Journal of Convergence for Information Technology
Vol. 10. No. 10, pp. 40-46, 2020
e-ISSN 2586-4440
DOI : https://doi.org/10.22156/CS4SMB.2020.10.10.040

비정형 문서에서 감정과 상황 정보를 이용한 감성 예측

김진수
안양대학교 아리교양대학 교수

Sentiment Prediction using Emotion and Context Information in Unstructured Documents

Jin-Su Kim
Professor, Ari Liberal Arts, Anyang University

요 약 인터넷의 발전으로 사용자들은 자신의 경험이나 의견을 공유한다. 영화평과 같은 비정형 문서의 전체적인 감정이나 장르 등의 정보를 고려하지 않고 연관된 키워드를 사용하기 때문에 적절한 감정 상황에 따른 감성 정확도를 저해한다. 따라서 사용자들이 작성한 비정형 문서가 속한 장르나 전반적인 감정 등의 정보를 기반으로 감성을 예측하는 시스템을 제안한다. 먼저, 비정형 문서로부터 기쁨, 화남, 공포, 슬픔 등의 감정 집합과 연관된 대표 키워드를 추출하고, 감정 특징단어들의 정규화된 가중치와 비정형 문서의 정보를 훈련 집합으로 CNN과 LSTM을 조합한 시스템에 훈련한다. 최종적으로 영화 정보와 형태소 분석기와 n-gram을 통해 추출한 정제된 단어들과 이모티콘, 이모지 등을 테스트함으로써 감정을 이용한 감성 예측 정확도와 F-measure 측면에서 향상됨을 보였다. 제안한 예측시스템은 슬픈 영화에서 슬픈 단어의 사용과 공포 영화에서 무서운 단어 등의 사용으로 인해 부정으로 판단하는 오류를 피함으로써, 감성을 상황에 따라 적절하게 예측할 수 있다.

주제어 : 감성 예측, 오피니언 마이닝, 상황 정보, 딥러닝, 자연어 처리

Abstract With the development of the Internet, users share their experiences and opinions. Since related keywords are used witho0ut considering information such as the general emotion or genre of an unstructured document such as a movie review, the sensitivity accuracy according to the appropriate emotional situation is impaired. Therefore, we propose a system that predicts emotions based on information such as the genre to which the unstructured document created by users belongs or overall emotions. First, representative keyword related to emotion sets such as Joy, Anger, Fear, and Sadness are extracted from the unstructured document, and the normalized weights of the emotional feature words and information of the unstructured document are trained in a system that combines CNN and LSTM as a training set. Finally, by testing the refined words extracted through movie information, morpheme analyzer and n-gram, emoticons, and emojis, it was shown that the accuracy of emotion prediction using emotions and F-measure were improved. The proposed prediction system can predict sentiment appropriately according to the situation by avoiding the error of judging negative due to the use of sad words in sad movies and scary words in horror movies.

Key Words : Sentiment Prediction, Opinion Mining, Context Information, Deep Learning, NLP

1. 서론

인터넷 기술의 빠른 발전으로 인해 다양한 분야에서 정보를 공유하고 다양한 의견을 서로 교환할 수 있다. 특히, 자연어로 표현된 텍스트로부터 데이터 분석을 통해 의견이나 성향 등을 예측하여 유의미한 정보를 획득하려는 오피니언 마이닝 연구가 진행되고 있다. 오피니언 마이닝은 상업 분야뿐만 아니라, 정치, 경제, 사회 분야 등에서 널리 사용되며, 수치화 및 시각화를 통해

*Corresponding Author : Jin-Su Kim(kjspace@anyang.ac.kr)
Received August 30, 2020
Revised October 3, 2020
Accepted October 20, 2020
Published October 28, 2020

직관적인 감성 판단으로 의사결정에 반영할 수 있으며, 최근까지 자연어 처리 분야에서도 활발히 연구되고 있다[1,2]. 영화평, 상품평, 인터넷 댓글, SNS 등의 내용은 작성자가 상대에게 정보나 견해를 신속하게 표현하기 위해 짧고 핵심적인 의미의 단어를 표현하여 작성한다. 신속하게 자신의 감정이나 정보 전달로 신조어나 이모지나 이모티콘 등을 많이 사용하고 오타나 파괴된 문법 등도 빈번히 발생한다. 기존의 전통적인 기계학습인 Naive Bayes, Support Vector Machine(SVM) 등을 이용하여 감성을 분석하였을 때, 학습한 특정 영역에서는 성능은 좋으나, 다른 분야에 적용하였을 때 성능이 저하됨을 보였다[3]. 최근에는 딥러닝 기술을 활용하여 단어 임베딩을 이용함으로써 단어들 간의 연관도를 측정하여 높은 정확도를 보인다[4]. 영화평에는 일반적으로 기쁜 내용이면 긍정, 슬픈 내용이면 부정으로 판별하지만, 슬픈 영화평에 “울었다”와 같은 슬픈 감정의 키워드가 있다고 부정으로 속단할 수 없다. 따라서 영화 감상평의 경우, 감성 키워드뿐만 아니라, 영화의 장르, 감독, 배우 등의 전반적인 상황 정보를 고려하여 감성을 예측하여야 한다.

본 연구에서는 사용자가 작성한 영화평과 같은 비정형 문서로부터 데이터를 획득하여 영화에 대한 감정을 예측하고 예측된 감정과 영화의 상황 정보를 고려하여 해당 영화에 대한 긍정 및 부정을 예측하는 시스템을 구축하고자 한다. 제안하는 방법은 감정 및 감성 상호작용을 위한 특징 추출을 위해 먼저 비정형 문서에 표현된 감정 키워드를 형태소분석을 통해 추출한다. 형태소 분석과정에 제외되는 이모지, 이모티콘, 신조어 등은 전처리과정에서 의미가 있는 대표 키워드로 변환하며, 맞춤법이나 띄어쓰기 오류인 경우에는 n-gram을 사용한다. 후보 감정 키워드들로부터 빈도수를 이용하여 대표 감정 키워드 및 긍정 및 부정의 감성 키워드를 CNN과 LSTM을 이용하여 감정을 예측하고 최종적으로 감성을 판별한다.

2. 관련 연구

오피니언 마이닝은 텍스트에 표현된 개인적인 느낌이나 관점, 감정, 신념 등을 분석하여 객관화 및 정량화하여 긍정 및 부정적 감성을 분석하는 것이다[5]. 다른 사용자들에게 자신의 경험, 의견, 느낌 등에 대해 좋고 싫은 다양한 감정을 표현하여 긍정적 또는 부정적 평가에 사용된다. 감성 분석을 위해 감성 사전이나 감성 의미망과 같은 자원을 수집하고 구축한다. Thayer가 제시한 감정은 긍정과 부정의 척도에 따른 밸러스와 High, Low의 활성 정도에 따라 Excited, Happy, Pleased, Relaxed, Peaceful, Calm, Sleepy, Bored, Sad, Nervous, Angry, 그리고 Annoying으로 분류한다[6]. Ekman[7]는 지역이나 문화 등에 관련 없이 남녀노소 누구나 일반적으로 느끼는 기본 감정으로 Anger, Disgust, Fear, Joy, Sadness, 그리고 Surprise로 분류한다. Ekman이 제시한 감정에서 긍정의 범주에는 Joy, 부정에 속하는 감정들은 Anger, Disgust, Fear, Sadness이며, Surprise는 긍정 및 부정에 함께 사용된다[8].

영화평, 상품평, SNS, 채팅 등의 다양한 서비스에서 표현되는 텍스트는 글자 수의 제한으로 인해 자신의 생각이나 감정 표현에 있어 기존의 언어로 표기하기 불가능한 표현과 창의적 변형 등을 통한 긍정적인 측면과 국어 어문 규정의 파괴, 문법 의식 약화 등의 부정적인 측면이 나타난다. 인터넷의 발전에 따라 빠르게 변화하고 생성되는 신조어, 약어, 준말, 외계어, 이모지, 이모티콘 등은 사용자의 감정을 표현하는 중요한 정보이기 때문에, 감정을 내포하는 신조어, 약어, 이모지 등을 활용하여 작성자가 표현하려는 감정을 유추해야 한다.

전통적인 기계학습 기법에서의 분류는 특징 집합을 따로 추출하지만, 최근의 딥러닝 신경망 구조는 복잡한 특징을 자동으로 추출하기 때문에 음성인식, 내용요약, 감성분석, 텍스트 분류 등의 자연어처리에 많이 사용한다. 딥러닝 기반의 방법에서는 CNN(Convolutional Neural Network), RNN(Recurrent Neural Network), LSTM(Long Short-Term Memory) 등을 사용하여 모델을 구성한 후 데이터 집합을 이용해 모델을 학습시키는 방식으로 텍스트나 음성 처리를 수행한다. CNN은 이미지 분류에 특화되었지만, 최근에는 자연어처리에서도 성능이 뛰어남을 보였다[9]. LSTM은 장기 의존성 문제, 즉 이전에 학습한 정보를 현재의 문제에 활용하도록 입출력 게이트와 망각 게이트로 구성된 특별한 형태의 RNN이다[10]. LSTM은 연산속도가 느리고 기존의 메모리가 덮어 쓰일 가능성은 있으나 메모리와 결과값 제어가 가능하다는 장점이 있다.

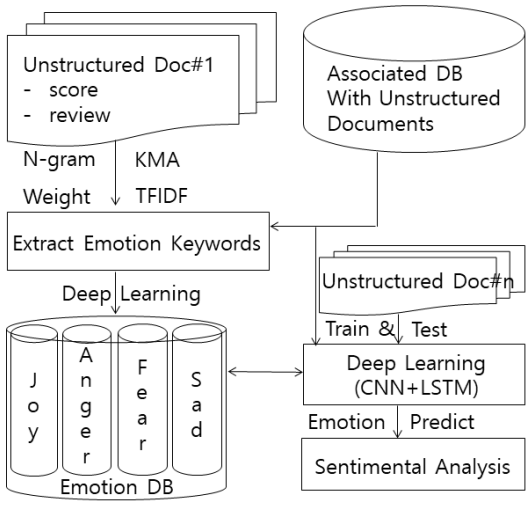

Fig. 1. Sentimental Analysis Architecture using Emotion Prediction

3. 비정형 문서에서 감정과 상황 정보를 이용한 감성 예측

비정형 문서에 표현된 감정을 이용한 감성 분석 시스템은 Fig. 1과 같이 크게 두 부분으로 구성된다. 수집된 문서들로부터 훈련 집합에 사용될 감정 데이터베이스를 생성하는 부분과 테스트 집합과 상호 정보를 적용하여 감정 예측 및 감성을 분석하는 부분이다.

감정 정보 데이터베이스를 구축하기 위해 비정형 문서에 포함된 텍스트, 평점 등의 데이터를 수집하고 전처리된 감정 문서들로부터 한국어 형태소 분석기(Korean Morphological Analyzer, KMA)를 이용하여 키워드를 추출한다. 추출된 키워드 중 가중치를 이용하여 대표 키워드를 선별하여 감정에 따른 단어들의 연관 정도를 딥러닝을 적용하여 데이터베이스를 생성한다. 이때, 형태소분석에서 제외된 단어들은 n-gram을 통해 신조어 데이터베이스를 만들어 추가한다. 생성된 감정 데이터베이스와 CNN을 이용하여 콘텐츠 정보를 고려한 감정별로 훈련 문서를 훈련한다. 마지막으로 예측된 감정과 콘텐츠 정보를 고려하여 리뷰에 포함된 극성을 LSTM을 통해 예측한다.

3.1 감정 데이터베이스 생성을 위한 전처리 과정

비정형 문서에 내포된 감정을 판단하기 위해서는 먼저 감정별 대표 키워드를 추출한다. 문서는 자연어, 숫자, 이모지, 이모티콘 등의 다양한 문자 기호들의 조합으로 구성된다. 특히, 네티즌이 작성한 영화평의 경우, 문서의 길이가 짧고 문법파괴, 띄어쓰기 미준수, 약어, 신조어 사용, 이모지 등을 사용하여 자신의 느낌을 신속하게 전달한다. 먼저, 형태소 분석기를 이용하여 의미를 지닌 키워드를 일차적으로 추출한다. 짧은 글을 통해 자신의 감정을 대중에게 실시간으로 정확하게 표현하기 위해 약어, 신조어, 이모지, 외계어 등의 인터넷 언어는 형태소분석에서는 추출되지 않지만, 감정에 영향을 주는 키워드들도 있다. 먼저 인터넷 언어들은 축적된 문서의 전처리과정에서 축약 이전의 문장이나 감정 대표 단어로 변환한다. 이러한 과정을 거친 후에도 처리되지 않은 띄어쓰기 오류, 단어 우월 효과, 오타 등의 경우에는 n-gram이나 음절의 조합을 통해 감정에 사용되는 키워드를 추출한다. 음절의 조합이나 n-gram을 적용하여 의미있는 키워드를 찾는 것은 전체적인 부하 증가로 인해 성능을 저하하지만, 최근 처리장치의 빠른 발전을 통해 해결할 수 있다. 문서에 나타난 감정을 지닌 키워드를 추출하기 위해 KoNLPy[11]를 사용하며, 명사, 동사, 형용사, 감탄사 등을 추출하고, 조사, 한정사 등은 추출에서 배제한다. 각 감정을 대표하는 키워드를 먼저 추출하여 감정을 예측하는 것은 시간적 측면이나 공간적 측면에서 효율적이므로 성능을 향상하기 위해 매우 중요하다. 즉, 대표 키워드 추출은 감정 예측을 위해 훈련에 필요한 차원을 감소시키는 데 유용하게 사용된다. 따라서 본 연구에서는 비정형 문서의 감정을 대표하는 키워드의 빈도 및 감정 가중치 기반의 키워드를 추출하고 감정 데이터베이스를 구성한다. 영화평의 경우 긴 문서와 달리 짧은 문단으로 작성되어 역문단빈도수($TF \cdot IPF$)[12]를 사용하여 각 감정 집합의 대표 감정 키워드를 추출한다. 역문단빈도수는 대표 키워드 추출을 위한 단어 빈도(TF, Term Frequency)와 역문단빈도수(IPF, Inverse Paragraph Frequency)의 곱으로 식 (1)과 같다.

$$TF \cdot IPF = TF \cdot [\log_2(N) - \log_2(PF) + 1] \quad (1)$$

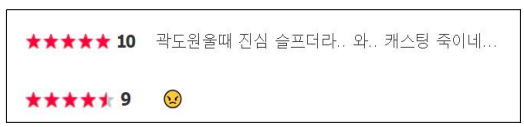

Fig. 2. Example of Movie reviews

함축된 영화평으로부터 긍정 및 부정적 감성을 판별

하기 위해 작성자가 평가한 점수를 사용하며, 평점이 1~4점이면 부정, 5~8점은 중립, 9~10점은 긍정으로 판별하여 훈련하면 약간의 예측 오류가 발생한다. 일반적인 기본 감정과 긍정과 부정의 감성과의 관계는 대체로 일치하지만, 장르에 따라 댓글의 내용 및 평가가 달라질 수 있다.

Fig. 2의 첫 번째 예는 영화 "남자가 사랑할 때"의 리뷰이다. 문서의 내용만을 가지고 감정을 판단한다면, '슬프더라'와 '죽이네'는 슬픈 감정 또는 두려움 감정의 부정적 단어로 구성되어 있다. 이를 긍정으로 인식(10점)하고 훈련 집합으로 사용한다면 긍정적인 측면의 오류가 높아질 수 있다. 그러나 영화 콘텐츠의 장르가 슬픈 드라마라는 정보를 고려하면 평가글의 슬픈 감정은 긍정적인 극성으로 볼 수 있다. 콘텐츠의 장르나 전체적인 감정을 고려하여, 평가글의 내용을 분석하면 오류가 아닌 정상적인 극성을 예측할 수 있다. 따라서 콘텐츠 의견의 경우 감정 및 극성을 예측하기 위해서는 콘텐츠의 다양한 정보를 고려하여 댓글의 내용을 분석해 감정을 판별해야 한다. 두 번째 예의 경우, 평가한 평점이 9점인 이모지로 표현한 경우, 평점은 긍정이지만, 이모지의 감정은 Anger인 부정적 감성이기 때문에 긍정과 부정 사이에 감성 혼란이 발생할 수 있다. 감정을 가진 인터넷 언어로 구성된 신조어 데이터베이스를 생성하고 전처리과정에서 변환한다. 예를 들어, 유니코드로 작성된 이모지(😀(grinning face, 😀) 😠(angry face, 😠))나 이모티콘(ღ, ღ)은 콘텐츠의 감정을 판별하는데 중요한 의미를 가진다. 이뿐만 아니라, '죠타', '행복햐다' 등의 오타나 단어 우월 효과와 같은 음절의 전치로 인해 감정 판별의 중요한 키워드를 학습에 사용하지 못하는 경우가 있다. 이러한 문제를 해결하기 위해 음절의 재조합과 n-gram 기법을 통해 신조어 데이터베이스에 추가하고, 오타로부터 변환된 정상적인 키워드는 다시 형태소 분석과정을 거친다. 따라서 콘텐츠 의견과 같은 짧은 문단으로 작성된 텍스트는 사용자들은 자신의 감성을 표현하기 위해 연역적 혹은 귀납적 방법 등 다양한 방법으로 설명하기 때문에 키워드 간의 연관성과 신조어까지 고려하여 감정별 데이터베이스를 생성한다. 대표 키워드에 가중치를 적용하며 식(2)와 같다.

$$Weight_{key} = Norm(score) \bullet \frac{len(key))*freq(key)}{len(paragraph)} \quad (2)$$

영화평의 평점은 일반적으로 10점까지의 값으로 콘텐츠에 대해 긍정 및 부정적 평가를 한다. 평점의 폭이 크지 않기 때문에 계산 부하가 적은 최대-최소 정규화를 사용하여 0과 1 사이 값으로 변환한다. 짧은 글로 작성자의 감정을 압축하여 표현하기 때문에 기존의 긴 문서와는 달리 키워드의 가중치를 재계산하여 감정을 표현하는 키워드에 적용한다. 즉, 영화 평점의 정규화 점수($Norm(score)$), 키워드의 길이($len(key)$), 키워드의 빈도수($freq(key)$), 그리고 영화 리뷰의 전체 길이($len(paragraph)$)를 사용하여 키워드에 대한 가중치($Weight_{key}$)를 재계산한다.

3.2 딥러닝을 적용한 감정 및 감성 예측

감정을 예측하기 위해 감정 데이터베이스를 이용하며, 각 감정 예측에 필요한 단어들 사이의 연관된 특징들을 생성하기 위해 CNN과 LSTM을 조합하여 학습한다. 딥러닝에 사용되는 특징은 정규화된 평점, 콘텐츠 의견에 포함된 대표 키워드, 그리고 해당 콘텐츠의 정보를 혼합하여 각 감정별 키워드들의 연관성이 내재된 훈련 집합을 구성한다. 이때 콘텐츠의 감정 정보는 수동으로 작성하며, 각 감정별 연관된 키워드의 가중치를 적용하여 새로운 콘텐츠 의견에 대한 감정 및 감성을 판단한다. 단어 임베딩을 수행하여 감성 분석에서 문장간 유사성과 감정, 긍정 그리고 부정의 레이블을 훈련하고, 단어 임베딩의 출력을 통해 텍스트별 벡터를 만드는 과정에서 딥러닝을 적용한다. 감정을 예측하기 위해 CNN을 사용하며, Convolution layer에서 kernel의 크기를 조정하여 다른 단어와의 연관 관계 구조를 파악하고 지역적 특징을 추출하기 쉽게 한다. 각 단계마다 경사 하강법 소실 문제를 완화하기 위해 ReLU를 활성화 함수로 사용하며, 각 과정에서 ReLU를 진행한 후, Max pooling으로 Convolution Layer의 결과를 축소하고 과적합 방지를 위해 Drop out 과정을 거친다. CNN을 통해 예측한 감정과 LSTM을 통해 예측한 감성을 이용하여 최종적인 감성을 판별한다.

4. 실험 및 분석

영화평과 같은 비정형 문서로부터 감정 및 감성을

예측하기 위해 CNN-LSTM을 혼합하여 감정 분석 정보에 필요한 특징을 추출하고, 추출된 감정과 콘텐츠의 장르 및 상황 정보 등을 적용하여 감정 및 감성을 예측하는 시스템을 구축한다. 다양한 감정 텍스트들로부터 분류된 감정 집합들과 감정 대표 키워드들을 CNN의 훈련 집합으로 입력하기 위해 콘텐츠 의견들과 해당 콘텐츠의 기본적인 상황 정보를 수집한다. 본 논문에서는 Joy, Anger, Fear, Sad 의 4가지 감정만을 분류하며, Surprise는 긍정도 부정도 아닌 선택적 감정이고, Disgust는 표현 빈도가 낮게 분포되어 제외한다. 네이버와 다음 영화 사이트로부터 무작위로 영화를 선별하여 감정별로 300개의 네티즌 영화평을 분류한 후, 각 감정 문서 중 무작위로 추출한 80%의 문서는 훈련 집합으로 사용하고, 감정별 20%의 문서는 테스트 집합으로 사용한다. 이러한 방법으로 5회 반복하여 테스트하였다. 훈련에 사용될 문서의 전처리과정에 사용되는 단어는 네이버 오픈 사전에 있는 일반어, 유행어, 신조어, 사투리, 채팅어 등과 유니코드 이모지[13]를 감정 및 신조어 데이터베이스에 구축에 활용한다. 감정 및 감성을 예측하기 위한 환경은 Python에서 keras로 CNN, LSTM 모델을 사용하며, 형태소분석을 위해 KoNLPy를 사용한다. CNN을 이용한 감정을 예측하기 위해 128개의 임베딩 차원, 128 크기의 필터 4개, 배치 크기는 64, 에포크는 200, 드롭아웃은 0.5로 설정한다. 긍정 및 부정 감성을 예측하기 위해 LSTM은 128개의 뉴런, 배치 크기는 100, 에포크는 10, 그리고 활성화 함수로 sigmoid를 사용하며, 4개의 감정을 예측하기 위해 CNN에서는 Softmax를 사용한다.

Fig. 3. Prediction of Emotion and Sentimental Analysis in review

Fig. 3은 특정 영화평("인생의 마지막을 담담하고도 아름답게 준비하는 과정을 담은 영화. 너무 감동적이에요")에 대한 감정을 수치화하고 감성 예측한 결과를 보인다. '아름답게', '감동적' 등의 단어에 의해 기쁜 감정이 높게 예측되고, 이를 통해 긍정의 결과를 보여준다.

감정 예측을 위한 평가 척도로 정확도(Accuracy)를 사용하고, 긍정 및 부정의 감성을 예측에는 F1-measure를 사용한다.

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F-measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

여기서, TP, TN, FP, 그리고 FN은 각각 True Positive, True Negative, False Positive, 그리고 False Negative를 나타낸다.

Table 1. Comparison of Emotion Prediction Accuracy(%)

Emotion	SVM	Proposed Method - (n-gram & emoji)	Proposed Method
Joy	81.4	83.2	83.3
Anger	81.8	83.4	84.5
Fear	79.3	86.5	87.4
Sad	83.1	88.1	88.6

Table 1은 SVM을 이용한 전통적인 방법과 제안한 방법 중 n-gram과 이모지를 제외하고 딥러닝 하였을 때의 감정 예측 정확도를 비교한 결과이다. 전체적으로 제안한 예측시스템의 평균 성능이 SVM, n-gram과 이모지를 적용하지 않았을 때를 비교했을 때, 약 4.55%과 0.65% 향상되었다.

Table 2. Comparison of F-measure(%)

	SVM	Proposed System (- movie info)	Proposed Method
Positive	90.4	93.7	94.5
Negative	91.3	93.5	93.7

Table 2는 영화평에 대한 감성 예측을 정밀도와 정밀도를 이용하여 F-measure를 계산한 결과를 보여준다. 전통적인 방법 중 하나인 SVM을 이용한 감성 예측과 관련 콘텐츠의 감정 정보를 고려하지 않았을 때를 비교하면, 제안한 시스템의 긍정 예측은 각각 4.1%, 0.8%, 부정 예측은 2.4%, 0.2% 높았다. 이러한 결과는 높고 낮은 평점이 아닌 중립적인 극성에서 콘텐츠의 내용과 콘텐츠 정보를 고려함으로써 긍정 감성을 높이고 훈련

집합에 사용되는 문서들의 잡음을 줄여 정확도를 높일 수 있었다. 즉, 부정적 감성을 지닌 감정의 감성 예측 오류를 해당 영화의 전체적 감정과의 연관성을 고려함으로써 효과적인 감성 예측에 도움을 준 것으로 판단된다.

5. 결론 및 향후 연구

영화평, 댓글, 상품평 등과 같은 비정형 문서로부터 딥러닝을 활용하여 감정을 통해 긍정적이거나 부정적인 감성의 성향을 파악할 수 있다. 본 연구에서는 영화가 지닌 상황 정보와 영화평을 딥러닝으로 분석하여 감정 및 감성을 예측한다. 먼저 최종적인 감성 예측을 위해 각 감정별로 수집한 비정형 문서들로부터 형태소를 분석하여 감정 키워드를 추출하고 정규화된 연관 가중치를 계산한다. 이때 신조어, 오타, 띄어쓰기 등과 같은 오류인 경우에도 신조어 데이터베이스와 n-gram방식을 적용하여 감정 의미를 내포한 키워드를 추출한다. 또한, 비정형 문서의 평가된 평점만을 고려하지 않고, 영화 자체의 감정이나 장르 정보를 영화 게시글과의 감정 연관성을 비교하여 정규화된 감정 키워드 벡터를 생성한다. 대용량의 순서를 가진 감정 단어들의 데이터 처리에 적합한 CNN-LSTM을 조합하여 감정을 예측하고 감성을 판별함으로써, 기존의 전통적인 방법에 비해 정확도와 F-measure에서 성능이 향상되고, 영화평 자체의 감정만으로 감성을 파악하지 않고 영화에 대한 정보를 고려함으로써 감성 예측에 효과적임을 보였다. 그러나 영화 정보를 수동으로 삽입하고, CNN과 LSTM의 혼합된 딥러닝, 형태소분석, n-gram, 그리고 글자 조합을 통해 오타를 수정하는 과정에서 시스템의 부하가 증가함으로써 성능을 저하시키는 요인이 발생하였다.

향후에는 영화의 감정 정보를 시나리오나 영상 인식 등을 통해 자동화 및 효율적인 감정 대표 키워드의 가중치를 통해 감정 예측을 높이고 감정에 따른 감성 정확도를 향상하고 비정형 문서에 나타난 감정 및 감성의 변화와 영화의 흥행도, 시간 흐름과의 상호 관계를 연구하고자 한다.

References

[1] S. D. Kim, E. B. Park, S. J. Lee & K. Y. Kim. (2010). A Syllable Kernel based Sentiment Classification for Movie Reviews. *Journal of Korean Institute of Intelligent Systems, 20(2),* 202-207.
DOI : 10.5391/JKIIS.2010.20.2.202

[2] K. Y. Kim & C. S. Kim. (2009). A String Kernel based Sentiment Classification for Blog Text. *Proceedings of KIIS Fall Conference 2009, 19(2),* 199-201.
DOI : 10.5391/JKIIS.2012.22.5.563

[3] S. Seo & J. Kim, (2016). Sentiment Analysis Research Trend Based on Deep Learning. *The Korea Multimedia Society, 20(3)*, 8-22.

[4] A. Rexha, M. Kröll, M. Dragoni & R. Kern. (2016). Polarity Classification for Target Phrases in Tweets: A Word2Vec Approach. *ESWC 2016. LNCS, 9989,* 217-223.
DOI : 10.1007/978-3-319-47602-5_40

[5] M. Kang, J. Ahn & K. Lee. (2018). Opinion mining using ensemble text hidden Markov models for text classification. *Expert Systems with Applications, 94,* 218-227.
DOI : 10.1016/j.eswa.2017.07.019

[6] R. Thayer. (1989). The Biopsychology of Mood and Arousal. *Oxford University Press.*

[7] K. R. Scherer & P. Ekman. (2014). Approaches to Emotion. *Psychology Press, New York.*

[8] M. Chang. (2012). Empirical Sentiment Classification Using Psychological Emotions and Social Web Data. *Journal of Korean Institute of Intelligent Systems, 22(5),* 563-569.
DOI : 10.5391/JKIIS.2012.22.5.563

[9] Y. Kim. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746-1751.
DOI : 10.3115/v1/D14-1181

[10] F. Abid, M. Alam, M. Yasir & C. Li. (2019). Sentiment analysis through recurrent variants latterly on convolutional neural network of Twitter. *Future Generation Computer Systems, (95),* 292-308.
DOI : 10.1016/j.future.2018.12.018

[11] E. Park & S. Cho. (2014). KoNLPy: Korean natural language processing in Python. *Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology, Chuncheon, Korea.* 133-136.

[12] J. Kim. (2014). Emotion Prediction of Document using Paragraph Analysis. *Journal of Digital Convergence, 12(12),* 249-255.
DOI : 10.14400/JDC.2014.12.12.249

[13] *Unicode Emoji*. http://www.unicode.org/reports/tr51

김 진 수(Jin-Su Kim) [정회원]

· 1998년 2월 : 인천대학교 전자계산공학과 (학사)
· 2001년 8월 : 인하대학교 컴퓨터공학과 (석사)
· 2010년 2월 : 인하대학교 정보공학과 (박사)

· 2011년 3월 ~ 현재 : 안양대학교 아리교양대학 교수
· 관심분야 : 융복합, 딥러닝, 빅데이터, 자연어처리
· E-Mail : kjspace@anyang.ac.kr