Naïve Bayes 와 SVM 을 이용한 트위터 데이터의 긍정/부정 의견 자동분류 결과 분석†

조희련*, 김성국**
*연세대학교 미래융합기술연구소
**연세대학교 글로벌융합공학부
e-mail : {heeryon, songkuk}@yonsei.ac.kr

Initial Analysis of Positive/Negative Opinion Classification of Twitter Data Using Naïve Bayes and SVM

Heeryon Cho*, Songkuk Kim**
*Yonsei Institute of Convergence Technology, Yonsei University
**School of Integrated Technology, Yonsei University

요 약

'나꼼수 비키니 시위'에 대한 긍정적(지지), 부정적(비판) 의견을 담은 트위터 데이터를, 단어의 출현에 주목하여 Naïve Bayes (NB)와 Support Vector Machine (SVM)을 적용하여 자동분류 한 결과, NB 가 75.98%로, 73.65%인 SVM 보다 약간 더 나은 성능을 보였다. 본 실험을 통해, 기계학습을 이용한 대중의견(opinion) 자동분류 시스템을 실용화할 때의 고려사항에 대해 살펴 본다.

1. 서론

개인이 자신의 의견을 인터넷(World Wide Web) 상에 쉽게 개진 할 수 있게 되면서, 대중의견을 자동으로 분류・분석하려는 연구가 최근 들어 활발히 진행되고 있다[1]. 특히, 개인의 생각을 140 자의 짧은 텍스트에 담아 공유할 수 있게 하는 '트위터[2]'와 같은 온라인 소셜 서비스의 보급으로, 대중의견 데이터의 입수가 과거에 비해 용이해 졌는데, 이것이 대중의견 데이터 분류・분석 연구에 박차를 가하고 있다.

한편, 하드웨어와 소프트웨어 기술의 발달로 계산기의 연산능력이 크게 증가하면서, 이전에는 비현실적이던 기계학습의 실질적인 활용이 가능해 졌다.

이러한 호재를 맞아, 우리는 인터넷 상의 대중의견 데이터를 수집・통합・가시화 하는 시스템을 연구개발하고 있는데, 그 첫 걸음으로, 특정 관심 대상에 대한 긍정/부정 의견 데이터를 수집하여, 대표적인 기계학습 기법을 적용해 자동분류 성능을 검증해 보았다. 이 때 기계학습 기법으로는 Naïve Bayes (NB)와 Support Vector Machine (SVM)을 선택했고, 자동분류 대상 데이터로는 긍정적/부정적 의견을 고루 담고 있는 '나꼼수 비키니 시위' 트위터 데이터를 수집했다.

'나꼼수 비키니 시위' 논쟁은, '나는 꼼수다[3] (줄여서 '나꼼수')'라는 인터넷 팟캐스트가, BBK 사건 관련 허위사실 유포 혐의로 구속 수감된 민주통합당 정봉주[4] 전 의원을 석방하기 위해 1 인 시위를 독려하는 과정에서 불붙은 논쟁이다.

† 본 연구는 지식경제부 및 정보통신산업진흥원의 "IT 명품인재양성사업"의 연구결과로 수행되었음 (NIPA-2010-C1515-1001-0001)

이 과정에서 한 여성이 비키니 차림의 1 인 시위 사진을 인터넷에 올렸고, 이에 대해 나꼼수 진행자들이 몇 가지 발언을 했는데, 나꼼수의 시위 독려 행위와 발언을 문제 삼는 비판 세력과, 이를 문제 삼지 않는 지지 세력이 트위터 상에서 첨예하게 대립했다.

그러나 트위터 상의 대중의견이 늘 찬성 또는 반대, 긍정 또는 부정으로 팽팽하게 맞서는 것은 아니다. 실제로 '나꼼수 비키니 시위' 데이터를 분석하기 앞서, 우리는 '서울시 무상급식[5]' 데이터를 분석해 보았는데, 90% 이상의 트위터 데이터가 무상급식을 지지하는 내용을 담고 있음을 확인할 수 있었다. 이렇듯 다수의 의견이 한 쪽으로 쏠리는 경우, 단순히 모든 데이터를 다수의 의견으로 분류하는 것만으로도 높은 자동 분류 성능을 달성할 수 있다. 그러나 이렇게 동일한 의견이 범람하는 경우에는, 다수의 의견을 파악하는 것 보다는, 다수와는 다른 소수의 의견을 자동으로 선별(분류)해 보여주는 것이 더 값질 수 있다. 두 경우 모두, 정확한 의견 분류가 결정적이다.

이후 본고에서는 긍정(지지)과 부정(비판)의 의견이 팽팽히 맞서는 경우의 의견분류에 초점을 맞춰, 대표적인 기계학습 기법의 성능을 검증함으로써 대중의견 자동분류 시스템 실용화의 고려사항에 대해 알아본다.

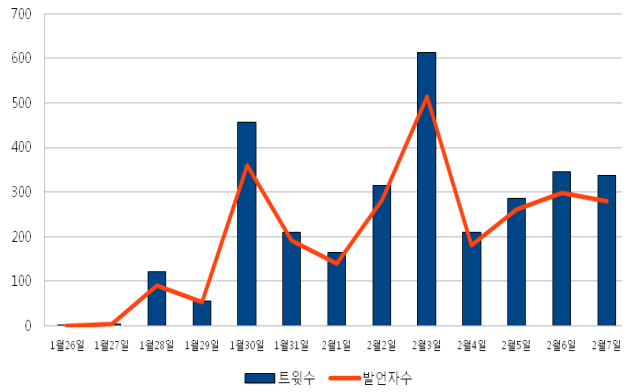
2. 실험 데이터

'나꼼수 비키니 시위' 트위터 데이터를 수집하기 위해, 과거 트위터 데이터 검색 서비스인 Topsy[6]의 Otter API[7]를 이용하여 '나꼼수'와 '비키니'의 문자열을 포함하는 트윗(1 개의 트위터 데이터)을 수집했다.

데이터 수집 대상기간은 해당 논쟁이 한창 진행 중이던 2012 년 1 월 26 일부터 2012 년 2 월 7 일까지의 13 일로 했다. (이후에도 논쟁은 계속됐으나 본고의 데이터는 이를 포함하지 않는다.)

데이터 수집 결과, 총 4,363 개의 트윗이 수집됐고, 각각의 트윗을 사람이 읽은 후, 해당 트윗이 '나꼼수 비키니 시위'를 (1) 비판하는 내용을 담고 있는지, (2) 지지하는 내용을 담고 있는지, (3) 중립적이거나 관련이 없는 내용을 담고 있는지를 판단해 레이블을 달았다. 그 후 (3)의 중립 데이터를 배제하고, 데이터를 선별하여, 총 1,377 개의 비판(negative, 44.11%) 트윗과 총 1,745 개의 지지(positive, 55.89%) 트윗을 더해, 모두 3,122 개의 트윗을 자동분류 실험 데이터로 삼았다.

(그림 1)에 '나꼼수 비키니 시위' 데이터의 일일 트윗수와 고유 발언자수를 표시한다. 그래프를 보면 2012 년 2 월 3 일에 514 명의 고유 발언자가 모두 614 개의 '나꼼수'와 '비키니'를 포함하는 트윗을 날리는 것으로 (한 사람이 여러 개의 트윗을 날리는 경우 포함) '비키니 시위' 논쟁이 절정에 달하고 있는 것을 알 수 있다. 또, 그 이전인 1 월 30 일 (트윗 456 개) 에 먼저 논쟁이 피크에 이르는 것을 확인할 수 있다.



(그림 1) 일일 트윗수와 고유 발언자수

그렇다면 무엇이 논쟁을 가열시켰을까? 개별 트윗에 출현하는 단어를 분석하는 것으로 우리는 힌트를 얻을 수 있다. <표 1>은 개별 트윗을 한국어 형태소 분석기[8]에 돌려, 보통 명사와 고유 명사를 먼저 수집하여 명사 리스트를 작성한 후, 리스트에서 고유 단어(unique words)를 추출하여 일별로 집계한 결과이다. 대상 기간에 걸쳐 출현빈도의 합계가 많은 순으로 단어를 정렬하여, 상위 30 위의 단어를 제시한다. '나꼼수'와 '비키니'를 검색어로 하여 데이터를 수집했기 때문에, 당연히 '나꼼수'와 '비키니'가 각각 1, 3 위를 차지하고 있다. (19 위의 '비키니시위'도 고려.)

주목해야 할 부분는 중간에 자리잡고 있는 '공지영'과, 밑에서 3 번째에 위치한 '이보경'이다. '공지영'은 1 월 30 일에 갑자기 114 회나 출현하고, '이보경'도, '비키니 시위' 논쟁 시작 이후 한 번도 언급되지 않다가, 2 월 3 일에 갑자기 회자되고 있다. 분석 결과, 두 경우 모두 매스컴이 트윗 내용을 보도하면서, 트위터 상의 논쟁이 가열된 것으로 파악됐다.

<표 1> 일일 고유 단어의 출현 빈도

단어/날짜	1/26	1/27	1/28	1/29	1/30	1/31	2/1	2/2	2/3	2/4	2/5	2/6	2/7	합계
나꼼수	1	3	112	53	419	187	151	300	563	200	270	328	284	2871
시위	0	0	30	10	110	21	41	70	191	48	152	91	46	810
비키니	0	1	3	5	61	32	24	100	116	47	105	123	66	683
논란	0	0	4	5	81	38	15	119	166	29	29	63	67	616
정봉주	0	0	15	8	125	20	31	53	78	31	23	14	13	411
성희롱	0	1	3	1	46	14	30	27	44	17	136	59	17	395
사과	0	0	17	4	31	19	4	33	127	32	15	33	31	346
발언	0	1	5	4	35	26	13	27	30	18	97	51	19	326
여성	0	1	15	11	42	27	20	23	38	10	23	43	17	270
문제	0	0	8	6	26	10	8	30	31	26	30	27	27	229
말	0	0	9	3	25	22	13	35	35	12	20	17	21	212
사진	0	1	16	11	39	17	9	11	34	11	14	32	10	205
생각	0	0	10	7	41	10	13	23	36	13	8	21	21	203
응원	0	0	9	2	21	12	17	27	46	18	6	12	12	182
공지영	0	0	2	8	114	16	3	3	10	1	2	4	4	167
사건	0	0	6	4	27	15	20	20	15	8	13	19	19	166
요구	0	0	16	6	10	3	5	16	21	6	18	32	20	153
표현	0	1	4	2	29	9	2	9	43	4	23	19	5	150
비키니시위	1	0	4	0	18	10	6	13	36	8	16	15	9	136
인증샷	0	0	7	1	8	6	2	17	20	7	31	21	14	134
주진우	0	0	9	6	5	3	14	10	41	17	6	11	4	126
가슴	0	1	6	5	40	10	11	7	17	8	6	7	5	123
김어준	0	0	0	1	7	3	5	11	19	2	26	34	15	123
사람	0	0	3	6	17	15	8	8	23	8	9	16	8	121
불쾌	0	1	18	7	56	9	3	5	7	3	4	1	6	120
아니다	0	0	5	0	1	2	0	0	5	0	78	28	0	119
진보	0	0	15	1	22	7	8	14	17	4	6	11	14	119
이보경	0	0	0	0	0	0	0	0	60	14	2	27	15	118
방송	0	0	2	1	3	3	3	58	16	1	9	12	9	117
입장	0	0	4	2	4	1	1	6	51	9	16	14	6	114

논쟁 가열의 두 경우에 차이가 있다면, 전자는 소설가 공지영씨가 '비키니 시위'를 비판하는 트윗을 1 월 28 일에 날린 후, 매스컴이 이 사실을 1 월 30 일에 대대적으로 보도한 점이고, 후자는 MBC 여기자가 '비키니 시위' 동참 사진을 2 월 3 일에 올리자 마자 매스컴이 이를 당일 보도한 점이다. 즉, 트위터 상의 발언과 매스컴을 통한 보도가 순차적으로 이루어졌는가, 아니면 즉각적으로 이루어졌는가의 차이이다. 이러한 차이도 <표 1>의 단어 패턴에서 유추할 수 있다.

3. 실험 방법

대표적인 기계학습 기법인 NB 와 SVM 을 '나꼼수 비키니 시위' 트위터 데이터의 긍정/부정 의견 자동분류에 적용하기 위해 Ruby 언어[9]로 구현된 NB classifier [10]와 libsvm [11]을 사용했다. 훈련과 테스트를 위한 데이터의 특징(feature)으로는, 트윗별로 단어의 출현여부(presence)에 주목했다. 아래는 실제 데이터의 변환예이다.

트윗 문장: "나꼼수가 비키니 사건에서 청취자들을 대상으로 여성의 성을 상품성으로 보는것을 단순 해프닝으로 치부해버리면 권력을 가지지 못한 약자들의 설움을 대변하고자 했던 나꼼수의 문제의식과 맞지않는 행동입니다."

데이터 변환결과: ["나꼼수", "사건", "청취자들", "대상", "여성", "성", "상품성", "단순", "해프닝", "치부", "권력", "약자들", "설움", "대변", "문제의식", "행동"]

특기할 사항은 (1) 실험 데이터의 구성을 명사 vs. 모든 단어로 나눈 것과, (2) 실험 방법을 3-fold cross validation (CV) 이외에 초기 데이터의 학습 vs. 대상 기간 전반에 걸친 데이터의 학습으로 설정한 점이다.

<표 3> Naïve Bayes 와 SVM 의 긍정(지지) / 부정(비판) 의견 자동분류 성능 비교

분류 기법	학습 방법	고유 단어수 (8,173)							
		훈련 성능 (%)				테스트 성능 (%)			
		error pos	error neg	error all	accuracy	error pos	error neg	error all	accuracy
Naïve Bayes	첫 1,041개	2.62 (13/496)	2.20 (12/545)	2.40 (25/1041)	97.60 (1016/1041)	42.19 (527/1249)	33.89 (282/832)	38.88 (809/2081)	61.12 (1272/2081)
	3의 배수	3.54 (21/593)	3.13 (14/448)	3.36 (35/1041)	96.64 (1006/1041)	29.51 (340/1152)	18.84 (175/927)	24.75 (515/2081)	75.25 (1566/2081)
	3-fold CV	***4.31***	***2.23***	***3.40***	***96.60***	***31.63***	***17.38***	***25.30***	***74.70***
SVM	첫 1,041개	1.81 (9/496)	4.04 (22/545)	2.98 (31/1041)	97.02 (1010/1041)	43.23 (540/1249)	35.46 (295/832)	40.12 (835/2081)	59.88 (1246/2081)
	3의 배수	1.52 (9/593)	6.47 (29/448)	3.65 (38/1041)	96.35 (1003/1041)	17.27 (199/1152)	33.48 (311/929)	24.51 (510/2081)	75.49 (1571/2081)
	3-fold CV	***2.12***	***6.03***	***3.84***	***96.16***	***19.49***	***32.49***	***25.21***	***74.79***
분류 기법	학습 방법	고유 단어수 (9,509)							
		훈련 성능 (%)				테스트 성능 (%)			
		error pos	error neg	error all	accuracy	error pos	error neg	error all	accuracy
Naïve Bayes	첫 1,041개	2.02 (10/496)	1.65 (9/545)	1.83 (19/1041)	98.17 (1022/1041)	43.23 (540/1249)	29.57 (246/832)	37.77 (786/2081)	62.23 (1295/2081)
	3의 배수	3.04 (18/593)	1.79 (8/448)	2.50 (26/1041)	97.50 (1015/1041)	26.91 (310/1152)	22.07 (205/929)	24.75 (515/2081)	75.25 (1566/2081)
	3-fold CV	***3.63***	***1.53***	***2.69***	***97.31***	***28.53***	***18.37***	***24.02***	***75.98***
SVM	첫 1,041개	3.63 (18/496)	3.55 (19/545)	3.55 (37/1041)	96.45 (1004/1041)	47.56 (594/1249)	32.33 (269/832)	41.47 (863/2081)	58.53 (1218/2081)
	3의 배수	2.36 (14/593)	3.79 (17/448)	2.98 (31/1041)	97.02 (1010/1041)	18.66 (215/1152)	30.46 (283/929)	23.93 (498/2081)	**76.07 (1583/2081)**
	3-fold CV	***2.30***	***5.67***	***3.78***	***96.22***	***21.36***	***32.67***	***26.35***	***73.65***

실험 데이터를 이와 같이 구성한 이유는, (1) 명사에 주목하는 것이 분류 성능에 과연 도움이 되는지를 알아 보고, (2) 시간의 흐름에 따른 데이터 변화에 분류기(classifier)가 얼마나 강건(robust)한지를 파악하기 위함이다. 이를 위해 실험 데이터를 발언 시간에 따라 오름차순으로 정렬한 뒤(early data first), 상기 (1)과 (2)의 조건에 맞게 훈련 데이터와 테스트 데이터를 구분했다. 실험 데이터의 긍정(지지)/부정(비판) 의견의 구성(긍정/부정 데이터 수)을 아래 <표 2>에 명시한다.

<표 2> 실험 데이터 구성 (트윗수)

고유 단어수: 9,509				
훈련 데이터 수		테스트 데이터 수		
고유 단어수: 8,173				
훈련 방법	훈련 데이터 수		테스트 데이터 수	
	긍정(지지)	부정(비판)	긍정(지지)	부정(비판)
첫 1,041개	496	545	1,249	832
	1,041		2,081	
3의 배수	593	448	1,152	929
	1,041		2,081	
3-fold CV	582	459	1,163	918
	1,041		2,081	

<표 2>에서 '첫 1,041 개'는 데이터 대상기간의 가장 초기의 1,041 개의 데이터를 훈련 데이터로 사용했다는 뜻이며, '3 의 배수'는 시간 축으로 오름차순으로 정렬된 데이터의 0,3,6,9…번째 데이터를 골라 훈련 데이터로 사용했다는 뜻이다. 그리고 '3-fold CV'는 랜덤으로 훈련/테스트 데이터를 만들어 사용한 것을 나타낸다. (<표 2>의 긍정/부정 구성은 3 세트 평균.)

4. 실험 결과

'나꼼수 비키니 시위' 트위터 데이터에 NB 와 SVM 을 적용하여 <표 2>의 실험 조건으로 긍정/부정 의견을 자동 분류한 결과를 본 페이지 상단의 <표 3>에 정리한다. 훈련과 테스트 성능을 각각 error pos, error neg, error all (오분류율)과 accuracy(분류 정확도)로 나누어 측정했는데, error pos 는 긍정 의견이 부정으로 오분류된 경우이고, error neg 는 부정 의견이 긍정으로 오분류된 경우이다. 테스트 성능의 경우, 모든 단어를 포함하고(고유 단어수 9,509 개), 대상기간 전반에 걸친 훈련 데이터(3 의 배수)로 SVM 을 적용한 경우가 분류 정확도 76.07%로 가장 높았다(<표 3> 굵은 글씨 & 밑줄).

또, 3-fold CV 의 경우, 모든 단어를 이용하여 NB 분류기를 적용한 경우가, 75.98%로 가장 높았다(<표 3> 기울임꼴 & 굵은 글씨 & 밑줄). 이는 Pang & Lee[12] 의 결과보다 낮은 성능이지만(3-fold CV, NB: 81.0%, SVM: 82.9%. [12] p. 83, Figure 3 (2) 참조, 고유 단어수: 16,165 개), 데이터의 종류와 고유 단어수의 차이를 감안하면, 비슷한 수준의 분류 성능을 보인다고 할 수 있겠다. 한편, 본 실험 결과는 Pang 들과는 반대로, NB 가 SVM 보다 더 나은 성능(3-fold CV 의 경우)을 보였다. 실험 결과를 요약하면 다음과 같다.

- 명사만 사용한 경우(고유 단어수 8,173 개)와 문장의 단어 전체를 사용한 경우(9,509 개)에 성능의 큰 차이는 없었다.
- 대상기간 전반에 걸친 훈련 데이터(3 의 배

수)로 분류기를 생성한 경우가, 대상기간 초반의 데이터(첫 1,041 개)로 분류기를 생성한 경우보다 훨씬 좋은 성능을 보였다(약 60%의 정확도 vs. 약 75%의 정확도).

5. 결론 및 논의

대상기간 전반에 걸친 데이터로 훈련한 경우가 초반의 데이터로만 훈련한 경우보다 좋은 결과를 보인 이유는, 시간이 지남에 따라 토론 내용이 크게 변하면서, 단어의 구성이 변화한 때문으로 보인다. 따라서 의견 자동분류 시스템을 도입하여 장시간 운영할 경우, 초기 데이터로 생성된 분류기를 그대로 장시간 사용하게 되면, 시간이 지나면서 분류성능이 저하될 우려가 있다. 이 같은 현상은 평가 대상의 구체적 항목이 어느 정도 정해진 상품평(review)보다는(예컨대, 맛집의 경우 가격, 맛, 분위기 등), 토론 내용이 점차 변하는 사회적 이슈를 다루는 대중의견(opinion) 에서 더욱 두드러진다. 이를 방지하려면, 주기적으로 분류기를 경신하는 등의 방법으로, 시간의 흐름에 따른 단어 구성의 변화에 대처해야 할 것으로 보인다.

한편, 명사에 주목한 경우와 단어 전체를 이용한 경우에서 성능에 큰 차이가 없었던 것은, 개개 트윗이 짧은 문장으로 이루어졌고, 짧은 문장 안에서 명사가 차지하는 비율이 높아, 명사와 단어 전체간의 차이가 별로 없었기 때문인 것으로 추정된다. 그러나 일부 트윗은 블로그, 외부 트윗(예컨대 TwitLonger[13]), 신문 기사 등 외부로 가는 링크를 다는 것으로 의견을 대신하는 경우도 있어, 좀 더 정확한 의견 분류를 위해서는 이와 같은 외부 컨텐츠를 기존 트윗 문장들과 함께 적절하게 처리할 필요가 있다.

본고에서 우리는 대표적인 기계학습 기법인 NB 와 SVM 을 트위터 대중의견 데이터에 적용하여 이들 기법의 자동분류 성능을 확인했다. 그리고 대중의견 자동분류 시스템을 실용화하는 데 있어 가장 먼저 해결해야 할 과제가, 시간에 따른 토론 내용의 단어 구성의 변화라는 점을 파악했다. 앞으로는 본 실험의 오분류 결과에 대한 분석과, 시간에 따른 단어 구성의 변화에 대한 분석을 실시하고, 분류성능을 향상시키기 위한 기술 개발에 착수할 계획이다.

참고문헌

[1] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” Foundations and Trends in Information Retrieval, 2(1-2), pp. 1–135, 2008.

[2] http://twitter.com

[3] http://ko.wikipedia.org/wiki/나는_꼼수다

[4] http://ko.wikipedia.org/wiki/정봉주

[5] http://ko.wikipedia.org/wiki/서울시의_무상급식_정책_논란

[6] http://gadgetwise.blogs.nytimes.com/2011/07/26/a-better-way-to-search-twitter/

[7] http://code.google.com/p/otterapi/

[8] 최기선 외, 한나눔 한국어 형태소 분석기, http://semanticweb.kaist.ac.kr/home/index.php/HanNanum

[9] http://www.ruby-lang.org/

[10] http://classifier.rubyforge.org/

[11] C. Chang and C. Lin, “LIBSVM: A library for support vector machines,” ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011. (Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm)

[12] B. Pang, L. Lee and S. Vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques,” Proceedings of EMNLP, 2002.

[13] http://twitlonger.com