LDA를 이용한 트윗 유저의 연령대, 성별, 지역 분석

이호경O, 천주룡, 송남훈, 고영중
동아대학교
hogay88@gmail.com, balendia@gmail.com, nh.song.89@gmail.com, youngjoong.ko@gmail.com

Analyzing ages, gender, location on Twitter using LDA

Ho-Kyung LeeO, Ju-Ryong Chun, Nam-Hoon Song, Youngjoong Ko
Donga University, Computer Engineering

요 약

요즘 많은 사람들은 트위터를 통해 짧은 문장의 트윗을 작성하여 자신의 의견이나 생각을 표현한다. 사람들이 작성한 트윗은 사용자의 연령, 성별, 지역에 따라 다른 특성이 담겨있다. 이러한 정보를 이용하여, 기업에서는 연령대, 성별, 지역에 따라 각기 다른 마케팅 전략을 세울 수 있을 것이다. 본 논문에서는 트위터 사용자들의 트윗을 분석하여 연령대, 성별, 지역을 예측하려 한다. 네이버 오픈사전의 자질, 한국전자통신연구원(ETRI)의 개체명 사전을 이용한 자질 및 한국어 형태소 분석, 음절 단위의 bigram을 클래스별 의미 있는 자질로 선택하고 LDA를 이용하여 예측된 확률분포를 활용하여 분류한 결과, 연령 72%, 성별 75%, 지역 43%의 납득할만한 예측 정확도 결과를 얻게 되었다.

주제어: Twitter, LDA(Latent Dirichlet Allocation), 연령대, 성별, 지역

1. 서론

최근 들어, 소셜 네트워크 서비스(Social Network Service)가 확산 되면서 사람들이 자신의 의견, 생각, 개인적인 경험을 공유하고 표현할 수 있게 되었다. 트위터(twitter)는 블로그의 인터페이스와 미니홈페이지 기능, 메신저의 기능을 한데 모아놓은 소셜 네트워크 서비스이다. 트위터는 하나의 트윗(tweet)을 140자 이내로 제한하고 있으며, 그 트윗은 사용자의 의견이나 생각을 포함하고 있다. 트위터 뿐만 아니라, 전 세계적 SNS에 대한 관심 및 스마트폰 보급률의 증가에 따라 SNS를 사용하는 사용자들의 수가 증가하고 있으며 또한 트위터와 관련된 많은 연구가 수행되고 있다.

그리고 이러한 트위터 분석 연구는 트위터 사용자의 성별, 연령대, 거주지역 그리고 기업에서는 관심있는 보험상품을 조회하는 패턴을 파악하여 고객의 요구사항을 분석하고 신규 보험 상품 개발 등에 이용되는 사례에 적용 되고 있다.

사람들은 개인마다 자신만의 특성이 드러나는 글쓰기 방식을 가지고 있다. 이러한 개개인의 글쓰기 방식 특성을 이용하여 연령대를 기준으로 집단의 문체를 분석, 판별하고자 하는 연구들이 많이 이루어지고 있다. 이들 연구는 같은 연령대의 사람들은 비슷한 시기에 동일한 공감되는 일들을 겪으면서 다른 연령대와 구별되는 성향이 글속에 잘 나타나게 된다는 점에서 착안한 것이다. 연령대를 분류하는데 좋은 자질로는 이모티콘 자질이 있다. 10/20대의 경우 30/40대에 비해 이모티콘의 사용 빈도가 높다. 성별을 분류하는데 에는 호칭에 관한 단어들이 좋은 자질로 이용된다. 그리고 지역에 따른 좋은 자질로는 각종 지역 명칭이나 사투리 사전을 이용 할 수 있다.

본 논문에서는 네이버의 오픈 사전, ETRI 개체명 사전과 같은 자원을 활용하여 연령대, 성별, 지역을 분류하는데 자질들을 선정하여 사용하였으며 트윗 길이 정보[2], 자음 및 모음으로 구성된 한글 이모티콘 개수 정보[3] 또한 자질로 이용하였다. 연령대, 성별, 지역별로 구성한 자질들을 LDA를 활용하여 분포를 계산하고, 계산된 분포를 기반으로 트위터 사용자의 연령대, 성별, 지역을 분류하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 제안하는 방법과 연관된 관련 연구에 대해 기술하고 3장에서는 제안하는 기법의 전체적인 구성과 각 부분에 대해 설명한다. 4장에서는 LDA를 이용한 실험방법 및 결과를 살펴보고 5장에서는 결론 및 향후 연구에 대해 말한다.

2. 관련 연구

트위터에 관한 초기 연구들은 트윗 행태와 이용자에 관한 기초통계 분석 위주의 연구들이었으나, [1]에서는 한국어 트위터에서 연령대에 따라 문체가 어떻게 달라지는가를 비교적 적은 규모의 자질을 통해 분석하고 예측하였다. [1]에 더하여 [2]에서는 트윗의 문체뿐 아니라 내용을 다루고 있는 부분들을 계량화 할 수 있게 자주

이 논문은 2013년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2013R1A1A2009937)

1) 각 사용자의 한 트윗당 길이 정보

2) 한글 모음 및 자음으로 구성된 이모티콘 개수
예) ㅎㅎㅎㅎ, ㅋㅋㅋㅋ

쓰이는 n-gram 방식과 같은 자질들을 추가로 구성하여 연령대 및 성별 예측을 하였다.

다른 트윗 분석 연구로는 다양한 시간 스케일에서 주제를 추출할 수 있는 NMF(Non-negative Matrix Factorization) 클러스터링 기법을 적용하여 트위터의 트랜드를 [3]에서 분석하였고, [4]에서는 n-gram을 이용한 자질 추출 기법과 슬라이딩 윈도우(sliding window)를 이용하여 자질을 추출하여 트윗을 분석하였다. [5]에서는 100여개에서부터 많게는 5,700가지의 자질들의 분포를 토대로 연령대를 예측하는 방법을 제안하였다.

마지막으로 [6]에서는 LDA를 기반으로 트위터 데이터를 분석하여 토픽의 변화 시점 및 패턴을 파악하는 연구를 진행하였다.

3. 제안 방법

본 논문에서는 한국어 형태소 분석 결과[가]와 음절 단위 bigram[나]을 기본 자질로 사용하고, 여기에 연령대, 성별, 지역을 분류하는데 좋은 정보를 주는 자질들을 추가로 사용한다. 한국어 형태소 분석 정보는 명사 정보만을 이용한다. 연령대, 성별, 지역을 분류하기 위해 여러 자질들을 기반으로 LDA를 이용하여 분포를 추정한다. 추정된 분포를 활용하여 사용자가 작성한 여러 트윗의 연령대, 성별, 지역을 분류함으로써 효과적으로 사용자의 연령대, 성별, 지역을 분류하는 방법을 제안한다.

3.1 데이터 구축

트위터 사용자들의 개인 프로필 및 국내 트위터 관련 사이트를 이용하여 트위터 사용자의 연령대, 성별, 거주지역에 따른 트윗의 가장 최근의 데이터만을 수집한다. 성별은 남, 여로 분류, 연령대는 10대~20대, 30대~40대로 분류, 그리고 거주지역은 수도권, 충청도권, 전라도권, 경상도권으로 분류하여 모든 트위터 사용자를 균등하게 수집한다. 트윗 하나당 문서 하나로 구성하여 외국어로 된 트윗, 반복된 트윗, 광고성 트윗을 제거하는 정제과정을 거친다.

3.2 자질 선택 및 추출

표 1. 자질 분류

구분		예
연령대 분류 자질 [다]	[다-1] 채팅어, 유행어, 신조어	쩐다, 헐랭, 멘붕, 근자감, 귀요미, 본좌, 득템 ...
연령대 분류 자질	[다-2] 정치, 경제, IT, 연예 뉴스	4대강 국회 부동산 금리 보조금 KT 오자룡 ...
[다]		
연령대 분류 자질 [다]	[다-3] 트윗 길이 정보	밥 먹자!/S0, 태안 근흥초등학교에서 ...(중략)... 맛보았습니다. 새삼 SNS의 위력을 절감했습니다./S1
	[다-4] 자음 및 모음 한글 이모티콘 개수 정보	ㅎㅎㅎㅎ ㅋㅋㅋㅋ ㅇㅋㅇㅋ/E0
성별 분류 자질 [라]	[라-1] 남녀 호칭, 특수기호	언니, 오빠, 누나 형, 누님, 형님 ... ※, ★, ☆, ♥, ♡, ♠, ♤, ▲, △, ◈ ...
지역별 분류 자질 [마]	[마-1] 지역명, 관광지명 및 사투리	수원, 인천, 대구, 광주, 청주, 창원, 청담동, 이태원, 종로, 해운대, 사하구, 월미도, 북한산, 지리산 ... 가구웁는, 개갈, 산찬하다, 부앙부앙하다, 천지빽가리, 포분들리다 ...

* S0 : 트윗의 길이 5자 이하
* S1 : 트윗의 길이 81자 이상
* E0 : 자음 및 모음 한글 이모티콘 개수 10개 초과

연령대 자질으로는 10대~20대와 30대~40대를 구분할 수 있는 자질로 네이버 오픈 사전을 이용하여 [다-1]에 수록된 단어들 중 추천수 10이상만 추출, 또한 네이버의 기사 중 정치, 경제, IT, 연예 뉴스[다-2]들의 주제문의 명사만을 추출하여 총 1354개의 단어로 이루어진 연령대 자질 사전을 구축하였다. 10대~20대와 30~40대를 구분할 수 있는 또 다른 자질로 연령대마다 다를 것이라고 판단되는 [다-3]를 자질로 이용하였다. 또한 'ㅋ'나 'ㅎ' 또는 'ㅠㅠ'와 같은 [다-4]를 자질로 이용하였다.

[다-3]는 트윗의 길이를 5자 이하 'S0'태그, 81자 이상은 'S1'태그를 부착하여 사용하였으며 [다-4]는 트윗의 자음 및 모음 한글 이모티콘 개수가 10개 초과하는 트윗에 'E0'태그를 부착하여 사용하였다.

성별 자질 선택으로는 남, 여를 구분할 수 있다고 판단되는 [라-1]를 자질로 하여 총 82개의 성별 자질 사전을 구축하였다.

지역 자질 선택으로는 한국전자통신연구원(ETRI)의 개체명 사전에서 각 지역을 구분할 수 있다고 판단되는 수도권, 충청도, 전라도, 경상도의 지역명 및 관광지명을 추출하고 또한 네이버 오픈사전의 사투리 사전에서 충청

도, 전라도, 경상도의 사투리를 추천수 10이상만 추출하여 총 8255개의 [마-1]을 구축하였다.

3.3 LDA의 적용 및 계산

사용자의 각 트윗에서 제안한 자질들을 추출하고 LDA를 적용하여 분포를 계산하였다. LDA에서 연령대, 성별은 토픽을 2개로, 지역은 토픽을 4개로 설정하였다. 아래는 연령대별 각 주제에서 가중치 기준으로 상위 N 개의 단어의 목록의 예이다.

표 2. 연령대별 자질 사전 토픽 결정 예

토픽1	토픽2
학교	기업
출첵	퇴근
멘붕	부동산
헐	회사
아이돌	맛집
문상	회장
...	...

본 논문에서는 LDA를 이용하여 계산된 토픽-단어 분포를 기반으로 트윗의 토픽을 결정한다. 예를 들어, 연령대별 토픽 분포에서 토픽1에 10대~20대의 자질이, 토픽2에 30~40대의 자질이 각각 많이 등장하였다고 판단되면 토픽1은 10대~20대, 토픽2는 30~40대라고 직접 결정한다.

표 3. 문서별 가중치에 따른 토픽 예측 예

	문서	토픽1	토픽2
사용자	문서1	**3.244**	2.756
	문서2	0.352	**1.648**
	문서3	**21.528**	6.472
	문서4	**5.616**	2.384
	문서5	**5.000**	0.000
	대소 비교	4	1

사용자의 토픽은 LDA를 적용하여 나온 문서당 토픽 분포 가중치의 결과를 이용하여 예측한다. 한 트위터 사용자의 토픽을 예측하는 방법은 다음과 같다.

1. 한 트위터 사용자당 트윗 문서에서 토픽들의 가중치를 대소 비교한다.

2. 개수가 많이 나온 토픽을 선택하여 그 트위터 사용자를 예측한다.

예를 들어 한 사용자가 n개의 트윗 문서 $\{d_1, d_2, \cdots, d_n\}$를 가질 때, j번째 토픽 t_j 에 대한 문서 d_i의 스코어를 $Score_i(t_j)$라고 하면, $Vote_i(t_j)$는 수식 (1)과 같이 정의한다.

$$Vote_i(t_j) = \begin{cases} 1 & \text{if } j = (\text{argmax}_x Score_i(t_x)) \\ 0 & else \end{cases} \quad (1)$$

그리고 그 사용자는 최종적으로 수식 (2)에서 산출된 토픽 t_y 에 매핑시킨다

$$\text{argmax}_y \sum_{i=1}^{n} Vote_i(t_y) \quad (2)$$

즉, 앞에서 클래스 별로 결정한 토픽과 한 트위터 사용자의 트윗 문서에서 예측한 토픽을 비교하여 정답을 판단한다.

4. 실험 및 결과

각 클래스 별로 형태소 분석기반 자질, bigram 자질을 baseline로 설정하고 본 논문에서 제안한 연령대, 성별, 지역별로 분류된 자질 정보 및 사전을 조합하여 트위터 사용자를 예측하였다.

예측 정확도의 계산 방법은 기본적으로 f1-measure을 사용하였고 각 클래스 별 micro-평균으로 계산하였다.

4.1 데이터 집합

말뭉치는 최근 트위터를 활발히 이용하는 100명의 트윗 사용자당 190~200개씩, 총 19567개의 트윗으로 구성하였다. 그리고 개인 프로필 및 국내 트위터 관련 사이트에 기록된 것을 기준으로 10대, 20대, 30대, 40대별로 25명씩, 남자 51명, 여자 49명, 지역을 수도권 30명, 충청도권 21명, 전라도권 22명, 경상도권 27명으로 각 클래스 별로 균등하게 구성하여 트윗을 수집하였다.

4.2 실험 결과

실험은 한 트위터 사용자를 예측하는 결과이다.

표 4. 실험 방법에 따른 연령대 예측 결과

방법	10대/20대	30대/40대	micro
[가]	0.429	0.586	0.520
[나]	0.505	0.515	0.510
[다]	0.632	0.512	0.580
[다] + [다-3]	0.568	0.661	0.620
[다] + [다-3] + [다-4]	0.750	0.625	0.700
[가] + [다] + [다-3] + [다-4]	**0.763**	**0.659**	**0.720**
[나] + [다] + [다-3] + [다-4]	0.775	0.592	0.710

[가] : 형태소 분석 명사 자질 [baseline1]

[나] : bigram 자질 [baseline2]
[다] : 연령대 자질 사전
[다-3] : 트윗 길이 정보
[다-4] : 한글 이모티콘 개수 정보

연령대 예측 결과는 형태소 분석 명사 자질과 연령대 자질 사전, 트윗 길이 정보, 한글 이모티콘 개수 정보를 모두 조합한 방법이 가장 높은 72%의 정확도로 예측할 수 있었다.

표 5. 실험 방법에 따른 성별 예측 결과

방법	남	여	micro
[가]	0.598	0.434	0.530
[나]	0.608	0.592	0.600
[라]	**0.771**	**0.725**	**0.750**
[가] + [라]	0.452	0.603	0.540
[나] + [라]	0.432	0.667	0.580

[가] : 형태소 분석 명사 자질 [baseline1]
[나] : bigram 자질 [baseline2]
[라] : 성별 자질 사전

표 6. 실험 방법에 따른 지역별 예측 결과

방법	수도권	충청도	전라도	경상도	micro
[가]	0.286	0.244	0.333	0.222	0.270
[나]	0.320	0.324	0.340	0.333	0.330
[마]	**0.233**	**0.421**	**0.379**	**0.633**	**0.432**
[가] + [마]	0.413	0.308	0.235	0.171	0.320
[나] + [마]	0.381	0.316	0.233	0.171	0.300

[가] : 형태소 분석 명사 자질 [baseline1]
[나] : bigram 자질 [baseline2]
[마] : 지역 자질 사전

그에 반해 성별 예측 결과와 지역별 예측 결과는 형태소 분석 명사 자질이나 bigram자질과 함께 적용한 자질 사전을 모두 조합한 방법의 예측 정확도가 생각보다 높지 않았다. 성별이나 지역별 예측에서는 필요하다고 판단되는 자질이 한정적이기 때문에 많은 자질이 포함된 형태소 분석 명사 자질 및 bigram 자질은 역효과가 나지 않았나 판단된다.

따라서 성별 예측 결과에서는 성별 자질 사전만을 사용한 방법이 가장 높은 75%의 정확도를 보여주었고 지역별 예측 결과에서는 지역별 자질 사전만을 사용한 방법이 약 43%의 정확도로 가장 높게 예측되었다.

baseline로 적용했던 형태소 분석 명사 자질과 bigram 자질은 토픽을 2개로 나눈 연령대, 성별 예측 결과는 50%가 조금 넘는 일반적인 정확도를 보여주며 마찬가지로 토픽 4개로 나눈 지역별 예측 결과는 25%가 조금 넘는 정확도를 보여주었다. 본 논문에서 제안한 각 클래스별 자질 정보와 사전을 이용하였을 때 확실히 예측 정확도가 높아지는 것을 볼 수 있었다.

5. 결론

본 논문에서는 각 클래스 별 특성에 맞는 자질 사전을 구축하여 LDA를 이용, 트위터 사용자를 예측하였다. [2]에서 제시한 자질 구축 방법 및 SVM을 이용한 예측과 대응하여, 학습을 하지 않고도 각 클래스별로 판단되는 자질을 추출하여 LDA를 이용함으로써 납득할 만한 정확도를 예측하였기 때문에 의미 있는 연구가 되었다고 생각한다. 그리고 향후 연구로는 트위터의 리트윗(Retweet)[1], 리플라이(Reply)[2]기능을 이용하여 사용자간의 관계를 이용하여 정확도를 높이고 나아가 사용자의 직업, 관심사 등의 토픽을 늘려 예측 할 수 있는 시스템을 구축하는 연구를 진행 할 예정이다. 이러한 연구들이 활발히 진행된다면 마케팅 분야에 적절히 활용 될 수 있을 것이라 생각된다.

참고문헌

[1] 김상채, 박종철, “한국어 트윗의 문체 기반 자질 분석을 통한 연령대 예측” , HCI 2012 학술대회
[2] 김상채, 박종철, “문체 분석을 활용한 한국어 트위터 사용자의 연령대 및 성별 예측”, 2012 한국컴퓨터종합학술대회 논문집 Vol.39, No.1(B)
[3] 하용호, 임성원, 김용혁, “내용기반 트윗 클러스터링을 통한 트랜드 분석”, 2012년 가을 학술발표논문집 Vol.30, No.2(B)
[4] 홍초희, 김학수, “트윗 분류를 위한 효과적인 자질 추출”, 2011 한국컴퓨터종합학술대회 논문집 Vol.38, No.1(A)
[5] J. D. Burger, J. Henderson, G. Kim and G. Zarrella,“Discriminating Gender on Twitter”, Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 1301-1309, 2011.
[6] 전설아, 허고은, 정유경, 송 민, “트위터 데이터를 이용한 네트워크 기반 토픽 변화 추적 연구”,정보관리학회지 30(1), 285-302. 2013.

1) 리트윗(Retweet)이란 자신의 스트림에 나타난 다른 사용자의 트윗을 자신의 팔로워들에게 전파하는 기능

2) 리플라이(Reply)란 자신의 스트림에 나타난 다른 사용자의 트윗에 대한 사용자의 답변을 작성한 트윗을 사용자에게 보내는 기능