한국어 트위터의 감정 분석 도구

서형원○, 전길호, 최명길, 남유림, 김재훈
한국해양대학교 컴퓨터공학과
wonn24@gmail.com○, asone7784@naver.com, cmg5478@naver.com,
zin1987@nate.com, jhoon@hhu.ac.kr

A Sentiment Analysis Tool for Korean Twitter

Hyung-Won Seo○, Kil-Ho Jeon, Myung-Gil Choi, Yoo-Rim Nam, Jae-Hoon Kim
Department of Computer Engineering, Korea Maritime University

요 약

본 논문은 자동으로 한글 트위터 메시지(트윗: tweet)에 포함된 감정을 분석하는 방법에 대하여 기술한다. 제안된 시스템에 의하여 수집된 트윗들은 어떤 질의에 대해 긍정 혹은 부정으로 분류된다. 이것은 일반적으로 어떤 상품을 구매하기 원하는 고객이나, 상품에 대한 고객들의 평가를 수집하기 원하는 기업에게 유용하다. 영문 트윗에 대한 연구는 이미 활발하게 진행되고 있지만 한글 트윗, 특히 감정 분류에 대한 연구는 아직 공개된 것이 없다. 수집된 트윗들은 기계 학습(Naive Bayes, Maximum Entropy, 그리고 SVM)을 이용하여 분류하였고 한글 특성에 따라 자질 선택의 기본 단위를 2음절과 3음절로 나누어 실험하였다. 기존의 영어에 대한 연구는 80% 이상의 정확도를 가지는 반면에, 본 실험에서는 60% 정도의 정확도를 얻을 수 있었다.

주제어: 트위터, 기계 학습, 감정 분석, 감정 분류

1. 서론

현재 전 세계적으로 화제를 모으고 있는 트위터는 편리하고 간단한 인터페이스와 기능을 제공하여 뉴스, 미디어, 대중 매체보다 더 쉽고 빠르게 자신이 원하는 정확한 정보를 공유할 수 있게 해준다. 더군다나 모바일 환경에서도 쉽게 트윗을 작성하고 공유할 수 있기 때문에 그 어느 매체보다 더욱 더 빠르게 성장하고 있고 파급력 또한 막대하다.

트위터는 트윗이라고 하는 일종의 메시지(status message)들을 작성할 수 있는데, 실제 트위터에는 다양한 주제에 대한 트윗들이 존재한다. 이것은 어떤 상품을 구매하기 원하는 사람이나, 어떤 상품을 새롭게 발매하는 기업이 기존 비슷한 상품에 대한 사용자들의 상품평을 원할 때 유용한 자원이 될 수 있다. 이를 위해 자동으로 트윗들을 수집하여 감정 분석을 하는 것이 본 논문의 핵심이라고 할 수 있다.

감정 분석은 주로 기계 학습을 이용하여 분석하는데, 이미 다양한 방법과 도메인에 한글에 대한 감정 분석을 한 연구들이 있었다[1]. 하지만 기존의 감정 분석[2,3]은 주로 구단위, 문장 단위, 문서 단위의 감정 분류를 하거나 영화 리뷰처럼 대량의 코퍼스를 대상으로 하는 연구였다. 리뷰와 같은 문서는 일단 그 문서를 구성하는 글의 양이 많고 그것을 작성한 저자의 생각이나 느낌을 자세하게 기록하는 반면에, 트윗은 그 글의 길이가 140 음절로 제한이 되어있기 때문에 많은 내용을 포함할 수 없으며 글의 형식 또한 맞춤법에 크게 얽매이지 않는 편이다.

본 논문은 영어 트윗들을 대상으로 하는 감정 분류 도구[7]를 참고하여, 한글을 대상으로 하는 감정 분석 도구를 제안하고 자세히 기술한다.

1.1 감정 정의

본 논문에서는 기존의 연구[7]와 마찬가지로 감정을 '개인의 긍정 혹은 부정적인 느낌'이라고 정의한다. 예를 들어, 있는 사실을 그대로 알리는 메시지나 뉴스의 헤드라인을 따오는 경우의 대부분은 긍정 혹은 부정적인 느낌을 포함하지 않는다. 따라서 본 논문은 이런 경우를 제외하고 오로지 긍정과 부정의 트윗만을 연구 대상에 포함한다고 정의한다. 그러나 대부분의 트윗은 심지어 감정조차 포함하고 있지 않는 것이 많기 때문에 '중립'이란 클래스를 학습 말뭉치에 포함시킬 수 없는 것이 현재 하나의 연구 제약이다.

1.2 트윗

트위터 메시지는 몇 가지 고유한 특징을 가지는데, 이 절에서는 이전 감정 분류에 대한 연구와 차별되는 것들을 나열한다.

길이: 트위터 메시지의 최대 길이는 140자이다. 영어 트윗의 경우 평균 14개의 단어, 78음절로 구성되었다. 한국어의 경우 평균 12개의 단어, 63개의 음절로 구성되었다.

데이터 가용성: 지난 연구에 따르면 실험 평가를 위해 굉장히 적은 양의 학습 말뭉치를 구축한 반면, Twitter API를 이용하여 수천만 개의 트윗을 쉽고 편하게 수집할 수 있다.

언어 모델: 트윗은 단지 데스크톱에서뿐만 아니라, 여

러 디바이스를 통해 작성될 수 있다. 단적으로, 최근에 스마트 폰이 굉장히 활성화되면서 이제는 어디서나 손쉽게 문자 메시지를 보내듯이 트윗을 작성할 수 있다. 이로 인해 주로 사적일 때 사용하는 상당히 짧으면서도 함축적인 문장, 채팅 용어, 은어 등을 사용하는 경우가 많기 때문에 그 어떤 언어 처리 분야보다 문법적 오류를 많이 포함하고 있다.

2. 관련 연구

2.1 기계학습을 이용한 감정 분석

문서를 긍정 혹은 부정으로 분류하기 위해 가장 일반적인 방법 중 하나는 기계 학습을 사용하는 것이다. 물론 기계 학습에 적용할 여러 가지 알고리즘들이 있지만, 그 중에서도 SVM이 문서 범주화에 가장 좋은 성능을 보였다[2,3]. 학습에 이용할 자질은 단어 혹은 음절 unigram과 bigram, 단어 위치, 그리고 품사 정보 등 다양한 것들이 있다. 영어는 그 중에서 단어 unigram(1어절)이 83%로 가장 높은 성능을 보였다는 연구가 있지만[2,3], 한국어의 경우 형태학적 분석에서 오류를 범할 수 있는 다양한 변칙적 단어들을 많이 포함하기 때문에 이것을 그대로 적용하기 어렵다. 따라서 본 논문에서는 한국어 단어의 80% 이상이 2음절 혹은 3음절로 구성되어 있다[4]는 점과, 이것들이 정보 검색 분야에서 충분히 좋은 자질로 사용될 수 있다는 연구 결과[5,6]에 따라 한글 2음절과 3음절을 자질로 사용하였다.

2.2 이모티콘

본 논문에서는 학습 말뭉치에 포함된 모든 이모티콘을 제외하였다. 이모티콘을 제외하지 않는다면 정확도 측면에 있어서 MaxEnt와 SVM 분류기의 성능이 떨어지고 Naive 분류기에는 약간의 좋은 영향을 미치기 때문이다[7]. 이것은 하나의 부작용을 야기할 수 있다. 만약 실험 말뭉치에 이모티콘이 포함될 경우, 학습 말뭉치에는 이모티콘이 포함되어 있지 않기 때문에 이것들이 성능 향상에 아무런 영향을 끼치지 못한다는 점이다. 이모티콘을 제외시킨 이유는 그것이 완벽하게 하나의 감정을 표현하지 않는 경우가 많기 때문이다. 예를 들어, “어제 무한도전 소지섭 패션 간지 ㅠㅠㅠㅠㅠ” 라는 문장을 보면, ‘ㅠㅠ’ 이것이 부정적인 측면을 뜻하지 않고 긍정적인 표현을 나타내고 있다.

2.3 자질 축소

일반 문서와는 다르게 트위터 메시지는 그것만의 고유한 속성을 가진다. 예를 들어, 사용자의 이름과 해시 태그 앞에는 각각 '@'와 '#'을 붙인다. 기존 연구의 경우[7], 이런 자질과 URL을 간소화 시키고 단어 내 음절의 중복(e.g. huuuuungry)을 제거하여 전체 학습 말뭉치의 크기를 45.85% 줄였다. 하지만 본 시스템의 경우 영어가 아닌 한글을 대상으로 하기 때문에 그대로 적용시키는 것은 무리가 있다. 단지 중복되는 기호의 개수를 제한시켰다.

3. 시스템 흐름도

이 장에서는 전체 시스템 흐름에 대하여 기술한다. 그림1에서 보듯이 웹 페이지에서 질의어를 보내면 Twitter API를 통하여 질의어에 대한 트윗들을 수집한다. 그 다음은 모아진 트윗들을 분류기에 적합한 입력 형태로 바꾸기 위해 자질 추출기를 이용한다. 여기서 각각 2음절과 3음절로 분리된 자질들을 이용하여 분류기의 입력 형태로 바뀌게 된다.

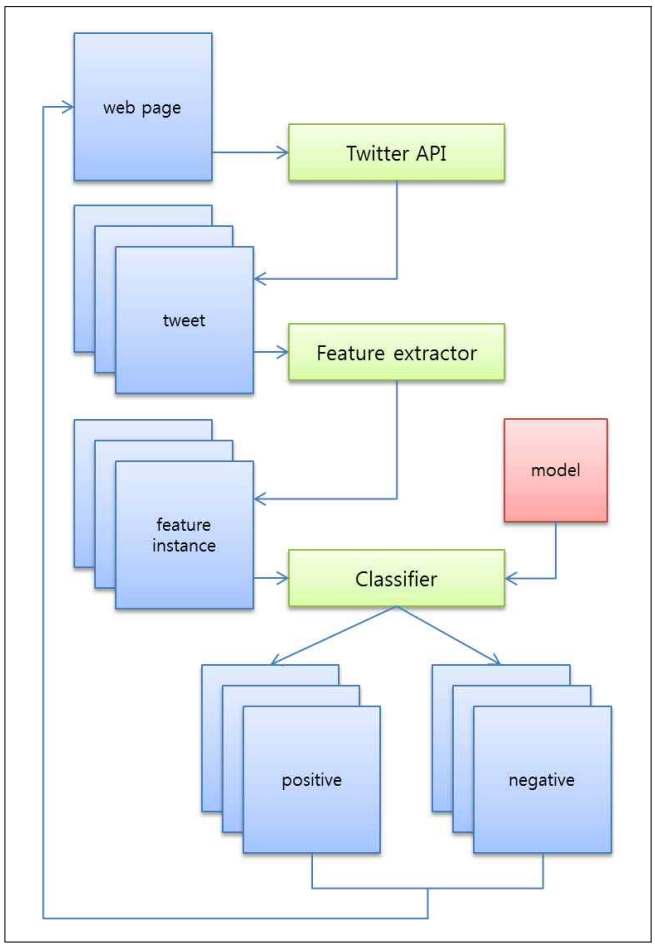

그림 1 시스템 흐름도

분류기에서는 이미 학습된 모델을 이용하여 각 인스턴스가 긍정인지 부적인지 분류하게 된다. 그 결과 값들을 모아서 웹 페이지에 결과로 보이게 된다.

4. 실험 준비

이 장에서는 실험과 평가를 위한 말뭉치 구축에 대해 기술한다.

4.1 말뭉치 수집

감정 분류를 하기 위한 다량의 트위터 메시지는 영어도 마찬가지지만 한글 말뭉치 역시 아직 공개된 것이 없다. 따라서 직접 트위터로부터 말뭉치를 수집해야 한다.

트위터 홈페이지에서는 API를 제공하기 때문에 비교적 수월하게 데이터를 수집할 수 있다. 여기서 제공 받은 트위터 메시지들은 특수한 매개변수들을 제공하는데, 이 중에 언어 설정을 한글로 하여 그에 대응되는 트윗들만 수집하였다.

트위터 API는 여러 제한이 있다. 한 번에 가져올 수 있는 트윗의 양에 제한이 있기 때문에 일정 시간을 두어 수집하였다. 그리하여 약 1주일간 대략 3만개의 트윗을 수집하였다. 수집된 트윗은 총 6개의 카테고리로 나누어지며 그것은 표 1과 같다.

표 1 : 수집된 트윗 카테고리

분류	종류
인물	'아이유', '임재범', '빅뱅', '박지성' 등
회사	'농협', '롯데마트', '애플', '현대' 등
장소	'연평도', '대구', '서울' 등
제품	'갤럭시s2', '아이폰4', '아이패드2' 등
영화	'해리포터', '트랜스포머', '쿵푸팬더' 등
이벤트	'한진중공업 직장폐쇄', '중부 집중호우' 등

말뭉치를 수집하고 난 후, 몇 가지 후처리를 하였는데, 먼저 리트윗의 경우는 중복을 피하기 위해 모두 제거하였다. 또한 이모티콘, 영어 단어 혹은 URL 등 한글을 제외한 문자를 모두 제거시켰다.

4.2 말뭉치 구축

모아진 말뭉치에 라벨을 붙이기 위해 기존의 긍정 부정 단어 리스트[8]를 이용하였다. 이 리스트는 정규표현식을 이용한 3,589개의 긍정 표현과 3,117개의 부정 표현으로 이루어져 있는데, 각 문서 당 감정을 포함하는 단어의 빈도수가 높은 쪽으로 라벨을 붙였고 그 예를 들면 표2와 같다.

표 2 긍정과 부정의 트윗 예

극성	트윗 예
긍정	난 아까 죽음의 성물1봤어.. 성물2 후기 다 좋다고 난리 ㅠㅠ♥
	ㅋㅋ 요즘 정형돈(정재형)이 제일 잘 나가!!!!! 웃겨 죽어 죽어 오늘거 얼릉 보고싶어
	'역시 월드스타' 박지성, 하버드대 여심도 녹였다!
부정	사실 엑페 아크를 지른건 그동안 앱등력에 눌려 억압받던 소니빠력이 폭발한 느낌... (2년 전에 엑페 X1 샀다가 개지뢰 밟았던 기억은 이미 삭제)
	싸이틱에서 비리보고왔는데 괜히짜증나네ㅠㅠ증거도없으면서...
	한진중공업 사태 지겹다. 노사가 합의했다는데 진보정치권에서 군불을 때는 모양이다.

어떤 문서에 긍정을 나타내는 단어가 2개, 부정을 나타내는 단어가 3개 포함되었다면 'negative'라고 라벨을 붙인다. 이것은 학습 말뭉치에 속한 트윗이 모두 한 가지 감정만을 가진다는 기존 연구의 전제와는 상반된 것이지만 자료 부족 문제에 따른 선택이었다. 향후 연구에서는 기존과 같이 한쪽으로만 쏠리는 순수 감정 문서를 대상으로 할 수 있을 것이다. 긍정 단어의 빈도수와 부정 단어의 빈도수가 같을 경우 중립이라고 판단하여 학습 말뭉치에서 제외시켰다. 이렇게 하여, 최종 학습 말뭉치 9,846개(긍정 5,037, 부정 4,809)를 얻었다. 하지만 이것은 사전을 통하여 긍정과 부정이라고 판단된 것일 뿐이지, 실제는 이 중에도 사람이 봤을 때, 중립인 경우가 많이 포함되어 있다. 이것은 향후에 학습 말뭉치를 정제해 나아가는 연구가 필요한 이유이기도 하다.

테스트 말뭉치는 수집한 전체 트윗 중, 학습 말뭉치를 제외한 1,096개(긍정 596, 부정 500개)의 instance를 사람이 직접 분류하였다.

5. 실험 및 평가

5.1 기계 학습

본 논문에서 테스트를 위해 사용한 기계 학습 알고리즘은 Naive Bayes, Maximum Entropy, Support Vector Machines 3가지이다. Naive Bayes와 MaxEnt는 open source python module인 Natural Language Toolkit(NLTK[1])를 사용하였고 SVM은 svm[light][2] (kernel: radial basis function)를 사용하였다.

5.2 실험 결과

트윗이 아닌 일반 영어 문서에 대한 기존 감정 분류[2]는 Naive Bayes, MaxEnt, SVM 분류기를 이용하여 총 16,165개의 자질에 대하여 각각 81.0%, 80.4%, 82.9%의 정확도를 가졌다. 그리고 트윗에 관련한 기존 감정 분류 연구[7]는 총 160만개의 트윗(긍정, 부정 각각 80만 개)에 대하여 앞선 결과와 비슷한 81.3%, 80.5%, 82.2%의 결과를 보였다. 본 시스템의 경우 자질을 2음절과 3음절로 나누어 사용해본 결과, 정확도는 표 3와 같다. 여기서 말하는 정확도는 정답에 비해 시스템이 내놓은 결과가 얼마나 일치하는 지에 대한 확률을 계산한 것이다.

표 3 실험 결과

	NB	MaxEnt	SVM
2음절	60.49%	60.76%	53.38%
3음절	57.29%	57.39%	52.92%

하지만 이것은 학습 말뭉치가 이전 연구의 1/160에 달하는 극도로 적은 양이었기 때문이라고 할 수 있다. 이것은 현재의 한계점이라고 할 수 있으며 향후 보완해야할 문제이기도 하다.

1) http://www.nltk.org/
2) http://svmlight.joachims.org/

6. 결론 및 향후 연구

본 논문에서는 한글 트윗에 포함된 감정을 자동으로 분류하는 시스템을 기술하였다. 기존 연구와는 다르게 한국어에 대한, 한글 특성에 맞는 자질(2음절, 3음절)을 이용하였다. 하지만 자료 부족의 문제로 정확도가 기존 연구에 비해 크게 떨어진다. 이는 향후 보완해야할 문제이며, 이를 위해 비지도학습을 이용하여 좀 더 쉽게 학습 말뭉치를 늘려가는 연구를 진행 중이다. 그 외, 자동으로 수집한 말뭉치에 라벨을 정할 때 이중 부정문이나 긍정과 부정 모두를 포함한 문서가 등장할 경우, 사람이 개입할 수 있는 pooling method를 도입하는 방법도 가능할 것이다. 또한 정확한 형태소 분석이나 띄어쓰기 검사 등이 가능하다면 좀 더 높은 정확도를 얻을 수 있을 것이다.

참고문헌

[1] 이공주, 김재훈, 서형원, 류길수, “뉴스 댓글의 감정 분류를 위한 자질 가중치 설정”, 한국마린엔지니어링학회, 제34권, 제6호, 871-879, 2010.

[2] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up? Sentiment classication using machine learning techniques." In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, Philadelphia, PA, 2002.

[3] 10. Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts." In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), Barcelona, Spain, 2004.

[4] C.-S. Kim, and Y.-B. Kim, "Statistical Information of Korean dictionary to construct an enormous electronic dictionary", Journal of Korean Contents Society, vol. 7, no. 6, pp. 60-68, 2007.

[5] J.-H. Lee, H.-R. Park, H.-J. Park, J.-A. Ahn, and M.-H. Kim, "An effective indexing methods for hangul texts", Proceegings of the Korean Society for Information Management Conference, pp. 11-14, 1995.

[6] C.-Y. Jung, "An indexing method based on the mixed n-gram for Korean information retrieval", Master Thesis in Department of Computer Engineering, Korea Maritime University, 2004.

[7] http://www.stanford.edu/~alecmgo/papers/TwitterDistantSupervision09.pdf

[8] 김재훈 외, “오피니언 마이닝을 위한 유사 감성 표현 인식에 대한 연구”, 한국전자통신연구원, 2011.