博 士 學 位 論 文

기계학습 기반의 한국어 단문 감정분류 기법에 관한 연구

高麗大學校 大學院

컴퓨터學科

鄭 暎 熹

2017年 8月

鄭 舜 榮 敎 授 指 導

博 士 學 位 論 文

기계학습 기반의 한국어 단문 감정분류 기법에 관한 연구

이 論文을 工學 博士學位 論文으로 提出함.

2017年 8月

高麗大學校 大學院

컴퓨터學科

鄭 暎 憙

鄭 暎 熹의 工學 博士學位 論文

審査를 完了함

2017년 8월

委員長 정 순 영 (印)

委 員 박 두 순 (印)

委 員 이 원 규 (印)

委 員 김 현 철 (印)

委 員 임 희 석 (印)

차 례

표 차례

그림 차례

수식 차례

요 약

오피니언 마이닝 분야 중 감성분석은 텍스트에 포함된 내용이 주관적인지 객관적인지 판별하고 작성자의 주관이 드러난 내용에 대해 감성의 극성을 분석하여 긍정(positive), 부정(negative), 중립(neutral) 중 하나로 분석하는 연구 분야이다. 초기 감성분석 연구는 대부분 웹사이트와 소셜 미디어 서비스에 나타난 의견들을 자동으로 분석하여 '긍정/부정' 또는 '좋다/싫다'의 분석 결과를 제공하였다면, 최근 연구에서는 데이터로부터 단순한 긍/부정이 아닌 기쁨, 슬픔, 기대, 공포 등 소비자의 다양한 감정을 인식하는 감정분석에 대한 연구가 시도되고 있다. 단순한 긍정, 부정의 감성분석은 "핸드폰 액정 사이즈가 작아서 손에 쏙 들어오니 글을 입력하기가 편하다", "핸드폰 액정 사이즈가 작아서 앙증맞고 귀엽다"와 같은 문장에서 긍정 이외의 정보를 추출하기는 어렵다. 그러나 보다 세분화된 감정분석에서는 위의 문장에서 '편하다', '귀엽다' 등과 같은 보다 고차원의 감정 추출이 가능하며, 이는 보다 정확한 정보 추출 및 이러한 의견을 분석하여 상품의 품질을 높이는데 기여할 수 있을 것이다. 본 연구에서 '긍정/부정' 또는 '좋다/싫다'와 같이 극성 분류가 아닌, 텍스트 데이터로부터 세분화된 감정을 분류 하는데 목적이 있다. 따라서 본 연구에서는 긍/,부정의 극성 분류를 감성분석이라 정의하고, 세분화된 다양한 감정을 분류하는 것을 감정분석으로 정의하였고, 한국어 기반 텍스트 데이터에서 세분화된 감정분류를 위한 표층적 자질 집합과 의미적 자질 집합을 설계하고, 감정분류의 정확도를 높이고자 하였다.

단순한 긍정, 부정이 아닌 '기쁨', '편안', '슬픔', '불안' 등과 같은 다양한 감정에 대해 인식하는 분류체계는 적용하는 응용 시스템에 따라 감점 범주의 종류가 제각각인 경우가 많다. 특히 국내의 경우, 감정에 대한 극성 분류 연구가 대부분이었으며, 여러 감정으로 분류한 연구는 아직 미비한 상태이다. 국외 감정에 대한 대표적인 분류체계는 플러칙(Plutchik)의 8 개의 분류체계(기쁨, 신뢰, 두려움, 놀람, 슬품, 혐오, 화남, 기대)이다.

플러칙은 인간의 기본 감정을 8개로 구분하였으며, 우리가 일반적으로 경험하는 대부분의 감정들은 이 여덟 가지 감정들이 서로 혼합하여 나타난다고 주장하였다. 기분상태검사(profIle of mood States, POMS)의 6개 분류체계(tension, depression, anger, vigor, fatigue, confusion)도 자주 사용되는 감정 범주이다. 그러나 이와 같은 감정 분류체계는 영어로 작성되어 있기 때문에 그 의미에 맞게 한국어로 번역해서 접근해야 하는 한계점이 존재한다. 본 연구에서는 이러한 문제점을 해결하기 위해 한국어 감정분류를 위해 한국어 감정어휘 목록을 통해 한국어 감정범주 25개를 선별하였다. 한국인들의 실정에 맞는 434개의 한국어 감정어휘 목록 중 친숙성 평가기준을 근거로 상위 25개의 감정단어를 선별하여 사용하고, 이를 토대로 학습 말뭉치를 구축하였다.

감정분류를 위한 접근방식은 최근 연구에서도 자주 사용되는 기계학습 기법을 사용하였다. 25개의 세분화된 감정들에 대해 지도학습을 수행하기 위해서는 충분한 학습 말뭉치가 필요하다. 그러나 한국어 감정범주에 대한 학습 말뭉치의 부재는 세분화된 감정분류를 어렵게 하고 있으며, 공개된 한국어 감정 말뭉치는 긍정, 부정에 대한 극성 정보만 담고 있는 한계점이 있었다. 따라서 본 연구에서는 다양한 감정분류를 위해 25개 감정범주에 대한 학습 말뭉치를 구축하였고, 표층적 자질과 의미적 자질을 설계하였다. 표층적 자질은 텍스트 데이터에서 감정분류를 위한 즉각적이고 일차원적인 근거를 제공한다. 그러나 텍스트 데이터가 내포하고 있는 의미를 담아내기에는 한계가 있다. 기존 연구에서는 이러한 의미를 표현하기 위해 다양한 자질들을 조합하거나 새로운 자질을 개발하여 학습에 사용하기도 하였다. 그러나 자연어 처리 분야에서, 단어를 '의미' 벡터 공간으로 임베딩하는 기법들이 소개되면서 감정분류 분야에서도 이러한 기법들을 사용하여 자질을 자동으로 학습하는 접근법이 시도되고 있다. '의미'는 단어 뿐만을 대상으로 하지 않고 구, 문장, 문서 등 자연어처리 과정에서 발생하는 모든 처리 단위가 대상이 된다. 대량의 데이터에서 단어나 문장을 의미 벡터 공간으로 임베딩하여 주변 단어와 문장과의 문맥 정

보를 활용하여 학습모델을 만들고, 기계학습 알고리즘을 통해 감정을 분류한다. 따라서 본 연구에서는 표면적 자질 뿐만 아니라 워드임베딩 기법을 활용한 doc2vec, skip-thought vector와 같은 의미적 자질을 사용하였으며, 표면적 자질과의 성능을 비교 분석하였다.

감정분류에 대한 또 다른 문제로 학습에 사용되는 어휘 자질의 과도한 사용으로 인해 데이터 부족 문제를 유발한 다는 것이다. 예를 들면, “나 정말 화가나(화남)”, “진짜 재밌다(재미)” 등에 대해서는 학습 말뭉치에 나타난 고빈도 감정 어휘를 사용하였기 때문에 잘 동작하지만, 학습 말뭉치에 나타나지 않은 “드디어 성공했어(기쁨)” 등의 표현에 대해서는 적절한 감정을 결정할 수 없게 된다. 본 연구에서는 25개의 세분화된 감정범주를 사용하기 때문에 이처럼 특정 감정범주를 위한 데이터 부족 문제가 발생한다. 감정범주의 개수가 많아질수록 해당 감정의 데이터를 수집하기는 점점 어려질 것이다. 따라서 본 연구에서는 데이터 오버샘플링 기법을 통해 이 문제를 해결하여 학습 데이터 부족 문제를 해결하고자 하였다.

본 연구는 한국어 문장에서 25개의 다양한 감정범주로 감정을 분류한 그 자체로도 의미가 있지만 다음과 같은 이유에서 더욱 중요하다.

첫째, 본 연구는 한국어로 작성된 텍스트 데이터를 분석하여 세분화된 25개의 감정범주를 분류하였고 높은 정확도를 제공한다. 세분화된 25개의 감정범주는 한국인 실정에 맞는 감정범주로 정의하였고, 다양한 감정범주를 분류하였다.

둘째, 최신 기술인 워드 임베딩 학습 기법을 사용하여 의미적 자질을 설계하고, 이를 한국어 감정분류에 적용하였다. 워드 임베딩 학습 기법은 학습 속도가 매우 빠르며, 표면적 자질을 사용한 학습결과와 비슷한 성능을 제공한다.

셋째, 감정분류를 위해 사용된 표면적 자질과 의미적 자질에 성능평가를 통해 각각의 자질에 대한 성능을 비교 분석하였으며, 이는 감정분류에 있어서 표면적 자질과 의미적 자질에 대한 실증적 실험결과를 제공하

고 이를 통해 한국어 감정분류에 있어서 앞으로 연구방향을 설정함에 있어 기초자료로 사용될 수 있다.

넷째, 본 연구에서는 25개의 감정을 분류하기 위한 학습 말뭉치를 전문가를 통해 직접 구축하였고, 감정분석 연구와 실무에 활용이 가능한 자원을 공유한다는 차별성을 가진다.

제 1 장 서 론

오피니언 마이닝 분야 중 감성분석은 텍스트에 포함된 내용이 주관적인지 객관적인지 판별하고 작성자의 주관이 드러난 내용에 대해 감성의 극성을 분석하여 긍정(positive), 부정(negative), 중립(neutral) 중 하나로 분석하는 연구 분야이다[1]. 본 장에서는 연구 배경과 기존 감정분류 문제점 및 연구 목적에 대한 내용을 설명한다.

제 1 절 연구의 배경

오피니언 마이닝 분야 중 감성분석은 텍스트에 포함된 내용이 주관적인지 객관적인지 판별하고 작성자의 주관이 드러난 내용에 대해 감성의 극성을 분석하여 긍정(positive), 부정(negative), 중립(neutral) 중 하나로 분석하는 연구 분야이다[1].

최근 인터넷 상의 소셜 미디어 서비스(트위터, 페이스북, 인스타그램, 마이스페이스 등)의 확산은 이러한 서비스를 누구나 손쉽게 접할 수 있게 도와주었으며, 이를 통해 생산된 텍스트, 이미지, 동영상들은 인터넷을 사용하는 여러 소비자 계층에 의해 읽히고 공감되는 자원이 되었다. 셀 수 없이 쏟아지는 이러한 자원들은 자동화된 정보 분석기법, 그 중 감성분석 기법에 의해 소비자들의 동향을 파악하는데 사용되어 왔다. Kim, S. Y., Park, S. T., & Kim, Y. K.[2] 연구에 따르면 기업들은 자사의 이미지뿐만 아니라 타 회사의 제품 및 서비스에 대한 소비자들의 성향을 분석하고 벤치마킹하여 수익 창출에 활용하고자 하였으며, Choi, S., & Choi, K.[3] 에 따르면 개인들은 관심 있는 분야에 대한 이미 구입한 다른 사람들의 의견을 듣고 파악하여 자신의 생각과 비교하거나 구매 전에 자신과 유사한 성향의 사람들의 의견을 신뢰하고 이를 구매에 반영한다고 하였다.

초기 감성분석 연구는 대부분 웹사이트와 소셜 미디어 서비스에 나타난 의견들을 자동으로 분석하여 ‘긍정/부정’ 또는 ‘좋다/싫다’의 분석 결과를 제공하였다면[4-9], 최근 연구에서는 데이터로부터 단순한 긍/부정이 아닌 기쁨, 슬픔, 기대, 공포 등 소비자의 다양한 감정을 인식하는 감정분석에 대한 연구가 시도되고 있다[10]. 단순한 긍정, 부정의 감성분석은 “핸드폰 액정 사이즈가 작아서 손에 쏙 들어오니 글을 입력하기가 편하다”, “핸드폰 액정 사이즈가 작아서 앙증맞고 귀엽다”와 같은 문장에서 긍정 이외의 정보를 추출하기는 어렵다. 그러나 보다 세분화된 감정분석에서는 위의 문장에서 ‘편하다’, ‘귀엽다’ 등과 같은 보다 고차원의 감정 추출이 가능하며, 이는 보다 정확한 정보 추출 및 이러한 의견을 분석하여 상품의 품질을 높이는데 기여할 수 있을 것이다. 본 연구에서 감성분석이라 함은 ‘긍정/부정’ 또는 ‘좋다/싫다’와 같이 극성 분류를 위한 의미로 정의 하였으며, 감정분석 및 감정분류라 함은 문장을 보다 세분화된 감정으로 분류하는 것으로 정의하였다.

최근 트랜드 마케팅 분석, 상품 리뷰 데이터 분석, 특정 인물에 대한 선호도 분석에서도 보다 자세하고 세분화된 감정분석 결과를 요구하고 있으며, Kim, M., & Park, S. O.[11]연구에서 자동차, 전자제품 등 상품 설계 및 디자인 분야에서도 사용자의 모바일 사용패턴 속에서 감정적 요소를 분석하여 대중의 감정을 상품 디자인에 반영한 실제 디자인 사례를 제시하였다. 감정 지향적 소비자는 연령, 성별, 교육, 건강의 개인적 요인과 사회문화적 요인의 영향을 받으며 자신의 감정에 맞는 제품을 소비하려는 경향이 높게 나타나며, 이러한 감정이 반영된 상품에 대한 소비자 만족도가 높게 나타난다. 특히 4차 혁명으로 인공지능에 대한 관심이 높아지면서 인공지능 로봇 또는 실생활 전 분야에서 사용자의 감정을 이해하고 공감하려는 연구가 다양한 형태로 진행되고 있다[12, 13].

단순한 긍정, 부정이 아닌 ‘기쁨’, ‘편안’, ‘슬픔’, ‘불안’ 등과 같은 다양한 감정에 대해 인식하는 분류체계는 적용하는 응용 시스템에 따라 감점 범주의 종류가 제각각인 경우가 많다. 특히 국내의 경우, 감정에 대한 극

성 분류 연구가 대부분이었으며, 여러 감정으로 분류한 연구는 아직 미비한 상태이다. 국외의 경우, Alm, C. O., & Sproat, R.[14]연구에서 이미 세분화된 감정분류(categorizing anger, disgust, fear, joy, sadness, positive surprise, and negative surprise into positive, negative, and neutral) 연구가 진행되었기 때문에 이에 비하면 국내의 감정분석 연구는 한참 저조한 실정이다. 국외 감정에 대한 대표적인 분류체계는 플러칙(Plutchik)의 8 개의 분류체계(기쁨, 신뢰, 두려움, 놀람, 슬품, 혐오, 화남, 기대)이다[15, 16]. 플러칙은 인간의 기본 감정을 8개로 구분하였으며, 우리가 일반적으로 경험하는 대부분의 감정들은 이 여덟 가지 감정들이 서로 혼합하여 나타난다고 주장하였다. 기분상태검사(profIle of mood States, POMS)의 6개 분류체계(tension, depression, anger, vigor, fatigue, confusion)도 자주 사용되는 감정 범주이다[17, 18]. 그러나 이와 같은 감정 분류체계는 영어로 작성되어 있기 때문에 그 의미에 맞게 한국어로 번역해서 접근해야 하는 한계점이 존재한다.

감정분류를 위한 접근방법은 어휘기반 기법과 기계학습 기법이 있다. 어휘기반 기법은 시소러스, 온톨로지, 감정기반어휘목록, 감정사전 등과 같은 리소스를 필요로 하며, 문장이나 문서에 포함된 단어들을 감정어 리소스에서 찾아 해당 단어의 감정을 추출하는 방법이다[19, 20]. 어휘기반 접근방식은 이미 보유한 어휘자원을 바탕으로 범주화 작업을 선행하여 감정어 리소스를 구축하고, 감정어 리소스를 바탕으로 분석하고자 하는 텍스트 데이터에 대한 감정점수를 계산하여 분류하는 기법을 일컫는다. 이처럼 어휘기반 감정분석은 감정어 리소스 등의 어휘자원에 대한 의존도가 높기 때문에 성능향상을 위해서는 감정어 리소스의 구축에 많은 노력이 필요하다. 즉, 감정어 리소스에 포함되지 않은 어휘가 실제 문장에서 사용될 경우, 이를 찾을 수 있는 방법은 없다. 특히 소셜 미디어 서비스에서 사용되는 신조어, 약어, 의미 없는 문장, 광고 등 예측할 수 없는 내용과 구어체로 이루어진 완전하지 않은 문장은 감정분류의 어휘기반 기법을 더욱 어렵게 하고 있다. 영어권에서는 다양한 외국어 어휘접인

SentiWordNet[21]이 다양한 연구에서 활용되고 있다. 한국어의 경우, 감정어 리소스 부재로 SentiWordNet과 같은 리소스를 한국어로 번역하여 사용하고 있기 때문에 감정 분류체계처럼 영어권의 감정을 한국어의 감정으로 그대로 번역하여 반영하기에는 어려움이 존재한다.

기계학습 접근방법은 지도학습과 자율학습으로 구분되며, 서포트 벡터 머신, 신경망 또는 네이브 베이지안 알고리즘과 같은 다양한 기법들을 통해 수행된다. 지도학습의 감정분석에서는 다양한 자질(feature)을 사용하여 데이터를 학습시킨다[22-25]. 이때 학습에 사용되는 학습 말뭉치는 해당 도메인에 맞춰서 구축되었기 때문에 해당 도메인에서 높은 성능 향상을 기대할 수 있다. 그러나 학습 말뭉치 구축을 위해서는 해당 분야의 전문가가 필요로 하며, 말뭉치 구축을 위한 시간과 비용이 소요된다. 이러한 학습 말뭉치는 해당 도메인의 데이터의 특성에 많은 영향을 받는다. 예를 들면, 컴퓨터 도메인에서 '자바'와 커피 도메인에서 '자바'는 다른 의미를 갖기 때문에 학습 말뭉치 구축 시에 다른 속성으로 태깅해야 한다. 이러한 도메인 종속적인 특징 때문에 특정 도메인에서 구축된 학습 말뭉치는 다른 도메인으로 전이하여 사용될 때 성능이 떨어질 수밖에 없다. 감정분석 분야에서도 이러한 도메인 종속적인 문제를 해결하기 위한 연구가 진행되었다[22].

본 연구에서는 한국어 감정분류를 위해 Park, I. J., & Min, J. K.[26]에서 한국어 감정어휘 목록을 통해 한국어 감정범주 25개로 선별하였다. Park, I. J., Min, J. K.[26]에서는 한국인들의 실정에 맞는 434개의 한국어 감정어휘 목록을 제공하였으며, 이 중 친숙성 평가기준을 근거로 상위 25개의 감정단어를 선별하여 사용하였다.

감정분류를 위한 접근방법은 최근 연구에서도 자주 사용되는 기계학습 기법을 사용하였다. 25개의 세분화된 감정들에 대해 지도학습을 수행하기 위해서는 충분한 학습 말뭉치가 필요하며 이를 위해서 본 연구에서는 25개 감정에 대한 학습 말뭉치를 구축하였고, 이모티콘, 감정어휘, 품사정보, 구문구조정보 등을 주요 자질로 제안하였다. 이러한 표층적 자질은

텍스트 데이터에서 감정분류를 위한 즉각적이고 일차원적인 근거를 제공한다. 그러나 텍스트 데이터가 내포하고 있는 의미를 담아내기에는 한계가 있다. 기존 연구에서는 이러한 의미를 표현하기 위해 다양한 자질들을 조합하거나 새로운 자질을 개발하여 학습에 사용하기도 하였다. 그러나 자연어 처리 분야에서, 단어를 '의미' 벡터 공간으로 임베딩하는 기법들이 소개되면서 감정분류 분야에서도 이러한 기법들을 사용하여 자질을 자동으로 학습하는 접근법이 시도되고 있다. '의미'는 단어 뿐만을 대상으로 하지 않고 구, 문장, 문서 등 자연어처리 과정에서 발생하는 모든 처리 단위가 대상이 된다. 대량의 데이터에서 단어나 문장을 의미 벡터 공간으로 임베딩하여 주변 단어와 문장과의 문맥 정보를 활용하여 학습모델을 만들고, 기계학습 알고리즘을 통해 감정을 분류한다. 따라서 본 연구에서는 표면적 자질 뿐만 아니라 워드임베딩 기법을 활용한 doc2vec[27], skip-thought vector[28] 을 의미적 자질로 사용하였고, 표면적 자질과의 성능을 비교 분석하였다.

기존 한국어 감정분류 연구는 한국어 감정범주에 대한 공식적인 분류체계가 아직 마련되어 있지 않으며, 영어권의 SentiWordNet과 같은 한국어 감정어 리소스 또한 부족한 실정이다. 이는 기계학습 접근방법에서 사용되는 학습 말뭉치의 부재로 이어지며 한국어 감정분류의 어려움을 시사하고 있다. 또한 학습에 사용되는 어휘 자질의 과도한 사용은 데이터 부족 문제를 유발한다. 예를 들면, "나 정말 화가나(화남)", "진짜 재밌다(재미)", 등에 대해서는 학습 말뭉치에 나타난 고빈도 감정 어휘를 사용하였기 때문에 잘 동작하지만, 학습 말뭉치에 나타나지 않은 "드디어 성공했어(기쁨)" 등의 표현에 대해서는 적절한 감정을 결정할 수 없게 된다. 본 연구에서는 25개의 세분화된 감정범주를 사용하기 때문에 이처럼 특정 감정범주를 위한 데이터 부족 문제가 발생한다. 감정범주의 개수가 많아질수록 해당 감정의 데이터를 수집하기는 점점 어려질 것이다. 따라서 본 연구에서는 데이터 오버샘플링 기법을 통해 이 문제를 해결하여 학습 데이터 부족 문제를 해결하고자 하였다.

제 2 절 연구의 목적

본 연구에서는 한국어로 작성된 소셜 미디어 데이터에서 감정을 25개로 범주화하고 각각의 감정으로 분류하기 위한 접근방식을 제안한다. 감정분류를 위해 한국어 감정 말뭉치를 구축하였으며, 기계학습에 사용되는 자질을 추출하기 위해 한국어 어휘의 특징을 고려한 표층 자질 집합과 워드 임베딩을 통해 자동으로 추출한 의미적 자질 집합을 소개한다.

한국어 감정분류에 영향을 미치는 표층 자질 집합을 문장레벨과 문서레벨로 구분하였고, 문장레벨에서 한국어의 형태학적 특징을 자질로 사용하였다. 실험을 통하여 자질들의 최적 조합을 구성하여 감정분류를 위해 SVM 분류기를 사용하였다. 이러한 기계학습에서는 학습 코퍼스의 역할이 매우 중요하다. 영어권에서는 WordNet, SentiWordNet[16], LKB 등 다수의 프로젝트들이 존재한다. 그래서 한국어의 많은 감정 분석 연구들이 영어권의 학습 코퍼스를 한/영으로 번역하여 이용하거나, 인터넷 사용자의 의견 및 감정 정보를 제공하는 사이트로부터 학습 코퍼스를 구축한다. 최근 국내의 경우에도 연구와 실무에 사용하도록 '한글 감성어 사전'[29]이 제공되었지만, 긍정, 중립, 부정에 대한 극성 정보만을 담고 있다. 본 연구에서는 25개의 세분화된 감정들에 대해 지도학습을 위한 학습 말뭉치를 구축하였으며, 이를 통해 명시적으로 정의한 표층 자질 집합들 중에서 한국어 감성분류에 효율적인 자질이 무엇인지 파악하고자 하였다.

의미적 관계를 학습에 사용하여 감정분류의 효율성을 높이기 위해 본 연구에서는 워드 임베딩 기법 중 두 가지 알고리즘을 적용하여 자질 벡터를 추출하고 이를 학습하여 감정을 분류하였다. 한국어의 복잡한 특징을 언어모델을 통해 학습하여 자동으로 의미적 자질 집합을 추출하였고, 학습을 위한 감정 말뭉치를 사용하지 않기 때문에 말뭉치를 구축하는데 드는 비용과 시간을 절약 할 수 있었다. 그러나 의미적 자질 집합이 벡터 공간에 표현될 때 주변 어휘와의 관계로 표현되기 때문에 의미가 정반대인 어휘가 벡터 공간에서 유사한 어휘로 표현되기 때문에 표층 자질 집합

이 갖고 있는 형태론적, 구문적 정보를 조합하여 감정을 분류에 사용하였다.

본 연구는 한국어 문장에서 25개의 다양한 감정범주로 감정을 분류한 그 자체로도 의미가 있지만 다음과 같은 이유에서 더욱 중요하다.

첫째, 본 연구는 한국어로 작성된 텍스트 데이터를 분석하여 세분화된 25개의 감정범주를 분류하였고 높은 정확도를 제공한다. 세분화된 25개의 감정범주는 한국인 실정에 맞는 감정범주로 정의하였고, 다양한 감정범주를 분류하였다.

둘째, 최신 기술인 워드 임베딩 학습 기법을 사용하여 의미적 자질을 설계하고, 이를 한국어 감정분류에 적용하였다. 워드 임베딩 학습 기법은 학습 속도가 매우 빠르며, 표면적 자질을 사용한 학습결과와 비슷한 성능을 제공한다.

셋째, 감정분류를 위해 사용된 표면적 자질과 의미적 자질에 성능평가를 통해 각각의 자질에 대한 성능을 비교 분석하였으며, 이는 감정분류에 있어서 표면적 자질과 의미적 자질에 대한 실증적 실험결과를 제공하고 이를 통해 한국어 감정분류에 있어서 앞으로 연구방향을 설정함에 있어 기초자료로 사용될 수 있다.

넷째, 본 연구에서는 25개의 감정을 분류하기 위한 학습 말뭉치를 전문가를 통해 직접 구축하였고, 감정분석 연구와 실무에 활용이 가능한 자원을 공유한다는 차별성을 가진다.

본 연구의 구성은 다음과 같다. 2장에서는 감정 분석과 관련된 연구에 대해 소개하고, 3장에서는 본 논문에서 제안한 감정분류를 위한 프레임워크에 대해 소개하고 4장에서는 구축한 말뭉치 분석과 감정분류 실험결과를 소개하고, 5장에서는 본 연구의 요약 및 연구의 의의를 제시한다.

제 2 장 관련연구

한국어 문장을 대상으로 하는 감정분류는 대부분 긍정, 부정에 대한 연구가 대부분이었다. 그러나 국외에서는 이미 긍정, 부정, 중립 이외의 다양한 감정범주로 분류하는 연구가 진행 되었으며, 영어뿐만 아니라 자연어 처리가 어려운 교착어(그리스어)를 대상으로 하는 연구도 수행되고 있다. 본 장에서는 먼저 한국어 감정분류를 위한 기존 감정범주에 대해 살펴보고, 지도학습과 자율학습을 적용한 감정분류 기법과 본 연구에서 학습을 위해 사용되는 의미적 자질 설계를 위해 사용되는 워드 임베딩을 소개한다. 이러한 관련연구 분석을 통해 기존 한국어 감정분류에 있어서 문제점이 무엇인지 살펴보고자 한다.

제 1 절 감정범주 정의

감정분류에 대한 연구는 긍정, 부정, 중립에 대한 극성 판단에 대한 연구가 대부분이다. 감정분류에 대한 연구는 영어권에서 먼저 활발히 연구가 진행되었기 때문에 한국어보다 영어권에서 감정분류범주가 더 다양하게 존재한다.

김윤석[30] 연구에서는 한글 텍스트 감정 분류를 위해 감정 사전을 구축하였다. 이때 Ekman[19]이 정의한 기본 감정 여섯 가지(‘기쁨’, ‘슬픔’, ‘공포’, ‘분노’, ‘혐오’, ‘놀람’)와 HCI(인간-컴퓨터상호작용)[31]에서 활용도가 높은 세 가지(‘흥미’, ‘지루’, ‘통증’) 감정 범주를 포함하여 9개의 감정 범주를 제시하였다. 감정 범주를 긍정(‘기쁨’, ‘놀람’, ‘흥미’)과 부정(‘공포’, ‘분노’, ‘혐오’, ‘통증’, ‘지루’, ‘슬픔’)으로 나누어 감정 사전을 구축하였다.

표 1 한국어 감정 도메인 (Kim, M. K.[32])

긍정	기쁨 (a)	감개무량, 감격, 감명, 감흥, 강추, 경애, 경쾌 ..
	안심 (b)	고객감동, 공감하다, 괜찮다. 교감하다. 느까다. 맘 놓다, 미소, 믿다, 안도하다, 안락하다. 안심하다…
	만족 (c)	가격만족, 가격저렴, 값싸다, 경탄하다, 굿, 귀엽다, 깔끔하다, 깨끗하다, 친절하다…..
	재미 (d)	매료되다, 매혹적이다, 박진감, 반하다, 살맛나다, 설레다, 신나다, 열광하다, 열렬하다, 열애…
	긍지 (e)	가슴뿌듯하다, 감복하다, 감사하다, 경외하다, 감탄하다, 긍지, 떳떳하다, 보람차다, 뿌듯하다….
부정	분노 (f)	가증스럽다, 개탄하다, 격노하다, 노기, 격분하다, 고깝다, 괘씸하다, 골나다, 굴욕, 어이없다
	공포 (g)	강압, 경악하다, 곤혹스럽다, 공포, 기겁하다, 끔찍하다, 무섭다, 두렵다, 무시무시하다, 섬뜩하다
	혐오 (h)	개쓰레기, 거부감, 경멸하다, 가소롭다, , 기만하다, 농락하다, 기고만장하다, 꺼리다, 당혹하다, 만행, …
	슬픔 (i)	가련하다, 가엾다, 고뇌하다, 고독하다, 고적하다, 고롭다, 낙담하다, 그립다, 낙망하다, 낙심하다, 망연자실하다..
	불만 (k)	가입거부, 가짜, 감언이설, 거부, 거절당하다, 결점, 결함, 계약위반, 고발하다, 고소하다, 과실, 권태롭다.

Kim, M. K.[32] 연구에서는 10개의 감정범주를 정의하고, 문서에 대한 리뷰(review)로부터 개체 대상, 감정 언어, 극성을 나타내는 언어, 강조어, 동사를 문장에서 추출하여 크게는 긍정, 부정 요소로 분류하는 것으로 세부적으로는 긍정 부정 요소를 10개로 소분류 하여 문서상에 나타난 의견을 평가하였다. <표 1>는 Kim, M. K.[32] 연구에서 사용된 한국어 감정 도메인을 보여준다.

이철성, 최동희, 김성순, & 강재우[17] 연구에서는 한글 문서를 기반으로 기계학습 모델을 적용하여 7개의 감정으로 분류하고 그 결과를 영화평에 적용하여 영화 장르별 감정특성을 분석하였다.

<표 2>는 국내에 소개된 감정범주의 개수이다. 이외의 연구에서는 감정의 극성(긍정, 부정, 중립)에 대한 분류가 대부분이며 아직 한글 문서에서 세분화된 감정을 식별하고 분류하는 연구는 국외 연구에 비해 미비하다.

표 2 국내 소개된 감정범주의 개수

연구명	감정범주의 개수	방법	정확도
김윤석[30]	9개(긍정3, 부정6)	SVM	64%
Kim, M. K.[32]	10개 (기쁨, 안심, 만족, 재미, 긍지, 분노, 공포, 혐오, 불만, 슬픔)	통계적 방법	
이철성[17]	7개(분노, 혼란, 우울, 피로감, 친근감, 긴장감, 생동감)	베이즈확률모델	52%

본 연구에서는 한국어 사정에 맞는 감정 범주를 결정하기 위해 Park, I. J., & Min, J. K.[33] 연구에서 제안한 한국어 감정어휘 목록을 활용하였다. Park, I. J., & Min, J. K.[33] 연구에서는 한국인들의 실정에 맞는 434개의 한국어 감정어휘 목록을 4가지 평가기준인 원형성, 친숙성, 쾌-불쾌, 활성화로 선별하여 제시하였다. 이 목록에서 소셜 미디어에 사용되는 감정어휘와 가장 관계가 깊은 평가기준인 친숙성을 기준으로 상위 25개의 감정 단어를 선별하여 본 연구의 감정 범주로 사용하였다.

제 2 절 감정분류 기법

감정분류는 주어진 텍스트를 감정에 따라 분류하는 것이다. 이러한 텍스트 분류 문제를 해결하기 위해 여러 가지 방법들이 제안되었다. 데이터의 정량적인 실험을 통한 통계기반 기법과 기계학습을 통한 감정을 분류하는 연구가 진행되고 있으며, 자연어처리 기법과 통계적, 기계학습법을 혼합하여 분석을 진행하기도 한다. 기계학습은 학습 말뭉치를 사용하는 지도학습과 학습 말뭉치 없이 학습을 진행하는 자율학습으로 구분되며, 감정분류를 위한 지도학습에 관한 연구와 자율학습에 관한 연구를 살펴보고자 한다.

2.1 지도학습

지도학습은 학습 말뭉치로부터 함수를 만들어내는 기계학습을 의미하며, 학습 말뭉치는 입력 대상의 쌍과 원하는 출력으로 구성된다. 지도학습기는 단지 소수의 훈련(입력 쌍과 목표 출력)만으로 유효한 입력값에 대한 함수의 값을 예측한다. 이는 함수의 출력으로 연속적인 값, 분류명 등을 예상할 있다. 이를 위해 학습기는 논리적이고 이성적인 방법으로 학습 말뭉치로부터 보이지 않는 상황까지 일반화해야 한다. 지도학습에서 주어진 문제를 해결하기 위해서는 일반적으로 다음 과정을 거친다.

1. 학습 데이터의 유형을 결정한다. 사용되는 데이터가 어떤 종류의 데이터 인지를 결정해야 한다. 한 개의 단어인지, 학습 데이터의 전체 단어인지, 학습 데이터 내 전체 문장인지와 같은 것이다.

2. 학습 말뭉치를 구축한다. 일반적으로 전문가에 의해 구축된 학습 말뭉치를 사용한다.

3. 학습 함수의 자질을 결정한다. 학습 함수의 정확성은 입력 대상이 어떻게 표현되느냐에 의해 크게 좌우된다. 보통 입력 대상은 자질 벡터로 바뀌고, 대상을 묘사하는 특징적인 수를 포함한다. 특징의 개수는 차원의

한계 때문에 너무 커서는 안 되지만, 출력을 예상할 수 있을 정도로 충분해야 한다.

4. 학습 함수의 구조와 동등한 학습 알고리즘을 결정한다. 대표적으로 지지기반 벡터 모델, 은닉 마르코프 모델, 회귀 분석 등이 있다.

5. 설계를 완성하고, 학습 코퍼스 상에서 선택한 학습 알고리즘을 통해 알고리즘에서 사용하는 인수들을 교차-검증데이터(cross-validation set)에서 성능을 최적화함으로써 조정된다. 학습 후, 알고리즘 성능은 학습 코퍼스에서 분리된 테스트 집합을 사용하여 측정한다.

이철성, 최동희, 김성순, & 강재우[17] 연구에서 한글 문서를 기반으로 기계학습 중 다항 네이브 베이지안 모델을 사용하여 인간의 감성을 7개의 감성으로 분류하였다. 정서 이론에서 사용하는 기분상태검사 이론을 적용하여 긴장감(tension), 우울(depression), 분노(anger), 활력(vigor), 피로(fatigue), 혼란(confusion), 친근감(friendliness)으로 분류하였으며, 그 결과를 영화평에 적용하여 영화 장르별 감성특성을 분석하였고, 실제 응용분야에 적용 가능함을 보여주고 있다.

이동엽, 조재춘, & 임희석[34] 연구는 아마존 패션 상품 리뷰 데이터를 학습하여 형성된 워드임베딩 공간을 이용하여 사용자의 감성을 분석하는 모델을 구축하였다. SVM 분류기 모델을 사용하여 해당 상품에 대한 긍정, 부정에 대해 극성을 분류하였으며, 88.0%의 정확도를 나타내었다.

Lee, G. H., & Lee, K. J.[35] 연구는 최근 동향을 나타내는 키워드를 신문기사로부터 추출하고, 추출된 키워드를 이용하여 수집된 트윗의 감성을 분석하였다. 토픽 키워드는 신문기사를 k-means 알고리즘을 이용하여 군집화한 후, 군집 내 단어 출현 빈도를 사용하여 토픽 키워드를 추출하였다. 감성분류를 위해 사용된 자질은 자연어 처리 관련 자질인 unigram, bigram, 트위터와 관련된 자질로는 긍정, 부정, 해쉬태그를 사용하였으며, 긍정, 부정의 이모티콘, 본문내 URL 존재 여부 등을 사용하였다. 이 중 unigiram과 bigram, 트위터 관련 자질을 모두 사용했을 때 74.46%의 정확도를 보였다. 이 논문은 트윗의 주제를 고려하지 않았을 때와 주제를 고려했을 때 감성분석 결과를 비교하였다. 토픽 주제를 고려하였을 때, 감

성 분석의 정확도가 그렇지 않은 경우보다 낮게 나왔으며 이는 해당 토픽과 감성 사이의 관계를 고려할 필요가 있음을 나타낸다.

김유영, & 송민[36] 연구는 네이버 영화 리뷰 데이터를 사용하여 기계학습 기반의 감성 분류기를 구축하고, 이를 통해 리뷰의 감성점수를 계산하고 데이터를 수치화하였다. 이 논문에서 사용된 자질은 단어 기반의 n-gram 사전과 글자 기반의 n-gram 사전이다. 자질 벡터에 많은 단어가 들어가 있기 때문에 차원수가 높을 수밖에 없으며, 이를 축소하기 위해 감성 분류 에서 가장 높은 성능을 보인 지지벡터 분류기로 감성을 분류하였다.

류진걸, & 신동민[37] 연구에서 감정분류의 정확도를 높일 수 있는 형태소를 이용한 자질을 지지기반벡터모델을 통해 추출하였으며, 추출된 특성을 조건부랜덤필드에 은닉상태를 도입하여 입력값의 잠재구조를 파악하여 특정 단어와 인접 단어들의 연관성을 고려한 긍정과 부정의 문장을 분류하는 모델을 제안하였다.

김윤석, & 서영훈[30] 연구에서 나이브 베이지안 알고리즘을 이용하여 한극 텍스트의 감정분류를 수행하였다. 감정 범주화는 Ekman(1971) 이 제안한 기본 감정 여섯 가지(기쁨, 슬픔, 고포, 분노, 혐오, 놀람)와 HCI에서 활용도가 높은 세 가지(흥미, 지루, 통증) 감정 범주를 정의하였다. 자주 출현한 감정 단어가 포함된 데이터는 감정분류가 될 가능성이 높다는 가정 하에 감정지수를 계산하고 점수가 높은 데이터는 감정분류가 잘 되도록 판단한다. 영화평 데이터에 대하여 약 64%의 정확률을 보이며, 학습 데이터가 특정 감정분류에 치우쳐 있는 데이터 희소 문제가 발생하고 있음을 시사하고 있다.

Kiritchenko[38]는 통계 텍스트 분류 접근법을 사용하여 트위터와 단문메시지(SMS)의 감성을 분석하였다. 메시지 수준 작업 즉, 메시지에서 긍정적, 부정 또는 중립적인 감정을 감지하고, 용어 수준 작업, 즉 주어진 용어에서 긍정적이거나 부정적인 감정을 감지하는 작업을 수행했습니다.

학습에 사용 된 데이터는 SemEval-2013 데이터 세트이며 부정적인 단어를 구별하고 극성을 올바르게 분류하기 위해 트윗 별 어휘집을 사용했습

니다.

한국의 마이크로 블로그 텍스트에서 감정 분석 연구는 한국의 문서를 기반으로 기계 학습 모델을 적용하고 인간 감성을 7가지 감정으로 분류했다. 감정 이론에 사용 된 POMS 이론을 적용하여 감정을 분류하고 그 결과를 영화 리뷰에 적용하고 영화 장르별 감정을 분석하였다[39]. 이 방법은 비교적 높은 정확도를 제공하지만 학습 말뭉치를 필요로 하기 때문에 말뭉치가 없으면 적용하기가 쉽지 않다.

지도학습에서 학습 말뭉치 구축이라는 한계를 극복하기 위해 대중 소셜 미디어에서의 정서 분석 연구는 셀프지도 알고리즘을 적용하여 '정서 표시가 있는 데이터'의 라벨을 결정하기 위해 '정서 라벨이 있는 데이터'를 사용하고, 이 레이블을 사용하여 데이터를 확장하여 학습에 사용하였다[40]. 학습에 필요한 데이터를 학습하는 대신에, 긍정과 부정적인 정서를 분류하기 위해 '정서적 라벨이 있는 데이터'에 자체 학습 알고리즘을 적용한다. 그러나 이 방법은 처음에는 '감정 라벨이 있는 데이터'가 필요하다는 제한점을 갖는다.

트윗 데이터의 감정분류 연구는 데이터 특성 때문에 높은 성능의 감정 분류 시스템 구현이 어렵다. 이에 기본 어휘 자질(명사, 동사, 형용사, 부사, 보조용언, 미등록어, 외국어)뿐만 아니라 트윗 데이터로부터 추출할 수 있는 네 가지 자질(이모니콘의 극성, 리트윗의 극성, 사용자 극성, 대체 어휘)을 사용한 감정분류 시스템을 구현하였고 성능 비교 실험을 소개하였다. 이모티콘 극성 자질은 가장 큰 성능 향상을 제공한다. 그러나 리트윗 극성 자질은 내용이 짧다보니 추출 가능한 자질이 적어 리트윗된 트윗과 다른 감정을 표현하는 경우가 의외로 많아 가장 낮은 성능 향상을 보였다. 대체 어휘 자질은 문자 사이의 유사도만을 사용하여 추출하였기 때문에 성능 향상에 큰 영향을 미치지는 못했다. 그러나 사용자 사이의 관계 정보를 더 정확하게 추출하고, 의미적으로 동일한 대체 어휘를 효율적으로 추출한다면 감정분류의 정확도를 높일 수 있을 것이다.

문서에 포함된 단어의, 인접 단어들이 갖는 각 감정의 연관성 정보를

활용하기 위해 선별된 단어를 특성으로 하는 지지기반벡터모델을 학습시켜 얻은 특성 계수를 은닉조건랜덤필드(hidden conditional random field)의 입력 값으로 사용하여, 감정을 분류하였다.

2.2 자율학습

자율학습은 정답 집합이 주어지지 않고, 입력 값에 근거하여 학습을 진행하는 기계학습의 한 방법이다. 지도학습의 실용성에도 불구하고, 현실적으로 정답 집합에 대한 학습 말뭉치를 구하기 힘들다는 단점 때문에, 자율학습 방법이 제안되었다. 텍스트 데이터 내에서 동시 출현하는 단어의 공기정보(co-occurrence)와 같은 확률적 속성을 기반으로 한다. 자율학습에서는 입력값의 통계적 특징을 파악하고, 비슷한 입력값이 비슷한 출력값을 생성할 수 있도록 연결 가중치를 조정한다. 즉, 자료나 샘플 없이 모든 데이터를 계속 확인하면서 비슷한 것을 연결하는 방식으로 출력값을 얻게 된다. 자율학습은 통계의 밀도 추정(density estimation)과 깊은 연관이 있다. 적용되는 확률 밀도는 그 데이터 집합에 의해 만들어진다. 이러한 자율학습은 데이터의 주요 특징을 요약하고 설명할 수 있다. 자율학습의 예로는 클러스터링, 독립 성분 분석, 연관성 규칙, 데이터 축소 등이 있다.

어휘 기반 감성분류 연구[40]는 자율 학습 기술에 의한 어휘 사전을 자동으로 습득한다. A. Go, R. Bhayani, & L. Huang [41] 연구는 학습 데이터를 자동으로 생성하는 거리 감독 접근법을 사용하였다. 그러나 이 방법에서 초기 학습 데이터의 잘못된 라벨링은 잘못된 학습 결과를 초래하기가 때문에 초기 학습 데이터에 대란 신중한 라벨링 작업이 필요하다.

Shelke, N., Deshpande, S., & Thakare, V.[22] 연구는 아마존 리뷰데이터를 활용하여 상품에 대한 극성을 분류하였다. 한 개의 문장을 the product features(atomosphere, location, drinks, food, jajita. connon powershot g3 등)와 의견을 의미하는 semantic features(uncomfortable, substandard, very good, too

sweeet 등)으로 구분하였다. semantic features에 해당하는 단어를 SentiWordNet에서 매치시켜 극성에 대한 점수를 계산하였다.

Cheng, K., Li, J., Tang, J., & Liu, H.[42] 연구에서는 signed social networks 개념을 사용하여 자율학습 기법으로 감성을 분석하였다. 아이템 r에 대하여 사용자 A에 의해 작성된 데이터가 있을 경우, positive linked set $P(t_i)$는 사용자 B에 의해 게시된 동일 아이템에 대해 게시된 t_j에 대한 전체 데이터 집합으로 정의하였으며, 사용자 B는 사용자 A로부터 연결되어 있다. 이러한 연결정보를 통해 아이템에 대한 긍정, 부정의 방향성을 계산하고 비교하고 업데이트하여 감성을 분류하였다.

홍태호, 김은미, & 차은정 연구[43]는 뉴스정보를 활용한 감성분석을 이용하여 주식시장을 예측하였다. 주식시장 예측은 시장 내부의 변화와 외부의 사건들을 모두 반영할 수 있어야 하지만 기존 연구들은 기술적 분석 방법만을 고려하거나 뉴스 혹은 소셜 네트워크 서비스를 통한 감성분석만을 고려하여 진행되었다. 이에 본 연 구에서는 주식사장의 추세를 확인하기 위해 기술적 분석기법을 사용하고, 시장의 외부요소를 반영하기 위해 뉴스의 감성분석기법을 활용하여 주식시작을 예측하였다. 감성사전 SentiWordNet을 사용하여 각각의 단어에 대해 긍정인지 부정인지를 구분하였다.

Schouten, K., van der Weijde, O., Frasincar, F., & Dekker, R.[44] 연구는 notional words간의 co-occurrence association rule을 제안하였다. 텍스트 데이터 집합으로부터 seed word를 찾아내고, co-occurrence digraph를 그린 후, spreading activation을 적용하여 rule을 생성한다. 생성된 rule을 데이터 집합에 적용하여 [notional word → category] 형태로 분류하는 기법을 제안하였다.

Hu, X., Tang, J., Gao, H., & Liu, H.[45] 연구에서는 이모티콘과 상품 rating과 같은 emotional signals은 포스트나 단어에서 감성을 표현하는 수단으로 보았으며, 이러한 emotional signals은 emotion indication과 emotion correlation으로 구분하였다. Emotion indication에서 post-level은 이모티콘과

상품 rating 정보를 사용하였으며, word-level에서는 sentiment lexicon을 활용하였다. Emotion correlation은 동시에 출현하는 단어의 극성이 유사한지 계산하였다. Stanford Twitter Sentiment, Obama-McCain Debate 데이터를 사용하여 감성을 분석하였다.

자율학습 기반의 국내·외 연구는 대부분 긍정, 부정과 같은 극성분류에 대한 감성분석 연구가 대부분이다. 이는 다양한 감정 분류를 위한 개개의 감정범주 특징을 반영한 규칙 생성이 어렵고, 세분화된 각 감정범주를 대표하는 seed word에 대한 연구가 미비하기 때문이다.

제 3 절 분산표현 학습

Hinton은 1986년 뉴론이 어떻게 개념을 나타내는가를 설명하기 위해 분산표현을 제안하였다. 한 개 뉴론이 한 개의 개념을 나타내면 이를 국소표현이라 하고, 벡터형태로 나타내서 이를 one-hot vector이라고도 부른다. 분산표현은 여러 뉴론을 작동시켜 한 개의 개념을 표현한다. 즉 개념을 표현하기 위해 여러 특징(자질)의 조합으로 나타내는 것을 의미한다. 단어의 분산 표현은 단어 자체가 가지는 의미를 다차원 공간에서 벡터로 표현하는 것을 의미하며, 워드 임베딩이라고도 한다.

3.1. 워드 임베딩

문서를 벡터 공간으로 임베딩하는 연구 중 가장 최근에 관심을 받고 있는 것은 Paragraph Embedding(Doc2Vec)이다[27]. Mikolov는 문맥상 비슷한 의미를 지닌 단어가 가까운 벡터공간으로 임베딩되는 신경망 기반 언어모델인 Word2Vec[46]을 문서, 문단, 문장 단위로 확장한 신경망 및 학습알고리즘을 제안하였다. 이러한 연구는 기존 토픽모델인 Latent Dirichlet Allocation[47], probabilistic latent semantic analysis(PLSA)[48] 등에 비해 우

수한 성능을 보여주었다. Word2Vec의 주요 목적은 뉴럴 네트워크를 기반으로 대량의 문서 데이터 집합을 벡터 공간에 고차원의 의미 벡터를 가지도록 효율적으로 단어의 벡터 값을 예측하기 위함이다. Word2Vec은 10억 개의 단어 시퀀스와 같은 대량의 학습 데이터일지라도 이를 학습하는데 하루가 걸리지 않을 정도로 아주 빠른 학습 속도를 유지한다. Word2Vec의 학습 결과는 단어들의 의미가 비슷할 경우 같은 벡터 공간에 표현된다. 예를 들면 '4차산업'과 '인공지능' 이란 단어는 실제로 벡터 공간에서 비슷한 공간에 위치하게 되며, 이는 단어의 의미에 따른 벡터 공간에서 군집화를 가능하게 하며, 벡터 연산을 통해 데이터 추론을 가능하게 한다.

자연어 처리에서 딥러닝을 적용하는 방법 중 분산 단어 표현(distributed word representation)이 있다. 분산 단어 표현은 워드 임베딩 벡터와 같은 의미이다. 워드 임베딩은 대용량의 말뭉치를 자율학습 방식으로 학습하여 차원 축소 및 추상화를 통해 문서에 등장하는 단어를 수십에서 수백 차원의 자질 벡터(feature vector)로 표현하는 것이다[49]. 워드 임베딩 학습을 끝내면 하나의 단어를 벡터 공간 상 하나의 벡터로 표현 할 수 있다. 워드 임베딩 학습 방법으로는 대표적으로 Mikolov가 은닉층(hidden layer)을 제거하고 신경망 모델을 단순화하는 방법인 Word2Vec 모델을 제안하여 단어 자질의 학습 시간을 비약적으로 단축시켰다. Word2Vec은 주변 단어로부터 현재 단어를 예측하는 continuous bag-of-words(CBOW) 방식과 현재 단어로부터 주변 단어를 예측하는 skip-gram 방식이 있다[46].

Continuous bag-of-words(CBOW)와 skip-gram 방식은 언어모델을 뉴럴 네트워크를 이용하여 구현한 것으로, CBOW 방식은 예측하고자 하는 특정 단어의 이전과 이후 단어들을 신경망의 입력으로 받아 사상층(projection layer)을 거쳐 대상 단어의 모든 후보 단어들의 확률을 구한다. 반면, skip-gram 방식은 현재 단어를 신경망의 입력으로 받아 사상층을 거쳐 이전 단어들과 이후에 나타나는 단어들의 모든 후보 단어들의 확률을 구하는 방식이다.

워드 임베딩 방법이 제안된 후, 단어 레벨에서 확장된 문장, 문단 그리

고 문서에 대한 표현 연구가 진행되었다. 문단(paragraph) 단위로 확장된 대표적인 임베딩 학습 방법은 Mikolov가 제안한 Doc2Vec 모델이다[27].

Doc2Vec 방식은 distributed memory(DM) 방식과 DBOW(Document Embedding with Distributed Bag of Words) 방식 두 가지가 있다. Word2Vec 알고리즘과 같은 원리로 학습 하지만, 문단 정보를 기억하기 위해 문단 벡터(paragraph vector)가 추가 되었다.

하나의 문단에 해당되는 모든 단어들을 순서대로 5개씩 읽으면서 다음 단어를 예측하는 과정을 통해 문단 벡터가 학습된다. 예를 들면, *"비가 너무 많이 와서 습하고 꿉꿉하다....."* 에서 (비가 너무 많이 와서 습하고), (너무 많이 와서 습하고 꿉꿉하다), (많이 와서 습하고 꿉꿉하다 ...) 등을 통해 다음 단어를 예측한다. 결국 문단 안에 있는 모든 단어들을 거치면서 문단 벡터가 학습되고 벡터 공간에 위치되면서 문단의 의미가 결정된다. 즉, 단어들이 학습될 때, 각각의 학습 단계를 벡터에 기억시키고 학습된 최종 벡터를 문단 벡터로 정의한다. 이 문단과 유사한 단어 열을 가지고 있는 문단도 이와 같은 학습과정을 통해 문단 벡터의 위치가 앞에서 제시된 문단 벡터의 위치에 근접하게 나타난다. 이는 입력되는 단어 벡터들이 유사하기 때문이다. 문단 벡터는 하나의 문단에 있는 모든 단어 열을 사용하여 문단의 의미를 얻는다. 문장의 의미를 얻기 위한 방법으로 사용되는 단어 벡터들의 범위를 하나의 문장으로 제한하여 학습시킨다면 벡터는 문장의 의미를 갖는 문장 벡터가 될 것이며 문서로 확장하여 학습시킨다면 문서 벡터가 될 것이다.

제4절 감정분류 연구의 문제점

영어에 비해 한국어를 대상으로 하는 감정분류 및 접근 방식에 대한 연구는 아직 미비한 상태이다. 이는 한국어가 갖고 있는 교착어적 특징으로 인해 한국어에 대한 연구 자체가 영어권에 비해 복잡하고 어렵기 때문이

다. 교착어는 어근에 조사나 어미와 같은 문법 형태소들이 결합되어 문법 관계를 표시하거나 단어를 형성하는 언어를 말한다. 예를 들어, *"할머니-께서-는 이야기-를 즐겁-게 하-시-더-구나"* 문장에서, 어근은 *'할머니', '이야기', '즐겁-', '하-',* 조사는 *'-께서', '-는', '-를',* 어미는 *'-게', '-시-', '-더-', '-구나'* 가 덧붙어 문장이 형성된다. 이처럼 어근에 붙는 문법 형태소의 수와 종류가 매우 풍부하여 이를 해석하고 분석하는데 어려움이 존재한다.

기존 한국어 감정분류 연구는 긍정, 부정에 대한 극성 분류가 대부분이었다. 이는 감정분류를 위해 사용되었던 학습 데이터 자체가 긍정, 부정으로 태깅된 영화 리뷰나 상품 리뷰 데이터가 전부였기 때문이다. 오픈되어 있는 한국어 감성어 리소스나 기계학습에 사용되는 감성어 학습 말뭉치의 부재는 세분화된 감성분류에 대한 연구를 어렵게 하고 있다. 최근 집단지성을 활용한 '한글 감성어 사전[29]'이 구축되어 공개되었으나 긍정, 부정, 중립에 대한 극성 정보만을 담고 있어 세분화된 감정분류를 위한 데이터로 사용하기에는 한계가 있다. 세분화된 감정분류 기법이 여러 형태의 기술 분야에 적용됨에 따라 극성 분류만을 다루는 시스템은 큰 제약으로 작용할 수 있다. 예를 들어, *'여성스러운-남성스러운'* 등과 같은 감정은 특정 상품이나 인물에 대한 이미지 파악에 있어 중요한 감정이지만, 단순히 '중립'으로 분류될 경우, 세부 감정 파악이 불가능하기 때문이다. 감정범주에 대한 정의도 응용프로그램에 따라 영어권에서 사용되는 감정범주를 재정의하여 사용하고 있는 실정이다. 또한 SentiWordNet, Stanford Sentiment Treebank 등과 같이 공개된 리소스는 영어권이 아닌 다른 언어를 사용할 경우 한국어의 감정을 그대로 전달하기에는 적합하지 않다. 이처럼 한글 감정어 사전이나 관련 학습 자료가 충분하지 않기 때문에 감정분류의 성능뿐만 아니라 감정분류를 위한 연구의 어려움이 존재한다.

감정을 정확하게 분류하기 위해서는 문장의 표층적 분석뿐만 아니라 문장이 표현하고자 하는 의미에 대한 적절한 평가를 내릴 수 있어야 한다. 예를 들면, *'크다'* 라는 서술어가 어떤 감정을 의미하는지 분야별로 정확

히 분석하기 위해서는 각 분야에 대한 기본 어휘 등의 언어 자원이 미리 구축되어야 하며, 학습에 사용되는 자질을 효율적으로 설계할 필요가 있다. 표층 자질 벡터는 한국어 특성을 고려하여 설계되어야 한다. 한국어 문장은 어절로 구성되어 있다. 영어에서는 각 단어 뒤에 띄어쓰기를 해서 단어를 구분하기 때문에, 한국어에서와 같은 문제가 발생하지 않는다. 그러나 한국어에서는 어절 별로 띄어쓰기를 해야 한다. 기계가 어절 안의 단어를 식별하지 않으면 안 되기 때문이다. 어절 안의 단어를 형태소라고 부르며 구문분석 전에 이러한 형태소 분석이 먼저 수행된다. 예를 들면, *"오늘 비가 왔다"* 와 같은 문장에 대한 형태소 분석은 다음과 같다.

오늘 비 가 왔다
[부사] [명사] [조사] [동사(완료형)]

한국어 문장의 구문분석은 영어의 구문분석과 다르다. 영어는 구(phrase)의 집합이기 때문에 영어의 해석은 문장을 구성하는 구를 분석하는데 집중되었다. 그러나 한국어 문장의 경우는 영어와 다르게 낱말 사이의 수식관계가 문장구조의 기본을 이룬다. 위 예제에서 *'오늘', '비가'*는 모두 *'왔다'* 와 수식관계를 이룬다. 이러한 형태소 분석과 구문분석을 반영한 표층적 자질을 설계해야 한다. 자질의 종류가 다양하고 많을수록 그렇지 않은 경우보다 학습이 잘 이루어진다. 그러나 학습하는데 걸리는 시간이 많이 소요될 뿐만 아니라 과잉적합이 된 나머지 새로운 데이터가 입력으로 들어올 때 이를 올바르게 분류하지 못하는 문제가 발생할 수 있다. 따라서 감정분류를 위한 효율적인 자질 설계가 필요하며 과적합 문제를 효율적으로 해결할 수 있어야 한다.

소셜 미디어와 같이 서로 상호작용하면서 의견을 주고받는 데이터일 경우, 이전 문장의 감정이 현재 문장에 반영될 수 있다. 다음 이어지는 대화 예제에서 화자 B의 *"응"* 에서 나타난 감정은 *"감정없음"* 이 아니라 이전 발화에서 이어지는 *"슬픔"*의 감정이 더 적절하다고 볼 수 있다.

A: 어제 혼났다며
B: 담임 선생님한테 지각했다고 혼났지.
A: 많이 혼났니?
B: 응

본 논문에서는 한국인 실정에 맞는 세분화된 감정범주를 결정하고 기계학습 방법을 사용하여 자동으로 범주를 할당하는 방법을 제안한다. 감정범주를 결정하기 위해서는 Park, I. J., & Min, J. K.[33] 연구에서 제시한 한국인 감정어휘 목록을 바탕으로 25개의 감정범주를 추출하고 기계학습을 위해 감정분류에 필요한 자질 집합을 표면적 자질, 의미적 자질로 구분하여 최적의 자질 집합을 제안한다. 그리고 제안한 자질을 사용하여 감정을 자동으로 분류하기 위해 대표적인 기계학습 방법 중 하나인 서포트 벡터 머신과 선형 회귀 분석을 사용한다.

제 3 장 감정분류 기법

본 연구는 텍스트 고유의 형태학적·어휘적 특징과 문장에서 단어가 담고 있는 의미를 특징으로 사용하는 감정분류 기법을 제안한다. 형태학적/어휘적 특징은 텍스트가 갖고 있는 표층적 자질(surface feature)로 정의하며, 문장 내 단어가 담고 있는 의미적 관계 정보를 의미적 자질(semantic feature)로 정의하였다. 이러한 자질들은 데이터로부터 감정분류의 단서를 제공하기 때문에 새로운 데이터를 예측하거나 분류할 때 매우 중요하다[49]. 표층 자질과 의미적 자질은 워드 임베딩 조합기법을 통해 결합되며, 학습 단계에서 감정분류기를 학습할 때 사용된다. [그림 1]은 본 연구에서 제안한 감정분류 프레임워크이다.

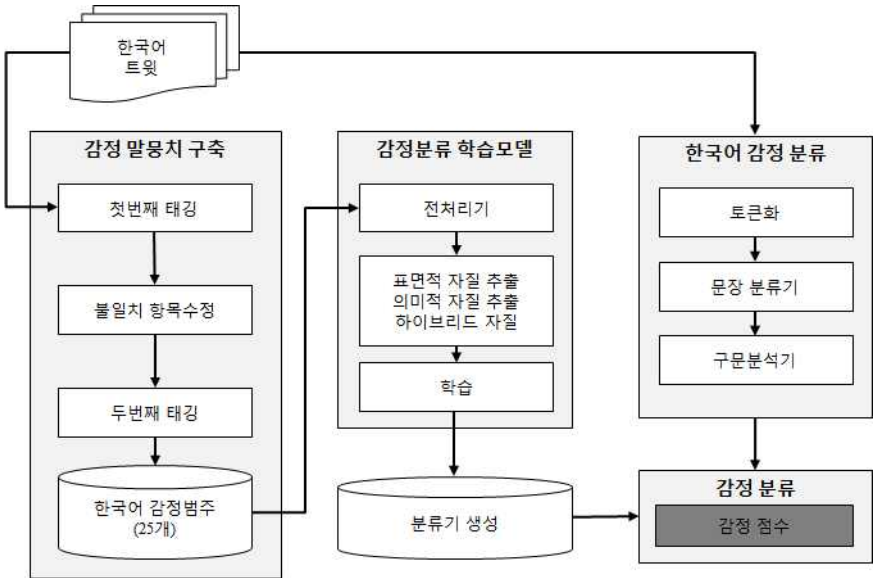


그림 1 제안 시스템 프레임워크

■ 감정분류 학습모델 : 감정 말뭉치로부터 문장 표상(sentence representation) 단계를 통해 추출된 자질을 사용하여 감정을 탐지할 수

있는 모델을 학습한다.

■ 기계학습기반 한국어 감정분류: 학습을 통해 만들어진 감정 모델을 사용하여 감정을 분류한다.

본 연구는 워드 임베딩 조합기법을 사용하여 표층 자질과 의미적 자질을 결합하였고, 감정 모델 학습과 감정분류 예측을 위해 모두 사용된다. 3장에서는 이러한 명시적 자질과 의미적 자질을 추출하는 방법과 이 두 자질을 결합하는 방법에 대해 설명하며, 학습 말뭉치의 희소성 문제를 해결하기 위한 오버샘플링 기법에 대해 설명한다.

제 1 절 전처리 작업

전처리 작업은 수집된 트위터 데이터를 대상으로 정제되지 않은 텍스트 데이터로부터 필요한 정보들을 추출하는데 필수적인 작업이다. 전처리 작업에서는 띄어쓰기 교정, 문장 분할, 토큰화 과정, 구문 분석을 거친다.

트위터와 같은 소셜 미디어는 띄어쓰기가 잘 지켜지지 않는다. 띄어쓰기 오류는 형태소분석기에서 형태소 별로 분할 처리되기도 하지만 문장 전체의 띄어쓰기가 무시된 경우, 처리가 어렵기 때문에 가정 먼저 띄어쓰기 교정을 통해 데이터를 처리한다.

문장이란 기본적으로 생각이나 감정을 말로 표현할 때 완결된 내용을 나타내는 최소의 단위이다. 문장의 길이가 길어질수록 구조가 복잡하여 문장을 분석하는데 시간이 오래 걸리며 낮은 정확성을 보인다. 따라서 정교한 문장 분리는 형태소 및 구문 분석 등 언어처리 작업에 활용될 때, 그 성능에 큰 영향을 미치기 때문에 중요한 전처리 작업 중 하나이다. 한국어는 영어와 마찬가지로 '.','?','!' 등을 구두점으로 사용한다. 일반적으로 구두점을 기준으로 문장을 분리하며, 복잡하고 긴 문장은 몇 개의 짧은 문장으로 분할하는 방법이 연구되고 있다. 그러나 트위터는 짧은 단문이 대부분이며, 이모티콘이 문장 내 중간 중간에 자주 사용되기 때문에, 기존 문장 분할 방법으로는 문장을 분할하는데 어려움이 있다.

특히, 소셜 미디어에서 이모티콘은 감정이나 느낌을 말로 표현하거나 전달하기 힘들 때 자주 사용되며, 문장의 일부 또는 문장 전체에 이모티콘이 사용되기도 한다. 한국어에서는 한글 자모를 이용한 이모티콘이 자주 사용되고 있다[50]. 이모티콘은 그 활용 사례가 아주 다양하기 때문에 이를 정확하게 찾아내어 문장으로 분리하기는 쉽지 않다. 따라서 이모티콘 기준이 아닌, 일반 문장을 식별하는 정규 표현식을 사용하여 일반 문장과 이모티콘 분할하였다. (식 1)은 문장 분할을 위한 정규 표현식이며, 정규 표현식으로 일반 문장을 찾아내었고, 포현이 불가능한 문장은 이모티콘으로 식별하여 사용하였다.

[^가-힣|a-z|A-Z|0-9|\s\(\)\:\-#@_,]{1,}　　(식 1)

분할된 문장은 토큰(어절) 단위로 구분하는 토큰화 과정을 거친다. 토큰화 과정에서는 트위터에 사용되는 예약어를 그 역할에 맞게 변환하거나 제거하는 과정을 포함한다. 파싱 단계에서 사용될 파서(parser)는 일반적인 구조의 문장에 최적화 되어 있기 때문에 일반적인 문장에 나타나지 않는 구성요소는 최대한 제거하거나 단순화하였다. <표 3>는 트위터에서 사용되는 예약어와 처리방법에 대한 설명을 보여준다.

표 3 트위터에서 사용되는 예약어 예제

예약어	설명
“RT”	바로 뒤에 나오는 문자열은 트윗 작성자가 직접 만든 문장이 아니므로 제거한다.
“@”	바로 뒤에 나오는 문자열은 특정 사용자를 지칭 하므로 제거한다.
“httpL//”	바로 뒤에 나오는 문자열은 LINK를 나타내므로 제거한다.

구문 분석은 문장을 이루고 있는 구성 성분으로 분해하고, 그들 사이의 관계(위계, 의존)를 분석하여 문장의 구조를 결정하는 작업이다. 구문 분석을 통해 입력된 문장에 대한 구조를 해석하고 그 구조를 명백히 하면 문장에 대한 해석이 용이하기 때문이다. 트윗 작성자의 감정을 찾아낸다는 것은 문장의 의미를 해석해야 하는 고도의 복잡한 작업이다. 따라서 뜻을 가진 가장 작은 말의 단위인 형태소를 기반으로 주어, 동사, 목적어 등 기본적인 역할을 분석하고 수식어와 수식 대상을 파악하여 문장의 전체적인 구조 파악이 수반되어야 한다.

한국어는 다른 언어에 비해 비교적 어순이 자유롭기 때문에 그 동안 의존문법에 기반한 구문 분석연구가 주를 이루어왔다[51]. 그러나 세종 트리뱅크[52]를 통해서 양질의 구문구조를 제공하고 있어 이를 활용한 구문분

석기의 연구가 최근 이루어지고 있다[52, 53]. 그 중에서도 [52]에서는 다양한 언어에서 state-of-art의 성능을 보이고 있는 berkeley parser[53]를 세종 트리뱅크에 활용하여 한국어 구문분석기를 제공하고 있어 본 연구에서는 이를 활용하였다.

제 2 절 감정 말뭉치 구축

지도학습을 통한 기계학습은 감정 범주가 부착된 감정 말뭉치를 필요로 한다. 감정 말뭉치는 언어 현상의 통계 정보나 규칙을 학습하는데 사용되며, 본 연구에서는 10명의 대원학생의 교차검증을 통해 감정 범주를 부착하였다.

2.1 감정 어휘 수집

인간의 감정에 대한 견해는 학자들에 따라 차이는 있으나 학자들이 정의한 인간의 기본 감정은 동서양이 거의 유사하다. 초기 감정 분석은 긍정/부정 등의 극성을 사용하였고, 단어의 극성 정보를 제공하는 한국어 시소러스가 없었기 때문에 SentiWordNet과 같은 극성 정보를 제공하는 영어 시소러스를 사용하여 극성 정보를 얻은 후, 영한사전으로 번역하여 사용해 왔다. 그러나 영어 단어의 극성 정보를 한국어로 번역하는 과정에서 영어와 한국어의 의미 차이가 발생할 수 있어서 극성 정보를 제대로 반영할 수 없는 단점이 있다. 또한 단어의 극성 정보만을 사용하면 다양한 감정을 단순화하게 된다. 따라서 극성뿐만 아니라 다양한 감정 상태를 반영할 수 있는 감정 어휘의 목록이 필요하다. 심리학에서는 인간의 언어가 갖는 포괄성과 다양성 때문에 어휘 분석을 통해 성격이나 정서의 구조를 연구해왔다. 성격의 5요인 연구를 제시한 GoldBerg[54]는 기본 어휘 가설(fundamental lexical hypothesis)을 제안하였다. 이 가설은 인간의 관계적 삶

에서의 중요한 개인차들이 세계의 대부분의 언어들 속에 어휘들로 자리 잡고 있기 때문에 감정 단어들을 추출하여 정리하고, 이들을 분석하여 감정 구조를 탐색하는 작업은 이와 같은 감정의 기본 어휘 가설에 근거하여 수행된 것이다. 따라서 본 연구에서도 텍스트 데이터로부터 감정 단어를 추출하고 분석하여 감정을 분류하고자 하였다.

본 연구에서는 최종 목표인 세분화된 다양한 감정분류를 위해 Park, I. J., & Min, J. K.[33] 연구에서 제안한 한국어 감정 어휘 목록을 사용하였다. 이 연구에서는 감정 어휘 목록에 수록될 단어들을 선별하고, 감정 현상을 연구하는 연구원 10명을 통해 선별된 감정 단어들이 감정 단어로서 얼마나 적절한지(원형성)와 친숙한지(친숙성)를 파악하도록 하고, '쾌-불쾌'와 '활성화 수준'을 제공한다.

이러한 감정 어휘 목록 434개 중 친숙성을 기준으로 상위 25개의 감정 단어를 선정하였다. 이는 소셜 미디어에 언급된 감정 단어들이 사용자들이 작성하기 쉽고 친숙한 어휘로 표현되기 때문이다.

선정된 한국어 감정 단어는 문장의 서술어 기능을 담당하는 용언으로 구성되어 있다. 한국어의 용언은 단어의 개념적 의미를 갖는 어간과 문법적 기능을 표시하는 어미로 구성된다. 어간은 한 단어의 개념적 의미를 나타내기 때문에 변화하지 않고 고정된 요소로 나타난다. 감정 단어에서 두 글자로 된 어간을 추출하여 본 연구의 감정 범주로 사용하였다.

또한 문서 수집을 위한 키워드로써 많은 데이터를 수집하기 위해 어간을 확장하여 검색 키워드로 사용하였다. 확장된 어간은 꼬꼬마 형태소 분석기[55]의 말뭉치 검색 기능을 활용하였으며 <표 4>는 본 연구에서 사용하는 감정 단어와 어간, 확장된 어간을 나타낸다.

표 4 감정단어와 어간, 확장된 어간

감정단어	원형성	친숙성	쾌-불쾌	활성화	어간	확장된 어간
기쁘다	5.98	**6.26**	5.94	5.56	기쁘	기쁘, 기뻐, 기뻤, 기쁜
재미있다	5.70	**6.19**	5.72	4.79	재미	재미, 재밌
반갑다	5.86	**6.14**	5.91	5.29	반갑	반갑, 반가
고맙다	5.39	**6.13**	5.50	3.77	고맙	고맙, 고마
사랑스럽다	5.77	**6.12**	6.09	4.52	사랑	사랑
좋다	5.90	**6.12**	5.54	5.13	좋다	좋
즐겁다	5.87	**6.09**	5.89	5.54	즐겁	즐겁, 즐거
행복하다	5.88	**6.08**	6.16	4.70	행복	행복
미안하다	5.02	**5.97**	2.91	4.01	미안	미안
편안하다	5.47	**5.95**	5.40	2.33	편안	편안
만족하다	5.34	**5.92**	5.64	3.98	만족	만족
슬프다	5.08	**5.85**	2.69	3.44	슬프	슬프, 슬퍼, 슬펐, 슬픈
아쉽다	5.20	**5.84**	3.07	3.56	아쉽	아쉽, 아쉬
그립다	5.68	**5.82**	4.37	3.18	그립	그립, 그리
무섭다	5.06	**5.82**	2.58	5.39	무섭	무섭, 무서
후회하다	4.74	**5.82**	2.45	4.06	후회	후회
우울하다	4.94	**5.81**	2.49	2.80	우울	우울
부럽다	5.23	**5.77**	4.00	4.37	부럽	부럽, 부러
불안하다	4.69	**5.75**	2.51	4.88	불안	불안
우습다	4.46	**5.71**	3.72	4.44	우습	우습, 우스
속상하다	4.67	**5.70**	2.32	4.48	속상	속상
창피하다	4.76	**5.69**	2.54	4.20	창피	창피
싫다	5.03	**5.68**	2.24	4.76	싫다	싫
실망하다	4.64	**5.66**	2.19	3.30	실망	실망
외롭다	5.02	**5.66**	2.28	2.76	외롭	외롭, 외로

2.2 감정 범주 부착

한국어 감정 분석을 위해 감정 모델을 학습시키고 분류하기 위한 감정 말뭉치를 직접 구축하였다. 감정 말뭉치는 그 문서에 해당하는 감정 범주가 부착된 문서 집합을 의미하며, 감정 분석을 위한 학습 단계와 평가 단

계에 사용된다. 한국어의 경우, 긍정/부정 이외의 다양한 감정에 대한 학습 말뭉치 부재로 감정분류기 성능의 비교 판단이 어려울 뿐만 아니라 기존 연구의 문제점을 보완한 후속 연구의 진행과 그에 따른 성능 향상 결과를 비교하기도 어려웠다.

한국어 감정 코퍼스 구축을 위해 트위터에서 제공하는 REST APIs를 사용하여 각 감정 단어의 '확장된 어간'을 부분 문자열로 갖는 트윗을 수집하였다. 각각의 트윗에는 작성자마다 '기쁨'이라는 동일한 감정이라도 자신의 감정을 표현하는 방법이 다르게 나타난다. 따라서 다양한 표현 방법으로 기술된 데이터를 수집하기 위해 다음과 같은 규칙을 사용하여 중복된 표현 방법을 제거하였다.

첫째, 리트윗을 제거한다. 이는 동일한 방식으로 기술된 데이터 표현 방법을 제거하기 위함이다.

둘째, 사용자별로 최대 3개의 트윗만 허용한다. 사용자마다 감정을 표현하는 방식이 다르기 때문에 많은 사용자로부터 기술된 다양한 표현 방식을 수집하기 위함이다.

수집된 트위터는 K대학교 소속 10명의 대학원생들에 의해 3단계에 걸쳐 25개의 감정 범주로 태깅되었다. 각 감정 범주에 대한 오리엔테이션을 충분히 하고, 사전에 제작한 온라인 주석 도구를 활용하여 감정 단어로 검색된 트윗을 대상으로 해당 감정이 들어 있는지 태깅하였다. 각 감정 범주는 420개의 트윗으로 구성되며, 신뢰성을 위해 한 개의 트윗은 3명의 대학원생이 태깅하도록 설계하였다. 한 개의 트윗에 대한 감정 범주가 만장일치로 나오지 않을 경우, 토론을 통해 다시 태깅하도록 하였으며, 오판의 위험성을 줄이기 위해 여러 단계에 걸쳐 신중하게 감정 범주에 대한 학습 말뭉치를 구축하였다.

학습 말뭉치는 트윗 내용의 표면적인 해석보다는 그 의미를 이해하고 판단하여 해당 감정 범주에 속하면 'acceptable', 그렇지 않을 경우 'unacceptable'로 태깅한다. 만약 해당 감정이 분명하게 드러나지 않을 경

우, 논의를 통해 ‘acceptable’과 ‘unacceptable’로 판단한 수를 비교하여 다수결로 결정하였다. 뉴스나 소설의 일부 내용이 담긴 트윗은 작성자가 직접 기술한 감정이 아니므로 ‘unacceptable’로 판단하였고, 광고와 인사말도 ‘unacceptable’로 판단하였다.

<표 5> 는 감정 범주가 트윗 내용에 의미적으로 포함되어 있는지 여부를 판단하는 예를 보여준다. 감정 단어가 포함되었더라도 뉴스 기사, 다른 사람의 감정, 조건문, 감정 대상이 불명확한 경우 ‘unacceptable’ 판단하였다.

표 5 감정 판단(acceptable, unacceptable) 예시

감정 범주	트윗 내용	판 단	설 명
기쁘	오모오모ㅠㅠㅠㅠㅠㅠㅠ 성스러운 짤에 겁나기쁘지만 저장이 안되능 못난 현실ㅠㅠㅠㅠ	✓ acceptable	기쁘다는 감정을 느낌
기쁘	#' 체전 도마 4연패' 양학선 " 자존심 지켜 기쁘다" - 중앙일보 LINK	unacceptable	뉴스 기사
기쁘	나는 왠지 기뻐져 기쁘기 때문에, 작업이 진행되지 않아.」 - 켄토	unacceptable	다른 사람의 감정
재미	아..... 진짜 페달 엔솔들이 보배롭다 ㅠㅠㅠㅠㅠ 너무 재미있어 ㅠㅠㅠㅠ	✓ acceptable	재미있다는 감정을 느낌
재미	REPLY 아 그렇군요 가보고 싶네요 코스 조금 까다롭지만 오늘 재밌게 잘 쳤어요.^^	✓ acceptable	재미있다는 감정을 느낌
재미	REPLY 봇에게 재미있는 말을 알려주면 재미있는 말을 합니다	unacceptable	조건문
좋다	기↗ 분↓ 이↓ 좋↓ 다↗↗↗ 기분이 ↗↗↗↑↗↑↑ 다↓	✓ acceptable	좋다는 감정을 느낌
좋다	REPLY 그래! 좋-은 생각이야..!	unacceptable	생각이 좋은 것임

2.3 오버샘플링 기반 학습 말뭉치 확장

학습 말뭉치는 일정 규모 이상의 크기를 갖추고 내용적으로 다양성과 균형성이 확보될 때 학습 말뭉치를 사용하는 지도학습의 성능을 높일 수 있다. 그러나 분류 대상의 범주 종류가 많으면 많을수록 데이터 수집이 어려워지며 이로 인해 데이터 불균형 문제가 발생한다. 데이터 불균형은 긍정 샘플과 부정 샘플 수가 비슷하지 않거나 한쪽의 수가 다른 쪽의 수보다 극히 작거나 많을 때 발생한다. 본 연구에서는 감정 범주의 수가 25개로 많고, 학습 데이터가 균형적으로 존재하지 않기 때문에 샘플 데이터의 균형을 맞추기 위해 샘플 수가 부족한 범주에 SMOTE(synthetic minority over-sampling technique)을 적용하였다[56]. SMOTE은 불균형 데이터의 균등화 방법 중 하나로 적용 분야의 전문지식을 필요로 하지 않고 분류기의 제한이 없다는 장점을 갖고 있다.

SMOTE 알고리즘은 기준이 되는 샘플을 한 개 선택하고, 해당 샘플과 가장 가까운 k개의 샘플을 추려낸다. 이때 k개의 샘플은 모두 소수 집단에서 추출한 샘플이다. 일반적으로 k는 5정도로 설정한다. 그리고 기준 샘플과 이들 k개 이웃 간의 차이(difference)를 구하고, 이 차이에 0~1 사이의 임의의 값을 곱하여 원래 샘플에 더한다. 이렇게 만든 새로운 샘플을 훈련 데이터에 추가한다. SMOTE는 오버샘플링처럼 단순히 복제하는 것이 아니고, 소수 집단과 비슷한 새로운 케이스를 생성하여 오버피팅 문제를 보완하였다. SMOTE 알고리즘은 <표 6>과 같다.

SMOTE 방법은 생성된 데이터와 기존의 소수 범주 데이터를 합쳐서 새로운 소수 범주로 설정하고, k는 일반적으로 $k = M/N$ 으로 구한다. 이때 M은 전체 학습 데이터에서 가장 큰 범주의 샘플 수, N은 현재 소수 집단의 샘플 수이다.

표 6 SMOTE 알고리즘

1. 소수 집단의 샘플 중 무작위로 하나의 샘플(V_s)을 선택한다.
2. 선택된 샘플을 기준으로 소수 집단의 샘플들과의 거리를 계산한다.
3. 기준이 된 샘플과 가장 가까운 k개의 소수 집단 샘플을 설정한다.
4. k개의 샘플 중 임의의 샘플(V_{s_n})을 선택한다.
5. 기준 샘플(V_s)과 임의의 샘플(V_{s_n})을 사용하여 새로운 샘플(V_{nw})을 계산한다.

   $V_{nw} = V_s + R_{0\sim1} \times V_{S_n}$

   $R_{0\sim1}$: 0~1 사이의 임의의 값,

   V_s: 문장 S의 벡터,

   V_{s_n}: V_s의 k개의 이웃 중 한 문장 벡터
6. step1에서 선택한 소수 집단을 제외한 나머지 중에서 다시 무작위로 하나를 선택하여 추출하고, step2 ~ step5를 반복한다.

제 3 절 감정분류 학습모델

효과적인 문서 감정 분류를 위해서 가장 중심이 되어야 하는 부분이 자질의 선정 방법이다. 본 논문에서는 문서의 감정 분류를 위해서 사용될 충분한 양의 감정 자질을 표층 자질과 의미적 자질로 정의하여 추출한다. 생성된 감정 자질을 이용하고 제4절의 감정분류 알고리즘을 적용하여, 문서에 대한 감정을 분류한다.

표 7 연구에서 사용된 자질의 종류

자질 구분	내 용
표층 자질	어절, 형태소 분석, 품사태깅, 이모티콘
의미적 자질	Doc2Vec, Skip-Thought
하이브리드 자질	표층 자질 + 의미적 자질

3.1 표층 자질 추출

감정분류를 위해서는 트윗 데이터를 벡터화하여 트윗을 대표할 수 있는 형태로 변환하는 작업이 필요하다. 한국어 어휘의 자질 벡터를 명시적으로 정의한 자질 집합을 사용한다. 만약 모든 단어를 자질 정보로 사용할 경우, 기계학습으로 처리하기 힘든 수준의 방대한 벡터가 생성되어 정확성과 효율성을 떨어뜨릴 수 있다. 따라서 분류기를 학습하는데 적절한 단어들을 자질로 선택하는 것은 중요하다.

3.1.1 문장 수준 자질

문장수준에서 사용되는 용어들은 대부분 감정을 분류하는 것을 목적으로 사용하기 때문에 문서수준에서의 분석과는 달리 빈도수를 고려하는 것보다 문장의 용어들이 표현하는 감성을 분석한다. 그러나 문서수준의 분석에서는 입력변수로 사용될 용어들을 적절하게 추출하여 사용함으로써 문서를 분류하는데 효과적이다. 입력변수의 선정은 모형의 정확성에 많은 영향을 미치며, 입력변수가 잘못 선정된 경우 모형의 예측정확성은 현저히 낮아지며 모형의 성과가 달라진다[57].

한국어 문장 (a)“하트 같이 만들어서 기뻤어”와 (b)“오늘도 건강하고 기쁘게” 에 대한 구문분석은 [그림 3]과 같다. 한국어는 영어와 달리 비교적 자유로운 어순과 구성성분의 잦은 생략 등으로 인해 구구조 표현보다는 주로 의존구조 표현을 사용하여 구문분석을 수행한다. 의존 구문분석이란 단어와 단어 간의 비대칭 의존관계를 파악하는 과정으로 문장을 구성하는 각 단어의 지배소를 찾는 과정이라고 할 수 있다. [그림 3]은 구문분석 결과를 트리로 표현한 것이다. 의존 구문분석을 이용하여 한국어 구문 분석을 하는 이유는 첫째로 어순의 자유성에 의한 어절의 위치 문제가 의존 구문분석에서는 쉽게 해결된다. 둘째, 구성요소의 불연속성이나 구성요소의 생략 등과 같은 현상에 큰 영향을 받지 않으며 따라서 매구 견고성이 있는 파싱 방법을 구축할 수 있기 때문이다.

파스 트리(parse tree)는 ROOT로부터 시작한다. 문장(a)에서 확인 할 수 있듯이 한국어의 경우 ROOT의 자식으로 Sentence(S)가 오지 않는 경우도 있다. 문장(a)의 작성자는 하트 같이 만들어서 “기쁘” 다는 감정을 표현했으며, 문장(b)의 작성자는 본인보다는 상대방이 건강하고 “기쁘”기를 희망하고 있다. 이처럼 한국어의 경우, 주변 구문구조에 따라 서로 다른 의미를 전달하기 때문에 문장의 구문정보는 중요하다.

(a) 하트 같이 만들어서 기쁘었어

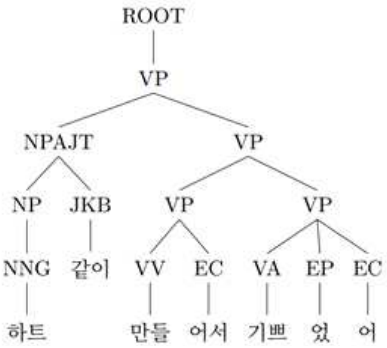


그림 2 구문 분석(a)

(b) 오늘도 건강하고 기쁘게

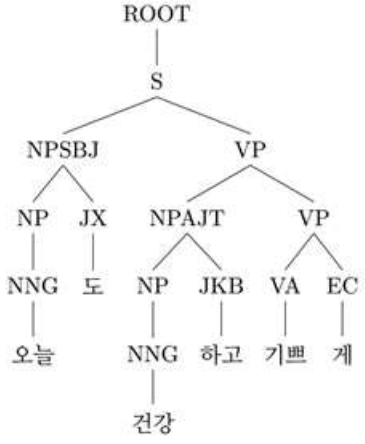


그림 3 구문 분석(b)

본 연구에서는 문장을 표현하는 특징으로 기호문자, 어절, 품사, Rewrite Rules(RR), Emotion Keyword(EK) 사용하였다. 기호문자는 한국어 단어의 형태소 분석을 위한 가장 작은 단위의 특징이고, 어절은 간단하면서도 전 세계적으로 가장 널리 사용되는 특징이다. 한국어 단어의 경우 문장(a)의 “기뻤어” 같이, 복수의 단어가 합쳐져서 하나의 단어가 될 수 있으므로, 분리해서 사용한다.

문서나 문장 내에 출현하고 있는 어절들을 추출하여 감정분류를 위한

자질로 사용합니다. 이는 어휘기반 접근방식에서 감정용어와 잘 알려진 감정어휘를 사용하여 감정을 분류하기 때문에 어절 자체로도 감정분류를 위해 의미있는 자질이 될 수 있기 때문입니다[58].

품사는 문법 정보를 표현하기 위해서 일반적으로 사용하는 자질이다. 형태소란 의미기능을 부여하는 언어의 형태론적 수준에서의 최소 단위를 의미한다. 문장을 구성하는 단어들로부터 최소 의미단위인 형태소를 분리해 내고 각 형태소들의 문법적 기능에 따라 적절한 품사를 부착해 주고, 필요한 경우 단어의 원형도 복원하는 기술이다. 영어와는 달리 한국어 형태소 분석은 훨씬 복잡하다. 한국어는 교착어적 특성을 지닌 언어로 하나 이상의 형태소와 결합되어 한 어절(한국어 띄어쓰기 단위)을 형성할 수 있기 때문이다. 즉 영어에서의 형태소 분석에서는 하나의 단어가 하나의 형태소에 해당되므로 형태소를 분리하는 일이 없는데 한국어에서는 어절을 구성하는 형태소들을 분리해 내고 각 형태소의 품사를 결정해야 하므로 분석의 복잡도가 상당히 높아진다. 품사 태그에 대한 정보는 부록 Ⅰ의 Penn Treebank part-of-speech tags를 참고한다.

RR은 최근 문장의 구문구조를 표현하는 용도로 빈번히 사용되는 자질로써, 트리를 더 작은 서브트리로 잘게 나누어 사용하는 자질이다. 예를 들어 문장(a)의 파스 트리는 [그림 4]과 같이 표현 할 수 있다.

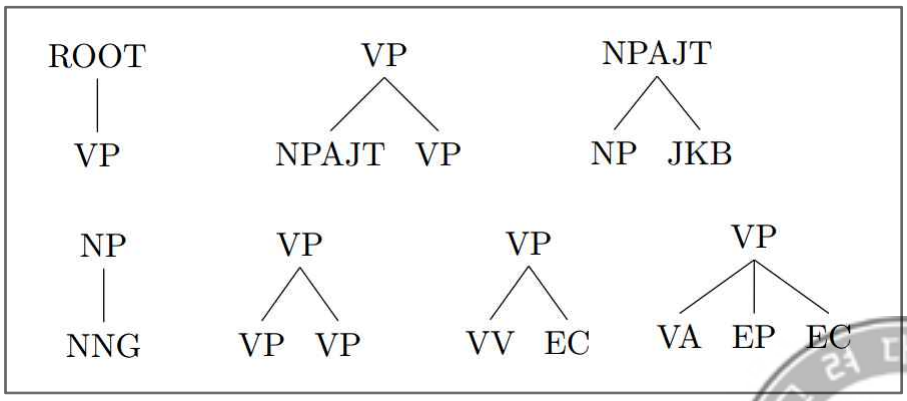


그림 5 RR 자질

EK는 감정분류를 위한 감정키워드 자질이다. 감정 자질로 사용되는 감정키워드는 일반적으로 한국어 감정 분류 시스템에서 자주 사용되는 대표 어휘이다. 보통 긍정 자질, 부정 자질로 구분하여 어휘를 수집하고 대표 어휘를 선정하여 자질 확장하여 사용하기도 한다. 그러나 본 연구에서는 감정범주로 선정한 키워드만 사용하였다[59].

어절과 품사는 순서 정보를 포함하는 n-gram으로 사용하였으며, RR와 EK의 경우 bag-of-tree의 형태로 사용하였다. 문장(a)에 대한 표층 자질은 <표 8>와 같다.

표 8 표층 자질 예제

자질	N그램	예제
어절	1	하트:1, 같이:1, 만들:1, 어서:1, 기쁘:1, 었:1, 어:1
어절	2	하트 같이:1, 같이 만들:1, 만들 어서:1, 어서 기쁘:1, 기쁘 었:1, 었 어:1
어절	3	하트 같이 만들:1, 같이 만들 어서:1, 만들 어서 기쁘:1, 어서 기쁘 었:1, 기쁘 었 어:1
품사	1	NNG:1, JKB:1, VV:1, EC:2, VA:1, EP:1
품사	2	NNG JKB:1, JKB VV:1, VV EC:1, EC VA:1, VA EP:1, EP EC:1
품사	3	NNG JKB VV:1, JKB VV EC:1, VV EC VA:1, EC VA EP:1, VA EP EC:1
RR	–	ROOT VP:1, VP NPAJT VP:1, NPAJT NP JKB:1, NP NNG:1, VP VP VP:1, VP VV EC:1, VP VA EP EC:1
EK	–	기쁘:1

이모티콘은 작성자의 감정이나 상태를 시각적으로 강조해 준다. 소리와 모습을 대표한 'ㅋ', 'ㅎ' 'ㅠ/ㅜ'는 생동감 있게 자신의 감정을 표현한다[60] 전달력이 글보다 빠르며, 말로 표현해 내야 하는 수고로움 없이 간결하고 분명하게 감정을 표현할 수 있다.

트위터에 사용된 이모티콘은 해당 트윗에 감정이 포함되어 있는지 판단하기 좋은 자질이다. 이모티콘은 길이도 무척 다양하게 변형되어 사용되기 때문에 단어 단위의 자질도 유용하지 않다. 이모티콘을 위한 자질은 기호문자 수준의 n-gram이 사용되었다. <표 9>는 "ㅠㅜㅠㅠㅜ" 이모티콘이 어떻게 기호문자(n-gram)으로 표현되는지 보여준다.

표 9 기호문자 자질 예제

자 질	N그램	예제
기호문자	1	ㅠ:3, ㅜ:2
기호문자	2	ㅠㅜ:2, ㅜㅠ:1, ㅠㅠ:1
기호문자	3	ㅠㅜㅠ:1, ㅜㅠㅠ:1, ㅠㅠㅜ:1

3.1.2 문서 수준 자질

하나의 트윗은 복수의 문장과 복수의 이모티콘으로 되어 있다. 트윗에 포함된 다수의 문장과 이모티콘을 문서 수준의 자질로 사용하였다. 하나의 트윗은 감정 단어를 포함한 문장과 그렇지 않은 문장으로 나눌 수 있다. 감정 단어가 포함된 문장 수준의 자질들은 그렇지 않은 문장보다 감정을 예측하는데 큰 영향을 주기 때문에 감정 단어가 포함된 문장가 그렇지 않은 문장을 구분할 필요가 있다. 또한 이모티콘은 앞서 문장을 나누는 역할을 한다. 특히 이모티콘은 바로 앞에 위치한 문장에 대한 감정 정도를 나타내므로, 이모티콘은 이전 문장에 의존적이다. 이를 고려한 문서 수준의 자질은 다음과 같이 구성하였다.

$$F_{i,inc} = Word_{i,inc} + POS_{i,inc} + RR_{i,inc} + CH_{i,inc}$$
$$T_i = F_{i,true}F_{i,false}$$

(식 2)

T_i 는 i번째 트윗의 자질을 의미한다. $F_{i,true}$은 i번째 트윗에서 감정단어를 포함한 문장의 자질이며, $F_{i,false}$ 는 감정단어를 포함하지 않은 문장의 자질이다. $Word_{i,n}$, $POS_{i,n}$은 i번째 트윗에서 감정단어 포함 유무에 따른 Word와 POS의 n-gram 자질이다. $RR_{i,n}$은 i번째 트윗에서 감정단어 포함 유무에 따른 RR의 bag-of-tree 자질이다. $CH_{i,n}$은 i번째 트윗에서 감정단어 포함 유모에 따른 이모티콘의 기호문자 수준의 n-gram 자질이다.

3.2 의미적 자질 추출

전통적인 기계학습의 경우, 특정 도메인에 맞는 학습 데이터를 통해 사람이 명시적으로 정의한 자질 집합을 사용하여 학습이 진행된다. 일반적으로 충분한 학습 데이터를 구축하는 것이 매우 어려울 뿐만 아니라 어휘 자질의 과도한 사용은 데이터 부족 문제를 유발한다. 즉, 학습 데이터에 나타난 고빈도 감정 어휘에 대해서는 잘 동작하지만, 학습 데이터에 나타나지 않은 “이것 밖에 안돼(실망)” 등의 표현에 대해서는 적절한 감정을 결정할 수 없기 때문에 학습에 효율적인 자질을 설계하는 것 또한 매우 중요하다.

본 연구에서 설계한 명시적 자질 집합들은 텍스트에 대한 문장 구조와 구문적 요소를 표현하는 데는 큰 문제가 없지만 구, 문장, 문맥(context) 등 의미(semantic) 정보를 표현하는 데는 한계가 있다. 본 연구에서는 대량의 학습 데이터 확보의 어려움, 문맥을 통한 감정의 지속성 등의 문제를 풀기 위해 단어나 문장을 그 의미와 맥락을 고려하여 벡터 공간으로 매핑시키는 임베딩 벡터를 감정분류 모델의 의미적 자질로 사용하였다. 하나의 단어에 대해 수백 개 정도의 저차원의 실수 벡터로 표현하고, 이와 같은 단어에 대한 분산(distribution) 표현은 다양한 벡터 연산을 통해 단어들 간의 관계를 유추할 수 있게 해준다. 분포가설에 의하면 비슷한 문맥을 가진 단어는 비슷한 의미를 갖는다[61]. 텍스트 문서 내에서 문장 내 ‘한 단어’와 같이 출현하는 다른 단어들을 ‘주변 단어’로써 인공 신경망에서

학습을 시킨다. 연관된 단어들은 문서 내에서 가까운 곳에 출현할 가능성이 높아지기 때문에 학습을 반복해 나가는 과정에서 '주변 단어'가 비슷한 단어는 벡터 공간에 비슷한 위치에 놓이게 된다. 텍스트에서 감정분류를 위해 현재 문장과 주변 문장에 대한 벡터 정보를 갖고 있다면 그 의미를 정확하게 파악하는데 도움이 될 것이다. 본 연구에서는 인공 신경망 기반의 언어 모델 학습인 Doc2Vec과 Skip-Thought vectors을 사용하여 의미적 자질로 사용하였다. 이 두 모델은 단어의 순서나 문장의 순서를 고려하여 벡터 공간으로 임베딩하기 때문에 문맥 정보를 보다 효율적으로 표현할 수 있다.

3.2.1 Doc2Vec

Doc2Vec은 단어 자체가 가지는 의미를 다차원 공간에서 벡터화하는 방식으로 단어들을 실수 공간에 매핑 시키고, 이들 사이의 유사도를 계산함으로써 단어 사이의 의미적 관계를 표현하는 워드 임베딩 기법 중 문맥 정보를 획득하기 위해 Mikolov가 제안한 모델이다[27]. Doc2Vec 모델은 Word2Vec모델과 거의 유사한 학습 방법을 갖고 있으며, 대표적으로 DM(distributed memory), DBOW(distributed bag of words) 알고리즘이 있다[42]. Doc2Vec은 단어의 순서와 의미를 내포한 벡터 표현을 생성해낸다. 즉, 문서에 포함된 어휘는 동일하나 순서가 다르면 다른 벡터를 생성하고, 의미가 유사한 문서들은 벡터 공간상에 가까이 위치하도록 벡터를 생성하는 것이다. 이렇게 생성된 벡터들은 감정분류 학습모델에 사용된다.

DM 알고리즘은 문서행렬 D가 추가된 것을 제외하면 Word2Vec의 CBOW(continuous bag of words)와 유사하다. "the cat sat on"이라는 문장이 존재할 때, "on"을 예측하기 위하여, "the cat sat" 3개의 단어 벡터와 문서 벡터를 평균 또는 연결(concatenation)한 벡터를 사용한다. Word2Vec과 마찬가지로 학습을 마치게 되면, 가중치 행렬을 문서 벡터와 단어 벡터로 사용할 수 있게 된다. 결국 문서에 포함된 단어를 예측하는 작업의 간접

적인 결과로 순서와 의미를 내포함 문서 행렬 D와 단어 행렬 W를 얻게 된다. [그림 5]는 CBOW, Skip-gram 모델에서 입력과 출력의 차이를 보여준다.

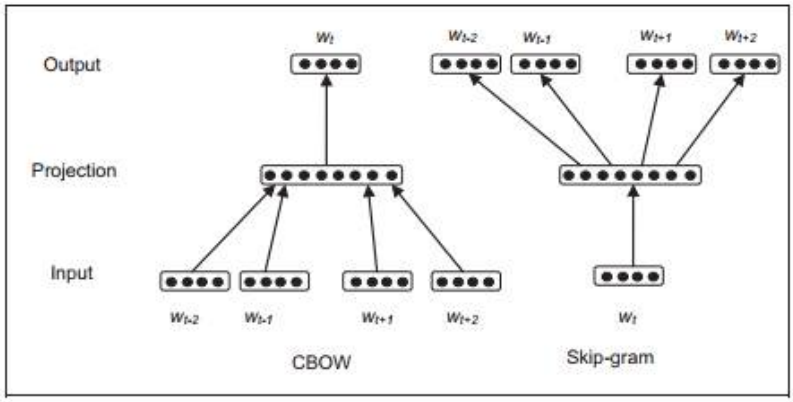


그림 5 CBOW, Skip-gram 모델

Doc2Vec은 문서 벡터를 단어 벡터와 같이 하나의 또다른 단어처럼 학습을 위한 입력으로 받아, 현재 문맥에서 누락된 정보를 기억하는 역할을 수행한다. 학습이 끝나면 Doc2Vec 모델에서 문서 임베딩 행렬을 얻을 수 있으며, [그림 6]처럼 문서 임베딩 벡터 공간을 얻게 된다.

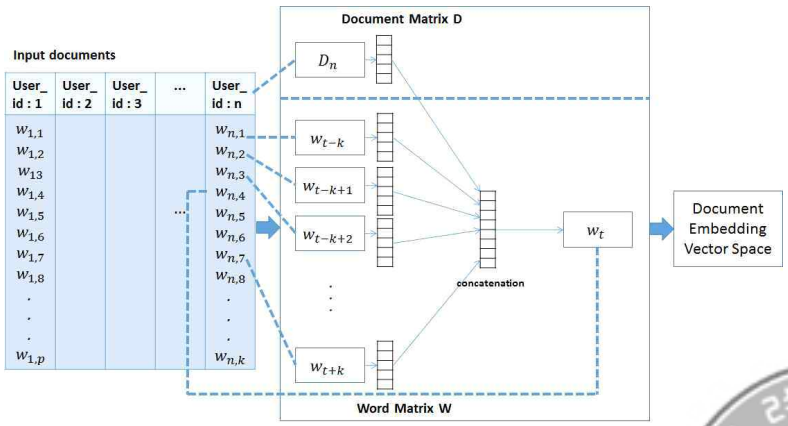


그림 6 Doc2Vec 처리과정

3.2.2 Skip-Thought Vector

본 연구에서는 감정분류를 위해 연속적인 문장의 내재된 자질을 학습하기 위해 Skip-Thought vector를 사용하였다. Skip-Thought Vector는 Skip-Thought 모델에서 사용되는 벡터이다.

Skip-Thought 모델은 워드 임베딩 기법 중 Skip-Gram 모델을 문장 단위로 확장한 기법으로 Skip-Gram 모델이 특정 단어의 주변 단어들을 예측하기 위해 단어를 사용하는 대신, Skip-Thought 모델은 단어나 문장을 인코딩하여 주변 문장을 예측한다. Skip-Thought 모델은 여러 순환 신경망 기법 중 GRU-인코더, GRU-attention 디코더로 구현되었다. GRU(gated recurrent unit)[62]는 순환 뉴럴 네트워크(RNN, Recurrent Neural Network) 종류 중 하나로 노드간 거리가 멀어질 경우, 그 정보를 현재 노드에 반영하기 어렵기 때문에 원하는 노드의 정보를 학습에 반영하고자 고안된 방법이다. 이전 노드의 출력이 현재 노드의 입력으로 사용되기 현재 노드의 출력에 영향을 줄 수 있다. [그림 7]는 GRU-인코더, GRU-attention 디코더를 나타낸다. 본 연구에서는 단어나 문장을 문장 벡터로 인코딩하는 Skip-Thought 모델의 GRU-인코더만 사용하여 문장에 대한 의미적 자질을 추출하였다. GRU-인코더 마지막 시퀀스의 출력이 skip-thought vector이다.

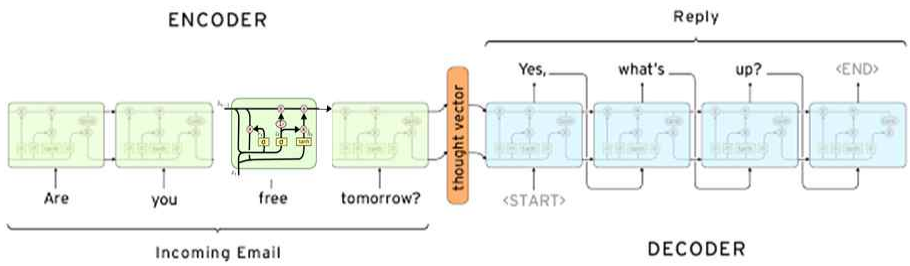


그림 7 GRU-인코더 / GRU-디코더

GRU-인코더는 순환 신경망으로 학습한다. 순환 신경망은 단어와 단어

사이의 연관성과 시퀀스 즉, 시계열 특성을 반영한 모델이다. 특정 시점의 노드는 그 시점의 입력값과 더불어 이전 시점의 은닉층의 값을 활용하여 상태를 결정한다. 이전 단어에 대한 정보가 계속해서 누적되어 현재 시점의 상태에 영향력을 발휘할 수 있는 모델이며, 일반적으로 그 구조는 [그림 8]와 같다.

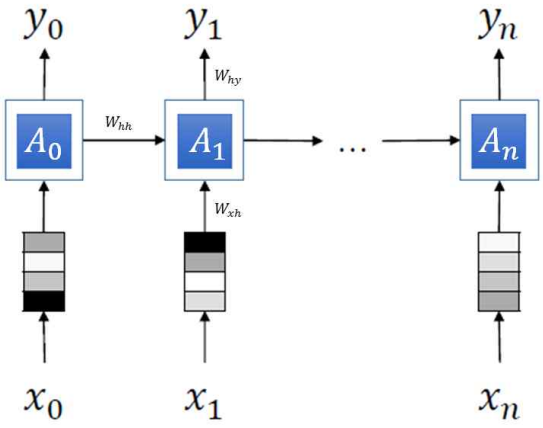


그림 8 순환 신경망 구조

A노드가 은닉층에 해당되며, x_1의 은닉층은 현재 입력층에서 계산된 값과 이전 노드인 x_0에서 계산된 은닉층의 값으로 계산되며 수식으로 표현하면 (식 3)과 같다.

$$h_t = f_w(h_{t-1}, x_t) \qquad \text{(식 3)}$$

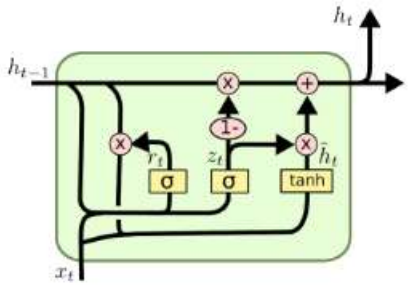
t는 단어의 순서 또는 시계열 순서이다. 은닉층에서 활성함수(activation function)가 적용되기 때문에 (식 4)로 표현되며, 각 시점에서 출력층은 (식 5)로 표현된다.

$$h_t = tanh(W_{hh}h_{t-1} + W_{xh}x_t) \quad \text{(식 4)}$$

$$y_t = W_{hy}h_t \quad \text{(식 5)}$$

시간 t일 때, W_{hh}는 이전 은닉층$(t-1)$에서 현재 은닉층(t) 사이의 가중치 값, W_{xh}는 입력층t에서 은닉층 사이의 가중치 값, W_{hy}는 은닉층에서 출력층 사이의 가중치 값을 의미한다.

GRU는 갱신(z), 리셋(r)의 두 개의 게이트를 통해 정보의 흐름이 조절된다. 장단기 기억 네트워크의 잊기와 입력 게이트들을 하나의 단일 "갱신(z) 게이트"로 합치고, 그것은 또한 셀 상태와 숨겨진 상태(hidden state)를 합친 모델로 장단기 기억 네트워크를 단순화 시킨 구조를 갖는다.



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

그림 9 GRU 구조

3.3 하이브리드 자질

본 연구에서는 명시적으로 정의한 표층 자질과 임베딩 방법으로 자동으로 추출된 의미적 자질을 결합하여 학습에 사용하였다. 본 연구에서는 표층 자질과 의미적 자질을 모두 감정분류 모델에 사용된다. 두 개의 자질은 합성곱(convolution function)을 이용하여 결합한다.

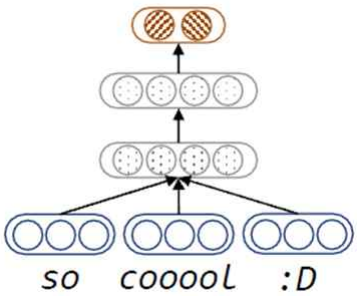


그림 10 합성곱

제4절 기계학습 기반 감정분류

4.1 모델 생성

문서 임베딩을 이용한 학습 모델 생성을 위해 트위터에서 최신 트윗을 수집한다. 트위터는 리플라이(@replies)와 멘션(@mentions)으로 대화가 이루어지며, 90%가 이 리플라이와 멘션으로 특정 트위터 사용자에게 대화를 전달한다. 리플라이는 상대방 트위터 이름을 적고 자신의 의견을 표현하는 방식이며, 멘션은 상대방 트위터의 이름이 맨 앞에 언급하는 것이 아니라 글 사이에 높인다는 차이점이 있다. 즉, 상대방의 말에 대답/응답할 때는 이름을 맨 앞에 놓고, 누군가를 언급할 때는 글 사이에 이름을 적어 표현한다. <표 10>은 트위터에서 리플라이와 멘션의 예를 나타낸

다.

표 10 트위터에서 리플라이와 멘션의 예

리플라이	@coolof 핸드폰 케이스가 참 맘에 드네요.
멘션	오늘 이곳 날씨가 참 좋군요. @aa 님이 계신 곳은 어떤가요?

본 연구에서는 doc2vec 모델의 학습 효율을 높이기 위해 동일 주제나 인물 등에 대해 언급한 데이터를 수집한다. 이를 위해 리플라이 형태로 서로 트윗을 주고받는 사용자 위주로 데이터를 수집하였다. 수집된 데이터는 동일한 전처리 과정을 거쳐 노이즈를 제거하고, 한 개의 트윗을 한 개의 문서로 간주하여 사용한다. 한 개의 트윗은 140자 이내의 문자로 이루어지며, 문장 개수는 가변적이다. 각 트윗에 있는 문장들은 마침표를 제거하여 한 개의 라인으로 연결한다. <표 11>처럼 트윗 한 개는 한 개의 문서가 되고 이는 한 개의 라인으로 나열된 형태로 사용된다.

표 11 입력 예제

1	어제 친구랑 문자하면서도 느꼈다 '맥주 한잔 하겠냐' 물으니 친구는 '내일 중요한 강연과 술자리가 있어 다음주에 보자'했다 단 한 문장 안에 답변과 이유와 비전제시가 있는 이런 의사소통을 좋아한다 그러다 보니 여러 번 묻게 하는 사람이 참 힘들다
2	#아파트분양원가공개 좌절 이후... 2004년하반기부터 부동산가격 폭등! 서민의 삶은 피폐해지고, 강부자들만 살기좋은 #헬조선 강남복부인들의 노무현 칭송이 자자해짐...ㅠㅠ

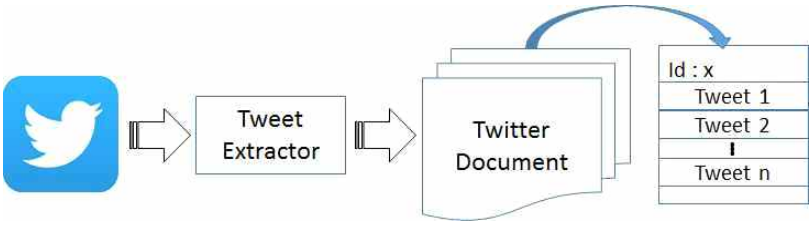


그림 11 트위터 문서 생성 과정

4.2 입력층과 출력층

본 연구에서는 단락 임베딩 기법 중 문서 정보와 단어의 순서 정보를 모두 사용하는 DM기법을 사용하여 학습하였다. 이는 의미 분석 실험에서 DBOW보다 좋은 성능을 보였기 때문이다. Word2Vec은 한 개의 단어를 한 개의 벡터로 표현한다. 그러나 Doc2Vec은 이뿐만 아니라 문서에 대한 정보를 입력 단어와 같이 사용하여 현재 문맥에서 누락된 정보를 기억하도록 한다. 현재 단어와 예측 단어 사이의 거리인 windows size가 5인 경우, 입력층에 사용되는 데이터는 다음과 같다.

```
[['word1', 'word2', 'word3', 'word4', 'word5', 'lastword'], 'label1']]
```

4.3 모델학습

문서 벡터와 단어 벡터는 역전파를 이용한 확률적 기울기 하강 기법으로 학습이 수행된다. 문장의 각 문장의 감정 레이블과 함께 학습된다. 이 방법으로 각 문장 레이블이 문장의 의도를 나타내도록 DM을 학습시켰다. 너무 적은 수의 학습 데이터는 오히려 학습에 방해가 될 수 있기 때문에 전체 데이터에서 5번 이상 나타났던 단어들만으로 학습하였다. 주어진 데이터로 W(word vctors), U(softmax weights), b(softmax weight), D(paragraph vectors)를 학습한다. 새로운 문장에 대해서 이미 학습했던 W, U, b는 유

지하면서 D에 column을 추가하고 기울기 하강 기법을 적용하여 새로운 문장에 대한 벡터(D')를 만들게 된다. 이 단계에서 얻은 D'는 인식기를 통해 미리 정의된 25개 감정 중 하나로 분류된다. 즉, 학습된 doc2vec 모델은 새로운 문장이 입력으로 들어오면, 새로운 문장에 대한 벡터를 생성할 수 있다. 여기서 추출한 doc2vec의 벡터로 감정분류를 위한 인식기로 사용하였다. 본 연구에서는 여러 인식기를 사용하여 감정을 분류하였다.

4.4 감정분류 알고리즘

4.4.1 서포트 벡터 머신

서포트 벡터 머신은 Vapnik(1995)에 의해 처음 소개된 이진 선형 분류기(binary linear classifier)로서 적은 학습데이터, 고차원의 자질 공간(feature space)에서 높은 일반화 성능을 보인다는 특징이 있다. SVM과 같은 기계학습은 학습 데이터를 과하게 잘 학습한다. 일반적으로 학습 데이터는 실제 데이터의 부분 집합인 경우가 대부분이기 때문에 학습을 할수록 학습 데이터에 대해서는 오차가 감소하지만, 실제 데이터에 대해서는 오차가 증가하는 시점이 존재할 수 있다. 기계학습에서 이러한 과적합(overfitting) 문제를 해결하기 위한 방법이 계속 연구 중이며, 그 중 가장 정교한 방법이 바로 최적화 기법을 사용하는 것이다. 인공 신경망, 에너지 기반 모델(energy-based model), 서포트 벡터 머신(SVM) 등이 있다. 이 중 서포트 벡터 머신은 기존 인공 신경망이나 에너지 모델과는 다르게 데이터를 분류하는 기준점(기준 데이터)을 찾는 것과 동시에 각 데이터 집합과 기준점과의 거리(margin)을 최대화 하는 방식으로 학습을 진행한다. 서로 다른 데이터 집합이 [그림]와 같이 존재한다고 할 경우, 기계학습 알고리즘은 A와 B라는 두 직선에 해당하는 분류기를 생성하여 데이터를 분류한다. 이때 서포트 벡터 머신은 항상 두 그룹으로부터 마진이 최대인 A를 선택함으로써 새로운 데이터가 입력될 때 오류의 가능성을 최소화한다.

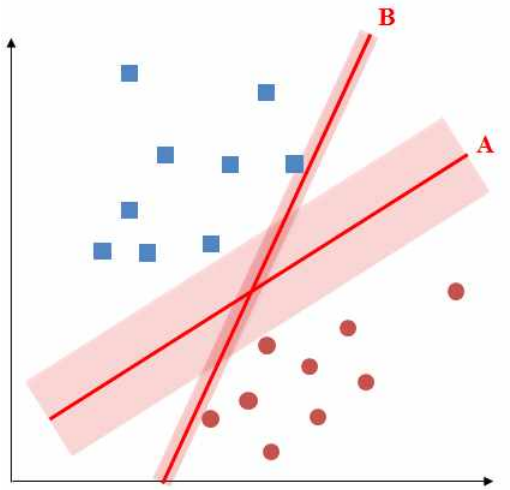


그림 12 SVM 분류기

이러한 학습 방식때문에 서포트 벡터 머신은 현재까지 기계학습 알고리즘 중에서도 뛰어난 성능으로 평가받고 있으며 자연어처리뿐만 아니라 다양한 분류 문제에 널리 사용되고 있다[63, 64]. 본 연구에서도 서포트 벡터 머신의 이러한 특징 때문에 감성분류를 학습하는데 사용하였다.

(1) $f(x) = sign(\sum_{(y_i,x_i)\in SV_s} \alpha_i y_i K(x_i,x) + b)$ (식 6)

(2) $polybomial : K(x_i, x_j) = (x_i \bullet x_j + 1)^d$ (식 7)

SVM의 학습과정은 L개의 학습데이터 $\{(x_i, y_i) | 1 \le i \le L, x_i$ 는 n차원의 자질벡터, $y_i \in (+1, -1)\}$ 로부터 최대 마진을 가지는 초평면(hyperlpane) w•x + b = 0을 찾는 것으로 최적의 초평면 f(x)는 (식 6)와 같이 표현된다. α_i는 각 제약조건에 대응하는 라그랑지 곱수(lagrange muliplier)이며, K는 선형분리가 불가능한 경우에도 자질벡터를 고차원의

자질공간으로 사상시킴으로써 선형분리가 가능하게 만들어주는 커널 함수이다. 데이터의 성격에 따라 활용 가능한 여러 커널함수(K) 중 하나를 선택하여 사용할 수 있는데, 본 논문에서는 (식 7)의 다항 커널함수(polynomial kernel function)를 사용하였다. 이 커널함수는 계산량의 큰 증가 없이 d개의 자질들을 조합하여 사용한 것과 같은 효과를 가진다.

4.4.2 로지스틱 회귀분석

로지스틱 회귀분석(Logistic Regression, LR)은 통계학 및 데이터 마이닝 분야의 다양한 분야에서 널리 사용된다. 이러한 로지스틱 회귀분석은 출력 및 입력 변수들 간의 관계를 탐구하는 모델로서 이진 혹은 다항의 종속 변수(Y)가 어떻게 독립적인 입력 변수의 집합($X = \{X_1, X_2, ...\}$) 과 서로 관련되는 지를 나타낸다. 이진 로지스틱 회귀는 주어진 관측 값들에 따른 출력 클래스의 사후 확률을 모델링하며 이때에 로지스틱 회귀는 입력과 출력 사이의 관계가 입력의 선형 조합을 이용하는 로지스틱 방정식의 형태로 추정된다고 가정한다. 이러한 로지스틱 회귀분석에서 입력의 각 속성은 각자의 가중치와 결합된다. 감정분류를 위한 알고리즘으로 로지스틱 회귀분석을 사용하여 25개의 감정을 분류하였다.

제 4 장 실험 및 결과

본 연구는 기계학습 기반의 한국어 감정분류를 위한 연구방법을 제안하였다. 세분화된 감정분류의 정확성을 높이기 위해서는 한국어 감정 말뭉치가 필요하다. 본 연구에서는 감정분류의 정확성을 높이기 위해 직접 한국어 감정 말뭉치를 구축하였고, 수집된 트윗 데이터에 10명의 언어 심리학에 능숙한 대학원생을 통해 수동으로 감정 범주를 부착하였다. 각 감정 범주별로 420개, 총 10,500개의 트윗에 감정 범주를 부착했으며, 이를 대상으로 학습을 진행하였다. 발생 빈도가 낮은 표면적 자질은 일반적으로 사용 사례가 적기 때문에 감정분류에 유용한 정보가 아니라고 판단하여 5번 이상 나타나지 않은 자질은 제거하였다.

본 장에서는 첫째, 직접 구축한 한국어 감정 말뭉치에 대한 신뢰도 평가를 먼저 수행한다. 한국어 감정 말뭉치는 기계학습에서 실제로 학습에 사용되는 데이터이기 때문이다. 본 연구에서는 감정의 범주가 25개로 다양하기 때문에 모든 범주에 대하여 동일한 수준과 동일한 양의 학습 데이터를 수집하기는 쉬운 일이 아니다. 소셜 미디어 상에 자주 언급되는 감정에 대해서는 데이터가 많지만, 그렇지 않은 감정에 대해서는 학습 데이터가 많지 않다. 이를 해결하기 위해 오버샘플링 기법을 적용하였고, 이에 대한 실험결과를 먼저 평가하였다.

두 번째로, 본 연구에서는 한국어 감정분류에 기반이 되는 자질을 표면적 자질과 의미적 자질로 정의하였고, 이에 대한 유용성을 평가하였다. 한국어 감정 자질은 감정을 지닌 대표어휘로부터 시작하여 확장할 수 있으며, 형태학적, 구문론적 감정 자질들과의 조합을 통해 감정분류 성능을 평가하였다. 이뿐만 아니라 워드 임베딩 기법을 사용한 의미적 자질을 사용한 감정분류 성능평가와 표면적 자질, 의미적 자질의 조합에 따른 감정분류에 대한 성능을 평가하였다. 이는 10-교차검증으로 진행되었다.

성능평가를 기법은 정확률(Precision), 재현율(Recall), F-measure, 정확도

(Accuracy)을 사용하였으며, (식 8)과 같다.

$$\mathrm{Precision} = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F = \frac{2 \times Recall \times \mathrm{Precision}}{Recall + \mathrm{Precision}} \qquad \text{(식 8)}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

제 1 절 한국어 감정 말뭉치 분석

본 연구에서 구축한 한국어 감정 말뭉치는 25개의 감정 범주에 대한 태그가 부착되어 있으며, 한 명의 주관적인 견해로 감정이 결정되는 것을 막기 위해 여러 명의 전문가에 의한 판단을 교차 검증하여 감정 범주를 부착하였다.

말뭉치 구축 과정에서 감정 어휘가 포함된 문장이 해당 감정 범주에 속할 경우, 'acceptable', 그렇지 않을 경우 'unacceptable'로 태깅하였으며, 각각에 대한 빈도수를 계산하여 acceptable ratio를 계산하였고, 그 식은 다음과 같다.

$$Accept^{i}_{ratio} = \frac{\left|Sentence^{i}_{accept}\right|}{\left|Sentence^{i}_{accept}\right| + \left|Sentence^{i}_{unaccept}\right|} \qquad \text{(식 9)}$$

(식 9) 에서 i는 25개 감정 범주 중 i번째 감정 어휘, $Sentence^{i}_{accept}$ 는 i 번째 감정 어휘를 포함한 문장이 해당 감정범주로 '*acceptable*' 된 문장의

개수, $Sentence^{i}_{unaccept}$는 i번째 감정 어휘가 포함된 문장이 해당 감정범주로 '*unacceptable*' 된 문장의 개수를 의미한다.

표 12 감정범주별 acceptable ratio

positive		negative	
감정 범주	*Accept Ratio*	*감정 범주*	*Accept Ratio*
고맙	89.05%	싫다	83.81%
반갑	78.33%	아쉽	76.74%
기쁘	60.95%	미안	70.95%
좋다	60.48%	창피	61.37%
사랑	52.09%	속상	60.48%
만족	29.28%	부럽	55.95%
재미	28.16%	무섭	55.71%
행복	24.11%	슬프	51.91%
즐겁	18.28%	우울	47.14%
편안	17.85%	불안	39.29%
		실망	39.05%
		후회	20.00%
		외롭	13.57%
		그립	3.33%
		우습	2.85%

<표 12>는 (식 8)의 계산 결과를 보여준다. '*acceptable*' 비율은 감정 범주별로 최소 2.85%에서 최대 89.05%까지 매우 다양하게 나타나고 있다. '고맙(89.05%)', '반갑(78.33%)', '싫다(83.81%)', '아쉽(76.74%)', '미안(70.95%)' 에 해당하는 감정 범주는 단어의 어휘적 특징이 해당 감정을 표현하는데 크게 무리가 없지만, '즐겁(18.28%)', '편안(17.85%)', '외롭(13.57%)', '그립(3.33%)', '우습(2.85%)'에 해당하는 감정범주는 학습 데이터가 균형적이지 않음을 확인할 수 있었다. 실제로 'acceptable' 비율이 낮은 감정 범주는 자신의 감정이 아닌 상대방의 감정이나 중의적인 표현

으로 더 많이 사용되고 있으며, <표 13>는 '우습', '그립', '편안', '즐겁' 감정 범주에 대한 실제 사례를 보여준다.

표 13 상대방 감정 또는 중의적으로 표현된 문장

우습	@Yui_DBR 살아있을 땐 따분하고 짜증나는 놈이라고 생각했었는데, 알면 알수록 우습네.
	@Yui_DBR 굳이 '네가' 라기보단. 죽은 사람한테 우습다니 너무해! 가 일반적인 반응...
	검차를 넘 우습게 보는거네요..ㅎㅎ@son5959
그립	내가그리웠니 ?
	토르로키 잘 안그리는 이유 : 갑옷시발
	벨바는 시간이넘쳐날때 그림마니그리는데..요새그림안그리는걸보면 정말 답없이살고있단것을나타낸다...
편안	편안히 주무세요~ ^^ http://t.co/J7t32k2gS0
	오늘도 좋은 하루 되시길. 당신의 영혼에 편안한 안식처를..
즐겁	@aura_ion ㅋㅋㅋㅋㅋㅋㅋㅋ우라님 피카츄 즐겁게 즐기고 오셔요ㅠㅠㅠㅠㅠㅠㅠ
	이제부터 즐거운 주말이네요. 즐거운 주말 보내세요~

트위터에서 '우습', '편안', '즐겁' 감정범주는 현재 자신의 감정 상태를 표현하기보다는 상대방에 대한 감정 상태를 언급할 때 자주 사용되었으며, '그립' 감정범주는 한글의 자모로 인해 '그리다'라는 용언과의 구분이 어려워 '그림을 그리다'는 행위를 표현한 문장들이 많이 나타났다.

표 14 확장된 학습 말뭉치

감정범주	데이터 크기	accept class	unaccept class	total
즐겁	기본 말뭉치 (SMOTE percentage=0)	76 18.28%	344 81.72%	420
	SMOTE percentage=200	152 30.65%	344 69.35%	496
편안	기본 말뭉치 (SMOTE percentage=0)	74 17.85%	346 82.15%	420
	SMOTE percentage=200	148 29.96%	346 70.04%	494
외롭	기본 말뭉치 (SMOTE percentage=0)	57 13.57%	363 86.43%	420
	SMOTE percentage=200	114 23.90%	363 76.10%	477
그립	기본 말뭉치 (SMOTE percentage=0)	13 3.33%	407 96.67%	420
	SMOTE percentage=800	104 20.35%	407 79.65%	511
우습	기본 말뭉치 (SMOTE percentage=0)	11 2.85%	409 97.15^	420
	SMOTE percentage=800	88 17.71%	409 82.29%	497

학습 말뭉치는 일정 규모 이상의 크기를 갖추고 내용적으로 다양성과 균형성이 확보될 때 학습 말뭉치를 사용하는 지도학습의 성능을 높일 수 있기 때문에 본 연구에서는 acceptable ratio 20% 미만인 감정 범주에 대해 오버샘플링 기법을 사용하여 <표 14>와 같이 학습 말뭉치를 확장하였다. 확장된 학습 말뭉치는 감정분류를 위해 사용되었다.

확장된 말뭉치를 사용한 '즐겁', '편안', '외롭', '그립', '우습' 범주는 기본 말뭉치를 사용했을 때 보다 감정분류의 성능을 향상시켰다. <표 15>는 이에 대한 실험결과를 보여준다. '우습' 감정의 경우, 기본 말뭉치 'accept'인 데이터가 11개뿐이어서 '우습' 감정에 대한 다양한 유형의 데이터를 학습하는데 한계가 존재한다. 소셜 미디어에는 더 다양한 형태의 문장들이 존재한다. 따라서 이러한 불균형적인 학습 말뭉치를 확장하면 감

정분류의 성능을 높일 수 있을 것이다.

표 15 확장된 말뭉치 성능 비교

	기본 말뭉치 SVM(F)	확장된 말뭉치 SVM(F)
즐겁	78.80	85.70
편안	78.30	86.10
외롭	83.30	88.40
그립	65.40	96.80
우습	65.70	98.00

제 2 절 표층적 자질을 사용한 감정분류 실험결과

한국어 감정분류를 위한 표층 자질을 위해 어절(n-gram), 품사(n-gram), EK, RR, 기호문자(n-gram)을 실험을 통해 비교하였다.

표 16 긍정에 해당하는 감정범주의 최적화된 자질 조합의 성능비교

	어절			*품사*			*EK*	*RR*	*기호문자*			*P(%)*	*R(%)*	*F(%)*
	1	*2*	*3*	*1*	*2*	*3*			*1*	*2*	*3*			
고맙		✓				✓	✓		✓	✓	✓	*92.1*	*92.4*	*90.9*
반갑	✓			✓					✓			*84.1*	*85.0*	*84.3*
기쁘		✓		✓					✓	✓		*77.6*	*77.9*	*77.6*
좋다	✓						✓					*73.7*	*74.0*	*73.5*
사랑		✓		✓			✓		✓	✓		*75.5*	*75.4*	*75.4*
만족		✓				✓		✓	✓	✓		*81.8*	*82.1*	*81.9*
재미	✓					✓			✓	✓		*74.0*	*75.4*	*74.4*
행복		✓										*82.2*	*83.2*	*82.2*
즐겁		✓		✓				✓				*85.7*	*86.7*	*85.7*
편안	✓			✓			✓					*86.3*	*87.4*	*86.10*

<표 16>는 긍정에 속하는 감정분류에 대한 최적조합을 나타내며, <표 17>는 부정에 속하는 감정분류에 대한 성능이 가장 높은 자질들의 조합

에 대한 실험결과이다. 여러 표층 자질 중 어절, 품사, 기호문자(이모티콘) 자질이 감정을 분류하는데 자주 사용되고 있음을 알 수 있다.

표 17 부정에 해당하는 감정범주의 최적화된 자질 조합의 성능비교

	어절			*품사*			*EK*	*RR*	*기호문자*			*P(%)*	*R(%)*	*F(%)*
	1	*2*	*3*	*1*	*2*	*3*			*1*	*2*	*3*			
싫다		✓				✓		✓				*86.4*	*87.6*	*85.7*
아쉽	✓			✓				✓	✓			*80.9*	*82.0*	*81.1*
미안		✓		✓					✓	✓		*76.4*	*77.6*	*76.2*
창피		✓					✓					*75.4*	*75.6*	*74.7*
속상			✓	✓					✓	✓		*73.2*	*73.6*	*73.1*
부럽	✓				✓				✓			*89.8*	*89.8*	*89.7*
무섭		✓		✓				✓				*75.8*	*75.7*	*75.4*
슬프												*74.8*	*74.8*	*74.7*
우울	✓			✓					✓	✓		*75.8*	*75.7*	*75.7*
불안			✓	✓				✓	✓	✓		*75.3*	*75.2*	*75.3*
실망		✓				✓	✓		✓	✓		*84.6*	*84.5*	*84.5*
후회		✓		✓			✓		✓			*84.1*	*85.2*	*84.3*
외롭	✓						✓		✓	✓	✓	*88.4*	*88.3*	*88.4*
그립	✓						✓		✓			*97.0*	*97.4*	*96.8*
우습		✓						✓	✓			*98.4*	*98.3*	*98.0*

긍정에 해당하는 감정범주에서 어절(2-gram)과 기호문자(1-gram) 자질의 조합이 10개 범주 중 6번 이상 사용되었다. 부정에 해당하는 감정범주에서는 어절(2-gram), 품사(1-gram), 기호문자(1-gram)의 조합이 15개 범주 중 7번 이상 사용되었다. 각 감정범주를 분류하기 위한 자질은 조합은 다양한 형태로 나타나고 있음을 보여준다.

<그림 13>는 25개의 감정을 분류하는데 사용된 표층 자질들의 빈도를

나타낸다. 25개의 감정범주 중 '기호문자(1-gram)' 자질은 18번 사용되었다. 이는 트위터에 사용된 이모티콘에 대한 자질로써 텍스트 내에서 감정을 판별하는 주요 자질로 사용됨을 실험을 통해 확인할 수 있었으며, 한 개의 이모티콘 만으로 감정을 분류하는데 큰 영향을 미침을 알 수 있었다. '기호문자(1-gram)' 자질이 사용되지 않은 감정범주로는 '좋다', '행복', '즐겁', '편안', '창피', '무섭' 이다. 해당 감정범주에서는 '어절(1-gram)', '어절(2-gram)', 'EK'가 주로 사용되었으며, 한국어 어절이 해당 범주를 분류하는데 더 큰 영향을 주고 있음을 알 수 있다.

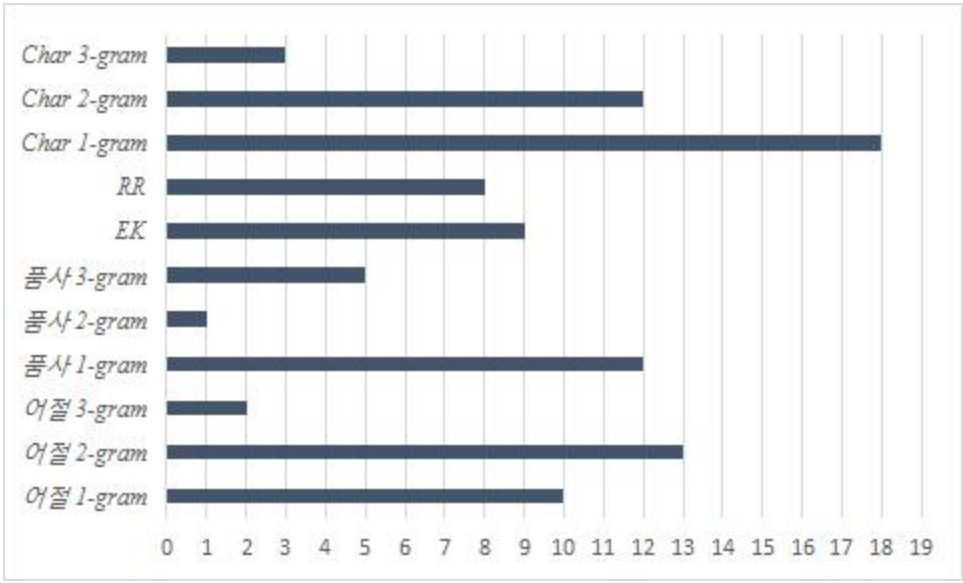


그림 13 최적 성능에서 사용된 자질들의 사용 빈도

가장 빈번하게 사용되는 상위 5개의 표층 자질은 '기호문자(1-gram)', '기호문자(2-gram)', '어절(1-gram)', '어절(2-gram)', '품사' 이다. 이모티콘은 직관적으로 감정 상태를 표현해주기 때문에 이를 제외하면, '어절'과 '품사' 자질이 감정분류에 기여함을 알 수 있다.

'어절'은 최적화된 자질 조합의 성능비교에서 보듯이 감정범주를 분류하는데 빠짐없이 사용되고 있지만, 어절의 1-gram, 2-gram, 3-gram이 동시에

같이 사용되고 있지는 않다. 각 감정범주 별로 하나의 n-gram 어절이 다른 자질들과 결합하여 최적의 성능을 내고 있다. <표 18>는 각 감정범주에서 사용되고 있는 n-gram 어절 중 최적의 성능 조합으로 사용된 어절만 사용한 감정분류 결과이다. '우습', '행복' 감정범주의 경우, 어절(2-gram)만으로 최적의 성능을 보이고 있다.

표 18 n-gram 어절에 따른 성능비교(긍정)

감정범주	*어절*		*F(%)*	*최적조합 F(%)*	*오차*
	1-gram	*2-gram*			
고맙		✓	*88.7*	*90.9*	*2.2*
반갑	✓		*81.4*	*84.3*	*2.9*
기쁘		✓	*74.2*	*77.6*	*3.4*
좋다	✓		*72.8*	*73.5*	*0.7*
사랑	✓		*74.3*	*75.4*	*1.1*
만족		✓	*77.3*	*81.9*	*4.6*
재미		✓	*71.4*	*74.4*	*3.0*
행복		✓	*82.2*	*82.2*	*0.0*
즐겁	✓		*84.1*	*85.7*	*1.6*
편안		✓	*81.9*	*86.1*	*4.2*

표 19 n-gram 어절에 따른 성능비교(부정)

감정범주	어절		F(%)	최적조합 F(%)	오차
	1-gram	2-gram			
싫다	✓		82.8	85.7	2.9
아쉽	✓		79.7	81.1	1.4
미안	✓		72.0	76.2	4.2
창피		✓	71.2	74.7	3.5
속상		✓	69.8	73.1	3.3
부럽	✓		88.0	89.7	1.7
무섭		✓	71.5	75.4	3.9
슬프		✓	71.9	74.7	2.8
우울	✓		72.6	75.7	3.1
불안	✓		67.4	75.3	7.9
실망		✓	82.8	84.5	1.7
후회		✓	81.9	84.3	2.4
외롭	✓		86.5	88.4	1.9
그립	✓		96.4	96.8	0.4
우습		✓	98.0	98.0	0.0

최적 성능에 사용된 자질 중 모든 감정범주에 사용된 '어절' 자질과 상위 5개의 자질에 대한 성능비교는 <표 20>와 같다. '어절' 자질이 감정분류의 성능에 크게 영향을 미치더라도 다른 자질들과의 조합이 무조건 성능을 높이지는 못하며, 또한 감정범주마다 '어절(1-gram)' 또는 '어절(2-gram)'의 영향력이 다르게 나타남을 살펴볼 수 있다.

표 20 다양한 조합에 따른 성능비교

	F-measure (%)				
	어절(1)	어절(1)+기호문자(1)+기호문자(2)+품사(1)	최적 조합	어절(2)	어절(2)+기호문자(1)+기호문자(2)+품사(1)
고맙	*87.3*	***89.5***	***90.9***	*88.7*	***89.7***
반갑	*81.4*	***82.8***	***84.3***	*78.1*	***81.8***
기쁘	*70.0*	***73.2***	***77.6***	*74.2*	***76.3***
좋다	***72.8***	*66.7*	***73.5***	***68.3***	*67.6*
부럽	*88.0*	***89.0***	***89.7***	*81.1*	***84.2***
사랑	*74.3*	***73.9***	***75.4***	*72.7*	***73.3***
만족	*76.2*	***80.4***	***81.9***	*77.3*	***79.9***
재미	*69.8*	***74.4***	***74.4***	*71.4*	***71.7***
행복	***78.6***	*75.6*	***82.2***	***82.2***	*77.9*
즐겁	***84.1***	*82.2*	***85.7***	*83.8*	***83.9***
편안	***85.1***	*83.9*	***86.1***	*81.9*	***82.2***
그립	***96.4***	*95.8*	***96.8***	***96.1***	*95.8*
우습	*95.7*	***96.1***	***98.0***	***98.0***	*96.3*
싫다	*82.8*	***84.3***	***85.7***	*82.5*	***83.1***
아쉽	*79.7*	***82.2***	***81.1***	*77.3*	***81.3***
미안	***78.6***	*70.3*	***76.2***	***82.2***	*70.8*
창피	*67.9*	***73.2***	***74.7***	*71.2*	***73.2***
속상	***68.6***	*65.6*	***73.1***	***69.8***	*68.7*
무섭	*69.2*	***70.3***	***75.4***	*71.5*	***71.8***
슬프	***71.0***	*69.5*	***74.7***	*71.9*	***72.5***
우울	***72.6***	*72.2*	***75.7***	*67.8*	***69.8***
불안	*67.4*	***69.9***	***75.3***	*63.8*	***68.2***
실망	*77.9*	***80.3***	***84.5***	*82.8*	***84.0***
후회	*79.3*	***81.2***	***84.3***	*81.9*	***83.5***
외롭	***86.5***	*86.2*	***88.4***	***86.2***	*86.1*

제 3 절 의미적 자질을 사용한 감정분류 실험결과

본 연구에서 사용한 의미적 자질은 Dov2Vec과 Skip-Thought이며, 데이터 크기와 학습하는데 걸리는 시간은 <표 21>과 같다. 표면적 자질의 학습 시간에 비하면 상당히 빠르게 학습이 이루어졌음을 알 수 있다.

표 21 실험 데이터 정보

의미적 자질	학습 시간	데이터크기	테스트크기
Dov2Vec	60.442 secs	트윗 301,750	12,571
Skip-Thout	20807.481secs	트윗 301,750개	12,571

본 연구에서 사용된 Dov2Vec과 Skip-Thought은 단어가 아닌 문장과 문단으로 데이터를 확장하여 감정분류를 위한 자질로 사용되었다. 의미적 자질로 자주 언급되고 사용되는 Word2Vec은 한 개의 문장 안에 두 개의 단어가 빈번하게 동시에 출현할 경우, 이 두 단어는 같은 의미로 보고 서로 근접한 벡터 공간에 표시된다. 본 연구에서 사용되는 '기쁘' 감정과 '무섭' 감정은 서로 다른 감정범주일 뿐만 아니라 다른 감정을 의미한다. 그러나 Word2Vec의 유사도 계산에서 이 두 단어는 높은 유사도를 보여주었다. 이는 워드 임베딩 기법의 입력과 출력에 관계가 있다. 연속적인 단어 w_{n-1}, w, w_{n+1}가 있을 때, 입력값 w는 출력값으로 w_{n-1}, w_{n+1}를 갖게 된다. 이때 해당 단어의 주변 단어가 비슷하면 해당 단어의 벡터 값도 비슷해진다. 실제로 '기쁘' 감정의 주변단어와 '무섭' 감정의 주변단어의 비슷한 단어들로 인해 성능이 떨어질 수 있음을 의미한다. 실제 <표 22>를 살펴보면, '기쁘' 범주와 '무섭' 범주에서 유사도가 높은 단어목록의 상당부분이 겹치는 것을 알 수 있으며 이는 '기쁘' 범주와 '무섭' 범주가 유사한 의미를 갖고 있는 것으로 해석될 수 있다. 다른 감정범주에 대한 Word2Vec 실험결과는 부록 Ⅱ에 제시하였다.

표 22 벡터 공간에서 '기쁘', '무섭'에 대한 유사도가 높은 어절목록

	유사도가 높은 단어목록
기쁘	고맙,0.776 **부끄럽,0.769 놀랍,0.735 슬프,0.727 안타깝,0.722** 멋지,0.709 반갑,0.706 **아름답,0.685** 예쁘,0.677 사랑스럽,0.666 이렇,0.641 기대되,0.617 **외롭,0.614 두렵,0.612** 즐겁,0.612 아깝,0.611 기뻐,0.610 **아쉽,0.608 귀엽,0.608 괴롭,0.607** 춥,0.605 기뻤,0.601 이쁘,0.600 **흥미롭,0.598** 재밌겠,0.584 길어졌,0.577 **시끄럽,0.577 만족스럽,0.576** 사랑한,0.569 어렵,0.566 기엽,0.561 친절하시,0.561 **부럽,0.559** 불편하겠,0.557 넘치,0.552 스럽,0.551 무겁,0.549 정말로,0.548 무섭,0.548 행복했,0.541 웃기,0.536 멋있다,0.535 느껴지,0.531 잘생겼,0.528 졸리,0.524 배고프,0.524 어지럽,0.523 뿌듯하,0.522 새롭,0.520 그렇,0.518
무섭	어렵,0.820 아깝,0.788 **부끄럽,0.782 슬프,0.774 시끄럽,0.748** 웃기,0.736 **아름답,0.736 외롭,0.725 귀엽,0.723 괴롭,0.712** 스럽,0.708 **아쉽,0.705 안타깝,0.689 부럽,0.689 두렵,0.672** 무겁,0.661 낫,0.660 가깝,0.653 더럽,0.649 힘들,0.649 이쁘,0.647 잘생겼,0.645 빡치,0.637 밉,0.636 어리,0.636 그립,0.635 기엽,0.634 춥,0.633 쉽,0.633 걱정되,0.633 사랑스럽,0.623 다르,0.619 멋지,0.617 길어졌,0.616 재미없,0.614 **놀랍,0.613** 귀찮,0.612 졸리,0.605 서글프,0.605 까다롭,0.604 반갑,0.601 이렇,0.600 **흥미롭,0.598** 떠오르,0.596 멀,0.586 **만족스럽,0.585** 컸,0.584 대단하시,0.583 편했,0.581 걱정된,0.581

따라서 본 연구에서는 Word2Vec 기법을 문장과 문단으로 확장한 Dov2Vec과 Skip-Thought 기법을 사용하여 의미적 자질로 사용하였으며, 실험결과는 다음과 같다. <표 23>, <표 24> 은 Dov2Vec로 학습모델을 만들어 Logistic Regression과 SVM 분류기로 실험한 결과이다. SVM 분류

기보다 Logistic Regression 분류기의 성능이 더 높게 나왔다.

표 23 Dov2Vec을 사용한 감정분류(긍정)의 정확도

극성	감정범주	LogisticRegression 정확도(%)	SVM 정확도(%)
긍정	고맙	*91.43*	*90.24*
	반갑	*80.00*	*78.33*
	기쁘	*76.19*	*73.10*
	좋다	*65.95*	*66.19*
	사랑	*77.64*	*76.90*
	만족	*73.10*	*68.33*
	재미	*71.90*	*71.90*
	행복	*78.26*	*76.36*
	즐겁	*82.14*	*79.76*
	편안	*80.71*	*79.05*

표 24 Dov2Vec을 사용한 감정분류(부정)의 정확도

극성	감정범주	LogisticRegression 정확도(%)	SVM 정확도(%)
부정	싫다	*84.29*	*82.14*
	아쉽	*82.21*	*81.50*
	미안	*75.48*	*75.24*
	창피	*72.59*	*69.43*
	속상	*67.14*	*66.43*
	부럽	*83.57*	*82.38*
	무섭	*65.00*	*61.90*
	슬프	*72.62*	*71.90*
	우울	*70.00*	*69.05*
	불안	*67.14*	*61.67*
	실망	*73.33*	*71.43*
	후회	*82.38*	*79.76*
	외롭	*89.05*	*86.43*
	그립	*96.67*	*96.19*
	우습	*97.14*	*96.90*

<표 25> <표 26>는 Skip-thought로 학습모델을 만든 후, 다양한 알고리즘을 적용한 실험한 결과를 보여준다. 다른 알고리즘보다 Logistic

Regression 알고리즘의 분류기 성능이 가장 높게 나왔다.

표 25 Skip-Thought을 사용한 감정분류(긍정)의 정확도

감정 범주	*LogisticRegression*	*SVM*	*GaussianNB*	*MLPClassifier*	*DecisionTreeClassifier*	*RandomForestClassifier*
고맙	***91.43***	*90.71*	*84.29*	*89.05*	*85.24*	*90.00*
반갑	***84.76***	*81.67*	*74.76*	*78.33*	*77.62*	*81.90*
기쁘	***74.76***	*73.81*	*72.86*	*60.95*	*65.00*	*73.81*
좋다	***77.86***	*77.62*	*71.43*	*60.48*	*62.62*	*72.14*
사랑	***84.78***	*83.30*	*81.12*	*73.66*	*70.52*	*83.07*
만족	***80.48***	*80.00*	*71.67*	*70.71*	*68.57*	*78.10*
재미	*74.45*	***75.9***	*70.24*	*66.90*	*64.76*	*72.62*
행복	***82.08***	*80.41*	*70.62*	*75.88*	*68.73*	*78.98*
즐겁	***82.62***	*82.14*	*73.57*	*80.71*	*69.76*	*80.95*
편안	***84.76***	*84.29*	*75.71*	*82.14*	*75.24*	*83.57*

표 26 Skip-Thought을 사용한 감정분류(부정)의 정확도

감정 범주	*LogisticRegression*	*SVM*	*GaussianNB*	*MLPClassifier*	*DecisionTreeClassifier*	*RandomForestClassifier*
싫다	*85.48*	***85.95***	*79.76*	*60.00*	*70.00*	*81.90*
아쉽	***82.22***	*81.98*	*76.46*	*76.69*	*72.89*	*79.56*
미안	***78.10***	*77.62*	*69.29*	*70.95*	*66.67*	*77.38*
창피	***76.28***	*75.31*	*73.32*	*61.35*	*66.00*	*72.86*
속상	***71.43***	*69.76*	*67.38*	*60.48*	*60.24*	*65.95*
부럽	***86.43***	*84.52*	*72.38*	*83.81*	*78.57*	*85.71*
무섭	*70.24*	*68.57*	*69.52*	*56.19*	*61.90*	***70.71***
슬프	*74.05*	*69.76*	*71.67*	*70.24*	*60.71*	***74.52***
우울	***74.52***	*72.86*	*71.90*	*63.10*	*63.81*	*71.43*
불안	***74.52***	*73.57*	*73.57*	*60.71*	*65.24*	*73.81*
실망	***76.67***	***76.67***	*70.95*	*60.95*	*63.57*	*73.33*
후회	***84.05***	*83.81*	*77.14*	*80.00*	*74.05*	*80.48*
외롭	*88.33*	***87.38***	*81.67*	*86.43*	*78.57*	*86.67*
그립	***96.67***	***96.67***	*94.52*	***96.67***	*92.86*	***96.67***
우습	***97.14***	***97.14***	*96.90*	***97.14***	*95.00*	***97.14***

<표 27>는 표층 자질과 의미적 자질로 학습 후, 동일한 SVM 분류기로 실험한 결과이다. 의미적 자질로 문장을 표현한 기법보다 표층 자질로

문장의 특징을 학습한 경우, 성능이 조금씩 더 높게 나왔다.

표 27 감성분류 성능 비교

	감정 범주	표면적	Doc2Vec	Skip-thought
		SVM	Logistic Regression	Logistic Regression
긍정	고맙	90.90	**91.43**	**91.43**
	반갑	84.30	80.00	**84.76**
	기쁘	**77.60**	76.19	74.76
	좋다	73.50	65.95	**77.86**
	사랑	75.40	77.64	**84.78**
	만족	**81.90**	73.10	80.48
	재미	74.40	71.90	**74.45**
	행복	**82.20**	78.26	82.08
	즐겁	**85.70**	82.14	82.62
	편안	**86.10**	80.71	84.76
부정	싫다	**85.70**	84.29	85.48
	아쉽	81.10	82.21	**82.22**
	미안	76.20	75.48	**78.10**
	창피	74.70	72.59	**76.28**
	속상	**73.10**	67.14	71.43
	부럽	**89.70**	83.57	86.43
	무섭	**75.40**	65.00	70.24
	슬프	**74.70**	72.62	74.05
	우울	**75.70**	70.00	74.52
	불안	**75.30**	67.14	74.52
	실망	**84.50**	73.33	76.67
	후회	**84.30**	82.38	84.05
	외롭	88.40	**89.05**	88.33
	그립	**96.80**	96.67	96.67
	우습	**98.00**	97.14	97.14

표면적 자질을 해당 도메인의 특성을 고려하여 설계되었기 때문에 워드 임베딩 기법을 활용한 의미적 자질보다 감정분류 성능이 높게 나타났다.

그러나 의미적 자질은 학습 속도가 매우 빠르며, 학습 데이터의 양에 상관없이 신속하게 문장 내 의미정보를 학습을 할 수 있는 장점이 여전히 존재한다.

제 4 절 자질 조합에 따른 감정분류 실험결과

의미적 자질을 사용하여 학습모델을 구축할 경우, 학습 말뭉치를 구축하는데 드는 비용과 시간을 줄일 수 있지만, 감정분류의 정확성을 높이기에는 한계가 있다. 본 논문에서는 이러한 표층 자질과 의미적 자질을 결합한 학습모델을 구축하였고, 이를 분류기에 사용하였다. <표 27> 이에 대한 실험결과를 보여준다.

<표 28>, <표 29>의 비교를 통해 표층 자질과 워드 임베딩 자질을 같이 사용했을 때 기존 성능보다 나은 성능을 얻을 수 있었다. 이는 기존 벡터에서 표현하지 못했던 정보를 표층 자질 벡터가 보완하여 좋은 성능을 얻었음을 의미한다.

표 28 표층 자질과 Doc2Vec 자질 조합에 따른 실험결과

감정범주	표면적	Doc2Vec	표면적 Doc2Vec
고맙	90.90%	90.24%	91.74%
반갑	84.30%	78.33%	85.33%
기쁘	77.60%	73.10%	80.60%
좋다	73.50%	66.19%	75.19%
사랑	75.40%	76.90%	78.40%
만족	81.90%	68.33%	83.33%
재미	74.40%	71.90%	74.70%
행복	82.20%	76.36%	83.36%
즐겁	85.70%	79.76%	85.71%
편안	86.10%	79.05%	87.05%
싫다	85.70%	82.14%	86.64%
아쉽	81.10%	81.50%	83.00%
미안	76.20%	75.24%	76.74%
창피	74.70%	69.43%	74.93%
속상	73.10%	66.43%	73.93%
부럽	89.70%	82.38%	89.88%
무섭	75.40%	61.90%	77.40%
슬프	74.70%	71.90%	75.40%
우울	75.70%	69.05%	76.55%
불안	75.30%	61.67%	76.67%
실망	84.50%	71.43%	84.93%
후회	84.30%	79.76%	85.26%
외롭	88.40%	86.43%	88.93%
그립	96.80%	96.19%	97.69%
우습	98.00%	96.90%	98.40%

표 29 표층 자질과 Skip-Thought 자질 조합에 따른 실험결과

감정범주	표면적	Skip-Thought	표면적 Skip-thought
고맙	90.90	91.71	91.81
반갑	84.30	85.67	85.77
기쁘	77.60	78.81	79.11
좋다	73.50	77.62	77.92
사랑	75.40	83.30	83.60
만족	81.90	81.00	81.30
재미	74.40	75.95	76.25
행복	82.20	82.41	82.71
즐겁	85.70	86.14	86.44
편안	86.10	87.29	87.59
싫다	85.70	85.95	86.25
아쉽	81.10	81.98	82.28
미안	76.20	77.62	77.92
창피	74.70	75.31	75.61
속상	73.10	69.76	74.06
부럽	89.70	84.52	90.82
무섭	75.40	68.57	77.87
슬프	74.70	69.76	77.06
우울	75.70	72.86	76.16
불안	75.30	73.57	78.00
실망	84.50	76.67	85.20
후회	84.30	84.81	85.11
외롭	88.40	88.38	89.71
그립	96.80	96.67	96.77
우습	98.00	98.14	98.24

제 5 장 결론 및 향후과제

본 논문에서는 텍스트 데이터에 대한 세분화된 감정분류를 위한 연구방법을 제안하였다. 세분화된 감정분류를 위해 25개 감정범주에 대한 분류체계를 마련하였고, 기계학습에 사용되는 감정어 학습 말뭉치를 구축하였으며, 기계학습에 사용되는 표면적 자질과 의미적 자질을 정의하여 감정범주의 정확도를 높이고자 하였다.

감정분류를 위해 문장의 형태론적 정보를 담고 있는 표층적 자질과 단어나 문단의 의미정보를 학습에 사용하기 위한 의미적 자질을 제안하였다. 표층적 자질은 한글의 특성이 반영되고 감정범주를 잘 구분할 수 있는 자질들로 설계하였고, 어절(2-gram), 기호문자(1-gram)의 조합이 감정분류의 정확도를 높이는데, 크게 기여하고 있음을 실험을 통해 확인할 수 있었다. 트위터는 입력 가능한 글자 수가 제한되어 있기 때문에 입력된 문장은 비교적 단순한 구조로 작성되므로 복잡한 문장구조를 표현하는 자질보다는 직관적인 정보를 표현한 자질이 감정분류의 정확도를 높이는 결과를 보여주었다. 기존 감정분류 연구에서 사용되었던 자질들은 학습하는데 많은 시간과 비용이 소요되었다. 또한 이러한 자질들을 설계하기 위해서는 해당 도메인에 대한 지식이 바탕이 되어야 한다. 그러나 이런 표면적 자질은 단어가 내포하고 있는 의미 정보를 표현하는데 한계가 있다.

본 연구에서는 의미적 자질을 추출하기 위한 방법으로 Doc2Vec과 Skip-Thought 기법을 사용하였다. Doc2Vec과 Skip-Thought 기법은 문단에 포함된 단어와 문단에 대한 의미를 의미 벡터 공간에 표현함으로써 표면적 자질로 표현할 수 없는 의미를 표현해준다. 의미적 자질은 자질의 종류에 대한 설계를 따로 할 필요가 없으며, Doc2Vec과 Skip-Thought 기법을 통해 자동으로 학습할 수 있었으며, 학습하는데 걸리는 시간 또한 표면적 자질에 비하면 매우 빠르다. 감정분류의 정확도는 표층적 자질이 감정분류에 더 큰 기여를 하고 있으며, 이 두 개의 자질을 조합하여 사용하

였을 경우 더 높은 정확도를 보인다.

본 논문에서 제안한 한국어 텍스트 감정분류 시스템의 의의는 다음과 같다. 첫째, 본 연구는 한국어로 작성된 텍스트 데이터를 분석하여 세분화된 25개의 감정범주를 분류하였고 높은 정확도를 제공한다. 세분화된 25개의 감정범주는 한국인 실정에 맞는 감정범주로 정의하였고, 다양한 감정범주를 분류하였다.

둘째, 최신 기술인 워드 임베딩 학습 기법을 사용하여 의미적 자질을 설계하고, 이를 한국어 감정분류에 적용하였다. 워드 임베딩 학습 기법은 학습 속도가 매우 빠르며, 표면적 자질을 사용한 학습결과와 비슷한 성능을 제공한다.

셋째, 감정분류를 위해 사용된 표면적 자질과 의미적 자질에 성능평가를 통해 각각의 자질에 대한 성능을 비교 분석하였으며, 이는 감정분류에 있어서 표면적 자질과 의미적 자질에 대한 실증적 실험결과를 제공하고 이를 통해 한국어 감정분류에 있어서 앞으로 연구방향을 설정함에 있어 기초자료로 사용될 수 있다.

넷째, 본 연구에서는 25개의 감정을 분류하기 위한 학습 말뭉치를 전문가를 통해 직접 구축하였고, 감정분석 연구와 실무에 활용이 가능한 자원을 공유한다는 차별성을 가진다.

워드 임베딩 기술을 활용한 기법들은 도메인에 의존적인 연구 분야에서 자율학습의 성능을 높이는데 크기 기여하고 있다. 문서 내 감정분류 뿐만 아니라 대화형 시스템에서 발화자의 감정을 추출하는 연구로 확장이 가능하고, 서로 주고받는 문장 속에서 문맥의 의미를 보다 효율적으로 파악하여 감정을 분석해 낼 수 있을 것이다.

본 연구는 한국어 문장 속에서 25개의 감정범주에 대한 감정분류 프레임워크를 제시하고, 표층 자질과 의미적 자질에 대한 실증적인 데이터를 제공하였다. 실험결과를 통해 감정의 범주가 다양해질수록 한 가지 기법으로 감정을 분류하는데 한계가 있다는 사실을 알게 되었고, 최적의 성능을 위한 자질들의 조합과 분류기들의 조합에 대한 향후 연구가 필요하다.

부록 Ⅰ

Tag	Description	Tag	Description
CC	Coordinating conjunction	SYM	Symbol
CD	Cardinal number	TO	to
DT	Determiner	UH	Interjection
EX	Existential there	VB	Verb, base form
FW	Foreign word	VBD	Verb, past tense
IN	Preposition or subordinatingconjunction	VBG	Verb, gerund or present participle
JJ	Adjective	VBN	Verb, past participle
JJR	Adjective, comparative	VBP	Verb, non-3rd person singular present
JJS	Adjective, superlative	VBZ	Verb, 3rd person singular present
LS	List item marker	WDT	Wh-determiner
MD	Modal	WP	Wh-pronoun
NN	Noun, singular or mass	WP$	Possessive wh-pronoun
NNS	Noun, plural	WRB	Wh-adverb
NNP	Proper noun, singular	$	dollar sign
NNPS	Proper noun, plural	#	pound sign
PDT	Predeterminer	“	left quote
POS	Possessive ending	”	right quote
PRP	Personal pronoun	(	left parenthesis
PRP$	Possessive pronoun	)	right parenthesis
RB	Adverb	,	comma
RBR	Adverb, comparative	.	sentence-final punc
RBS	Adverb, superlative	:	mid-sentence punc :
RP	Particle		

Penn Treebank part-of-speech tags

부록 Ⅱ

1. ‘편안’ 감정과 유사한 어절

ㅁ,0.750 ＊,0.744 The,0.740 ＊ ◕ᴗ◕＊,0.732 짝짝,0.729 。◕,0.729 ㄱ,0.729 ㄷ,0.726 ノ✺,0.726 ≧ㅁ≦,0.724 ノ,0.722 사라짐,0.721 ◕。,0.720 tweet,0.718 Happiness,0.717 흡족,0.717 이징,0.717 바둥바둥,0.716 찡끗,0.715 착잡,0.715 상쾌,0.714 ´；ω；｀,0.713 ✺ㄴ,0.713 쑤담,0.713 Happy,0.713 추기,0.712 느림,0.711 ˙,0.711 날아가기,0.710 Kim,0.710 Of,0.710 ˘ω˘,0.710 ´▽｀,0.709′,0.709 죠타,0.708 51,0.707 아등바등,0.706 창궁,0.705 ﻭ✧,0.704 〃,0.704 기립박수,0.703 ＊´ᴗ,0.703 루넷님,0.703 ＼,0.703 날숨,0.703 Time,0.702 검열,0.701 해파리,0.701 스담스담,0.701 ノ⁺♡,0.701

2. ‘창피’ 감정과 유사한 어절

쭈글,0.873 쓰러짐,0.857 덜그럭,0.852 쭈굴,0.841 스담,0.838 숙연,0.837 죤,0.835 쥬륵,0.835 왈칵,0.833 드러눕,0.833 머슥,0.832 훌찌락,0.831 코슥,0.829 머쓱,0.829 쿨럭,0.828 절레,0.827 쥬금,0.825 멍댕,0.825 터덜,0.821 츄욱,0.819 벽뿌숨,0.818 투닥투닥,0.817 뻔뻔,0.815 옆눈,0.812 입막,0.812 납득,0.809 속닥속닥,0.809 부들,0.808 진지,0.807 서성서성,0.806 촤악,0.805 뿌듯,0.804 나쁨,0.804 아련,0.804 땅침,0.803 흐뭇,0.802 손꼭,0.802 피토,0.802 ´°̩ω°̩｀,0.800 뽀듬,0.800 털석,0.799 소곤,0.799 넙죽,0.797 빠안,0.796 두둥,0.796 우럭,0.795 먼산,0.794 둑흔,0.793 심각,0.793 깨달음,0.792

3. ‘우습’ 감정과 유사한 어절

새롭,0.802 드물,0.763 유쾌하,0.748 자유롭,0.745 평화롭,0.745 갑작스럽,0.744 격하,0.741 차갑,0.738 흥미롭,0.735 빡세,0.733 힘겹,0.732 쉽,0.731 짓궂,0.726 지나치,0.726 따숩,0.725 까맣,0.724 어설프,0.719 재미나,0.717 서툴,0.711 뜨겁,0.711 꿋꿋하,0.709 구렇,0.707 유일하,0.707 여유롭,0.706 뒤늦,0.706 느슨하,0.706 눈부시,0.704 둥글,0.701 뻘하,0.700 끈질기,0.699 느긋하,0.699 흐릿하,0.696 시끄럽,0.691 짖궂,0.688 괴롭,0.687 빡시,0.683 거칠,0.675 서럽,0.659 빠르,0.658 #비투비,0.657 드럽,0.656 즐겁,0.652 서글프,0.652 말갛,0.648 아름답,0.647 너답,0.646 번거롭,0.642 대차,0.642 능숙하,0.640 느껴진,0.639

4. ‘혐오’ 감정과 유사한 어절

차별,0.777 동성애,0.754 인권,0.740 발언,0.729 보수,0.721 주의,0.712 동성애자,0.709 소수자,0.703 신체,0.683 정치,0.681 타인,0.681 상대,0.676 여성,0.675 종교,0.675 인간,0.672 시위,0.667 집단,0.665 행위,0.664 성소수자,0.661 태도,0.660 지지,0.659 정의,0.656 배신,0.651 특정,0.650 정치인,0.647 거부,0.644 판단,0.639 정당,0.638 존재,0.637 개념,0.636 적폐,0.635 부정,0.634 권리,0.631 부모,0.631 논리,0.630 폭력,0.628 진보,0.628 위협,0.628 거짓,0.626 홍준표,0.624 스스로,0.624 대상,0.624 추구,0.623 공격,0.623 인물,0.621 예의,0.618 비난,0.617 변화,0.616 행동,0.614 물질,0.614

5. ‘싫다’ 감정과 유사한 어절

싫고,0.672 피곤하다,0.623 어쩌려,0.619 싫어,0.591 싫어지,0.586 싫으,0.582 미안하다,0.581 도크,0.564 안자,0.562 하려,0.560 중요하다,0.558 괴롭히,0.555 아프,0.550 불쌍하다,0.550 힘들,0.550 자려,0.548 맛있다,0.547 이냐,0.547 밉,0.545 소리치,0.543 박차,0.543 슬프,0.542 필요하다,0.533 잘생기,0.532 풀리,0.530 춥,0.526 무겁,0.524 졸리,0.522 시끄럽,0.522 어렵,0.521 싫,0.519 재밌다,0.518 닉넴보,0.515 부러지,0.515 틀리,0.515 귀찮,0.514 움직이,0.514 기대려,0.513 사려,0.512 늘어지,0.509 웃기,0.509 쓰러지,0.508 아깝,0.508 많다,0.508 사랑스럽,0.508 울리,0.507 떨어지,0.506 산다,0.506 빡치,0.506 놀리,0.504

6. ‘사랑’ 감정과 유사한 어절

애정,0.678 행복,0.666 삶,0.651 고통,0.629 희망,0.626 칭찬,0.620 소원,0.595 자신감,0.591 죽음,0.578 행운,0.569 실망,0.557 욕심,0.556 마음,0.554 의견,0.553 생명,0.550 욕망,0.547 남편,0.539 애인,0.538 축복,0.535 인연,0.535 기적,0.531 힘,0.529 즐거움,0.529 최선,0.528 솔직함,0.525 진실,0.524 미움,0.521 이별,0.520 영광,0.515 외로움,0.513 인생,0.510 운명,0.507 감정,0.506 추억,0.505 영혼,0.505 꽃,0.505 안목,0.496 목숨,0.496 요정,0.494 비타민,0.493 선물,0.492 슬픔,0.492 기운,0.491 실력,0.489 구원,0.488 이모티콘,0.487 불이익,0.486 관심,0.485 작품,0.485 역할,0.485

7. ‘소름’ 감정과 유사한 어절

웃김,0.689 돋는,0.666 돋,0.624 미쳤,0.611 웃겨,0.595 미친,0.592 화난,0.590 돋았,0.583 웃긴,0.573 짜증,0.563 욬,0.560 돋아,0.559 잘생겻,0.557 쩐다,0.554 미쳣,0.553 빡친,0.547 깜짝,0.546 쩔어,0.545 놀랬,0.544 당황,0.538 극혐,0.537 쩔,0.530 핵,0.524 미쳐,0.521 귀여웠,0.518 멋있,0.516 핵귀,0.516 터졌,0.514 놀랐,0.513 웃기,0.513 졸라,0.511 존나,0.510 돋네,0.509 무서웠,0.508 설레,0.505 진쟈,0.505 ;;;;;,0.502 터짐,0.502 조음,0.502 잘생겼,0.499 엍,0.497 기네,0.495 반할,0.492 레알,0.492 ;;;;,0.491 쳣,0.491 현웃,0.490 노잼,0.490 딲,0.489 뭉,0.489

8. ‘불안’ 감정과 유사한 어절

지식,0.711 불신,0.707 슬픔,0.695 혼란,0.692 욕망,0.690 인식,0.685 의식,0.679 권력,0.678 호기심,0.677 감각,0.676 불만,0.673 두려움,0.672 의문,0.669 영역,0.667 습관,0.663 지상,0.662 죽음,0.662 통증,0.661 환경,0.660 신체,0.659 접촉,0.659 중심,0.657 집단,0.656 죄책감,0.655 주름,0.654 눈빛,0.653 이성,0.651 목록,0.651 인력,0.650 개혁,0.647 재능,0.644 시각,0.643 어둠,0.643 석상,0.640 부끄러움,0.639 기술,0.638 야권,0.638 마을,0.637 일말,0.637 이론,0.636 노출,0.636 기색,0.635 직종,0.635 내부,0.635 현실,0.635 범주,0.634 세월,0.633 침묵,0.632 공간,0.632 머릿속,0.629

9. ‘즐겁’ 감정과 유사한 어절

새롭,0.697 쉽,0.678 빠르,0.675 사이좋,0.671 평화롭,0.663 자유롭,0.660 우습,0.652 예쁘,0.640 아름답,0.640 멋지,0.637 힘겹,0.635 느긋하,0.633 여유롭,0.632 시끄럽,0.626 격하,0.622 뻘하,0.622 외롭,0.620 따숩,0.619 유쾌하,0.612 기쁘,0.612 재미나,0.611 반갑,0.598 지나치,0.590 빡세,0.586 번거롭,0.584 까다롭,0.581 뜨겁,0.580 서툴,0.575 다르,0.574 이렇,0.572 이쁘,0.572 이뿌,0.570 괴롭,0.569 흥미롭,0.558 즐거운,0.557 거칠,0.556 낫,0.552 밤늦,0.551 바쁘,0.551 유일하,0.544 부끄럽,0.543 눈부시,0.541 둥글,0.540 뒤늦,0.537 배부르,0.536 고맙,0.536 드물,0.534 부드럽,0.533 건강하시,0.532 안타깝,0.532

10. ‘미안’ 감정과 유사한 어절

미안해,0.731 윽,0.610 하아,0.600 으응,0.594 읏,0.568 자기야,0.567 인마,0.566 아으,0.551 닥쳐,0.549 미워,0.538 그만,0.533 -..,0.529 흐흥,0.527 흐읏,0.521 끅,0.517 속상하게,0.516 그래그래,0.516 켁,0.514 하하,0.504 상현,0.504 흐응,0.501 하읏,0.498 졸려,0.496 푸흐,0.495 응,0.492 창피해,0.486 미안하네,0.481 -"",0.480 어쩌지,0.475 착해,0.474 그러니까,0.470 그랬구나,0.468 아하하,0.466 속상하다,0.464 시끄러워,0.464 후우,0.463 어어,0.461 피곤,0.459 망쳐,0.457 아냐,0.457 아윽,0.457 미안하다,0.454 어엉,0.454 히잉,0.454 흐,0.450 으으,0.450 미안하구,0.449 속았,0.448 젠장,0.446 큭,0.445

11. ‘부럽’ 감정과 유사한 어절

대단하시,0.755 재밌겠,0.743 맛있겠,0.726 아쉽,0.703 안타깝,0.699 부끄럽,0.695 슬프,0.692 무섭,0.689 잘생겼,0.670 그립,0.658 놀랍,0.650 기엽,0.649 고통스럽,0.644 미쳣,0.634 귀엽,0.630 반갑,0.630 기대된,0.629 어렵,0.622 아깝,0.620 사랑스럽,0.611 미쳤,0.606 기대되,0.606 두렵,0.599 만족스럽,0.598 행복하,0.598 오진,0.596 밉,0.593 멋지,0.589 이겼,0.585 웁니,0.585 나댔,0.583 웃기,0.581 잘하시,0.580 아름답,0.577 이쁘,0.574 걱정된,0.572 외롭,0.571 슬푸,0.571 불쌍하,0.570 신난,0.569 잘생겻,0.565 무서웠,0.565 땡기,0.562 망했,0.562 기쁘,0.559 배고프,0.559 고맙,0.555 오졋,0.555 재미없,0.553 힘듭니,0.552

12. ‘좋’ 감정과 유사한 어절

귀찮,0.748 똑같,0.693 거지같,0.688 좁,0.686 괜찮,0.676 조,0.643 나잖,0.530 힘들,0.526 좋아지,0.514 갠찬,0.506 하잖,0.505 맞잖,0.499 즐겁,0.490 중요하잖,0.489 낫,0.488 놓,0.486 놀잖,0.482 무섭,0.480 예쁘,0.478 뛰어놀,0.472 다르,0.467 기쁘,0.465 쩔잖,0.464 같잖,0.458 좋아졌,0.458 귀엽,0.458 나쁘,0.456 위험하잖,0.453 불쌍하잖,0.452 주저앉,0.448 멋지,0.445 가라앉,0.444 잇잖,0.441 오잖,0.438 어렵,0.437 죽잖,0.435 살아남,0.435 슬프,0.434 이쁘,0.429 많잖,0.428 싫잖,0.424 알,0.422 내놓,0.422 아쉽,0.422 행복한,0.421 살,0.420 재밌잖,0.418 이렇,0.418 받잖,0.417 편했,0.416

13. ‘행복’ 감정과 유사한 어절

즐거움,0.692 기쁨,0.666 사랑,0.666 희망,0.654 고통,0.644 행운,0.634 삶,0.625 죽음,0.620 슬픔,0.618 축복,0.603 응원,0.599 애정,0.596 흐름,0.593 운명,0.592 요정,0.588 영광,0.583 소원,0.581 생명,0.575 욕망,0.566 솔직함,0.564 환영,0.556 시련,0.554 자신감,0.554 이별,0.554 비타민,0.554 추억,0.553 진실,0.549 의식,0.548 심신,0.545 하나님,0.544 미학,0.544 발상,0.542 외로움,0.540 최상,0.539 야생,0.538 시대,0.537 천추,0.537 정답,0.536 화분,0.535 명복,0.532 칭찬,0.530 가책,0.530 힘,0.527 지식,0.525 최후,0.524 드드득,0.522 상냥함,0.521 구원,0.521 태양,0.521 독서,0.521

14. ‘고맙’ 감정과 유사한 어절

기쁘,0.776 반갑,0.775 부끄럽,0.737 기뻤,0.680 안타깝,0.674 사랑스럽,0.655 아름답,0.654 받겠,0.654 감사했,0.636 만족스럽,0.635 놀랍,0.632 고마웠,0.631 멋지,0.630 사랑한,0.629 귀엽,0.623 두렵,0.615 행복했,0.614 예쁘,0.610 잊었,0.608 아깝,0.608 괴롭,0.606 재밌겠,0.605 고마워,0.600 슬프,0.599 잘생겼,0.597 찾아왔,0.596 빌겠,0.588 시끄럽,0.585 흥미롭,0.583 이렇,0.583 즐거웠,0.582 길어졌,0.580 잘했,0.572 멋있다,0.569 외롭,0.565 실례하겠,0.565 성장했,0.564 오겠,0.563 춥,0.560 해보겠,0.559 수고했,0.559 맛있다,0.558 늦었,0.556 부럽,0.555 이겼,0.554 힘내겠,0.553 밉,0.551 실망했,0.549 좋아하겠,0.548 자랑스럽,0.545

15. ‘걱정’ 감정과 유사한 어절

부담,0.613 다행,0.581 걱정하지,0.568 안심,0.553 고민,0.517 욕심,0.509 대답,0.496 거짓말,0.492 미련,0.490 실망,0.485 큰일,0.475 노력,0.473 건강,0.470 농담,0.468 말,0.458 무리하지,0.458 악몽,0.454 금방,0.447 오랜만,0.442 상관,0.434 실망하지,0.434 -.,0.426 상상,0.424 긴장하지,0.422 고민하지,0.420 약속,0.418 절대,0.418 후회,0.416 죽지,0.411 신경,0.411 그러니까,0.410 도움,0.409 생각,0.409 아프,0.407 몸,0.406 생각하지,0.406 유감,0.401 마음,0.400 괜찮,0.399 슬퍼하지,0.393 동정,0.392 기운,0.392 적당히,0.389 버릇,0.389 위안,0.386 믿기,0.386 어른,0.386 그만,0.386 대하지,0.384 거절,0.384

16. ‘아쉽’ 감정과 유사한 어절

안타깝,0.796 슬프,0.778 부끄럽,0.770 어렵,0.738 놀랍,0.731 무섭,0.705 부럽,0.703 아깝,0.690 외롭,0.667 두렵,0.663 슬펐,0.654 늦었,0.652 만족스럽,0.636 괴롭,0.625 재밌겠,0.622 이렇,0.608 기쁘,0.608 가깝,0.598 그렇,0.598 낫,0.593 반갑,0.592 흥미롭,0.586 졸리,0.584 이겼,0.579 아름답,0.577 이겠,0.577 길어졌,0.574 춥,0.573 바쁘,0.570 힘들,0.568 까다롭,0.566 고통스럽,0.565 무서웠,0.565 스럽,0.564 그립,0.559 아쉬웠,0.557 배고프,0.554 맛있겠,0.552 힘들었,0.549 대단하시,0.542 시끄럽,0.541 서글프,0.540 맞았,0.539 멀,0.536 덥,0.535 불편하겠,0.535 받겠,0.533 찾았,0.531 재미없,0.528 실례했,0.527

17. ‘외롭’ 감정과 유사한 어절

괴롭,0.803 아깝,0.785 슬프,0.757 무섭,0.725 부끄럽,0.724
안타깝,0.723 춥,0.720 시끄럽,0.713 어렵,0.711
아름답,0.697 이렇,0.677 평화롭,0.670 낫,0.667
아쉽,0.667 힘들,0.666 무겁,0.663 아프,0.663 떠오르,0.661
배고프,0.661 쉽,0.659 서글프,0.648 흥미롭,0.646
두렵,0.643 거슬리,0.642 덥,0.636 서툴,0.635 놀랍,0.634
대수롭,0.633 가깝,0.633 맞겠,0.628 많겠,0.628
졸리,0.625 번거롭,0.624 밉,0.623 우습,0.623 어지럽,0.622
걱정되,0.621 즐겁,0.620 걱정된,0.617 기쁘,0.614 뜨겁,0.613
많아졌,0.612 바쁘,0.612 느껴지,0.612 차갑,0.611 나쁘,0.609
뻘하,0.605 다치,0.604 더럽,0.602 길어졌,0.600

18. ‘그립’ 감정과 유사한 어절

안타깝,0.722 아름답,0.717 아깝,0.715 스럽,0.709
길어졌,0.672 대단하시,0.662 두렵,0.660 부럽,0.658
만족스럽,0.654 부끄럽,0.647 놀랍,0.646 어렵,0.640
무섭,0.635 편했,0.634 커졌,0.633 이겠,0.626 슬프,0.626
땡기,0.624 사랑스럽,0.619 아름다웠,0.616 행복했,0.616
가깝,0.610 보도했,0.608 잘생겼,0.607 멋지,0.602
떨리,0.601 그려졌,0.600 예뻤,0.594 힘들었,0.591
식었,0.590 싫었,0.590 스러웠,0.588 이겼,0.585 괴롭,0.585
실례했,0.584 넘치,0.584 떠오르,0.583 밉,0.583
즐거웠,0.582 재밌었,0.576 컸,0.575 역겹,0.574 재밌겠,0.573
느껴지,0.572 열렸,0.569 나가셨,0.569 반갑,0.568
웃겼,0.563 귀여웠,0.562 찾았,0.561

19. ‘우울’ 감정과 유사한 어절

사망,0.775 왈칵,0.767 우럭,0.758 피토,0.757 쓰러짐,0.754 쥬금,0.750 감격,0.749 주르륵,0.749 숙연,0.746 쭈글,0.746 쥬륵,0.738 절레,0.735 ╥ﾌ╥,0.734 쿨럭,0.733 허름,0.732 륵,0.729 미침,0.721 쥴쥴,0.719 드르렁,0.719 덜그럭,0.715 터덜,0.711 먼산,0.710 훌찌락,0.709 ⌒,0.709 둑흔,0.709 아련,0.708 광광,0.707 창피,0.707 주륵,0.707 쭈굴,0.705 옆눈,0.705 시름시름,0.702 푸닥,0.701 주금,0.700 입틀막,0.697 소곤,0.697 드러눕,0.696 땅침,0.693 죤,0.690 츄욱,0.688 오열,0.686 주륵주륵,0.685 찔림,0.683 입막,0.683 와장창,0.682 또르르,0.682 속닥,0.680 울망,0.680 심각,0.680 :?),0.680

20. ‘만족’ 감정과 유사한 어절

잡덕,0.596 상상,0.542 완성,0.538 타입,0.529 경험,0.527 해석,0.523 조건,0.523 의문,0.519 설명,0.519 표현,0.518 행복,0.512 비용,0.511 영광,0.510 성적,0.510 성향,0.510 유감,0.509 수준,0.508 실망,0.505 유품,0.505 정답,0.505 이론,0.504 연상,0.503 자연,0.502 시작,0.500 즐거움,0.500 넘사벽,0.499 욕망,0.499 공감,0.499 행운,0.498 목적,0.498 농담,0.498 감성,0.497 시장님,0.496 정의당,0.496 특권,0.495 실감,0.494 조합,0.493 폭력,0.493 뜻,0.493 해피엔딩,0.493 방식,0.492 설정,0.492 인식,0.492 발음,0.490 비판,0.489 존못,0.489 빡침,0.489 착각,0.488 자신감,0.487 판단,0.487

21. ‘반갑’ 감정과 유사한 어절

고맙,0.775 부끄럽,0.723 기대되,0.710 기쁘,0.706 만족스럽,0.704 뵙겠,0.699 안타깝,0.698 힘내겠,0.690 아름답,0.684 기뻤,0.681 즐거웠,0.677 기대하겠,0.673 감사했,0.671 빌겠,0.671 재밌겠,0.668 멋지,0.663 받겠,0.663 다가가겠,0.658 수고하셨,0.657 실례했,0.648 놀랍,0.646 찾아왔,0.645 행복했,0.643 고생하셨,0.642 이렇,0.637 참여했,0.636 감사하겠,0.631 부럽,0.630 가겠,0.629 이겼,0.628 흥미롭,0.627 늦었,0.619 오겠,0.619 찾았,0.615 찾아가겠,0.613 잘생겼,0.612 사랑스럽,0.610 해주셨,0.608 다녀오겠,0.608 멋있,0.607 보도했,0.606 가보겠,0.606 두렵,0.604 말씀하셨,0.604 무섭,0.601 잘했,0.601 즐겁,0.598 실례하겠,0.598 노력하겠,0.597 해보겠,0.596

22. ‘짜증’ 감정과 유사한 어절

화가,0.619 난리,0.613 큰일,0.608 고장,0.570 소름,0.563 존나,0.546 눈물,0.532 웃기,0.522 박살,0.502 지랄,0.500 냄새,0.497 벗어,0.488 울었,0.486 실감,0.480 슬프,0.478 허탕,0.474 솔직히,0.472 졸라,0.466 죽었,0.464 슬퍼,0.462 슬펐,0.461 돌았,0.456 빡쳐,0.455 싫다,0.455 혼자,0.452 걍,0.452 무섭,0.451 시발,0.448 안달,0.447 쫓,0.446 좆,0.446 나댔,0.441 똥,0.437 무서워,0.435 멘탈,0.435 극혐,0.435 씨발,0.434 눈치,0.433 미쳐,0.432 미쳤,0.429 미친,0.428 웃겨,0.427 착했,0.424 기억,0.423 회복되,0.422 코피,0.420 쳐,0.416 피곤했,0.415 무서웠,0.414 쳤,0.412

23. ‘재미’ 감정과 유사한 어절

문제,0.601 어이,0.598 갈수,0.567 살수,0.554 지구,0.549 쓸모,0.548 효과,0.535 여유,0.530 멋,0.529 필요,0.524 확률,0.523 공포,0.517 애니,0.514 탱커,0.513 날수,0.512 이유,0.510 의미,0.509 장르,0.505 볼품,0.495 흥미,0.492 수,0.491 취미,0.491 관계,0.490 스토리,0.488 일본어,0.487 인기,0.485 후회,0.481 경우,0.477 대사,0.476 분위기,0.475 남캐,0.471 뜬금,0.461 영어,0.460 여러 가지,0.460 죄,0.457 윅스,0.448 본적,0.447 상관,0.446 수위,0.445 실제,0.442 목표,0.441 재수,0.440 마이너,0.439 요소,0.438 컴퓨터,0.435 자세,0.434 소용,0.434 사투리,0.433 눈치,0.432 힐러,0.432

24. ‘속상하’ 감정과 유사한 어절

커엽,0.552 깊다,0.548 창피하,0.547 뿌듯하,0.546 좋아하더,0.546 애쓰,0.542 반하면,0.539 됀,0.539 보태,0.538 숨도,0.536 불안해서,0.533 들어오,0.527 답하는,0.524 팅겨,0.524 웃프,0.522 예브,0.520 갚는,0.519 부지런해,0.518 털려,0.516 알아보시,0.516 찢어지,0.514 생각되는,0.514 죄송스럽,0.512 힘들어하는,0.511 쿠키런,0.510 놀랏,0.510 울려,0.509 어지럽,0.509 시끄럽,0.508 사가겠,0.508 나타나셔,0.507 우후,0.507 어지러,0.505 기어이,0.504 부담스럽,0.504 슬푸,0.503 고통스럽,0.503 교자,0.502 내겠,0.502 자랐,0.502 친해진,0.501 즉사,0.501 이선,0.501 삼즈,0.501 빡,0.500 넘어오시,0.500 넣겠,0.499 취저,0.499 매웠,0.498 피곤하다,0.497

25. ‘불쌍’ 감정과 유사한 어절

진쟈,0.734 웃김,0.727 최고다,0.690 조타,0.687 낱,0.685 무서움,0.678 허잉,0.677 킄,0.675 불상,0.672 흑흐,0.672 웅앵웅,0.668 기엽네,0.665 진쨔,0.665 증말,0.664 행벅,0.662 창피하다,0.659 어흑,0.656 흐구,0.655 허이잉,0.653 킄,0.653 스발,0.653 스바,0.651 찌통,0.650 시카,0.648 쿠쿠,0.647 쩔어요,0.646 으어,0.645 이쁨,0.644 헝,0.644 기엽구,0.643 퓨,0.642 쿠,0.639 허윽,0.638 튜,0.636 히잉,0.635 실성,0.635 앜,0.635 이잉,0.634 호오옥,0.633 젠장,0.633 진자,0.632 비싸,0.631 졸귀,0.631 귀염,0.629 답답해,0.629 지쨔,0.629 zzz,0.629 서러움,0.628 흐윽,0.628 시바,0.627

26. ‘후회’ 감정과 유사한 어절

미련,0.602 문제,0.592 필요,0.574 거침,0.559 고민,0.557 의지,0.550 이유,0.549 변함,0.541 부담,0.541 경우,0.534 군말,0.528 쓸모,0.528 재수,0.528 기회,0.526 이야기,0.526 얼마,0.525 죄,0.522 믿음,0.519 상관,0.518 상상,0.518 선택,0.510 공부,0.507 일이,0.505 가능성,0.504 확률,0.504 단어,0.503 증거,0.499 가차,0.499 지지,0.498 존재,0.498 소용,0.497 생각,0.496 일도,0.496 볼품,0.491 용기,0.489 의미,0.488 대화,0.487 이해,0.482 기대,0.482 메리트,0.481 재미,0.481 방해,0.481 합의,0.478 오해,0.478 안중,0.474 시간,0.473 공포,0.472 적도,0.470 준비,0.466 권리,0.466

27. ‘슬프’ 감정과 유사한 어절

부끄럽,0.803 어렵,0.783 아깝,0.780 아쉽,0.778 무섭,0.774 안타깝,0.757 외롭,0.757 힘들,0.732 괴롭,0.731 기쁘,0.727 놀랍,0.715 시끄럽,0.707 웃기,0.704 아름답,0.699 스럽,0.698 춥,0.697 졸리,0.693 부럽,0.692 사랑스럽,0.690 귀엽,0.667 바쁘,0.665 슬펐,0.651 두렵,0.648 기엽,0.646 걱정되,0.645 멋지,0.645 아프,0.640 이쁘,0.640 무겁,0.638 그립,0.626 잘생겼,0.626 배고프,0.625 밉,0.619 이렇,0.616 서글프,0.616 가깝,0.611 예쁘,0.607 슬퍼,0.605 기대되,0.603 낫,0.602 고맙,0.599 떠오르,0.597 길어졌,0.593 고통스럽,0.593 쉽,0.590 덥,0.589 느껴지,0.587 늦었,0.585 까다롭,0.576 걱정된,0.571

28. ‘실망’ 감정과 유사한 어절

영광,0.720 유감,0.719 동정,0.678 착각,0.671 농담,0.663 의문,0.654 정답,0.651 기적,0.644 죽음,0.637 명언,0.631 죄인,0.629 최악,0.624 솔직함,0.623 이론,0.617 진실,0.615 어른,0.610 행운,0.607 조건,0.605 오랜만,0.601 감동,0.601 다행,0.599 운명,0.598 반칙,0.597 욕심,0.596 기쁨,0.595 안심,0.590 마련,0.590 존못,0.589 결말,0.588 닝겐,0.587 따름,0.585 비밀,0.582 믿음,0.579 뜻,0.577 동감,0.576 축복,0.576 넘사벽,0.575 죄책감,0.574 희망,0.571 지름길,0.570 선택,0.569 결론,0.568 위안,0.567 뒷북,0.567 사냥꾼,0.566 편견,0.565 대목,0.564 고통,0.56 동문,0.563 자랑,0.563

참고문헌

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

[2] Kim, S. Y., Park, S. T., & Kim, Y. K. (2015). Samsung-Apple patent war case analysis: focus on the strategy to deal with patent litigation. Journal of digital convergence, 13(3), 117-125.

[3] Choi, S., & Choi, K. (2015). Achievement and satisfaction research of the undergraduate orchestra club activities-A convergent aspects of statistical method and opinion mining. Journal of the Korea Convergence Society, 6(4), 25-31.

[4] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of computational science, 2(1), 1-8.

[5] DiGrazia, J., McKelvey, K., Bollen, J., & Rojas, F. (2013). More tweets, more votes: Social media as a quantitative indicator of political behavior. PloS one, 8(11), e79449.

[6] Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In LREc (Vol. 10, No. 2010).

[7] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of twitter data. In Proceedings of the workshop on languages in social media (pp. 30-38). Association for Computational Linguistics.

[8] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(2009), 12.

[9] Kim, S. M., & Hovy, E. (2004). Determining the sentiment of opinions. In Proceedings of the 20th international conference on Computational Linguistics (p. 1367). Association for Computational Linguistics.

[10] Yadollahi, A., Shahraki, A. G., & Zaiane, O. R. (2017). Current State of Text Sentiment Analysis from Opinion to Emotion Mining. ACM Computing Surveys (CSUR), 50(2), 25.

[11] Kim, M., & Park, S. O. (2013). Trust management on user behavioral patterns for a mobile cloud computing. Cluster computing, 16(4), 725-731.

[12] 김유신, 김남규, 정승렬. (2012). 뉴스와 주가: 빅데이터 감성분석을 통한 지능형 투자의사결정모형. 지능정보연구, 18(2), 143-156.

[13] 서민송, & 유환희. (2017). 재난 관련 SNS 데이터를 이용한 감성도 분석. 한국지형공간정보학회 학술대회, 3-6.

[14] Alm, C. O., & Sproat, R. (2005, October). Emotional sequencing and development in fairy tales. In International Conference on Affective Computing and Intelligent Interaction (pp. 668-674). Springer, Berlin, Heidelberg.

[15] Plutchik, R., & Kellerman, H. (1980). Emotion, Theory, Research, and Experience: Theory, Research and Experience. Academic press.

[16] Plutchik, R. (2003). Emotions and life. Washington, DC: American Psychological Association.

[17] 이철성, 최동희, 김성순, & 강재우. (2013). 한글 마이크로블로그 텍스트의 감정 분류 및 분석. 정보과학회논문지: 데이타베이스, 40(3), 159-167.

[18] Wang, W., Li, Y., Huang, Y., Liu, H., & Zhang, T. (2017). A Method for Identifying the Mood States of Social Network Users Based on Cyber Psychometrics. Future Internet, 9(2), 22.

[19] Kumar, A., & Joshi, A. (2017, March). Ontology Driven Sentiment Analysis on Social Web for Government Intelligence. In Proceedings of the Special Collection on eGovernment Innovations in India (pp. 134-139). ACM.

[20] Agarwal, B., & Mittal, N. (2016). Sentiment analysis using conceptnet ontology and context information. In Prominent Feature Extraction for Sentiment Analysis (pp. 63-75). Springer International Publishing.

[21] Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In LREC (Vol. 10, pp. 2200-2204).

[22] Shelke, N., Deshpande, S., & Thakare, V. (2017). Domain independent approach for aspect oriented sentiment analysis for product reviews. In Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications (pp. 651-659). Springer, Singapore.

[23] Manek, A. S., Shenoy, P. D., Mohan, M. C., & Venugopal, K. R. (2017). Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. World wide web, 20(2), 135-154.

[24] Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In LREc (Vol. 10, No. 2010).

[25] Mohammad, S. M., Kiritchenko, S., & Zhu, X. (2013). NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. arXiv preprint arXiv:1308.6242.

[26] Park, I. J., & Min, J. K. (2005). Making a List of Korean Emotion Terms and Exploring Dimensions Underlying Them, Korean journal of social and personality psychology 19(1), 109-129.

[27] Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on Machine Learning (ICML-14) (pp. 1188-1196).

[28] Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip-thought vectors. In Advances in neural information processing systems (pp. 3294-3302).

[29] 안정국, & 김희웅. (2015). 집단지성을 이용한 한글 감성어 사전 구축, 한국경영정보학회 학술대회, 527-532, 2015.

[30] 김윤석, & 서영훈. (2013). 기계 학습을 이용한 한글 텍스트 감정 분류, 한국엔터테인먼트산업학회 학술대회 논문집, 206-210.

[31] Ekman, P., & Friesen, WV (1971). Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17(2), 124-129.

[32] 김명규, 김정호, 차명훈, & 채수환. (2009). 텍스트 문서 기반의 감성 인식 시스템. 감성과학, 12(4), 433-442.

[33] Park, I. J., & Min, J. K. (2005). Making a List of Korean Emotion Terms and Exploring Dimensions Underlying Them, Korean journal of social and personality psychology 19(1), 109-129.

[34] 이동엽, 조재춘, & 임희석. (2017). 워드 임베딩을 이용한 아마존 패션 상품 리뷰의 사용자 감성 분석. 한국융합학회논문지, 8(4), 1-8.

[35] Lee, G. H., & Lee, K. J. (2013). Twitter Sentiment Analysis for the Recent Trend Extracted from the Newspaper Article. Korea Information Processing Society, 2(10), 731-738.

[36] 김유영, & 송민. (2016). 영화 리뷰 감성분석을 위한텍스트 마이닝 기반 감성 분류기 구축. 지능정보연구, 22(3), 71-89.

[37] 류진걸, & 신동민. (2014). SVM과 HCRF를 이용한 텍스트 문서 감정 분류 모델," 대한산업공학회 추계학술대회 논문집, 897-903.

[38] Kiritchenko, S., Zhu, X., Mohammad, S. M.. (2014). Sentiment analysis of short informal texts, Journal of Artif. Intelligence Research, 723-762.

[39] Lee, Cheolseong, et al. (2013). Classification and analysis of emotion in korean microblog texts, Journal of Korean Institute of Information Scientists and Engineers: Databases 40(3), 159-167.

[40] 홍소라, 정연오, & 이지형. (2014). 대용량 소셜 미디어 감성분석을 위한 반감독 학습 기법. Journal of Korean Institute of Intelligent Systems, 24(5), 482-488.

[41] Alec Go, Richa Bhayani, & Lei Huang (2009). Twitter sentiment classification using distant supervision, CS224N project report, Stanford.

[42] Cheng, K., Li, J., Tang, J., & Liu, H. (2017). Unsupervised Sentiment Analysis with Signed Social Networks. In AAAI, 3429-3435.

[43] 홍태호, 김은미, & 차은정. (2017). 뉴스 감성분석과 SVM을 이용한 다우존스 지수와 S&P500 지수 예측. 인터넷전자상거래연구, 17(1), 23-36.

[44] Schouten, K., van der Weijde, O., Frasincar, F., & Dekker, R. (2017). Supervised and Unsupervised Aspect Category Detection for Sentiment Analysis With Co-Occurrence Data. IEEE Transactions on Cybernetics.

[45] Hu, X., Tang, J., Gao, H., & Liu, H. (2013, May). Unsupervised sentiment analysis with emotional signals. In Proceedings of the 22nd international conference on World Wide Web (pp. 607-618). ACM.

[46] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[47] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[48] Hofmann, T. (1999, July). Probabilistic latent semantic analysis. In Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence (pp. 289-296). Morgan Kaufmann Publishers Inc..

[49] Lewis, D. D. (1992, February). Feature selection and feature extraction for text categorization. In Proceedings of the workshop on Speech and Natural Language (pp. 212-217). Association for Computational Linguistics.

[50] 박자람, & 차미영. (2012). 온라인 소셜미디어에서의 이모티콘 사용, 한국 HCI 학회 학술대회.

[51] DongHyun Choi, Jungyeul Park, & Key-Sun Choi., (2012). Korean Treebank Transformation for Parser Training, ACL – SPMRL.

[52] Park, K. M., & Lim, H., S.(2014). Acquiring lexical knowledge using raw corpora and unsupervised clustering method," Cluster computing, 17(3), 901-910.

[53] Slav Petrov & Dan Klein. (2007). Improved inference for unlexicalized parsing, In Proceedings of HLT NAACL.

[54] Goldberg, & Lewis R. (1990). An Alternative "deScription of perSonality": The Big-Five FactOr Structure, Journal of personality and social psychology 59(6), 1216.

[55] 이동주, 연종홈, 황인범 & 이상구. (2010) 꼬꼬마: 관계형 데이터베이스를 활용한 세종 말뭉치 활용 도구. 정보과학회논문지: 컴퓨팅의 실제 및 레터, 16(11), 1046-1050.

[56] N. Chawla, K. Bowyer, L. Hall, & WP Kegelmeyer, (2002). "SMOTE: synthetic minority oversampling technique," Journal of artificial intelligence research 16, 321-357.

[57] 홍승현. (2003). 유전자 알고리즘을 활용한 인공신경망 모형최적입력변수의 선정: 부도예측 모형을 중심으로. 한국지능정보시스템학회논문지, 9(1), 225-249.

[58] Turney, P.D. & Littman, M.L. (2003). Measuring Praise and Criticism : Inference of Semantic Orientation from Association, ACM Transactions on Information Systems(TOIS), 21(4), 315-346.

[59] A. Esuli & F. Sebastiani, (2005). Determining the Semantic Orientation of Terms through Gloss Classification, In Proceedings of the CIKM, pp. 617-624.

[60] 박자람 & 차미영. (2013). 온라인 소셜미디어에서의 이모티콘 사용, 한국HCI학회 학술대회, 537-539.

[61] Harris, Z. S. (1954). Distributional structure. Word, 10(2-3), 146-162.

[62] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

[63] Yamada, H., & Matsumoto, Y. (2003, April). Statistical dependency analysis with support vector machines. In Proceedings of IWPT (Vol. 3, pp. 195-206).

[64] Nivre, J., Hall, J., Nilsson, J., Eryiğit, G., & Marinov, S. (2006, June). Labeled pseudo-projective dependency parsing with support vector machines. In Proceedings of the Tenth Conference on Computational Natural Language Learning (pp. 221-225). Association for Computational Linguistics.

Abstract

In the opinion mining field, the emotional analysis determines whether the contents contained in the text are subjective or objective and analyzes the polarity of the emotions with respect to the contents of which the subject of the author is revealed, and analyzes it as one of positive, negative and neutral. In the initial emotional analysis research, if most of the opinions on the website and the social media service were automatically analyzed and the analysis result of 'positive / negative' or 'good / no' was provided, , Emotional analysis that recognizes various emotions of consumers such as sadness, expectation, and fear are being tried. It is difficult to extract information other than affirmative in sentence analysis of simple positive and negative sentiment analysis. However, in the more detailed emotion analysis, it is possible to extract higher-order emotions such as 'comfort' and 'cute' in the above sentence, which can contribute to enhancement of product quality by extracting more accurate information and analyzing such opinions. The purpose of this study is to classify emotions subdivided from text data, not polarity categories such as 'positive / negative' or 'good / no'.

In this study, we define emotional analysis as positive / negative polarity classification and define emotional analysis to classify various subdivided emotions. In the Korean - based text data, the surface feature set and the semantic feature set And to improve the accuracy of emotion classification.

Classification systems that recognize various emotions such as simple affirmation, not joy, pleasure, comfort, sadness, and anxiety are often of different kinds depending on the application system applied. Especially in Korea, polarity classification studies on emotions were mostly conducted, and the researches classified into various emotions are still insufficient. A representative classification system for overseas emotions is Plutchik's eight

classification systems (joy, trust, fear, surprise, sadness, aversion, anger, expectation). Plucksch distinguished eight basic human emotions, and most of the emotions we experience in general suggest that these eight emotions are mixed. The six classification systems (tension, depression, anger, vigor, fatigue, confusion) of the profiler of mood states (POMS) are also frequently used emotional categories. However, since such an emotion classification system is written in English, there is a limit to be translated into Korean in accordance with its meaning. In order to solve these problems, this study selected 25 Korean emotional categories through Korean emotional vocabulary list for Korean emotional classification. Based on the familiarity evaluation criteria of 434 Korean emotional vocabulary lists corresponding to Korean people 's situation, the top 25 emotional words were selected and used to construct a learning corpus.

The approach for classifying emotions is based on machine learning techniques, which are frequently used in recent studies. Sufficient learning corpus is needed to perform instructional learning on 25 subdivided emotions. However, the absence of a learning corpus for the Korean emotion category makes it difficult to classify the emotion classification, and the disclosed Korean emotion corpus contains only polarity information about affirmation and negation. Therefore, in this study, we constructed a learning corpus for 25 emotion categories for various emotion categories, and designed the surface qualities and semantic qualities. Surface features provide an immediate, one-dimensional basis for emotion classification in textual data. However, there is a limit to the meaning of text data. In previous studies, various qualities were combined or new qualities were developed to express these meanings. However, in the field of natural language processing, techniques for embedding words into 'semantic' vector space have been introduced, and in the field of emotional classification, approaches are being

tried to automatically learn the qualities using these techniques. 'Meaning' refers not to words alone but to all processing units that occur in natural language processing such as phrase, sentence, and document. Embedding a word or sentence in a large amount of data into a semantic vector space creates a learning model using context information of surrounding words and sentences and classifies emotions through machine learning algorithm. Therefore, in this study, semantic qualities such as doc2vec and skip-thought vector using word embedding technique as well as surface qualities were used and the performance with surface feature was compared and analyzed.

Another problem with emotion classification is that it causes data shortage due to excessive use of lexical qualities used in learning.

In this study, because of the use of 25 subdivided emotional categories, there is a lack of data for specific emotional categories. As the number of emotional categories increases, it will be increasingly difficult to collect data on emotions. Therefore, in this study, we tried to solve the problem of lack of learning data by solving this problem through data oversampling technique.

This study is meaningful in itself by classifying emotions into 25 different emotion categories in Korean sentences, but it is more important for the following reasons.

First, this study classifies 25 classified emotion categories by analyzing text data written in Korean and provides high accuracy. Twenty - five sub - categories of emotion were defined as emotional categories appropriate for Korean people, and various emotional categories were classified.

Second, semantic qualities are designed using the latest word - embedding learning techniques and applied to Korean emotion

classification. Word-embedded learning techniques are very fast to learn and provide similar performance to those using surface features.

Third, the performance of each qualities is compared and analyzed through the performance evaluation on the surface and semantic qualities used for emotion classification. This provides empirical test results on surface qualities and semantic qualities in emotion classification, Can be used as a basis for setting future research directions in Korean emotion classification.

Fourth, in this study, a learning corpus to classify 25 emotions is constructed directly by experts, and the emotional analysis research and the sharing of resources that can be utilized for practical purposes are differentiated.