http://dx.doi.org/10.3745/KTSDE.2013.2.10.731

Twitter Sentiment Analysis for the Recent Trend Extracted from the Newspaper Article

Gyoung Ho Lee† · Kong Joo Lee††

ABSTRACT

We analyze public opinion via a sentiment analysis of tweets collected by using recent topic keywords extracted from newspaper articles. Newspaper articles collected within a certain period of time are clustered by using K-means algorithm and topic keywords for each cluster are extracted by using term frequency. A sentiment analyzer learned by a machine learning method can classify tweets according to their polarity values. We have an assumption that tweets collected by using these topic keywords deal with the same topics as the newspaper articles mentioned if the tweets and the newspapers are generated around the same time. and we tried to verify the validity of this assumption.

Keywords : Twitter, Sentiment Analysis, Tweet Topic, Clustering

신문기사로부터 추출한 최근동향에 대한 트위터 감성분석

이 경 호† · 이 공 주††

요 약

본 논문은 사회의 최근 동향에 대한 여론의 반응을 관찰하기 위한 방법을 나타낸다. 최근 동향을 나타내는 키워드를 신문기사로부터 추출하고, 추출된 키워드를 이용하여 수집된 트윗의 감성 분석을 통해 최근 동향에 대한 여론을 분석한다. 수집된 신문기사를 k-means알고리즘을 이용하여 군집화하고, 군집내의 단어의 출현 빈도를 이용하여 토픽 키워드를 선정하였다. 각 토픽에 대하여 수집된 트윗은 그 토픽 대한 트윗이라는 가정하에 기계학습 방법을 이용하여 긍/부정을 판별하여 감성을 판단하게 하였다. 그리고 이와 같은 가정에 대한 타당성을 검증해 보았다.

키워드 : 트위터, 감성분석, 트윗 토픽, 군집화

1. 서 론

대중이 관심을 가지거나 대중에게 알려야 하는 정보를 정제된 표현으로 나타낸 글이 신문기사이다. 또한 인터넷 신문기사는 인터넷에 올라온 수많은 정보 중 신뢰할 수 있고 중요하다고 여겨질 수 있는 정보 중 하나이다[1]. 인터넷의 발달로 기사의 생성과 유통이 쉬워지고 독자의 반응을 쉽게 얻을 수 있게 되어 많은 인터넷 신문사가 설립되었고, 사회의 이슈가 되고 있는 다양한 주제에 대하여 빠르게 기사화 하고 배포할 수 있게 되었다. 국내의 경우 인터넷 포탈 네이트(www.nate.com)의 연예면 뉴스 페이지에 하루 평균 약 3000건 정도의 인터넷 기사가 올라온다. 이런 양적인 측면과 신문기사가 가지는 특징을 고려했을 때, 인터넷 신문기사가 최신 동향에 대한 토픽을 포함한다고 생각할 수 있다. 인터넷 신문기사 분석을 통해 이러한 토픽을 추출할 수 있다면, 추출된 토픽이 수많은 인터넷 상의 정보 중에서 대중들이 관심을 가지고 이야기하고 있는 가치가 있는 토픽이 될 것이다.

이전까지의 여론 조사는 주로 오프라인 상에서 직접 대면을 통한 설문이나 대상에게 전화를 걸어 설문 조사하는 방법이 주류였다. 하지만 최근에 트위터의 발달로 일반 대중들이 자신의 생각이나 자신이 접한 이슈에 대하여 자신의 트위터에 올려 여러 사람들과 소통하는 경우가 많아졌다. 많은 사람들이 많은 주제에 대하여 저마다의 생각을 작성하고, 이를 공유하고 토론함으로써 트위터는 이미 기성 언론이나 인터넷 게시판이 가지던 여론 형성 기능과 맞먹는 정도의 여론의 장이 되어가고 있다[2]. 그렇기 때문에 어떠한 토픽에 대하여 대중들의 생각을 알아보기 위한 장소로 트위터가 큰 위력을 발휘할 수 있다.

기존의 트위터 감성 분석 연구의 대부분은 해당 트윗의 주제에 대한 고려 없이 트위터에 나타난 긍, 부정의 감성추출에 중점을 두었다. 이와 다르게 본 논문에서는 트윗에 나타난 주제를 파악하고, 이 주제에 대한 트윗의 감성을 판별하고자 한다. 특정 시기의 신문기사 그룹에서 중요하게 다루어지는 단어들은 그 시기에 많은 사람들이 언급하는 화제가 되는 키워

※ 이 논문은 2013년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(이공분야기초연구, No.2012004132).
† 준 회 원 : 충남대학교 정보통신공학과 박사과정
†† 정 회 원 : 충남대학교 정보통신공학과 부교수
논문접수 : 2013년 4월 22일
수 정 일 : 1차 2013년 5월 25일
심사완료 : 2013년 6월 28일
* Corresponding Author : Kong Joo Lee(kjoolee@cnu.ac.kr)

드일 것이다. 그렇기 때문에 화제가 되는 단어들(토픽 키워드)을 포함하고 있는 트윗이 있다면, 그 트윗은 해당 화제에 대하여 언급하고 있는 트윗일 것이라고 가정하였다. 본 논문에서는 이와 같은 가정과 함께 인터넷 신문기사와 트위터가 가지는 특징을 이용하여 화제가 되는 최근 동향에 대한 일반 대중의 여론을 분석하는 방법을 설명하고 있다.

본 논문에서 수행하는 작업은 다음과 같다. 첫째, 정해진 일자에 등록된 신문기사를 인터넷으로부터 수집한다. 둘째, 수집한 신문기사를 자동으로 분류하고, 분류된 신문기사의 각 분류마다 가지고 있는 토픽 키워드를 추출한다. 셋째, 추출된 토픽을 포함한 트윗 데이터를 수집하고 수집된 트위터의 감성을 분석하여 해당 토픽에 대한 트위터상의 여론을 파악해 본다. 마지막으로 토픽과 트윗의 감성 분석결과 간의 관계에 대하여 분석한다.

2. 관련 연구

문서나 문장에서 어떤 대상에 대한 감성을 찾아내는 것을 감성 분석이라고 한다. 감성 분석 연구는 주로 영화, 제품에 관한 리뷰[3], 블로그[4], 뉴스기사[5] 등에서 활발히 이루어져 왔다. 하지만 이러한 분야의 연구들은 주로 연구의 대상이 되는 글이 많은 줄(large text)로 이루어져 있다. 반면에 본 논문에서 수행하고 있는 트위터 감성 분석은, 트위터에서 140자까지의 짧은 글만 등록할 수 있게 하기 때문에 분석하는 대상이 짧은 길이를 가지고 있고, 짧은 글의 제약사항으로 인한 비속어, 축약, 비아냥 등 분석을 어렵게 만드는 특징을 가지고 있는 분석대상이다.

(Akshat et al., 2012)에서는 2가지 트위터 데이터 집합에 대하여 자신들이 정의한 알고리즘과 여러 자질을 혼합한 기계학습 방법을 이용한 감성분석 실험 결과를 보여주고 있다[6]. 실험에 사용된 데이터 셋은 Stanford Dataset[7]과 Mejaj Dataset[8]이다. 실험에서 사용한 데이터 셋에 포함된 학습 데이터 안의 unigram들의 긍정, 부정 트윗에서 나타난 발생빈도 간의 차를 이용하여 긍정, 부정을 분류하는 1)Baseline과 2)긍정, 부정, "!"문자의 전처리, 3)각 단어들의 어근 추출, 4)불용어 제거, 5)철자 교정, 6)미리 정의된 감성 표현들을 이용하는 감성어휘 자질, 7)명사 제거, 8)unigram의 긍정, 부정에서의 출현 빈도 간의 차의 크기에 따라 긍, 부정의 점수에 차등을 두는 *Popularity Score* 방법을 가지고 각 단계를 조합하는 실험을 수행하였다. Stanford Dataset을 이용한 실험에서 1)Baseline방법은 78.8%의 정확도를 보였고, 각 단계의 모든 조합 결과 87.2%의 정확도를 보였다. Mejaj Dataset에서는 Baseline 방법이 77.1%, 모든 조합에서 88.1%의 정확도를 보였다.

이 논문에서는 또한 여러 자질을 조합하여 감성을 분석한 실험 결과도 나타내고 있다. 트윗에서 추출되는 자질은 자연언어 처리 관련 자질인 f1)unigram, f2)bigram과 트위터와 관련된 자질인 f3)긍정,부정의 HashTag, f4)긍정, 부정의 이모티콘, f5)본문 내에 URL 존재 여부, f6)@마크를 이용한 대상 지칭 여부, f7)본문에서의 "!"문자 출현 여부를 나타내는 자질이 있다. 이러한 자질들을 Stanford Dataset을 이용하여 실험한 결과, 자연언어 처리 관련 자질인 f1과 f2의 조합에서는 85.34%의 정확률을, 트위터 관련 자질인 f3, f4, f5, f6, f7을 조합한 결과 60.12%의 정확률을 보였다. 자연언어 처리 관련 자질과 트위터 관련자질을 결합한 결과 87.64%의 정확도를 보였다. 본 논문에서는 [6]의 논문에서 사용한 트윗의 전처리 방법과 자질의 종류를 차용하여 실험에 사용하였다.

긍정, 부정의 감성을 판별한 (Akshat et al., 2012)와는 다르게, (Apoorv et al., 2011)에서는 긍정, 부정, 중립의 트윗을 판별하는 실험을 수행하였다[9]. 이들은 자신들이 수집한 트윗에 긍정, 부정, 중립의 감성을 할당하였고, 각각 1709개의 트윗을 선정하여 각 감성 트윗의 수적인 균형을 맞추도록 하였다. 이 논문에서는 unigram을 자질로 사용하는 것을 Baseline으로 하고, 자신들이 정의한 11가지 종류의 자질과의 조합을 통한 감성값 분류의 실험 결과를 나타내고 있다. 긍정, 부정의 분류에서 unigram만을 이용한 Baseline실험은 71.35%의 정확도를 보였고, 자질의 모든 조합을 이용할 경우 75.39%의 정확도를 보였다. 긍정, 부정, 중립의 분류에서 Baseline실험의 결과는 56.58%의 정확도를 보였고 모든 자질의 조합의 경우 60.50%의 결과를 나타내었다. 긍정, 부정의 분류가 긍정, 부정, 중립의 분류보다 더 나은 분류 결과를 나타내는 것을 볼 수 있다. [9]의 실험 결과에서 긍정, 부정의 분류 결과가 긍정, 부정, 중립의 분류 결과보다 우수함을 보였기 때문에 본 논문의 실험에서도 긍정, 부정의 감성 분류를 수행하였다.

이전까지의 연구들은 사용자가 감성 분석을 원하는 쿼리를 제공하면 그 쿼리를 포함한 트윗을 수집하고, 수집된 트윗은 해당 쿼리에 대한 감성을 표현하는 트윗이라 가정하고 분석을 수행하였다. 하지만 (Jiang et al., 2011)에서는 쿼리를 포함한 트윗을 수집하고, 감성 분석과정에서 검출된 감성 표현이 해당 쿼리와 의존관계를 가지고 있는지 확인한 후 의존 관계가 있을 경우 감성을 부여하도록 한다[10]. 이 논문의 연구에서는 교사학습 방법인 SVM 알고리즘을 사용하면서, 자질의 분류를 검출 대상에 독립적인 자질과 검출 대상과 비독립적인 자질로 구분하였다. 검출 대상에 독립적인 자질 검출 대상과의 관계를 찾기 위하여 "Maximum Spanning Tree dependency parser"를 사용하였다. 실험 결과, 검출 대상에 독립적인 자질과 비독립적인 자질 모두 사용하였을 경우 85.6%의 정확도를 보였다. 이 논문에서는 트윗에서 나타난 감성 표현이 그 트윗의 구문분석 결과 감성 분석을 수행하려는 대상을 수식하는 경우 해당 감성 표현의 감성값이 대상의 감성 판별 자질로 사용한다. 본 논문은 이와 다르게 트윗을 수집할 때 수집 키워드가 화제성을 가지는 토픽 키워드라면, 해당 트윗에 나타난 감성은 시의성과 화제성 등을 고려했을 때 그 토픽 키워드에 대한 것이라는 가정을 도입하였다

앞서 소개한 연구와 다르게 (Batista & Ratte, 2012)은 다중분류시스템(Multi-Classifier System, MCS)을 사용하여 트윗의 감성분류를 수행하였다[11]. 이 논문에서는 [6]의 실

험에서 사용한 Stanford Dataset을 실험에 활용하였다. 데이터 셋을 학습 데이터와 실험 데이터로 나누고, 학습 데이터에서 unigram, bigram, trigram을 학습 자질로 추출하였다. 추출된 각각의 자질을 이용하여 3개의 naive bayes 분류기를 학습 시키고, 각 분류기의 분류 결과를 수신자 조작 특성(Receiving Operating Characteristics, ROC)공간[12]에 나타내었다. ROC공간에서 나타나는 분류기들의 결과 통합에 우수한 성능을 보이는 반복적 불리언 결합(Iterative Boolean Combination, IBC)[13]을 이용하여 3가지 종류의 자질로 학습한 3개의 naive bayes 분류기의 결과를 통합하였다. 그 결과 unigram, bigram, trigram을 자질로 사용한 분류기의 Area Under Curve(AUC) 값은 각각 0.8330, 0.6861, 0.5484로 unigram을 자질로 선택하였을 때 가장 좋은 성능을 보였다. 하지만 3가지 분류기를 통합한 IBC의 AUC값은 0.8579로 unigram만을 사용한 분류기의 분류 결과보다 우수한 성능을 보여 bigram 자질을 사용한 분류기와 trigram자질을 사용한 분류기를 함께 사용하는 MCS의 유효성을 입증하였다.

앞서 살펴본 연구 결과들은 다양한 자질과 학습 방법을 이용하여 감성분석을 수행하는 방법에 대하여 나타내고 있다. 하지만, 특정 주제 없이 트위터 자체에서 나타난 감성을 분석하거나 사용자가 입력한 검색어에 대한 감성 분석을 수행하는 한계가 있다. 이와 다르게 본 논문에서는 신문기사에서 자동으로 화제가 되는 키워드를 추출하고, 이 키워드를 통해 감성을 분석하여 사용자에게 제공하는 방법에 대하여 연구하고 있다. 이러한 기술을 토픽탐지 및 추적 기술(Topic Detection and Tracking, TDT)연구와 결합하여 정치나 연예, 증시관련 이슈에 대한 일반 대중들의 반응을 트위터를 통해 추적하여 활용할 수 있다. 본 연구는 이러한 연구의 시금석이 될 수 있다.

3. 토픽 키워드 추출

3.1 전체 개요

본 논문에서 제안하는 시스템은 특정 기간에 발생한 사회적 이슈들에 대하여 각 이슈에 대한 일반 대중의 동향을 트위터를 통해 파악하여 사용자에게 제공하는 것이다. 신문기사가 사람들이 알아야 할 것, 알았으면 하는 것을 나타내는 것이기 때문에 본 논문에서는 인터넷 신문기사를 통해 사회적 관심사가 되는 표현들을 추출한다. 트위터는 다양한 사람들이 많은 주제에 대하여 쉽게 이야기하는 공간이므로 일반 대중의 여론과 유사한 흐름이 있을 것이라 판단하였다. 그렇기 때문에 신문기사로부터 추출된 키워드(토픽)들과 이를 통해 수집된 트윗을 이용하여 트윗에 나타난 감성을 분석한다. 감성 분석의 결과는 해당 토픽에 대한 트위터에 나타난 긍정, 부정 여론으로 간주할 수 있다. 이를 그림으로 표현하면 Fig. 1과 같다.

본 논문의 단계별 실험을 위하여 인터넷 포털 네이트에 2013년 2월 5일 올라온 연예면 기사 3,447건의 기사를 수집하여 실험에 활용하였다. 수집된 인터넷 기사를 k-means 알고리즘을 이용하여 주제별로 군집화하고, 각 군집에서 자주 등장하는 표현들을 토픽 키워드로 결정하였다. 트위터 API를 이용하여 각 토픽 키워드를 포함한 트윗을 수집하고, 수집된 트윗의 일부에 긍정, 부정, 감성 없음의 3가지 감성 값을 할당하였다. 이를 SVM 알고리즘을 이용하여 학습하는 학습 데이터로 이용하여 감성 분석 시스템에 활용하였다.

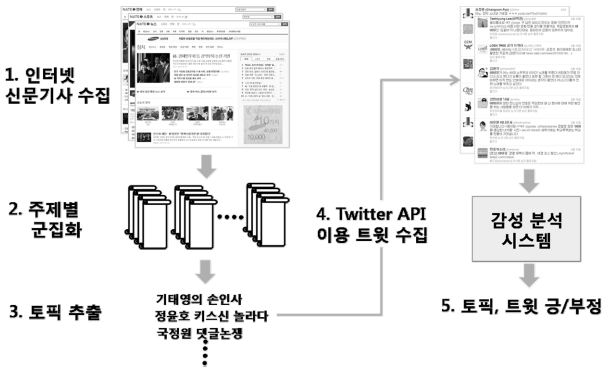


Fig. 1. Overall Process

3.2 자질 추출

인터넷 신문기사는 기사하나에 제목이 하나씩 할당되는 특징이 있다. 신문기사의 제목은 독자들의 주의를 끌고 기사내용을 압축하여 제공하기 위하여 신문기사를 대표할 수 있는 키워드로 구성된다. 그렇기 때문에 본 논문에서는 군집화와 토픽추출에 필요한 자질을 신문기사의 제목을 통하여 추출한다. 신문기사 제목 분석을 위하여 추출하는 자질은 다음과 같다.

1. 정규화한 개체명
2. 기능어를 제외한 실질 형태소

1) 개체명 정규화

각 신문기사의 제목에 대해 개체명 정규화작업을 수행한다. 신문기사 제목이 시스템 입력으로 들어오면 미리 정의된 개체명 사전을 이용하여 신문기사 제목이 가지고 있는 개체명을 정규화된 형태로 바꾸어 준다. 개체명 사전에는 인물(Person) 정보 약 30,000건, 기관 및 단체(ORG) 정보 약 23,000건, 제목(Title) 정보 약 35,000건을 가지고 있다. 사전의 형태는 Table 1와 같다.

2) 형태소 분석

1)의 과정을 거친 신문기사 제목을 형태소 분석기를 통해

Table 1. Example of Named Entity Dictionary

ENTRY	TAG	NORMALIZED
뉴스 데스크	TITLE	뉴스데스크
뉴스데스크		
영진위	ORG	영화진흥위원회
영화진흥위원회		
동방신기	PERSON	동방신기

형태소 분석한다. 분석된 결과 지정사, 어미, 조사, 동사파생접미사, 형용사파생접미사, 기호문자, 글자수가 1개인 단어 등을 제외한 형태소를 자질로 선정하였다. 추출된 형태소 중 “8/시”, “연예/계”와 같이 “숫자+단위성의존명사”, “명사+명사형파생접미사”와 같은 경우에 하나의 자질로 취급하였다. 1)과 2)의 과정을 통해 처리된 신문기사 제목의 예는 Table 2와 같다.

Table 2. Result of Feature extraction

Original
‘뉴스 데스크’ 8시 편성 3개월, MBC 내린 평가는?
Result
<NE tag="TITLE">뉴스데스크</NE> <MA tag="SN">8시</MA> <MA tag="NNG">편성</MA> <MA tag="SN">3개월</MA> <NE tag="ORG">MBC</NE> <MA tag="VV">내리</<MA> <MA tag=“NNG”>평가</MA>

이러한 과정을 통해 추출된 자질은 군집화와 토픽 추출의 입력으로 사용된다.

3.3 군집화

신문기사 군집화 알고리즘으로는 문서 군집화에 대해 보편적으로 사용되고 있는 k-means알고리즘을 이용하였다[14]. 신문기사 군집화를 위하여 각각의 신문기사에서 추출한 자질들을 k-means 알고리즘의 입력 형식에 맞게 바꾸어 주어야 한다. 각 자질의 점수 계산은 문서를 term vector로 표현할 때 일반적으로 사용되는 tf·idf 점수를 사용하였다[15]. 수집된 신문기사에서 추출한 자질을 term vector로 표현하고 이를 군집화와 토픽 추출에 활용한다.

k-means 알고리즘의 경우 데이터를 몇 개의 군집으로 나눌 것인지 결정하는 k값의 결정이 성능에 많은 영향을 줄 수 있다[16]. 일반적으로 군집화 대상이 되는 분야를 관찰하고 군집화의 목적에 맞는 k값을 할당하게 된다. 본 논문에서 사용할 k값을 결정하기 위하여 군집실험을 수행하기 전 7일간의 인터넷 신문기사를 수집하고, 수집된 신문기사의 제목을 가장 작은 주제를 가지는 단위로 분류 하였다. 분석결과, 하나의 주제에 대해 주로 3~4건의 기사가 작성되는 것을 볼 수 있었다. 이에 따라 본 논문의 실험에 사용되는 신문기사의 개수를 고려하여 k값을 1000으로 결정하여 군집화를 수행하였다. 각 군집의 크기를 작게 하여 신문기사 군집으로부터 좀 더 명확한 토픽 키워드를 추출하기 위하여 k값을 크게 잡았으며, 실용적인 측면과 더 정확한 토픽 키워드 추출을 위한 k값의 결정 또는 추출 방법론에 대한 추가적인 연구가 필요하다.

군집화 수행 결과 군집이 가지고 있는 신문기사가 1개인 경우, 화제성이 없는 신문기사 군집으로 판단하고 군집을 제거하였다. 이를 통해 총 273개의 군집화 결과를 얻었다. 군집화 결과의 일부가 Table 3과 같다.

Table 3. Clustering result

	Titles of newspapers
1	- '무릎팍도사' 스타강사 김미경, 강호동과'토크의 달인' 겨룬다 - 스타강사 김미경, '무릎팍도사' 출연… 멘토의 고민은 - 김미경, '무릎팍도사' 출연…'강호동과 입담대결' - 스타강사 김미경, 사람끄는 인기비결 무엇
2	- '커밍아웃' 홍석천" 여자친구와 1년동안 뽀뽀NO, 슬슬 떠나더라" - '힐링' 홍석천 "커밍아웃 행복하게 살고 싶었다" - 홍석천" 커밍아웃 왜 행복하고 싶었다" - 홍석천 커밍아웃 후회, "내가 사랑하는 사람도 동성애자로 낙인"
3	- '청엘' 최성준 브라우니 안고 '훈훈한 자태' - '청엘' 최성준, 브라우니와 인증샷 "유명인사 만났다" - 최성준, 브라우니와 인증샷 "순수남 매력 폭발"
4	- '토크클럽 배우들' 4%대 고전 - '토크클럽 배우들' 신성일, '삼백다이어트' 비법공개 - '토크클럽 배우들' 신성일, 예명 비화 공개 "뉴스타 넘버원" - 선우일란 안소영… 파격 고백에도 '토크클럽 배우들' 시청률은

3.4 토픽 키워드 추출

본 논문에서는 각 신문기사 군집 내에서 단어의 출현 빈도를 이용하여 토픽키워드를 선정하였다. 군집화 결과 신문기사 개수가 2개인 군집의 경우 출현 빈도가 가장 높은 하나의 단어를 토픽 키워드로 선정하였고, 군집 내의 기사 개수가 3개인 경우 출현 빈도 상위 3개의 단어를, 신문기사 4개 이상인 경우 상위 5개의 단어를 군집의 토픽 키워드로 선정하였다. 같은 출현 빈도의 경우 개체명과 명사를 우선순위에 두었다. 각 군집에서 추출된 토픽 키워드의 결과는 Table 4와 같다.

Table 4. Examples of extracted Topics

Topic	Clusters
스타강사 김미경 강호동 무릎팍도사 인기비결	- '무릎팍도사' 스타강사 김미경, 강호동과' 토크의 달인 겨룬다 - 스타강사 김미경, '무릎팍도사' 출연… 멘토의 고민은 - 김미경, '무릎팍도사' 출연…'강호동과 입담대결' - 스타강사 김미경, 사람끄는 인기비결 무엇
홍석천 커밍아웃 힐링캠프 동성애자 행복	- '커밍아웃' 홍석천" 여자친구와 1년동안 뽀뽀NO, 슬슬 떠나더라" - '힐링' 홍석천 "커밍아웃 행복하게 살고 싶었다" - 홍석천" 커밍아웃 왜 행복하고 싶었다" - 홍석천 커밍아웃 후회, "내가 사랑하는 사람도 동성애자로 낙인"
광고천재 이태백 시청률	- '광고천재 이태백' 불안한 출발…첫 회 시청률 4.3% - '광고천재 이태백' 첫회, 시청률 4.3% 출발 - '광고천재 이태백', 첫 회 시청률 4.3%…' 마의' 자체 최고 경신

추출한 토픽 키워드가 신문기사 군집에 대한 주제 키워드로서 적합한지에 대하여 평가하였다. 평가를 위하여 2명의 평가자로 하여금 각 군집의 신문 기사 제목들을 보고, 이 군집을 대표할 수 있다고 생각되는 5개의 단어를 선정하도록 하였다. 자동 추출된 키워드와 평가자가 수동으로 작성한 키워드들 간의 상관관계를 평가하기 위해 정보검색분야

에서 평가 척도 중 하나로 사용되는 R-Precision을 통해 평가하였다[17]. 검색을 통해 얻어진 문서 집합을 R, 검색을 통해 나온 문서 중 찾고자 했던 적합한 문서의 집합을 r이라고 한다면 R-Precision은 |r|/|R|이다. 이러한 개념을 토픽 키워드 평가에 적용하였다. 추출된 키워드를 검색 결과 문서 R로 간주하고, 평가자에 의해 선정된 키워드를 전체 적합 문서 집합으로 간주한다. 이때 자동으로 추출된 키워드와 평가자에 의해 선정된 키워드의 교집합이 r이 된다. 군집으로부터 자동으로 추출되는 토픽 키워드의 경우 길이가 가변적이기 때문에 이러한 R값의 크기의 변화를 반영할 수 있는 평가척도인 R-Precision을 사용하였다. 평가자1, 2가 선정한 각 군집의 키워드와 자동으로 추출된 키워드 간의 R-Precision 평균을 Table 5에 나타내었다.

Table 5. R-Precision of the extracted topic words

	Tester 1	Tester 2
R-Precision	0.866	0.908

4. 트위터 감성 분석

4.1 트위터 수집

3.3절에서 구성한 군집 중 신문기사 군집화의 적절성, 군집의 주제에 대한 트윗 사용자들의 긍, 부정에 대한 편파성, 주제의 시의성 등을 고려하여 31개의 군집을 선정하고, 각 군집의 토픽 키워드로부터 수집된 트윗 중 5,122개의 트윗에 대하여 *긍정(Positive), 부정(Negative), 감성 없음(Neutral)*의 값을 할당하였다. 트윗이 언급한 내용에 아무런 감성이 나타나 있지 않거나 단순한 신문기사의 링크 또는 기계에 의해 작성되어 감성분석의 의미가 없는 경우 “감성 없음”으로 분류한다. 수집된 트윗의 극성값의 분포와 각각의 예를 Table 6와 Table 7에 나타내었다.

4.2 감성 없음 판별

Table 6에서 극성값이 할당된 비율을 보면 감성 없음으로 분류된 트윗의 비율이 높다는 것을 알 수 있다. 본 연구에서는 트윗의 감성 분석을 두 단계에 거쳐서 수행한다. 우선, 트윗에 나타난 감성이 존재하는지의 여부를 판별하고, 감성이 존재하는 트윗에 대해서만 다시 긍/부정의 판별을 수행한다. 이렇게 하기 위해서는 감성이 없는 트윗을 우선 적절하게 제거시키는 방법이 필요하다. 평가자에 의해 생성된 트윗들을 분석해 본 결과 감성 없음 트윗이 공통으로 가지는 몇 가지 규칙을 찾을 수 있었다. 그 규칙은 Table 8과 같다.

Table 6. Collected tweets

polarity	size	ratio
Positive	1277	24.9%
Negative	803	15.6%
Neutral	3042	59.3%
total	5,122	

Table 8. Example of neutral Tweets

1. 신문기사 제목 또는 기사의 첫줄의 일부 + 신문기사 URL로만 구성된 트윗
ex) - *홍석천 잇따른 방송 출연, TV 말하려는 건? URL*
2. [SS영상], [포토], [인터뷰]와 같이 [단어]로 시작하는 트윗
- *[포토] 씨스타19 보라, 이기적 각선미 (쇼!음악중심) URL*
3. 포털 또는 트위터 관련업체에서 제공하는 실시간 검색어 순위
- *네이버 실시간 검색어 1)한서대학교 URL / 2)대구한의대 / 3)백석대학교 / 4)충남대학교 통합정보시스템 / 5) 가천대 / 6)홍석천 / 7)울산대학교 / 8)인디애나 시카고 / 9)고두림 / 10)광고천재 이태백*
4. 스팸키워드를 포함하는 경우
- *tanganillasv고객서비스1위┼성인용품1위쇼핑몰3바나나몰┬URL ┌'이웃집 꽃미남' 윤시윤, 애절한 '이마 손'으로 女心사로잡았다*
5. 유튜브, 음악 스트리밍 서비스 업체 등에서 동영상 및 음원 재생시 자동으로 등록되는 트윗
- *super soothing. listen to this --> [이웃집 꽃미남 OST] 박신혜 (Park Shin Hye) - 새까맣게 (Pitch-Black) MV: URL via TAG_ID*
6. 실제 사용자가 작성하였지만 대상에 대한 긍, 부정의 감성을 나타내지 않은 경우

Table 7. Example of Collected tweets

Polarity	Topic	tweets
Positive	이웃집 꽃미남 박신혜 윤시윤 러브라인	- 이웃집꽃미남 재미짐 ㅋ 어제 우연히보고 급꽂혀서 본방사수중ㅋㅋ 깨금이가 얼른 독미 위로해주길...
	신보라	- 앜ㅋㅋ종캌ㅋㅋㅋㅋ신보라너무귀여웤ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ - 신보라 왜 개그맨해여.. 가수해..
Negative	광고천재 이태백 시청률	- 지금 저 이태백 어쩌고 하는 드라마도 봐라. 첫 회인데 광고 계열 용어 딱 한 두개 나왔다가 지금 또 진구랑 한채영 연애 썰이나 풀고 앉아있잖아.
	홍석천 커밍아웃 힐링캠프 행복 동성애자	- 홍석천 힐링캠프는 생각보다 별로....
Neutral	씨스타	- 갑자기 지난여름 씨스타 러빙유 춤 연습했던거 생각난다.
	이웃집 꽃미남 박신혜 윤시윤 러브라인	- 이웃집 꽃미남 첨부터 보니까. 익숙한 자국이 보여.;; 박신혜도 비염.ㅎㅎ
	광고천재 이태백 시청률	- 1)이제석 광고 URL / 2)한채영 아역 / 3)어르신 폐지 / 4)싸이 슈퍼볼광고 / 5)정유경 / 6)이동욱 감사패 / 7)갤럭시Q / 8)고두림 해명 / 9)최성준 / 10)백제예대

각각의 감성 없음 타입에 대하여 검출 패턴을 작성하고, 이를 통해 “감성 없음” 트윗의 검출 결과를 분석하였다.

2,080개의 감성을 가지고 있는 트윗과 3,042개의 감성 없음 트윗을 분석하였다. Table 8을 규칙을 적용한 결과는 Table 9과 같다.

Table 9. Accuracy of Detection of neutral Tweets

		Experimental result	
		neutral	not neutral
Correct actual	neutral	2733	309
	not neutral	274	1806
precision		0.909	
recall		0.898	

본 논문의 분석 대상이 연예 분야이고 수집시기에 일반 대중에게 화제가 되는 키워드로 트윗을 수집하였기 때문에 홍보성 신문기사나 실시간 검색어 등의 트윗이 많은 것으로 보인다. 이런 트윗들은 신문기사 DB안의 정보와 정규식 등을 조합하여 검출할 수 있다. 하지만 실제 사용자가 작성하였지만 감성을 나타내지 않은 경우에는 검출에 어려움이 있다. (Table 9)의 결과 중 대부분의 오류는 이와 같이 사용자가 직접 작성한 트윗 중에 감성이 나타나지 않은 트윗에서 발생하였다. 이에 대한 검출을 위한 방법과 자질의 연구가 더 필요한 부분이다.

4.3 트위터 분석 단계

수집된 트윗의 감성을 분석하는 단계는 다음과 같다.

1) 전처리 과정

트위터에는 RT, @, #등 트위터에서 사용되는 기능을 위한 태그들이 존재한다. 또한 사용자들이 외부의 사진이나 신문기사들을 전파하기 위하여 본문 내용에 URL이 들어 있는 경우가 있다. 이러한 태그들을 TAG_RT, TAG_ID, TAG_HASH, TAG_URL로 정규화 시켜준다.

2) 자질 추출

전처리된 트윗에서 학습에 사용할 자질을 추출한다. 본 논문에서는 3.2절에서 제안한 방법과 동일한 방법을 이용하여 추출한 unigram 자질과 bigram 자질, 트위터 관련 자질을 실험에 사용하였다. 이러한 자질들은 [6]의 실험에서 사용한 자질을 응용하여 설계하였다. 트위터 관련 자질 중 긍정표현 출현(Positive expression), 부정표현 출현(Negative expression)은 트윗 본문에 긍정의 감성을 나타내는 이모티콘이나 단모음(^^, ㅋ, ㅎ, 기타) 또는 부정의 감성을 나타내는 이모티콘이나 단모음(-_-, ㅜㅜ, ㅠㅠ, 기타)이 나타나는지 여부에 대한 자질이다. 자질의 종류와 설명은 Table 10과 같다.

3-1) 학습 단계

트윗으로부터 추출된 자질 벡터를 SVM 알고리즘에 적용하여 학습시킨다. 해당 트윗에 대한 극성값이 할당되어 있는 학습 데이터를 SVM 알고리즘에 적용하여 학습 모델을 생성한다.

Table. 10. Feature list

Natural Language	Unigram(f1)	Unigram
	Bigram(f2)	Bigram
Twitter	URL(f3)	Presence of URLs
	HASH TAG(f4)	Presence of HASH Tags
	TAGET(f5)	Presence of "@" Tag
	Positive expression(f6)	Presence of Positive expression
	Negative expression(f7)	Presence of Negative expression
	!(f8)	Presence of "!"

3-2) 예측 단계

학습 단계와는 반대로, 트윗의 극성값을 예측해야 하는 경우이다. 해당 트윗의 자질 벡터를 학습 단계로부터 생성된 학습 모델에 적용한다. 그 결과 학습모델로부터 예측값을 받게 된다. 이 값이 해당 트윗의 극성값으로 결정된다.

4.4 트윗 긍정, 부정 분류

실제 감성값 판별이 필요한 새로운 트윗 데이터가 입력되었을 때 극성값을 판별하는 방법은 1) 트윗의 감성 포함 여부 평가, 2) 감성 포함 트윗의 긍정, 부정 분류 단계를 거치게 된다.

평가자에 의해 극성값이 할당된 트윗 중 감성 없음 트윗을 제외한 2,080개의 트윗(긍정 1,277개, 부정 803개)을 이용하여 감성값 할당 실험을 수행하였다. 감성 분석 대상인 연예분야의 트윗을 분석 결과 긍정의 트윗이 부정의 트윗보다 더 많이 나타나는 특징을 발견하였다. 실험에 사용한 데이터는 이러한 편향성을 반영하고 있다. Table 10 자질의 감성 분류 성능을 보기 위하여 각 자질들의 조합에 대하여 실험을 수행하였으며, 10-fold cross validation을 통해 평가하였다. 각 자질 조합의 실험 결과를 Table. 11에 나타내었다.

모든 트윗을 긍정으로 분류하는 분류기를 baseline으로 정하고 실험하였을 때 분류기의 분류정확도 62.8%, 자질로 unigram만을 이용한 실험에서는 72.64%의 분류 정확도를 보였다. unigram과 bigram을 함께 사용한 결과 73.29%의

Table 11. Result of Sentiment analysis

Feature List	Accuracy
baseline	62.8%
unigram(f1)	72.64%
unigram(f1)+bigram(f2)	73.29%
f3+f4+f5+f6+f7+f8 (TWF)	60.92%
f1+f2+TWF	74.46%

성능을 보였다. 트위터 관련 자질인 f3, f4, f5, f6, f7, f8(TWF)만을 사용한 경우 60.92%의 결과를 보였고 unigram과 bigram, 트위터 관련 자질을 모두 사용한 결과 74.46%의 분류 정확도를 보였다.

5. 토픽과 트위터와의 관계

본 논문에서는 다음과 같은 가정을 도입하였다.

1. 화제성을 가진 키워드로 수집된 트윗은 시기와 키워드의 중요성을 고려할 때 그 키워드에 관련된 이슈에 대해 언급하고 있을 것이다.
2. 그렇기 때문에 해당 트윗이 감성을 나타내고 있다면 그 감성의 대상이 토픽 키워드일 것이다.

이 가정의 타당성을 검증하기 위하여 감성분석 실험에 사용된 트윗을 이용하여 토픽 키워드와 트윗간의 관계를 평가하였다. 평가 방법은 토픽 키워드와 그로부터 추출된 트윗에 대하여 1)트윗이 토픽 키워드에 대하여 언급하고 있는지, 2)감성분석 결과가 토픽 키워드에 대한 감성과 일치하는지에 대하여 평가하였다.

토픽 키워드"베를린"에 대하여 *"류승완 감독 '베를린' 잘 만들었네. 리얼한 액션 죽인다~"*와 같은 경우 이 트윗은 *"베를린"* 이라는 토픽에 대하여 언급하고 있으며, 토픽에 대하여 긍정의 감성을 가지고 있다. 반면 *"베를린 후기2: 전지현 머릿결은 진짜 최고다."*같은 경우 *"베를린"*에 대하여 언급하고 있지 않지만 긍정의 감성을 나타내고 있다. 또한 토픽 "최고다 이순신 웃기다 이지훈"에 대하여 *"최고다 이순신 진짜 기대된다~"*같이 키워드 전체가 포함되지는 않지만 의미적으로 토픽의 넓은 의미의 주제에 대하여 관계가 있을 경우도 토픽에 대하여 말하고 있다고 간주한다. 평가 결과를 Table 12에 나타내었다.

Table 12. To pic, tweet, sentiment analysis relation

topic-tweet relation	tweet-senti relation	topic-tweet-senti relation
88.42%	74.46%	65.97%

Table 12의 topic-tweet relation은 트윗의 주제가 토픽 키워드이며 해당 트윗에 나타난 감성이 토픽과 관련된 경우의 비율을 나타낸다. tweet-senti relation은 트윗에서 토픽을 고려하지 않는 감성 분석의 결과로 Table 11의 실험 중에서 가장 좋은 결과를 나타낸다. topic-tweet-senti relation은 트윗의 주제가 토픽이며, 감성분석 실험 결과의 결과와 실제 트윗에서 나타난 토픽에 대한 감성이 일치하는 비율을 나타낸 것이다. 분석에 사용된 트윗 중 토픽 키워드와 관련이 있는 트윗의 경우가 88.42%로 나타났다. 트윗과 토픽간의 관계를 고려하지 않은 경우 감성 분석의 정확도는 74.46%를 보였다. 트윗이 토픽에 대하여 말하고 있으며, 감성 분석 결과가 트윗에서 나타난 토픽에 대한 감성과 일치하는 경우가 65.97%로 나타났다.

위 실험 결과에서 토픽 키워드로 수집된 트윗 중 그 토픽에 대하여 말하고 있는 트윗의 비율이 88.42%로 나타나 앞서 가정하였던 토픽 키워드와 그로부터 수집된 트윗 간의 관계에 대한 가정의 타당성을 보이고 있다. 이는 트윗의 감성 분석을 수행 할 때 트윗이 화제성이 있는 키워드로 수집된 트윗이라면 구문분석단계 등을 거치지 않아도 해당 트윗의 감성을 주제 키워드에 대한 것으로 간주할 수 있다는 결론을 이끌어 낼 수 있다.

트윗의 주제를 생각하지 않고 긍, 부정을 분석할 때는 74.46%의 결과를 나타내었다. 하지만 해당 토픽의 주제를 고려하였을 때 감성 분석의 정확도는 65.97%까지 떨어진다. 이는 앞으로 더 적절한 토픽 키워드 추출과 트윗의 감성 분석 성능을 높이면서 올려야 할 과제가 된다.

6. 결 론

본 논문에서는 신문기사의 모음으로부터 사회 이슈를 찾아내고, 그 이슈에 대한 트위터상의 감성을 파악할 수 있는 시스템에 대하여 설명하였다. 그리고 실험을 통해 그 가능성을 보였다. 이를 통해 해당 주제에 대한 대중의 긍정, 부정의 감성을 파악할 수 있게 되어 주제에 대한 여론의 향방을 살펴 볼 수 있게 된다. 이러한 기술과 토픽탐지 및 추적 기술을 결합하여 정치, 연예, 증시관련 이슈의 변화와 그에 따른 대중들의 반응을 추적하여 활용할 수 있을 것이다.

본 논문에서 군집화에 사용한 k값 결정 방법 이외에도 군집화 결과를 평가하여 자동으로 k값을 할당하는 방법같이 k값을 결정하는 다른 방법이 존재한다. 시스템의 사용성과 도메인 특성에 맞는 적절한 k값을 선정하는 것은 어려운 일이고 앞으로 계속 연구해 나가야 할 내용이다. 군집으로부터 추출한 토픽의 경우 기본적인 단어의 출현 빈도만을 추출하였기 때문에 문맥정보 등을 고려한다면 더 나은 토픽 키워드를 추출할 수 있을 것으로 보인다.

본 논문에서 가정한 화제성과 트윗의 감성분석 결과에 대한 상관관계 실험에서 65.97%의 결과를 보였다. 이는 앞으로 트위터 감성분석의 성능을 높이고, 신문기사 군집화와 토픽 추출의 성능을 향상시킴으로써 높여나갈 수 있으리라 생각하고 계속 연구해야 할 주제이다.

본 논문에서 수행한 일련의 작업은 각 단계마다 다양한 접근방법과 연구주제를 가지고 있다. 신문기사 군집화와 토픽 추출, 트위터 분석 등 다양한 연구 분야와 주제가 혼합되어 있어 각각의 주제에 대하여 심화된 연구를 통해 본 시스템의 효용성을 높일 수 있을 것이라 판단하고 이는 향후 연구과제로 남아 있다.

참 고 문 헌

[1] J. Dimmick., Y. Chen, and Z. Li, "Competition between the Internet and traditional news media: The gratification-

opportunities niche dimension," in *The Journal of Media Economic*, 17.1, pp.19-33, 2004.

[2] D. A. Shamma, L. Kennedy, and E. F. Churchill, "Tweet the debates: understanding community annotation of uncollected sources," in *Proceedings of the first SIGMM workshop on Social media*. ACM, pp.3-10, 2009.

[3] Turney and D. Peter, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics*, pp.417-424, 2002.

[4] G. Dray, M. Plantié, A. Harb, P. Poncelet, M. Roche, and F. Trousset, "Opinion mining from blogs," in *IJCISIM'09: International Journal of Computer Information Systems and Industrial Management Applications*, 1, pp.205-213, 2009.

[5] N. Godbole, M. Srinivasaiah, and S. Skiena, "Large-scale sentiment analysis for news and blogs," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, Vol.2, 2007.

[6] A. Bakliwal, P. Arora, S. Madhappan, N. Kapre, M. Singh, and V. Varma, "Mining sentiments from Tweets," in *WASSA 2012*, pp.11-18, 2012.

[7] A. Go, B. Richa, and H. Lei. "Twitter sentiment classification using distant supervision," in *CS224N Project Report, Stanford*, pp.1-12, 2009.

[8] N. N. Bora, "Summarizing Public Opinions in Tweets," in *Journal Proceedings of CICLing*, 2012.

[9] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in *Proceedings of the Workshop on Languages in Social Media. Association for Computational Linguistics*, pp.30-38, 2011.

[10] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent twitter sentiment classification," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Vol.1, pp.151-160, 2011.

[11] L. B. Batista, and S. Ratte, "A Multi-Classifier System for Sentiment Analysis and Opinion Mining," in *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining*. pp.96-100, 2012.

[12] R. C. Prati, G. E. A. P. A. Batista, and M. C. Monard, "A survey on graphical methods for classification predictive performance evaluation," in *Knowledge and Data Engineering, IEEE Transactions,* pp.1601-1618, 2011.

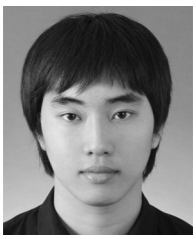
[13] W. Khreich, E. Granger, A. Miri, and R. Sabourin, "Iterative Boolean combination of classifiers in the ROC space: An application to anomaly detection with HMMs," in *Pattern Recognition*, 43(8), pp.2732-2752, 2010.

[14] J. Li, and K. Zhang, "Keyword extraction based on tf/idf for Chinese news document," in *Wuhan University Journal of Natural Sciences*, pp.917-921, 2007.

[15] M. J. Pazzani, J. Muramatsu, and D. Billsus, "Syskill & Webert: Identifying interesting web sites," in *Proceedings of the national conference on artificial intelligence*. pp.54-61, 1996.

[16] P. S. Bradley, and U. M. Fayyad, "Refining initial points for k-means clustering," in *Proceedings of the fifteenth international conference on machine learning*. Vol.66, pp.91-99, 1998.

[17] C. D. Manning, P. Raghavan, and H. Schütze, "*Introduction to information retrieval*," Vol.1. Cambridge: Cambridge University Press, ch8, pp.148, 2008.



이 경 호

e-mail : lee6boy@empal.com

2011년 충남대학교 정보통신공학과 (학사)

2013년 충남대학교 정보통신공학과 (석사)

2013년~현 재 충남대학교 정보통신공학과 박사과정

관심분야 : opinion mining & social data

이 공 주

e-mail : kjoolee@cnu.ac.kr

1992년 서강대학교 전자계산학과 (학사)

1994년 한국과학기술원 전산학과(공학석사)

1998년 한국과학기술원 전산학과(공학박사)

1998년~2003년 한국마이크로소프트(유) 연구원

2003년 이화여자대학교 컴퓨터학과 대우전임강사

2004년 경인여자대학 전산정보과 전임강사

2005년~현 재 충남대학교 정보통신공학과 부교수

관심분야 : 자연언어처리, 기계번역, 정보검색, 정보추출