졸업프로젝트
라쿠텐 상품 리뷰의 사용자 감성 분석

이상훈
20120716
숭실대학교 정보통계보험수리학과

User sentiment analysis on Rakuten Product Review

요약

국내 및 해외에서 지속적으로 성장하고 있는 패션 시장에서 상품에 대한 소비자의 평가 데이터는 매우 중요하다. 전자 상거래를 이용하는 소비자에게 다른 소비자들이 작성한 상품에 대한 평가 데이터는 그 상품의 구입 여부를 결정하는데 큰 영향을 미친다. 기업의 입장에서도 상품에 대한 소비자의 평가 데이터를 분석하여 소비자의 피드백을 반영한다면 기업의 성과에 긍정적인 영향을 미칠 수 있다. 이에 본 연구에서는 라쿠텐 패션 상품의 리뷰 데이터를 바탕으로 오피니언마이닝(OpinionMining)에 사용되는 여러 가지 방법의 감성분석을 진행한다. 데이터는 2018년 6월경 판매된 남성, 여성의류 카테고리에서 상품에 대한 구매리뷰 약 7만 4천건이 추출되었으며. 해당 데이터를 전처리하고 분류 학습을 위한 모델을 구축하여 감성분석을 진행했다. 그리고 분석을 통해 도출된 결과를 바탕으로 비정형데이터를 분석한 결과가 마케팅 전략으로 확장될 수 있을지에 대해 고찰해보았다.

서론

현대 사회에서 패션 시장의 규모는 해외와 국내 모두 지속적으로 증가하고 있다. 패션 시장에서 유행의 흐름은 빠르며 회사들은 고객들의 니즈를 맞춰 제품을 출고하며 상품에 대한 고객들의 반응을 분석해야한다. 소비자는 상품을 구매하기 전에 다른 고객이 작성한 다양한 의견 및 평가들을 접할 수 있고, 이 과정에서 소비자들은 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. 소비자들이 상품에 대한 정보를 얻을 때 상품을 판매하는 기업이 제공하는 정보 보다는, 소비자 자신과 유사한 일반 소비자들의 정보를 더 신뢰하며 의견을 수용하는 것이다. 소비자들이 작성한 상품에 대한 평가 데이터는 상품을 판매하는 기업 입장에서도 기업의 성과에 긍정적인 영향을 미친다는 연구가 있다. 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 의견은 소비자가 해당 제품에 대해 어떤 성향을 보이는지 분석하는 데에 이용될 수 있다.

그 예시로 라쿠텐의 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 여러 가지 방법을 활용할 수 있다. 본 연구에서는 라쿠텐 마켓 사이트에서 판매된 남성, 여성 의류 카테고리에 있었던 상품 리뷰 데이터를 기반으로 상품에 대한 소비자의 성향을 예측하려고 한다.

Collecting Data

본 연구에서 사용된 데이터는 2018년 6월 라쿠텐에서 판매된 여성, 남성의류 카테고리에 있던 1위~10위까지의 상품을 조회하여 그 리뷰들을 크롤링하여 얻어졌다. 상품의 점수와 구매 고객의 성별, 연령 분포를 그래프를 통해 확인할 수 있다. 수집한 데이터는 총 74700개(여성 51030, 남성 22410)이며 상품코드와 점수, 리뷰, 성별, 상품 타입 속성으로 설정하였다. 이때 주의해야 했던 점은 디폴트로 평점이 높은 리뷰를 기준으로 첫 번째 페이지에 그 리뷰가 나타나기 때문에 편향을 막기 위해 각 상품에 등록된 리뷰 전체 표본을 확보했다

[그림1]

[그림2]

[그림3]

[그림1]~[그림3]을 통해 라쿠텐 마켓사이트에서 여성상의 카테고리에서 판매된 인기상품 리스트와 평점분포, 상품에 대한 고객의 구매평을 볼 수 있다. [그림4]는 크롤링한 데이터로 만들어진 데이터프레임으로 상품코드, 평점, 리뷰, 성별, 의류타입과 같은 칼럼을 가진다.

	ProductCode	Rating	Review	sex	type_
0	200367_10007602	5	前回MとLサイズと購入しました。ピチピチ履きたかったので、今回はMサイズのみ購入です。もしか...	0	여성하의
1	200367_10007602	2	かなりのリピーターです。何年も前からずっと気に入って購入し続けて来ましたが、今回購入した商品...	0	여성하의
2	200367_10007602	2	レビューがよかった為2本購入しました。169センチ、55キロ、Lサイズにして大きかったです。...	0	여성하의
3	200367_10007602	3	皆さんのレビューを読み漁り、悩みに悩んでLを注文しましたが、やっぱり大きかった(>人<;)ス...	0	여성하의
4	200367_10007602	4	ブラックとブラックグレープを購入しました。伸縮性はめっちゃあるって訳では無いです。でも履いて...	0	여성하의
5	200367_10007602	4	１５２センチ、４３キロ。太ももがわりとしっかりした、ややO脚気味です。サイズは履いた感じはぴ...	0	여성하의
6	200367_10007602	5	履きやすくてきにいりました。ありがとうございました	0	여성하의
7	200367_10007602	1	170cm入らなかった。返品したい。	0	여성하의
8	200367_10007602	4	レビューを見て、ストレッチが効いてるならとMサイズ購入。162ｃｍ、53~4ｋｇ辺り。ぽっこ...	0	여성하의
9	200367_10007602	2	158cm、43kg、ヒップヌードサイズ84cmくらいです。普段Sサイズのボトムを着用してい...	0	여성하의

[그림4]

Datacleaning [Window10, Python 3.6.5, Jupyter Notebook]

[Step1] 정규화(Normalization)

데이터 전처리를 위해 리뷰 텍스트를 정규화하였다. 먼저 중복되는 상품 리뷰 데이터를 제거하고, 각 상품에 해당하는 리뷰 데이터를 그룹화 하는 작업을 진행하였다. 정규화의 결과는 용어이며, 용어는 정규화된 단어타입으로 색인 대상이 되기 때문에 먼저 일본어 문장의의 특징을 이해해야하고 그 특징에 맞게 전처리해야한다.

예시문장은 다음과 같다. 문장을 보면

お友達の紹介で、女子２人で三時のティータイムに利用しました。2人用のソファに並んでいただきまゝす　v(^^)v　なかよし（笑）最後に出された,モンブランのｹｰｷ。やばっっっ！！これはうまーーーい!!とってもＤｅｌｉｃｉｏｕｓで、サービスもGoodでした□これで2,500円はとってもお得です☆h

1. 전각과 반각의 혼재 ("２"와"2", "Ｄｅｌｉｃｉｏｕｓ"と"Good", "モンブラン"と"ｹｰｷ")
일본어는 숫자와 영문자, 카타카나가 사용 될 경우 폰트 사이즈가 차이가 난다.
2. 숫자의 자리수(,) 桁区切りの"," ("3,000")
이 상태에서 형태소분석을 하면 3과 000이 분리되어 표시된다.
3. 문자 중복표현 ("やばっっっ！！" と "うまーーーい!!")
4. 이모티콘 사용 ("v(^^)v" と "（笑）"
5. 그림문자사용 ("□")

다음은 위의 특징을 제거한 결과이다.

お友達の紹介で、女子　人で三時のティータイムに利用しました。　人用のソファに並んでいただきますv　　vなかよし 笑 最後に出された モンブランのケーキ。やばっっっ　これはうまーい　とってもDeliciousで、サービスもGoodでしたこれで 円はとってもお得です

1. 전각과 반각을 통일. (Delicious, Good)
2. 숫자를 제거하고 ‘ ’ 공백 삽입. (女子２人 -> 女子 人)
3. 정규화 (やばっっっ！！これはうまーーーい!! -> やばっっっ　これはうまーい)
4. 이모티콘 제거 ("v(^^)v" と "（笑）"　-> v　　v
5. 그림문자제거 (" ")

[Step2] 일본어 형태소분석(Morphological analysis, POS tagging)

일본어 문장은 영어, 한국어와는 다르게 문장에 띄어쓰기가 나타나지 않는 특징을 가지기 때문에 이에 앞서 구문분석 및 의미분석을 위해 띄어쓰기를 할 필요성이 있다.
형태소 분석을 위해 打田氏가 만든 일본어 형태소분석 패키지(Python) Janome를 사용하였다. (MacOS, Linux의 경우 Mecab 패키지 설치가 간단하여 형태소 분석에 폭넓게 사용되지만 Window에서 사용하기 위해서는 패키지 설치를 위한 Path설정, Mecab 폴더 내부에 있는 Setup.py 코드수정과 단어 사전이 따로 필요하므로 설치 과정이 복잡하다.)

Tokenization(토큰화)

토큰화는 문서를 쪼개서 작은 단위의 토큰으로 만드는 작업이다. 토큰화의 시작은 특정 문자 제거, 문장부호 제거 등으로부터 시작하는 데, 토큰화는 해당 언어 특징에 매우 종속적이며 올바른 토큰화 과정을 정의하는 것은 간단하지 않다.

Stop Words(불용어)

문서 내에서 별다른 뜻이 없는 단어들을 Stop words라고 하며 영어의 경우 a,an,and,are,as,at,be와 같은 단어들이 있고 일본어의 경우 あそこ(저기),あたり(~쪽),あちら(저쪽),あっち(저쪽),あと(나중),あなた(당신),あれ(저거),いくつ(몇개),いつ(언제),いま(지금)의 경우로 본 연구에서는 명사(動詞), 형용사(形容詞), 동사(動詞), 부사(副詞)를 추출하여 불용어를 제거한다.

Lemmatization(표제어 추출)

Lemma는 한글로 번역하면 '표제어' 또는 '기본 사전형 단어'의 뜻으로 Lemmatization은 단어들로부터 Lemma(표제어)를 찾아가는 과정으로 단어들이 서로 다른 모습을 가지더라도, 그 뿌리 단어(root)를 가지기 때문에 그 단어들의 모습을 어근화 해주는 과정이라 할 수 있다. 예를들어 am,are,is는 서로 다른 모습을 하지만 그 뿌리 단어는 be라고 볼 수 있다. Lemmatization은 문맥을 고려하며, 수행했을 때 결과는 해당 단어의 품사 정보를 보존한다.

Stemming(어간추출)

Stem(어간)을 추출하는 작업을 Stemming이라고 한다. 어간 추출은 형태소분석을 단순화한 버전이라고 볼 수 있고, 정해진 규칙만 보고 단어의 어미를 자르는 어림짐작의 작업이라고 볼 수 있다. 섬세한 작업이 아니기 때문에 어간 추출 후에 나오는 결과 단어는 사전에 존재하지 않는 단어일 수도 있다. 예를들어 this -> thi, was -> wa 처럼 어간의 s를 제거하는 Stemming의 경우 사전에 없는 단어들이 추출된다.

본 연구에서는 일본어 형태소분석에 사용되는 janome 라이브러리를 설치하여 상품리뷰 텍스트에 대해 형태소분석을 진행했다. 해당 라이브러리는 토큰화, 품사태깅, 표제어추출, 어간추출과 같은 자연어 처리 기법을 포함한다.

예시로 [그림4]의 6번 인덱스에 있는 문장에 대해 형태소분석을 진행해보았다.

'履きやすくてきにぃりました。ありがとうございました'
(신발을)신기 편했습니다. 감사합니다.

履き 動詞,自立,*,*,五段・カ行イ音便,連用形,履く,ハキ,ハキ
やすく 形容詞,非自立,*,*,形容詞・アウオ段,連用テ接続,やすぃ,ヤスク,ヤスク
て 助詞,接続助詞,*,*,*,*,て,テ,テ
き 動詞,非自立,*,*,カ変・クル,連用形,くる,キ,キ
に 助詞,格助詞,一般,*,*,*,に,ニ,ニ
いり 動詞,自立,*,*,五段・ラ行,連用形,いる,イリ,イリ
まし 助動詞,*,*,*,特殊・マス,連用形,ます,マシ,マシ
た 助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
。 記号,句点,*,*,*,*,。,。,。
ありがとう 感動詞,*,*,*,*,*,ありがとう,アリガトウ,アリガトー
ござい 助動詞,*,*,*,五段・ラ行特殊,連用形,ござる,ゴザイ,ゴザイ
まし 助動詞,*,*,*,特殊・マス,連用形,ます,マシ,マシ
た 助動詞,*,*,*,特殊・タ,基本形,た,タ,タ

문장을 토큰으로 나누면 여러 품사들로 토큰이 나뉜다. 여기서 조사, 조동사, 기호 등을 제외한 명사(動詞), 형용사(形容詞), 동사(動詞), 부사(副詞)를 추출해 전처리를 진행하였다.

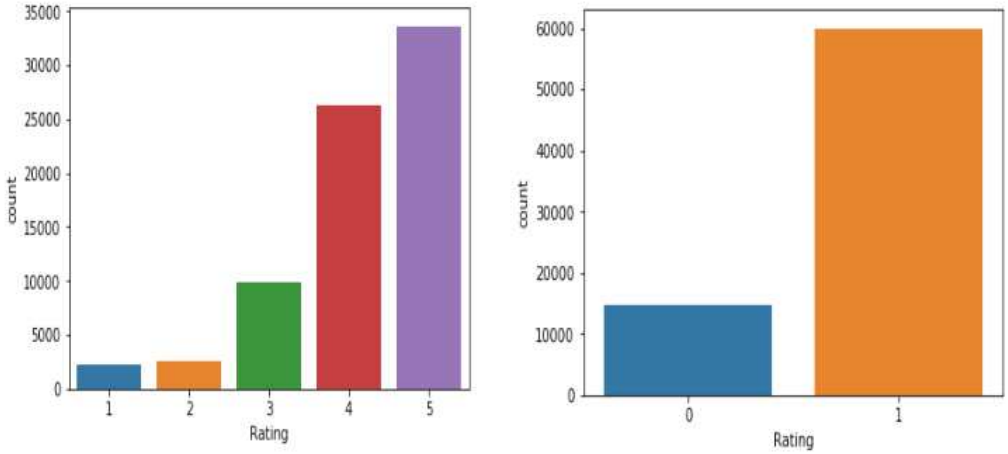
전처리전 문장
'履きやすくてきにぃりました。ありがとうございました' [신기 편했습니다.감사합니다.]

전처리후 문장
[履く, やすぃ, くる, いる] = [신다(동사),싸다(형용사),오다(동사),있다(동사)]

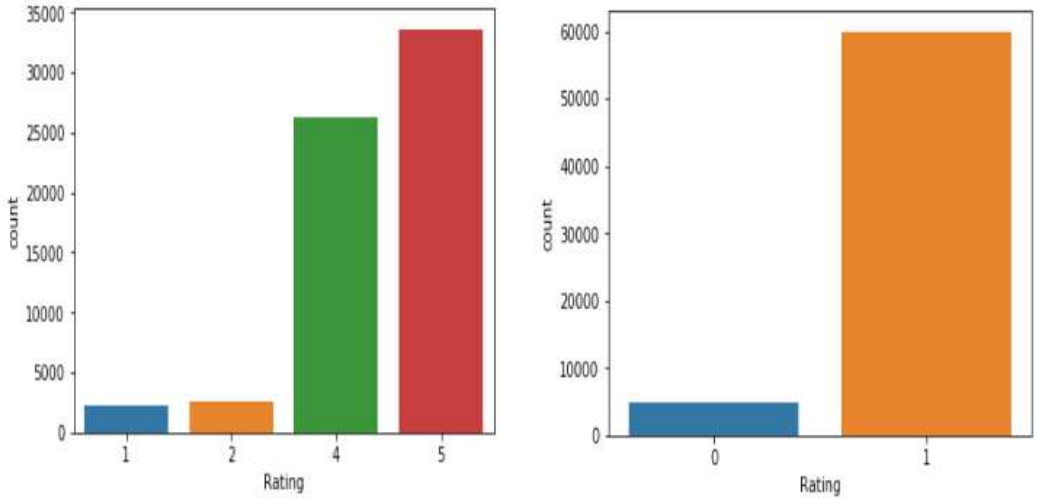
결과적으로 [履く, やすぃ, くる, いる]라는 배열이 추출 되었지만 전처리전 문장의 의미인 [신기 편했습니다. 감사합니다]와는 다른 결과가 나왔는데 이는 janome 라이브러리의 토큰화, 표제어추출, 어간추출 알고리즘을 통해 履きやすく(신기 편한) 에서 [履く,やすぃ]를 추출하였기 때문이다. 이 결과는 분명히 문제가 있는데 의미적으로 본다면, 動詞(동사) (ます형) +やすぃ가 결합되면 ~하기 쉽다, ~하기 편하다는 의미를 가져 履き + やすく(신기 편한)의 의미가 되지만 토큰화로 분리된 단어 각각의 의미를 본다면 [신다, 싸다]라는 결과가 나오게 된다. 또, 문장에 나타난 きにぃり의 경우 본래 きにぃる[気に入る] 마음에 들다, 기분에 맞다 라는 의미를 가지지만 き(動詞),に(助詞),ぃり(ぃる)로 나눴기 때문에 본래의 의미가 사라졌으며 만약 고객이 き(気,마음)를 한자로 작성하지 않고 히라가나로 썼을 경우 본래의 의미를 잃어버리는 한계를 가진다. 이는 언어의 특성 때문인데 일본어는 띄어쓰기가 없으며 동일한 발음에 대해 히라가나, 카타카나, 한자로 다르게 표현할 수 있기 때문에 의미론적으로 완벽히 텍스트를 분류하는 형태소 분석기를 구현하기 어려운 점이 있다.

[Step3] 이진화

본 연구의 목적은 일본어로 작성된 고객의 리뷰가 긍정인지 부정인지 분류 하는데 있어 높은 정확도를 보이는 좋은 성능의 모델을 구축하는데 있다. 이를 위해서 분류 문제에 맞는 전처리가 필요하다. 먼저 1점~5점까지의 분포 중 1~3점은 0, 4~5점은 1로 점수를 이진화 해보았다. 그 결과 [그림 5]는 74700개의 Rating 데이터를 리뷰를 기반으로 변환된 점수의 분포를 보여준다. [그림6]은 64888개의 Rating 데이터 분포이다. 여기서는 [그림5]에서 3점인 리뷰를 제거하였고 결과적으로 [그림6]에서 더욱 뚜렷한 분포 차이를 확인 할 수 있었다. 본 연구에는 3점을 제거한 뒤 (1점,2점) -> 부정의견, (4점,5점) ->긍정의견으로 간주하고 이를 각각 0과 1로 이진화 시켰다. 각각 4916건, 59972건의 데이터를 가진다.



[그림5]



[그림6]

1. 단어사전을 이용한 극성분석

극성 분석은 단어들이 가지는 의미지향성(Semantic orientation)을 분석해 그 단어가 긍정인지 부정인지 추출하는 연구이다. 예를들어 生(삶), 幸福(행복). 優秀(우수)와 같은 단어들은 긍정적인 의미로 죽음(死), 불행(不幸). (열악)劣悪의 경우는 부정의미를 가진다. 이처럼 "단어들은 긍정과 부정의 의미지향성(Semantic orientation)을 가지며 단어식별에 있어서 극성(spin polarization of electron)분석을 통해 단어들의 의미지향성을 추출하고 가치화하는 것은 중요한 작업이다"[1] 이를 활용해 실제 모든 텍스트를 읽지 않고도 사람들의 태도를 읽을 수 있을 것이다.영문 텍스트의 경우 오래전부터 극성분석 연구가 진행되었는데[2] 그 중 English Lexicion 프로젝트는 아주 유명하다. 총 816명의 참가자와 6개 대학의 공동 연구를 통해 40,481 words and 40,481 nonwords를 가지는 단어사전이 만들어져 자연어연구에 널리 이용되고 있다[3].

본 연구에서는 단어사전에 근거한 감성분석을 진행하기 위해 일본어 자연어처리연구에 널리 사용되는 단어감정극성대응표(単語感情極性)을 이용하였다. 이는 도쿄 공업대학의 高村大也, 乾孝司, 奥村学가 이와나미국어사전(岩波国語辞書)을 기반으로 진행한 "スピンモデルによる単語の感情極性抽出"(스핀모델에 따른 단어의 감정극성추출)이란 연구 결과로 배포된 대응표이며 명사 49,002개, 동사 4,254개, 형용사 665개, 부사 1207개의 단어 수를 가지며 각 단어별로 -1점부터 1점으로 수치화 되어있다.

긍정단어의 단어점수
優れる(훌륭하다) : すぐれる:動詞:1
良い(좋다) :よい:形容詞:0.999995
喜ぶ(기뻐하다) : よろこぶ:動詞:0.999979
褒める(칭찬하다) : ほめる:動詞:0.999979
めでたい(경사스럽다) : めでたい:形容詞:0.999645
. . .
부정단어의 단어점수
ない(없다) : ない:助動詞:-0.999997
酷い(심하다) : ひどい:形容詞:-0.999997
病気(아픈) : びょうき:名詞:-0.999998
死ぬ(죽다) : しぬ:動詞:-0.999999
悪い(나쁘다) : わるい:形容詞:-1

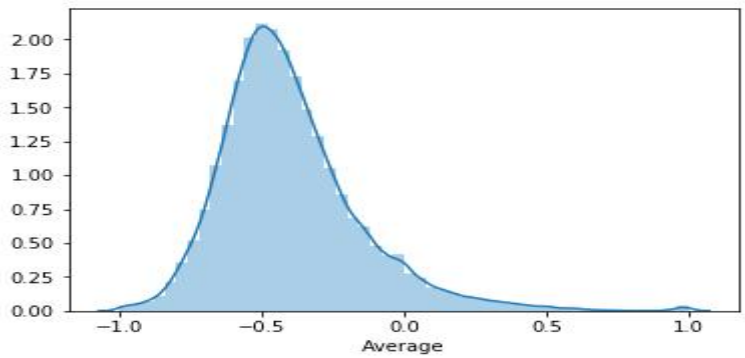
총 64888건의 토큰화된 리뷰 데이터를 대응표의 단어와 1:1로 매칭하여 각각의 단어들이 가지는 극성점수를 평균화하여 순위를 나타내보았다. 추출된 품사는 동사, 명사, 형용사, 부사이며 1점에 가까울수록 긍정적, -1에 가까울수록 부정인 리뷰이다. [테이블1]과 [테이블2]에서 상위 3개, 하위 3개 리뷰와 평균값을 확인할 수 있는데 의미를 본다면 점수가 단어의 의미에 맞게 맵핑됐음을 보여준다.

Idx	Japanese	Korean	Score
0	履く やすい 良い 気持ち 良い	신다 싸다 좋은 기분 좋은	0.999990
1	すごい 物 良い しっかり する いる 着心地 いい 良い 買い物 する	엄청 물건 좋다 똑바로 하다 착용느낌 좋다 좋은 쇼핑 하다	0.999990
2	ちょうど 良い お気に入り 一つ	마침 좋은 마음에 들다 하나	0.999985

[테이블1]

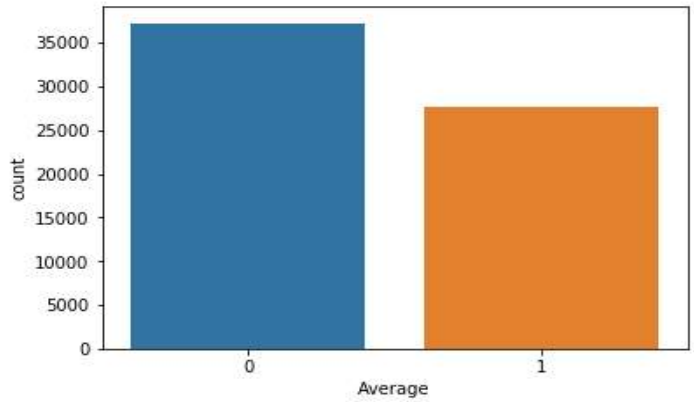
Idx	Japanese	Korean	Score
64885	色合い 悪い かわいい 匂い きつい 臭い 残念	색조 나쁜 귀여운 냄새 힘든 냄새나는 안타까운	-0.996661
64886	あと 色あせる 心配 する いる リピ する	자국 빛이바래다 걱정 하다 재구입 하다	-0.999297
64887	ものすごい 履く やすい ぼつ キツ くい ない	엄청나다 신다 싸다 바가지 씌우다 힘들다 후회 없다	-0.999987

[테이블2]



[그림7]

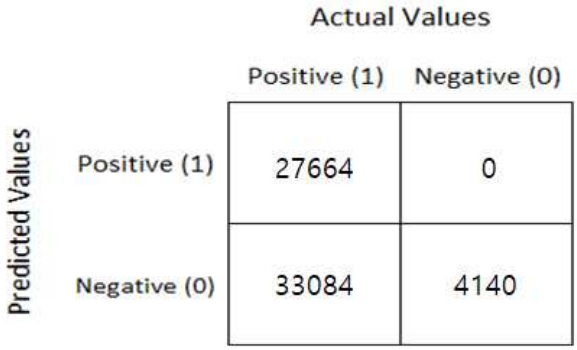
[그림7]에서 Average의 분포를 확인했을 때 평균은 -0.400967이었고 이 값을 기준으로 이보다 작은 감정점수를 가진 리뷰에는 0을, 크다면 1을 설정해 클래스를 구분하였다. [그림9]는 변환된 클래스로 총 64888개의 데이터 가운데 0은 37224개 1은 27664개의 count를 가진다.



[그림8]

분류 모델평가 및 결과

단어사전을 이용한 이진 분류의 성능을 평가하기위해 오차 행렬을 사용해 [그림6]에서의 분류된 데이터를 실제값이라고 하였을 때 [그림9]의 데이터와 비교해보았다.



[그림9]

Accuracy	Precision	Recall	F1-Score
0.49	1.00	0.455	0.63

[테이블3]

정밀도(Precision)는 100%로 모델이 긍정으로 분류한 모든 데이터에 대해 실제값을 정확히 예측했지만 총 데이터 64,888개 중 33,084개를 오분류하여 예측 정확도(Accuracy) 49%를 보였다. 정밀도(Precision)는 높았지만 모델이 부정으로 분류한 데이터가 실제 값과 불일치하는 케이스가 많아져(33,084개) 재현률(Recall)이 0.455로 낮아졌기 때문에 F1점수가 낮아졌다. 총 64,888개의 데이터의 평점 평균이 4.19이고 실제 긍정 값이 59,972개, 부정 값이 4,916개를 가져 극단적으로 긍정리뷰가 많았기 때문에 평균이 양값을 가질거라 예상했지만 결과는 음값이 나왔다. 이는 일본어는 히라가나, 카타카나, 한자를 통해 동일한 내용을 다르게 표현할 수 있기 때문에 리뷰데이터에는 있지만 단어감정극성대응표(単語感情極性)에는 그 단어가 없었기 때문에 분석에서 제외됐거나 긍정적, 중의적인 단어들에 음값을 매겨 실제 그 단어가 가지는 의미지향성(Semantic orientation)을 올바르게 해석해내지 못했기 때문이라고 해석할 수 있다. 영문 텍스트를 전처리할 때는 정규화 과정에서 대문자를 소문자로 통일하지만 일본어는 3가지의 다른 표기체계를 하나로 통일한다면 분석은 간단해질 수 있지만 한자가 없어질 경우 가독성이 떨어지고 본래의 의미가 훼손될 수 있기 때문에 극성 점수에 의미지향성(Semantic orientation)을 올바르게 반영하기 어려울 것이다. 단어감정극성대응표(単語感情極性)는 이와나미국어사전(岩波国語辞書)을 기반으로 만들어졌기 때문에 새롭게 생겨나는 신조어를 점수에 반영하지 못하는 단점도 확인할 수 있다.

2. TF-IDF(Term Frequency-Inverse Document Frequency)를 이용한 감성분석

Bag of Words(BoW)

Bag of Words란 단어들의 순서를 전혀 고려하지 않고, 단어들의 출현 빈도(frequency)에만 집중하는 텍스트 데이터의 수치화 표현 방법이다. 단어들의 가방이란 뜻으로 단어의 중복을 허용하고 순서를 무시한다. Bow를 만드는 과정은 다음 2가지 방법이다.

1. 각 단어에 고유한 인덱스를 부여한다.
2 각 인덱스의 위치에 토큰의 등장 횟수를 기록한 벡터를 만든다.

(예시)
문서1 : [정부가 발표하는 취업률과 구직자가 느끼는 취업률은 다르다.]

1. 각 단어에 고유한 인덱스를 부여
('정부': 0, '가': 1, '발표': 2, '하는': 3, '취업률': 4, '과': 5, '구직자': 6, '느끼는': 7, '은': 8, '다르다': 9)
2. 각 인덱스의 위치에 토큰의 등장 횟수를 기록한 벡터를 만든다.
[1, 2, 1, 1, 2, 1, 1, 1, 1, 1]

Bow는 각 단어가 등장한 횟수를 수치화하는 텍스트 표현 방법이기 때문에, 주로 어떤 단어가 얼마나 등장했는지를 기준으로 문서가 어떤 성격의 문서인지를 판단하는 작업에 쓰인다.

단어문서개수행렬(Term Document Count Matrix)

단어문서개수행렬이 다수의 문서에서 등장하는 각 단어들의 빈도를 행렬로 표현한 것을 말한다. 즉 각 단어들의 빈도를 행렬로 표현한 것으로 각 문서에 대한 Bow를 하나의 행렬로 만든 것으로 생각할 수 있다. 예를 들어 다음 3개의 문서가 있다고 할 때, 단어문서개수행렬을 만들면 [테이블4]로 나타낼 수 있다.

문서1 : 먹고 싶은 과자
문서2 : 먹고 싶은 라면
문서3 : 저는 우동이 좋아요

	우동이	먹고	싶은	과자	라면	저는	좋아요
문서1	0	1	1	1	0	0	0
문서2	0	1	1	0	1	0	0
문서3	1	0	0	0	0	1	1

[테이블4]

TF-IDF(Term Frequency-Inverse Document Frequency)

단어문서개수행렬은 여러 문서에 등장하는 모든 단어에 대해서 빈도 표기를 하는데 이런 방법은 문서들의 비교, 분석을 어렵게 한다. 분석은 고려해야할 변수들이 많을 수록 어려워지는데 예를들어 the의 빈도수가 많다고 해서 the의 중요도가 높다고 할 수 없기 때문이다. 이때 단어문서개수행렬에 불용어와 중요한 단어에 대해서 가중치를 줄 수 있는 방법으로 TF-IDF라는 방법이 고안되었다.

TF-IDF는 단어의 빈도와, 역 문서 빈도(문서의 빈도에 특정 공식을 취함)를 사용하여 단어 행렬 내의 각 단어들마다 중요한 정도를 가중치로 주는 방법이다. 우선 행렬을 만든 후에, 거기에 TF-IDF 가중치를 주면 된다. TF-IDF는 TF와 IDF를 곱한 값을 의미하는데 이를 식으로 표현해보겠다. 문서를 d, 단어를 t, 문서의 총 개수를 n이라고 표현할 때 TF, DF, IDF는 각각 다음과 같이 정의할 수 있다.

(1) tf(d,t) : 특정 문서 d에서 특정 단어 t의 등장 횟수
앞의 예시에서 [테이블4]에 있는 각 단어들이 가진 값들이다.

(2) df(t) : 특정 단어 t가 등장한 문서의 수
특정 단어가 각 문서, 또는 문서들에서 몇 번 등장했는지는 관심을 가지지 않으며 오직 특정 단어 t가 등장한 문서의 수에만 관심을 가진다. 앞의 예시에서 '먹고','싶은'의 두 단어는 각각 문서1, 문서2에서 각각 2번씩 출현했으므로 df는 2이다. 문서1 또는 문서 2에서 '먹고'라는 단어가 100번 출현했다고 하더라도 단어의 등장 횟수는 중요하지 않고 출현한 문서의 수에 관심을 가진다.

(3) idf(t) : d(f)에 반비례하는 수 $idf(d,f) = \log(\frac{n}{1+df(t)})$

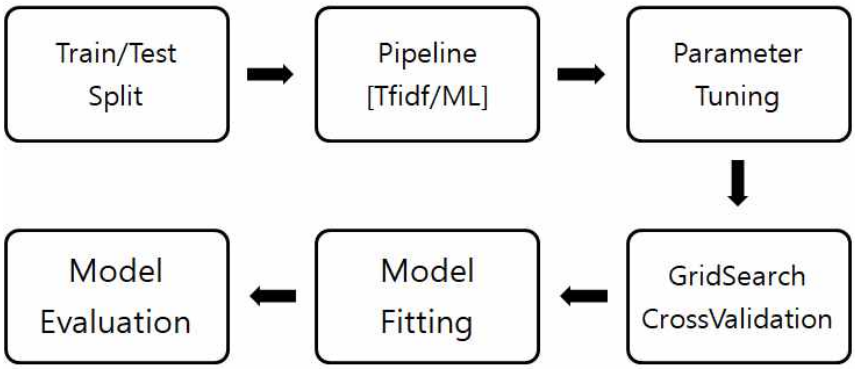
log를 사용해 d(f)의 역수를 표현하고자 하는데 분모에 +1을 더해주는 이유는 d(f)값이 0이 되지 않게 하기 위함이다. log를 취하지 않았을때 idf(d,f)값이 총 문서의 수가 커질수록 기하급수적으로 커지게 되기 때문에 log를 취해준다. 이 값은 여러 문서에서 등장한 단어의 가 중치를 낮추는 역할을 하게된다. TF-IDF는 (1)에서 구한 tf(d,f)값과 idf(t) 값을 곱한 값으로 모든 문서에서 자주 등장하는 단어는 중요도가 낮다고 판단하며, 특정 문서에서만 자주 등장하는 단어는 중요도가 높다고 판단한다. 즉 the나 a와 같이 자주 등장하는 불용어의 TF-IDF값은 다른 단어의 TF-IDF에 비해서 낮아지게 된다.

Ngram

BOW 표현 방식은 단어의 순서가 완전히 무시된다는 큰 단점이 있다. 의미가 완전 반대인 두 문자열 "it's bad, not good at all"과 “it's good, not bad at all"이 완전히 동일하게 변환되며 이러한 문제를 해결하기 위해 토큰 하나의 횟수만 고려하지 않고 연속된 n개의 토큰을 함께 고려하는 방법이 생기게 됐다.

TF-IDF를 이용한 감성분석 설계

단어행렬의 자질 추출을 위해(feature_extraction) 싸이킷런의 TfidfVectorizer 클래스를 이용한다. 해당 클래스는 리뷰 텍스트에 있는 단어들의 TF-IDF 점수를 특징벡터에 투영 시킨다 연구에 사용된 데이터는 형태소분석을 마친 원 데이터(A)와 평점이 3점인 리뷰를 제거한 데이터(B)로 나누어 모델의 정확도를 비교한다. [그림10]은 분석 설계도로 데이터(A,B)를 각각 훈련(70%) /테스트(30%) 셋으로 랜덤샘플링하고 Pipeline을 사용해 비정형 데이터 분석에 널리 이용되는 다양한 분류모델(ML)과 전처리(Tfidf)방법을 하나의 추정기로 연결하였으며, 그리드서치 교차 검증을 통해 매개변수들을 대상으로 가능한 여러 가지 조합을 시도해 일반화 성능을 개선하고자 하였다. 모델 적합 후 평가 단계에서는 모델의 일반화 성능 및 최적 파라미터를 점검한다. [테이블5]는 분석에 이용된 하이퍼파라미터를 보여주고 있다.



[그림10]

ML	Parameters	Explanations	value
RandomForest	max_depth n_estimators	The maximum depth of the tree The number of trees in the forest.	[10,30,50,70,100] [100,150,200,300,500]
Multinomial NaiveBayes	alpha	Additive (Laplace/Lidstone) smoothing parameter	[0.0,0.5,1.0]
Stocastic GradientDescent	alpha	Constant that multiplies the regularization term.	[7e-06~ 6e-05]
Logistic Regression	C	Inverse of regularization strength; must be a positive float. Like in support vector machines, smaller values specify stronger regularization.	[0.1,1,10]
TermFrequency Index DocumentFrequency	min_df ngram	When building the vocabulary ignore terms that have a document frequency strictly lower than the given threshold The lower and upper boundary of the range of n-values for different n-grams to be extracted	[3,5,7] [1,2,3]

[테이블5]

분류 모델평가 및 결과

사용자 감성 분석을 위한 긍정, 부정 분류모델을 학습하기 위해 TF-IDF를 이용하여 텍스트 분류에 사용되는 다양한 기계학습 알고리즘을 적용해보았다. [테이블6]는 교차검증(Cross Validation)을 통해 선정된 최적의 파라미터와 분류기별 모델평가에 사용되는 정밀도 (Precision), 재현률(Recall), F1-score과 정확도(Accuracy) 결과치를 보여준다. 이 4가지 지표는 데이터 A를 사용해 훈련모델 적합 후 도출한 최적 매개변수를 테스트 데이터와 다시 적합하여 도출한 최종교차검증 점수이다. 평가에 사용되는 모델은 모든 척도에서 LogisticRegression 분류기가 가장 좋은 성능을 보였기 때문에 이를 기준으로 사용한다. [테이블7]은 데이터 B를 사용해 같은 분류 방법으로 나타난 결과를 보여주는데 모든 모델의 성능이 큰 폭으로 향상되었고 특히 Logistic Regression의 경우 모든 지표에서 95%가 넘는 성능을 보였다. 이는 분류모델이 평점 3점을 가진 리뷰데이터를 정확히 분류하지 못했기 때문이다. LogisticRegression은 선형모델분류기로 고차원데이터의 이진 분류문제에서 강력한 성능을 보인다. TF-IDF 단어행렬의 텍스트 분류 문제는 매우 많은 자질을 가지는 고차원의 분류문제이기에 이러한 분류기를 사용하기 적합하다. 매개변수로 사용된 C는 규제의 강도를 결정하는데 C값이 변함에 따라 클래스를 구분하는 결정경계가 이동한다. C값이 작을수록 규제가 많이 적용되며 모델의 복잡도는 낮아진다. 모델이 과소적합된 경우 C값을 높혀 성능 향상을 기대할 수 있는데 [테이블6]과 [테이블7]을 보면 C값이 10으로 증가해 성능이 향상됐음을 확인할 수 있다.

[Data A : N=74.700 Train =52,290 Test = 22,410]

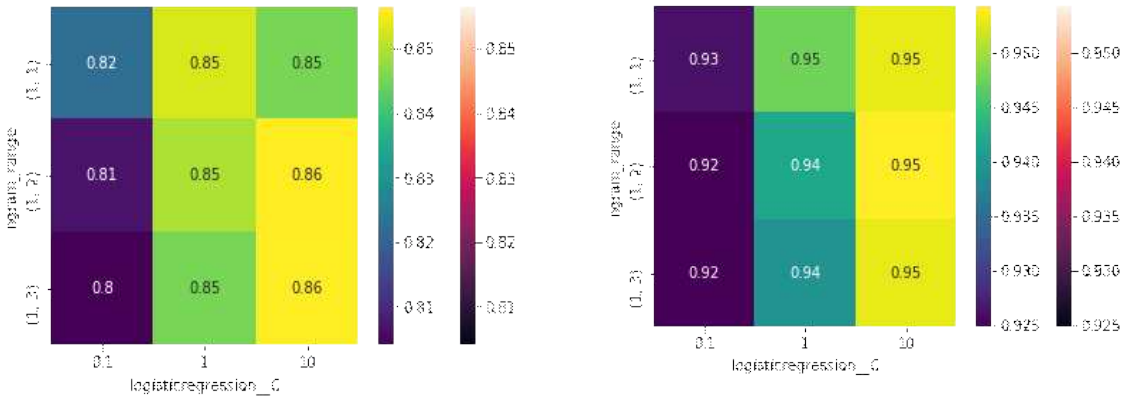
ML	ML_Best parameters	TF_IDF_Best parameters	Precision	Recall	F1score	Accuracy
RandomForest	max_depth:100 n_estimators:150	min_df: 7 ngram: 1	0.8251	0.8277	0.7801	0.828
Multinomial NaiveBayes	alpha: 0.5	min_df: 7 ngram: 3	0.8344	0.8457	0.8196	0.846
Stocastic GradientDescent	alpha: 6e-05	min_df: 7 ngram: 3	0.8458	0.849	0.8189	0.849
Logistic Regression	C: 1	min_df: 7 ngram: 2	0.845	0.8548	0.8359	0.855

[테이블6]

[Data B: N=64.888 Train=45,421 Test=19,467]

ML	ML_Best parameters	TF_IDF_Best parameters	Precision	Recall	F1score	Accuracy
RandomForest	max_depth:100 n_estimators:500	min_df: 7 ngram: 1	0.9375	0.9380	0.9178	0.938
Multinomial NaiveBayes	alpha: 0.5	min_df: 7 ngram: 2	0.9446	0.9453	0.9312	0.945
Stocastic GradientDescent	alpha: 6e-05	min_df: 7 ngram: 3	0.946	0.9488	0.9393	0.949
Logistic Regression	C: 10	min_df: 7 ngram: 3	0.9505	0.9544	0.9502	0.954

[테이블7]



[그림11]

TF-IDF 점수가 낮은 자질		TF-IDF 점수가 높은 자질	
자질	의미	자질	의미
'乗る ため'	타기 위해	'プレゼント'	선물
'ただ ドレス'	그냥 드레스	'オレンジ'	오렌지
'角度 変わる'	각도 변화	'いい'	좋은
'予定 ベビー'	예정 아기	'あったかい'	따뜻한
'状態 ファスナー'	상태 지퍼	'気に入る'	마음에든
'少し 汚れる'	조금 더러운	'仕事'	일
'多少 入れる'	다소 넣다	'身長'	키
'多少 広げる'	다소 넓히다	'彼氏'	남자친구

[테이블8]

교차검증을 통한 성능개선

[그림11]는 min_df값을 고려하지 않았을 때, C의 규제 정도와 Ngram에 따라 변화된 정확도를 보여주는 히트맵 그래프로. 좌측의 그래프는 C값이 10이고 N값이 2,3일 때 정확도가 가장 높았으며, 우측의 그래프는 C값이 10이고 N값이 2,3일 때 정확도가 가장 높다. 이 결과는 파라미터 조정을 통해 성능을 기대할 수 있음을 보여준다. [테이블8]을 볼 때, 자질 추출에서 바이그램(n=2)과 유니그램(n=1)이 사용되었고 낮은 점수의 자질들은 부정클래스로 높은 점수의 자질들은 긍정클래스로 분류되며 점수가 낮을 때는 의미가 불분명하고 부정적인 단어가 나타나지만 높은 점수를 가지는 자질의 경우 의미가 분명하며 긍정적인 의미의 단어가 나타난다.

불균형데이터

분석에 사용된 데이터는 [그림6]과 같이 긍정 리뷰의 비율이 극단적으로 높은 불균형 클래스의 데이터셋이다. 이러한 데이터셋을 분류하는 문제는 정확도만으로 모델의 성능을 평가하는데 무리가 있다(Accuracy paradox). 불균형한 데이터를 사용한 모델의 성능을 평가할 때는 정밀도와 재현율의 조화 평균인 F1점수가 더 나은 평가지표가 될 수 있는데. 정밀도(Precision)과 재현율(Recall)을 같이 고려하기 때문이다. 불균형 문제에서 최종 모델을 선택하는 것은 어려운 작업이며 이를 해결하기 위한 방법으로 데이터를 추가하여 균형을 맞추거나 ROC curve를 그려 결정 함수의 임계값을 조정하며 지표의 변화를 확인할 수 있다.

3. 토픽모델링과 문서군집화

소셜 미디어의 확산과 텍스트, 음성과 같은 비정형 데이터를 처리하는 빅데이터 분석 기술의 발달에 따라 데이터로부터 소비자의 의견 및 트렌드를 이해하고 이를 마케팅 전략으로 사용하는 오피니언마이닝(OpnionMining)기법이 널리 활용되고 있다. 토픽모델링은 이러한 오피니언 마이닝 및 소비자 동향 분석에 사용될 수 있는 감성분석의 방법 중 하나이다. 토픽모델링은 텍스트 데이터에 자주 적용되는 비지도학습으로 문서를 하나 또는 그 이상의 토픽으로 할당하는 작업을 말하며, 한 문서가 하나의 토픽에 할당되면 같은 문서를 군집화 하는 방식이다. 가장 대표적인 토픽모델링 방법은 Blei et al.(2003)[4]의 LDA(Latent Dirichlet Allocation)로 다수의 문서에서 잠재적으로 의미 있는 토픽을 찾는 절차적 확률 분포 모델 이다. LDA는 단어들의 집합이 어떤 토픽들로 묶인다고 가정하고, 이 단어들이 각각의 토픽 에 구성될 확률을 계산하여 결과 값을 토픽에 해당할 가능성이 높은 단어들의 집합으로 추출한다. 이렇게 추출된 토픽은 의미가 있다 하더라도 단어 집합의 주제가 아니기 때문에 결과를 이해하는데 있어 주관적인 의견이 개입된다.

이전 감성분석에 사용된 데이터(n=64888)를 LDA를 적용해 10개의 토픽을 얻었지만 이 중 단어를 통해 주제를 유추할 수 있는 결과는 다음 4가지 토픽이었다. 먼저 [테이블9]의 토픽5의 단어를 보면 상품의 싼 가격, 꼼꼼한 상품, 좋은 디자인이란 요소에 대해 고객만족도가 높음을 알 수 있다. 토픽6은 드레스, 결혼식, 귀여운, 디자인과 같은 중요 단어를 통해 이 토픽이 웨딩드레스라는 주제를 가지고 있음을 유추 할 수 있다. 마지막으로 토픽9는 상품 판매자의 대응, 얼마나 빨리 상품을 받을 수 있는지 같이 배송이라는 토픽과 연관된 단어들로 구성되었다. 나머지 7개의 토픽모델은 着心地(입기편한), 体型(체형), 暖かい(따뜻한), 浴衣(욕의)와 같이 의류카테고리에서 나올 수 있는 단어들로 구성되었지만 같은 주제로 묶기는 어려웠다.

토픽5	
단어	의미
安い	싸다
満足	만족
できる	가능하다
価格	가격
しっかり	꼼꼼히
商品	상품
よい	좋은
もの	물건
いい	좋은
デザイン	디자인

토픽6	
단어	의미
結婚式	결혼식
見える	보이다
可愛い	귀여운
少し	조금
れる	~되다
そう	그렇게
長い	길다
ドレス	드레스
レビュー	리뷰
デザイン	디자인

토픽9	
단어	의미
届く	도착하다
商品	상품
注文	주문
対応	대응
発送	발송
せる	시키다
満足	만족
すぐ	바로,금방
早い	빠른
いただく	받다

[테이블9]

논의 및 결론

패션시장은 변화의 속도가 굉장히 빠르다. 불과 며칠 전 유행했던 아이템의 매출이 갑자기 증가하거나 감소하는 일은 주변에서 어렵지 않게 찾아볼 수 있다. 때문에 기업들은 빠르게 변화하는 시장의 트렌드에 맞춰 상품에 소비자의 취향을 반영해야한다. 전자 상거래를 이용하는 소비자에게 다른 소비자들이 작성한 상품에 대한 평가 데이터는 그 상품의 구입 여부를 결정하는데 큰 영향을 미치기 때문에 기업의 입장에서도 상품에 대한 소비자의 평가 데이터를 분석하여 소비자의 피드백을 반영한다면 기업의 성과에 영향을 미칠 것이다.

라쿠텐(楽天) 마켓은 일본 최대의 쇼핑몰로 출점 상점 수 4만개, 상품 수 1억 4천만점, 연간 유통 1.4조엔에 이르는 대형 쇼핑 플랫폼으로 본 연구에서는 라쿠텐 마켓에서 2018년 6월~7월경 여성, 남성의류 카테고리에 판매된 상품 리뷰 데이터를 수집해 사용자 감정분석에 적용해보았다. 소비자가 작성한 상품 리뷰의 내용을 바탕으로 소비자가 상품에 대해 긍정적인 의견을 가지고 있는지, 부정적인 의미를 가지고 있는지를 분류할 수 있는 모델을 구축함으로써 소비자가 등록한 평점 이외에 보다 객관적인 방식으로 소비자의 의견을 파악할 수 있는 방법을 제안해보았다.

먼저 단어사전을 이용해 감성분석을 진행했다. 일본어 자연어 처리연구에 널리 사용되고 있는 단어감정극성대응표(単語感情極性)에 나타난 극성점수를 리뷰 데이터와 대응시켜 평균화된 점수를 구했다. 정밀도는 높았지만 오분류율이 높아 49%의 낮은 정확도를 보였다.
일본어는 히라가나, 카타카나, 한자의 3가지 다른 표기 체계를 가지는데 이들을 하나로 정규화하기 어렵기 때문에 모든 조합을 단어감정극성대응표(単語感情極性)에 반영하기는 어렵다. 그리고 새롭게 생겨나는 신조어를 결과에 반영하지 못한다. 이러한 문제들은 일본어 자연어처리 연구가 해결해 나아가야할 방향일 것이다.

다음으로 형태소분석을 마친 리뷰 데이터를 원 데이터(A)와 평점이 3점인 리뷰를 제거한 데이터(B)로 나누고 TF-IDF 단어문서행렬을 만들었다. TF-IDF는 각 문서를 구성하고 있는 단어들을 빈도와, 역 문서 빈도(문서의 빈도에 특정 공식을 취함)를 사용하여 단어행렬 내의 각 단어들마다 중요한 정도를 가중치로 주는 방법이다. 이렇게 형성된 자질을 교차검증(Cross Validation)을 하여 분류 모델별 최적의 파라미터와 성능지표를 얻은 결과 소비자의 긍정,부정 의견을 파악하는 모델의 성능은 LogisticRegression 분류기가 95%의 정확도를 나타내 가장 좋은 성능을 보였다. 그리고 결정함수의 규제(C)값과 N-gram조정을 통해 자질 구성을 다양화시켜 모델의 성능을 개선할 수 있었다.

마지막으로 토픽모델링을 통해 라쿠텐을 이용한 소비자들이 어떤 상품, 서비스를 중요하게 생각하는지 확인할 수 있었다. 기업은 토픽 모델링을 통해 낮은 평점을 준 고객이 남긴 텍스트를 분석해 핵심 용어를 추출하여 원인을 요약하고 의사결정에 반영할 수 있다.
"비정형 텍스트인 사용자의 의견 정보에 따른 패션 트랜드를 분석하고 의류 마케팅에 활용하는"[5] 연구는 다양한 곳에서 진행 중에 있으며 기업은 이를 마케팅 전략으로 활용해 고객들이 만족을 느끼는 상품을 추천하고 문제점을 빠르게 해결할 수 있을 것이다.

참고문헌

[1] 高村大也, 乾孝司, 奥村学
"スピンモデルによる単語の感情極性抽出", 情報処理学会論文誌ジャーナル, Vol.47 No.02 pp. 627--637, 2006.
Hiroya Takamura, Takashi Inui, Manabu Okumura, "Extracting Semantic Orientations of Words using Spin Model", In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL2005), pages 133--140, 2005.
[2] Philip J Stone; et al
The General Inquirer : a Computer Approach to Content Analysis
Cambridge : The M.I.T. Press, 1966.
[3] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll and Manfred Stede
Lexicon-Based Methods for Sentiment Analysis
Posted Online May 26, 2011 https://doi.org/10.1162/COLI_a_00049
© 2011 Association for Computational Linguistics
Computational Linguistics Volume 37 | Issue 2 | June 2011 p.267-307
[4]Blei et al.(2003)
Latent Dirichlet Allocation
Article in Journal of Machine Learning Research 3(4-5):993-1022 · May 2003
[5]이윤주, 서지훈, 최진탁(2014)
SNS 텍스트 콘텐츠를 활용한 오피니언마이닝 기반의 패션 트랜드 마케팅 예측 분석
Journal of KIIT. Vol. 12, No. 12, pp. 163-170, Dec. 31, 2014. pISSN 1598-8619, eISSN 2093-757
[6]Andreas Muller
Introduction to Machine Learning with Python