



Word2vec과 앙상블 분류기를 사용한 효율적 한국어 감성 분류 방안

Effective Korean sentiment classification method using word2vec and ensemble classifier

저자 (Authors)	박성수, 이건창 Sung Soo Park, Kun Chang Lee
출처 (Source)	한국디지털콘텐츠학회 논문지 19(1) , 2018.1, 133-140 (8 pages) Journal of Digital Contents Society 19(1) , 2018.1, 133-140 (8 pages)
발행처 (Publisher)	한국디지털콘텐츠학회 Digital Contents Society
URL	http://www.dbpia.co.kr/Article/NODE07321791
APA Style	박성수, 이건창 (2018). Word2vec과 앙상블 분류기를 사용한 효율적 한국어 감성 분류 방안. 한국디지털콘텐츠학회 논문지, 19(1), 133-140.
이용정보 (Accessed)	서울대학교 147.46.182.*** 2018/03/26 14:07 (KST)

저작권 안내

DBpia에서 제공되는 모든 저작물의 저작권은 원저작자에게 있으며, 누리미디어는 각 저작물의 내용을 보증하거나 책임을 지지 않습니다. 그리고 DBpia에서 제공되는 저작물은 DBpia와 구독 계약을 체결한 기관소속 이용자 혹은 해당 저작물의 개별 구매자가 비영리적으로만 이용할 수 있습니다. 그러므로 이에 위반하여 DBpia에서 제공되는 저작물을 복제, 전송 등의 방법으로 무단 이용하는 경우 관련 법령에 따라 민, 형사상의 책임을 질 수 있습니다.

Copyright Information

Copyright of all literary works provided by DBpia belongs to the copyright holder(s) and Nurimedia does not guarantee contents of the literary work or assume responsibility for the same. In addition, the literary works provided by DBpia may only be used by the users affiliated to the institutions which executed a subscription agreement with DBpia or the individual purchasers of the literary work(s) for non-commercial purposes. Therefore, any person who illegally uses the literary works provided by DBpia by means of reproduction or transmission shall assume civil and criminal responsibility according to applicable laws and regulations.

Word2vec과 앙상블 분류기를 사용한 효율적 한국어 감성 분류 방안

박 성 수 · 이 건 창*

성균관대학교 경영대학

Effective Korean sentiment classification method using word2vec and ensemble classifier

Sung Soo Park · Kun Chang Lee*

SKK Business School, Sungkunkwan University

[요 약]

감성 분석에서 정확한 감성 분류는 중요한 연구 주제이다. 본 연구는 최근 많은 연구가 이루어지는 word2vec과 앙상블 방법을 이용하여 효과적으로 한국어 리뷰를 감성 분류하는 방법을 제시한다. 연구는 20 만 개의 한국 영화 리뷰 텍스트에 대해, 품사 기반 BOW 자질과 word2vec를 사용한 자질을 생성하고, 두 개의 자질 표현을 결합한 통합 자질을 생성했다. 감성 분류를 위해 Logistic Regression, Decision Tree, Naive Bayes, Support Vector Machine의 단일 분류기와 Adaptive Boost, Bagging, Gradient Boosting, Random Forest의 앙상블 분류기를 사용하였다. 연구 결과로 형용사와 부사를 포함한 BOW자질과 word2vec자질로 구성된 통합 자질 표현이 가장 높은 감성 분류 정확도를 보였다. 실증결과, 단일 분류기인 SVM이 가장 높은 성능을 나타내었지만, 앙상블 분류기는 단일 분류기와 비슷하거나 약간 낮은 성능을 보였다.

[Abstract]

Accurate sentiment classification is an important research topic in sentiment analysis. This study suggests an efficient classification method of Korean sentiment using word2vec and ensemble methods which have been recently studied variously. For the 200,000 Korean movie review texts, we generate a POS-based BOW feature and a feature using word2vec, and integrated features of two feature representation. We used a single classifier of Logistic Regression, Decision Tree, Naive Bayes, and Support Vector Machine and an ensemble classifier of Adaptive Boost, Bagging, Gradient Boosting, and Random Forest for sentiment classification. As a result of this study, the integrated feature representation composed of BOW feature including adjective and adverb and word2vec feature showed the highest sentiment classification accuracy. Empirical results show that SVM, a single classifier, has the highest performance but ensemble classifiers show similar or slightly lower performance than the single classifier.

색인어 : 감성분류, 자질 표현, Bag-of-word(BOW), word2vec, 앙상블 분류기

Key word : Sentiment classification, feature representation, Bag-of-words(BOW), word2vec, ensemble classifier

<http://dx.doi.org/10.9728/dcs.2018.19.1.133>



This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received 20 October 2017; **Revised** 08 November 2017
Accepted 29 January 2018

***Corresponding Author; Kun Chang Lee**

Tel: +82-2-760-0505

E-mail: kunchanglee@gmail.com

I. 서 론

감성분석(Sentiment Analysis)은 텍스트로 표현된 사람들의 의견, 정서, 태도 및 감정을 분석하는 연구 분야이다[1]. 블로그, 리뷰, 포럼 및 소셜 네트워크 서비스와 같이 사용자가 작성한 독창적인 콘텐츠가 폭발적으로 증가함에 따라 감성분석의 활용이 점점 더 중요 해지고 있으며, 자연어 처리(Natural Language Processing) 분야에서 가장 활발한 연구 분야 중 하나이다. 감성 분석은 사람들이 기술한 의견에 대한 잠재적인 정보를 담고 있는 자료에서 주관적인 정보를 확인하는 과정을 의미한다[2]. 여러 가지 감성분석의 과제 중에서 중요한 과제의 하나는 텍스트가 내포하고 있는 긍정, 부정의 정서를 분석하는 감성 분류이다[3].

감성 분류 기술은 상품, 서비스에 대한 사용자 리뷰 등을 긍정적 인 등급과 부정적 등급으로 자동으로 분류하는 것으로, 최근에는 기계학습 기반의 감성 분류가 탁월한 성능으로 널리 사용되고 있다. 기계학습 모형에 사용되는 자질(Feature) 추출 방법은 단순한 기능보다는 복잡한 자질을 추출하고, 어떤 유형의 자질이 더 가치 있는지를 찾는 것은 기계학습 기반 감성 분류에서 두 가지 중요한 문제이다[4].

전통적인 감성 분류는 문서에 포함된 단어의 출현 여부나 빈도를 One-Hot 벡터로 표현하는 방식이 많이 사용되고 있다. 단일 단어(unigram), 다중 단어(n-gram), TF-IDF 등 다양한 BOW(Bag-of-Words) 자질 추출 방법이 제안되었다[5]-[7]. 그러나 BOW는 단어순서가 제거되어 같은 단어가 사용되는 다른 의미의 문장이 동일한 자질로 표현되어[8], 사람들의 감성을 표현하는 맥락이 생략되어 잠재적인 감성 표현을 포착하지 못하는 단점을 가지고 있다. 최근의 연구는 벡터 공간에 단어를 표현하는 워드 임베딩(Word Embedding)이 많이 연구된다[9][10].

기계 학습 알고리즘을 사용하여 감성을 분류하는 많은 연구가 수행되었다. 예를 들어 SVM(Support Vector Machines)과 NB(Naive Bayes), NN(Neural Networks) 등이 감성을 식별하는데 일반적으로 사용된다[11][12]. 최근 몇 가지 기본 분류 방법의 출력을 결합해서 분류 출력을 형성으로 분류 정확도를 향상시키는 앙상블 학습 방법의 사용에 대한 관심이 높다[13][14].

영어 및 중국어 분야는 워드 임베딩과 앙상블 학습기를 사용한 많은 연구가 진행되고 있지만[15][16], 한국어 감성 분류에 대한 실증 연구는 전무한 실정이다. 이에 본 연구는 워드 임베딩과 앙상블 학습기를 사용하여 효과적인 한국어 감성 분류 방법을 제안한다. 연구를 통하여 다음과 같은 연구문제를 밝힌다.

RQ1. 한국어 감성분류에서 BOW와 word2vec 중 더 효과적인 자질 추출 방법은 무엇인가?

RQ2. 또한, BOW와 word2vec을 결합한 자질 추출이 더 효과적인 감성분류 성능을 나타내는가?

RQ3. 한국어 감성분류에서 단일 분류기와 앙상블 분류기 중 더 효과적인 분류 방법은 어떤 것인가?

II. 관련 연구

2-1 단어 벡터 표현(Word Vector Representation)

단어의 벡터로 표현하는 대표적인 방법은 단어를 하나의 벡터로 나타내는 것으로, 텍스트에 단어가 포함되면 1로 표현하고, 누락 된 단어는 0으로 나타내는 것이다. 이 방법을 BOW라 하고 감성 분류를 위한 자질 표현 방법으로 널리 사용되고 있다. Pang et al.은 감성 분류를 위해 감성 자질을 포함하는 BOW를 구성하고, 기계 학습 방법을 적용한 긍정, 부정 감성 분류 연구를 수행하였다[11]. 이 연구는 700 개의 긍정 영화 리뷰와 700 개의 부정 영화 리뷰에서 감성적인 단어를 추출하고 정서 분류를 수행했다. 연구는 BOW의 감성 분류에 대한 입력으로 Unigram, Bigram, 품사 및 단어 위치를 사용하였다. 또한, 기계 학습의 성능을 최대화하도록 다양한 BOW자질을 확장 구성하는 방법에 대한 연구도 수행되었다[5], [6]. BOW 방법은 간단하고 강력하지만 두 가지의 주요한 단점이 있다. 한 가지는 “행복”과 “사랑”과 같은 단어 사이의 관계를 파악 할 수 없고, 다른 하나는 텍스트에 포함 된 비표준 단어로 인한 BOW의 크기 증가 및 희박성(Sparsity) 문제로 인해 견고한 감성 분류기를 만드는 것이 어렵다[17].

BOW 방법과 다르게 단어 임베딩은 단어의 분산 표현으로 [18], 신경 확률론적 언어 모델에 기초한 학습 벡터 표현을 사용한다. 단어 임베딩 방법의 하나인 Word2vec은 신경망 기반의 알고리즘이다[10]. Word2vec은 텍스트 내의 단어를 벡터 공간으로 투영하는 단어 임베딩을 생성하는 데 특히 효율적인 계산 모델이다. Word2vec 알고리즘은 그림 1과 같이 단어의 벡터 표현을 학습하는 두 가지 모델을 제공한다. 연속적인 BOW(CBOW; continuous Bag-of- Words) 모델은 주변 단어의 맥락을 기반으로 목표 단어를 예측하고, Skip-gram은 현재 단어의 주변 단어를 예측하는 모델이다. CBOW 또는 Skip-gram 알고리즘을 텍스트에 적용하면 텍스트의 단어를 D차원 공간의 벡터로 표현할 수 있다. Word2vec은 단어 자체의 의미를 벡터 형식으로 표현하는 방법으로 복잡한 개념 표현은 물론 다른 단어를 추론하는 추론을 구현할 수 있다.

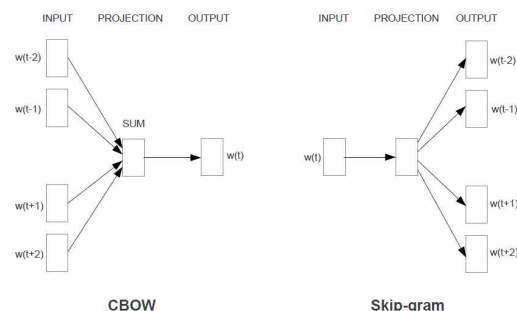


그림 1. Word2vec 모형 아키텍처
Fig. 1. Word2vec model architecture

Zhang et al.은 word2vec과 SVM을 사용하여 제품 리뷰 텍스트에 대한 소비자의 감성 분류를 제안하였다[3]. 제안된 방법은 어휘(Lexicon) 기반 자질, 품사 기반 자질을 추출하고, word2vec의 Skip-gram을 사용하여 단어의 벡터 표현자질을 학습했다. 연구 결과, word2vec을 자질로 하고 SVM을 사용하는 감성 분류 방법이 가장 높은 성능을 보였다. 또한, word2vec 자질과 어휘집 자질을 결합한 하이브리드 벡터도 연구되었다[2]. 하이브리드 벡터는 BOW와 워드 임베딩의 단점을 보완해서, 텍스트의 맥락 및 의미론적 표현을 포함하여 문서의 전반적인 감성의 방향을 캡처하는 벡터 표현이 만들어진다. 결과적으로 하이브리드 벡터를 사용한 감성 분류 방법이 사전 기반 및 워드 임베딩 자질에 비해 높은 정확도를 보였다.

2-2 한국어 감성 분류 연구

한국어로 작성된 텍스트의 감성 분류를 위해 다양한 감성 분석 연구가 수행되었습니다. 표 1에 한국어 감성 분류에 관한 최근의 연구를 요약하였다. Ahn et al.[19]은 두 문자열 간의 차이 계산에 레빈슈타인 거리를 적용하여 BOW를 생성하고, 이를 SVM과 NB의 입력으로 감성 분류기를 학습하였다. 연구결과 레빈슈타인 거리가 3일 때 다항NB의 감성분류 성과가 가장 높았고, 제안된 방법이 철자 오류에 대해 견고한 감성분류가 가능했다. Kim et al.[20]은 BOW를 입력으로 하는 다층 감성 분류기를 제안했다. 이 방법은 SVM과 NN의 두 단계의 지도학습을 기반으로, 영화 리뷰의 감성 점수를 예측한다. 제안된 감성 분류기가 단일 NB 분류기를 능가하는 결과를 보였다. Jung et al.[21]은 한국어 트위터 메시지에서 다양한 감정을 분류하기 위해 지도학습에 기반한 감정 분석 연구를 수행했다. Unigram, Bigram, Trigram, POS의 수, 감성 사전 및 낱글자 수와 같은 다양한 언어 자질의 조합을 평가하고, SVM으로 한국어 감성 분류기를 구성하여 감성 분류에 최적화된 자질을 도출하였다. 이 방법으로 얻은 F 값은 25 개의 감성 등급에 대해 73.1~98.0 %의 범위를 보였다. 대부분의 감성 등급에 최적화 된 자질은 일반적으로 단어 Bigram, POS Unigram 및 낱글자 Unigram 자질이 선정되었다.

Word2vec를 이용한 한국어 감성 분석은 주로 관련 단어 분석을 위해 수행되었다. Lee et al.[22]는 자동차 품질 이슈를 밝히기 위해, 비정형적인 텍스트에 대한 상관된 정보를 추출하는 것을 목적으로 word2vec을 적용해 차량 품질과 자동차 부품간의 유사도가 높은 단어를 발견했다. 그 외 한국어 감성분석에서 Word2vec은 주로 효율적인 감성사전 구축 연구에 많이 적용되었다. 다양한 연구에서 단어들 간의 연관성을 파악하는데 많이 사용되는 PMI(Point-wise Mutual Information)의 한계를 극복하고자 word2vec에 기반한 단어들 간의 관계를 파악하는 방법을 제안했다[23], [24].

이상의 연구들을 정리하면, 한국어에 대한 감성 분석 연구는 BOW 자질을 기반으로 다양한 기계 학습 방법을 사용해 감성을 분류하는 연구가 이루어졌지만, word2vec으로 벡터화된

감성 자질을 구성하여 감성을 분류하는 연구는 거의 이루어지지 않았다. 또한, 다양한 모델의 결과를 통합하는 양상블 학습 방법을 분류기로 적용한 감성 분류 연구는 없다.

표 1. 최근 한국어 감성분류 연구 요약

Table 1. Recent studies for Korean sentiment classification

Author	Year	Source Data	Feature set	Methods
Ahn et al. [19]	2013	Movie review	BOW	SVM, NB
Kim et al. [20]	2016	Movie review	BOW	SVM, NN
Jung et al. [21]	2017	Twitter message	BOW(uni/bi/tri-gram), # of POS, emotion lexicon, # of Characters	SVM
Lee et al. [22]	2017	Vehicle Quality Assessment text	BOW(TF-IDF)	NB, SVM, RF
Kim et al. [23]	2017	Movie review	BOW	NB
Heo et al. [24]	2017	Movie review	BOW(Lexicon)	NB, LR, NN, RF

III. 연구방법

영어[15], 그리스어 [2] 및 중국어 [3]와 같은 다양한 언어에 대해 Word2vec을 감성 자질로 사용하는 감성 분류 연구가 높은 성능을 나타내고 있다. 본 연구는 기존의 연구에 따라, 한글 리뷰의 감성 분석에 word2vec과 기계 학습을 접목하고 효율적인 감성 분석 방법을 제안하는 것을 목적으로 한다. 본 연구는 그림 2와 같이 5단계로 구성된다.

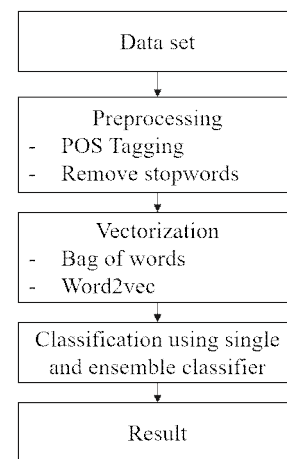


그림 2. 본 연구의 블록 다이어그램

Fig. 2. Block diagram of the proposed approach

1 단계: Maas et al.[9]의 방법으로 제작된 한국 영화 리뷰 데이터 코퍼스(1)를 연구에 사용하였다. 원 데이터는 15만 건의 학습

1) <https://github.com/e9t/nsmc/>

데이터와 5만건의 테스트 데이터로 구성되어 있으나, 연구 목적에 따라 두 데이터를 통합하여 20만 건의 리뷰 데이터로 연구를 진행하였다.

2 단계: 텍스트의 전처리 단계로, 한국의 자연 언어 처리를 위한 오픈 소스 패키지 인 Konlpy[25]를 사용하여 수집 된 한국 영화 리뷰를 단어로 분리하고 적절한 품사를 태깅한다. 품사 태깅 후, 잡음을 발생 시키거나 감성 분석에 영향이 없는 불필요한 텍스트를 제거한다. 제거된 텍스트는 다음과 같다.

- 중지 단어(Stop words): 외래 문자, 전자 메일, 어미, 조사, 해시 태그, 선어말어미 및 접미사와 같이 감성분류에 영향이 없는 어휘를 제거한다.
- 숫자 및 특수 문자: 영화 리뷰에 포함된 숫자와 마침표, 쉼표 등의 특수 문자를 제거한다.

3 단계: 전처리 된 리뷰 텍스트를 숫자 벡터화 한다. 두 가지 벡터화 방법을 사용하여 한글 리뷰를 감성 자질로 변환한다.

- BOW자질: 텍스트에 의미 있는 단어 모음을 만들고 리뷰 텍스트에서 단어의 출현 빈도를 계산하여 행렬로 변환한다. 이 과정에서, 자질 벡터에 포함되는 품사에 따라 분석의 결과가 달라지게 된다. 본 연구는 이전 연구에서 주요한 자질로 다루어 진 형용사, 부사, 동사와 같은 품사를 연구에 사용했다[3]. 또한, 선정된 품사들을 두 가지 다른 조합으로 BOW자질을 생성했다.

1) 형용사와 부사 BOW : BOW(A, D)

2) 형용사, 부사 및 동사 BOW : BOW(A, D, V)

- Word2vec자질: Word2vec는 비지도 학습적인 워드 임베딩 방법으로, 주어진 리뷰 텍스트에서 단어의 동시 발생을 모델링하여 단어 사이의 의미 및 의미 관계를 나타내는 벡터 공간이 생성한다. 전처리 과정을 거친 모든 리뷰와 단어들을 word2vec를 사용하여 학습했다. Word2vec의 모델 중 Skip-gram이 의미론적 정확도가 높고 학습 시간이 짧기 때문에[10], 본 연구는 CBOW 대신에 Skip-gram을 적용했다. 자질 차원은 300으로 설정하고, 다른 파라미터는 기본값으로 설정했다. Word2vec으로 생성한 단어별 벡터를 리뷰 수준의 자질 벡터 표현으로 변환한다. 식 1과 같이 리뷰 텍스트에 포함된 모든 단어 벡터의 평균을 계산한다. \vec{v}_i 는 단어 벡터이고, \vec{V}_d 는 리뷰의 벡터이다.

$$\vec{V}_d = \frac{\sum_{i=0}^n \vec{v}_i}{n} \quad (1)$$

- 통합 자질(BOW + Word2vec): 품사 기반의 BOW자질과 word2vec자질을 모두 감성 분류에 활용하는 통합 자질 표

현 방법이다. 각 리뷰 식별번호를 사용하여 BOW자질과 word2vec자질을 통합한다.

4 단계: 벡터화 된 자질 집합으로 변환 된 리뷰는 감성 분류를 위해 기계학습 기반의 단일 및 앙상블 분류기에 대한 입력 데이터로 사용된다. 본 연구에 사용 된 분류기는 표 2와 같다.

표 2. 연구에 사용된 분류기 유형

Table 2. Type of classifier used in this study

Single Classifier	Ensemble Classifier
Logistic Regression(LR)	Adaptive Boost(AB)
Decision Tree(DT)	Bagging(BG)
Naïve Bayes(NB)	Gradient Boosting(GB)
Support Vector Machine(SVM)	Random Forest(RF)

5 단계: BOW(A, D) 및 BOW(A, D, V), word2vec, BOW + word2vec을 자질로 하는 다양한 단어 벡터 표현의 효과를 단일 및 앙상블 분류기의 분류 성능으로 측정한다.

IV. 감성 분류 실험

4-1 감성 데이터

본 연구의 실험 데이터인 한글 영화 리뷰는 100,000 건의 긍정적 리뷰와 100,000 건의 부정적인 리뷰로 구성된다. 모든 리뷰에는 단어가 두 개 이상 포함되어 있으며, 각 리뷰는 1에서 10 점 사이로 리뷰 작성자가 평점 했다. 이 점수를 1에서 4 점은 부정 감성, 9 점과 10 점은 긍정 감성으로 사용한다. 연구 데이터 중에서 전처리 단계에서 불필요한 텍스트를 제외하고 197,008 개의 리뷰가 word2vec의 학습과 기계학습 모형에 사용되었다. 전처리 및 벡터변환 후, 각 자질 표현별 자질 수는 표 3과 같다.

표 3. 각 자질 표현에 따른 자질 수

Table 3. Number of features by each feature representation

Feature representation	# of features
BOW(A, D)	1209
BOW(A, D, V)	2089
Word2vec	300
BOW(A, D) + Word2vec	1509
BOW(A, D, V) + Word2vec	2389

4-2 감성 분류 성능 평가

본 연구에서 제안한 방법의 감성분류 성능 평가를 위해 5 fold Cross-Validation을 사용하고, 식 2와 같이 감성 분류 정확도(Accuracy)를 측정하였다.

$$\frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

TP(True Positive)는 긍정 리뷰를 분류기가 긍정으로 분류한 리뷰의 건수이고, FP(False Positive)는 긍정 리뷰를 분류기가 부정으로 분류한 리뷰의 건수이다. TN(True Negative)은 부정 리뷰에 대해 분류기가 부정으로 분류된 리뷰의 건수이고, FN(False Negative)은 부정 리뷰를 분류기가 긍정으로 분류한 건수이다.

V. 실험결과

개별 자질 표현 방법에 대한 단일 분류기 및 앙상블 분류기의 성능 평가 결과를 표 4에 요약하였다. 세 가지 자질 표현 방법, 형용사와 부사로 구성된 BOW자질(BOW(A, D)), 형용사, 부사, 동사로 구성된 BOW자질(BOW(A, D, V))과 word2vec자질 중에서 word2vec자질과 SVM을 사용한 감성 분류의 정확도가 83.29%로 가장 높았다. BOW(A, D, V)자질에 NB를 사용한 감성 분류 정확도는 66.36%로 가장 낮은 성능을 보였다. BOW(A, D)자질과 BOW(A, D, V)자질에서 가장 높은 정확도를 나타낸 분류기는 LR이고, 두 가지 자질 중에서 BOW(A, D)자질의 정확도는 75.87%로 1.15% 더 높았다. LR를 사용한 BOW(A, D)자질과 SVM을 사용한 Word2vec자질 간의 정확도를 비교하면, word2vec자질이 7.42% 더 높은 정확도를 나타냈다.

가장 높은 감성분류 성능보이는 분류기는 단일 분류기 중 SVM이지만, 앙상블 분류기의 성능은 정확도가 AB는 75.25 %, GB는 74.51 %, BG는 74.05 %, RF는 74.79 %로, 단일 분류기의 정확도에 근접하거나 약간 낮은 정확도를 보인다.

표 4. 개별 자질 표현별 정확도

Table 4. Accuracies by each of feature representation

Methods	Classifier	BOW (A, D)	BOW (A, D, V)	Word2vec
Single	LR	75.87	74.72	83.17
	DT	73.57	69.99	71.35
	NB	66.56	66.36	77.46
	SVM	75.45	74.43	83.29
Ensemble	AB	75.25	73.02	81.44
	BG	74.51	73.36	77.08
	GB	74.05	71.4	80.95
	RF	74.79	73.43	82.49

통합 자질 표현 방법(BOW+Word2vec)의 감성 분류 성능은 표 5로 정리하였다. 가장 높은 정확도를 나타내는 통합 자질과 분류기는 BOW(A, D) + Word2vec 자질에 SVM을 사용한 것으로, 정확도는 84.70%이다. BOW(A, D, V) + Word2vec자질은 83.36 %로 두 가지 통합 자질 표현의 최고 정확도 차이는

1.23%로 BOW(A, D) + Word2vec자질이 더 높게 나타났다. 앙상블 분류기 중 최고 성능은 RF가 83.54%로 나타났고, AB는 83.35 %, GB는 82.69 %, BG는 78.36 %이다. 개별 자질 표현 방법과 유사하게 앙상블 방법을 이용한 감성 분류는 단일 분류기에 비해 조금 낮은 정확도를 보이지만, DT(High=72.96%, Low=71.76)나 NB(High=70.38%, Low=68.86%)와 같이 상대적으로 현저하게 낮은 분류 성능을 보이지 않았다.

표 5. 통합 자질 표현별 정확도

Table 5. Accuracies by integrated feature representation

Methods	Classifier	BOW (A, D) + Word2vec	BOW (A, D, V) + Word2vec
Single	LR	84.52	83.36
	DT	72.96	71.76
	NB	70.38	68.86
	SVM	84.70	83.47
Ensemble	AB	83.35	81.99
	BG	78.36	77.56
	GB	82.69	81.34
	RF	83.54	82.37

개별 자질 표현 방법과 통합 자질 표현 방법에 대한 모든 감성 분류 실험에서 가장 좋은 성능은 BOW(A, D) + Word2vec 자질과 SVM을 사용한 것이다. BOW(A, D) + Word2vec 자질과 LR을 사용한 감성 분류가 다음으로 높은 성능을 보였다. 앙상블 학습 방법의 경우, BOW(A, D) + Word2vec 자질과 RF분류기를 사용한 감성 분류 성능이 83.54%로 가장 높게 나타났다.

VI. 결론 및 향후 연구 방향

본 연구는 한국어 리뷰 텍스트에 대한 효율적인 감성 분류 방법을 제시하는 것을 목적으로 한다. 연구는 한국 영화 리뷰 텍스트에 대해 품사 기반의 BOW자질과 word2vec자질을 사용하여 단일 및 앙상블 분류기를 사용하여 감성 분류를 수행했다. 연구 결과는 다음과 같다. 한국어 감성 분류에서 BOW자질과 word2vec자질 중 word2vec자질이 더 높은 성능을 보였다. 이것은 word2vec자질에 텍스트의 단어 사이의 의미론적 관계가 반영되기 때문으로 볼 수 있다. 또한, word2vec자질의 벡터 수가 BOW자질의 벡터 수에 비해 매우 적기 때문에 보다 적은 자원과 시간 소모로 효율적인 감성 분류 모델이 생성된다. 또한, BOW자질과 word2vec자질을 결합한 통합 된 자질 벡터 표현이 가장 높은 한국어 감성 분류 성능을 보였다. 분류 성능은 BOW자질이 가장 낮고, word2vec자질, BOW + word2vec자질 순으로 증가했다. 이것은 BOW자질에 포함된 단어의 구성적 속성과 word2vec자질이 가지고 있는 의미론적 속성을 결합해서 보다 효과적인 감성 분류가 가능함을 보였다.

본 연구는 단일 분류기와 앙상블 분류기 중에서 가장 좋은 성능을 나타내는 분류기는 단일 분류기인 SVM이다. 그러나 앙상블 분류기의 성능은 단일 분류기에 비해 근접하거나 조금 낮은 성능을 보여, 앙상블 방법은 또한 한국어 감성 분류의 주요 방법으로 사용될 수 있다.

본 연구는 한계는 연구에 사용된 데이터가 영화에 대한 리뷰 텍스트를 사용하여, 온라인 쇼핑, 국가정책, 기업 브랜드와 같은 다른 영역의 감성 분류에 적용하는 데 문제가 발생할 수 있다는 것이다. 본 연구를 바탕으로 다양한 영역의 감성 분류에 적용 가능한 방법을 연구 할 필요가 있다. word2vec 자질의 생성 시, 텍스트에 포함된 단어의 벡터들을 산술평균해서 리뷰의 벡터로 변환하였는데, 감성 분류에 더 높은 성능을 보이는 벡터 집합(Vector Aggregation)방법에 대한 연구가 필요하다. 또한, 본 연구의 실험은 기계 학습 분류기의 파라미터를 기본값으로 설정하여 실행되었다. 향후 보다 정교한 감성 분류기를 위한 파라미터 값 설정에 대한 연구가 필요하다.

감사의 글

이 성과는 2017년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2017R1A2B4010956).

참고문헌

- [1] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1-167, 2012.
- [2] M. Giatsoglou, M. G. Vozalis, K. Diamantaras, A. Vakali, G. Sarigiannidis, and K. C. Chatzisavvas, "Sentiment analysis leveraging emotions and word embeddings," *Expert Systems with Applications*, vol. 69, pp. 214-224, 2017.
- [3] P. Zhang and Z. He, "Using data-driven feature enrichment of text representation and ensemble technique for sentence-level polarity classification," *Journal of Information Science*, vol. 41, no. 4, pp. 531-549, 2015.
- [4] Z. Zhai, B. Liu, H. Xu, and P. Jia, "Clustering product features for opinion mining," in *Proceedings of the fourth ACM international conference on Web search and data mining*, 2011, pp. 347-354.
- [5] N. F. F. Da Silva, E. R. Hruschka, and E. R. Hruschka, "Tweet sentiment analysis with classifier ensembles," *Decision Support Systems*, vol. 66, pp. 170-179, 2014.
- [6] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Systems with Applications*, vol. 57, pp. 117-126, 2016.
- [7] S. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Vol. 2*, 2012, pp. 90-94.
- [8] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1188-1196.
- [9] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning Word Vectors for Sentiment Analysis," *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142-150, 2011.
- [10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," in *Proceedings of workshop at ICLR*, pp. 1-12, 2013.
- [11] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, Vol. 10, 2002, pp. 79-86.
- [12] D. Chatzakou and A. Vakali, "Harvesting opinions and emotions from social media textual resources," *IEEE Internet Computing*, vol. 19, no. 4, pp. 46-50, 2015.
- [13] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," *Information Sciences*, vol. 181, no. 6, pp. 1138-1152, 2011.
- [14] G. Wang, J. Sun, J. Ma, K. Xu, and J. Gu, "Sentiment classification: The contribution of ensemble learning," *Decision Support Systems*, vol. 57, no. 1, pp. 77-93, 2014.
- [15] H.-S. L. Dong-yub Lee Jae-Choon Jo, "User Sentiment Analysis on Amazon Fashion Product Review Using Word Embedding," *Journal of the Korea Convergence Society*, vol. 8, no. 4, pp. 1-8, 2017.
- [16] J. Lilleberg, Y. Zhu, and Y. Zhang, "Support Vector Machines and Word2vec for Text Classification with Semantic Features," *Cognitive Informatics & Cognitive Computing (ICCI* CC), 2015 IEEE 14th International Conference on. IEEE, 2015.*, pp. 136-140, 2015.
- [17] Y. Ren, R. Wang, and D. Ji, "A topic-enhanced word embedding for Twitter sentiment classification," *Information Sciences*, vol. 369, pp. 188-198, 2016.
- [18] Y. Bengio, H. Schwenk, J.-S. Senécal, F. Morin, and J.-L. Gauvain, "Neural Probabilistic Language Models," in *Innovations in Machine Learning: Theory and Applications*, D. E. Holmes and L. C. Jain, Eds. Berlin,

- Heidelberg: Springer Berlin Heidelberg, 2006, pp. 137–186.
- [19] K.-M. Ahn, Y.-S. Kim, Y.-H. Kim, and Y.-H. Seo, “Sentiment Classification of Movie Reviews using Levenshtein Distance,” *Journal of Digital Contents Society*, vol. 14, no. 4, pp. 581–587, Dec. 2013.
 - [20] Y. Kim and M. Song, “A Study on Analyzing Sentiments on Movie Reviews by Multi-Level Sentiment Classifier,” *Journal of Intelligence and Information Systems*, vol. 22, no. 3, pp. 71–89, 2016.
 - [21] Y. Jung, K. Park, T. Lee, J. Chae, and S. Jung, “A corpus-based approach to classifying emotions using Korean linguistic features,” *Cluster Computing*, vol. 20, no. 1, pp. 583–595, 2017.
 - [22] C. Lee, K. Hyun, Y. Byeong, M. Mun, and S. Joo, “Informal Quality Data Analysis via Sentimental analysis and,” *Journal of the Korean Society for Quality Management*, vol. 45, no. 1, pp. 117–127, 2017.
 - [23] Y. Kim and H. Shin, “Finding Sentiment Dimension in Vector Space of Movie Reviews: An Unsupervised Approach,” *Journal of Cognitive Science*, pp. 85–101, 2017.
 - [24] S.-Y. O. Chan Heo, “A Novel Method for Constructing Sentiment Dictionaries using Word2vec and Label Propagation,” *Journal of Korean Institute of Next Generation Computing*, vol. 13, no. 2, pp. 93–101, 2017.
 - [25] E. L. Park and S. Cho, “KoNLPy: Korean natural language processing in Python,” in *Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology*, 2014, pp. 133–136.



박성수(Sung Soo Park)

2010년 : 성균관대학교 경영대학원(경영학석사)

2010년~현 재: 성균관대학교 경영대 박사과정

※관심분야 : 인공지능(Artificial Intelligence), 빅데이터(Big Data), 감성분석(Sentiment Analysis) 등



이건창(Kun Chang Lee)

1984년 : 카이스트 경영과학과(공학석사-의사결정지원)

1988년 : 카이스트 경영과학과(공학박사-인공지능)

1990년 ~ 1995년: 경기대학교 경영정보학과 교수

1995년 ~ 현 재: 성균관대학교 경영대학 교수

※관심분야 : 창의성과학, 인공지능, 헬스 인포매틱스, 빅데이터 분석, 감성분석