1. 비지도학습 기반의 오피니언 마이닝이란

○ 감성 분석이란 문서의 주관적인 감성, 의견, 감정, 기분 등을 파악하기 위한 방법임.

- 문서 내 텍스트가 나타내는 여러가지 주관적인 단어와 문맥을 기반으로 감성 수치를

계산하는 방법을 이용함.

- 긍정 감성 지수와 부정 감성 지수로 구성되며 이를 합산해 긍정, 부정 감성을 결정함.

- 머신러닝 관점에서 지도 학습과 비지도 학습 방식으로 나눌 수 있음.

- 비지도 감성 분석은 Lexicon(감성 어휘 사전)을 기반으로 함.

• 감성 사전은 감성 지수를 가지고 있는데, 이 감성 지수는 단어의 위치, 문맥 등을

참고하여 결정됨.

• NLTK의 감성 사전에는 SentiWordNet, VADER, Pattern이 있음.

2. 한국어 감성사전

○ 한국어 감성사전에는 KNU 감성사전과 KOSAC 감성사전이 있음.

1) KNU 감성사전

- <https://github.com/park1200656/KnuSentiLex>

- 표준국어대사전을 구성하는 형용사, 부사, 동사, 명사의 모든 뜻풀이에 대한 긍정, 중립, 부정으로 분류하기 위해 Bi-LSTM 딥 러닝 모델을 사용함.

- Bi-LSTM 모델은 각 뜻풀이의 확률 값을 계산하여 최종적으로 300,000개에 달하는

뜻풀이를 그룹에서 top-2500 긍정어와 부정어를 추출하여 최소 3명의 평가자들이

각 단어의 긍정, 중립, 부정을 판별하는 방식으로 감성사전을 구축함.

2) KOSAC 감성사전

- [http://word.snu.ac.kr/](http://word.snu.ac.kr/kosac/data/lexicon.zip)**[kosac](http://word.snu.ac.kr/kosac/data/lexicon.zip)**[/data/lexicon.zip](http://word.snu.ac.kr/kosac/data/lexicon.zip)

- 1-gram, 2-gram, 3-gram의 어휘로 구성되어 있으며 형태소 단위로 감성어휘를 제공함

- 기존 감성사전은 일반용이라는 약점을 가지며, 영어를 단순 번역하였다는 단점이 있음.

3. 한국어를 바탕으로 감성사전을 구축하는 방법 및 구축 프로세스

○ 감성어 사전 구축 블로그

- <https://projectlog-eraser.tistory.com/19>

- (Step1) 형태소 분석

• 파이썬 한국어 처리 패키지 KoNLPy의 형태소 분석기 Okt는 의미를 직관적으로

파악할 수 있고, 결과 데이터의 크기가 작지만 분석 시간이 오래 걸리므로 Mecab으로

형태소를 분석함.

- (Step2) 활용할 품사 선택

• Mecab의 품사 태그 분류 <https://docs.google.com/spreadsheets/d/1OGAjUvalBuX-oZvZ_-9tEfYD2gQe7hTGsgUpiiBSXI8/edit#gid=0>

* Mecab 기준 VCP, VCN, IC, MAG, VA, VV, XR로 태깅되는 형태소들, 이 형태소가 포함되어 있는 태그들(VX+EC+VCP 등)을 감성어 사전에 포함함.

- (Step3) 사전 구축 근거

• KNU 한국어 감성사전 활용

- 사전에 존재하는 단어들을 기반으로 단어들을 추가하며 감성사전을 구축함.

1. KNU 감성사전에 있는 단어를 Mecab으로 형태소 단위로 분석한 뒤, 연구에 활용할 품사가 들어있는 형태소들만 선택하여 긍정어 사전에 있는 형태소는 1, 부정어 사전에 있는 형태소는 -1을 태깅함.
2. 태깅 결과를 확인하며 똑 같은 형태소가 긍정, 부정 두 가지 모두 태깅되는 경우 태깅을 제거함.
3. 태깅 결과를 기존의 형태소 데이터셋에 합치고, 가나다 순으로 정렬함.

* 판단 근거의 우선 순위는 표준국어대사전과 네이버 국어사전의 뜻풀이와 예문으로 하고, 태깅 작업을 혼자의 판단으로 하지 않고 여럿이 체크하는 방식으로 함.



그림 1 감성사전 구축 결과

○ 이재웅, 윤현노, & 문남미. (2017). 소비자 분석을 위한 감성사전 모델링. *한국정보처리학회 학술대회논문집*, *24*(1), 850-853.

■ 연구 목적

- 기존 텍스트 마이닝 기법은 특정 분야에서 발생하는 감성단어에 대해서는 높은 감성

분석 성능을 기대할 수 없음.

• 기존에는 형태소 분석 후 정규화된 단어를 이용해 감성사전을 구축함.

- 따라서, 본 연구는 특정 분야를 미리 정했을 때를 가정하여 발생하는 감성단어를 카테

고리화 하여 기존 범용적인 감성사전에 비해 특정 요소에서 발생할 수 있는 긍부정

반응을 추출 가능하게 하는 감성사전을 제안함.

■ 연구 방법

- 특정 상품에 대한 의견 및 평가 (리뷰) 수집하여 감성사전을 구축함.

- (Step1) 형태소 분석

• 한국어 정보처리를 위한 파이썬의 오픈소스 라이브러리 ‘KoNLPy’를 사용함.

• ‘Konoran’ 클래스를 사용하여 형태소를 분석함.

- ‘Konoran’ 클래스는 총 42가지 tag로 데이터를 분류하여 적절히 세분화된 단어들을

제공하며, 빠른 처리 속도를 보이기 때문에 본 연구에 적합함.

- (Step2) 분석 단어 가공

• 형태소 분석기에서 수집한 태그의 빈도수를 추출하기 위한 전처리 작업을 함.

- 같은 의미를 표현하지만 표현 방식이 상이한 여러 가지 단어에 대한 단일화

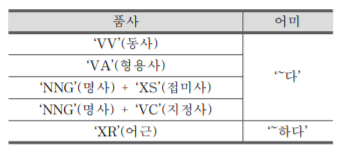


표1 감성 사전 형태소 분석 예시

• 부정어 처리를 위해 단어 앞뒤에 ‘~지 않다’ 처럼 연결어미(EC) + 보조용언(VX)인

경우, ‘~이 아니다’ 처럼 부정지정사(VCN)이 활용되는 경우, ‘안 ~하다’, ‘별로 ~하다’

처럼 단어 앞에 일반 부사(MAG)가 나오는 경우 등 여러 가지 경우를 고려함.

- (Step3) 카테고리화 및 감성 처리

• 분석한 형태소들의 빈도수를 기반으로 가중치를 두러 선별된 단어들을 이용해 긍정

과 부정으로 나누어 카테고리화 함.

• 긍정적인 단어는 ‘+’ 가중치, 부정적인 단어는 ‘-‘ 가중치를 두어, 단어의 극성을 판단

- 각 카테고리의 극성을 전부 0으로 초기화 한 다음, 분류된 단어들이 긍정일 경우

카테고리 극성에 +1, 부정일 경우 -1을 함.

- 리뷰 전체의 극성은 모든 카테고리의 극성을 더한 값을 이용해 양수일 경우 긍정,

음수일 경우 부정, 0일 경우 중립으로 구분함.

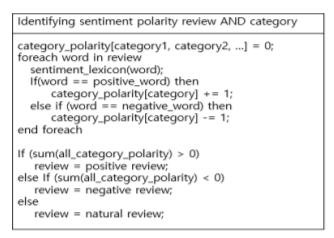


그림 2 극성 판별 알고리즘

* 연구 결과

- 추출한 데이터에서 카테고리를 만들기 위해 각 상품에 자주 등장하는 단어의 빈도수 측정한 결과



표 2 단어 빈도수 측정 결과

- 선별한 카테고리에 단어의 극성을 판단하여 사전을 구축함.

• 빈도수를 바탕으로 카테고리를 만들고, 세분화할 필요가 있는 카테고리는 세부 항

목을 만들어 해당하는 단어의 극성을 판별하여 분류함.

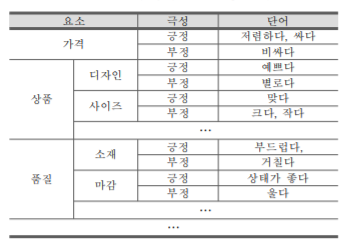
 

표 3 감성사전 모델링 결과 표 4 구축된 감성사전 바탕 결과

* 결론 및 시사점

- 범용감성사전을 사용하는 것이 아니라 각기 다른 분야마다 맞는 감성사전을 구축하는 방법을 제시한다는 점에서 의미가 있음.