기계 학습을 이용한 한글 텍스트 감정 분류 및 분석

Classification and Analysis of Emotion in Korean Texts using machine learning

저자
(Authors)

한명호, 류주현, 서수영
Myung-Ho Han, Joo-Hyeon Ryu, Su-Young Seo

출처
(Source)

한국정보과학회 학술발표논문집 , 2014.6, 1722-1724(3 pages)

발행처
(Publisher)

한국정보과학회
The Korean Institute of Information Scientists and Engineers

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE02444514

APA Style

한명호, 류주현, 서수영 (2014). 기계 학습을 이용한 한글 텍스트 감정 분류 및 분석. 한국정보과학회 학술발표논문집, 1722-1724

기계 학습을 이용한 한글 텍스트 감정 분류 및 분석

한명호 ○, 류주현, 서수영
명지대학교 컴퓨터공학과
promh90@gmail.com, ssy_angel@naver.com, rjhz90@naver.com

Classification and Analysis of Emotion in Korean Texts using machine learning

Myung-Ho Han○, Joo-Hyeon Ryu, Su-Young Seo
Dept. Computer Engineering, Myongji University

요 약

스마트폰의 보급으로 인해 시간과 공간의 제약을 받지 않고 SNS을 통해 자신의 의견이나 감정을 표현할 공간이 넓어졌다. 이로 인해, 텍스트 데이터의 양이 급증함에 따라 텍스트 데이터 활용의 중요성이 대두되고 있다. 본 연구에서는, 텍스트를 긍정과 부정을 판단하는 감성적인 연구(sentiment analysis)가 아닌, 여러 감정으로 분류하는 감정분석(emotion analysis) 연구를 제안한다. 또한, 제안하는 기법을 영화평 데이터로부터 감정 분류하는 방법에 적용하였다. 감정 분류를 위한 인간의 기본적인 7가지 감정으로 분류하기 위해, 각 감정마다 감정 단어 사전을 구축하고, 통계기반 기계학습 정보를 이용하여 한글 텍스트 감정 분류에 적용하였다. 본 연구에서 제안한 방법이 실제 응용분야에서 적용 가능함을 보여준다.

키워드: 감정분석, 영화평분석, 기계학습

1. 서 론

최근 인터넷 통신과 정보검색 등의 기능을 갖춘 스마트폰의 보급으로 인해 사용자들의 관심 분야나 활동들을 공유할 수 있는 SNS의 선풍적인 인기를 끌게 되었다. 텍스트와 사진 또는 이미지를 작성할 수 있는 SNS는 정치, 사회 분야에서 큰 영향을 끼치게 되었고 이러한 영향이 중요시 됨으로 텍스트를 분석해 사용자들의 감성과 의견을 통계/수치화하는 오피니언 마이닝(Opinion Mining)[1] 분야가 각광을 받게 되었다.

과거에는 오피니언 마이닝 분야가 특정 오브젝트에 대한 의견이 긍정인지 부정인지 또는 중립인지에 대한 감성 분석연구가 주를 이루었지만 근래에 들어서는 사용자들의 감정을 분석하는 쪽으로 방향을 돌리고 있다.

본 연구에서는 '칠정(七情)'이라는 인간의 기본적인 일곱 가지 감정 희(喜), 노(怒), 애(哀), 락(樂), 애(愛), 오(惡), 욕(欲)에 바탕을 두어 비교적 사용자들의 감정표현이 잘 드러난 영화평에 적용하여 감정의 특성을 분류하고자 한다. 기본적인 7개의 감정은 심리학에서 분류하는 여러 정서이론 중에서 가장 적합하다고 판단되어 사용하게 되었다. 본 연구는 국내 대형 포털 사이트인 Daum과 Naver의 영화평 데이터를 사용하여 기계학습[2]을 통해 모델을 만들고, 학습된 모델을 이용하여 영화에 대한 영화평들의 감정 특성을 보여준다. 또한, 기계학습에 의한 방법을 위해 Bag of Word[3] 기법을 사용하고자 Raw Data로부터 감정자질[4]에 대한 출현 횟수를 추출하였다. 또한, 결과값을 이용하여 SVM[5][6] (Support Vector Machine) 알고리즘을 적용하여 학습을 시켰다.

2. 기계학습 분류(Classification)

2.1 전처리(Preprocessing)

본 연구의 전처리 과정은 다음과 같다..

Jsoup 외부 라이브러리를 이용한 크롤러(crawler)를 통해 포털 사이트 Daum, Naver에서 제공하는 영화평을 20만개 이상을 수집하였다. 그 후, 영화평 필터(Filter)를 거쳐서 광고 텍스트 및 비속어 등의 감정[7]분류에 적합하지 않은 데이터를 제외한 나머지 텍스트 약 15만개 정도를 추출하였다. 또한 본

연구에서는, 한글 텍스트 분류라는 점에서 한글을 제외한 다른 언어들은 빈칸 처리함으로써 한글 영화평만을 사용하였다. 추출된 영화평 데이터들을 형태소 분석기를 통해 태깅 된 형태소로 추출하였다.

2.2 감정 자질 추출

기계학습의 과정은 다음 그림 1과 같다.

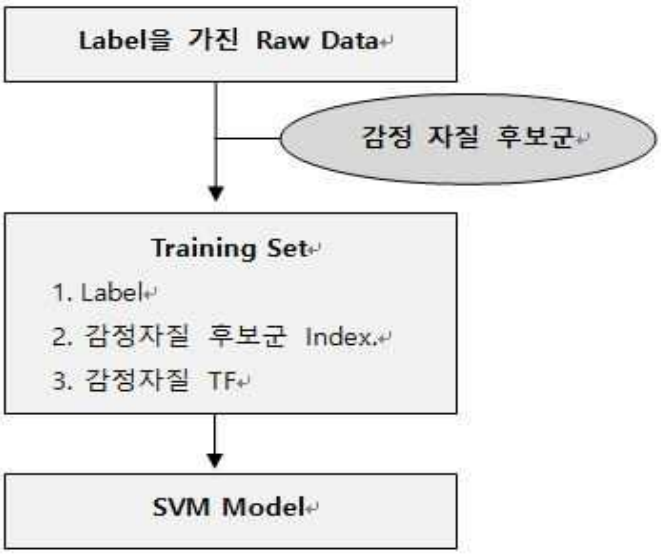

그림 1. Raw Data는 감정 자질 후보 군에서 Training Set을 추출한다. 추출된 Training Set으로 Model을 만든다.

본 연구에서는 "KOMORAN 형태소 분석기"를 이용하여 감정 자질(Emotion Feature)이 될 수 있는 VA(형용사), NNG(일반 명사), NNP(고유 명사) 형태소를 추출하였다. 여기서 말하는 감정 자질이란, 감정이 될 수 있는 형태소의 품사이며, 우리는 분리된 형태소들을 직관적으로 보고, 관찰 한 결과 대부분의 감정들이 VA, NNG, NNP에서만 출현하였다는 것을 발견하였다. 사전에 대량의 문서(document)에서 감정자질들을 추출하였다. 이 때, 각 감정자질의 TF(Term Frequency) 값을 사용하여 최소 10번 이상 나온 감정자질만을 저장해 놓았다. 저장해 놓은 자료를 토대로 Bag of Words 기법을 이용해 감정자질들을 한 문서로부터의 TF값을 측정하였다. 추출된 감정자질의 Index와 해당 감정자질의 TF값을 기계학습의 Training Set으로 사용하였다.

2.3 기계학습을 이용한 분류

본 연구에서는 학습시키려는 데이터와 이에 대한 Label을 같이 학습 시키는 Supervised Learning을 시도한다.

여기서 학습시키려는 데이터는 사전에 Bag of Words 기법을 통해 만들어진 Training Set이며, 이에 대한 Label는 7개의 감정범주(Class) 중 하나이다. 본 연구에서는 기계학습 알고리즘으로 SVM (Support Vector Machine)을 사용하였다. SVM 알고리즘을 사용한 이유는 감정범주에 해당하는 사전(Dictionary) 기반으로 감정을 분류할 때 한계가 있다. 사전 데이터의 양이 많아짐에 따라 정확성이 높아지긴 하지만, 감정사전에 들어가 있지 않은 형태소가 출현하게 된다면, 감정분류를 하지 못하게 된다. 문서(Document)와 이에 해당하는 답지(Label)를 SVM Model에 학습시킨다면, 감정사전(Emotion Dictionary)이 없어도, 새로운 데이터를 분류 가능하다는 점에서 SVM알고리즘을 적용하였다.

또한 데이터가 많을 경우, SVM은 과도하게 에러를 줄이려 하지 않고, 불필요한 자질을 걸러내면서 학습하는 능력이 뛰어나기 때문에 사용하였다. 이후, 학습된 모델을 통해 새로운 데이터의 예측(Predict)을 하였다. 기계학습에 사용될 형태소의 자질을 선별 하기 위해서 출현빈도가 낮은 형태소를 제거하였다. 현재 사용하고 있는 형태소 분석기는 사전에 없는 단어를 분석 할 경우 형태소로 나누지 못하고 입력단어 그대로 출력하기 때문에 띄어쓰기가 없거나, 여러 단어가 붙어 있거나, 맞춤법에 어긋난 단어들을 새로운 형태소로 인식하게 된다. 때문에 점유율(occupancy ratio)이 최소 하나 이상인 경우에만 적합한 자질을 갖췄다고 판단하였다[8]. 점유율 R은 아래의 식과 같이 정의하고 여기서 E는 감정, m은 형태소를 뜻한다. :

$$R(m_i|E_n) = \frac{\mathrm{Count}(m_i|E_n)}{\mathrm{Count}(E_n)}$$

이 점유율의 최솟값을 기준으로 출현빈도가 낮은 형태소를 제거하였고, 이때의 최솟값을 최소점유율(Minimum Occupancy Ratio, MOR)이라 한다. 어떤 형태소가 적어도 하나의 감정에서 최소점유율 이상의 출현빈도를 가지게 되는 경우 자질로 선별하여 분석에 사용하였다.

3. 실험 및 결과

3.1 테스트 데이터

테스트 데이터는 각 감정에 해당 된다고 판단한 7개의 영화 리뷰 총 6만5천개의 데이터를 사용하였다. 해당 문서들의

모든 VA, NNG, NNP 에 대해 최소 10번 이상 출현하고 2-gram이상의 형태소들만을 추출하였다. 그 결과 총 2542개의 형태소를 확보 하였다. 이 형태소들을 기준형태소라 하겠다.
그 후, 각 범주에 해당하는 문서에서 리뷰(영화평 데이터) 한 줄씩 기준형태소와 비교하여 출현 형태소의 위치와 TF값을 측정하였다. 그 결과들을 훈련 데이터로 사용해서 SVM Model을 만들었다.

3.2 결과 분석

임의의 테스트 데이터 2000개를 수작업을 통해 해당 데이터의 답을 미리 작성하여 만들어진 <표1>을 SVM Model에 적용하여 예측한 결과 <표2>와 같은 결과를 얻었다.

<표1> 수작업을 통해 미리 작성한 감정 개수

희(喜)	노(怒)	애(哀)	락(樂)	애(愛)	오(惡)	욕(欲)
351	236	189	332	284	387	221

<표2> SVM 분류 결과 감정 개수

희(喜)	노(怒)	애(哀)	락(樂)	애(愛)	오(惡)	욕(欲)
241	279	264	449	339	265	163

그 결과, 정확도(accuracy)는 약 69%가 나왔다. 낮은 정확도의 이유로는 해당 감정범주를 나눌 때, 희(喜)와 락(樂)의 감정이 너무 비슷해 감정의 경계선이 불투명 하고, 보통 사람들이 영화를 보고 난 후, 애(哀)와 애(愛)의 감정을 함께 느낌으로 인해 이 또한 경계선이 불투명 하였다. 하지만, 더 많고 정확한 데이터를 사용하고, 경계선을 좀 더 투명하게 한다면 정확도의 수치는 많이 올라 갈 것으로 판단한다.

4. 결론

본 연구는 영화평 데이터를 7개의 감정(“희(喜), 로(怒), 애(哀), 락(樂), 애(愛), 오(惡), 욕(慾)”)으로 분류 하기 위해서, 기계학습 모델을 적용하였다. Training Set을 구축하기 위해 Bag of words 기법을 이용하여 감정 자질이 될 수 있는 형태소인 VA, NNG, NNP와 TF를 사용하였다. Training Set으로 학습된 모델을 기준으로 영화평 데이터를 분류 하였다. 분류 시 보다 많고, 정확한 Training Set을 이용하면 감정 분류의 정확성이 더 높아질 것으로 보인다. 만약, 이 연구를 실제 영화관련 사업에 적용을 한다면, 사용자들은 단순히 평점만이 아닌, 직관적으로 볼 수 있는 감정 결과를 통해 쉽게 영화의 재미 정도를 알 수 있을 것으로 보인다. 이와 같은 연구를 통하여 텍스트 데이터가 많이 사용되고 있는 SNS 분야에도 응용이 가능할 것으로 기대된다.

5. 참고문헌

[1] Seung Youp Lee, Kwan Ho In Ung Mo Kim "Analyzing University Bulletin Board Data by using Opinion Mining" Korean Institute of Information Scientists and Engineers. 2012

[2] Yun-Suk Kim, Young-Hun Seo ”Hangul text sentiment classification using machine learning” Journal of Korea Entertainment Industry Association. 2013’

[3] Hanna M. Wallach"Topic Modeling : Beyond Bag of Words" Proceedings of the 23rd international Conference on Machine Learning. 2006

[4] 황재원, 고영중 "감정 분류를 위한 한국어 감정 자질 추출 기법과 감정 자질의 유용성 평가" Korean Journal of Cognitive Science. 2008

[5] Cho-Hee Hong, Hark-Soo Kim "Comparative Study of various Machine-learning Features for Tweets Sentiment Classification" The Korea Contents Associations. 2010

[6] Joa-Sang Lim, Jin-Man Kim "An Empirical Comparison of Machine Learning Models for Classifying Emotions in Korean Twitter" Journal of Korea Multimedia Society. 2014

[7] Michael W. Morris and Dacher Keltner "How Emotions Work : The Social Functions of Emotional Expression in Negotiations" Elsevier Science Inc. 2000

[8] Cheolseong Lee, Donghee Choi, Seongsoon Kim, Jaewoo Kang ”Classification and Analysis of Emotion in Korean Microblog Texts” Korean Institute of Information Scientists and Engineers. 2012’