[UCI]I804:24011-200000265217

２０１６年 2月
석사학위논문

SNS 특징정보를 활용한 단문텍스트의 카테고리 분류에 관한 연구

조 선 대 학 교 대 학 원

소프트웨어융합공학과

나 성 희

2016년 2월

석사학위논문

SNS 특징정보를 활용한 단문텍스트의 카테고리 분류에 관한 연구

나 성 희

SNS 특징정보를 활용한 단문텍스트의 카테고리 분류에 관한 연구

A study on the categorization of the short text using the SNS feature informations

2016년 2월 25일

조 선 대 학 교 대 학 원

소프트웨어융합공학과

나 성 희

SNS 특징정보를 활용한 단문텍스트의 카테고리 분류에 관한 연구

지도교수 김 판 구

이 논문을 공학석사학위신청 논문으로 제출함.

2015년 10월

조 선 대 학 교 대 학 원

소프트웨어융합공학과

나 성 희

나성희의 석사학위논문을 인준함

위원장 조선대학교 교수 반 성 범 (인)

위 원 조선대학교 교수 이 민 창 (인)

위 원 조선대학교 교수 김 판 구 (인)

2015년 11월

조 선 대 학 교 대 학 원

목 차

표 목 차

그 림 목 차

ABSTRACT

A study on the categorization of the short text using the SNS feature informations

Sunghee NA
Advisor : Prof. Pankoo Kim Ph.D.
Department of SoftWare Convergence Engineering
Graduate School of Chosun University

Recently, developments in wireless internet and the spread of mobile devices, social networking services (SNS) are rapidly progressing. SNS allows the user to place one's offline social connections onto an online platform, which in turn, strengthens and widens the user's personal network. As the number of SNS users rapidly increases, SNS is being transformed into a platform that allows the user to share one's interests with an unspecified mass, which, in the process, is changing the objective of SNSs from strengthening user's personal networks into an informational platform that freely shares interests or information.

Due to some characteristics of SNS, including the real-time generation and rapid spread of data, there has been a phenomenon of decreased efficiency in information acquisition using search engines, largely because of the overload of available information. To alleviate the problem of information overload, there have been experiments that explore the realm of document sorting, search engines, text translation, spam mail filtering, and many others. Data that generated on the internet or by social networking do not have a consistent standard, and because the contents of said data include short

sentences, it is hard to sort the data based just on the content of the document.

Therefore, this dissertation attempted to solve the phenomenon of overload of information due to the large amount of data generated real time, and to separate the texts used on the internet and other SNSs for effective data management. Short texts were collected from Twitter, which most often displays sentences in short formats. The collected short texts were analyzed to extract certain features. Words were weighted based on the extracted features to examine the correlation between the collected dataset and unsorted data, and the short texts were categorized. From the results of the experiment proposed in this dissertation, it was found that text categorization showed better performance when features were weighted using feature frequencies. Result of short text classification utilizing the feature frequency , it shows the 90.7 percent .

Ⅰ. 서론

A. 연구 배경 및 목적

최근 무선인터넷의 발달과 모바일 기기의 보급으로 인해 다양한 소셜 네트워크 서비스(SNS, Social Network Service)들이 발달하고 있다. 소셜 네트워크 서비스란, 사용자의 오프라인 인맥을 온라인으로 옮겨와 사용자의 인맥을 강화해주는 서비스 플랫폼을 의미한다[1]. 소셜 네트워크 서비스는 인터넷이 발달함에 따라 이용하는 사용자가 계속해서 증가하고 있다[1]. 소셜 네트워크 서비스는 미니홈피와 블로그와 같은 형태로 나타났으나, 인터넷과 모바일기기의 발달로 인해 단문 형태로 메시지를 주고받는 형식의 트위터(Twitter)와 페이스북(Facebook)의 소셜 네트워크 서비스가 활발히 진행되고 있다. 인터넷과 모바일 기기의 발달이 활발하지 않을 때의 소셜 네트워크 서비스는 인맥 강화를 중점으로 다른 사용자들에게 안부를 묻는 글이 많이 작성되었다. 그러나 모바일기기의 보급과 무선인터넷의 발달 이후에는 불특정 다수와 인맥을 맺게 되면서 웹상에서는 다양한 주제를 바탕으로 이야기를 나눌 수 있게 되었고[2], 자신의 관심사 및 정보를 공유하는 글이 작성되게 되었다. 이는 소셜 네트워크 서비스의 발달과 함께 소셜 네트워크 서비스의 목적이 인맥강화 중심에서 개인의 관심사 및 정보 공유로 변화함을 나타내고 있다. 단문 형태의 소셜 네트워크 서비스는 빠른 데이터 재생산성과, 빠른 확산성[3]으로 실시간으로 생성되는 데이터의 양이 증가하게 되었다. 실시간으로 생성되는 데이터의 양이 증가함에 따라 업무 처리 시 검색의 효율성을 저하하는 정보의 과부하 현상을 초래하였다[4]. 정보의 과부하 현상을 해결하기 위해서는 사용자 개인의 시간과 노력이 필요하다[5]. 따라서 실시간으로 생성되는 데이터들에 대한 효율적 처리, 분석, 관리 하는 연구에 대한 관심이 높아지고 있다.

본 연구에서는 실시간으로 생성되는 대량의 데이터들로 인한 정보의 과부하 현상을 해결하기 위해 데이터들의 효율적 처리, 분석, 관리를 위해 단문텍스트를 카테고리별로 분류하고자 한다. 기존의 문서분류 방법은 단어의 빈도수를 이용하는 방법이 주로 이용되었으나 단문의 경우 나타나는 단어의 수가 한정이기 때문에 기존의 방법을 이용하기에 어려움이 있다. 따라서 단문을 이용하는 소셜 네트워크 서

비스인 트위터를 이용하여 데이터를 수집하고, 기계학습을 이용하여 단문텍스트가 가지는 특징을 바탕으로 가중치를 부여하여 단문텍스트를 카테고리별로 분류하는 것을 돕는다. 특징을 활용하여 단문텍스트를 카테고리별로 분류한다면, 소셜 네트워크 서비스 또는 웹상에서 단문텍스트를 분류하는 데 도움을 주어 정보의 효율적 처리와 관리를 가능케 하여 다양한 소셜 검색 서비스에서 활용할 수 있을 것이다.

B. 연구 내용 및 구성

본 연구의 주된 내용은 단문텍스트의 특징들을 기반으로 특징 가중치를 부여하고 단문텍스트 간의 상관성 분석을 통해 단문텍스트를 카테고리별로 분류하는 것이다. 본 논문의 구성은 다음과 같이 구성되어 있다.

본 장인 서론에 이어 2장 관련 연구에서는 기존의 문서분류와 단문텍스트 분류를 위한 기존 연구들을 통해 본 연구의 이해를 돕는다.

3장에서는 SNS의 특징들을 활용하여 단문텍스트를 카테고리별로 분류하기 위한 카테고리 선정방법과 특징 가중치 값을 구하는 방법을 제시하고 상관성 분석을 통해 단문텍스트를 카테고리별로 분류하는 방법을 제시한다.

4장에서는 제안하는 방법의 실험을 위한 데이터수집방법과 이용하는 학습데이터셋과 실험데이터 셋에 대해 설명한다. 또한, 단문텍스트의 분류방법에 대한 정확도를 측정하여 성능을 평가한다.

마지막으로 5장에서는 본 연구에 대한 전체적인 결과와 향후 연구를 제시하며 마무리한다.

Ⅱ. 관련 연구

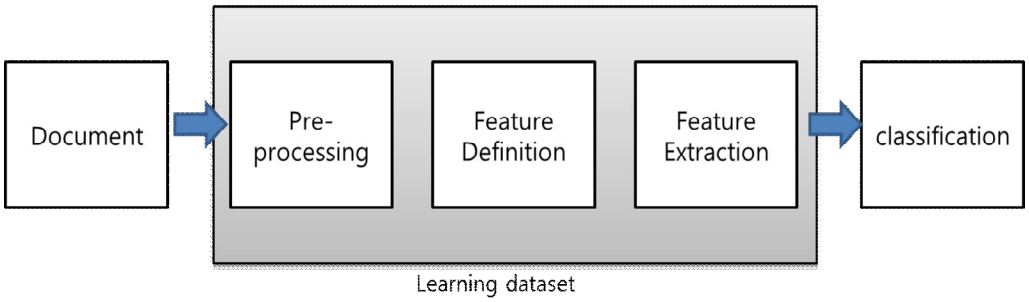
A. 문서분류(Document Classification)

문서분류(Document Classification)의 정의는 “문서의 내용을 바탕으로 미리 정해진 범주에 문서를 분류하는 것”으로 나타낸다. 문서분류는 문서의 유사성을 비교하여 비슷한 문서의 집단을 형성하는 방법과 분류체계를 가지고 있는 데이터를 적합한 범주에서 분류하여 문서의 집단을 형성하는 방법 두 가지로 나누어진다[6]. 기존의 문서분류 방법은 문서의 내용을 읽고 전문가의 견해에 따라 수작업으로 분류하여 관리하는 방법을 이용하였다. 그러나 인터넷이 발달하면서 기존의 문서가 전산화되고, 인터넷에서 사용하는 웹 문서들이 증가하게 되었다. 웹 문서는 실시간으로 많은 양이 생성되어 기존의 방법인 모든 문서를 읽고 수작업으로 분류하기에 많은 시간과 비용이 소요되었고, 전문가의 견해만으로는 대량의 문서를 객관적으로 분류할 수 없다는 단점이 나타나게 되었다. 이후 실시간으로 생성되는 대량의 문서들을 보다 효율적이고 객관적으로 분류하기 위해 자동 문서분류방법이 나타나게 되었다[7, 8].

자동 문서분류(Automatic document classification)는 기계학습 (Machine Learning) 알고리즘을 이용하여 문서를 분류하는 것을 나타낸다. 기계학습은 문서분류 시 주로 사용되는 알고리즘으로 로봇이나 프로그램 등의 과거 경험을 바탕으로 주어진 문제에 대해 최적의 의사결정을 끌어내는 방법론이다[9]. 1980년부터 연구 분야의 한 부분으로 자리 잡기 시작한 문서분류는 서비스업, 제조업, 정보검색 등 여러 분야에서 핵심기술로 자리 잡고 있으며, 단어의 빈도수와 분류기를 활용하여 문서를 분류하는 방법 등이 연구되고 있다[10]. 기계학습은 입력 데이터를 받아들여 문제해결모델을 만드는 방식에 따라 크게 지도학습(Supervised Learning)과 비지도 학습(Unsupervised Learning) 두 가지 방법으로 구분하여 설명할 수 있다[30].

지도학습은 소속과 범주를 가진 데이터들을 학습데이터로 이용하여 범주의 특성을 학습하여 학습 데이터 셋을 구축하여 데이터를 그룹화하는 방법을 나타낸다. 비지도 학습은 지도학습과 달리 학습 데이터 셋을 따로 구축하지 않고 수집한 데이

터들의 특징을 기준으로 데이터를 그룹화하는 방법으로 사용자에 의해 데이터의 범주가 부여되지 않는 특징을 가지고 있다[11]. 본 논문에서는 기계학습 알고리즘 중 학습 데이터 셋을 구축하는 지도학습 방법을 이용한다.



[그림 2-1] 지도학습 문서분류 프로세스

[그림 2-1]은 지도학습 방법을 이용한 문서분류 프로세스이다. 지도학습을 이용한 문서분류는 학습데이터를 바탕으로 미분류 데이터를 분류한다. 문서분류는 주로 뉴스의 분야를 나누는 것, 스팸메일 분류, 논문의 연구 분야를 나누는 등을 나타내기 때문에 문서의 내용을 이해하는 것이 중요하다[32].

지도학습 방법의 문서분류는 데이터를 수집하고 전처리과정에서 문서분류 시 불필요한 숫자, 특수문자, 관사, 조사 등의 문자를 제거하고 형태소를 분석하는 등의 작업을 거친다. 전처리 과정을 거친 문서들에 대한 특징을 정의하고 특징을 추출하여 학습데이터 셋을 구축한다. 학습데이터 셋 구축 뒤에는 미분류 데이터들과의 분류를 위해 사용되는 분류알고리즘의 종류는 Naïve Bayes, SVM(Support Vector Machine), Decision Tree(결정 트리), K-NN(K-Nearest Neighbor) 등을 적용하거나 TF-IDF방법을 이용하여 문서를 분류한다.

문서분류 프로세스의 가장 큰 특징은 문서가 가지고 있는 특징을 찾는 것이다. 특징을 이용한 문서분류방법을 살펴보면 [28]과 같이 특징을 선택하는 방법을 찾기 위한 연구가 있다. [28]은 Retures-22173과 OHSUMED 문서 데이터를 이용하여 DF(Document Frequency), IG(Information Gain), MI(MutualInformation), CHI(ChiSquare Statistics), TS(Term Strenght) 방법 중 텍스트 문서에 적합한 특징 선택방법을 찾아내는 특징에 관련된 연구가 진행되었다. 또한, 추출한 특징의 값을 주기 위해서는 단어빈도수(TF, Term Frequency), 역문헌 빈도수(IDF,

Inverse Document Frequency) 등의 용어 가중치를 적용할 수 있다[7].

TF- IDF 방법은 문서분류 시 이용되는 방법으로 TF는 단어의 빈도수를 나타내고 IDF는 역 문서 빈도로써 특정단어를 포함하는 문서의 수를 나타내어 여러 문서로 이루어진 문서 군에서 단어의 중요도는 나타내는 방법이다[12]. 그러나 무선 인터넷과 모바일 기기의 보급으로 인해 소셜 네트워크 서비스가 활성화 되었고, 단문으로 자신의 관심사 및 정보를 공유하는 단문형태의 데이터들이 생성되었다. 소셜 네트워크에서 사용되는 단문 형태의 데이터들은 빠른 속도로 재생산되고, 재생산된 데이터들의 확산속도가 빠르다는 장점을 가지고 있어, 단문으로 이루어지는 데이터들의 양이 증가하게 되었다. 각각의 하나의 문서로 생각할 수 있는 단문텍스트는 웹상에서 사용되는 신조어, 특수문자, 이모티콘 등의 사용으로 단문텍스트 자체에서 얻을 수 있는 단어의 수가 적다. 그로인해 기존의 문서분류 방법에서 사용하는 방법을 적용하기에는 어려움이 있다.

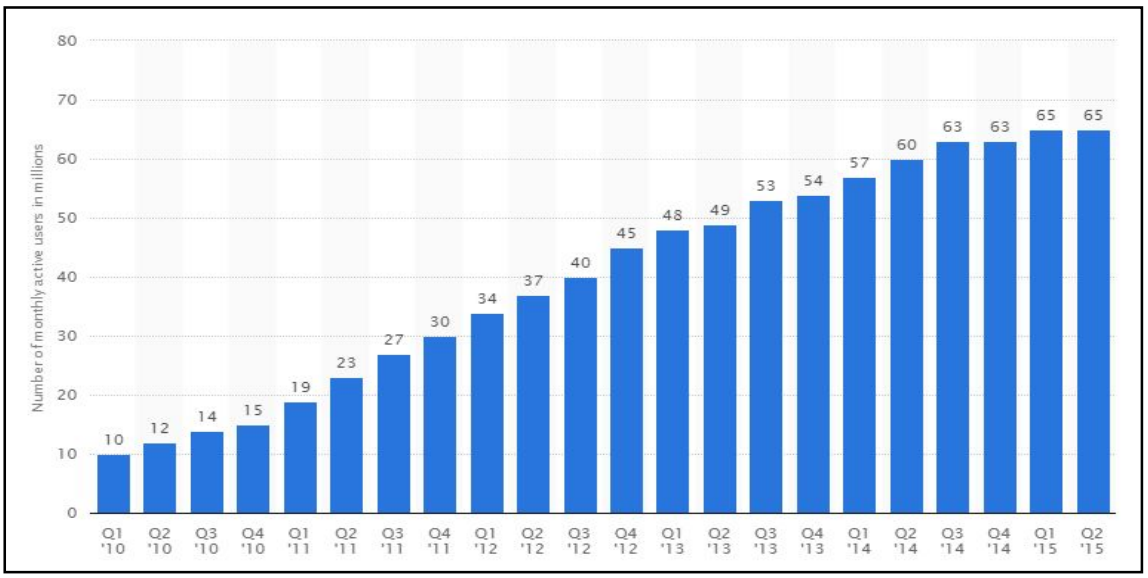
B. 단문텍스트 문서분류

소셜 네트워크 서비스는 사용자의 사회적 관계망을 생성, 유지, 강화해주는 플랫폼으로 블로그나 미니홈피 등의 형태로 나타났다. 이후 무선인터넷과 모바일 기기의 보급이 활발해지며 단문을 통해 다수의 사용자와 메시지를 주고받는 형태의 소셜 네트워크 서비스들이 나타나게 되었다. 소셜 네트워크 서비스는 프로필 기반, 비즈니스 기반, 블로그 기반, 버티컬 기반, 협업기반을 바탕으로 구분할 수 있다[2]. 그중 프로필 기반의 소셜 네트워크 서비스는 자신의 프로필을 간략히 작성하여 개인정보를 많이 드러내지 않고, 쉽게 이용할 수 있어 다수의 사용자가 이용하고 있다. 기존의 소셜 네트워크 서비스는 사용자의 오프라인 인맥을 온라인으로 옮겨와 인맥을 강화해주는 목적으로 이용되었다. 그러나 소셜 네트워크 서비스가 다수의 사용자와의 메시지를 주고받을 수 있게 되면서 소셜 네트워크의 이용 목적이 인맥 강화를 넘어 자신의 관심사 및 다양한 정보를 공유하는 것으로 변화하게 되었다. 소셜 네트워크 서비스의 발달과 목적의 변화로 채팅 메시지, 제품 설명서 및 안부 글 등 여러 가지 단문의 형태를 가진 텍스트들이 실시간으로 생성되며 단문텍스트들이 증가하게 되었다. 실시간으로 생성되는 단문텍스트들을 효율적으로 이용하고 관리하기 위해 문서분류의 방법이 이용된다. 그러나 단문텍스트는 문서의 내용이 짧고 추출할 수 있는 단어의 수도 적어 기존의 문서분류 방법을 적용하는 데에는 어려움이 있다. 따라서 단문텍스트를 분류할 때 단문텍스트가 가지는 특징을 이용하여 단문텍스트를 분류한다면 분류하는 데 도움을 줄 수 있을 것으로 생각된다. 본 논문에서는 단문텍스트의 카테고리 분류를 위해 사용자들이 많이 사용하는 소셜 네트워크들 중 서비스를 시작한 이후로 사용자들이 꾸준히 이용하고 단문텍스트를 가장 잘 나타내는 소셜 네트워크 서비스인 트위터를 대상으로 하였다.

단문텍스트를 이용하는 대표적 소셜 네트워크 서비스인 트위터는 일상의 순간을 기록하기 위한 서비스 플랫폼으로 2006년 서비스를 시작했다. 트위터는 다양한 주제의 단문텍스트가 하루 4억 건 이상 생성되고 있다[3,13,14]. 트위터는 다른 소셜 네트워크 서비스와 달리 140자 글자 수를 제한하는 특징이 있다. 이는 단문텍스트의 간결함을 부여하고 효율적으로 단어를 사용하게 하는 것이다[15]. 이 외에 트위터는 단방향 구조로 되어있어, 사용자가 여러 사용자에게 이야기를 전달할 수 있는 구조로, 사용자가 전하고자 하는 이야기, 정보 등을 빠르게 전달할 수 있다

[16,17].

트위터에서 작성되는 글은 글을 작성하는 형태에 따라 명칭이 모두 다르다. 사용자가 직접 트위터에 작성하는 단문텍스트는 트윗(Tweet)이라고 하며, 사용자가 작성한 트위터의 트윗은 사용자가 따르는 사용자들의 트윗들과 함께 모아 볼 수 있으며 볼 수 있는 공간을 타임라인이라고 한다. 또한, 관심 있는 사용자를 따른다고 하여 관심 있는 사용자의 트윗을 받아볼 수 있는 것을 “팔로우(Follow)한다”고 표현한다[17]. 트위터에서 단문텍스트가 전달되는 것은 리트윗(Retweet)이라고 한다. 리트윗은 사용자가 긴급정보 및 유용한 정보를 여러 사람에게 알리고자 할 때 이용한다. 리트윗은 트위터 내에서 강력한 위력을 지니고 있다. 2007년 캘리포니아에서 산불이 났을 때와 2009년 유에스 에어웨이 비행기가 불시착했을 때, 트위터의 리트윗 기능을 통해 어느 뉴스 매체보다 빠르고 정확한 소식을 전해준 일화가 있다[16]. 이는 소셜 네트워크 서비스가 단순 인맥 강화를 위한 서비스를 넘어 정보전달의 역할을 하는 것을 나타낸다. 이에 소셜 네트워크 서비스의 데이터의 흐름을 분석하고 효과적인 데이터 분석방법에 대한 연구가 진행되었다[18,19].



[그림 2-2] 2010부터 2015 2분기까지의 활성 트위터 이용자수

[그림 2-2]는 2010년부터 2015년 2/4분기까지의 미국의 매달 활성 트위터 이용자 수를 나타내고 있다[15]. 2010년 1/4분기에는 천만 명이었던 트위터 이용자 수

는 점차 증가하여 2015년 2/4분기에는 미국 내 트위터를 이용하는 사용자는 6천5백만 명으로 나타났다. 소셜 네트워크 서비스를 이용하는 사용자가 증가함에 따라 이를 이용하여 효과적인 자질 추출에 관한 연구, 사용자의 감정을 분석하는 연구들, 데이터의 효율적인 분석, 분류에 관한 연구들이 활발하게 진행되고 있다[5,11,14,20,21].

[그림 2-3] 소셜 네트워크서비스의 특징이 나타나는 단문텍스트

[그림 2-3]은 트위터의 단문텍스트 일부이다. 트위터의 단문텍스트 구조는 사용자 정보를 알리는 부분과 사용자가 쓴 글 부분으로 나눌 수 있다. 사용자 정보를 알리는 부분에서는 사용자의 아이디와 닉네임을 확인할 수 있다. 사용자의 아이디를 나타낼 때는 '@' 기호와 함께 사용자 아이디(identity)를 적어 '@identity'형식으로 나타낸다. [그림 2-3]에서 굵게 표시되어 사용자의 닉네임을 나타내는 부분과 @RolliingStone, @Discovery, @natgeowild와 같이 사용자의 아이디를 나타내는 것을 확인할 수 있다. 사용자의 정보를 나타내는 부분 이외에 사용자가 작성한글을 나타내는 부분에서는 본문, 해시태그(HashTag, '#')와 URL(Uniform Resource Locator)이 나타나는 것을 확인할 수 있다. 해시태그(HashTag)는 '#'기호와 함께 특정 단어를 적는 것으로 최초의 목적은 소셜 네트워크 서비스에서 검색을 편리하게 하기 위한 기능이었다[22]. 그러나 최근 들어 해시태그는 검색의 편리함을 넘어 자신의 관심사를 나타내고 정보를 공유하는 용도로 이용되고 있다. URL은 중요

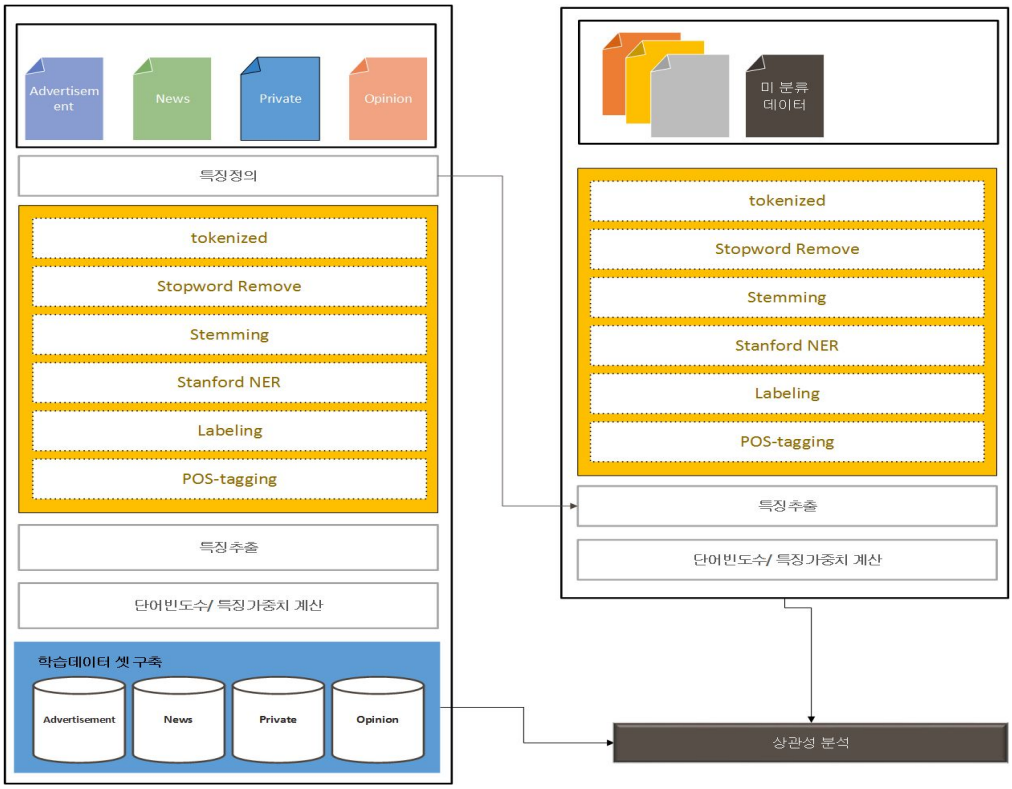
한 내용만을 전달하는 단문텍스트의 문제점을 보완하기 위해 사용된다. 그러나 URL의 길이로 인해 단문텍스트의 내용을 전달할 수 없는 것을 방지하기 위해 URL-Shortening 기능이라고 하여 URL 길이를 'rol.st/1W2qCul'과 같이 줄여서 작성하는 기능을 이용하여 표현하고 있다. 단문텍스트가 가지는 특징들은 해시태그, 사용자 아이디, URL 특징 이외에 단문텍스트의 주제에 따라 다른 특징을 나타낼 수 있다.

Ⅲ. 단문텍스트의 특징을 이용한 카테고리 분류방법

본 논문에서는 단문텍스트의 특징을 정의하고 추출하여 단문텍스트의 단어빈도수에 SNS에서 나타나는 특징들을 활용하여 특징 가중치라는 값을 구하여 학습데이터 셋을 구축한 뒤, 미분류 데이터와의 상관성 분석을 통해 단문텍스트를 카테고리별로 분류하는 방법을 제안한다.

A. 시스템 구성도

[그림 3-1]은 제안하는 단문텍스트의 특징을 고려한 단문텍스트 분류의 전체 시스템 구성도이다. 크게 학습 데이터 셋을 구축하는 단계와 미분류 데이터를 학습데이터 셋 기반으로 분류하는 단계로 나뉜다.



[그림 3-1] 시스템 구성도

시스템 구성도는 학습데이터 셋 구축을 위해 단문텍스트를 수집한 뒤, 단문텍스트가 가지고 있는 특징들을 정의한 뒤 전처리과정인 토큰화, 불용어 제거, 스테밍, NER 형태소 분석 과정을 거쳐 특징을 추출한다. 또한, 특징 외에 형태소 분석을 통해서 명사와 동사를 추출하여 단어의 빈도수와 단어가 등장할 때 함께 나타나는 특징빈도수를 계산하고 특징이 가중치로의 역할을 할 수 있는 특징 가중치를 계산하여 학습데이터 셋을 구축한다.

미분류 데이터 셋 역시 학습데이터 셋과 같은 방법을 통해 앞서 정의한 특징과 단어빈도수, 특징 가중치를 추출하기 위해 전처리과정을 거친 뒤 학습데이터 셋과 상관성 분석을 통해 미분류 데이터를 카테고리에 분류하게 된다.

B. 카테고리 선정

학습 데이터 셋을 구축하기 위해서는 학습 데이터를 수집할 카테고리를 선정해야 한다. 본 연구에서 카테고리를 선정한 방법은 다음과 같다. 기존의 단문텍스트 분류를 하는 선행연구들을 바탕으로 각 논문에서 사용되는 카테고리들을 분석하여 카테고리를 선정하였다[32-46].

[표 3-1] 선행연구 카테고리

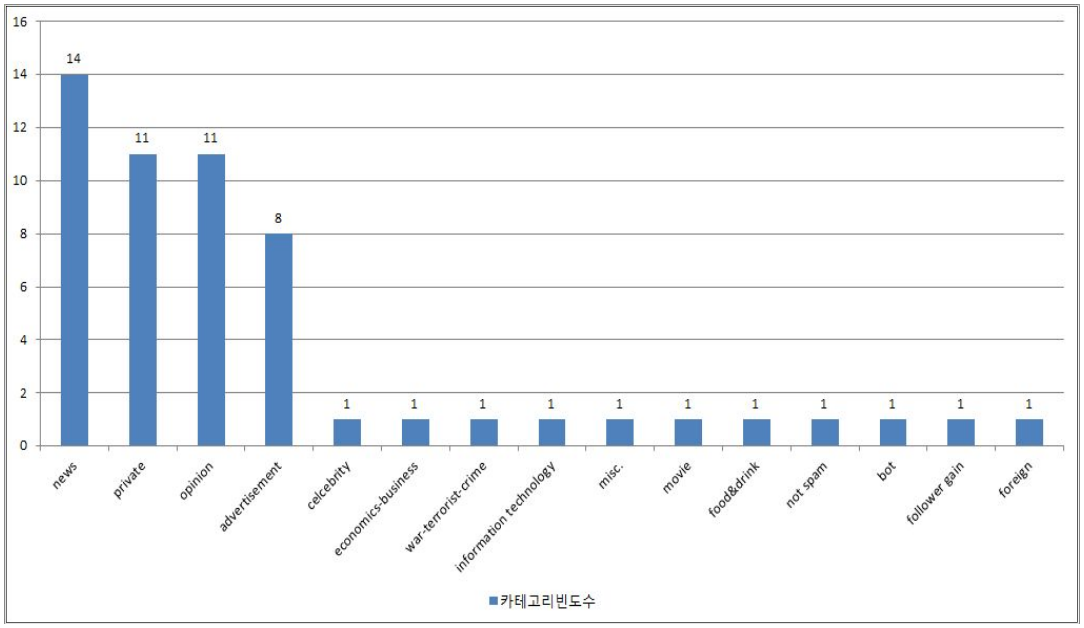
category1	category2	category3	category4	category5
- economic - entertainment - foreign - lifestyle - politic - social - sport	- news - opinion - deal - event - private message	- company - event - location - misc. - movie - person - product	-politics -sports -technology	- news - sports - politics - entertainment - education
category6	**category7**	**category8**	**category9**	**category10**
- crime - events - humor - movie - music - opinion - other - politic - sport	- article / multimedia sharing - technical disussions - new releases - satires - news - product promotions - community events - security updates - career - crowdsourcing requests - other	- health - environment - society	- advertising - informational - positive opinion - negative opinion - other	- news - promotion - survay - lifestyle
category11	**category12**	**category13**	**category14**	**category15**
- food & drink - travel - career & education -goods & services -event& activity - trifle	- economics-business - war-terrorist-crime - health - sport - development-government - politics - accident - entertainment - disaster-climate - education - society - international	- neutral news - personal news - opinionated news - opinions - deal - events - private message	- News and media - info_technology - reference - website and blogs - sport	- advertising - celebrity - not spam - bot - follower gain - explicit

[표 3-1]은 본 연구의 카테고리 선정을 위하여 단문텍스트를 대상으로 분류하는 15개의 논문에서 사용된 카테고리들을 나타낸다. 선행연구에 사용된 카테고리들은 분류하고자 하는 대상에 따라 다른 카테고리를 나타냈다. 선행연구에서 사용된 카테고리 중 promotions, deal 카테고리가 나타내는 의미는 'Advertisement'와 같은 의미를 나타냈다. 그 외에 같은 의미를 지니는 카테고리들을 묶어 하나의 카테고리로 지정하여 [표 3-2]에서 나타내고 있다.

[표 3-2] 선행연구 카테고리 선정 과정

category1	category2	category3	category4	category5
- economic - **opinion** - foreign - **private** - **news** - **news** - **news**	- **news** - **opinion** - **advertisement** - **private** - **private**	- company - **private** - location - misc. - movie - **private** - **advertisement**	- **news** - **news** -technology	- **news** - **news** - **news** - entertainment - education
category6	**category7**	**category8**	**category9**	**category10**
- crime - **private** - humor - movie - music - **opinion** - other - **news** - **news**	- article / multimedia sharing - technical disussions - new releases - satires - **news** - **advertisement** - community events - security updates - career - crowdsourcing requests - other	- health - environment - **news**	- **advertisement** - informational - **opinion** - **opinion** - other	- **news** - **advertisement** - survay - **private**
category11	**category12**	**category13**	**category14**	**category15**
- food & drink - travel -career& education -**advertisement** -event& activity - trifle	- economics-business - war-terrorist-crime - health - **news** - development-government - **news** - accident - entertainment - disaster-climate - education - **news** - international	- **news** - **private** - **private** - **opinion** - **advertisement** - **private** - **private**	- **news** - info_technology - reference - website and blogs - **news**	- **advertisement** - celebrity - not spam - bot - explicit - follower gain

그 결과 사용된 카테고리의 비중을 살펴본 결과 [그림 3-2]와 같이 나타났다.



[그림 3-2] 문서분류 시 사용되는 카테고리

[그림 3-2]는 [표 3-1]에서 의미가 같은 카테고리를 하나의 카테고리로 나타내고 그 외 하나의 묶어지지 못한 카테고리들을 나타내고 있다. 카테고리의 의미가 같게 나타난 'News', 'Private', 'Opinion', 'Advertisement'에 대한 카테고리의 빈도수가 높게 나타났으며, 다른 카테고리와 합쳐지지 못한 카테고리들의 빈도수는 1로 나타났다. 따라서 본 논문에서는 문서분류 시 많이 사용되는 4개의 카테고리를 바탕으로 단문텍스트 분류를 위한 카테고리 'Advertisement', 'News', 'Private', 'Opinion'을 선정하였다.

C. 학습데이터 구축

단문텍스트는 짧은 문장을 통해 다수의 사용자와 정보를 공유하거나 자신의 상황이나 관심사를 공유한다. 인터넷의 발달로 인해 실시간으로 생성되는 단문텍스트가 증가하면서 데이터의 효율적 관리의 중요성이 높아지고 있다. 본 절에서는 단문텍스트의 효율적 관리를 위해 단문텍스트의 특징을 활용하여 특징 가중치를 부여하여 카테고리별로 분류하는 방법에 대해 기술한다.

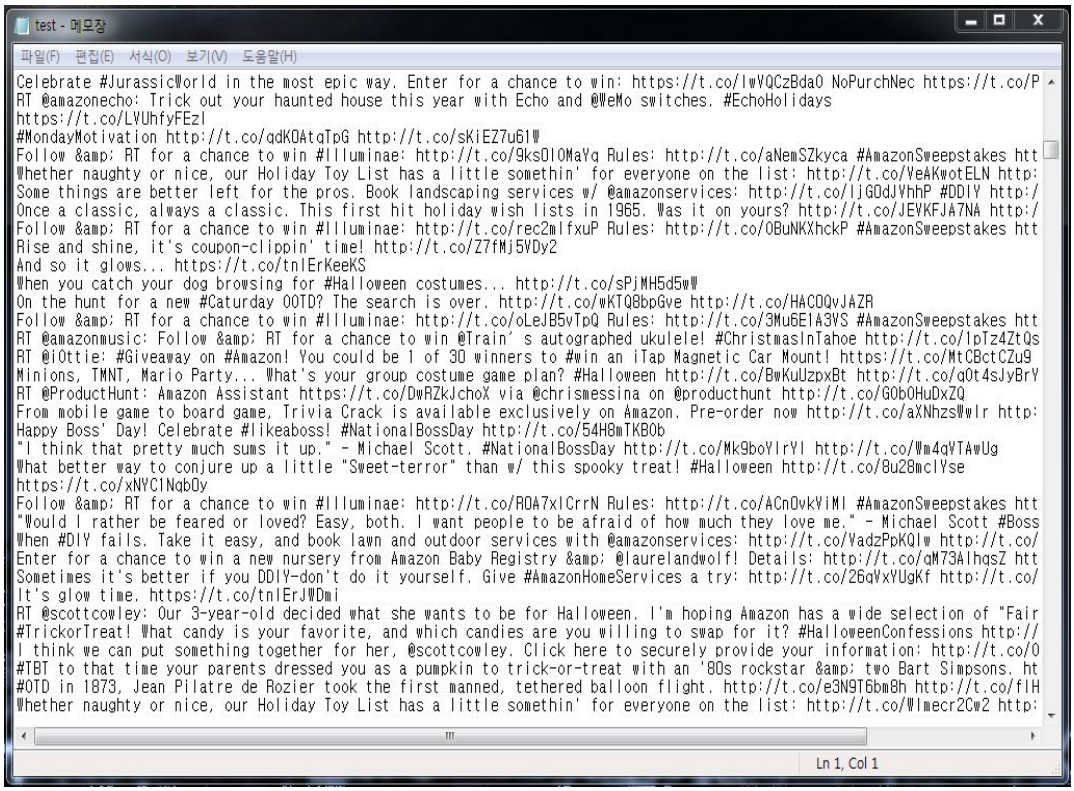
1. 단문텍스트 수집

본 절에서는 단문텍스트 분류를 위한 학습데이터 구축을 위해 앞 절에서 선정한 카테고리의 단문텍스트를 수집하는 과정을 다룬다. 앞 절에서 선정한 카테고리 중 'Advertisement'는 사용자들이 관심을 가지는 정보들을 나타내는 카테고리로 주로 제품의 정보, 이벤트 등을 나타낸다. 'Advertisement'에 해당하는 단문텍스트들은 미국의 대표적 쇼핑몰인 'Amazon' 트위터에서 데이터를 수집하였다.

'News'카테고리는 대부분의 선행연구에서 사용자들이 문서분류 시 이용하는 대표적 카테고리이다. 'News'의 경우 정치, 사회, 스포츠 등 여러 가지 주제를 포함하고 있어 문서분류에 적합한 카테고리이다. 'News'에 해당하는 단문텍스트는 미국의 방송국인 'ABC' 트위터에서 데이터를 수집하였다.

'Private'카테고리는 사용자가 다른 사용자와 사적으로 이야기를 나누는 등 개인적인 이야기를 나타내는 카테고리이다. 'Private'의 데이터 수집은 트위터의 팔로우 수가 2만 명이 넘고 꾸준히 활동하는 사용자인 'Julia Frakes'의 트위터에서 데이터를 수집하였다.

'Opinion'카테고리는 'News', 'Private'와 성향은 비슷하지만, 사용자의 의견을 확실히 드러내는 카테고리이다. 'Opinion'의 데이터 수집은 다양한 주제에 대해 이야기를 나누는 트위터 'Daily Review'에서 수집하였다. [그림 3-3]은 실험을 위하여 수집한 단문텍스트 일부를 나타내는 그림이다.



[그림 3-3] 수집된 단문텍스트

2. 특징 정의

본 논문에서는 단문텍스트의 특징을 활용하여 카테고리 분류를 위해 선행연구를 통해 얻을 특징과 카테고리의 특성을 통해 얻은 특징을 표로 정리하였다. 선행연구를 통해 나타난 특징들은 사용자를 언급할 때 '@'와 함께 아이디를 적어 '@ 아이디'형태로 나타나는 것과, '#'(해시태그)를 이용하여 정보를 나타내고 다른 사용자들과 관심을 공유하는 것과 단문텍스트의 단점이 짧은 문장을 보조하기 위한 URL이 나타나는 것을 알 수 있었다.

카테고리의 특성을 통해 나타난 특징들로는 'Advertisement' 카테고리에서는 홍보 및 제품을 꾸미는 형용사의 출현이 많았으며, 제품의 값을 나타내는 통화 '$'기호가 나타나기도 하며, 값의 할인을 나타내는 '%'와 자세한 설명을 나타내는 URL 등의 특징이 나타났다. 'News' 카테고리에서는 인물 명, 지역명, 회 명 등이 나타나며 News의 자세한 내용을 포함하는 URL이 나타났다. 'Private' 카테고리에서는

일상의 글, 개인적인 글, 약속을 나타내는 시간 등을 다루기 때문에 자신을 표현하는 일인칭 대명사가 주로 나타났으며, 인터넷에서 사용되는 신조어 등이 사용되는데 신조어는 고유명사로 분류할 수 있어 명사의 빈도수가 높게 나타났다. Opinion 카테고리의 경우 대상에 따라 다른 특징이 나타났지만, 특정 대상에 대한 의견을 나누기 위한 사용자를 언급하는 경우가 많았다. [표 3-3]은 선행연구를 바탕으로 단문텍스트의 특징을 정의한 것이다.

[표 3-3] 단문텍스트의 특징 정의

no	Feature	example tweet	labeling data
F1	일인칭 대명사	I, my, me	1-Person Pronouns (1PP)
F2	num%	30%	Precent (PCT)
F3	$num	$15	Dollar (DOL)
F4	url	http://ow.ly/KF5fu	url (URL)
F5	@id	@DisneyPixar	@id(ID)
F6	인명	William	Person (PER)
F7	지명	Colorado	location (LOC)
F8	회사명	Amazon	organization (ORG)
F9	해시태그	#HappyPrimeDay	hashtag (HASH)

[표 3-3]에서 정의한 단문텍스트의 특징은 9가지로 '일인칭 대명사', ' num%', '$num', ' url', '@id', '해시태그', '인명', '지명' '회사명'을 특징으로 정의하였다.

단문텍스트에서 특징은 늘 같은 형태로 나타나지 않는다. 일인칭 대명사만 해도 3가지 종류로 나타나기 때문에 특징들을 쉽게 나타낼 수 있도록 라벨링(Labeling)을 통해 나타낸다. 라벨링은 여러 유형의 단어들을 알기 쉬운 하나의 단어들로 변화하여 특징 추출을 쉽게 하였다.

3. 특징 추출

앞 절에서 정의한 특징들을 추출하기 위해서는 단문텍스트에서 불필요한 정보를 제거하고 필요한 정보를 추출할 수 있는 전처리 과정이 필요하다.

[표 3-4] 단문텍스트가 가지는 특징들

구분	원본 게시글
ad	We're celebrating 10/10 with Windows 10, are you? Today only, take 10% off this Acer desktop: http://t.co/extVqDwnb2 http://t.co/4ZVNpCOoDd
news	Donald Trump refuses to apologize over comments about John McCain's record in Vietnam: http://abcn.ws/1KexEvh
private	It's funny because I say I need everything. I have multiple cameras. I just want another one...
review	i cried and laughed out loud like my 6 year old self would have. @PixarInsideOut moved my heart.thank you @DisneyPixar!!

[표 3-4]는 특징을 나타내고 있는 단문텍스트이다. 특징은 정의한 모든 특징이 나타나는 것이 아니라, 몇 개의 특징들이 모여서 나타나는 것을 보이고 있다. [표 3-3]과 같은 형태를 지닌 단문텍스트에서 특징을 추출하기 위해서는 파이썬(Python)의 자연어처리 라이브러리인 NLTK (Natural Language Toolkit)을 이용한다. 특징추출과정은 (1) 토큰화 (2) 불용어 제거 (3) Stanford NER (Name Entity Recognizer) (4) Stemming (5) 라벨링 (Labelling) 단계를 거쳐 추출할 수 있다.

(1) 토큰화(Tokenizing)

토큰화는 문장 내에서 공백 또는 특정 단어를 기준으로 나누는 것을 의미한다. 본 연구에서는 수집한 단문텍스트를 라인별로 한 줄씩 불러와 단문텍스트의 공백을 기준으로 토큰화를 수행한다.

[표 3-5] 토큰화 데이터 결과

	Feature example tweet labeling data
원본 데이터	@1NL_by_ERO Bonsoir, je suis desole pour le retard de livraison, l'avez-vous signale a notre service client via : https://t.co/jhujRUq7sa ?
토큰화 데이터	'@1NL_by_ERO' 'Bonsoir' ',' 'je' 'suis' 'desole' 'pour' 'le' 'retard' 'de' 'livraison' ',' 'l'avez-vous' 'signale' 'a' 'notre' 'service' 'client' 'via' ':' 'https://t.co/jhujRUq7sa' '?'

(2) 불용어 제거(Stopword)

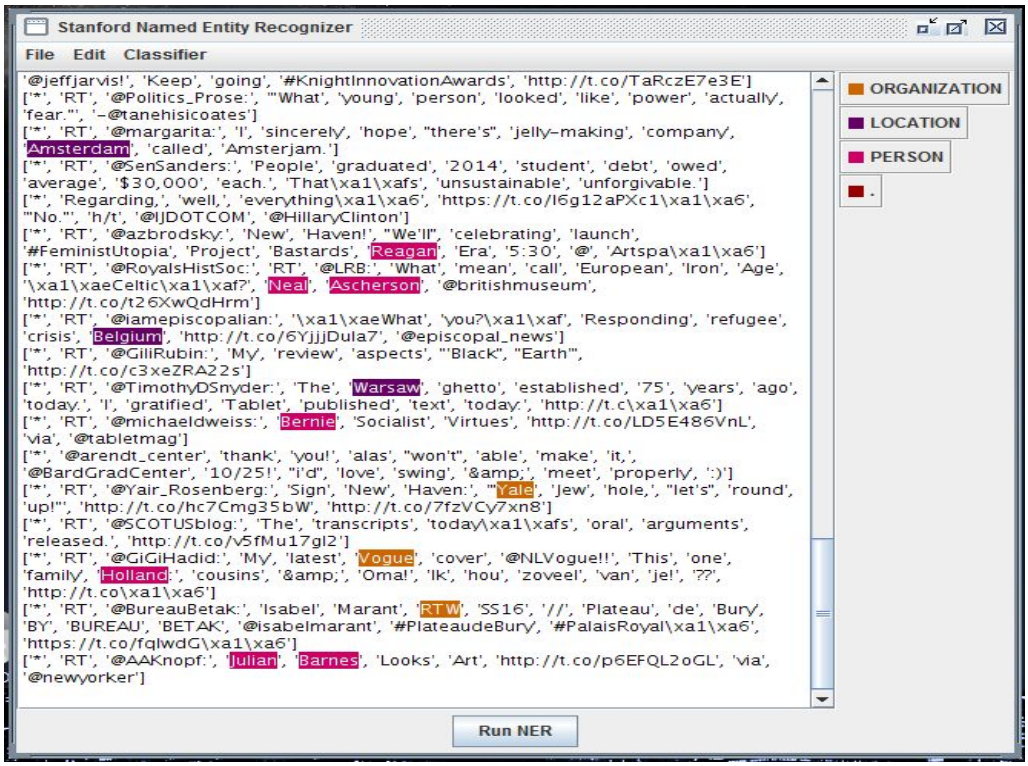
불용어 제거(Stopword)는 문장 분석 시 불필요한 단어들을 제거하는 것을 뜻한다. 불용어는 주로 숫자, 특수문자, 관사. 조사 등 문장 내에서 별다른 의미를 가지지 않고, 그 외 실험 시 사용자가 불필요하다고 생각되는 단어들을 추가하여 이용할 수 있다. 본 실험에서는 토큰화를 하면서 불용어를 제거하였고, 불용어는 특수문자 @', '#', '$', '%'를 제외한 특수문자들과 NLTK에서 제공되는 STOPWORDS 라이브러리를 이용하여 특수문자 이외의 불용어를 제거한다. 토큰화 데이터에서 불용어를 제거한 데이터는 [표 3-6]에서 나타내고 있다.

[표 3-6] 불용어제거 데이터 결과

	Feature example tweet labeling data
토큰화 데이터	'@1NL_by_ERO' 'Bonsoir'',' 'je' 'suis' 'desole' 'pour' 'le' 'retard' 'de' 'livraison'',' 'l'avez-vous' 'signale' 'a' 'notre' 'service' 'client' 'via' ':' 'https://t.co/jhujRUq7sa' '?'
불용어제거 데이터	'@1NL ERO', 'Bonsoir', 'je', 'suis', desole', 'pour', 'le', 'retard', 'de', 'livraison', 'l', 'ave', 'vous', 'signale', 'notre', 'service', 'client', 'via', 'https://t.co/jhujRUq7sa'

(3) Stanford NER (Name Entity Recognizer)

Stanford NER (Name Entity Recognizer)은 개체명 인식을 뜻하는 것으로 다양한 클래스들에 따라 다양한 개체명 인식이 가능하다. 개체명 인식을 할 때는 Stanford에서 제공하는 클래스를 기반으로 개체명인식이 가능하다. 3 class에서는 organization, person, location를 인식하고 4 class는 3 class에 misc를 추가로 인식한다. 7 class는 date, money, percent, time. organization, person, location의 개체명 인식이 가능하다. 본 연구에서는 3 class를 이용하여 지명, 인명, 회사명을 추출한다.



[그림 3-4] Stanford NER 실행 화면

[그림3-2]는 윈도우환경에서 Stanford NER을 실행시킨 것으로 3 class를 이용한 경우, ORGANIZATION, PERSON, LOCATION을 프로그램이 찾아내어 사용자가 보기 쉽게 제공하고 있고, 파일로 따로 저장할 수 있는 프로그램이다. Stanford NER을 이용한 과정은 [표 3-7]에서 나타내고 있다.

[표 3-7] Stanford NER 데이터 결과

	Feature example tweet labeling data
불용어제거 데이터	'@1NL ERO', 'Bonsoir', 'je', 'suis', desole', 'pour', 'le', 'retard', 'de', 'livraison', 'l', 'ave', 'vous', 'signale', 'notre', 'service', 'client', 'via', 'https://t.co/jhujRUq7sa'
Stanford NER	'@1NL', '<ORGANIZATION>ERO</ORGANIZATION>', 'Bonsoir', 'je', 'suis', 'desole', 'pour', 'le', 'retard', 'de', 'livraison', 'l', 'ave', 'vous', 'signale', 'notre', 'service', 'client', 'via', 'https://t.co/jhujRUq7sa'

(4) STEMMING

Stemming은 단어의 원형, 어근을 찾아주는 것으로 'thinking'이라는 단어가 있을 때, 이 단어의 원형은 'think'인 것을 나타내 주는 것이다. Stemming을 이용하는 이유는, 여러 가지 형태로 나타나는 단어를 하나의 형태로 바꾸어 같은 의미를 가지는 단어를 하나의 품사로 바꾸기 위해서 사용한다. 실험에서는 NLTK 라이브러리에서 제공하는 Porter Stemming을 이용하였다. [표 3-1]에서는 단문텍스트의 원본데이터가 Stemming 단계를 거치게 되면 몇 가지 복수형, 과거형 단어들이 단어의 원형으로 변경된 것을 볼 수 있다.

[표 3-8] Stemming 결과

	Feature example tweet labeling data
Stanford NER	'@1NL', '<ORGANIZATION>ERO</ORGANIZATION>', 'Bonsoir', 'je', 'suis', 'desole', 'pour', 'le', 'retard', 'de', 'livraison', 'l', 'ave', 'vous', 'signale', 'notre', 'service', 'client', 'via', 'https://t.co/jhujRUq7sa'
Stemming	'@1NL', '<ORGANIZATION>ERO</ORGANIZATION>','Bonsoir', 'je', 'suis', 'desole', 'pour', 'le', 'retard', 'de', 'livraison', 'l' 'ave', 'vous', 'signale', 'notre' 'service', 'client', 'via', 'https://t.co/jhujRUq7sa'

[표 3-8]에서는 토큰화시킨 데이터를 Stemming한 결과를 나타내고 있다. Stemming은 토큰화가 먼저 수행된 이후 수행되어야지만 정확한 결과를 나타낼 수 있으며, [표 3-1]에서의 Stemming 결과는 'tips'가 'tip'으로 'spooked'가 'spook'으로 바뀐 것을 보이고 있다.

(5) 라벨링 (Labeling)

라벨링(Labeling)은 특정 상품에 대한 간략한 내용을 작성하는 것으로 본 논문에서는 단문텍스트의 특징 추출을 간편히 할 수 있게 하는 작업이다. 앞서 특징 정의에서 나타낸 것과 같이 단문텍스트에서 나타나는 특징들을 간편하게 바꾸는 작업을 한다.

[표 3-9] Labeling 결과

	Feature example tweet labeling data
Stemming	'@1NL', '<ORGANIZATION>ERO</ORGANIZATION>', 'Bonsoir', 'je', 'suis', 'desole', 'pour', 'le', 'retard', 'de', 'livraison', 'l', 'ave', 'vous', 'signale', 'notre', 'service', 'client', 'via', 'https://t.co/jhujRUq7sa'
Labeling	'ID', 'ORG','Bonsoir', 'je', 'suis', 'desole', 'pour', 'le', 'retard', 'de', 'livraison', '1PP' 'ave', 'vous', 'signale', 'notre' 'service', 'client', 'via', 'URL'

[표 3-9]에서 나타나는 특징은 라벨링을 통해 'ID', 'ORG', '1PP', 'URL'가 나타나는 것을 확인할 수 있다.

(6) POS-Tagging (형태소 분석)

POS-Tagging은 문장 내의 단어들이 명사, 동사, 형용사 등 어떤 품사를 나타내는지를 알려주는 방법이다. POS-Tagging은 문서 분류의 정확을 높이기 위해 원하는 품사만을 추출할 때 이용한다. 명사, 동사, 형용사는 Stanford POS-Tagger를 적용하여 POS-tagging 표를 바탕으로 명사는 'NN', 'NNP', 'NNS', 'NNPS'를 추출하였고, 동사는 'VB', 'VBD', 'VBG', 'VBZ'를 추출하였다. 다음 [표 3-9]는

Stanford에서 제공하고 있는 POS-Tagger 표이다[22]. 명사와 동사만을 고려하는 이유는 앞 단계에서 Poter-Stemming을 통해 단어의 어근만을 고려하여 단어의 형태가 온전하지 않아, 형용사의 경우 고유명사로 검색되기 때문에 검색 시 잘 나타나는 명사와 동사를 고려하였다.

[표 3-10] POS(Part of Speech) Tagger

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

[표 3-9]에 의거하여 Stemming을 거친 단어들에 대한 형태소 분석결과는 다음 [표 3-10]을 통해 나타낼 수 있다.

[표 3-11] POS-Tagging 결과

	Feature example tweet labeling data
Labeling	'ID', 'ORG', 'Bonsoir', 'je', 'suis', 'desole', 'pour', 'le', 'retard', 'de', 'livraison', '1PP' 'ave', 'vous', 'signale', 'notre' 'service', 'client', 'via', 'URL'
POS -TAGGING	("'ID',", 'NN'), ("'ORG','Bonsoir',", 'NN'), ("'je',", 'NN'), ("'suis',", 'NN'), ("'desole',", 'NN'), ("'pour',", 'NN'), ("'le',", 'NN'), ("'retard',", 'NN'), ("'de',", 'NN'), ("'livraison',", 'NN'), ("'1PP'", 'NN'), ("'ave',", 'NN'), ("'vous',", 'NN'), ("'signale',", 'NN'), ("'notre'", 'NN'), ("'service',", 'NN'), ("'client',", 'NN'), ("'via',", 'NN'), ("'URL'", 'NN')

특징 추출을 위한 6단계를 거치고 난 뒤 본 논문에서 제안하는 특징 가중치를 부여한 문서분류를 위해 특징 가중치 값을 계산한다. 특징 가중치는 [표 3-10] POS-Tagging 방법을 통해 추출된 명사와 동사의 단어빈도수(Word Frequency)를 활용한다. 단어빈도수는 텍스트 마이닝에서 주로 이용되는 가중치 기법으로, 특정 단어가 문서 내에서 가지는 중요성을 수치로 나타내는 방법이다. 단어빈도수를 구하는 공식은 다음 수식과 같다.

$$Wf = \frac{\text{문서 내 단어 출현 횟수}}{\text{문서의 갯수}} \quad (1)$$

추출한 단어 빈도수와 특징을 이용하여 문서분류를 위해서 특징 가중치라고 하는 값을 측정한다. 특징 가중치를 이용하는 이유는 단문텍스트의 특성상 문서로 생각될 수 있는 텍스트가 단문으로 이루어져 있어 출현하는 단어의 개수가 적게 나타나기 때문이다. 특징 가중치는 전처리과정 추출한 명사, 동사, 형용사들의 빈도수를 구하고, 특징이 나타날 때 함께 등장하는 단어들에 대해서는 별도로 특징빈도수를 구한다. 특징 가중치를 구하는 수식은 다음과 같다.

$$Wfw = Wf + \frac{\sum(feature)}{Wf} \quad (2)$$

수식 (2)의 Wfw는 단어 특징 가중치를 나타내는 수식으로 학습데이터를 구축하기 위해 수집한 데이터에서 나타난 단어들의 빈도수에 단어가 등장할 때 나타나는 특징의 출현 빈도수 합의 평균을 더 하여 나타낸다. 특징의 출현 빈도를 더해주는 이유는 카테고리별로 나타나는 주요 특징이 다르므로 같은 단어라고 할지라도 각각의 카테고리에서 다른 값을 가지게 하여 단문텍스트를 분류하는데 정확도를 높이기 위해 사용한다.

전처리 과정을 거쳐 명사와 동사에 해당하는 단어를 추출하고 단어 빈도수와 단어의 특징빈도수를 계산하면 특징 가중치를 이용한 문서분류를 위한 학습 데이터 셋을 구축할 수 있으며, 미분류 데이터 셋 또한, 같은 과정을 거쳐 데이터 셋을 구축한다.

D. 단문텍스트 카테고리 분류

앞 절에서는 단문텍스트 분류를 위해 단문텍스트를 수집하기 위해 카테고리를 선정하고, 단문텍스트가 가지는 특징을 정의, 추출하여 학습데이터 셋을 구축하였다. 단문텍스트의 분류는 학습데이터 셋과 미분류 데이터의 데이터 셋의 상관성을 분석하여 단문텍스트를 카테고리별로 분류하였다.

1. 상관성 분석을 이용한 분류방법

본 절에서는 앞 절에서 구축한 학습 데이터 셋과 미분류 데이터 셋 간의 상관성을 비교하여 단문텍스트를 카테고리별로 분류하는 방법에 대해 기술한다. 상관성 분석(Correlation Analysis)은 확률론과 통계학에서 두 변수의 상관성을 분석하는 방법이다[24]. 상관성 분석에서 두 변수는 서로 독립적인 관계일 수도 있고, 서로 상관이 있는 관계일수도 있다. 이때 두 변수 간의 상관관계를 나타내는 것이 상관성 분석이다. 상관성 분석의 결과는 상관계수를 통해 나타내진다. 본 연구에서 이용하는 상관성 분석 방법 중 스피어만 상관계수 (Spearman rank-order correlation coefficient)를 이용한다.

기존의 문서는 일정한 규격을 가지고 있는 문서로 문서의 내용을 바탕으로 문서들을 분류하였다. 그러나 단문텍스트는 일정한 규격을 가지지 않았고, 단문으로 내용을 파악하기에 짧은 데이터를 가지고 있다. 따라서 기존의 문서분류 방법은 단문텍스트에 적용하기에 부적합하다. 이에 본 논문에서는 단문텍스트 내에 나타나는 특징들과 단어의 빈도수와 특징빈도수를 이용하여 학습데이터를 구축하고 실험 데이터와의 일치하는 단어의 빈도수 특징빈도수를 이용하여 두 데이터 간의 상관관계를 통해 카테고리 분류하기 위해 상관성 분석방법 중 하나인 스피어만 상관계수를 이용하여 단문텍스트를 카테고리별로 분류 하였다.

스피어만 상관계수는 두 데이터의 실제 값 대신 두 값의 순위를 이용하여 상관계수를 이용하는 방법으로 비선형 데이터들의 연관성을 파악할 수 있는 장점이 있다[25]. 본 연구에서는 스피어만 상관계수를 이용하기 위해 학습데이터 셋과 미분류 데이터 셋 간의 순위를 임의로 부여하여 스피어만 상관계수를 통한 상관성 분석하였다. 스피어만 상관계수는 수식 (3)과 같은 형태로 정의하고 있다

다음 수식(3)과 같이 정의하고 있다.

$$\rho = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \quad (3)$$

스피어만 상관계수의 범위는 [-1, 1]이며 0일 때는 서로 연관이 없음을 나타내기 때문에 상관성 분석을 위해 단어의 값을 정규화해준다. 정규화는 카테고리별로 가지고 있는 단어의 개수를 기준으로 한다. 본 연구에서는 스피어만 상관계수의 값을 계산하여 상관계수의 값이 크게 나타난 카테고리로 분류하였다. 본 논문에서는 학습 데이터 셋과 미분류 데이터를 각각 X, Y로 간주하여 계산하였다.

Ⅳ. 실험 및 결과

본 장에서는 본 연구에서 데이터 수집을 위해 트위터 Open API를 이용하여 구축한 학습데이터 셋과 실험데이터 셋에 대하여 설명한다. 또한, 단문텍스트의 특징을 기반으로 특징 가중치를 부여하여 상관성 분석을 통해 단문텍스트를 카테고리별로 분류하고 정확도 이용하여 성능을 평가한다. 실험은 PC상에서 python 2.7 버전을 이용하여 구현하였다.

A. 데이터 수집

본 절에서는 연구에서 사용된 트위터 단문텍스트를 수집하는 과정에 대해해 기술한다. 트위터 단문텍스트를 수집하기 위해서는 트위터 개발자 모드에서 제공하는 Open API를 이용하여 트위터에 접근할 수 있는 Access_token을 발급받아 데이터를 수집할 수 있다. 트위터에서 단문텍스트를 수집하는 방법은 다음 세 단계로 나타낼 수 있다.

(1) 트위터 개발자 모드에서 제공하는 Open API를 이용하기 위해 트위터 개발자 페이지에서 OAuth 인증절차를 거쳐 트위터에 접속할 수 있는 consumer_token, consumer_secret, access_token, acces_secret 값을 발급받는다.

(2) 학습 데이터로 이용될 단문텍스트는 미리 지정한 4개의 카테고리 'Advertisement', 'News', 'Private', 'Opinion'에 해당하는 트위터에서 단문텍스트를 수집한다.

(3) 수집한 단문텍스트는 각각의 카테고리 이름을 따라 a'Advertisement', 'News', 'Private', 'Opinion'의 텍스트 파일 (.txt)을 생성하고 게시 글 한 개를 기준으로 행 분리하여 저장한다.

[표 4-1]은 학습데이터 셋을 구축하기 위한 단문텍스트를 수집하는 PYTHON 코

드이며, 정답집단을 만들기 위해 각 카테고리에 해당하는 단문텍스트를 수집한다. 본 연구에서는 'Advertisement', 'News', 'Private', 'Opinion'에 해당하는 단문텍스트를 트위터에서 수집하는 python 코드이다.

[표 4-1] 단문텍스트 수집 코드

```
def twitter_fetch(screen_name ,maxnumtweets):
  #'Fetch tweets from @BBCNews'
  # API described at https://dev.twitter.com/docs/api/1.1/get/statuses/user_timeline

  consumer_token = 'OYhrBeVXr7idp3L2CJbittGxb' #substitute values from twitter website
  consumer_secret = 'ZjfYRQWKRFDi256OYUW98tud9QFl9gR4xNOxG1iRvL8lyCshok'
  access_token = '400719198-UcukbsydOtQPVnaOLcXkHHKE4QbXSXgSYLSrMKlU'
  access_secret = 'VNGujssz2RM90iJQWUuv4eg6Ldyx7FrJaV82oTzFAXJKq'
  # 트위터 개발자에서 입력값받기

  auth = tweepy.OAuthHandler(consumer_token,consumer_secret)
  auth.set_access_token(access_token,access_secret)

  api  = tweepy.API(auth)

  for status in tweepy.Cursor(api.user_timeline,id=screen_name).items(150):# 데이터 수

      temp= status
      f=open("C:₩₩Python27₩₩text₩₩test.txt",'a') #파일오픈, 저장
      f.write(temp.text)
      f.write('₩n')
      f.close()

      print temp.text+'₩n'
if __name__ == '__main__':
  twitter_fetch(amzon'',10)     # ' ' 안에 사용자이름입력
```

B. 데이터 셋

본 절에서는 본 연구에서 구축한 데이터 셋에 대해 설명한다. 데이터 셋은 지도 학습방법을 적용하여 학습데이터 셋과 실험에서 이용된 실험 데이터 셋으로 구성된다.

1. 학습 데이터 셋

본 연구에서 특징 가중치를 이용한 단문텍스트의 카테고리 분류방법을 제안하기 위해 단문텍스트의 명사와 동사에 해당하는 단어의 빈도수를 추출하고 단어빈도수의 특징 가중치를 [표 4-2]와 같이 학습데이터 셋을 구축하였다.

[표 4-2] 학습데이터 셋 예제

no.	Word	Word Frequency	Word feature wight
1	DEAL	8	9.875
2	SAVE	7	8.28571
3	happyprimeday	5	6.6
4	select	4	6.25
5	headphone	1	3
…	…	…	…
122	philip	1	3

학습에 사용된 트위터의 단문텍스트는 총 2,800건에 해당하며, 특징을 기반으로 선정한 카테고리를 기준으로 ‘Advertisement’, ‘News’, ‘Private’, ‘Opinion’ 별로 700개씩 나누어 수집하였다.

2. 실험 데이터 셋

본 논문에서 사용된 실험 데이터 셋은 카테고리를 기준으로 각 단문텍스트에 번호를 붙여 단문텍스트가 어느 카테고리로 분류되었는지 정확도 계산을 위해 1200건의 실험 데이터 셋을 구축하였다. [표 4-3]은 실험 데이터 셋을 표현 하고 있다.

[표 4-3] 실험 데이터 셋

카테고리	번호	게시물 건
Advertisement	1 ~ 300	300
News	301 ~ 600	300
Private	601 ~ 900	300
Review	901 ~ 1200	300

실험데이터 셋도 학습데이터 셋과 같은 방법으로 실험 데이터 셋을 구축한다. [표 4-4]와 같이 미분류 데이터 셋을 구축할 수 있다.

[표 4-4] 실험 데이터 셋 예제

no.	Word	Word Frequency	Word feature weight
1	HappyPrimeDay	2	5
2	deal	3	6
3	Save	2	3
4	select	1	5
5	Philip	1	5
6	headphone	1	5

C. 실험평가 방법 및 결과 분석

1. 실험 평가 방법

실험의 평가는 기계학습의 모델검증을 위해 사용되는 교차검증을 이용한다. 카테고리의 분류 정확성을 측정하기 위해 정확률을 이용하였다. 단문텍스트의 카테고리 분류에 대한 성능평가는 교차검증을 이용한다. 교차검증을 통해서는 정확률(Precision), 재현율(Recall), F-score를 구할 수 있다[31].

[표 4-5] 교차검증 표

		실제정답	
		참(True)	거짓(False)
실험결과	참 (True)	True Positive (tp)	False Positive (fp)
	거짓 (False)	False Negative (fn)	True Negative (tn)

[표 4-5]는 교차검증 표이다. 정확률은 전체 결과 중에 참으로 분류되었을 때만 나타내는 것이고, 재현율은 실험결과 참으로 분류된 결과 중 나온 결과 중에 참으로 분류되었을 때를 식(5), 식(6)을 통해 나타낸다. 본 실험에서는 Advertisement에 분류된 모든 데이터 중 실제 'Advertisement'로 분류된 데이터를 뜻한다.

$$\mathrm{Pr}ecision = \frac{tp}{tp+fp} \tag{5}$$

$$Recall = \frac{tp}{tp + fn} \tag{6}$$

$$F_1 = \frac{2*\mathrm{Pr}ecision*reall}{\mathrm{Pr}ecision + reall} \tag{7}$$

정확률과 재현율은 반비례 관계를 가지며, 정확률이 높으면 재현율이 낮아지는 형태를 가지고 있다. 이를 보조하기 위해 정확률과 재현율을 이용한 F-score 방법을 이용한다.

F1-score는 정확률과 재현율을 이용한 정확도 측정방법으로 정확률과 재현율의 중요성을 동일하게 두고 계산을 한다. 본 논문에서는 단문텍스트의 카테고리 분류의 정확도를 구하기 위해 식 (5), (6), (7)을 이용하여 단문텍스트의 카테고리 분류의 성능을 평가한다.

2. 실험 결과 분석

단어빈도수만을 이용하여 카테고리를 분류하는 방법의 결과와 본 논문에서 제안하는 방법을 이용하여 카테고리의 분류된 결과의 정확률을 평가한 결과는 다음과 같다. 스피어만 상관계수를 이용하여 미분류 텍스트를 분류하는 결과는 [표 4-6]과 에서 나타내고 있다.

[표 4-6] 단문텍스트 카테고리 분류 결과

	Advertisement	News	Private	Opinion
Advertisement (300)	**272**	7	12	9
News (300)	11	**276**	6	7
Private (300)	23	9	**232**	36
Opinion (300)	4	8	20	**268**

[표 4-6]은 미분류 데이터를 특징 가중치를 이용하여 카테고리별로 분류한 표를 나타낸다. 특징 가중치를 이용하여 카테고리별로 미분류 데이터를 분류하였을 때 90% 이상의 정확도를 나타내며 분류된 것을 확인할 수 있었다.

그러나 미분류 카테고리 중 'Private' 카테고리의 경우, 본 논문에서 정의하는 특징이 많이 나타나지 않았다. 'Private' 카테고리의 단문텍스트는 사용자의 직업 또는 주된 관심사에 따라 다른 카테고리의 성향을 나타내는 특징들이 나타났기 때문에 'Advertisement', ' Opinion' 카테고리 쪽으로 오 분류가 다소 나타난 것으로 보인다.

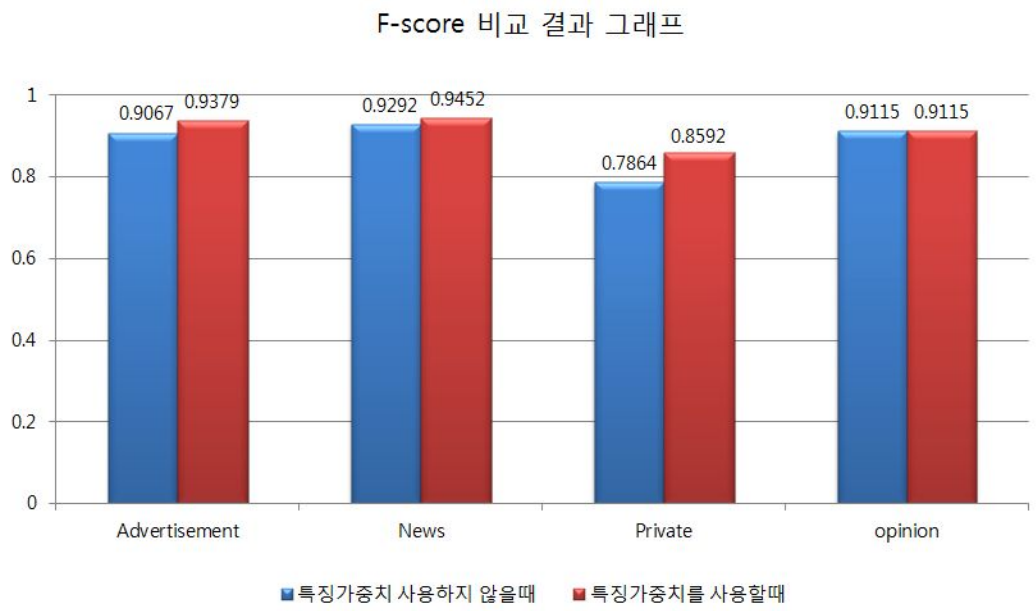
[표 4-7] 단어빈도수를 이용한 카테고리 분류 정확도 결과

단어빈도수를 이용한 정확도 결과				
	Advertisement	News	Private	Opinion
Precision	0.9067	0.92	0.7733	0.8933
Recall	0.9067	0.9387	0.8	0.9305
F-score	0.9067	0.9292	0.7864	0.9115

[표 4-8] 단어빈도수와 특징 가중치를 이용한 카테고리 분류 정확도 결과

단어빈도수와 특징 가중치를 이용한 정확도 결과				
	Advertisement	News	Private	Opinion
Precision	0.9067	0.92	0.7733	0.8933
Recall	0.9717	0.9718	0.9667	0.9305
F-score	0.9379	0.9452	0.8592	0.9115

[표 4-7]과 [표 4-8]은 특징 가중치의 적용하지 않았을 때와 특징가중치를 적용했을 때에 따른 정확도 결과로, 학습데이터 셋과 미분류 데이터 셋 간의 상관성을 비교하여 Precision, Recall, F-score 값을 구한 것이다. [표 4-7]과 [표 4-8]를 비교하면 단어 빈도수만을 이용해 단문텍스트를 카테고리에 분류한 결과 정확도는 'Advertisement'는 0.9067, 'News'는 0.9292, 'Private'는 0.7864, 'Opinion'은 0.9115의 값을 가지는 것을 알 수 있었다. 그러나 단어빈도수와 단어 특징빈도수를 이용한 상관성 분석결과 'Advertisement'는 0.9379, 'News'는 0.9452, 'Private'는 0.8592, 'Opinion'은 0.9115의 값으로 단문텍스트 카테고리 분류의 정확도가 향상된 것을 확인할 수 있다.



[그림 4-1] F-score 비교 결과 그래프

전체적인 분류 결과는 90% 이상의 정확성을 보여 본 논문에서 제안하는 특징 가중치를 이용하는 방법에 대한 효율성을 입증할 수 있는 결과를 보이고 있다. 앞서 [표 4-7]과 [표 4-8]에서 나타낸 F-Score 결과를 그래프로 [그림 4-1]에서 나타내고 있다.

[그림 4-1]을 보면 전체적으로 단어빈도수만을 사용했을 때보다 단어특징빈도수를 이용하여 카테고리별로 분류하였을 때 소량이지만 상승된 결과를 보여주고 있다. 특징 가중치를 함께 이용할 때, 분류의 정확률이 높아진 카테고리는 'Private'으로 0.7864에서 0.8592 으로 정확도가 증가한 것을 확인할 수 있었다. 그 외, 'Advertisement'는 0.9067에서 0.9379로 'News'는 0.9292에서 0.9452로 상승된 것을 확인할 수 있었다. 'Private'의 경우는 1인칭 대명사를 특징으로 정해짐에 따라 높게 분류가 된 것으로 보인다. 'Opinion' 카테고리는 특징 가중치를 이용하여 미분류 텍스트들을 분류한 결과, 특징 가중치를 이용하지 않았을 때와 차이가 나타나지 않았다. 'Opinion' 카테고리는 하나의 주제가 아닌 여러 주제를 대상으로 하고 있어 정확한 분류가 되지 않은 것으로 보인다.

향후 'News'와 'Advertisement' 카테고리에 해당하는 텍스트들의 효율적 관리를 필요로 할 때, 'News'와 'Advertisement' 카테고리에 해당되는 단문텍스트들을

이용하는 분야에서 유용하게 사용될 수 있을 것으로 사료된다. ‘Private’은 특징 가중치를 이용했을 때 가장 높은 변화폭을 보였으므로, 사용자의 관심사를 파악하거나, 개인서비스를 중심으로 하는 서비스에서 이용 가능할 것으로 보인다. 또한, ‘Opinion’ 카테고리의 정확도를 높이기 위해서는 ‘Opinion’ 카테고리에서 나타나는 특징들을 분석하여 특징 가중치의 값을 높이는 방법을 고려해야 하여 발전시킨다면 SNS를 통한 마케팅, 정보검색, 사용자 맞춤형 서비스 등 다양한 분야에서 활용될 것으로 기대된다.

Ⅴ. 결론 및 제언

본 논문에서는 단문텍스트에서 나타나는 특징들을 바탕으로 특징 가중치라고 하는 값을 이용하여 학습 데이터 셋을 구축하고, 미분류 데이터 셋과 상관성 분석(스피어만 상관계수)을 이용한 단문텍스트의 카테고리 분류 방법을 제안하였다.

본 논문에서 제안하는 방법은 기존에 사용되는 문서분류와 달리 실시간으로 생성되는 단문텍스트들을 이용하여 단문텍스트가 가지고 있는 특징을 바탕으로 특징을 정의하고, 정의한 특징을 이용하여 특징 가중치 값을 통해 단문텍스트의 카테고리 분류하는 것이다.

단문텍스트의 카테고리 분류를 위해 단문텍스트를 분류하는 선행연구를 통해 'Advertisement', 'News', 'Private', 'Opinion' 카테고리를 선정하였고, 해당 카테고리로의 분류에 적합한 단문텍스트가 가지고 있는 특징을 정의하였다. 특징 가중치의 경우 학습 데이터셋 구축 시 단문텍스트의 단어빈도수와 단어가 등장할 때의 특징 빈도수를 활용하여 나타내는 것으로, 본 연구를 통해 단문텍스트 분류 시 특징 가중치를 이용하여 카테고리에 분류하였을 때 'Advertisement'는 93%, 'News'는 94%, 'Private'는 85%, 'Opinion'은 91%로 평균 90. 7%의 결과로 좋은 성능을 나타냄을 확인하였다.

본 연구에서는 단문텍스트를 대상으로 'Advertisement', 'News', 'Private', 'Opinion' 카테고리를 선정하여 단문텍스트를 분류하였으나, 향후에는 다양한 카테고리를 선정하고, 9개의 특징 이외에 이모티콘, 신조어, URL의 종류 등 다른 특징을 추가로 정의하여 단문텍스트를 분류한다면, 보다 더 높은 성능을 보일 것으로 사료된다. 또한, 본 연구는 영어를 대상으로 연구하였으나, 향후 다른 언어를 대상으로 연구하고 단문텍스트의 카테고리 분류 체계를 갖추는 데 이용된다면 다양한 소셜 검색 서비스에 활용할 수 있을 것으로 기대된다.

참고문헌

[1] https://ko.wikipedia.org/wiki/소셜네트워크서비스.

[2] 이윤희, "국내 SNS 의 이용 현황과주요 이슈 분석.", INTERNET & SECURITY FOCUS 10, pp.56-78, 2012.

[3] 허상희, 최규수, "트위터에서 트윗 (tweet) 의 특징과 유형 연구." 한민족어문학 제 61권, pp.455-494, 2012.

[4] 이아람, "특징 선택과 가중치 부여를 통한 자동문서분류시스템의 성능 향상," 서울시립대학교, 석사학위논문, 2013.

[5] 홍초희, 김학수, " 트윗분류를 위한 효과적인 자질 추출", 2011 한국컴퓨터종합학술대회 논문집, 제38권, 제1호, pp.229-232, 2011.

[6] 임혜영, "SVM 분류기를 이용한 문서 범주화 연구", 연세대학교 문헌정보학과 석사학위 논문, 2000.

[7] 나윤재 "변형 나이브 베이즈 분류기를 이용한 자동 문서분류에 관한연구", 연세대학교 석사학위논문, 2008.

[8] 장병탁, "차세대 기계학습 기술", 정보과학회지 제 25권, 제 3호, pp.96-107, 2007.

[9] 조우승, "기계학습을 통해 예측한 유량을 반영하는 CEP 기반의 실시간 블록 누수 탐지 방법", 충남대학교, 석사학위논문, 2015.

[10] 박찬정, "기계학습을 이용한 특허문서의 다중 IPC 자동분류 방법", 경기대학교, 박사학위논문, 2013.

[11] 김병주, "문서분류에서의 SVM 및 나이브베이지안, EM 알고리즘의 특성 비교", 동국대학교, 석사학위논문, 2009.

[12] https://ko.wikipedia.org/wiki/TF-IDF

[13] 이재환, et al., "트위터에서의 토픽별 감정 패턴 분석.", 2014 한국정보과학회 제 41 회 정기총회 및 동계학술발표회, pp.204-206, 2014.

[14] 김예경, " 트위터 사용자의의 음악청취 행태와 음악트렌드의 관계성 분석 및 예측", 서울대학교, 석사학위논문, 2015.

[15] Oulasvirta, Antti, et al. "Making the ordinary visible in microblogs." Personal and ubiquitous computing vol.14, no.3, pp.237-249, 2010.

[16] 이의종, 김정동, 백두권, “트위터 특징에 기반한 콘텐츠 중요성 평가기법”, 정보과학회논문지 제41권 제12호, pp.1136-1144, 2014.

[17] Tumasjan, Andranik, et al. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." ICWSM 10 , pp.178-185, 2010.

[18] Cha, Meeyoung, et al, "Measuring User Influence in Twitter: The Million Follower Fallacy." ICWSM 10, pp.10-17, 2010.

[19] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?", Proc. of the 19th international conference on World wide web, pp.591-600, 2010.

[20] http://www.statista.com/statistics/274564/monthly-active-twitter-users-in-the-united-states/

[21] 이상연, 이건명, “토픽 모델링을 이용한 댓글 그래프 기반 소셜 마이닝 기법”, 한국지능시스템학회 논문지, 제24권, 제6호, pp.640-645, 2014.

[22] 남민지, “SNS 해시태그를 이용한 사용자 감정 분류 방법에 관한 연구”, 조선대학교, 석사학위논문, 2015.

[23] 송지민, “뉴스 트윗 탐지를 위한 기계학습 방법”, 한양대학교, 석사학위논문 , 2014.

[24] 박정식, 윤영선, 박래수, “현대통계학 제5판”, pp.332-339, 2012.

[25] 서민구, “R을 이용한 데이터 처리 & 분석 실무”, pp.342-348, 2014.

[26] 홍초희, 김학수, “트윗 감정분류를 위한 다양한 기계학습 자질에 대한 비교 연구”, 한국콘텐츠학회논문지, 제12권, 제12호, pp.471-478, 2012.

[27] 나성희, 김정인, 김판구 “워드넷 유사도를 이용한 뉴스 트윗 카테고리 분류”, 2015한국멀티미디어학회 춘계학술발표대회 논문집, 제18권,제1호, pp.180-183 , 2015.

[28] 나성희, 김판구“단문텍스트의 단어빈도수 및 특징빈도수를 활용한 분류 방법”, 한국 스마트미디어학회 2015 추계학술대회, 제4권, 제2호 pp .205-208 ,

2015.

[29] Yang Yiming and Jan O.Pedersen, "A comparativestudy on feature selection in text categorization.", 14th International Conference on Machine Learning(ICML), pp.412-420, 1997.

[30] 최지명, " 기계학습 알고리즘을 이용한 한국어 텍스트 저자 판별 : 블로그의 영화 리뷰를 대상으로", 연세대학교, 석사학위논문, 2015.

[31] Olson, David L and Dursun Delen. "Advanced data mining technique"s. Springer Science & Business Media, 2008.

[32] Sriram, Bharath, et al. "Short text classification in twitter to improve information filtering." Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2010.

[33] 퍼다나, 김수형, 나인섭, "나이브 베이지안 분류기를 이용한 트위터 트랜딩 주제 분류", 2013 한국정보과학회 제40회 정기총회 및 추계학술발표회, pp.879-881, 2013.

[34] Tare, Mohit, et al. "Multi-Class Tweet Categorization Using Map Reduce Paradigm." International Journal of Computer Trends and Technology (IJCTT), vol.9, no.2, pp.78-81, 2014

[35] Dilrukshi, Inoshika, and Kasun de Zoysa. "A Feature Selection Method for Twitter News Classification." International Journal of Machine Learning and Computing vol4, no.4, pp.365, 2014

[36] Jotikabukkana, Phat, et al. "Effectiveness of Social Media Text Classification by Utilizing the Online News Category."

[37] Wang, Jinpeng, et al. "Mining User Intents in Twitter: A Semi-Supervised Approach to Inferring Intent Categories for Tweets." Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015.

[38] Chun, Yang-Ha. "A SNS Message Type Classification System Using Language Independent Features and Dependent Features." International Journal of Software Engineering and Its Applications vol.9, no.7, pp.151-158, 2015.

[39] Go Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment

classification using distant supervision." CS224N Project Report, Stanford vol.1, no.12, 2009.

[40] Selvaperumal, P., and A. Suruliandi. "A short message classification algorithm for tweet classification." Recent Trends in Information Technology (ICRTIT) 2014 International Conference on. IEEE, 2014.

[41] Meles, Sanai. "Twitter and the Major News Network." , pp.1-11, 2015.

[42] Godea, Andreea Kamiana, et al. "An Analysis of Twitter Data on E-cigarette Sentiments and Promotion."

[43] Culotta, Aron, Jennifer Cutler, and Junzhe Zheng. "Finding Truth in Cause-Related Advertising: A Lexical Analysis of Brands' Health, Environment, and Social Justice Communications on Twitter." The Journal of Values-Based Leadership vol.8, no.2, pp.1-16, 2015.

[44] Yulianti, Evi, Sharin Huspi, and Mark Sanderson. "Tweet-biased summarization." Journal of the Association for Information Science and Technology, pp.1-12, 2015.

[45] Sharma, Abhishek, Yuan Tian, and David Lo. "What's Hot in Software Engineering Twitter Space?", IEEE, pp.541-545, 2015

[46] Edwards, Gordon, and Amy Guy. "Connections between Twitter Spammer Categories." Making Sense of Microposts (# Microposts2015), pp.22-25 , 2015.