx 0.902 \quad \cdots(4)$$

따라서 최대 사후 확률을 사용하면 성능 저하가 발생하는데, 이는 우리가 준비한 테스트 데이터 셋 6,822개에서도 알 수 있다. <표 3>은 테스트 데이터 셋으로 평가해본 결과이다.

<표 3> 최대 사후 확률과 최대 우도 추정에 대한 성능 비교 표

방법	불안 재현율	비불안 재현율	정확도	성능
최대 사후 확률	0.5069	0.9641	0.9371	0.458
최대 우도 추정	0.8550	0.8292	0.8317	0.590

최대 사후 확률의 경우 비불안 재현율과 정확도에서 우수한 성능을 보였지만, 불안 재현율에서 성능이 좋지 않아 종합 성능이 최대 우도 추정의 0.590보다 낮은 0.458이다. 따라서 본 논문에서는 최대 우도 추정을 사용한다.[12-14]

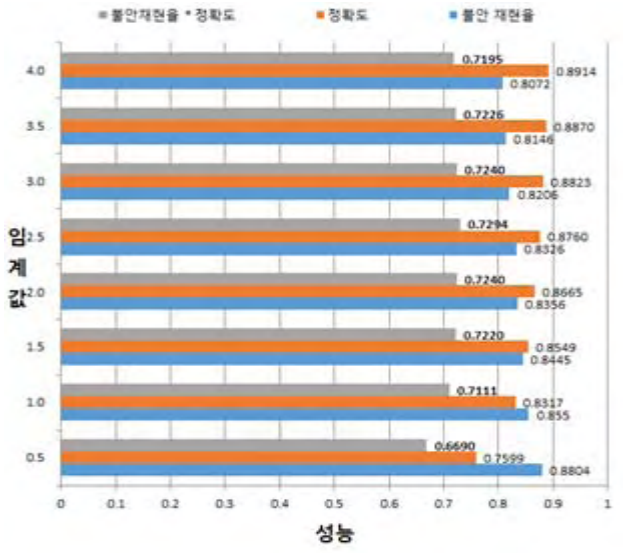
4.2 임계값

본 논문에서 사용한 최대 우도 추정을 이용한 나이브 베이즈 분류기의 성능을 최적화 시키기 위한 임계값을 찾는다.

$$\frac{P(\text{트윗}|\text{불안})}{P(\text{트윗}|\text{비불안})} > \text{임계값} \quad \cdots(5)$$

본 논문에서 사용한 분류기는 식 (5)가 참일 경우 불안으로 분류하고, 거짓일 경우 비불안으로 분류한다. 임계값을 변경해가며, 분류기의 성능을 평가한 결과는 (그림 3)과 같다.

본 논문은 불안을 분석하기 위한 논문이므로 지표 중 불안 재현율과 정확도를 중요시한다. 따라서 두 지표를 동시에 활용하는 (불안 재현율 * 정확도)를 성능 지표로 한다. 이 성능지표가 임계값 2.5에서 0.7294로 가장 높기 때문에 임계값을 2.5로 설정한다.



(그림 3)임계값에 따른 성능

5. 결론 및 향후 연구

본 논문에서는 SNS로부터 시공간 정보가 포함된 데이터를 수집 및 분류하여, 사회 불안의 분포를 시각화하는 기법을 소개했다. 사회 과학자들이 이를 활용하여 체계적으로 불안을 연구하고, 사회 불안 문제들의 원인 파악 및 해결을 할 수 있다.

본 논문에서 소개한 기법에 대한 향후 연구로는 두 가지 방향이 있다. 첫 째는 분류의 추가이다. 불안 뿐 아니라 분노, 행복, 슬픔 등 세분화된 감정 판별 기술을 개발한다. 두 번째는 분류기의 성능 개선이다. 트레이닝 데이터를 추가하고, 최대 우도 추정을 사용한 나이브 베이즈 분류기 외에 다른 알고리즘을 시도하여 더 성능이 좋은 분류기를 개발한다.

참고문헌

[1] Brader, Ted, George E. Marcus, and Kristyn L. Miller.: Emotion and public opinion, The Oxford Handbook of American Public Opinion and the Media (2011)

[2] Stopa, Lusia and Clark, David M,: Social phobia and interpretation of social events, Behaviour research and therapy, 38.3, 273-283 (2000)

[3] Nasreen, Hashima E., et al,: Low birth weight in offspring of women with depressive and anxiety symptoms during pregnancy: results from a population based study in Bangladesh, BMC public health, 10.1, 515 (2010)

[4] Zelkowitz, Phyllis, Claudette Bardin, and Apostolos Papageorgiou,: Anxiety affects the relationship between parents and their very low birth weight infants, Infant Mental Health Journal, 28.3 296-313, (2007)

[5] Montgomery, Scott M., et al.: Unemployment pre-dates symptoms of depression and anxiety resulting in medical consultation in young men, International Journal of Epidemiology, 28.1, 95-100, (1999)

[6] Wright, Kevin B.: Researching Internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services, Journal of Computer-Mediated Communication, 10.3, (2005)

[7] Efron, B. (2013). Bayes' theorem in the 21st century. Science, 340(6137), 1177-1178.

[8] Murphy, K. P. (2006). Naive bayes classifiers. University of British Columbia.

[9]Yusuke Yamamoto: TWITTER4J, http://twitter4j.org/en/index.html, (2007), Accessed 30 June 2017

[10] Junsoo Shin: KOMORAN, GitHub repository, https://github.com/shin285/KOMORAN, Accessed 30 June 2017

[11] James Beck, James Vere, and Kenneth J. Arnold.: Parameter estimation in engineering and science, (1977)

[12] Grobelnik, Marko.: Feature selection for unbalanced class distribution and naive bayes, ICML, (1999)

[13] Frank, Eibe, and Remco Bouckaert.: Naive bayes for text classification with unbalanced classes, Knowledge Discovery in Databases: PKDD 2006, 503-510, (2006)

[14] Chawla, Nitesh V.: Data mining for imbalanced datasets: An overview, Data mining and knowledge discovery handbook, Springer US, 853-867, (2005)

[15] Ng, Vincent, and Claire Cardie.: Improving machine learning approaches to coreference resolution, Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, (2002)