# 상품 리뷰 분석을 위한 텍스트 마이닝 기법의 비교

이지웅*, 진영택*
*한밭대학교 컴퓨터공학과
e-mail:jiw4422@nate.com
ytjin@hanbat.ac.kr

# A Comparison of Text Mining Algorithms for Product Review Analysis

Ji-Woong Lee*, Young-Taek Jin*
*Dept of Computer Engineering, Hanbat National University

요 약

오늘날 정보화 시대에서는 온라인 쇼핑의 상품리뷰 등 대용량의 텍스트 문서가 존재하며 제품에 대한 정서적인 의견뿐만 아니라 제품 선호도 및 상품 비교와 같은 유용한 정보를 제공한다.
본 논문에서는 사용자가 작성한 상품 리뷰로부터 제품의 특성을 비교하는 비교의견을 추출하기 위해 적용한 다양한 텍스트 마이닝 기법의 비교 결과를 제시한다.

## 1. 서론

만약 누군가가 저가의 헤드폰을 살지 고가 브랜드의 헤드폰을 살지 고민 중이라면, 아마도 웹에서 두 제품에 대한 비교 정보를 얻고자 할 것이다. 웹의 상품 리뷰, 블로그, 뉴스 기사 등에는 대상 제품에 대한 풍부한 비교 자료가 있으며, 이 자료들이 의사결정에 상당한 영향을 끼치게 될 것이다.

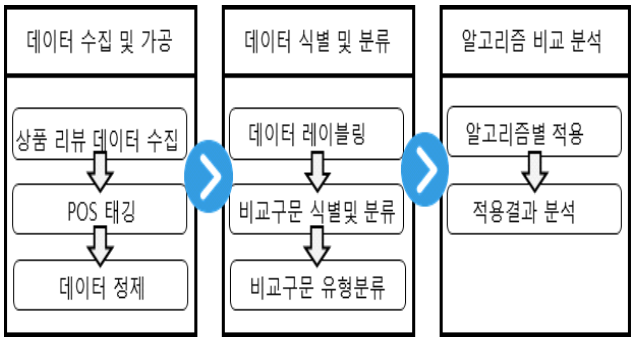
비교는 여러 평가 방법 중 가장 강력하고 확실한 평가 방법 중 하나이며, 더 나은 의사결정을 위해 웹 검색을 통해 더 많은 정보를 얻고자 한다. 웹 검색을 통해 비교 정보를 찾는 일은 기존의 설문 조사 방법이나 전통적 측정보다 훨씬 빠르고 쉬우며 경제적이다. 그러나 비교 리뷰를 직접 읽는 것은 완벽한 해결책이 될 수 없는데, 이유는 소량의 데이터로 인해 편향된 관점을 가질 수 있고, 많은 양의 데이터를 조사하는 것은 아주 많은 시간을 필요로 하기 때문이다. 이러한 상품리뷰로부터 비교의견을 추출하는 연구[1,3]는 많은 연구자에 의해 수행되었으나 어떤 알고리즘을 적용하는 것이 좋은지를 나타내는 비교 연구는 많지 않다.

본 논문에서는 제품의 사용경험이 있는 사용자가 작성한 리뷰로부터 제품의 특성을 비교하는 의견을 추출하기 위해 다양한 텍스트 마이닝 기법을 비교한 결과를 제시한다. 이 방법의 첫 단계로 문장이 비교구문인지 여부를 식별하고, 그 비교구문이 긍정 또는 부정의견인지 식별한다.

마지막으로 각 비교 문에 나타난 제품 및 브랜드 이름과 제품의 피처를 추출하였다. 최종 결과로 정밀도, 재현율(Recall), 정확도 등 알고리즘별 성능을 측정하고 결과를 비교하였다.

## 2. 시스템 개요 및 수행 절차

상품 리뷰 분석을 위해 머신러닝 분류 알고리즘인 랜덤 포레스트, 선형 SVM, 로지스틱 회귀를 이용하여 비교구문 추출, 피처 추출을 위해 적용한 절차는 (그림 1)과 같다.




(그림 1) 상품리뷰의 분석을 위한 절차

1) 상품리뷰 데이터 수집: 상품 리뷰 데이터는 대표적 웹 사이트인 아마존에서1만여건의 헤드폰 리뷰를 수집하였다.

2) POS 태깅: 형태소 분석(POS-tagging) 작업은 형태소의 뜻과 문맥을 고려하여 마크업을 하는 작업이다.본 연구에서는 nltk POS 태깅을 사용하였다.

3) 데이터 정제: html 태그, 불용어 제거 등과 같이 텍스트를 정제해 준다.

4) 데이터 레이블링: 데이터 레이블링은 데이터에 답을 주는 것으로 학습할 데이터에 평가를 하는 지도학습을 하기 위한 준비단계이다. 본 연구에서는 비교구문의 분류의

데이터 레이블링과 구별된 비교구문의 유형별 분류 및 카테고리 분류 등의 작업을 수행했다. 비교 식별 레이블링의 경우 0은 비교 문장이 아닌 경우이며, 1의 경우는 비교구문을 표시한다.

5) 비교구문 식별 및 분류

비교 유형은 네 가지 유형으로 분류한다. 우선 순위를 매길 수 있는 비교가 세 가지 유형으로 나뉘며, 비교가 명시적으로 나타나지 않는 경우는 는 한 가지 유형이 존재한다. 우선순위의 비교 유형은 크거나, 작거나, 같거나, 모든 비교대상보다 같거나 작은 관계를 기반으로 정의된다[2][4].

(1) 명시적 비교: 비교대상 개체들의 피처에 대해 개체의 전체적인 순서를 나타내는 크거나 작은 유형의 관계를 나타낸다. 특정 피처와 관련하여 어떤 객체의 순서를 표현하는 greater or less 유형의 관계이다.

(2) 동일 비교 : 어떤 피처에 대해 두 개체가 동등한 것을 서술하는 equal 유형의 관계이다.

(3) 최상급 비교 : 한 객체가 모든 다른 객체에 대한 순위를 매기는 모든 다른 것들과의 greatest, worst, best 유형의 단계이다

(4) 암묵적 비교: 두 개 이상의 객체의 피처를 비교하지만 등급을 매기지 않는 문장으로 “객체 A는 B와 같거나 다르다”, “객체 A는 피처 F1를 가지고 객체 B는 피처 F2를 가진다.”, “객체 A는 피처 F1을 가지지만 객체 B는 없다.”와 같은 형태를 가진다.

6) 비교구문 유형 분류

5번째 단계에서 구별된 비교구문을 미리 선정한 비교구문 유형별로 데이터 레이블링 작업을 한다. 비교구문은 여러 형태의 비교구문으로 분류할 수 있으며 이를 분류하여 비교 대상 A와 비교 대상 B의 관계 즉, 긍정, 부정에 대한 분류를 한다.

<표 1 > 비교구문 유형, 관계 및 구분

| 유형 | 조건 | 관계 | 구분 방법 |
|---|---|---|---|
| 타입 A | more, better than, A better than B increasing+ pos decreasing+ neg. | A 가 B 보다 좋다 | -POStags(JJR, RBR,JJS, RBS) -비교문(~ er than, few, least, most) -비교단어 (beat, inferior, outperform, superior 등) |
| 타입 B | worse than, A worse than B increasing+ neg. decreasing+ pos. | A 가 B 보다 나쁘다 | |
| 타입 C | equal, same | A 와 B 가 동일하다 | |

<표 1>은 비교 구문 분류 기준 유형의 조건과 유형의 분류시 대상 제품의 A, B의 관계에 따라 분류 조건을 표시하였다.

7) 비교구문으로 판별되고 유형별로 분류된 문장에서 비교 대상 A와 B의 제품과 어떤 피처를 비교구문에 사용하였는지를 추출하여 비교대상들이 어떤 피처를 가지고 긍정 부정을 나타내었는지 분석한다.

8) 제품의 피처 추출: 피처(Feature)란 제품의 특성을 나타내는 단어로 헤드폰의 리뷰의 경우 Sound, noise canceling, price, bettry등의 경우가 피처가 된다. 리뷰의 대상이 되는 상품의 피처는 제품마다 다양하게 존재하며 피처를 통해서 제품의 선호도를 알 수 있다. 이 과정에서는 제품의 피처와 대상 제품의 브랜드 또는 모델을 추출한다.

추출과정은 청킹(chunking)을 통한 룰을 작성하여 청킹 룰을 적용하여 문장을 추출하였다. 비교 기준이 되는 Pos tag인 JJR, JJS, RBR, RBS과 같이 <표 2>와 같이 청킹 룰 기준 비교 형용사와 비교 부사를 찾아 타입별 문장의 피처 선호도 A, B를 구별하기 위해 비교구문을 청킹 단어를 기준으로 앞 부분을 A, 뒷 부분을 B로 구별한다.

<표 2> 제품의 chunking rule

| chunking rule | | |
|---|---|---|
| POS tag | <JJR>, <JJS>, <RBR>, <RBS> | + <IN><br>+ <NN><IN><br>+ <JJ><IN><br>+ <RB> |

A,B 대상 제품의 위치를 인식 후 A, B 두 제품을 찾아준다. 명사의 Pos tag인 <NN>, <NNS>, <NNS> 를 딕셔너리를 이용하여 대상의 제품과 매핑 시켜준다. (′old′, ′JJ′), (′Bose′, ′NNP′)와 같은 경우 딕셔너리에 정의된 Bose를 찾는다. 명사가 존재하지 않는 경우 대명사 <PRP>로 대상 제품을 정의한다. 대명사의 경우 지정하는 대상의 판별이 명확치 않아 대상 제품의 별점을 기준으로 5점 만점에 3점 이상일 경우 비교구문 제품의 선호도에 따라 대명사 지정 대상 제품이 해당 제품인지 판별하여 정의한다. 두 가지 과정이후 대상 제품이 없는 경우는 유형별 chunking을 통해서 해당 제품으로 지정하거나 대상 제품 지정 단어가 없는 경우 해당 제품으로 지정한다. 마지막으로 피처의 경우 명사 <NN>, 형용사 <JJ>가 피처가 된다. 위와 같은 과정을 거쳐 피처 추출기를 통해서 실제 분류된 데이터 100건을 A, B, 피처를 추출하였다. 정확도는 A 87%, B 86%, 피처 81%로 추출이 가능하였다.

## 3. 알고리즘 적용 및 결과 분석

데이터를 BOW와 word2vec로 두가지의 데이터 벡터화 기법을 통해 랜덤 포레스트, 선형 SVM, 로지스틱 회귀로 판별하였다. 비교구문 유형별 분류 작업에 사용된 리뷰 데이터는 2천여건으로 비교구문으로 레이블링된 데이터를

훈련 집합, 검증 집합, 테스트 집합으로 나누어 모델을 학습시켰다. BOW로 벡터화된 데이터를 랜덤포레스트, 로지스틱 회귀, 선형SVM으로 학습 시켰을 때 각각 정확도는 0.91%, 0.90%, 0.90%로 측정되었다.

BOW를 이용한 알고리즘 성능 결과 모델 세부 평가 내역은 전체적으로 고르게 성능 측정값을 나타내었으며 유형별로 측정 성능이 알고리즘별 차이가 있었다. 재현율을 기준으로 A타입의 경우는 랜덤 포레스트, B 타입의 경우는 로지스틱 회귀. C타입의 경우는 로지스틱 회귀가 가장 측정 성능이 높았다. <표 3>은 비교 구문 유형의 식별을 위해 적용한 세가지 알고리즘의 성능측정 결과를 제시하고 있다. Word2Vec의 모델은 google news copus를 사용하였다. Word2Vec으로 벡터화된 데이터를 랜덤포레스트, 로지스틱회귀, 선형SVM로 학습 시켰을 때 각각 정확도는 0.77%, 0.76%, 0.84%로 측정되었다.

알고리즘별 비교구문의 식별 및 분류의 경우 알고리즘과 벡터화 방식의 변경으로 분류 성능 측정결과 종합적으로 가장 높은 성능을 보여준 알고리즘은 랜덤포레스트를 BOW로 벡터화한 측정 성능이 가장 높게 나타났다.

## 4. 결론 및 향후 과제

본 논문에서는 사용자가 작성한 상품 리뷰에 나타난 비교구문 식별 및 분류, 비교구문 유형분류, 비교 대상 및 피처를 추출하기 위한 다양한 텍스트 마이닝 알고리즘을 비교하는 연구를 수행하였다. 연구에서 랜덤 포레스트, 선형 SVM, 로지스틱 회귀 알고리즘과 BOW, word2vec의 벡터화 방법을 이용하여 성능을 측정 결과를 비교 분석하였다. 이처럼 리뷰에서 비교구문을 텍스트 마이닝의 다양한 알고리즘을 이용하여 성능을 비교한 결과 좀더 세부적 분류와 모델링을 선정하여 사용한다면 유용한 정보 제공이 가능할 것이다. 향후 과제로는 한글 리뷰에 적용과 리뷰 구문 패턴을 세밀히 분석하기 위한 방법이 필요하다.

<표 3> 비교구문 유형분류 식별 성능 측정 결과

| Classifier | Accuracy | type | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| RandomForest (BOW) | 0.91% | A | 0.88% | 0.97% | 0.92% |
| | | B | 1.00% | 0.84% | 0.91% |
| | | C | 1.00% | 0.91% | 0.95% |
| RandomForest (Word2Vec) | 0.77% | A | 0.73% | 0.93% | 0.82% |
| | | B | 1.00% | 0.57% | 0.73% |
| | | C | 1.00% | 0.51% | 0.67% |
| LogisticRegression (BOW) | 0.90% | A | 0.90% | 0.94% | 0.92% |
| | | B | 0.93% | 0.90% | 0.91% |
| | | C | 0.90% | 0.95% | 0.92% |
| LogisticRegression (Word2Vec) | 0.76% | A | 0.78% | 0.80% | 0.79% |
| | | B | 0.75% | 0.79% | 0.77% |
| | | C | 0.88% | 0.88% | 0.88% |
| Linear SVM (BOW) | 0.90% | A | 0.90% | 0.93% | 0.92% |
| | | B | 0.96% | 0.87% | 0.91% |
| | | C | 0.98% | 0.93% | 0.95% |
| Linear SVM (Word2Vec) | 0.84% | A | 0.86% | 0.83% | 0.84% |
| | | B | 0.89% | 0.85% | 0.87% |
| | | C | 0.96% | 0.89% | 0.93% |

## 참고문헌

[1] Jindal, N., & Liu, B. (2006, August). Identifying comparative sentences in text documents. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieva,l pp. 244-251.

[2] Ganapathibhotla, M., & Liu, B. (2008, August). Mining opinions in comparative sentences. In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1,pp. 241-248.

[3] Kasturi Dewi Varathan,Anastasia Giachanou and Fabio Crestani,Comparative Opinion Mining: A Review, Journl of Association for Informative Science and Technology, 68(4):811?829, 2017.

[4] Shiliang Sun, Chen Luo Junyu Chen, A review of natural language processing techniques for opinion mining systems, Information Fusion Vol.36, pp. 10-25. 2017.