오피니언 마이닝을 통한 뉴스 댓글의 감정 분석

Sentiment analysis of article comments using opinion mining

조형관, 김지훈, 하수민
(Hyoung Kwan Cho, Ji Hun Kim, Sue Min Ha)

지도교수 : 김유성

인하대 정보통신공학부

요 약: 최근 인터넷상으로 사람들의 의견 교환이 활발해지면서 이를 수집하고 분석하여 유용한 정보를 추출하는 오피니언 마이닝에 대한 관심이 높아지고 있다. 본 논문에서 그 중 뉴스 댓글의 감정을 분석하는 방법을 제안한다. 한글 텍스트를 분석할 때 어려운 점은 구축되어있는 한글 감정 사전이 없다는 점이다. 이를 위해 본 연구에서는 필요한 감정 분류를 위해 감정 사전을 구축하고, 지속적으로 단어가 추가될 수 있도록 서베이를 만들었다. 웹 크롤링을 통해 수집한 뉴스 댓글과 구축된 감정 사전을 비교하여 감정 분석을 진행하였다. 이 때 Levenshtein 거리를 통해 BoW(Bag of Word)를 생성하고 Bayesian 분류기를 사용하여 오차를 최소화 하였다. 분석 결과를 확인하기 위해 본교 재학생들을 대상으로 설문을 진행한 결과 분노 감정에 대해 93.5%의 신뢰도를 확인하였다. 분석 결과를 사용자들이 실시간으로 확인할 수 있는 User Interface(UI)를 디자인하였다.

주제어: 오피니언 마이닝, 댓글 분석, 감정 사전, Levenshtein 거리, Bayesian 분류기, UI

Abstract: Recently, sharing information online became common. Opinion mining, which is to gather information from it and to analyze it to become useful data, is getting people's attention. In this paper, we propose the emotion prediction method, which analyzes comments of online news articles. When analyzing Korean texts, the most difficult thing is that there is no good Korean sentiment dictionary that fits the needs. For this, we made a sentiment dictionary to classify the feelings into 7 levels and a survey was made to add more words to the dictionary. We compare the crawled news comments and sentiment dictionary to analyze the sentiments. The process has been made using Levenshtein distance and Bayesian classifier to reduce errors. We asked students to analyze the sentiments and as a result, got 93.5% of credibility. Also, we designed a User Interface(UI) so people could see the result of analysis in real time.

Keywords: Opinion mining, comment analysis, sentiment dictionary, Levenshtein distance, Bayesian classifier, UI

Ⅰ. 서론

최근 인터넷의 접근성이 증가하면서 다양한 웹사이트 혹은 소셜네트워크서비스를 통한 정보공유가 활발해졌다. 따라서 이러한 매체들을 통해 생산되는 방대한 양의 데이터로부터 의미 있는 정보를 추출해내는 기술의 필요성 또한 대두되고 있다. 그 중에서도 오피니언 마이닝은 사람들의 여론과 의견을 분석해서 유용한 정보로 재가공하는 기법으로, 이를 활용하여 텍스트에 포함된 내용을 분석할 수 있다. 오피니언 마이닝은 개인 연구뿐만 아니라 정부, 기업 측면에서도 대중의 의견을 분석하기 위해 사용되어 활용도가 매우 높다.

소셜 네트워크 서비스를 통해 공유되는 의견들은 신조어, 약어 또는 광고 등이 많이 포함되어 감정 분류 데이터로 적합하지 않다고 판단하여 본 논문에서는 네이버에서 제공하는 인터넷 뉴스기사의 댓글을 감정 분류 데이터로 선정하였다. 여러 분야의 뉴스를 제공하기 때문에 이에 대해 나타날 수 있는 사람의 감정이 다양할 것이라 생각된다. 따라서 본 연구에서는 오피니언 마이닝을 통하여 뉴스기사 댓글에 대한 감정분석을 진행하고 이를 기반으로 세부 감정 그래프를 가진 새로운 뉴스기사 플랫폼을 설계한다.

특히 이 연구에서는 감정분석을 진행할 때 기존 연구들에서 분석한 것처럼 긍정, 부정이 아닌 총 7가지의 감정으로 데이터를 분석한다. 7가지 감정은 뉴스 기사를 보고 느낄 수 있는 감정을 추출하여 결정되었으며 분노, 놀람(부정), 슬픔, 무관심, 흥미, 놀람(긍정), 행복이 이에 해당한다. 자세한 감정 분류를 위해 감정 사전을 구축하고 사전에 등재되지 않은 단어를 처리하기 위해서 survey 시스템을 설계하여 실시간으로 변화하는 사전을 만든다. 뉴스 댓글 감정 분석을 하는 과정에서 웹 크롤링, 형태소 분석, 오타 판별을 위한 레빈슈타인 거리, 동음이의어 분류를 위한 Bayesian 분류기를 사용한다.

본 논문의 구성은 다음과 같다. 2 장에서는 기존 관련 연구에 대해서 살펴보고 한계점을 확인한다. 3 장에서

저 작 권 양 도 서

(Copyright Transfer Form)

소속 : 인하대학교 정보통신공학과

성명 : 조 형관 , 김지훈, 하수민 학번 : 12104691 , 12101420 , 12101647

논문제목 : 오피니언 마이닝을 통한 뉴스 댓글의 세부 감정 분석
(Sentiment analysis of article comments focusing on the emotional stages)

본인은 상기 논문을 2015 학년도 1 학기 정보통신프로젝트 최종 보고서 겸 결과 논문으로 제출하고자 합니다. 본 논문의 내용은 저자가 직접 연구한 결과인 것과 이전에 출판된 적이 없음을 확인합니다. 또한 공저자와 더불어 인하대학교 정보통신공학부에서 발간하는 논문집에 본 논문을 수록하는 것을 허락하며 제반 저작권을 정보통신공학부에 양도합니다.

2015 년 6 월 22 일

주저자 : 조 형관 (서명)

공저자 : 김 지훈 (서명)

하 수민 (서명)

지도교수 김유정 (서명)

정보통신공학과장 귀하

오피니언 마이닝을 통한 뉴스 댓글의 감정 분석

Sentiment analysis of article comments using opinion mining

조형관, 김지훈, 하수민
(Hyoung Kwan Cho, Ji Hun Kim, Sue Min Ha)

지도교수 : 김유성

인하대 정보통신공학부

요 약: 최근 인터넷상으로 사람들의 의견 교환이 활발해지면서 이를 수집하고 분석하여 유용한 정보를 추출하는 오피니언 마이닝에 대한 관심이 높아지고 있다. 본 논문에서 그 중 뉴스 댓글의 감정을 분석하는 방법을 제안한다. 한글 텍스트를 분석할 때 어려운 점은 구축되어있는 한글 감정 사전이 없다는 점이다. 이를 위해 본 연구에서는 필요한 감정 분류를 위해 감정 사전을 구축하고, 지속적으로 단어가 추가될 수 있도록 서베이를 만들었다. 웹 크롤링을 통해 수집한 뉴스 댓글과 구축된 감정 사전을 비교하여 감정 분석을 진행하였다. 이 때 Levenshtein 거리를 통해 BoW(Bag of Word)를 생성하고 Bayesian 분류기를 사용하여 오차를 최소화 하였다. 분석 결과를 확인하기 위해 본교 재학생들을 대상으로 설문을 진행한 결과 분노 감정에 대해 93.5%의 신뢰도를 확인하였다. 분석 결과를 사용자들이 실시간으로 확인할 수 있는 User Interface(UI)를 디자인하였다.

주제어: 오피니언 마이닝, 댓글 분석, 감정 사전, Levenshtein 거리, Bayesian 분류기, UI

Abstract: Recently, sharing information online became common. Opinion mining, which is to gather information from it and to analyze it to become useful data, is getting people's attention. In this paper, we propose the emotion prediction method, which analyzes comments of online news articles. When analyzing Korean texts, the most difficult thing is that there is no good Korean sentiment dictionary that fits the needs. For this, we made a sentiment dictionary to classify the feelings into 7 levels and a survey was made to add more words to the dictionary. We compare the crawled news comments and sentiment dictionary to analyze the sentiments. The process has been made using Levenshtein distance and Bayesian classifier to reduce errors. We asked students to analyze the sentiments and as a result, got 93.5% of credibility. Also, we designed a User Interface(UI) so people could see the result of analysis in real time.

Keywords: Opinion mining, comment analysis, sentiment dictionary, Levenshtein distance, Bayesian classifier, UI

Ⅰ. 서론

최근 인터넷의 접근성이 증가하면서 다양한 웹사이트 혹은 소셜네트워크서비스를 통한 정보공유가 활발해졌다. 따라서 이러한 매체들을 통해 생산되는 방대한 양의 데이터로부터 의미 있는 정보를 추출해내는 기술의 필요성 또한 대두되고 있다. 그 중에서도 오피니언 마이닝은 사람들의 여론과 의견을 분석해서 유용한 정보로 재가공하는 기법으로, 이를 활용하여 텍스트에 포함된 내용을 분석할 수 있다. 오피니언 마이닝은 개인 연구뿐만 아니라 정부, 기업 측면에서도 대중의 의견을 분석하기 위해 사용되어 활용도가 매우 높다.

소셜 네트워크 서비스를 통해 공유되는 의견들은 신조어, 약어 또는 광고 등이 많이 포함되어 감정 분류 데이터로 적합하지 않다고 판단하여 본 논문에서는 네이버에서 제공하는 인터넷 뉴스기사의 댓글을 감정 분류 데이터로 선정하였다. 여러 분야의 뉴스를 제공하기 때문에 이에 대해 나타날 수 있는 사람의 감정이 다양할 것이라 생각된다. 따라서 본 연구에서는 오피니언 마이닝을 통하여 뉴스기사 댓글에 대한 감정분석을 진행하고 이를 기반으로 세부 감정 그래프를 가진 새로운 뉴스기사 플랫폼을 설계한다.

특히 이 연구에서는 감정분석을 진행할 때 기존 연구들에서 분석한 것처럼 긍정, 부정이 아닌 총 7가지의 감정으로 데이터를 분석한다. 7가지 감정은 뉴스 기사를 보고 느낄 수 있는 감정을 추출하여 결정되었으며 분노, 놀람(부정), 슬픔, 무관심, 흥미, 놀람(긍정), 행복이 이에 해당한다. 자세한 감정 분류를 위해 감정 사전을 구축하고 사전에 등재되지 않은 단어를 처리하기 위해서 survey 시스템을 설계하여 실시간으로 변화하는 사전을 만든다.

뉴스 댓글 감정 분석을 하는 과정에서 웹 크롤링, 형태소 분석, 오타 판별을 위한 레빈슈타인 거리, 동음이의어 분류를 위한 Bayesian 분류기를 사용한다.

본 논문의 구성은 다음과 같다. 2 장에서는 기존 관련 연구에 대해서 살펴보고 한계점을 확인한다. 3 장에서

세부적인 연구방법을 절차에 따라 기술한다. 총세 단계로 구성되어있으며 웹크롤링, 텍스트마이닝, 오피니언마이닝 순이다. 4 장에서 최종결과를 도출하고 5 장 에서 결과 에 대한 신뢰도, 결론 및 향후 발전 가능성에 대해 논의한다.

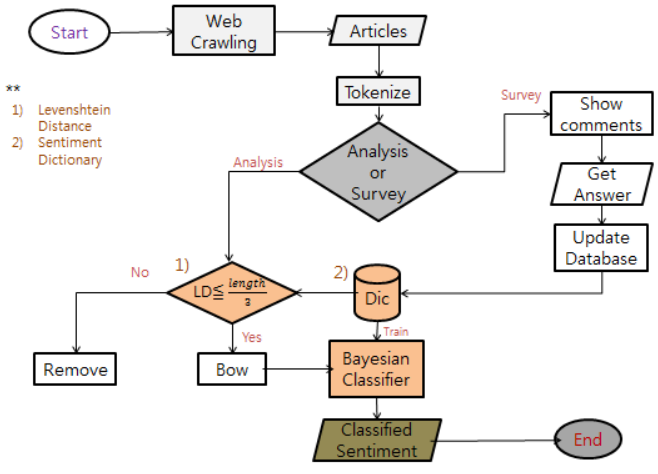
Ⅱ. 관련 연구

오피니언 마이닝을 활용할 수 있는 분야로는 감성을 분류하는 연구로 기사, 영화평, SNS 댓글 등을 중심으로 연구가 활발히 진행되고 있다. 그러나 기존 연구에서는 형태소 분석 결과를 통해 그 단어가 반드시 들어가야만 단순 matching 방식을 사용하고 있다. 따라서 예외처리할 부분이 많이 발생하게 된다. 예를 들어 “짜증나”와 “짜잉나”를 비교할 때 기존의 방식을 사용하면 감정으로 인식되지 않는다. 하지만 레빈슈타인 거리를 사용하면 단어를 각 자음과 모음으로 분류하여 음소 단위로 비교하여 거리로 나타낸다. 이를 통해 오타를 판별할 수 있다.

기존 연구의 또 다른 단점은 긍정과 부정, 많으면 중립까지 포함하여 세 가지로만 감정을 분석하고 있다는 점이다. 이러한 연구의 경우 긍정, 부정에 대한 초기 씨앗 단어를 설정하고 각 씨앗 단어와 단어들의 Pointwise Mutual Information(PMI)를 이용하거나, 영어 단어 사전인 WordNet 을 활용하여 동의어, 반의어 관계를 이용하고는 한다. 그러나 이러한 분류의 경우 감정 분류의 개수가 적어 자세한 감정을 파악할 수 없다는 한계를 지닌다. 따라서 본 연구에서는 총 7 가지의 감정 분류 방법에 대하여 기술한다. 이는 뉴스 기사를 보고 사람들이 느낄 수 있는 감정들로 구성되었으며, 손선주 외 3 인의 “한국어 감정표현단어의 추출과 범주화”의 부록으로 첨부된 감정 단어 목록을 참고하였다.

Ⅲ. 연구 방법

본 연구는 웹 크롤링, 텍스트 마이닝, 오피니언 마이닝 세 단계를 거쳐 진행되며 그림 1 은 연구 방법의 전체



과정을 나타낸 블록 다이어그램이다.

그림 1. 전체 블록 다이어그램

3.1 웹 크롤링

감정 분류 데이터로 선정한 네이버 뉴스 기사와 댓글을 모두 수집하는 것이 본 연구의 첫 번째 단계이며 이를 웹 크롤링이라 한다. 웹 크롤러는 조직적, 자동화된 방법으로 월드 와이드 웹을 탐색하는 컴퓨터 프로그램이다. 웹 크롤러가 하는 작업을 웹 크롤링이라고 하며 대체로 방문한 사이트의 모든 페이지의 복사본을 생성하거나 웹 페이지의 특정 형태의 정보를 수집하는데 사용된다.

본 연구의 목적은 분석 대상인 모든 뉴스 기사를 기사별, 날짜별, 카테고리 별로 분류하여 감정상태를 나타내는 것이다. 이를 위해 네이버 사이트의 소스코드를 가져오고 파싱이라는 과정을 거쳐서 한글 텍스트만 가져올 수 있어야 한다. 파싱이란 일련의 문자열을 의미있는 토큰으로 분해하고 이들로 이루어진 파스 트리를 만드는 과정을 말한다.

이 단계에서 ruby 라는 프로그래밍 언어를 사용한다. Ruby 는 사용자 친화적인 문법으로, 손쉬운 개발 환경을 제공하며 텍스트 처리에 뛰어난 언어이다.

우선, 웹사이트와 상호작용을 하기 위한 라이브러리인 mechanize gem 을 이용하여 네이버 뉴스 사이트의 전체 소스를 불러온다. 그 다음, json gem 을 사용하여 한글 텍스트만 파싱한다. 이렇게 모은 데이터를 분석 대상으로 쓰거나 User Interface(UI)에 불러올 수 있도록 데이터베이스에 저장한다.

3.2 텍스트 마이닝

텍스트 마이닝은 텍스트 데이터에서 자연 언어 처리 기술에 기반하여 유용한 정보를 추출, 가공하는 빅데이터 분석 기술이다. 본 연구에서는 형태소 분석이 이 단계에 해당한다. 많은 분야에서 텍스트 마이닝을 진행하고 있지만 대부분의 연구는 영어로 진행되며 상대적으로 한글 형태소 분석기 연구는 미흡하다. 따라서 한글과 가장 유사한 특징을 가진 일본어 형태소 분석기인 mecab 을 통해 형태소를 분석한다. 기존 연구에서 mecab 에 한글 문법을 추가하였고 이 연구에서는 그것을 바탕으로 뉴스 기사 댓글에서 볼 수 있는 예외까지 추가시켰다.웹 크롤링 단계를 통해 가져온 순수 한글 댓글을 형태소 분석기로 분석했다.

형태소는 의미기능을 부여하는 언어의 형태론적 수준에서의 최소 단위를 말한다.따라서 형태소 분석은 문장을 구성하는 단어들로부터 최소 의미단위인 형태소들을 분리해내고 각 형태소들의 문법적 기능에 따라 적절한 품사(명사, 동사, 부사, 형용사, 관사 등)를 부착하는 기술이다.

기본적으로 mecab 에서 제공되는 기능으로 각 단어에 대한 품사를 지정 할 수 있어 감정어의 어근이 될 수 있는 명사와 형용사를 필터링 한다. 그렇게 얻은 결과물을 mysql 데이터베이스에 저장한다.

형태소 분석 결과를 비교할 대상은 감정사전이다. 감정사전의 경우 기존 연구에서는 긍정, 부정 혹은 중립으로만 분류하기 때문에 새롭게 제작했다.

Thayer 의 감정 분류 모델은 경우 2 차원 공간에 사람의 감정 상태를 수치적으로 표현하는 차원 접근법을 제시한다. 이 모델은 감정을 Excited, Happy, Pleased, Relaxed, Peaceful, Calm, Sleepy, Bored, Sad, Nervous, Angry, Annoying 으 로 분류한다. 그러나 본 논문에서는 성능 향상과 정확한 감정 예측 오류를 최소화하기 위해 ‘분노’, ‘놀람(부정)’, ‘슬픔’, ‘무관심’,

'흥미', '놀람(긍정)', '행복' 7 가지로 감정을 분류하였다. 그림 2 의 7 단계 감정 분류 모델에서 확인할 수 있다.

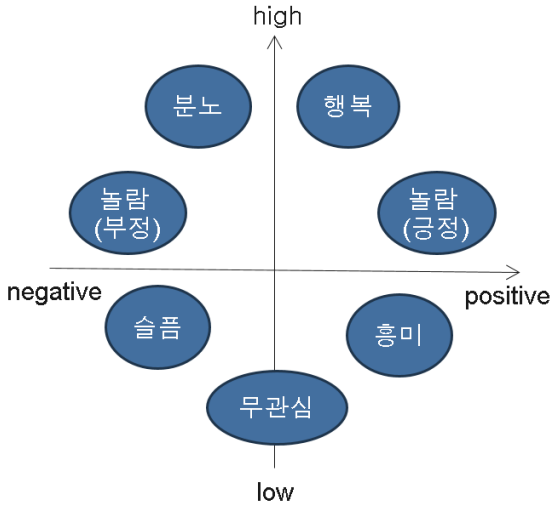


그림 2. 7 단계 감정 분류 모델

'놀람'이라는 감정의 경우 'startled', 'confused'에 해당하는 '부정적인 놀람'과 'amazed', 'excited'에 해당하는 '긍정적인 놀람'이 있기 때문에 이를 y 축을 기준으로 대칭이 되도록 표현하였다.

각 감정 집합의 키워드를 추출하여 감정 사전을 구축하고, 이를 통해 사용자들이 작성한 댓글에 대한 감정 분석을 수행하고자 한다.

감정 사전을 구축할 때는 기존 논문인 "한국어 감정표현단어의 추출과 범주화"의 부록을 참고하였다. 이를 참고한 이유는 어휘빈도를 고려한 단어들을 기초자료로 하여 일상생활에서 실제로 사용하는 단어들을 분류해놓았기 때문이다. 또한, 감정 단어 선별을 국문학 전공자와 감정연구자들이 담당하였는데, 이접근방식은 감정의 정의나 유형에 대한 지식이 있고감정을 다른 심리과정과 구분할 수 있는 능력을 지닌 전공자들의 분석을 통한 감정표현언어 추출작업이라는 점에서 신뢰성을 높인다. 전문가들에 의해 여러 단계에 시행을 거쳐 신중하고 일관성 있게 단어들을 선별한 방법은 부적절한 단어가 목록에 포함될 오류를 줄일 수 있는 합리적인 방법이다.

본 연구는 뉴스 댓글을 분석 대상으로 삼기 때문에 신조어나 은어 및 비속어 그리고 간단한 이모티콘을 감정 사전에 등록하여 감정 분석이 효과적으로 진행되도록

설계하였다.

그림 3. 서베이 화면

이 외에도 지속적으로 새로운 단어를 추가할 수 있도록 서베이를 진행한다. 서베이 페이지를 제작하여 실제 뉴스 댓글에 포함되어있는 단어들에 대한 감정을 선택할 수 있도록 하였다. 이는 실시간으로 진행되며 그렇게 감정과 연결된 단어들은 감정 사전 데이터베이스에 저장된다. 또한 확실하게 분류되어있는 감정 단어들이 서베이 대상의 잘못된 선택으로 잘못 분류되지 않도록 감정 단어별 웨이트를 설정하였다. 서베이 진행 화면은 그림 3 과 같다.

3.3 오피니언 마이닝

마지막 연구 단계는 오피니언 마이닝이다. 오피니언 마이닝이란 어떤 사안에 대한 사람들의 의견이나 평가, 태도, 감정 등을 분석하는 것을 말한다. 특정 주제에 대한 사람들의 의견을 모아 문장을 분석한다.문장 분석에서는 의견을 뽑아내어 긍정과 부정으로 나누고 그 강도를 측정하는데 본 연구에서는 3.2 에 제시된 7 단계 감정 분류 모델에 따라 분석을 진행한다. 분석을 진행하는 과정은 그림 3 과 같으며 레벤슈타인 거리와 Bayesian 분류기를 사용한다.

레벤슈타인 거리는 두 문자열 사이의 차이를 계산하기 위한 알고리즘이다. 하나의 문자열에 수정 연산을 하여 다른 문자열과 같게 할 때 가장 적은 수정 연산 횟수를 구하는데, 이 횟수가 레벤슈타인 거리이다. 수정 연산은 삽입, 삭제, 치환이있으며, 문자의 수정 횟수를 이용한다. 두 문자열 α와 β가 주어졌을 때 두 문자열의 레벤슈타인 거리를 구하는 알고리즘은 다음과 같다.

LD(α, *len*α, β, *len*β)
if *len*α= 0 then return *len*β
if *len*β= 0 then return *len*α
if αlenα- 1 = βl e n β- 1 then cost = 0
else cost = 1
return min(LD(α, *len*α - 1, β, *len*β) + 1
LD(α, *len*α, β, *len*β - 1) + 1
LD(α, *len*α - 1, β, *len*β – 1 + cost)

위 알고리즘에서*len*ω는 문자열 ω의 길이를 의미하며ω¡의 형태는 문자열 ω 의¡번째 문자를 의미한다. 위의 알고리즘은 문자열 α를 문자열 β와 같도록 하는 거리를 계산하는 알고리즘 을 의미한다.본 논문에서는 두 문자열 사이의레벤슈타인 거리가 작을 수록 유사한 문자열일 가능성이 많다고 가정을 하고 이를 이용하여 BOW를 생성하고 이를 기계학습에 적용한다.

형태소 분석 결과는 의미 없는 단어들이 많이 존재하므로 Bag of Words(이하 BOW)라는 전처리 과정을 통해 새로운 입력 배열을 생성한다. 즉,

형태소 분석된 단어의 길이<(사전 내 단어/3)

인 경우 가장 가까운 단어를 새로운 입력배열에 삽입하여 BOW 를 생성한다. 이 값은 여러 번의 시행착오에 의한 것으로서, 단어의 음절 수가 1 인 경우에는 오차가 없는 것으로 가정하고 음절 수가 2 이상인 경우부터 판별한다. 음절 수가 2 이고 *len*ω가 4 인

경우 예를 들어, 자다 의 경우 “ㅈ,ㅏ,ㄷ,ㅏ” 로 분리되고 $len_\omega / 3 = 1.33$ 이 되어 적어도 1 개의 자음 혹은 모음의 오타를 판별 할 수 있게 된다. 마찬가지로 음절 수가 2 이고 받침을 가지고 있어 len_ω가 6 인 경우는 최대 2 개까지의 오타를 판별 할 수 있는 값이 된다.

감정단어는 주관적으로 분류할 수 있는 기준이 존재하지 않기 때문에 확률 모델로써 분류에 접근한다. 기본적으로 Bayesian Classifier 는 조건부 확률에 의존한다. 즉사건 B 에 대한 A 의 조건부 확률은 다음과 같다.

$$P(A|B) = P(B|A) * P(A)/P(B)$$

이를 활용하여 한 기사에서 각 감정으로 분류될 확률을 구하면 다음과 같다.

$$P(sentiment|article) = P(article|sentiment) * P(sentiment)/P(article)$$

이는 각각의 감정에 대한 확률을 비교하기 위한 것이므로 공통 분모인$P(article)$을 생략할 수 있다.
따라서

$$P(sentiment|article) = p(article|sentiment) * P(sentiment)$$

이고

$$P(article|sentiment article) = \prod p(words|sentiment)$$

으로 나타낼 수 있다.

즉 사전의 각 감정에 나타나는 단어 각각의 확률을 계산하기 때문에 서베이를 통해 다수의 의견을 반영하여 weight 를 가진 감정사전을 그대로 적용할 수 있다.

하지만 여기에서 문제는 weight 가 0 인 감정의 값을 계산시에 누적 곱의 결과가 0 이 되어버린 다는 것이다. 따라서 분자에 +1 을 해주고 분모에는 학습데이터의 전체 개수 즉, 감정 사전 내 전체 단어 수를 더해준다. 이를 Laplace smoothing 이라고 한다. 이에 해당하는 식은 다음과 같다.

$$P(A|B) = \frac{count(A,B)+1}{\sum(count(A,B)+1)} + \frac{count(A,B)+1}{\sum(count(A,B)+|V|)}$$

확률 모델은 단어의 수가 많아질수록 더욱 세밀하게 분석을 할 수 있다는 장점이 있다. 하지만 input 단어의 개수가 너무 많아질 경우 float 데이터 타입의 최소값을 초과하여 0 이 되어버린다. 이를 Log under flow 라고 한다. 이를 방지하기 위해 확률 곱 연산을 모두 Log 처리하여 곱 연산을 합 연산으로 바꾸면 각 감정의 확률을 비교할 수 있다. 따라서 이 과정을 통해 단어의 개수가 많은 section 별 분류 단위에서 감정분석 결과를 낼 수 있다. 하지만 Log 처리를 거치는 과정에서 각 확률의 비율이 Equalized 되어 변화하기 때문에 대수비교에만 사용가능 할 뿐 전체적인 확률로 나타낼 수는 없다. 따라서 section 별 분류를 할 경우에는 각 기사에서 도출된 감정들이 신뢰도 있는 결과라 가정하고 그 비율이 누적된 값의 산술 평균을 통하여 결과를 도출한다.

Ⅳ. 실험 및 결과 분석

4.1 실험 데이터

네이버 뉴스 기사의 주요 3 개 분야 정치, 사회, 문화 분야 각각에 대해서 최신 270 개의 기사를 선별하였으며 각 기사에는 20 개 이상의 댓글이 존재한다. 분야별 분석 후 한 개의 기사에 대한 분석을 진행한다.

4.2 실험 결과

4.2.1 분야별 실험 결과

정치, 사회, 문화 분야별 분석을 했을 때 세 분야 모두 분노의 감정으로 분류되었으며 분류값은 표 1 과 같다.

표 1. 결과값

분야	감정	결과값
정치	분노	66.2%
	흥미	12.8%
	놀람(부정)	11.6%
사회	분노	46.5%
	흥미	23.3%
	놀람(부정)	16.3%
문화	분노	38.7%
	흥미	35.5%
	놀람(부정)	8.33%

이는 각 분야에 대한 상위 세 개의 감정과 결과값을 나타낸다.

4.2.2 기사별 실험 결과

문화 분야의 기사 중 “출산 10 시간만에 하이힐 신고 퇴원? 동서양의 산후조리”라는 기사를 분석한 결과는 다음과 같다.

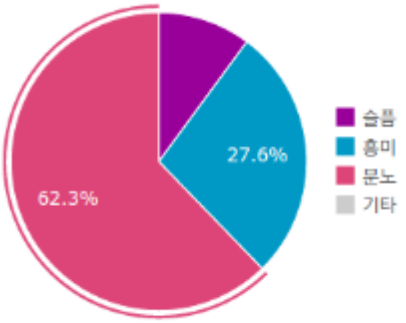


그림 4. 한 기사에 대한 감정 분포도

감정 퍼센트를 확인해본 결과 상위 세 개의 감정이 분노, 흥미, 슬픔으로 나타났다. 각각의 퍼센트는 분노 62.3%, 흥미 27.6%, 슬픔 10%이며 기타 감정으로 놀람(긍정)이 0.01%로 나타났다.

분석 대상이 된 댓글 중 “부럽다.. 난 평소 운동도 안하는 사람인데 마라톤도 좋아해서 그래도 애낳고 3 일 동안 못움직임.. 늙어서 그런가.. 슬프네..” 이 있다. 이

댓글이 형태소 분석되었을 때, '부럽다'와 '슬프다'가 동일한 비율로 존재하지만 Bayesian 분류기를 통과한 결과 문맥에 맞게 분석이 되었다는 것을 확인할 수 있다. 또한 디자인 적인 요소를 추가하여 분석된 단어의 크기에 따라 파이차트의 크기를 변화시켜 사용자들이 시각적으로 인지할 수 있게 하였다.

4.2.3 결과 페이지

실험 결과를 사용자들이 확인할 수 있도록 User Interface(UI)를 제작하였다. 웹 페이지를 통해 기사의 분야별 그리고 각 기사별 감정 분석 결과를 확인할 수 있다. 전체 화면은 그림 5 와 같으며 페이지 정치, 사회, 문화 분야를 선택했을 때 그 분야에 해당하는 기사들의 목록이 왼편에 나타나며 감정 분석 결과를 오른편에서 확인할 수 있다.

그림 5. UI 전체 화면

4.3 결과 분석

4.2.1 의 실험 결과를 분석한 결과 정치, 사회, 문화 모두 분노의 감정으로 분류되었다. 이는 전체 section 에 대한 분석과 각 기사들의 산술평균으로 계산된 값 모두 동일했다. 이는 분류 시에 상대적으로 작은 데이터의 특징을 가질 수 있는 각 기사 단위의 분석도 의미 있는 값을 만들어 낸다는 것을 나타낸다.

4.2.2 의 경우 특정 기사에 대한 댓글을 분석한 결과 상위 세 개의 감정이 분노, 흥미, 슬픔이라는 것을 확인할 수 있었다. 결과를 검증하기 위해 본교 재학생 40 명에게 설문조사를 시행하였다. 실험에 사용한 기사를 제공하고 댓글을 확인한 뒤 사용자들의 감정이 어떤 것이라 생각하는지 선택하게 하였다. 그 결과 분노 24, 흥미 10, 놀람(부정) 4, 슬픔 2 라는 결과가 나왔다. 이를 백분율로 계산해보면 분노 60%, 흥미 25%, 놀람(부정) 10%, 그리고 슬픔이 5%로 나타난다. 이 수치를 이용하여 신뢰도를 계산해보면 분노가 93.5%, 흥미 90.58%, 슬픔 50% 이다. 이를 통해 본 연구의 결과물에 대한 검증을 할 수 있었다.

마지막으로 본 논문의 방법으로 해결이 힘든 문제는 위에서 언급한 나머지 감정들에 대한 확률 처리 방법이다. 예를 들어 log 값으로 다른 감정 값들과 비교할 때 대수 비교는 할 수 있지만 그 비율이 달라지기 때문에 정확한 확률을 알 수 없다.

이러한 예를 정확히 분석하기 위해서는확률 값이 0 에 가까워지지 않도록 충분한 서베이 및 감정사전 개정을 통해 단어의 개수를 늘려야 한다.

이 외에도 형태소 분석이 완벽하지 않다는 점을 고려해야 한다. 문장이 너무 많은 단어로 쪼개져서 의미가 없는 부분이 감정 분석 대상으로 고려되는 경우가 있다. 이러한 문제를 해결하기 위해 지속적으로 형태소 분석 사전을 업데이트하고 테스트를 통해 의미 없는 단어들을 필터링 하는 과정이 필요하다.

Ⅴ. 결론

이번 연구를 통해 각 분야별 다양한 기사에 대해서 감정 분류를 진행하였다. 분야별로 분석을 했을 때 정치, 사회, 문화 분야가 모두 분노의 감정을 가장 크게 나타냈다. 이는 뉴스 기사를 보고 댓글을 다는 사람들 중 분노를 느끼는 사람들이 많다는 것을 보여주며 부정적인 뉴스 기사가 많다는 것을 알 수 있게 한다.

분야별 결과를 확인한 뒤 4.2.2 처럼 개별 기사의 댓글을 분석해보기도 했다. 그 결 4.2.2 의예시에서처럼 분류기 잘 되는 경우도 있지만 분류가 한쪽으로 편향되는 경우도 있다. 이는 확률 모델 한계로서 각 단어에 대한 총 모집단의 개수가 서로 다른 데에서 비롯된다. 위에서 언급했던 Laplace Smoothing 과정을 통해 처리하였지만 만약 가장 모집단이 작은(각 단어가 나타날 확률이 큰) 감정의 단어들이 너무 많이 나타나게 될 경우 다른 감정에 대한 확률들은 상대적으로 0 에 가깝게 된다. 하지만 이것이 바로 분류기의 본 목적인 하나의 감정으로 분류하여 결론을 내는 것이므로 바람직한 결과라고 할 수 있다.

또한 본교 재학생들을 대상으로 한 해당 기사에 대한 설문조사를 통해 결과물에 대한 신뢰도를 확인할 수 있었다.

하지만 여기서 문제는 "사전에 없는 단어들이 많이 등장 할 경우 어떻게 처리할 것인가?" 이다. 감정 단어들은 본래 하나의 감정으로 단정지어 결정될 수 없기 때문에 감정사전에 대해서는 survey 시스템을 통해서 사전이 갱신 될 수 있는 여지를 두었다. 뉴스 기사 댓글 전체로부터 형태소 분석을 다시 진행하여 그 형태소들을 바탕으로 survey 문제를 구성하였기 때문에 주관적으로 판단할 수 없는 감정 단어들일지라도 대중의 집단지성을 활용할 수 있다.

분류기 자체가 확률 모델이고 분류에 앞서 레벤슈타인 거리를 통한 BoW 를 생성하여 사전에 없는 단어에 대한 확률이 영향을 미치는 것을 방지했기 때문에 최종 도출되는 결과값 자체가 신뢰도를 의미한다고 볼 수 있다. 따라서 실제 의미있는 신뢰도 값은 단어가 매칭될 때 나타나는 비율 즉, 각 기사에 대한 BoW 생성률은

(형태소분석된 전체 단어 개수/BoW 의 개수)*100(%)

를 확률로 표현할 수 있다. 현재 평균 적인 Bow 생성율은 약 16.7%이다. 이는 신뢰도로 보기에는 낮은 값이지만, 서베이가 이제 막 시작되고 있다는 점에서 상당히 높은 값으로 볼수 있다. 또한 서베이의 근원 자체가 전체 기사의 형태소 분석된 단어 이므로 최종적으로는 BoW 생성률은 100%에 달할 것이다.

기계학습은 대표적으로 Supervised 와 Unsupervised 로 나눠지는데 이 연구에서 다룬 분류기는

Supervised 다. 이는 사용자가 미리 분류에 대한 값을 지정해주고 그 학습결과에 따라 분류가 되는 모델이다. 하지만 분류된 결과값에 대한 충분한 신뢰도가 있다면 Unsupervised machine learning 을 수행할 수 있다.

현재는 결과 값에 대한 절대적인 신뢰도가 없지만향후에는충분히 신뢰도 있는 결과를 얻어 각 분류결과를 threshold 값을 통해 하나의 감정으로 확실하게 결정하고 이를 다시 training 하는 Unsupervised training 을 수행한다. 이러한 과정은 독자들이 감정분석 결과를 구독할 때마다 그 값이 실시간으로 반영되어 보다 높은 분류 신뢰도를 얻을 수 있게 한다.

본 연구를 통해 뉴스 기사의 구독자들은 뉴스기사를 읽음과 동시에 수많은 댓글들에서 다른 구독자들이 어떠한 감정을 느끼고 있는지 알 수 있으며, 기업이나 단체 차원에서도 각 기사의 분야별 감정분석 결과를 활용할 수 있을 것이다.

VI. 참고 문헌

[1]서영훈 외 3 인, “Levenshtein 거리를 이용한 영화평 감성 분류 Distance”, 충북대학교 석사학위 논문, 2011.

[2]윤영선, 강점자, “레벤스타인 거리에 기초한 위치 정확도를 이용한 고립 단어 인식 결과의 비유사 후보 단어 제외”, 『말소리와 음성과학 제 1 권 제 3 호』 pp. 109~115, 2009

[3] 손선주 외 3 인, “한국어 감정표현단어의 추출과 범주화”,감성과학, Vol. 15, No. 1, pp.105-120, 2012

[4] 김윤석, 서영훈, “기계 학습을 이용한 한글 텍스트 감정 분류”, Journal of Korea Multimedia Society Vol. 17, No. 2, pp. 232-239, 2014

[5] 이강복 외 2인, “SNS에서 단어 간 유사도 기반 단어의쾌-불쾌 지수 추정”, 2013

[6] 김진수, “문단 분석을 통한 문서 내의 감정 분석”, 2014

[7] 황재원, 고영중, “감정 자질을 이용한한국어 문장 및 문서 감정”, 2008

[8] 이공주 외 3 인, " 뉴스 댓글의 감정 분류를 위한 자질 가중치 설정”, 2010

[9]김윤석,서영훈, “기계학습을이용한한글텍스트감정분류”,한국엔터테인먼트산업학회 2013 추계학술대회 논문집, pp. 206-210, 2013

[10] 황재원, 고영중, "감정 단어의 의미적 특성을 반영한 한국어 문서 감정분류 시스템”, 정보과학회논문지, 2010

[11] 조하나 외 3인, “인터넷 뉴스 댓글의 감성 분석을 통한 오피니언 마이닝”, 2013

[12] 안정국, 김희용, “한글 감성어 사전 API 구축및자연어처리의활용”, 한국지능정보시스템학회 2014년 추계학술대회, pp. 177-182, 2014.11

[13] 김은영, “현대 국어 감정동사의 범위와 의미 특성에 대한 연구”, 2005

조 형 관

인하대학교 정보통신공학과 재학중.

김 지 훈

인하대학교 정보통신공학과 재학중.

하 수 민

인하대학교정보통신공학과재학중.