Journal of the Korea Convergence Society
Vol. 8. No. 4, pp. 1-8, 2017
https://doi.org/10.15207/JKCS.2017.8.4.001

워드 임베딩을 이용한 아마존 패션 상품 리뷰의 사용자 감성 분석

이동엽, 조재춘, 임희석*
고려대학교 컴퓨터학과

User Sentiment Analysis on Amazon Fashion Product Review Using Word Embedding

Dong-yub Lee, Jae-Choon Jo, Heui-Seok Lim*
Dept. of Computer Science and Engineering, Korea University

요 약 현대 사회에서 패션 시장의 규모는 해외와 국내 모두 지속적으로 증가하고 있다. 전자상거래를 통해 상품을 구입하는 경우 다른 소비자들이 작성한 상품에 대한 평가 데이터는 소비자가 상품의 구입 여부를 결정하는데에 영향을 미친다. 기업의 입장에서도 상품에 대한 소비자의 평가 데이터를 분석하여 소비자의 피드백을 반영한다면 기업의 성과에 긍정적인 영향을 미칠 수 있다. 이에 본 논문에서는 아마존 패션 상품의 리뷰 데이터를 학습하여 형성된 워드임베딩 공간을 이용하여 사용자의 감성을 분석하는 모델을 구축하는 방법을 제안한다. 실험은 아마존 리뷰 데이터 570만건을 학습하여 형성된 워드임베딩 공간을 이용하여 긍정, 부정 리뷰 데이터의 개수에 따라 총 3개의 SVM 분류기 모델을 학습하는 방식으로 진행하였다. 실험 결과 긍정 리뷰 데이터 5만건, 부정 리뷰데이터 5만건을 이용하여 SVM 분류기를 학습하였을 때 88.0%로 가장 높은 정확도(accuracy)를 나타냈다.

• **주제어 :** 워드 임베딩, 감성분석, 오피니언 마이닝, 인공지능, 딥러닝, 융합기술

Abstract In the modern society, the size of the fashion market is continuously increasing both overseas and domestic. When purchasing a product through e-commerce, the evaluation data for the product created by other consumers has an effect on the consumer's decision to purchase the product. By analysing the consumer's evaluation data on the product the company can reflect consumer's opinion which can leads to positive affect of performance to company. In this paper, we propose a method to construct a model to analyze user 's sentiment using word embedding space formed by learning review data of amazon fashion products. Experiments were conducted by learning three SVM classifiers according to the number of positive and negative review data using the formed word embedding space which is formed by learning 5.7 million Amazon review data.. Experimental results showed the highest accuracy of 88.0% when learning SVM classifier using 50,000 positive review data and 50,000 negative review data.

• **Key Words :** Word Embedding, Sentiment Analysis, Opinion Mining, Artificial Intelligence, Deep Learning, Convergence technique

*Corresponding Author : 임희석(limhseok@korea.ac.kr)
Received February 9, 2017 Revised March 31, 2017
Accepted April 20, 2017 Published April 28, 2017

1. 서 론

현대 사회에서 패션 시장의 규모는 해외와 국내에서 모두 지속적으로 증가하고 있다. 2016년에는 국내 패션 시장에서 이루어지는 거래금액이 40조원에 이를 만큼 패션 상품의 거래는 오프라인 매장에서 뿐만 아니라 전자상거래를 통해서도 활발히 이루어지고 있다[1].

전자상거래에서는 온라인 환경을 통해 소비자는 상품에 관한 다양한 의견 및 평가들을 접할 수 있고, 이 과정에서 소비자들은 상품 구매 이전에 온라인 검색을 통해 상품에 대한 정보를 얻을 수 있다. 자신이 구매하고자 하는 상품에 대한 다른 소비자들의 의견 및 평가 데이터는 소비자가 해당 상품의 구매 여부를 결정하는 데에 중요한 영향을 미친다[2]. 소비자들이 상품에 대한 정보를 얻을 때 상품을 판매하는 기업이 제공하는 정보 보다는, 소비자 자신과 유사한 일반 소비자들의 정보를 더 신뢰하여 의견을 수용하는 것이다[3]. 소비자들이 작성한 상품에 대한 평가 데이터는 상품을 판매하는 기업 입장에서도 기업의 성과에 긍정적인 영향을 미친다는 연구가 있다[4]. 따라서 상품에 대한 소비자의 평가 데이터를 분석하는 것은 기업의 입장에서도 매우 중요하다고 볼 수 있다. 상품을 구입한 소비자들의 데이터를 분석한다면, 소비자의 욕구에 맞게 적절한 상품의 생산량을 조절할 수 있을 것이다. 소비자들의 반응이 좋은 상품의 경우 상품의 생산량을 늘리고, 반응이 좋지 않은 상품들의 경우에는 생산량을 줄임으로써 결과적으로 기업의 입장에서 상품 생산에 지불되는 낭비를 줄일 수 있을 것이다. 또한 소비자들의 의견을 분석하여 해당 상품에 대한 소비자들의 피드백을 반영한다면 상품의 품질을 높이는데 기여할 수 있을 것이다.

해당 상품에 대한 별점과 소비자가 작성한 의견글은 소비자가 해당 제품에 대해 어떤 성향을 보이는지를 분석하는 데에 이용될 수 있다. 대표적인 예시로 아마존의 경우 해당 상품에 대해 소비자들이 작성한 리뷰를 살펴보면 해당 상품에 대해 소비자 자신이 생각한 별점과 리뷰를 남길 수 있다. 하지만 현대 사회에서는 하루에도 수많은 상품들이 늘어나고 각 상품에 대해 많은 평가 데이터들이 생성되기 때문에, 상품을 판매하는 기업 입장에서 일일이 상품 별 소비자들의 성향을 파악하는 것은 많은 비용이 발생할 수 있다. 리뷰 데이터의 경우에는 소비자 개인마다 주관적으로 작성하는 것이기 때문에 객관적으로 해당 상품에 대한 소비자들의 성향을 파악하는 것이 매우 어려울 수 있다.

이에 대해 상품의 리뷰를 보다 객관적으로 평가하기 위해 리뷰 데이터를 긍정, 부정으로 분류하는 방법을 활용할 수 있다. 문장을 긍정, 부정으로 분류하는 선행 연구로는 키워드 추출 기반의 연구가 있고[5,6], 온톨로지(ontology)를 기반으로 한 연구가 진행되어 왔으며[7,8], 어휘(lexicon) 나 품사 태깅(part-of-speech) 를 활용하여 자질(feature)를 구성한 연구가 있다[9,10,11]. 또한 다양한 n-gram의 조합으로 자질를 구성하고 긍정, 부정을 분류하는 연구가 있다[12,13]. 하지만 온톨로지(ontology)를 기반으로 긍정, 부정 모델을 구축하는 경우 시간과 비용이 많이 발생하며, 어휘(lexicon) 나 품사 태깅(part-of speech) 또는 n-gram 을 이용하여 자질을 구성하는 경우 그 문장의 의미론적(semantic) 특성을 반영하지 않고 분류 긍정, 부정 분류 모델을 구축 한다는 단점이 있다.

이에 본 논문은 문장의 의미론적 특성을 반영하여 긍정, 부정을 분류할 수 있는 모델을 구축하는 방법을 제안한다. 실험은 아마존 쇼핑 사이트에서 의류(clothing), 신발(shoes) 그리고 쥬얼리(jewelry) 카테고리에 해당하는 상품들의 소비자 리뷰 데이터 약 570만건을 이용하여 진행하였다. 그 결과 패션 상품 카테고리에 해당하는 소비자들의 리뷰 데이터를 기반으로 상품에 대한 소비자의 성향을 예측할 수 있었다.

2. 관련 연구

2.1 Word2vec

Word2vec은 Google의 Tomas Mikolov와 그의 팀원들에 의해 연구된 신경망(Neural Network) 기반의 연속 워드 임베딩(continuous word embedding) 학습 모델이다[14,15]. Word2vec을 이용하여 각 단어들을 학습 할 경우, 워드 임베딩 공간에서 비슷한 문맥을 가진 단어들은 서로 가까운 공간 분포를 가지게 된다. Word2vec은 연속 워드 임베딩을 표현하는 방법으로 두 가지의 모델 구조를 가진다. 첫 번째는 continuous bag-of-words (CBOW)의 모델 구조이고, 두 번째는 continuous skip-gram의 모델 구조이다. CBOW의 모델 구조에서는 예측하려는 단어의 주변 단어들을 이용하여 단어를 예측하고, skip-gram 모델 구조에서는 현재 주어진 단어를 이용하여 그 단어 주변의 단어들을 예측한다. 학습은 크기가 T 인 코퍼스(corpus)에서 t번째에 해당하는 단어

w_t 에 대해 주어진 문맥(context) 에서 각 토큰의 로그 가능도(log likelihood)를 최대화 하도록 진행된다. 주어진 문맥에서 윈도우(window) 사이즈가 c, w_{t-c}^{t+c} 가 w_t 를 중앙으로 하는 앞과 뒤의 윈도우 사이즈 c 범위 안에 해당하는 단어들의 집합 이라고 할 때, 로그 가능도를 최대화 하는 식은 [Equation 1]과 같다.

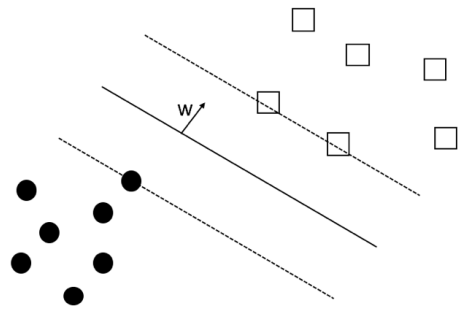
$$\max \frac{1}{T}\sum_{t=1}^{T} \log\ \mathrm{p}(w_t|w_{t-c}^{t+c}) \qquad (1)$$

본 논문은 아마존 리뷰 데이터의 내용을 이루고 있는 단어들의 의미론적 특성을 보존하기 위해 word2vec을 이용하여 총 570만건의 리뷰 데이터를 학습하였다

2.2 SVM (Support Vector Machine)

SVM(Support Vector Machine)은 [Fig 1]과 같이 학습 데이터들의 분포 안에서, 최대 마진(maximal margin)을 가지며 두개의 클래스(class)를 분류할 수 있는 초평면(hyperplane)을 찾아내는 알고리즘이다.

[Fig. 1]에서 실선을 1과 -1로 표현되는 두개의 클래스(class)를 구분하는 초평면이라 하고, W는 초평면의 법선 벡터(normal vector)라 한다. 클래스를 구분하는 초평면이 선택 될 때 이 초평면과 제일 가까운 두 클래스의 각 벡터들을 서포트 벡터(support vector)라 한다. 각 서포트 벡터를 지나는 초평면 사이의 거리를 마진이라고



[Fig. 1] Linear SVM

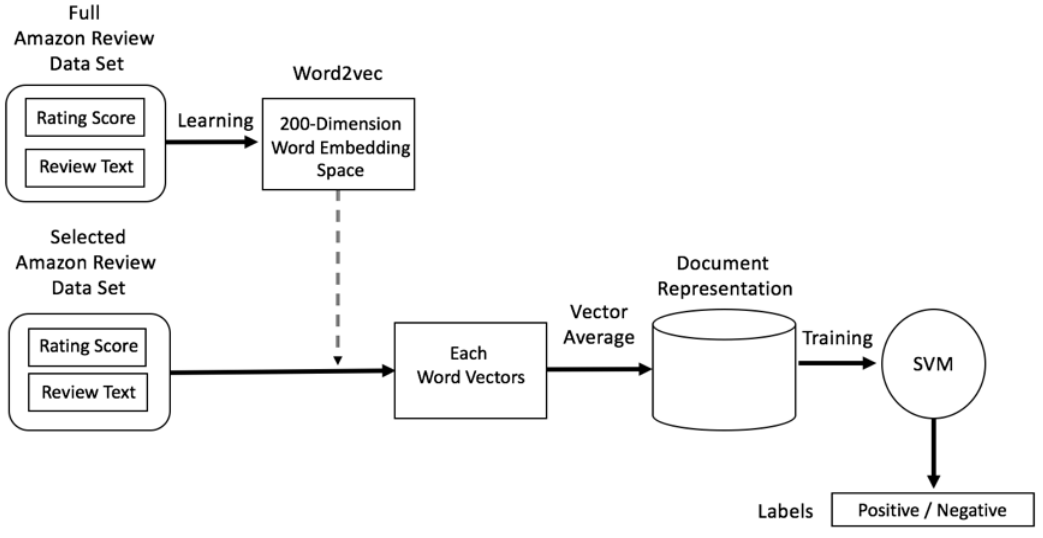
하며, 마진의 값은 $\frac{w}{2}$ 가 된다. 학습 데이터의 집합을 $D=(x_i, y_i)$ 라 하고, y_i 는 1 또는 -1의 값으로 x_i 가 어떤 클래스에 속해있는지를 나타내는 값이라고 할 때, 만약 y_i 가 선형적으로 분리 가능하다면 초평면은 [Equation 2]와 같이 나타낼 수 있다.

$$w \cdot x - b = 0 \qquad (2)$$

각 두 클래스에 해당하는 서포트 벡터 사이에는 데이터가 존재하지 않기 때문에 [Equation 3] 와 같이 나타낼 수 있다.

$$y_i(w \cdot x - b) \geq 1, \text{for all } 1 \leq i \leq n \qquad (3)$$

결국 마진의 최대값을 구하는 문제는 w를 최소화 하는 최적화 문제로 표현될 수 있다.



[Fig. 2] A Structure of Proposed Model

3. 소비자 감성 분석 모델 개발

[Fig 2]는 본 논문에서 제안하는 모델 구조를 나타낸다. 본 논문은 아마존 사이트에서 패션 카테고리에 해당하는 상품들의 리뷰를 분석하여 상품에 대한 소비자의 성향을 예측하였다.

상품 리뷰의 의미론적 특성을 표현하기 위해 word2vec 을 이용하여 의류, 신발, 쥬얼리 카테고리에 해당하는 아마존 상품 리뷰 데이터 570만건(Full Amazon Review Dataset)을 학습하였다. 학습된 word2vec 은 리뷰 데이터를 구성하는 각 단어들의 의미론적 특성을 반영한 워드 임베딩(word embedding) 공간을 형성한다. 소비자 감성 분석 모델 학습에 이용할 데이터(Selected Amazon Review Dataset)를 기반으로 형성된 워드 임베딩 공간을 이용하여 리뷰 데이터의 문장을 구성하고 있는 각 단어를 표현하는 벡터(vector) 값들의 평균화를 통해 자질(feature)을 구성한다. 이후 구성된 자질을 이용하여 SVM분류기 모델을 학습한다. 학습된 SVM 모델은 소비자의 리뷰 데이터를 분석하여 해당 상품에 대해 소비자가 긍정의 의견을 가지고 있는지, 부정의 의견을 가지고 있는지 판단하여 소비자의 성향을 예측할 수 있다.

3.1 Dataset

사용자의 감성을 분석하기 위한 모델을 구축하는 데에 570만건의 아마존 패션 상품 리뷰 데이터를 이용하였다. 아마존 데이터를 이용하여 상품에 관련된 연구를 시도한 것으로는 상품 이미지를 기반으로 비슷한 패션 스타일을 추천해주는 연구와[16], 상품 쿼리(query) 이미지와 함께 보조적으로 같이 사용할 수 있는 상품을 추천해주는 연구[17]가 있다. 본 논문에서는 아마존 상품 리뷰 데이터를 이용한 분석을 진행하기 위해 데이터 전처리 작업으로 중복되는 상품 리뷰 데이터를 제거하고, 각 상품에 해당하는 리뷰 데이터를 그룹화(grouping)하는 작업을 진행하였다.

본 논문에서 이용한 아마존 리뷰 데이터의 형태는 <Table 1>와 같다. Reviewer ID는 리뷰를 작성한 사용자의 ID를, Asin 은 아마존 상품의 고유 ID값을, Reviewer Name은 작성자의 이름, Helpful 은 상품의 유용함 정도를 별점으로 나타낸 것, Review Text는 리뷰의 내용, Overall은 소비자가 상품에 대해 평가한 평점, Summary는 리뷰의 결론, Unix Review Time은 unix 시스템 기준으로 리뷰가 작성된 시간, Review Time은 리뷰가 작성된 시간을 나타낸다.

〈Table 1〉 Amazon Product Review Data

Type	Contents
Reviewer ID	A2SUAM1J3GNN3B
Asin	00000013714
Reviewer Name	J.McDonald
Helpful	[2, 3]
Review Text	He is having a wonderful time playing these old hymns. The music at times hard to read because we think the book was published for singing from more than playing from. Great purchase though!
Overall	5.0
Summary	Heavenly Hymns
Unix Review Time	21252800000
Review Time	09 13, 2009

3.2 Word2vec 학습

Word2vec 학습에 필요한 코퍼스(corpus)를 생성하기 위해 <Table 1>과 같은 아마존 리뷰 데이터에서 Review Text로 표시되어있는 리뷰의 내용을 이용하였다. 리뷰 데이터 별로 리뷰의 내용은 짧게는 1줄에서 길게는 5줄 이상으로 이루어져있다. 본 논문에서는 리뷰 내용의 길이에 상관없이, 각 리뷰 데이터를 표현하는 행(row)이 해당 리뷰의 내용을 모두 담을 수 있도록 구성하였다. 이러한 방식으로 총 570만건의 리뷰 데이터의 내용을 담고 있는 코퍼스를 이용하여 word2vec 학습을 진행하였다.

<Table 2>는 word2vec 벡터의 학습에 사용된 하이퍼 파라미터를 나타낸다.

〈Table 2〉 Hyper-parameter Setting in Word2vec

Parameter	Explanations	Value
Size	Dimensionality of feature vectors	200
Window	Maximum distance within a sentence	5
Iteration	Number of iteration over the corpus	5
Alpha	Intial learning rate	0.025
Min_Count	Ignore all words if total frequency is lower than this value	5
Workers	Number of threads to train model	4

학습한 word2vec 결과 워드 임베딩 공간에 각 단어들은 의미론적 특성을 반영하며 클러스터링 된다. <Table 3>와 <Table 4>는 쥬얼리(jewellery)와 의류(clothing)

라는 패션 상품 카테고리를 나타내는 쿼리(query)단어에 대해 워드 임베딩 공간에서 유사 단어 검색 결과 비슷한 의미를 가진 단어끼리 클러스터링 된 결과를 나타낸다.

〈Table 3〉 Similar Word List with Jewellery

No	Similar Word List
1	Jewery
2	Jewlery
3	Sorrelli
4	Myia
5	Trinket
6	Necklace

〈Table 4〉 Similar Word List with Clothing

No	Similar Word List
1	Cloth
2	Cloths
3	Apparel
4	Outerwear
5	T-shirts
6	Swimwear

<Table 3>에서 쥬얼리(jewellery)라는 단어와 유사한 단어들을 보면 jewery 와 jewlery 와 같이 사용자들이 오타로 입력한 단어들도 제대로 입력한 쥬얼리(jewellery)라는 단어와 유사한 의미인 보석이라는 뜻을 가지므로 클러스터링 된 결과를 확인할 수 있다. Sorrelli와 myia 같은 단어들은 쥬얼리 상품을 판매하는 브랜드를 의미하는데 이 브랜드들도 워드 임베딩 공간에서 쥬얼리와 유사한 의미론적 특성을 가지고 있는것을 확인할 수 있다. 또한 장신구(trinket)이나 목걸이(necklace)와 같은 단어들은 쥬얼리라는 카테고리에 해당하는 단어들이기 때문에 쥬얼리와 함께 클러스터링 된 결과를 확인할 수 있다. <Table 4>에서 의류(clothing)라는 단어에 대해서도 살펴보면, cloth(천) 이나 cloths(천 조각들) 와 같이 의류를 구성하고 있는 소재에 대한 단어들도 같이 클러스터링 되는 결과를 확인할 수 있다. apparel(의복), outerwear(겉옷), t-shirts(티셔츠), swimwear(수영복) 와 같은 단어들은 의류라는 카테고리에 해당하는 상품들의 종류이기 때문에 의류와 함께 클러스터링 되는 결과를 확인할 수 있다.

<Table 5>와 <Table 6>는 감정 표현에 관련된 단어들에 대해 이와 유사한 단어들의 클러스터링 되는 결과를 나타낸다.

<Table 5>에서는 감정 표현에 관련된 단어인 좋다(great)라는 단어와 비슷한 의미를 가진 긍정 표현에 관련된 단어들이 각각 클러스터링 되는 결과를 확인할 수 있다. 좋다라는 단어와 유사한 의미를 가진 단어들을 살펴보면, 환상적인(fantastic), 좋은(good), 멋진(wonderful), 굉장한(terrific), 믿을 수 없는(fabulous) 과 같이 긍정 의미들을 가진 단어들끼리 클러스터링 되는 결과를 확인할 수 있다.

〈Table 5〉 Similar Word List with Great based on Emotion Expression

No	Similar Word List
1	Fantastic
2	Good
3	Wonderful
4	Terrific
5	Fabulous

〈Table 6〉 Similar Word List with Disappoint based on Emotion Expression

No	Similar Word List
1	Dissapoint
2	Disapoint
3	Bother
4	Fail
5	Suffocate

<Table 6>는 실망(disappoint)라는 단어와 비슷한 의미를 가진 부정 표현에 관련된 단어들의 클러스터링 결과를 나타낸다. 유사한 의미를 가진 단어들을 살펴보면 소비자들이 disappoint라는 단어를 입력하려고 할 때, 주로 오타로 잘못 입력하는 dissapoint, disapoint 와 같은 단어들이 실망이라는 단어와 유사한 단어로 클러스터링 되는 결과를 볼 수 있다. 또한 성가시게 하다(bother), 실패(fail), 곤란하게 하다(suffocate) 와 같이 부정적인 의미를 가진 단어들이 실망이라는 단어와 함께 클러스터링 되는 결과를 확인할 수 있다.

3.3 Document Representation

SVM 분류기 모델을 학습하는데 필요한 자질(feature)을 구성하기 위해 섹션[3.2]에서 word2vec 모델을 학습한 결과 형성된 워드 임베딩 공간을 이용하였다. Word2vec 학습 결과 형성된 워드 임베딩 공간을 이용하여 자질을 구성하는 방법으로는 각 문서(document)를 구성하고 있는 단어들의 벡터값들을 평균화하여 자질로 구성하는 연구가 있다[18,19]. 문서를 구성하고 있는 전체

단어의 개수가 N, 워드 임베딩 공간에서 문서의 i번째 단어를 벡터값으로 표현한 것을 v(i)라 할 때, 문서를 구성하고 있는 단어들의 벡터값의 평균은 [Equation 4]와 같이 표현된다.

$$\frac{1}{N}\sum_{i=1}^{N} v(i) \qquad (4)$$

[Fig. 3]은 [Equation 4]를 따라 본 논문에서 아마존 리뷰 데이터 문장을 구성하고 있는 단어별 벡터 값들의 평균을 통해 자질을 구성하는 알고리즘을 나타낸다.

```
1:word2vec ← trained word2vec model
2:feature_set ← empty list
3:document_set ← review text in review data
4:for each document in document_set do
5:   word_list ← split the string by space in document
6:   for each w in word_list do
7:     count ← 0
8:     vec ← initialize with zero vectors
9:     if w in word2vec then
10:      w_vec ← vector value of w in embedding space
11:      count ← count + 1
12:      add w_vec to vec
13:    else
14:      continue
15:    endif
16:  endfor
17:  vec_avg ← vec / count
18:  add vec_avg to feature_set
19:end for
```

[Fig. 3] Comprise feature using vector averaging

4. 성능 평가 및 결과

사용자 감성 분석을 위한 긍정, 부정 분류모델을 학습하기 위해 [섹션3.3]에서 구성한 문서 표현(Document Representation)을 이용하여 SVM 모델을 학습하였다.

아마존 리뷰데이터를 이용하여 긍정, 부정을 나타내는 클래스 레이블링(Labelling)을 진행하기 위해 <Table 1>과 같은 아마존 리뷰 데이터에서 소비자가 상품에 대해 평가한 평점정보를 이용하였다. 평점은 별 1개부터 별 5개까지 소비자가 상품에 대해 평가를 할 수 있는데, [10]의 연구와 같이 본 논문에서는 별 5개의 평점에 해당하는 리뷰 데이터는 긍정으로, 별 1개의 평점에 해당하는 리뷰 데이터에 대해서는 부정으로 간주하여 클래스 레이블링을 진행하였다.

〈Table 7〉 Performance of SVM depending on the number of Data

No	Train		Test		Accuracy
	Pos.	Neg.	Pos.	Neg.	
1	10000	10000	1000	1000	78.4%
2	20000	20000	2000	2000	83.1%
3	50000	50000	5000	5000	88.0%

<Table 7>은 학습 결과 SVM 학습에 이용한 긍정 부정 학습 데이터 개수와, 테스트 데이터 개수에 따른 SVM 분류기의 성능을 나타낸다. 긍정, 부정 리뷰 데이터의 개수에 따라 총 3개의 SVM 분류기 모델을 학습하였다. 긍정 리뷰 데이터 5만건, 부정 리뷰데이터 5만건을 이용하여 SVM을 학습하였을 때 88.0%로 가장 높은 정확도(accuracy) 를 나타냈다. 모델에 대한 평가는 학습에 사용하지 않은 긍정, 부정 리뷰 데이터 각각 5천건으로 구성된 테스트 데이터를 이용하여 진행하였다.

[Fig 4] Expected Effect of Sentiment Analysis Model

[Fig 4]은 학습된 소비자 감성 분석 모델을 아마존 패션 상품 리뷰 데이터에 적용할 경우 예상되는 기대효과를 나타낸다. [Fig 4]에서 별점 분포를 살펴보면 별점 5개와 별점 4개의 비율의 합이 60%이고 나머지 별점의 합 비율은 40% 이다. 별점이 3점인 경우의 리뷰 데이터의 내용을 살펴보면 상품에 대해 긍정적인 의견을 표현하고 있지만 3점으로 평가를 하는 소비자가 있는 반면, 부정적

인 의견을 표현하면서 3점으로 평가를 하는 소비자들이 있다. 이에 대해 소비자 감성 분석 모델을 이용하여 리뷰 데이터의 긍정, 부정 비율을 판단한 결과 긍정의 비율이 81%, 부정의 비율이 19% 을 나타낸다는 점을 활용한다면 보다 객관적으로 소비자나 기업 입장에서 상품에 대한 사용자들의 의견 분석을 진행하는데 도움이 될 수 있을거라 기대된다.

5. 논의 및 결론

국내와 해외에서 지속적으로 증가하고 있는 패션 시장에서 상품에 대한 소비자의 평가 데이터는 매우 중요하다. 소비자의 평가 데이터는 다른 소비자들이 해당 상품을 구매할 때 참고하여 상품의 구매 여부를 결정짓는 중요한 요소가 될 뿐만 아니라, 상품을 판매하는 기업 입장에서도 소비자의 평가 데이터를 활용하여 소비자의 니즈(needs)를 반영한 판매 전략을 세운다면 더욱 많은 이익을 창출할 수 있다. 소비자가 상품에 대해 평가하는 데이터로는 대표적으로 소비자가 작성한 리뷰 데이터를 활용할 수 있다.

이에 본 논문은 아마존 패션 상품 리뷰를 활용하여 상품에 대한 사용자의 감성 분석을 진행하였다. 소비자가 작성한 상품 리뷰의 내용을 기반으로 소비자가 해당 상품에 대해 긍정적인 의견을 가지고 있는지, 부정적인 의견을 가지고 있는지를 구분할 수 있는 모델을 구축함으로써 상품에 대한 별점 이외에 보다 객관적인 방식으로 소비자의 의견을 파악할 수 있는 방법을 제안할 수 있었다.

평가 데이터를 이용하여 word2vec 모델을 학습한 결과 워드 임베딩 공간에서 비슷한 문맥이나 의미를 가진 단어끼리 유사한 공간상의 분포를 가지는 것을 확인할 수 있었다. 이후 형성된 워드 임베딩 공간을 이용하여 각 문서를 구성하고 있는 단어들의 벡터값을 평균화하여 SVM 학습에 필요한 자질을 구성하고 문서 표현으로 활용하였다. 형성된 문서 표현을 이용하여 SVM 분류기를 학습한 결과 소비자의 긍정, 부정 의견을 파악하는 모델의 성능은 88%의 정확도를 나타냈다.

본 논문에서는 패션 상품 리뷰를 통해 의미론적 특성을 반영하여 사용자의 의견이 긍정인지 부정인지를 분석하는 모델을 구축하는 실험을 진행하였다. 향후 연구에서는 더 나아가 부정의 의견으로 판단된 리뷰들을 대상으로 주로 어떤 단어들이 리뷰에서 부정 의견을 형성하는지 또는 어떤 단어들이 주로 긍정의 의견을 형성하는지 등을 분석한다면 패션 상품 기업의 입장에서 더욱 더 객관적인 방법으로 상품에 대한 소비자의 의견 분석을 진행할 수 있을거라 기대된다.

ACKNOWLEDGMENTS

본 논문은 2016년도 정부 (미래창조과학부) 의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. R1610941).

본 연구는 문화체육관광부 및 한국콘텐츠 진흥원의 2016년도 문화기술 연구개발 지원 사업으로 수행되었음. [2016. 스마트 시니어세대의 문화향유를 위한 인지반응 맞춤형 UI/UX기술 개발]

REFERENCES

[1] S. Y. Jo. (2016, August 16). Fashion Journal&Textile Life. “Forecast of domestic fashion market in 2016”. Retrieved January,16 2017 from http://okfashion.co.kr/print_paper.php?number=44812&news_article=nm_news_article&target=print_paper

[2] W. James C. and A. L. Ostrom. “The Internet as information minefield: An analysis of the source and content of brand information yielded by net searches”. In Journal of Business Research, 56 (11), 907-14, 2003.

[3] B. Bickart and R. M. Schindler. “Internet forums as influential sources of consumer information”. In Journal of Interactive Marketing, pp. 314 40, 2001.

[4] J. A. Chevalier and D. Mayzlin. “The Effect of Word of Mouth on Sales: Online Book Reviews”. In NBER working paper, 2003

[5] J. S. Kim. “Emotion Prediction of Document using Paragraph Analysis”. In Journal of Digital Convergence, pp.249-255, 2014

[6] J. S. Kim. “Emotion Prediction of Paragraph using Big Data Analysis”. In Journal of Digital Convergence, pp.267-273, 2016

[7] L. Zaho and C. Li.“Ontology Based Opinion Mining

for Movie Reviews". In Springer, 2009.

[8] P. Baranikumar and N. Gobi. "Feature Extraction of Opinion Mining Using Ontology". In International Journal of Advances in Computer and Electronics Engineering,1,(pp. 18-22), 2016.

[9] B. Xue et al."A study on sentiment computing and classification of Sina Weibo with Word2vec". In IEEE Int. Cong on Big Data, pp. 358- 363, 2014.

[10] Zhang, D., Xu, H., Su, Z., & Xu, Y. "Chinese comments sentiment classification based on word2vec and svm perf". Expert Systems with Applications, 42(4), 1857 - 1863. 2016.

[11] Niu, T., Zhu, S., Pang, L., & El Saddik, "A. Sentiment analysis on multi-view social data". In Multimedia modeling (pp. 15 - 27) at Springer, 2016.

[12] Matsumoto, S., Takamura, H., & Okumura, M. "Sentiment classification using word sub-sequences and dependency sub-trees". In Advances in knowledge discovery and data mining (pp. 301 - 311) at Springer, 2005.

[13] Tripathy, A., Agrawal, A., & Rath, S. K. "Classification of sentiment reviews using n- gram machine learning approach". Expert Systems with Applications, 57, 117 - 126. 2016.

[14] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient Estimation of Word Representations in Vector Space". In Proceedings of Workshop at ICLR, 2013.

[15] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. "Distributed Representations of Words and Phrases and their Compositionality". In Proceedings of NIPS, 2013.

[16] J. McAuley, C. Targett, Q. Shi, and A. van den Hengel. "Image-based recommendations on style and substitutes". Proceedings of the 38st annual international ACM SIGIR conference., 2015.

[17] J. J. McAuley, R. Pandey, and J. Leskovec. "Inferring networks of substituable and complementary products". In KDD, 2015.

[18] R. K. Bayot and T.Gonçalves. "Author profiling using svms and word embedding averages". In CLEF, 2016.

[19] Y. Belinkov, M. Mohtarami, S. Cyphers, and J. Glass. "VectorSLU: A continuous word vector approach to answer selection in community question answering systems". In Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval, 2015.

저자소개

이 동 엽(Dong-Yub Lee) [정회원]

- 2017년 2월 : 인하대학교 컴퓨터정보 공학과 (이학사)
- 2017년 3월 ~ 현재 : 고려대학교 컴퓨터학과 소프트웨어 전공 석, 박사 통합과정

<관심분야> : 자연어 처리(NLP), 딥러닝(Deep Learning), 인공지능

조 재 춘(Jae-Choon Jo) [정회원]

- 2010년 2월 : 제주대학교 사범대 컴퓨터교육과 (이학사)
- 2012년 2월 : 고려대학교 컴퓨터교육과(이학석사)
- 2012년 2월 ~ 현재: 고려대학교 컴퓨터학과 박사수료

<관심분야> : 컴퓨터 교육, EDM, AI in Education

임 희 석(Heui-Seok Lim) [정회원]

- 1992년 2월 : 고려대학교 컴퓨터학과 (이학사)
- 1994년 2월 : 고려대학교 컴퓨터학과 (이학석사)
- 1997년 2월 : 고려대학교 컴퓨터학과 (이학박사)
- 2008년 3월 ~ 현재 : 고려대학교 컴퓨터 학과 교수

<관심분야> : 자연어 처리(NLP), 뇌신경 언어 정보 처리