2022 UROP 영역별 감정 사전 기반 분석 보고서

소속: 국민대학교 소프트웨어학부
학번: 20181703
이름: 평선호
지도교수: 박하명 교수
팀원: 윤상원 학우(20181651)

1. 서론

현재 인터넷상의 데이터가 급증함에 따라 사회 또는 기업에서 해당 데이터를 적절히 활용하여 서비스를 제공하고 있다. 데이터를 활용하는 분야 중에서 주위에서 제일 접하기 쉽고 상품을 구매할 때 참고를 많이 하는 리뷰데이터를 선정하였다.

리뷰데이터의 긍정,부정을 판별하고 상품별, 즉 영역별로 감성사전을 구축하는 과정을 통해 단순 긍정, 부정을 평가하는 것이 아닌 상품의 소비방법에 맞는 감성사전을 구축하였다. 이를 통해 리뷰를 입력했을 때 리뷰 속 긍정,부정 단어들을 판별하여 별점예측을 하는 서비스를 제공하고자 한다.

감성사전 기반 감정 분석을 통해 리뷰 속 어휘들의 극성을 판별하고 이를 계산하여 점수화를 진행하였다.

2. 연구배경

2.1 연구데이터 수집

본 연구에서는 다양한 종류의 물픔을 판매하는 전자상거래 사이트인 쿠팡(coupang) 리뷰데이터를 선정하여 영역별로 '가전 디지털', '생활 용품', '여성패션', '스포츠/레저', '완구/취미' 5가지 영영을 선정하여 상품 리뷰 데이터 크롤링을 진행하였다. 이때 상품 데이터 수집은 python의 selenium 라이브러리를 사용하여 크롤링을 통해 진행하였다. 이때 크롤링을 했을 때, 기존 사용자들이 입력한 별점을 바탕으로 데이터의 긍정,부정을 분류하였다. 리뷰 데이터의 약 80~90%정도가 별점 4,5점을 주면서 긍정적인 어휘를 바탕으로 리뷰를 작성하는 것을 확인하여 별점 4,5점을 긍정 리뷰로 판단하고 별점 1,2,3점의 리뷰들은 부정 리뷰로 판단하였다.

2.2 연구 절차

2.2.1 긍정 부정 분류 후 데이터 전처리

본 연구를 진행하는 과정 중, 크롤링한 리뷰데이터를 바탕으로 soynlp 오픈소스를 사용하여 한국어 분석을 진행했다. 크롤링한 리뷰데이터에는 기존 형식화 된 글들이 아닌 불용어가 포함되어 있는 리뷰데이터가 존재하였다. 감성 사전구축에 불필요하게 반복되는 이모티콘, 문구, 영어 리뷰, 오타 수정 등 리뷰 전처리 과정을 불필요한 요소들을 정리하였다.

2.2.2 데이터 태깅 및 어휘 통일 진행

한국어 특성상 하나의 어근에서 파생될 수 있는 많은 단어들이 존재한다. 이러한 한국어의 특징이 존재할 시 사전을 구축하는데 영향이 끼칠수 있으므로 어휘들을 토큰화를 진행하였다. 추가적으로 모든 어휘들을 형태소 분석을 통해 모든 어휘들에 형태소를 태깅을 진행하였다.
Ex) 맛있다, 맛있고, 맛있는 -> 맛있다 / Verb 로 통일을 하며 토큰화를 진행을 하였다.

2.2.3 기준 단어 선정

감성사전을 구축하기 위해 본 연구에서 긍정, 부정 리뷰 데이터속에서 기준 단어를 선정하였다. 해당 기준 단어들은 추후 리뷰 데이터들과 SO-PMI 계산을 통해 어휘들이 극성을 판별 할 수 있는 기준이 되어주기때문에 기준단어라 불린다.
형태소 만으로 극성을 판별하기 힘든 동사,부사 등은 제외하고 극성을 판별하기 쉬운 형용사를 기준단어로 선정하여 리뷰데이터 속 형용사인 어휘들의 빈도 수를 계산하였다. 긍정 리뷰 속 형용사 어휘의 빈도수와 어휘를 쌍으로 계산하고, 부정 리뷰 속 형용사 어휘의 빈도수와 어휘를 쌍으로 계산하였다.
특정 단어의 경우 긍정, 부정의 뜻을 가지는 것이 아니라, 단순히 자주 등장 하는 단어가 기준단어로 선정되는 것을 방지해야한다. 이를 해결하고자 긍정리뷰속에서 구한 기준단어들의 집합과 부정리뷰속에서 구한 기준단어들의 집합을 서로 빼주며 빈도수 문제를 해결하였다.

2.2.4 단어 극성 판별

리뷰데이터의 단어의 극성을 판별하기 위해 이전 과정에서 기준단어를 바탕으로 SO-PMI과정을 진행한다. PMI(Pointwise Mutual Information)란? 두 단어 간의 유사성을 분석하는 지표로 활용되는 방법이다.

$$PMI(w_1, w_2) = \log_2 \frac{p(w_1, w_2)}{p(w_1)p(w_2)}$$

여기서 w1과 w2는 분석하고자 하는 두 단어를 나타내며 PMI(w1,w2)는 한 문서내에서 w1과 w2가 함께 출현할 확률을 w1과 w2가 각각 출현할 확률의 곱으로 나눈 값이다. 따라서 PMI(w1,w2)는 두 단어

가 같은 문서에서 나타날 확률을 나타내며 값이 클수록 두단어 유사성, 즉 극성이 비슷한 것으로 판별 할 수 있다. 하지만 단어의 출현 빈도수만으로 감성극성 판별은 단어의 극성강도를 정확히 판별하기 어려운 문제가 있을 수 있고, 기준단어의 선정에 따라 단어극성 판별의 편차가 크게 달라지는 문제점이 발생한다. 이러한 문제점을 보완하고자 보다 극성을 정확히 판별하는 방법인 SO-PMI(Semantic Orientation From Pointwise Mutual Information)방식을 사용하였다.

$$SO-PMI(w) = \sum_{pw \in PW} PMI(w,pw) - \sum_{nw \in NW} PMI(w,nw)$$

여기서 PW는 긍정기준단어의 집합,NW는 부정기준단어의 집합을 의미한다. SO-PMI(w)는 단어 w와 긍정단어 집합과의 PMI합에서 부정단어 집합과의 PMI합을뺀 결과이다.
PMI를 계산하기 위해 soynlp의 모듈 중 하나인 pmi_func을 불러와 PMI계산을 진행하였다. 본 연구에서 감성 사전에 등재되는 단어는 형용사와 부사단어만을 활용하여 구축하였다. 따라서 전체 리뷰에서 등장하는 형용사와 부사 단어들을 추출하여 긍정 기준단어집합과 부정 기준단어집합과의 PMI를 계산한 다음 이를 빼면서 단어들의 SO-PMI를 계산한다.

2.2.5 감성사전 구축

전체 데이터에서 형용사와 부사들을 기준단어들과 PMI를 계산한다음 이를 활용하여 감성사전 구축을 진행하였다. 점수화된 단어들의 집합들을 SO-PMI 값에 따라 '강한 긍정' , '약한 긍정', '중립' , '약한 부정', '강한 부정' 5종류로 분류하였다. SO-PMI 점수를 기준으로 상위 20%가 '강한 긍정', 상위 20~40%가 '약한 부정', 상위 40~60%가 '중립', 상위 60~80% '약한 부정', 상위 80~100%가 '강한 부정'으로 분류하였다.

2.2.6 감성 사전을 활용한 평점 예측

테스트 데이터 리뷰 집합의 감성 점수를 측정한다. 해당 리뷰의 물품이 속한 영역의 감성 사전을 사용하여 '강한 긍정' 단어는 +2, '약한 긍정' 단어는 +1, '중립' 단어는 0, '약한 부정' 단어는 -1, '강한 부정' 단어는 -2점으로 계산한다. 이때 해당 사전을 json형식으로 파일화 하여 리뷰를 입력하면 json형식의 사전에 등재된 단어들을 점수로 변환하여 모두 합한 값을 단어의 수로 나누어 최대값이 2와 최소값이 -2로 나오게 고정한다.


```
"data":[
    {
        "index":0,
        "text":"좋다",
        "score":2
    },
    {
        "index":1,
        "text":"편하다",
        "score":2
    },
    {
        "index":2,
        "text":"고급스럽다",
        "score":2
    },
    {
        "index":3,
        "text":"만족하다",
        "score":2
    },
    {
        "index":4,
        "text":"깔끔하다",
        "score":2
    },
    {
        "index":5,
        "text":"예쁘다",
        "score":2
    },
    {
        "index":6,
        "text":"가볍다",
        "score":2
```

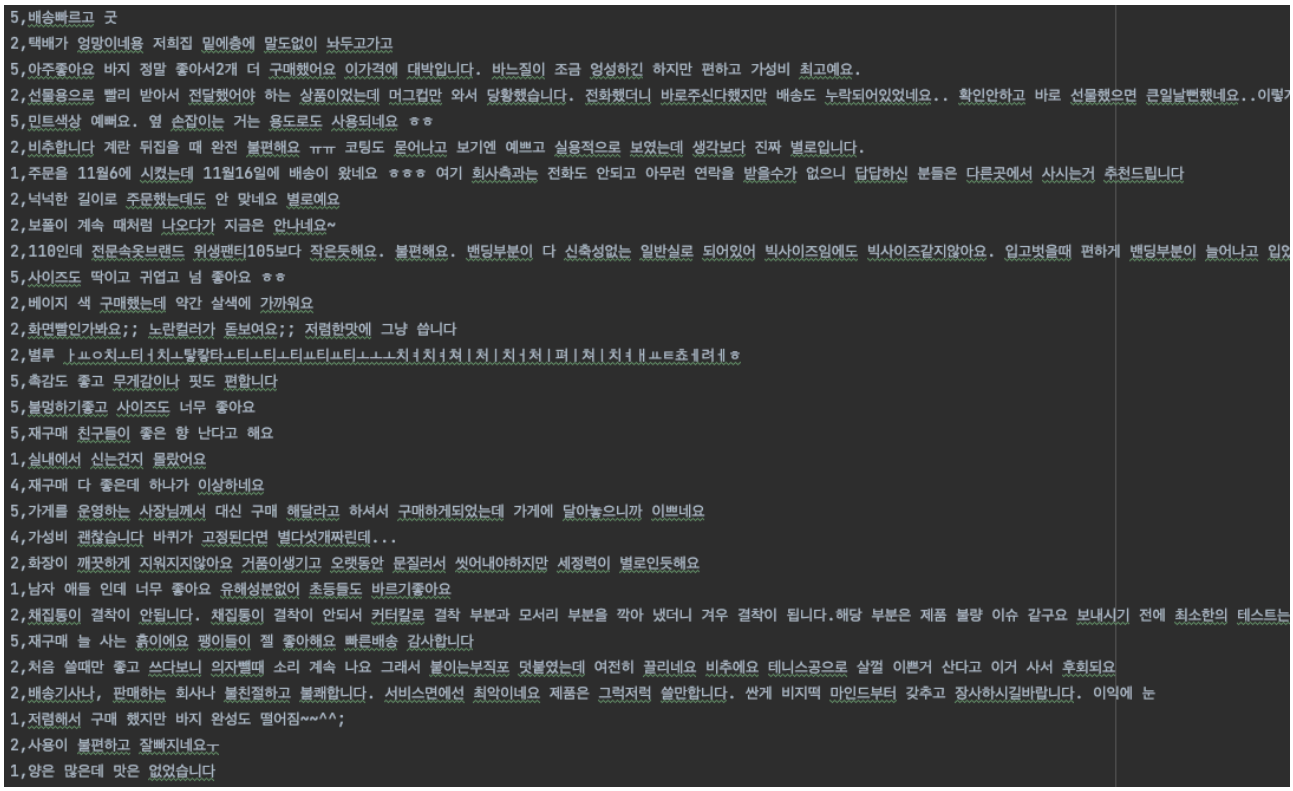


```
{
    "index":118,
    "text":"힘들다",
    "score":-2
},
{
    "index":119,
    "text":"좁다",
    "score":-2
},
{
    "index":120,
    "text":"답답하다",
    "score":-2
},
{
    "index":121,
    "text":"너무하다",
    "score":-2
},
{
    "index":122,
    "text":"작다",
    "score":-2
},
{
    "index":123,
    "text":"없다",
    "score":-2
},
{
    "index":124,
    "text":"짧다",
    "score":-2
},
```

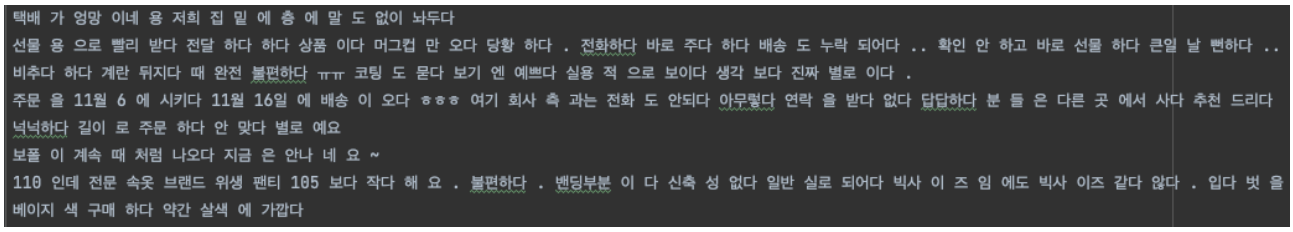

왼쪽 이미지가 사전에서 긍정 단어로 선정된 단어들의 일부이며, 오른쪽 이미지가 사전에서 부정 단어로 선정된 단어의 일부이다.

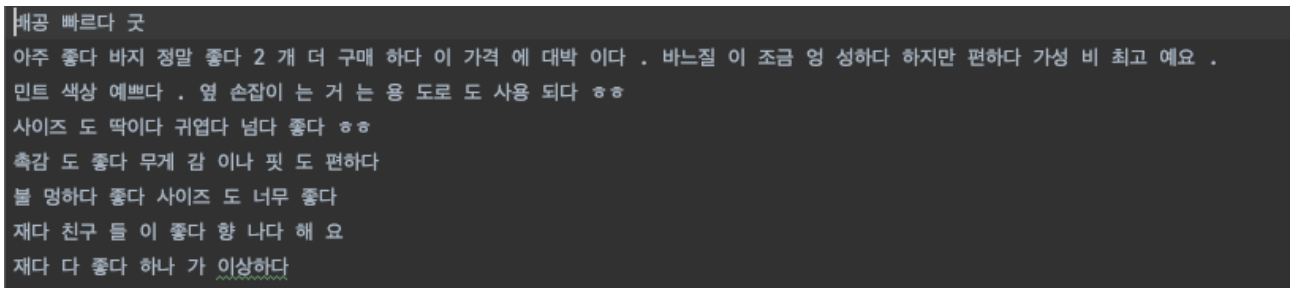
이후 평점 분포에 알맞게 감정 점수를 평점으로 변환한다. 본 연구를 진행함에 있어 수집한 리뷰 데이터를 분석 하였을 때, 리뷰의 대부분이 5점과 4점으로 분포하여있고, 3점 이하의 점수를 갖는 리뷰의 비중이 상당히 작다는 것을 확인 할 수 있었다. 5점과 4점의 분포를 계산하여 평점 분포를 평점 예측에 활용하였다. 훈련 데이터를 통해 만들어진 평점 비율을 앞서 구한 감성 점수들에 적용해 평점으로 변환한다.

3. 연구결과



selenium을 바탕으로 크롤링한 리뷰 데이터이다. 이때 리뷰별점과 같이하여 크롤링을 진행한다.





위 2개는 전처리 및 전체 리뷰데이터를 토큰화를 진행 한 후 각각 부정,긍정 리뷰데이터의 일부이다.

```
{
    "index":0,
    "text":"좋다",
    "score":2
},
{
    "index":1,
    "text":"편하다",
    "score":2
},
{
    "index":2,
    "text":"고급스럽다",
    "score":2
},
{
    "index":3,
    "text":"만족하다",
    "score":2
},
```

```
{
    "index":131,
    "text":"비싸다",
    "score":-2
},
{
    "index":132,
    "text":"기대하다",
    "score":-2
},
{
    "index":133,
    "text":"심하다",
    "score":-2
},
{
    "index":134,
    "text":"안좋다",
    "score":-2
},
{
    "index":135,
    "text":"불편하다",
    "score":-2
```


위 2개는 감성사전의 긍정부분과 부정 부분의 일부분이다.

```
사전에 단어가 없는 경우 결과가 None으로 나타납니다!!!
종료하시려면 #을 입력해주세요!!!
-2:매우 부정, -1:부정, 0:중립 , 1:긍정, 2:매우 긍정


리뷰를 입력하세요 : 가볍고 좋아요. 음식에 따우치도 만족으면 얇마나 좋았을까 아쉽네요. 따로 구매 했어요. 잘 사용할께요.
['가볍다/Adjective', '좋다/Adjective', './Punctuation', '음식/Noun', '에/Josa', '따우치/Noun', '도/Josa', '있다/Adjective', '얇마나/Noun', '좋다/Adjective', '아쉽다/Adjective', './Punctuation', '따로/Adverb', '구매/Noun', '하다/Verb', '
./Punctuation', '지다/Verb', '사용/Noun', '하다/Verb', './Punctuation']
['가볍다', '좋다', '있다', '좋다', '아쉽다']
어근 : 가볍다
극성 : 2
어근 : 좋다
극성 : 2
어근 : 있다
극성 : 0
어근 : 좋다
극성 : 2
어근 : 아쉽다
극성 : -2
total_score :  4
평균:  0.8
예측한 별점은 4점입니다.
```

이후 리뷰를 입력하면 해당 리뷰를 토큰화를 진행한다음 해당 어휘도에서 감성사전에 존재하는 극성 어휘들을 판별하여 계산한 후 , 별점을 예측하게 한다.

4. 결론

 본 연구에서는 SO-PMI를 활용해 영역별 감성 사전을 구축하고, 평점을 예측하는 연구를 수행하였다. 연구 결과 본 연구에서 구축한 영역별 감성 사전으로 평점을 예측하는 경우, 예측한 평점과 실제 평점이 유사하게 결과가 나왔다.

그러나, 본 연구를 진행하며 느낀 한계점은 다음과 같다.

쿠팡에 존재하는 리뷰의 90% 이상이 4점 또는 5점 리뷰이다보니, 부정적인 리뷰의 절대 적인 수가 부족했다. 따라서 리뷰 비율의 분포를 바탕으로 평점을 예측하다 보니 불안정한 점수측정이 가끔씩 결과로 이어졌다. 또한 직접 데이터를 크롤링을 통해 가져오다보니 데이터의 양의 수가 절대적으로 적어 감성사전을 구축하는데 있어 감성사전에 들어가는 단어의 수가 부족하였다. 그리고 한국어의 특성상 예를들자

면 ‘하지 않다’와 ‘못하다’를 토큰화를 진행하면 ‘하다’와 ‘않다’, ‘못’과 ‘하다’ 이런식으로 구분되는데 이때 SO-PMI를 측정하는 부분에 있어서 ‘하다’라는 단어가 부정리뷰데이터에서 높은 점수가 측정되어 이후 별점을 예측할때 문장속 ‘하다’라는 단어가 있으면 이를 부정 단어로 취급을 하는 경우가 발생하였다. 만약 한국어가 아닌 영어 데이터로 감성사전을 구축하였다면 이러한 문제가 발생하지 않았을 것 같다. 이를 개선하기 위해서는 ‘못’이나 ‘않다’를 갖고 있는 단어를 아예 반대말로 바꿔 치환한다면 해당 문제점을 바로 잡을 수 있을 것 같다. 예를들어 ‘행복하다’를 ‘불행하다’ 라는 단어로 치환을 하는 방식이다.

또한 쿠팡 리뷰데이터 등 이커머스 데이터가 아닌 좀 더 객관적인 지표를 얻을 수 있는 음식점과 호텔 등 소비자에게 직접적으로 노출되지 않는 업종의 리뷰에서는 부정적인 리뷰 데이터를 보다 쉽게 얻을 수 있을거라 생각한다.