2015년 8월
석사학위 논문

SNS 해시태그를 이용한 사용자 감정 분류 방법에 관한 연구

조선대학교 산업기술융합대학원

소프트웨어융합공학과

남 민 지

SNS 해시태그를 이용한 사용자 감정 분류 방법에 관한 연구

A Study on Classification Method of User Sentiment using SNS Hashtags

2015년 8월 25일

조선대학교 산업기술융합대학원

소프트웨어융합공학과

남 민 지

SNS 해시태그를 이용한 사용자 감정 분류 방법에 관한 연구

지도교수 신 주 현

이 논문을 공학석사학위신청 논문으로 제출함.

2015년 4월

조선대학교 산업기술융합대학원

소프트웨어융합공학과

남 민 지

남민지의 석사학위논문을 인준함

위원장 조선대학교 교수 金判九 (인)

위 원 조선대학교 교수 반성범 (인)

위 원 조선대학교 교수 신주현 (인)

2015년 5월

조선대학교 산업기술융합대학원

목 차

표 목 차

그 림 목 차

ABSTRACT

A Study on Classification Method of User Sentiment using SNS Hashtags

Minji Nam
Advisor : Prof. JuHyun Shin, Ph.D
Department of SoftWare Convergence Engineering
Graduate School of Chosun University

Recently, studies are being actively carried out for sentiment analysis, which is a type of natural language processing technologies for analyzing subjective data such as opinions, attitudes, and propensities of users expressed on Web, blogs, and social network services (SNSs). Conventionally, to classify the sentiments of texts, most studies were carried out in a form of determining positive/negative/neutral sentiment by assigning polarity values for sentiment vocabulary by using a sentiment Lexicon. However, in this study, the sentiments were classified on the basis of Thayer's model defined psychologically, unlike the polarity classification used in the opinion mining.

In this paper, as a method of classifying the sentiments, sentiment categories were proposed by extracting sentiment keywords for major sentiments by using hashtags, which are essential elements of Instagram. By applying this to user posts, sentiments can be determined through similarity measurement between the sentiment adjective candidates and sentiment category's sentiment keywords. Also, since sentiment distribution and major sentiments of users can be determined, the ambiguity of subject sentiments can be solved objectively.

As a test result for the proposed method, the average accuracy rate for the whole sentiment categories was 90.7%, showing a good performance.

In this study, the sentiment categories were proposed by selecting happy, angry, peaceful, and sad as categories as typical sentiments to improve the accuracy of sentiment classification and minimize the misclassification rate; but in the future, if classifiable sentiments are additionally selected and expended on top of the four sentiments, it is thought that user sentiments can be analyzed more specifically. Furthermore, this study was carried out only with sentiment adjectives extracted by using texts, but it should be extended to a study including a method of using emoticons or other parts of speech containing sentiments.

Through the proposed method, if a sentiment classification system of large capacity is prepared, it is expected that sentiment analysis will be possible for various fields such as major issues and social phenomena through SNS, and furthermore, it is expected to be used in user tailored services, recommendation services, or sentiment marketing on SNS.

Ⅰ. 서론

A. 연구 배경 및 목적

최근 소셜 네트워크 서비스(Social Network Service, SNS)가 스마트폰 사용과 더불어 사용자들의 생활 속 일부분으로 자리 잡으면서 다양한 SNS들이 등장하고 있다. 시대 흐름에 따른 SNS 등장을 살펴보자면, 먼저 1세대 SNS는 기존에 형성된 오프라인 인맥을 중심으로 온라인에서 소통하는 방식인 싸이월드(Cyworld)의 미니홈피나 블로그와 같은 형태였다. 2세대 SNS에서는 스마트 폰의 등장과 함께 트위터(Twitter)나 페이스북(Facebook)처럼 타임라인 형식으로 단문형태의 메시지를 실시간으로 주고받는 서비스가 인기를 끌고 있다. 이에 다양한 SNS들이 등장하면서 인맥 중심의 관계와 포괄적인 커뮤니케이션을 이어나가는 기존의 SNS에서 점차 관심사나 취미 등 특정한 주제를 중심으로 공유하는 사용자 맞춤형 서비스인 텀블러(Tumblr), 핀터레스트(Pinterest), 인스타그램(Instagram)과 같은 3세대 SNS가 떠오르고 있다[1,2]. 이러한 SNS들 중에서 사용자가 게시한 게시물 등을 이용하여 관심사나 감정을 분석하는 연구가 활발히 진행되고 있지만[3,12,13,14,15,16,17], 기존의 보편화된 SNS를 활용한 연구가 대부분이다.

따라서 본 연구에서는 3세대 SNS 중 하나인 인스타그램을 대상으로 하였다. 인스타그램에서는 사용자가 공유하고 싶은 이미지를 스마트 디바이스를 통해 촬영하거나 기존에 촬영한 이미지를 업로드 할 때 간단한 해시태그(Hashtag)를 덧붙여 글을 게시함으로써 자신의 감정을 표현하고 다른 사용자들과 교류한다.

본 연구에서는 인스타그램을 대상으로 오피니언 마이닝에서 활용하고 있는 극성 분류와는 달리 심리학적으로 정의된 감정을 기준으로 감정을 분류하였고, 실제 인스타그램에서 사용되는 감정을 분석하기 위해 인스타그램의 해시태그를 이용하여 감정 카테고리를 제안하였다. 기존 감정 분석에 대한 연구의 경우 텍스트의 감정을 분류하기 위해 감정사전을 이용하여 감정어휘에 대한 극성 값을 부여해 긍/부정을 판별하여 감성을 판단하였으나 제안하는 감정 카테고리를 통해 사용자의 주요 감정을 분석할 수 있어 주관적인 감정에 대한 모호함을 해결할 수 있다.

B. 연구 내용 및 구성

본 연구의 주요 내용은 SNS 중 하나인 인스타그램의 게시물을 활용하여 사용자의 감정을 분석하는 방법을 제시하기 위해 본 논문은 다음과 같은 구성으로 작성되었다.

본 장인 서론에 이어 2장 관련연구에서는 본 연구의 이론적인 배경과 관련하여 감정 분석의 기존 연구들을 살펴봄으로써 본 연구 내용의 이해를 돕는다.

3장에서는 인스타그램을 대상으로 사용자 감정 분석의 방법에 대해 기술한다. 본 연구에서 제안하는 사용자 감정 분석의 방법으로는 크게 해시태그를 이용한 감정 분류 방법과 제안한 해시태그 기반 감정 카테고리를 사용자의 게시물에 적용하여 감정을 분석하는 방법에 대해 제시한다.

4장에서는 데이터 수집에 관해 기술하고, 본 연구에서 이용한 학습 데이터 셋과 실험 데이터 셋에 대하여 설명한다. 또한 제안하는 감정 카테고리에 대한 정확성을 측정하여 성능을 평가한다.

마지막으로 5장에서는 본 연구에 대한 전체적인 결과를 요약하고, 향후 연구를 제시하며 마무리한다.

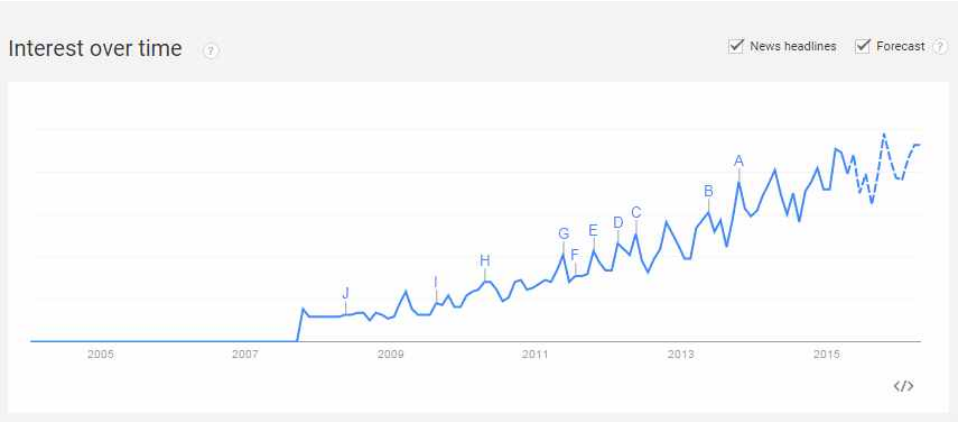
Ⅱ. 관련 연구

A. 감정 분석(Sentiment Analysis)

감정 분석이란 문장이나 문서 등 텍스트 정보에 표현된 사람들의 의견이나 태도, 감정 등을 분석하는 것을 말한다. 오피니언 마이닝(Opinion Mining)이라고도 불리는 감정 분석은 크게 3단계로 나누어질 수 있다[4].

첫 번째 단계는 데이터를 수집하는 단계이다. 블로그나 상품평 게시판 등 공개적인 데이터들뿐만 아니라 소셜 네트워크 서비스에서도 데이터를 수집할 수 있다. 두 번째 단계에서는 수집한 데이터에서 개인 정보나 감정과는 관련이 없는 부분을 배제시키고 감정 분석에 사용될 데이터만 분류한다. 마지막 세 번째 단계에서는 분리된 데이터의 '긍/부정'을 판단하는 단계로 프로그래밍을 사용하여 긍정적, 부정적인 단어를 탐지해 이를 정량화 한 뒤 통계적 기법을 적용한다.

최근 감정 분석에 대한 관심도가 높아지면서 연구 분야에서도 다양한 연구가 활발히 이루어지고 있다. 다음 [그림 2-1]은 구글 트렌드(Google Trends)[5]라는 서비스에서 감정 분석(Sentiment Analysis)을 검색한 그래프이다.



[그림 2-1] 감정 분석(Sentiment Analysis)에 대한 Google Trends 검색

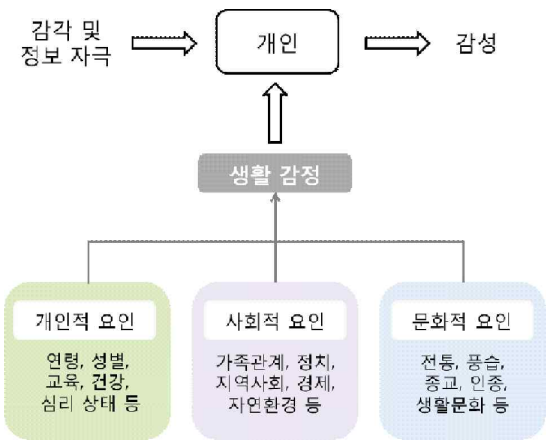
구글 트렌드란 검색되어지는 키워드와 관련 키워드와 연관도를 종합하여 지역별, 기간별 검색량을 보여주는 서비스이다. 대표적으로 구글 트렌드에서 미국의 대선 결과를 보여주어 어떤 대통령이 당선될 가능성이 큰지를 대선이 가까워질수록 보여주었고, 가장 우수한 예측 치수 중의 하나임을 증명하였다.

[그림 2-1]에서는 감정 분석을 키워드로 검색하여 전 세계를 기준으로 2015년 현재까지의 관심도와 앞으로의 예측까지를 보여주며 2007년을 기점으로 방대한 성장을 보여준다. 이러한 감정 분석을 이용하면 온라인상에서 사용자들의 특성이나 주관적인 생각들을 신속하게 파악할 수 있을 뿐 아니라 의사결정에 도움이 되는 유용한 정보를 제공할 수 있어 다양한 분야에 활용이 모색되고 있다.

본 절에서는 감정 분석의 연구 수단인 감정 정보를 살펴보고, 인스타그램을 대상으로 하게 된 배경과 소셜 네트워크 서비스(SNS) 데이터 기반의 감정 분석 연구 사례에 대해 기술한다.

1. 감정 정보

감정의 사전적인 의미는 '어떤 현상이나 일에 대하여 일어나는 마음이나 느끼는 기분'이라고 정의되며, 심리학적으로는 개인적, 심리적, 사회적, 문화적 원인으로 감정이 발생한다고 한다[6]. 다음 [그림 2-2]은 개인의 감정 발생에 미치는 요인들을 표현한 그림이다[7].



[그림 2-2] 개인의 감정 발생과 요인

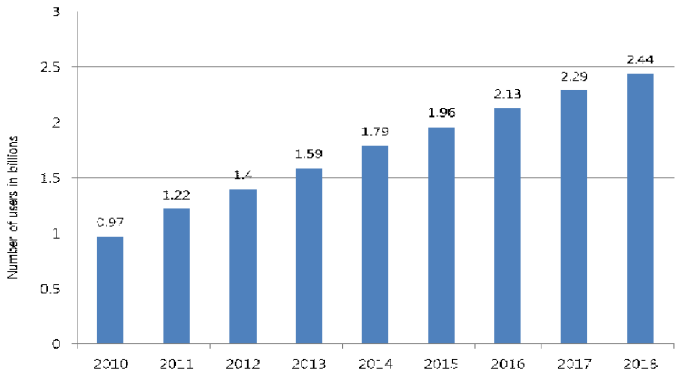
인간은 자신의 감정을 표현하기 위해 신체적 또는 언어적, 생리적 반응 등 다양한 방법으로 표현한다. 최근 얼굴 표정과 제스처의 영상정보, 맥박이나 심박수, 뇌파 등의 생체정보, 음성정보 등 인간의 신체적인 요소를 감지하여 감성정보를 추출해 인간의 감성을 일상생활에서 이용하는 제품이나 서비스에 융합함으로써 고부가가치를 창출할 수 있는 핵심기술인 감성 ICT 기술이 주목받고 있다[8].

이처럼 신체적 요소를 통해 감정을 연구하는 방법 외에 주로 사용되는 수단으로는 인간의 언어적 요소가 있다. 인간의 언어는 포괄성과 다양성의 특징을 가지므로 감정을 표현하는 데 있어서 가장 발달된 방식이다[9].

또한 감정을 측정하는데 있어서 언어적 표현 중 형용사 어휘를 사용하는 방법이 있는데, 형용사는 내적 상태를 기술하는 어휘로 측정 대상으로부터 감성적인 측면을 기술할 수 있는 모든 어휘의 대규모 집합을 구한 다음 어휘 간의 유사성 판단 등을 통해 감성의 구성 차원이나 범위를 추출한다[10]. 본 연구에서도 감정을 분석함에 있어서 형용사 어휘를 활용하였다.

2. 소셜 네트워크 서비스(SNS) 데이터 기반 연구

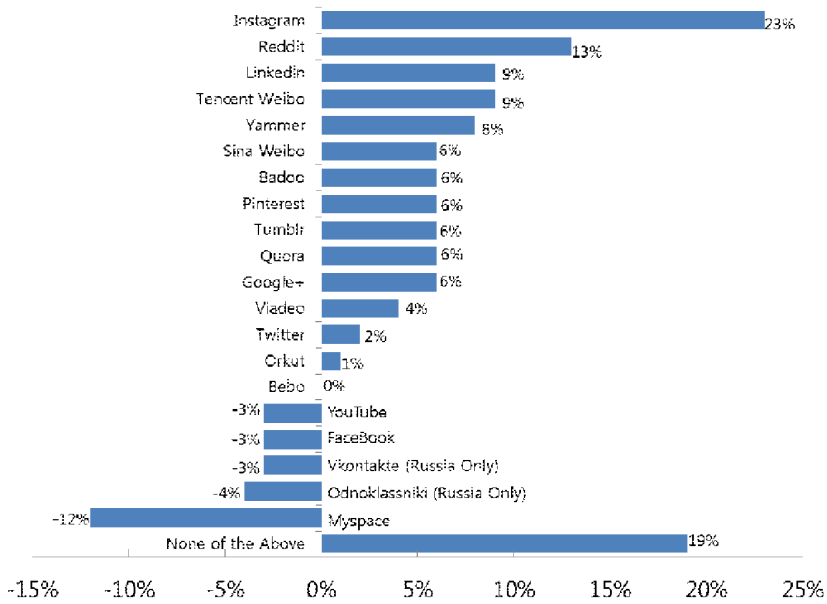
소셜 네트워크 서비스는 온라인상에서 공통적인 관심사를 공유하고 있는 사용자들 간의 관계를 형성하는데 정보를 공유하거나 여러 분야에 걸쳐 사용자들의 의견이나 감정을 공유하는 등 다양한 서비스를 제공함으로써 전 세계적으로 사람들의 생활 일부가 되고 있다. 최근 스마트 폰의 보급과 함께 소셜 네트워크 서비스의 사용자 수가 급증하고 있다.



[그림 2-3] 전 세계 소셜 네트워크 사용자 수

[그림 2-3]는 2010년부터 2018년까지의 전 세계적으로 소셜 네트워크 서비스의 사용자의 수를 보여준다[11]. 2010년부터 2014년까지의 소셜 네트워크 사용자 수를 통계로 2018년까지의 사용자 수를 예측한 것으로, 전 세계적으로 소셜 네트워크 침투는 계속 증가하고 있다. 실제 2012년의 사용자 수인 14억 명 중 인터넷 사용자의 63.1%가 소셜 네트워크 사용자로 분석되었고, 이 수치 또한 증가할 것으로 조사되었다. 2016년도에는 소셜 네트워크 사용자 수가 21억 명을 넘어설 것으로 추정되며, 앞으로도 전 세계적으로 스마트 폰과 모바일 기기의 사용으로 인한 소셜 네트워크의 사용 수치 또한 증가한다는 것을 알 수 있다. 편리한 접근성과 여러 분야에서 유용하게 사용되는 소셜 네트워크 사용자의 수가 증가함으로써 더욱 영향력이 증대되고 있으며 이에 따라 소셜 네트워크 서비스를 이용한 연구들도 활발하게 이루어지고 있다[12,13,14,15,16,17].

다음 [그림 2-4]는 SNS 서비스별 이용 성장률 현황을 보여준다[1,2]. 2013년도 2분기에서 4분기 사이에 실질적 사용률에 있어서 가장 높은 성장률을 보인 SNS는 2012년 페이스북(Facebook)이 인수한 사진 공유 서비스인 인스타그램이었다. 지난 6개월 동안 페이스북은 -3%의 성장률을 보인 반면 인스타그램은 23%의 사용률을 차지하며 전 세계 모든 소셜 미디어중 가장 빠르게 성장한 것으로 조사되었다.

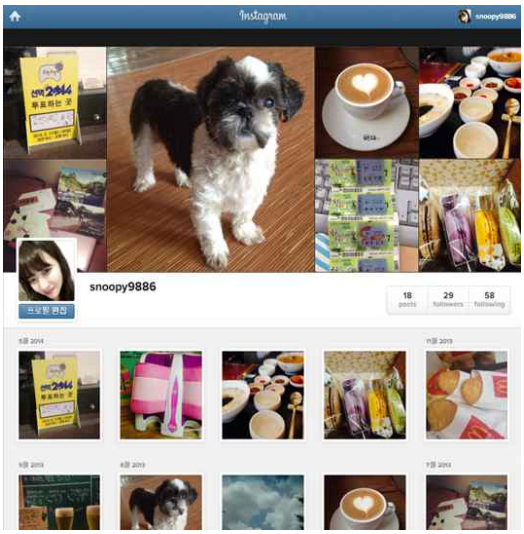


[그림 2-4] SNS 서비스별 이용 성장률 현황

[그림 2-4]에서도 볼 수 있듯이 페이스북은 여전히 전 세계적으로 강세를 보이나, 실질적 사용률을 보았을 때 관심사나 특정한 주제를 중심으로 공유하는 최신 소셜 네트워크 서비스인 인스타그램(Instagram)이나 텀블러(Tumblr), 핀터레스트(Pinterest) 등으로 옮겨가고 있는 상황이다. 또한 연구 분야에서도 기존의 보편화된 SNS를 활용한 연구가 대부분이고, 최신 소셜 네트워크 서비스를 분석하는 연구는 미흡한 실정이다. 따라서 본 연구는 최신 동향에 발맞춰 떠오르는 SNS인 인스타그램을 대상으로 하였다.

인스타그램은 온라인 사진 공유 및 소셜 네트워크 서비스로 2010년부터 서비스를 시작하였다. 사용자들은 인스타그램을 통해 사진을 찍거나 기존에 찍은 이미지를 공유하고 싶을 때 이미지를 선택하여 동시에 21가지의 사진 편집 내 필터 효과를 적용하여 간단한 문구와 해시태그를 통하여 다른 사용자들과 소통할 수 있다.

다음 [그림 2-5]는 인스타그램의 특징인 정사각형 모양의 직관적인 인터페이스를 보여준다.



[그림 2-5] 인스타그램(Instagram)

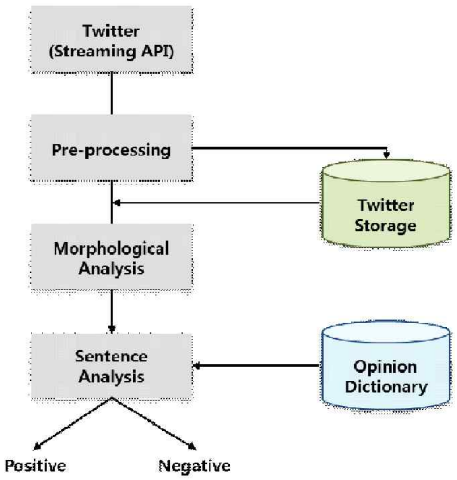
인스타그램의 특징은 친구를 맺지 않아도 팔로잉(Following)을 통해 다른 사람의 사진을 볼 수 있고, 이미지 관련 단어를 #과 함께 태그하면 카테고리가 반영된다. 또한 해시태그(Hashtag)를 통해 동일한 주제를 한 번에 모으고 관심을 유발시킬 수 있으므로 마케팅 수단으로 사용되기도 한다. 본 연구는 이러한 특징과 최근 실

질적 사용률이 높은 SNS인 인스타그램을 기반으로 감정 분석의 연구 대상으로 사용하고자 한다. 우선, 보편적인 SNS 데이터를 사용하여 감정 분석을 수행한 연구를 살펴보기로 한다. [12]에서는 트위터(Twitter)로부터 수집한 데이터를 이용하여 기계학습 모델을 적용해 7개의 감정인 '분노', '혼란', '우울', '피로', '친근감', '긴장감', '생동감'으로 영화평을 분류하여 영화 장르별 감정특성을 분석하였고, 직관적으로 알아볼 수 있는 결과를 통하여 감정 분석이 데이터의 성격을 넘어서 실제 응용분야에 적용이 가능함을 주장하였다. [13]에서는 트위터의 텍스트를 분석하여 다음 [표 2-1]의 정보를 추출한 후 기본적으로 정의해 놓은 감성을 수정하는 방법을 제안하여 감성의 극성뿐만 아니라 긍정과 부정의 근거가 되는 감성을 재구성하였다.

[표 2-1] 감성 정보 프레임의 구성 요소

구성 요소	내용	값	방향
감성도	감성의 강도	1~5	-
극성	감성의 극성	긍정/부정	-
앵커	화자의 감성을 직접적으로 표현	-	-
트리거	감성 표현의 이유 혹은 원인	-	-
수식어	감성의 강도 변화	1~3	강화/약화

[그림 2-6]은 [14]에서 트위터에서 지방선거와 관련된 데이터를 분석해 주어와 감정어휘의 거리 가중치를 적용해 제안하고 있는 오피니언 마이닝 흐름도이다.



[그림 2-6] 오피니언 마이닝 흐름도

B. 해시태그(Hashtag)

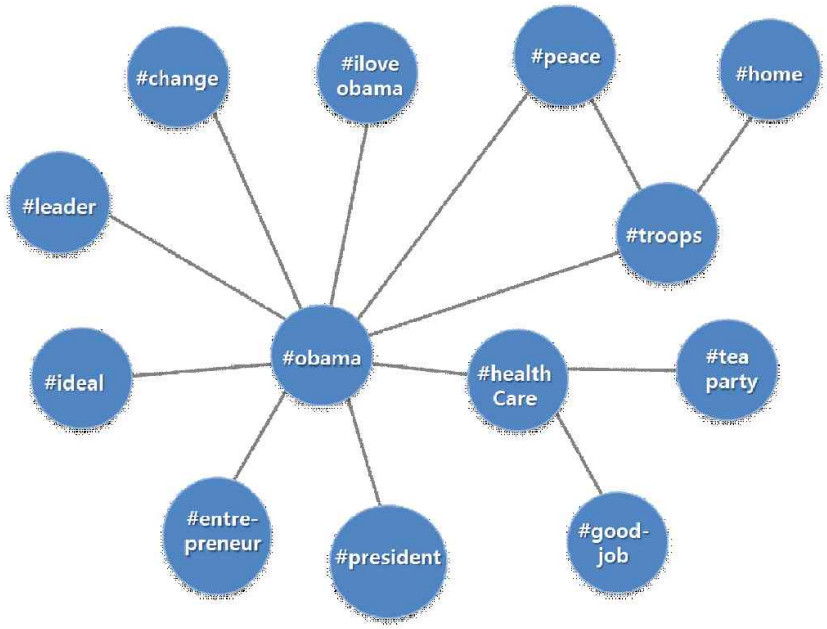
해시태그(Hashtag)란 '#'기호 뒤에 특정 단어를 써서 트위터나, 페이스북, 인스타그램 등 SNS상에서 특정 키워드를 편리하게 검색할 수 있도록 도입된 기능이다. SNS에서 사용자는 게시물에 해시태그를 함께 게시함으로써 특정 주제를 키워드로 관심사를 표현할 수 있고, 자신의 감정이나 상태에 대한 정보 또한 키워드로 공유할 수 있으며 이를 통해 사용자들 사이에서 공감을 이끌어낼 수 있다.

다음 [그림 2-7]은 해시태그의 사용 예를 보여주며 해시태그 기반으로 이미지 검색이 가능하여 동일한 주제나 관심사에 대해 빠르게 공유할 수 있다는 것을 보여준다.

[그림 2-7] 해시태그(Hashtag)의 사용 예

해시태그를 활용한 기존 연구들[15,16,17]중, [15]의 연구에서는 제안하는 트위터의 메시지를 감독분류(supervised classification) 방법론을 활용하여 해시태그가 주제를 표현하는 중요한 지표로 이용될 수 있다는 것을 실험결과로 보여주었다.

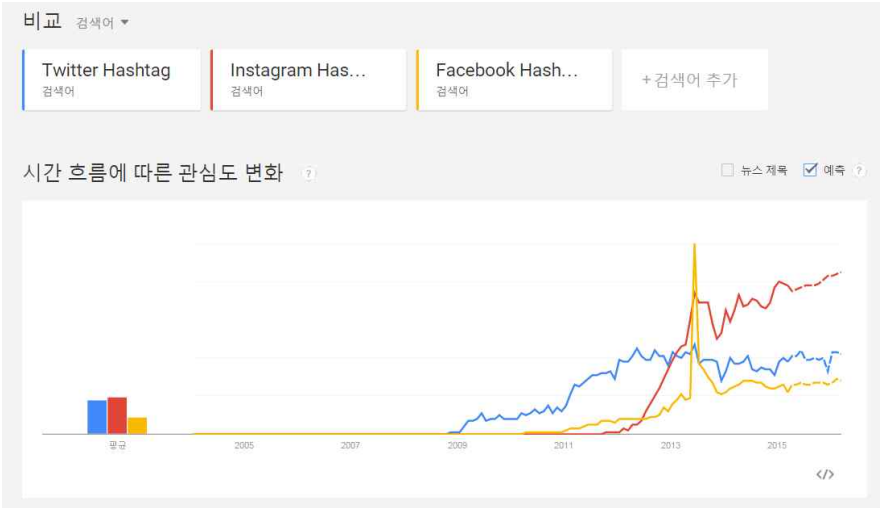
[16]에서도 트위터 상에서 감성 분석을 하는데 있어서 [그림 2-8]의 그래프 모델을 기반 하여 해시태그 감성 분류 접근법을 제시하였다. [그림 2-8]은 트위터에 한 번 이상 동시에 발생한 경우를 연결시킨 그래프의 예이며, 해시태그의 유형을 주제에 대해 주관적인 의견을 표현하는 감정 해시태그, 주제와 관련된 주제 해시태그, 표현하려는 대상과 감정이 함께 나타난 감정 주제 해시태그로 세 가지 범주로 정의하였다. 따라서 그래프 모델의 나타나는 동시 발생된 관계를 통해 감정의 극성을 결정하였고, 해시태그가 감성 분석의 중요한 요소로 작용한다는 것을 실험적으로 증명했다.



[그림 2-8] 해시태그 그래프 모델의 예

또한 [17]에서도 소셜 네트워크의 상승은 해시태그를 지원하는 기능이 중요한 역할로 작용함을 언급하였으며 인스타그램에서 나이브 베이즈 분류기(Naive Bayes classifier)를 사용하여 해시태그의 감정 분류를 시도하였다.

다음 [그림 2-9]는 구글의 트렌드 서비스[5]에서 해시태그를 사용하는 대표적인 SNS인 트위터, 인스타그램, 페이스북 각각의 해시태그를 검색어로 입력하여 전 세계에 대한 2005년부터 2015년 현재까지의 검색어에 대한 관심도와 앞으로의 예측을 보여준다.



[그림 2-9] 3가지 SNS 해시태그에 대한 Google Trends 비교 검색

이처럼 해시태그를 사용하는 대표적인 SNS 3가지를 비교 검색해본 결과 2015년 현재에는 인스타그램 해시태그가 가장 큰 관심을 보이고 있고, 이슈화가 되고 있는 것을 알 수 있다. 또한 기존 연구들에서 보았듯이 해시태그의 중요성과 감성 분석에서의 해시태그 활용 가치가 충분히 입증되었으므로 본 논문에서도 해시태그의 유용성에 따라 감정을 분류하는데 활용하였다.

C. 감정 분류(Sentiment Classification)

감정 분류는 문서나 문장에서 사용자들이 표현하는 감정을 추출하는 연구이다. 기존의 감정 분석의 연구에서는 감정어를 긍정과 부정으로 주로 극성 분류에 초점을 두었다면, 최근에는 감정어마다 극성을 부여하여 감정을 여러 가지로 분류하는 감정 분류에 대한 연구가 많이 진행되고 있다[22,23].

본 절에서는 인스타그램의 감정을 분류하는 방법에 있어서 심리학에서 분류하고 있는 몇 가지 감정 분류 체계를 알아보고, 기존 감정 분류 연구에 있어서 감정어 극성 분류 방법에 대해 설명한다.

1. 심리학적 감정 분류

감정에 대한 분류 체계로 심리학에서 분류하는 이론들[18,19,20]이 있다. 먼저 [18]의 POMS(Profile of mood States) 이론은 65개의 형용사로 구성된 형용사 단어를 다음 [표 2-2]와 같이 요인 분석하여 설문을 통해 평가하는 방법이다.

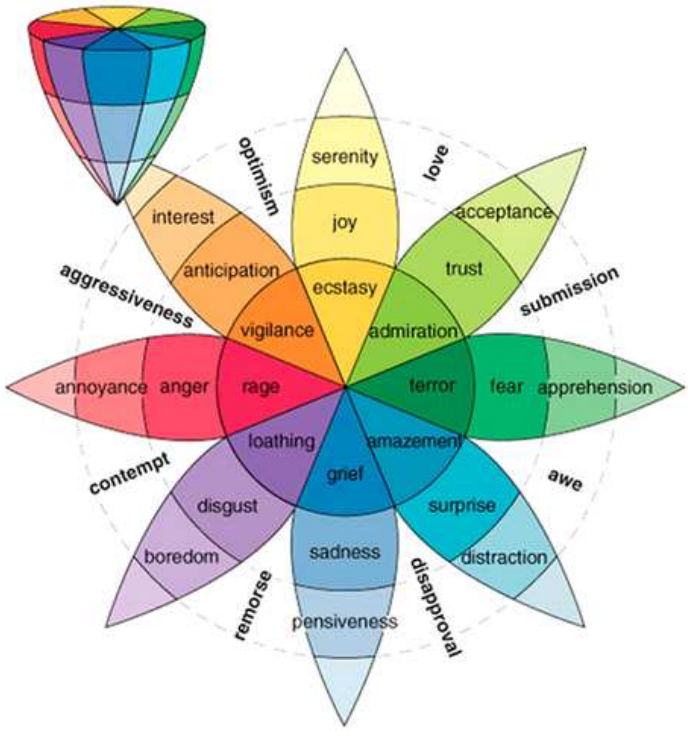
[표 2-2] POMS의 요인분석

POMS Standard
tension - anxiety
depression - dejection
anger - hostility
fatigue - inertia
vigor - activity
confusion - bewilderment

[12]의 연구에서는 [표 2-2]를 기초로 한국어로 번역한 후 감정형태소를 정의하고 기계학습 데이터로 이용하여 7개의 감정으로 분류하였다. 하지만 POMS 이론은 설문을 이용해 그 점수를 합계한 것에 기초하여 정서를 측정하는 데에는 유용하지만 감정을 분류하는 데는 모호한 부분이 있다.

Plutchik[19]은 동물과 인간에게 있어 공통적으로 나타나는 기본적인 감정을 8가지로 정의했으며, 인간의 감정을 [그림 2-10]과 같이 원반상에 감정 바퀴(Wheel of Emotions)라고 표현하였다.

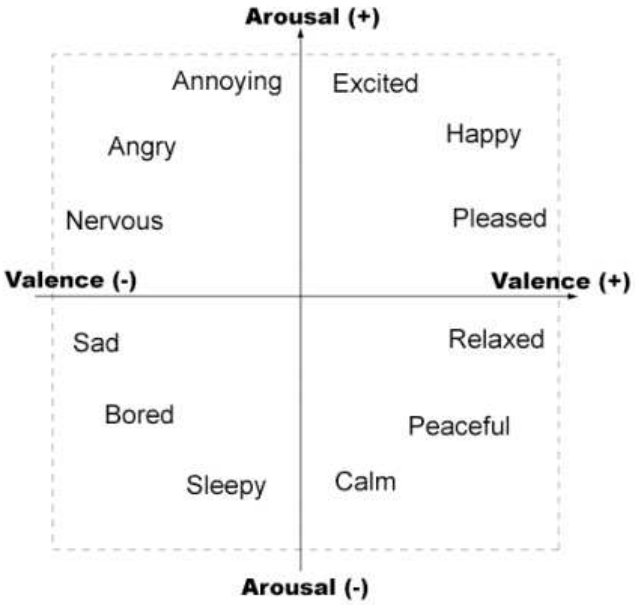
대체로 마주보고 있는 감정은 반대되는 감정을 나타내고, 8개의 기본 감정들이 존재하고, 강도에 따라 더 나뉘게 되며 각 기본 감정들이 혼합되어 여러 조합 감정을 이룬다. 또한, 감정의 세기에 따라 원반상에서 바깥쪽으로 갈수록 약한 감정들이 배열되어 있다.



[그림 2-10] Plutchik의 감정 바퀴

[13]의 연구에서는 [19]의 분류 체계를 인지적으로 분석해 오피니언 마이닝에 사용하기 위해 긍정과 부정으로 극성 판단을 하여 기본 감정 8개와 조합 감정 중 일부를 극성 분류를 하였다. 하지만 이 연구에서는 [그림 2-10]의 감정 바퀴에서 보았듯이 감정을 명사로 정의하였는데, 명사의 경우에는 그 단어의 의미에 종속되는 경향이 있어 대표 단어의 선택이 어렵게 된다[13].

Thayer 모델[20]은 생물심리학적인 관점으로 감정을 분류한 모델로 다음 [그림 2-11]과 같이 2차원의 공간에 긍정과 부정의 정도에 따른 Valence 축과 감성의 강도를 나타내는 Arousal 축을 기준으로 서로 다른 12가지의 감정들이 분류되어 있다. 또한 감정을 연구하는 분야에서도 주로 채택되어 사용하고 있다[21].



[그림 2-11] Thayer의 감정 모델

[21]에서는 감정을 형용사로 표현한 것과 특정 카테고리로 분류하는 모호성을 해결해준 [20]의 모델에서 제시하는 감정 형용사를 사용해 위치 기반 서비스에 감정을 적용하여 위치 카테고리를 추천해주는 감성 모델을 제안하였다.

본 논문에서는 [그림 2-11] 중 감정 분류의 정확성 향상과 오분류율을 최소화하기 위하여 4가지의 대표 감정인 Happy, Angry, Peaceful, Sad를 감정 형용사라고 정의하였으며, 각각의 감정 형용사가 포함된 해시태그들을 추출하여 인스타그램 사용자의 감정을 분류하고자 한다.

2. 감정어 극성 분류

감정어의 극성을 기반으로 감정을 분류하는 방법으로는 대부분 다음 [그림 2-12]과 같은 구조를 가진다[22,23]. 시스템 구성도의 순서에 따라 문서 또는 입력된 문장에서 단어를 품사별로 분석하여 분리해낸다. 형태소 분석을 통해 추출된 단어들은 감정 사전 등을 이용하여 해당 문장이나 문서의 특징 벡터를 구성한다. 이때 특징 벡터는 감정의 가중치로 사용될 수 있으며 최종으로 분류기에 의해 긍정이나 부정 또는 중립 중 3가지의 감정으로 분류가 된다.

[그림 2-12] 감정어 극성 기반 감정 분류 시스템 구성도

[22]에서는 [그림 2-12]에서 고려되지 않았던 한국어 문장의 문맥을 고려하여 분석하는 과정을 제안함으로써 기존의 극성 기반 감정 분류 시스템보다 우수한 성능을 실험적으로 보여주었다. [23]에서는 [그림 2-12]의 구성도 흐름대로 형식적인 감정 단어와 이모티콘 같은 비형식적인 감정 단어를 이용하여 문장의 감정을 분류하는 시스템을 제안하였다. 이와 같은 방법 외에도 감정어의 극성을 분류하는데 있어서 Pointwise Mutal Information(PMI)를 이용하는 방법과 감정 사전(Sentiment Lexicon)을 이용하는 방법이 있다.

a. Pointwise Mutal Information(PMI)를 이용한 방법

PMI는 확률론적인 방법에 기초한 방법으로 두 단어의 연관성을 측정하기 위해 사용되는 방법이다[24]. 다음 식 (1)은 단어 A와 B가 얼마나 연관성이 있는지를 확률적으로 계산하는 과정을 나타낸다.

$$MI\ A, B) = \log \frac{P(A, B)}{P(A)P(B)} \quad (1)$$

식 (1)에서 P(A, B)는 두 단어가 한 문장에 동시에 출현할 확률을 나타내며 PMI (A, B)의 값은 동시에 출현할 확률에서 각각의 두 단어가 독립적으로 출현할 확률로 나눈 값이다. PMI(A, B)의 값이 클수록 두 단어의 연관성이 높다는 것을 말한다. 즉, PMI의 값을 이용하여 단어의 극성을 분류하는 또 다른 방법으로는 SO-PMI(Sentiment Orientation from Pointwise Mutal Information)가 있다[24].

$$SO-PMI(word) = \sum_{pword} PMI(Word, pword) - \sum_{nword \in Nwords} PMI(Word, nword) \quad (2)$$

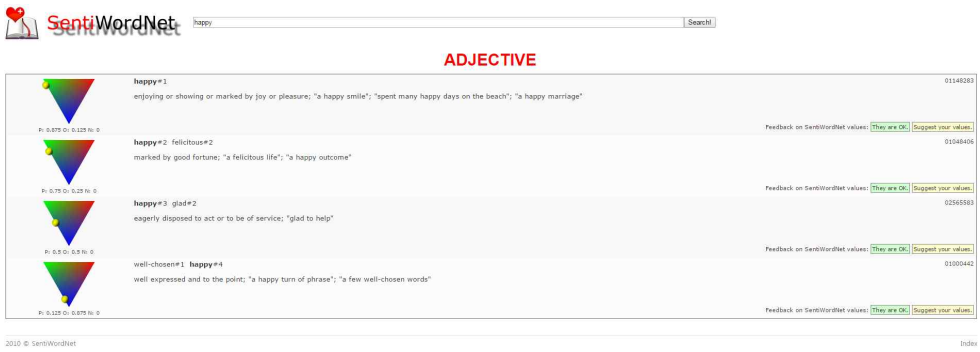
SO-PMI는 전문가가 사전에 긍정의 단어 집합인 pword와 부정의 단어 집합인 nword를 기준이 되는 Seed 집합으로 정의해 놓고, 식 (2)에서 극성 값을 알고자 하는 단어와 pword의 PMI 값의 합과 nword의 PMI 값의 합을 구해 둘의 차이를 구하여 계산한다. 식 (2)의 결과 값이 양수라면 긍정의 극성으로, 음수면 부정의 극성으로 분류된다. [25]에서는 상품평을 기반으로 상품에 대한 속성에 따른 감정어의 극성이 달라지는 특징을 고려해 속성별로 극성을 분류하는 방법을 제안하였는데, 속성별 감정어 집합을 생성해 식 (2)의 SO-PMI에 속성 정보를 추가하였다.

[26]에서는 SO-PMI는 학습 문서나 기준 용어로 인해 극성 분류가 달라진다는 점, 인터넷에서 쓰이는 단어와 잘못된 맞춤법이 오류를 범하는 점, 한국어 어미의 여러 가지 표현 방법으로 분석에 영향을 주는 점, 불필요한 단어의 극성이 분류되는 네 가지의 문제점을 해결하기 위해 제품의 특징을 반영하여 극성을 분류하는 방법을 제안하였다.

b. 감정 사전(Sentiment Lexicon)을 이용한 방법

감정의 분류나 오피니언 마이닝의 연구에 있어서 대부분 감정 사전을 바탕으로 극성을 결정하기 때문에, 감정 사전에 수록된 단어들의 집합이 중요한 요소로 작용한다. 이에 따라 감정 사전의 정확도를 높이기 위한 연구도 이루어지고 있다.

감정 분석의 연구에는 보통 공개된 영어권 감정 사전인 WordNet-Affect[27]나 SentiWordNet[28]을 이용한다. 그중 대표적으로 사용되고 있는 SentiWordnet은 영어 단어의 의미별로 긍정과 중립, 부정 3가지의 감정 강도 값을 부여해 둔 사전으로 다음 [그림 2-13]은 [28]이 제공하는 'happy'에 대한 감정 강도를 보여준다.



[그림 2-13] SentiWordNet이 제공하는 감정 강도

[26]에서는 제안하는 방법을 비교 실험하기 위하여 한국어 단어를 영어로 번역하여 [28]을 이용하여 극성 분류를 하였다. [29]에서는 영어 단어의 의미별로 긍정과 부정의 감정 강도를 저장해 둔 [28]을 기반으로 한 다양한 감성 자질 추출에 대한 기존 연구에 대하여 방법들을 비교 분석하였다.

한국어의 경우에는 영어 감정 사전을 번역해서 사용하거나 한국어의 특성에 따른 고려사항이 발생할 때 감정 사전을 수작업으로 구축하여 사용한다[30].

다음 [표 2-3]은 [30]에서 한국어 문법의 반어법 처리를 위해 오피니언 감정 사전을 구축하는데 적용한 반의어(Antonym) 규칙 알고리즘의 예를 보여준다.

[표 2-3] [30]의 반의어 규칙 알고리즘

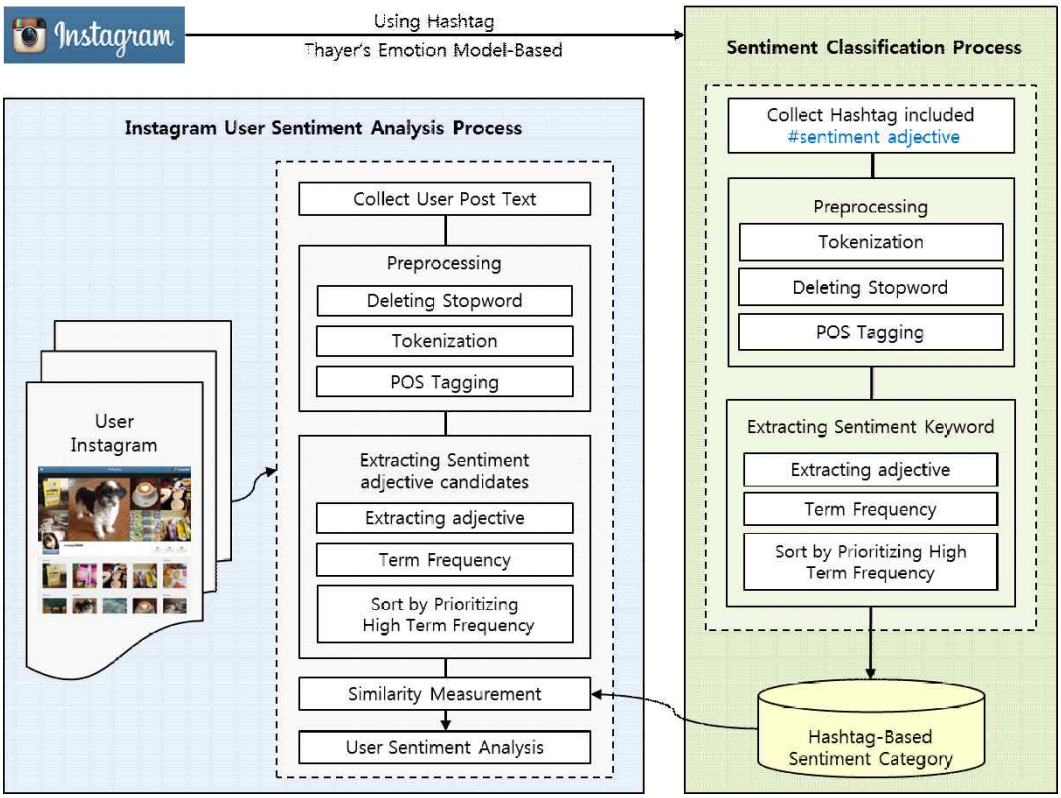
반의어		종결	
상승	하였지만	하락	하였다.
POS	NEG	NEG	NEG
POS + (Antonym)Neg = Neutrality		Neg + Neg = Neg	

Ⅲ. 인스타그램 기반 사용자 감정 분석

본 논문에서는 인스타그램에서 심리학적 감정 분류 체계 이론 중 하나인 Thayer의 모델 기반으로 해시태그를 이용하여 감정을 분류한 뒤 사용자가 작성한 게시글에 분류한 감정 카테고리를 적용하여 감정을 분석하는 방법을 제안한다.

A. 시스템 구성도

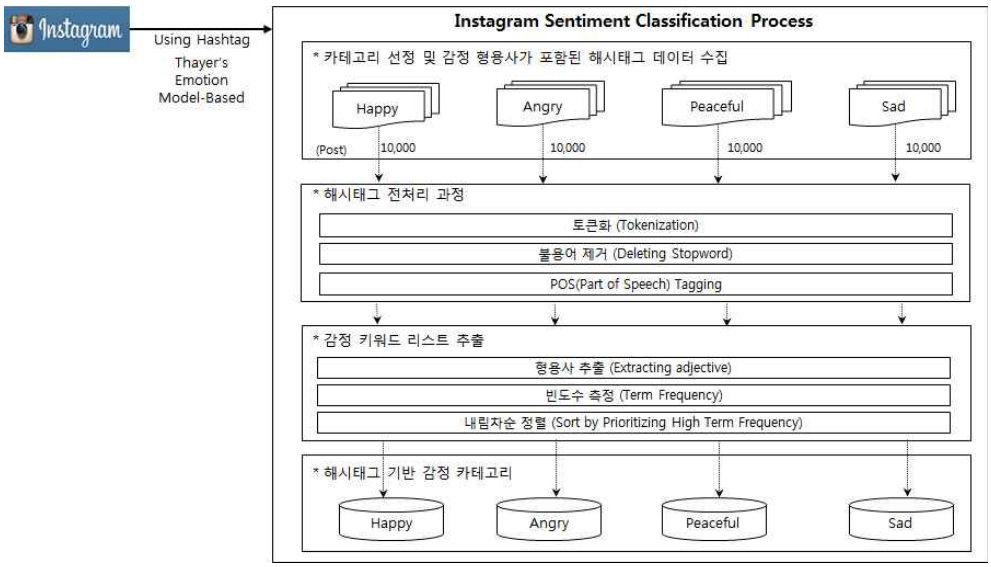
[그림 3-1]은 제안하는 감정 분석의 전체 시스템 구성도이다. 크게 감정을 분류하는 프로세스와 사용자의 감정을 분석하는 프로세스로 나뉜다.



[그림 3-1] 전체 시스템 구성도

B. 인스타그램 감정 분류 프로세스

인스타그램에서 사용자는 공유하고 싶은 이미지와 간단한 문구나 몇 가지의 해시태그로 자신의 기분이나 상황을 표현하고 다른 사용자들과 공유한다. 최근 인스타그램의 트렌드를 살펴보면 구구절절한 설명 대신 해시태그를 나열하는 방식으로 공유하고 싶은 게시물을 설명한다. 이때 해시태그는 어떤 게시물에 대해 주제를 담는 키워드가 될 수 있으며 자신의 기분이나 상태를 공유하는 감정을 담는 키워드가 될 수 있다. 본 절에서는 [그림 3-2]와 같이 해시태그를 이용하여 인스타그램에서 감정을 분류하는 방법에 대해 기술한다.



[그림 3-2] 인스타그램 감정 분류 프로세스

본 연구에서는 인스타그램 내에서 사용자들이 자주 공유하는 감정들을 분류하기 위해 해시태그를 활용한다. 먼저 감정들을 분류하기 위해 카테고리를 선정하여 기준을 세운다. 선정된 각각의 대표 카테고리를 감정 형용사로 표현하여 감정 형용사가 포함된 해시태그 데이터를 수집한다. 수집한 데이터들을 전처리 과정을 거쳐 감정 키워드 리스트를 추출한다. 추출된 감정 키워드 리스트에서 선정된 감정 키워드만을 해시태그 기반 감정 카테고리로 분류한다.

1. 카테고리 선정 및 감정 형용사가 포함된 해시태그 데이터 수집

본 절에서는 카테고리의 선정과 인스타그램의 게시물로부터 이미지와 게시된 시간정보나 댓글 등을 배제하고 해시태그 데이터만을 수집하는 과정을 다룬다.

[그림 3-3]은 인스타그램에서 사용자와 해시태그 기반의 검색이 가능한 것을 보여주는데, 그 중 해시태그 기반 검색의 예를 나타내고 있다. 그림에서 볼 수 있듯이 '#SPRING'이라는 해시태그를 통하여 그와 관련된 게시물들을 보여준다. 이처럼 해시태그는 관심 있는 분야를 키워드를 통해 한 번에 모아볼 수 있으며 다른 사용자들과 공유하고, 그에 대한 감정들을 교류하며 소통할 수 있다.

[그림 3-3] 해시태그 기반 검색의 예

해시태그 데이터를 수집하기 앞서 먼저 감정을 분류하기 위한 기준을 세우도록 한다. 기준을 세우기 위해 사용된 자료는 심리학적으로 정의된 Thayer의 감정 분류 체계이다. 본 연구에서는 감정 분류의 정확성 향상과 오분류율을 최소화하기 위해 Thayer의 대표 4가지 감정인 Happy, Angry, Peaceful, Sad를 카테고리로 선정하여 분류 기준을 세우고, 이를 감정 형용사라고 표현하였다.

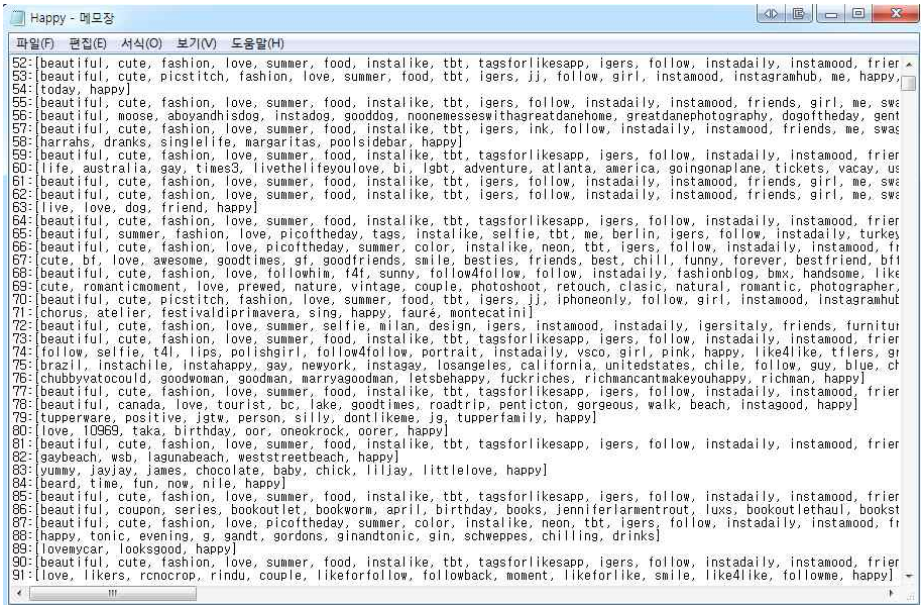
카테고리의 선정이 끝나면 인스타그램에서 정의된 감정 형용사가 포함된 해시태그 데이터를 수집하는데 그 과정은 다음과 같다.

(1) 인스타그램의 Open API(Application Platform Interface)를 사용하는 데 필요한 access_token을 발급받기 위해 인스타그램 개발자 페이지[31]에서 인증 절차를 거친다.

(2) #감정 형용사를 포함하고 있는 게시물의 정보들 중 해시태그만을 읽어온다.

(3) 읽어온 해시태그들을 게시물 하나를 기준으로 행 분리하여 Happy.txt 파일에 저장한다.

다음 [그림 3-4]는 위와 같은 과정을 통하여 'happy'라는 감정 형용사가 포함된 해시태그들 중 일부를 보여준다. 현재까지 추출된 부분은 게시물에서 이미지 등 다른 정보들은 배제하고 해시태그만을 수집한 결과이다. 이를 본 연구에 적용하기 위하여 다음 절의 전처리 과정을 거친다.



[그림 3-4] 감정 형용사가 포함된 해시태그 추출

2. 해시태그 전처리 과정

감정 형용사가 포함된 해시태그의 감정 키워드를 추출하기 위해서는 먼저 [그림 3-3]에서 전처리 과정이 필요하다.

본 절에서는 해시태그 데이터를 효과적으로 적용하기 위해 파이썬(Python)의 자연어 처리 라이브러리인 NLTK(Natural Language Toolkit)[32]를 이용하여 전처리 과정인 (1) 토큰화(Tokenizing), (2) 불용어 제거(Deleting Stopword), (3) POS(Part of Speech) 태깅 단계 순으로 기술한다.

(1) 토큰화(Tokenizing)

전처리 과정의 첫 단계로는 문장 내에서 단어별로 토큰화 시키는 것이다. 일반적으로 공백을 기준으로 하는 토큰화는 다음 [표 3-1]과 같다.

[표 3-1] 일반적인 토큰화 과정

```
>>> sentence = "At eight o'clock on Thursday morning“
>>> tokens = nltk.word_tokenize(sentence)
>>> tokens
['At', 'eight', "o'clock", 'on', 'Thursday', 'morning']
```

본 연구에서는 행 단위로 수집한 해시태그 단어들을 라인별로 한 줄씩 불러와 리스트 형태로 [표 3-2]와 같이 토큰화를 수행한다.

[표 3-2] 본 연구의 토큰화

```
>>> txtList=usertxt.readlines()
>>> for a in range(len(txtList)):
        tokens = nltk.word_tokenize(txtList[i])
```

(2) 불용어 제거(Deleting Stopword)

불용어는 단어를 분석할 때 제거가 되는 의미 없는 기능어를 뜻한다. 보통 불용어를 제거할 때에는 불용어 리스트를 이용하거나 상황에 따라 불용어를 정의해 사용한다. 이에 출현 빈도가 높은 영어의 관사나, 전치사, 접속사 등을 불용어로 취급하는 경우도 있고, 출현 빈도가 높지만 특정한 의미를 지니는 단어는 불용어에서 제외되는 경우도 있다. 다음 [표 3-3]은 파이썬의 NLTK 라이브러리에 포함된 영어단어의 불용어 리스트를 보여준다.

[표 3-3] NLTK 라이브러리에 포함된 불용어(Stopwords) 리스트

```
>>> from nltk.corpus import stopwords
>>> stopwords.words('english')
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', 'your', 'yours',
'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', 'her', 'hers',
'herself', 'it', 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves',
'what', 'which', 'who', 'whom', 'this', 'that', 'these', 'those', 'am', 'is', 'are',
'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does',
'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until',
'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into',
'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down',
'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here',
'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more',
'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so',
'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', 'should', 'now']
```

본 절에서는 최종 감정 키워드를 추출하는데 의미가 없는 숫자와 기호 등 알파벳 형태의 문자열을 제외한 나머지를 불용어로 정의한다. 해시태그는 보통 다른 사용자들과의 공유를 목적으로 게시하기 때문에 사용자들의 기본적인 표기법이 대부분 기본형태에 가까우며, 띄어 쓰거나 이모티콘은 해시태그로 인식 자체가 되지 않기 때문에 불용어 삭제 단계에서는 거의 숫자나 기호형태나 다른 나라의 언어들이 삭제된다. 불용어 제거 단계를 거치면 (3)의 POS 태깅 단계를 통하여 품사를 파악한다.

(3) POS(Part of Speech) 태깅

본 단계에서는 각각의 단어들에 대한 품사를 파악하는 POS 태깅 과정을 수행한다. 다음 [표 3-4]는 기본적인 POS 태깅 과정을 보여준다.

[표 3-4] 기본적인 POS 태깅 과정

```
>>> sentence = “At eight o'clock on Thursday morning”
>>> tokens = nltk.word_tokenize(sentence)
>>> tagged = nltk.pos_tag(tokens)
>>> tagged
[('At', 'IN'), ('eight', 'CD'), ("o'clock", 'JJ'), ('on', 'IN'), ('Thursday', 'NNP'), ('morning', 'NN')]
```

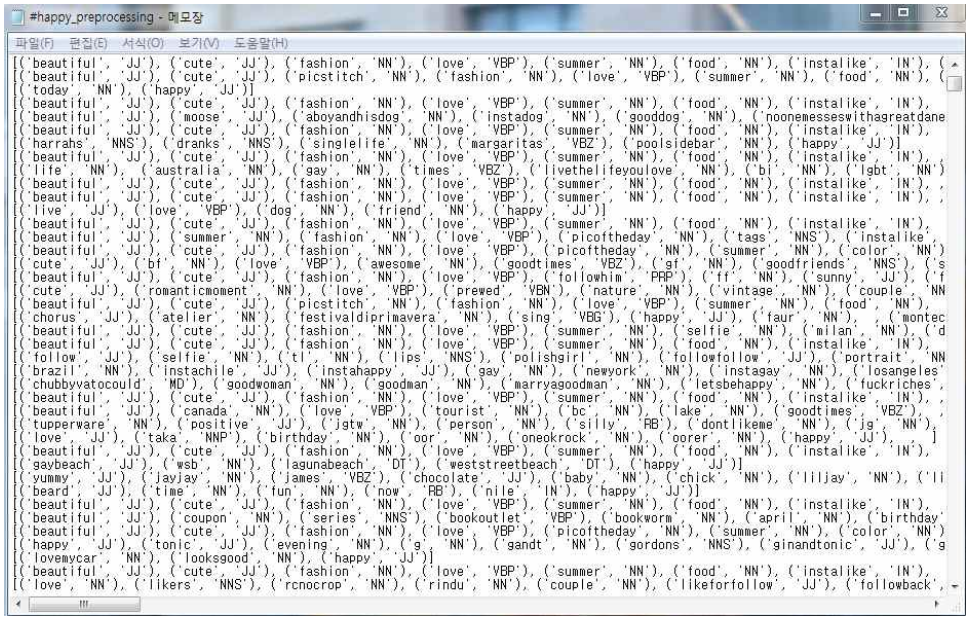
[표 3-5]은 품사를 파악하는데 이용한 Penn Treebank Project의 품사 태깅에 사용되는 태그들의 코드와 그에 대한 품사 목록을 나타내준다.

[표 3-5] POS(Part of Speech) Codes

CC	Coordinating conjunction	PRP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential there	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	to
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present
NNP	Proper noun, singular	VBZ	Verb, 3rd person singular
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP$	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

본 논문에서는 감정 형용사가 포함된 감정 키워드를 추출하기 위하여 태그가 JJ 즉, 형용사인 단어들을 이용한다. 또한 형용사(JJ)의 품사를 가지고 있는 JJR과 JJS도 함께 이용하여 태깅의 정확도를 높이기로 한다.

다음 [그림 3-5]는 전처리 과정을 거쳐 품사까지 파악하여 추출된 해시태그의 일부를 보여준다.



[그림 3-5] 전처리 과정을 거친 해시태그 추출

3. 감정 키워드 리스트 추출

본 절에서는 수집된 감정 형용사가 포함된 해시태그에서 전처리 과정을 거쳐 형용사 품사를 가진 단어들 중에서 빈도수(Term Frequency)를 측정하여 빈도수가 높은 순으로 내림차순 정렬해 감정 키워드를 추출한다.

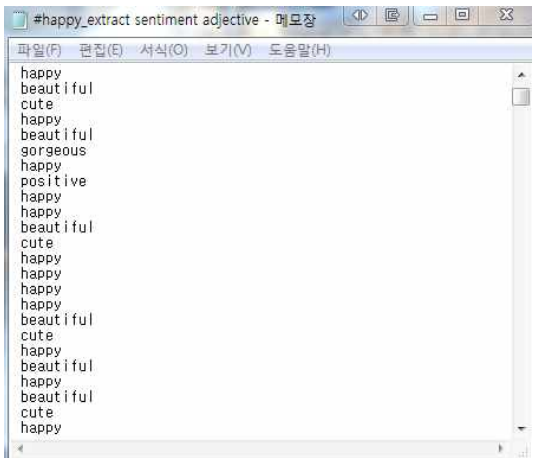
다음 [표 3-6]는 전처리 과정을 거쳐 추출된 감정 형용사와 동시에 출현하는 해시태그에서 형용사 단어만을 추출하는 알고리즘을 나타낸다. [그림 3-5]와 같이 리스트 형태로 품사까지 파악하여 저장된 데이터에서 한 행씩 불러들여 조건문을 사용하여 태깅된 값을 비교 연산자를 통해 형용사 품사코드인 JJ, JJR, JJS와 같으면 태깅된 단어를 저장시킨다.

[표 3-6] 감정 형용사와 동시에 출현하는 형용사 추출 알고리즘

```
for b in range(len(postag)):
if postag[b][1] == 'JJ' or postag[b][1] == 'JJS' or postag[b][1] == 'JJR':
        result.write(postag[b][0])
```

즉, [그림 3-5]의 한 행을 예로 들어 [('today', 'NN'), ('happy', 'JJ')] 가 태깅된 값이라면 'happy'라는 형용사가 추출되고 이를 저장시킨다.

[그림 3-6]는 [표 3-6]을 통해 생성된 감정 형용사 Happy의 리스트를 보여준다.



[그림 3-6] 감정 형용사 리스트

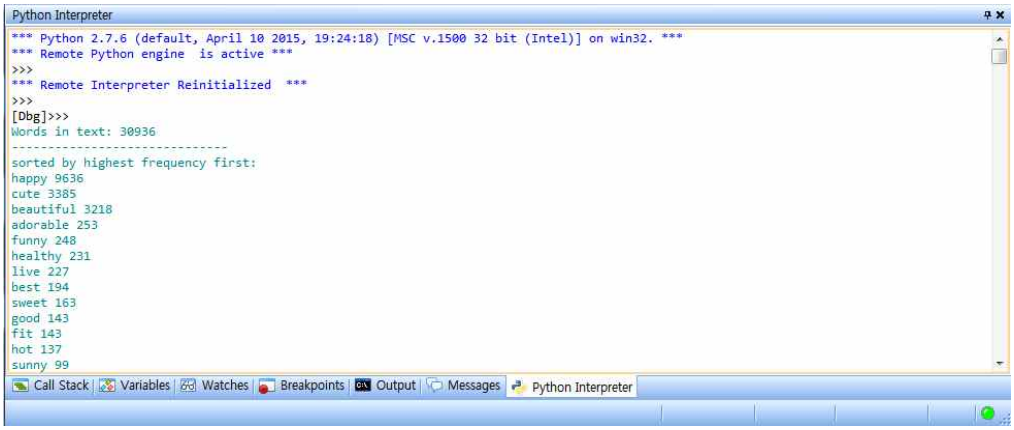
감정 형용사 리스트가 추출되면 빈도수(Term Frequency)를 측정하여 높은 빈도수가 나오는 단어를 우선순위로 내림차순 정렬한다.

[표 3-7]은 이에 대한 과정을 나타내는 알고리즘이다.

[표 3-7] 감정 형용사의 빈도수 측정과 내림차순 정렬 알고리즘

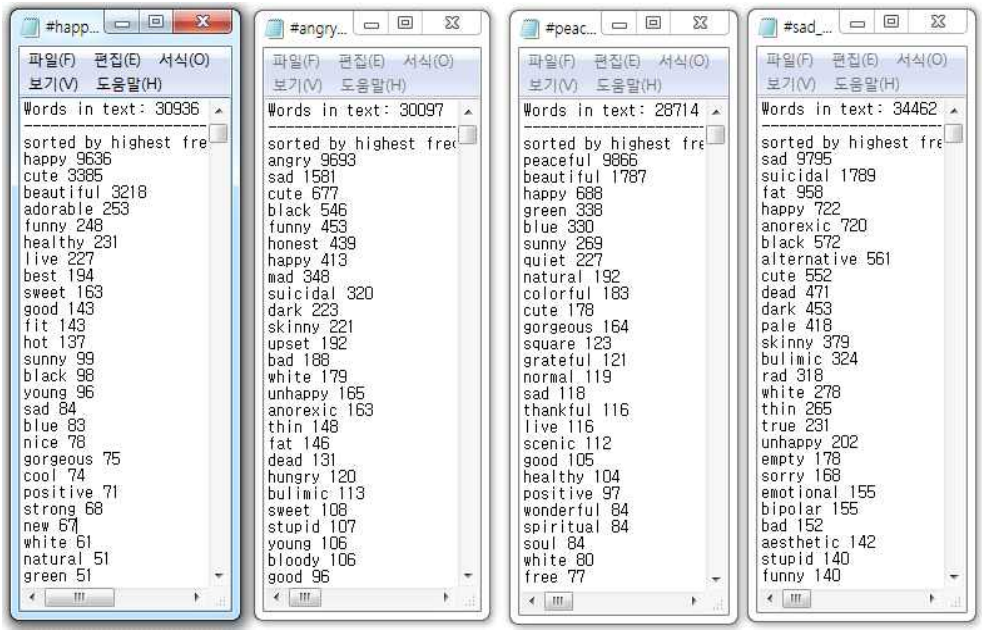
```
import nltk, re
filename = 'E:/#happy_extract sentiment adjective.txt'
word_list = re.split('\s+', file(filename).read().lower())
print 'Words in text:', len(word_list)
freq_dic = {}
## 단어와 빈도수를 넣을 빈 Dictionary를 생성
for word in word_list:
## 문자열 탐색
    try:
        freq_dic[word] += 1
        ## 문자 word가 Dictionary에 있으면 1을 증가
    except:
        freq_dic[word] = 1
        ## 문자 word가 Dictionary에 없으면 Key가 word이고 초기값
           1인 새 항목을 생성
print '-'*30
print "sorted by highest frequency first:"
freq_list = [(val, key) for key, val in freq_dic.items()]
## 튜플 쌍 (val,key)의 리스트를 생성
   Dictionary 메소드에서는 key를 이용해서 value를 추출.
   items()는 Dictionary 내에서의 모든 item(2개의 element로 구성된
   튜플)을 보여주며, 각각의 element는 key와 value를 뜻함
freq_list.sort(reverse=True)  ## 빈도수에 의한 정렬
## reverse의 인자값을 True로 주어 역순인 내림차순 정렬
for freq, word in freq_list:  ## 결과 보여주기
    print word, freq
f = open("E:/#happy_sentiment adjective TF.txt", "w")
f.write( str(freq_list) )
f.close()
```

[그림 3-7]은 [표3-7]에 대한 결과로 Happy에 대한 감정 키워드가 추출된 리스트 예를 보여준다.



[그림 3-7] 감정 키워드 추출

[그림 3-8]은 4가지의 주요 감정에 대해 추출된 감정 키워드 리스트들을 저장한 txt 파일을 보여준다.



[그림 3-8] 4가지 감정 키워드 리스트

4. 해시태그 기반 감정 카테고리 선정

본 절에서는 추출된 감정 키워드에서 빈도수의 횟수를 기준으로 선정된 감정 키워드를 해시태그 기반의 감정 카테고리로 분류한다.

[표 3-8]은 감정 키워드의 선정 기준을 나타낸 감정 키워드 필터링에 대하여 나타내고 있다.

[표 3-8] 감정 키워드 필터링

감정 키워드 선정 기준
frequency(감정 키워드 리스트) ≥ 100
해당 감정 키워드 리스트 단어 ≠ 감정 형용사 단어

감정 키워드 리스트의 빈도수가 100회 이상인 경우를 감정 키워드로 선정하였다. 또한 감정 형용사와 해당 감정 키워드 리스트의 각각의 단어가 일치하는 경우는 제외하였다. [표 3-9]는 감정 키워드 필터링을 적용하여 선정된 결과를 보여준다.

[표 3-9] 감정 키워드 선정

감정 키워드 리스트	Freq	감정 키워드
happy	9636	X
cute	3358	○
beautiful	3218	○
adorable	253	○
funny	248	○
healthy	231	○
live	227	○
best	194	○
sweet	163	○
good	143	○
fit	143	○
hot	137	○
sunny	99	X

[표 3-9]는 인스타그램의 게시물 10,000건 중 감정 형용사 'Happy'에 해당하는 해시태그 총 30,936개의 단어에서 감정 키워드는 2634개의 단어가 추출되었다. 그 중 감정 키워드 필터링을 적용한 후 빈도수 내림차순에 의하여 게시물건의 0.01% 이상의 빈도수를 가지는 11개의 감정 키워드가 선정된 결과를 '○' 표시하여 보여주고 있다. 최상위 빈도수인 happy가 감정 키워드 선정에서 제외된 이유는 'Happy'라는 감정 형용사의 단어가 일치하기 때문이며 나머지 감정 키워드 선정의 경우에서도 이와 마찬가지다. 이처럼 감정 형용사와 감정 키워드 리스트가 일치하는 경우 감정 형용사를 주요 감정으로 선정하고 이에 대한 범주로 선정된 감정 키워드로 분류하였다. [표 3-10]은 제안하는 감정 형용사에 따른 해시태그 기반의 감정 카테고리를 나타낸다.

[표 3-10] 해시태그 기반 감정 카테고리

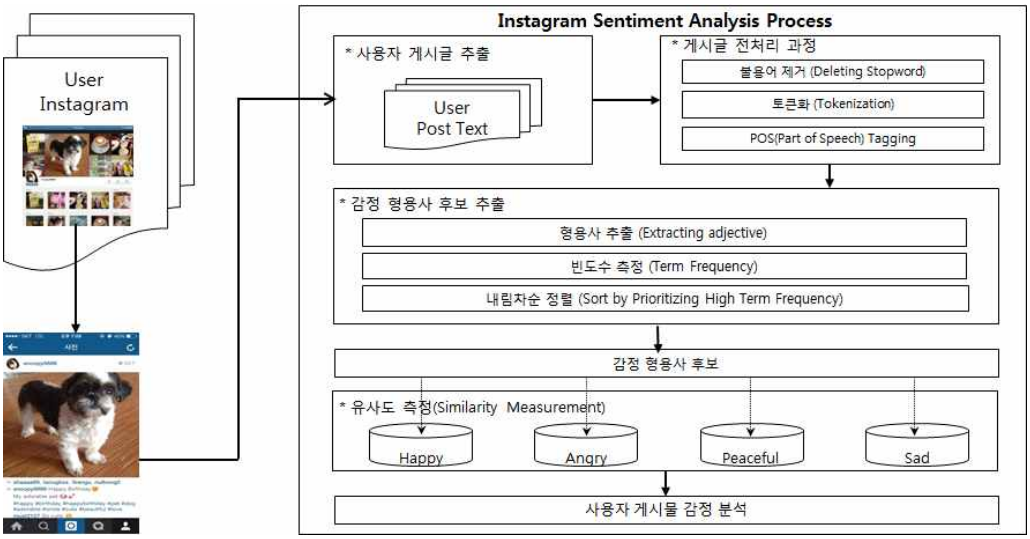
주요 감정	감정 키워드(Sentiment Keyword)
Happy	cute, beautiful, adorable, funny, healthy, live, best, sweet, good, fit, hot
Angry	sad, cute, black, funny, honest, happy, mad, suicidal, dark, skinny, upset, bad, white, unhappy, anorexic, thin, fat, dead, hungry, bulimic, sweet, stupid, young, bloody
Peaceful	beautiful, happy, green, blue, sunny, quiet, natural, colorful, cute, gorgeous, square, grateful, normal, sad, thankful, live, scenic, good, healthy
Sad	suicidal, fat, happy, anorexic, black, alternative, cute, dead, dark, pale, skinny, bulimic, rad, white, thin, true, unhappy, empty, sorry, emotional, bipolar, bad, aesthetic, stupid, funny, pathetic, anxious, mad, soft

제시하는 감정 카테고리는 최상위 감정 키워드를 주요 감정으로 선정하고 이에 대해 각각 선정된 감정 키워드를 볼 수 있다. 즉, 기존의 오피니언 마이닝에서 활용하고 있는 극성 분류와는 달리 심리학적으로 정의된 Thayer의 모델을 바탕으로 실제 인스타그램에서 공유되는 감정을 적용하기 위하여 인스타그램의 핵심적인 요소인 해시태그를 이용하여 주요 감정에 대한 세부적인 감정 키워드를 분류함으로써 인스타그램에서 주로 공유되는 감정을 파악할 수 있다. 다음 절에서는 감정 키워드를 핵심 단서로 이용하여 사용자의 감정을 분석하는 방법에 대해 다룬다.

C. 인스타그램 사용자 감정 분석 프로세스

인스타그램은 이미지와 동영상을 이용하는 스토리텔링 기법으로 많은 사용자들의 공감을 끌어내고 있으며 이를 마케팅의 수단으로 사용되기도 한다. 인스타그램 내에서 사용자들은 공유하고 싶은 게시물을 통해 다른 사용자들과 일상이나 의견, 정보 또는 감정 등을 공유하거나 댓글을 통해 공감하거나 공유할 수 있다.

본 절에서는 [그림 3-9]와 같이 앞서 제안한 해시태그 기반 감정 카테고리를 인스타그램 사용자 게시물에 적용하여 감정을 분석하는 방법에 대해 기술한다.

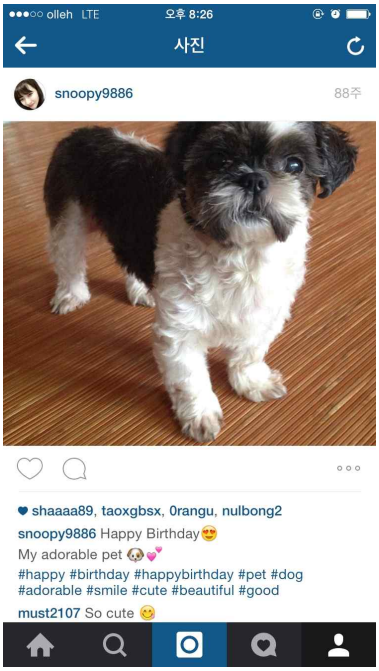


[그림 3-9] 인스타그램 감정 분석 프로세스

먼저 감정을 분석하기 위한 대상으로는 게시물 하나를 기준으로 두었을 때, 게시물에 있는 게시글과 댓글의 정보를 통해 사용자의 감정을 파악할 수 있다는 것을 전제로 인스타그램의 사용자 게시물에서 텍스트 정보만을 수집하여 감정 형용사 후보를 추출하여 정의한다. 이를 제안한 해시태그 기반의 각각의 감정 카테고리와의 유사도 측정을 통해 나온 수치 중 최댓값을 가지는 감정 카테고리로써 사용자의 감정을 분석한다.

1. 사용자 게시글 추출

[그림 3-10]는 [그림 3-9]의 인스타그램에 게시된 사용자의 게시물 중 하나를 예로 나타낸 것이며, 본 절에서 이를 예로 들어 설명한다.



[그림 3-10] 사용자 게시물

[그림 3-10]에서 볼 수 있듯이 게시물의 구성요소로 사용자의 아이디, 게시한 시간 정보, 이미지와 텍스트 정보, 다른 사용자가 올린 댓글 정보와 '좋아요'를 클릭한 정보 등이 존재한다. 하지만 본 절에서 감정을 분석하기 위해 사용될 게시물 중 객관적인 판단이 가능한 게시글 자료인 순수 사용자가 게시한 텍스트의 내용과 해시태그 정보, 댓글 정보를 데이터로 사용한다. 분석하는 기준은 사용자가 올린 게시물 하나를 기준으로 한다. 그 이유는 사용자들이 같은 감정으로 여러 개의 게시물을 게시하는 경우는 드물며 사용자의 게시물 전체나 부분적으로 데이터로 사용할 경우 그에 대한 데이터를 추출하는 데 있어서 기준뿐만 아니라 감정을 정의하는 데 모호한 문제가 있기 때문이다.

또한 각각의 게시물에 대한 시간 정보로 판단하였을 때에도 최근 게시물과 과거의 게시물에 담긴 감정이 지속되지 않기 때문이다. 따라서 각 하나의 게시물에 해당하는 주요 감정을 분석하기로 한다.

먼저, 사용자의 게시글을 추출하는 과정은 앞서 감정 형용사가 포함된 해시태그를 추출하는 과정과 유사하다. 달라진 점만 기술하자면 (1)에서 사용자의 게시물을 추출하기 위해서는 userId값의 고유번호가 필요한데 이는 랜덤으로 생성하였고, 비공개 사용자이거나 해당하는 userId 값의 계정이 없을 경우에는 접속이 되지 않고 그런 경우 재생성 후에 재접속을 하도록 하였다. 본 절에서는 [그림 3-10]의 userId 값을 이용하였다. (2)에서는 게시물의 정보에서 글의 내용과 태그정보, 댓글 정보를 포함하여 순수 텍스트 정보만 읽어오도록 하였다. (3)에서는 읽어온 정보를 한 줄씩 행 단위로 분리하여 Random_UserText.txt 파일에 저장하였다.

[그림 3-11] 임의의 사용자 게시글 추출

[그림 3-11]은 추출한 임의의 사용자 게시글의 텍스트 정보를 보여준다.

임의의 사용자의 게시물에서 10개의 게시물을 가져온 것이고 각 숫자는 해당 사용자의 최근 등록된 게시물의 순서를 나타낸다. 이때 숫자 옆의 텍스트 정보에서 사용자가 게시물을 업로드 할 때 최초 작성한 텍스트의 내용은 각 번호 옆에 '-'가 없는 행을 뜻하며 이때 해시태그의 정보도 담고 있다. '-' 와 함께 부여된 숫자는 해당 게시글의 댓글의 개수를 뜻하고 있으며 댓글의 정보와 함께 나타난다.

또한 해시태그 정보는 따로 '[ ]' 사이에 저장된 것을 볼 수 있는데, [그림 3-11]에서는 해시태그가 2번 중복되어 나오는 결과를 보여주고 있다.

따라서 사용자의 게시글을 추출하는 과정에서는 해시태그의 정보를 따로 가져오는 부분은 배제시켜 해시태그의 중복을 피하고자 한다.

다음 [표 3-11]은 [그림 3-10]의 게시글의 정보에서 해시태그의 정보를 가져오는 부분을 배제시켜 [그림 3-11]과 같이 추출된 데이터 형태로 나타낸 것이다.

[표 3-11] 추출된 게시글 정보

구분	원본 게시글
a	47-0 : So cute ??
b	47:::::Happy Birthday?? My adorable pet ???? #happy #birthday #happybirthday #pet #dog #adorable #smile #cute # beautiful #good

[표 3-11]은 [그림 3-10]의 userId 값을 이용하여 추출한 게시글 정보이며 b는 사용자가 이미지를 업로드할 때 최초로 작성한 텍스트를, a는 b에 대한 다른 사용자에 댓글을 뜻한다. 각각 텍스트 정보, 해시태그 정보, 댓글 정보가 포함된 것을 볼 수 있다. 이때 47이라는 숫자는 [그림 3-10] 사용자의 47번째의 게시물이라는 것을 뜻한다. 현재까지 추출된 부분은 사용자의 게시물에서 추출된 텍스트인 게시글 정보이며 이를 본 연구에 적용하기 위하여 다음 절의 전처리 과정을 거친다.

2. 게시글 전처리 과정

사용자의 게시글에서 감정 형용사 후보를 추출하기 위해 전처리 과정이 필요하다. 본 절에서는 (1) 불용어 제거(Deleting Stopword), (2) 토큰화(Tokenizing), (3) POS(Part of Speech) 태깅 단계 순으로 전처리 과정이 수행되며 앞서 감정 형용사가 포함된 해시태그의 감정 키워드를 추출하기 위한 전처리 과정의 각각의 내용은 같다. [표 3-12]는 사용자의 원본 게시글에서 각 단계에 대응하는 전처리 과정의 결과를 보여준다.

[표 3-12] 게시글 전처리 과정

원본 게시글 (Original Post)	47-0 : So cute ?? 47:::::Happy Birthday?? My adorable pet ???? #happy #birthday #happybirthday #pet #dog #adorable #smile #cute #beautiful #good
(1) 불용어 제거 (Deleting Stopword)	So cute Happy Birthday My adorable pet happy birthday happybirthday pet dog adorable smile cute beautiful good
(2) 토큰화 (Tokenization)	['So', 'cute'] ['Happy', 'Birthday'] ['My', 'adorable', 'pet'] ['happy', 'birthday', 'happybirthday', 'pet', 'dog', 'adorable', 'smile', 'cute', 'beautiful', 'good']
(3) POS 태깅 (Part of Speech Tagging)	[('So', 'IN'), ('cute', 'JJ')] [('Happy', 'JJ'), ('Birthday', 'NN')] [('My', 'PRP$'), ('adorable', 'JJ'), ('pet', 'NN')] [('happy', 'JJ'), ('birthday', 'NN'), ('happybirthday', 'NN'), ('pet', 'NN'), ('dog', 'NN'), ('adorable', 'JJ'), ('smile', 'NN'), ('cute', 'JJ'), ('beautiful', 'JJ'), ('good', 'JJ')]

3. 감정 형용사 후보 추출

전처리 과정을 거친 사용자의 게시글에서 감정 형용사 후보를 추출하기 위해 (1) [표 3-6]의 알고리즘을 이용하여 형용사 품사를 가진 단어들을 대상으로 (2) [표 3-7]의 빈도수(Term Frequency)를 측정하여 높은 순으로 내림차순 정렬한 뒤 선정된 감정 형용사 후보를 추출한다. 이때 분석하고자 하는 대상이 게시물 한 개이기 때문에, 빈도수에 따라 상위 5개까지를 감정 형용사 후보로 정의한다.

[표 3-13]은 선정된 감정 형용사 후보에 대한 (1), (2)의 과정과 결과를 보여준다.

[표 3-13] 감정 형용사 후보 추출

(1) 감정 형용사 (Extracting sentiment adjective)	cute happy adorable happy adorable cute beautiful good
(2) 빈도수 측정 (Term Frequency)	(2, 'happy') (2, 'cute') (2, 'adorable') (1, 'beautiful') (1, 'good')
감정 형용사 후보 (Sentiment adjective candidates)	happy, cute, adorable, beautiful, good

따라서 [그림 3-10] 사용자의 게시글에서 감정 형용사 후보로 'happy, cute, adorable, beautiful, good'이 추출된다.

4. 유사도 측정을 통한 사용자 게시물 감정 분석

본 절에서는 사용자의 게시글에서 추출한 감정 형용사 후보와 앞서 제안한 [표 3-10] 감정 카테고리와의 유사도를 측정함으로써 사용자의 감정을 분석하는 방법에 대해 기술한다. 유사도란 두 개체의 유사한 정도를 수치적으로 측정할 수 있는 척도를 말한다. 본 연구에서 사용자의 감정을 분석하기 위해 유사도를 측정하기 위한 이유는 분석하고자하는 사용자 게시글에서 추출한 감정 형용사 후보를 해시태그 기반으로 분류해놓은 감정 카테고리와 서로 어느 정도 유사하는지를 측정할 수 있는 기준이 필요하기 때문이다. 유사도 값을 측정하는 방법으로 코사인 유사도(Cosine Similarity)를 이용하였다. 코사인 유사도는 유사도 측정을 위한 대표적인 척도이며[33] 데이터 마이닝 분야에서 문서간의 유사도를 측정할 때 자주 사용된다.

식 (3)은 코사인 유사도를 계산하는 수식이며 이 값이 클수록 두 개체 사이의 유사도가 높다는 것을 말하고, -1에서 1사이의 값을 가지게 된다. 유사도의 값이 1일 경우는 서로 완전히 같은 경우를 의미하며, 0일 경우는 서로 독립적인 경우를, -1일 경우에는 서로 완전히 반대되는 경우를 뜻한다.

$$mialrity(A,B) = \cos(\theta) = \frac{A \cdot B}{A||B|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sum_{=1}^{n} A_i)^2 \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (3)$$

본 논문에서는 코사인 유사도가 문서나 텍스트에 적용될 경우 식 (3)의 벡터 A, B로 해당 텍스트의 단어 출현 빈도가 사용되는 특징을 가지고 있어 제안하는 감정 카테고리에 있는 감정 키워드와 추출된 감정 형용사 후보 각각의 단어 출현 빈도인 벡터 A, 벡터 B로 간주하여 유사도를 측정하기에 적합하다고 판단하였다.

따라서 식 (3)을 통하여 유사도의 값을 측정하고, 각각의 감정 카테고리의 감정 키워드와 유사도를 비교 측정하여 가장 높이 측정된 값을 최종적으로 사용자 게시물의 감정으로 판단하기로 한다. [표 3-14]는 식 (3)을 이용하여 유사도를 계산하는 과정을 보여주기 위해 (1) [표 3-10]의 감정 카테고리에서 Happy 카테고리의 감정 키워드와 (2) [그림 3-8]의 사용자의 게시글에서 추출된 감정 형용사 후보를 적용하여 계산한 과정을 나타낸 것이다.

유사도를 측정하는 과정으로는 먼저 측정하고자 하는 두 대상의 전체 단어를 식별하고 그 단어들의 출현 빈도를 파악해 벡터로 취급하여 벡터 A와 B를 생성하는 단계가 필요하다. 이를 식 (3)에 대입하여 코사인 유사도를 측정할 수 있다.

[표 3-14] 유사도 측정 과정

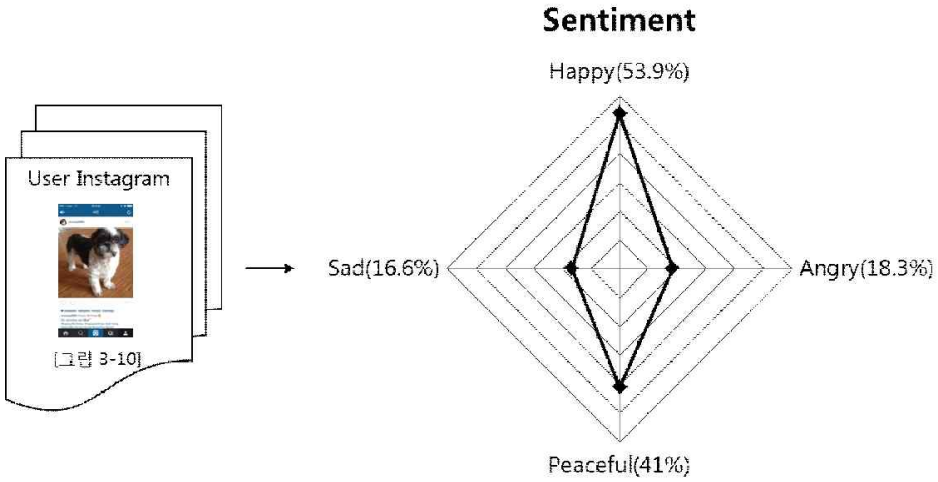
<table>
<tr><td>유사도
측정대상</td><td colspan="3">(1) Happy 카테고리의 감정 키워드
: cute beautiful adorable funny healthy
live best sweet good hit hot
(2) [그림 3-8]의 게시글에서 추출된 감정 형용사 후보
: happy cute adorable beautiful good</td></tr>
<tr><td>측정 대상의
전체 단어 식별</td><td colspan="3">happy cute adorable beautiful good
funny healthy live best sweet fit hot</td></tr>
<tr><td rowspan="15">단어 출현
빈도 파악</td><td>측정 대상 전체 단어</td><td>(1)의 빈도</td><td>(2)의 빈도</td></tr>
<tr><td>happy</td><td>0</td><td>1</td></tr>
<tr><td>cute</td><td>1</td><td>1</td></tr>
<tr><td>adorable</td><td>1</td><td>1</td></tr>
<tr><td>beautiful</td><td>1</td><td>1</td></tr>
<tr><td>good</td><td>1</td><td>1</td></tr>
<tr><td>funny</td><td>1</td><td>0</td></tr>
<tr><td>healthy</td><td>1</td><td>0</td></tr>
<tr><td>live</td><td>1</td><td>0</td></tr>
<tr><td>best</td><td>1</td><td>0</td></tr>
<tr><td>sweet</td><td>1</td><td>0</td></tr>
<tr><td>fit</td><td>1</td><td>0</td></tr>
<tr><td>hot</td><td>1</td><td>0</td></tr>
<tr><td>단어 출현 빈도를
통한 벡터 생성</td><td colspan="3">Vector A = {0,1,1,1,1,1,1,1,1,1,1,1}
Vector B = {1,1,1,1,1,0,0,0,0,0,0,0}</td></tr>
<tr><td>식 (3)에 대입</td><td colspan="3">$\sum_{i=1}^{n} B_i = 4$
$\sum_{i=1}^{n} A_i)^2 = 3.3166, \quad \sqrt{\sum_{i=1}^{n} (B_i)^2} = 2.2361$</td></tr>
<tr><td>유사도 값</td><td colspan="3">$Simialrity(A,B) = \cos(\theta) = \frac{4}{2.2361 \times 3.3166} = 0.5394$</td></tr>
</table>

[표 3-15]는 [그림 3-10]의 사용자의 게시글에서 추출된 감정 형용사 후보와 각각 감정 카테고리의 감정 키워드를 [표 3-14]와 같은 과정으로 유사도를 비교 측정한 결과이다. 유사도 값은 소수점 다섯 번째 자리에서 반올림하여 계산하였다.

[표 3-15] 감정 카테고리와 감정 형용사 후보 간 유사도 비교 측정

감정 카테고리(A)	감정 형용사 후보(B)	유사도(Similarity)
Happy 카테고리	happy cute adorable beautiful good	0.5394
Angry 카테고리		0.1826
Peaceful 카테고리		0.4104
Sad 카테고리		0.1661

[표 3-15]의 유사도 비교 측정의 결과를 보면 Happy 카테고리가 53.9%의 유사도, Angry 카테고리는 18.3%, Peaceful 카테고리는 41%, Sad 카테고리는 16.6%의 유사도를 보이고 있다. 이를 한눈에 알아볼 수 있도록 [그림 3-12]로 표현하였다.



[그림 3-12] 유사도 비교 측정을 통한 [그림 3-10]의 감정 분포도

[그림 3-12]를 통하여 한 게시물에 대한 네 가지의 감정이 분포되어 있는 것을 알 수 있다.

따라서 제안하는 네 가지의 감정 카테고리에 대한 각각의 유사도 값을 비교하여 가장 높게 나온 카테고리를 사용자 게시물의 주요 감정으로 판단한다.

[표 3-15]에서는 Happy 카테고리에 대한 유사도 값이 가장 높게 측정되었으므로 [그림 3-10]의 게시물에 해당하는 사용자의 주요 감정은 'Happy'라고 정의할 수 있다. 이를 통해 한 게시물 내에서도 크게 4가지의 감정들이 분포하지만 그 중에서도 한 가지의 주요 감정으로 판단할 수 있게 된다.

일반적인 문서 검색의 응용으로 문서간의 유사도를 구하기 위해 해당 문서를 분석하여 추출한 다수의 단어를 바탕으로 사전에 생성된 색인어와 유사도를 측정한다. 이에 본 논문에서는 인스타그램 내에서 사용되는 해시태그를 이용하여 제시한 감정 카테고리의 감정 키워드와 실제 인스타그램 사용자 게시글의 감정 형용사 후보 간의 유사도를 측정함으로써 감정 카테고리의 감정 키워드를 색인어로 사용하여 비교 대상이 적합하다고 판단되며, 감정 어휘를 나타내는 형용사 품사의 출현 빈도를 속성 값으로 측정한 유사도를 통해 해당 게시물의 사용자 감정을 판단할 수 있어 주관적인 감정에 대한 모호함을 객관적으로 해결할 수 있다.

Ⅳ. 실험 및 결과

본 장에서는 본 연구에서 데이터 수집을 위해 Java 기반의 Open API를 이용하여 구축한 데이터 셋인 학습 데이터 셋과 실험 데이터 셋에 대하여 설명한다.

또한 본 논문에서 제안한 해시태그 기반 감정 카테고리에 대한 효율성을 입증하기 위해 정확률(Precision)을 측정하여 성능을 평가한다. 실험은 PC상에서 python을 이용하여 구현하였다.

A. 데이터 수집

본 절에서는 본 연구에서 사용된 데이터인 감정 분류 프로세스에 필요한 해시태그 데이터와 임의의 사용자 게시글 데이터를 수집하는 부분에 대해 기술한다.

본 연구에서는 인스타그램에서 데이터를 수집하기 위해 Sachin Handiekar이 개발한 인스타그램의 Java 기반의 라이브러리인 jInstagram[34]을 이용하였다. 개발 환경으로는 JAVA JDK 1.8 버전을 사용하였고, Apache Maven 3.3.1 버전과 Eclipse Java EE IDE, LUNA(4.4.2)-M2E 버전을 연동하여 사용하였다.

[표 4-1] 해시태그 데이터 수집

```
                                        ⋮
public static void main(String[] args) {
InstagramSession is = new InstagramSession(new AccessToken(ACCESS_TOKEN));
try {
  BufferedWriter bw = new BufferedWriter(new FileWriter(tag+".txt"));
  PaginatedCollection<Media> media = is.getRecentMediaForTag(tag);
  int i = 1;
  for(Media _media: media) {
    String mg = _media.getTags().toString();
    bw.write(i+":"+mg);
    bw.newLine();
    if(i++ == 10000) {
      bw.flush();
      bw.close();
      break;
      } } } catch (Exception e) {
   e.printStackTrace();
   } } }
```

[표 4-1]은 제안하는 해시태그 기반 감정 카테고리의 학습 데이터로 사용된 해시태그 데이터를 수집하는 Java 코드이며 [그림 3-4]는 이에 대한 결과이다.

[표 4-2] 임의의 사용자 게시글 데이터 추출

```
                                    ⋮
    public static void main(String[] args) {
    InstagramSession is = new InstagramSession(new AccessToken(ACCESS_TOKEN));
    Random random = new Random();
    int userId = random.nextInt(1000000000);
      try {
        BufferedWriter bw = new BufferedWriter(new FileWriter("Random_UserText.txt"));
        PaginatedCollection<Media> feed = is.getRecentPublishedMedia(userId);
         while(feed.size() == 0){
           userId = random.nextInt(1000000000);
           feed = is.getRecentPublishedMedia(userId); }
    int i = 1;
    for(Media _feed: feed) {
     try{
       for(int j=0; j<_feed.getComments().size(); j++) {
         textFirstIndex=_feed.getComments().toString().indexOf(text, textLastIndex)+8;
         textLastIndex=_feed.getComments().toString().indexOf("\",", textFirstIndex);
         bw.write(i+"-"+j + " : ");
         bw.write(_feed.getComments().toString().substring(textFirstIndex, textLastIndex));
         bw.newLine(); }
        textFirstIndex = 0;
        textLastIndex = 0;
       } catch(Exception e){
            e.printStackTrace(); }
      try{
       bw.write(i+"::::::"+_feed.getCaption().getText());
       bw.newLine();
       bw.write(_feed.getTags().toString());
       bw.newLine();
       }catch(Exception e) {
         bw.write(i+"::::::");
         bw.newLine(); }
    if(i++ == 10) {
      bw.flush();
      bw.close();
      flag = true;
      break; } }
  if(!flag) {
    bw.flush();
    bw.close(); }
  } catch (Exception e) {
    e.printStackTrace();
  } } }
```

[표 4-2]는 제안한 해시태그 기반 감정 카테고리를 적용하여 사용자의 감정을 분석하는 방법에 있어 필요한 임의의 사용자 게시물 중 최근 등록된 10개 게시물을 기준으로 게시글 데이터를 추출하는 Java 코드이며, [그림 3-11]은 이에 대한 결과이다.

B. 데이터 셋

본 절에서는 본 연구에서 구축한 데이터 셋에 대하여 설명한다. 데이터 셋은 크게 제안한 방법에 사용된 학습 데이터 셋과 실험에 사용된 데이터 셋으로 구성된다.

1. 학습 데이터 셋 - 해시태그 기반

본 연구에서 해시태그 기반 감정 카테고리를 제시하기 위해 인스타그램의 게시물에서 해시태그만을 수집하여 [표 4-3]과 같이 학습 데이터 셋을 구축하였다.

[표 4-3] 학습 데이터 셋

해시태그	개수			
	게시물 건	총 단어	감정 키워드	감정 키워드 필터링 적용 후 선정된 감정 키워드
Happy	10,000	30,936	2,634	11
Angry	10,000	30,097	3,076	24
Peaceful	10,000	28,714	3,467	19
Sad	10,000	34,462	2,394	29
Total	40,000	124,209	11,571	83

학습에 사용된 인스타그램의 게시물은 총 40,000건에 해당하며, 분류 기준으로 선정한 카테고리의 주요 감정인 감정 형용사를 해시태그 기반 검색으로 10,000건씩 나누어 수집하였다. 전처리 과정을 거친 총 단어 124,209개에서 11,571개의 감정 키워드를 추출하여 필터링을 적용하여 선정된 83개의 감정 키워드를 감정 카테고리로 제시하였다.

2. 실험 데이터 셋 - 제안한 감정이 포함된 게시글

본 논문에서 실험에 사용된 인스타그램의 게시물은 Happy, Angry, Peaceful, Sad 네 개의 범주를 대상으로 제안한 감정 카테고리의 정확률 계산을 위해 번호를 부여하여 [표 4-4]와 같이 총 1,000건의 실험 데이터 셋을 구축하였다.

[표 4-4] 실험 데이터 셋

범주	번호	게시물 건
Happy	1 ~ 250	250
Angry	251 ~ 500	250
Peaceful	501 ~ 750	250
Sad	751 ~ 1000	250

[표 4-4]는 각 범주에 대한 게시물 건수를 나타내며 수집한 게시물에 순차적으로 번호를 부여하였다.

C. 실험 평가 방법 및 결과 분석

1. 실험 평가 방법

본 논문의 실험은 제안한 분류 방법인 해시태그 기반 감정 카테고리에 대한 효율성을 평가하기 위해 정확률(Precision)을 이용하였으며, 식 (4)와 같이 계산된다.

정확률은 제안한 감정 카테고리를 이용하여 분류한 결과의 정확성을 평가하는 것이며 본 논문에서는 제시한 감정 카테고리에 의해 분류된 게시물의 수 중 감정 카테고리에 의해 올바르게 분류된 게시물의 수로 Precision을 판단하였다.

$$\mathrm{Pr}\ cision = \frac{\text{감정 카테고리에 의해 올바르게 분류된 게시물의 수}}{\text{정 카테고리에 의해 분류된 게시물의 수}} \quad (4)$$

2. 실험 결과 분석

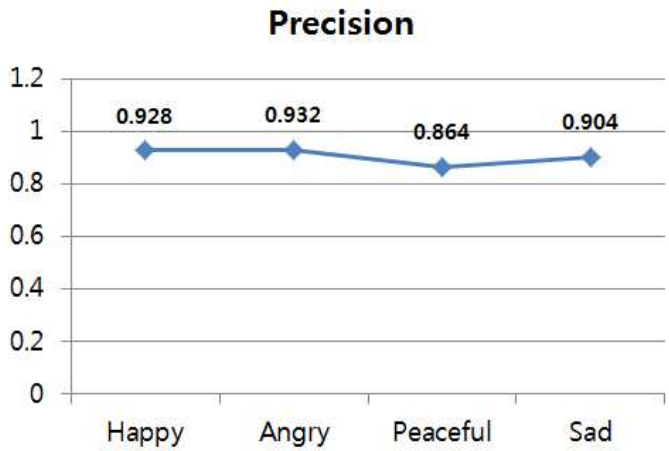
제안하는 해시태그 기반 감정 카테고리의 성능을 식 (4)의 정확률을 이용하여 평가한 결과는 다음 [표 4-5]와 같다.

[표 4-5] 제안한 감정 카테고리의 성능

구분	개수	올바르게 분류된 개수	Precision(%)
Happy	250	232	92.80
Angry	250	233	93.20
Peaceful	250	216	86.40
Sad	250	226	90.40
계	1,000	907	90.70

정확성 평가는 각 감정 카테고리별 정확률과 감정 카테고리에 대한 정답률을 평균화하여 도출하였다.

이를 [그림 4-1]과 같이 감정 카테고리별 정확률을 그래프로 표현하였다.

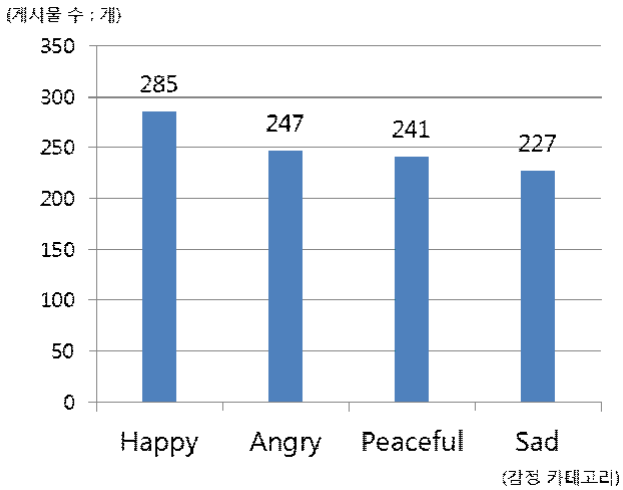


[그림 4-1] 감정 카테고리별 정확률

전체적으로 분류된 결과는 평균 90.7%의 정확성을 보여 제안한 감정 카테고리의 효율성을 입증하기에 만족할만한 결과를 보여주었다. 각 감정 카테고리별 결과에서는 Angry 카테고리가 가장 높은 정확률을 보였고, Happy 카테고리가 두 번째를 이은 것으로 보아 감정의 극성이 뚜렷할수록 높은 정확성을 보인다는 것을 알 수 있었다. 각각의 감정 카테고리들의 오분류된 결과를 분석해 본 결과 Happy 카테고리에서 올바르지 않게 분류된 정보는 대부분 Peaceful 카테고리에 해당되었으며, Peaceful 카테고리에서 올바르지 않게 분류된 정보는 대부분 Happy 카테고리에 해당하였다. 이를 통해 두 감정 카테고리가 어느 정도 상관관계가 있다는 것을 알 수 있었다. 그 중 Peaceful 카테고리에서 Happy 카테고리로 오분류된 정보에서는 Happy 카테고리의 감정 키워드인 'good'이라는 키워드가 크게 영향을 미친 것으로 분석되었다. Angry와 Sad 카테고리에서는 오분류된 결과를 분석한 결과 두 카테고리가 대부분 Happy 카테고리에 해당되었다. 이 두 감정 카테고리에서 오분류된 정보들은 대부분 사랑과 관련된 슬픈 글귀나 슬픈 노래가사 등에 자주 등장하는 지나간 사랑의 그리움, 아름다운 추억을 표현하는 'beautiful'과 'good'이라는 키워드가 크게 영향을 미친 것으로 작용하였다. 이러한 점들을 극복하기 위해서는 향후에는 제안하는 감정 카테고리의 감정 키워드를 선정하는 데 있어 'good'이나 'beautiful'과 같이 영향을 많이 미치는 키워드들에 대한 가중치를 부여하거나 더욱 세부적인

기준을 두어 오분류 사례를 줄이는 연구가 필요할 것으로 판단된다.

[그림 4-2]는 실험 데이터 셋에서 제안하는 감정 카테고리로 올바르게 분류된 감정 분포를 나타낸 그래프이다.



[그림 4-2] 실험 데이터 셋의 감정 분포

앞서 감정의 극성이 뚜렷할수록 높은 정확성을 보인만큼 [그림 4-2]에서도 감정의 극성이 뚜렷한 감정이 분포에 있어서도 큰 차이는 아니지만 비교적 Happy 카테고리에 대한 감정 분포가 가장 크며 Angry 카테고리가 그 뒤를 이었다. 이로써 본 논문에서는 감정 분류의 정확성 향상과 오분류율을 최소화하기 위해 Happy, Angry, Peaceful, Sad 4가지를 카테고리로 선정하여 분류기준을 세워 감정 카테고리를 제시하였고, 제시한 감정 카테고리에 대한 정확률을 측정하여 성능을 평가하였다. 향후에는 4가지 감정 외에 분류될 수 있는 대표 감정들을 추가하여 적용한다면 사용자의 감정을 보다 세밀하게 분석할 수 있는 연구가 될 것이라 생각한다.

이와 같이 본 연구에서 제시하는 감정 카테고리를 통해 사용자의 감정을 분석해 봄으로써 주관적인 감정에 대한 판단을 객관적으로 해결했다는데 큰 의미가 있다. 이를 통해 대용량의 감정 분류 체계가 갖춰진다면 SNS를 통한 주요 이슈나 사회적 현상 등 다양한 분야에 대한 감정 분석이 가능할 것으로 기대되며 나아가 SNS 상에서 사용자 맞춤형 서비스, 추천서비스 또는 감성마케팅 등으로 활용될 것으로 기대된다.

V. 결론 및 제언

본 논문에서는 인스타그램의 핵심적 요소인 해시태그를 이용하여 감정을 분류하여 주요 감정에 대한 감정 키워드를 추출하여 감정 카테고리를 제시하였고, 이를 사용자 게시물에 적용하여 게시글의 감정 형용사 후보와 감정 카테고리의 감정 키워드와의 유사도 측정을 통해 감정을 분석하는 방법을 제안하였다.

제안하는 방법의 특징으로는 오피니언 마이닝에서 활용하고 있는 극성 분류와는 달리 심리학적으로 정의된 Thayer의 모델을 기준으로 감정을 분류하였고 해시태그를 이용하여 감정 카테고리를 제안함으로써 실제 인스타그램에서 공유되는 감정을 적용하였다는 점이 있다. 또한 기존 감정 분석에 대한 연구의 경우 텍스트의 감정을 분류하기 위해 감정 사전을 이용하여 감정어휘에 대한 극성 값을 부여해 긍정/부정/중립의 감정으로 판단하는 것에 그쳤으나, 제안하는 감정 카테고리와의 유사도를 통해 감정 분포도와 사용자의 주요 감정을 분석할 수 있어 주관적인 감정에 대한 모호함을 객관적으로 해결할 수 있다. 제안 방법에 대한 실험 결과 전체 감정 카테고리에 대한 평균 정확률은 90.7%로 좋은 성능을 보였다. Angry 카테고리에서는 93.2%로 가장 높은 정확률을 보여주었고, Happy 카테고리는 92.8%, Sad 카테고리는 90.4%, Peaceful 카테고리는 86.4% 순으로 분석되었다.

본 연구는 감정 분류의 정확성 향상과 오분류율을 최소화하기 위해 대표 감정으로 Happy, Angry, Peaceful, Sad를 카테고리로 선정하여 감정 카테고리를 제시하였으나 향후에는 4가지 감정 외에 분류될 수 있는 감정을 추가로 선정하여 확장한다면 사용자의 감정을 보다 세밀하게 분석할 수 있는 연구가 될 것이라 생각된다.

본 논문의 향후 연구 방향으로는 감정 카테고리의 감정 키워드를 선정함에 있어 영향을 많이 미치는 키워드들에 대한 오분류율을 줄이는 연구가 필요하다.

또한 본 연구는 텍스트를 이용하여 추출한 감정 형용사만을 가지고 연구하였으나, 감정을 표현할 수 있는 이모티콘이나 감정을 품고 있는 다른 품사를 활용하는 방법 등의 확장된 연구로 이어져야 할 것이다. 제안하는 방법을 통해 대용량의 감정 분류 체계가 갖춰진다면 SNS를 통한 주요 이슈나 사회적 현상 등 다양한 분야에 대한 감정 분석이 가능할 것으로 기대되며 나아가 SNS상에서 사용자 맞춤형 서비스나 추천서비스 또는 감성마케팅 등으로 활용될 것으로 기대된다.

참고문헌

[1] 글로벌웹인덱스(GlobalWebIndex), https://www.globalwebindex.net/

[2] 박병선, “글로벌 SNS 이용 현황과 시사점,” 정보통신방송정책, 제 26권, 제 2호 통권 570호, pp. 22-34, 2014.

[3] J.I. Kim, D.J. Choi, B.K. Ko, E.J. Lee, P.K. Kim, "Extracting User Interests on Facebook," International Journal of Distributed Sensor Networks, Vol. 2014, pp. 1-5, 2014.

[4] ITWorld, http://www.itworld.co.kr/techlibrary/87940, "감정 분석의 이해," 2014.

[5] 구글 트렌드, http://www.google.com/trends/

[6] 이구형, “감성과 감정의 이해를 통한 감성의 체계적 측정 평가,” 한국감성과학회지, 한국감성과학회지, 제 1권, 제 1호, pp. 113-122, 1998.

[7] 학문명백과 : 복합학-감성과학-감성심리/생리, http://terms.naver.com/entry.nhn?docId=2098240&cid=44418&categoryId=44418

[8] 신현순, 함찬영, 엄남경, 김미경, 이석희, 김용선, “감성 ICT 기술 및 산업동향,” 전자통신동향분석, 제 29권, 제 5호, 2014.

[9] 박인조, 민경환, “한국어 감정단어의 목록 작성과 차원 탐색,” 한국심리학회지 : 사회 및 성격, 제 19권, 제 1호, pp. 109-129, 2005.

[10] 정찬섭, “감성과학의 심리학적 측면,” 한국감성과학회 학술발표대회자료, 제 1권, pp. 13-17, 1997.

[11] http://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[12] 이철성, 최동희, 김성순, 강재우, “한글 마이크로블로그 텍스트의 감정 분류 및 분석,” 정보과학회논문지:데이타베이스, 제40권, 제3호, pp. 159-167, 2013.

[13] 장문수, “심리학적 감정과 소셜 웹 자료를 이용한 감성의 실증적 분류,” 한국지능시스템학회 논문지, 제 22권, 제 5호, pp. 563-569, 2012.

[14] 김원상, 이종혁, 박제원, 최재현, “오피니언 마이닝을 통한 정당지지도 분석 기법,” 한국정보기술학회논문지, 제 12권, 제 10호, pp.133-141, 2014.

[15] K. D. Rosa, R. Shah, B. Lin, A. Gershman, and R. Frederking, "Topical

Clustering of Tweets," Proceeding of the ACM SIGIR:SWSM, 2011.

[16] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, "Topic Sentiment Analysis in Twitter:A Graph-based Hashtag Sentiment Classification Approach," Proceedings of the 20th ACM international conference on Information and Knowledge management(CIKM), ACM, pp. 1031-1040, 2011.

[17] N. Gunawardena, J. Plumb, N. Xiao, and H. Zhang, "Instagram Hashtag Sentiment Analysis," 2013.

[18] D. M. McNair, M. Lorr and L.F. Droppleman, "Manual for the Profile of Mood States," Educational and Industrial Testing Services, San Diego, 1971.

[19] R. Plutchik, "Emotions and Life : Perspectives From Psychology, Biology, and Evolution," American psychological Association, 2003.

[20] R. Thayer, "The Biopsychology of Mood and Arousal," Oxford University Press, 1989.

[21] 이홍석, "위치기반 서비스의 감성 적용에 관한 연구," 건국대학교, 석사학위논문, 2013.

[22] 최경훈, "문맥을 고려한 한국어 극성 기반 감정 분류 방법," 가톨릭대학교, 석사학위논문, 2013.

[23] 조상현, 강행봉, "형식적 및 비형식적 어휘 정보를 반영한 문장 감정 분류," 한국정보처리학회논문지, 제 18권, 제 5호, pp. 325-332, 2011.

[24] Turney and M. Littman, "Measuring praise and criticism: Inference of semantic orientation from association", Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics, pp. 417-424, July 2002.

[25] 임지연, "품사 패턴 기반 특징명사에 따른 감정어 추출 및 극성 분류," 숭실대학교, 석사학위논문, 2014.

[26] 서정열, 고찬, "제품 특징을 이용한 한국어 어휘의 극성 분류," 한국정보기술학회논문지, 제 12권, 제 12호, pp. 115-123, 2014.

[27] C. Strapparava and A. Valitutti, "Wordnet-affect : an affective extension of wordnet", In Proceedings of the 4th International Conference on Language Resources and Evaluation(LREC), pp. 1083-1086, May 2004.

[28] A. Esuli and F. Sebastiani, "SentiWordNet 3.0 : An Enhanced Lexical

Resource for Sentiment Analysis and Opinion Mining", Proceedings of the International Conference on Language Resources and Evaluation(LREC), May 2010.

[29] 강인수, "영어 트위터 감성 분석을 위한 SentiWordNet 활용 기법 비교," 한국 지능시스템학회 논문지, 제 23권, 제 4호, pp. 317-324, 2013.

[30] 서지훈, 조혜진, 최진탁, "한국 문법의 반의어 규칙을 적용한 오피니언 감성사전 설계," 한국정보기술학회논문지, 제 13권, 제 2호, pp. 109-117, 2015.

[31] https://instagram.com/developer/

[32] http://www.nltk.org/

[33] 권응주, 김종우, 허노정, 강상길, "소셜 네트워크에서 감정단어의 단계별 코사인 유사도 기법을 이용한 추천시스템," 정보기술아키텍처연구, 제 9권, 제 3호, pp. 333-344, 2012.

[34] http://jinstagram.org/