Sentiment Analysis

1. 정의
감성 분석(Sentiment Analysis)이란 텍스트에 들어있는 의견이나 감성, 평가, 태도 등의 주관적인 정보를 컴퓨터를 통해 분석하는 과정

2. 활용
-기업 내부적으로는 고객 피드백, 콜센터 메시지 등과 같은 데이터를 분석하며 외부적으로는 기업과 관련된 뉴스나 SNS 홍보물 등에 달린 댓글의 긍/부정을 판단하는 곳에 사용
-개인 단위에서는 영화를 보기 전에 리뷰를 참고하는 것과 같이 특정 제품이나 서비스를 이용할 지를 결정하는 데에 사용
-우리는 단지 머신러닝 방법론을 사용하지 않았을 뿐 은연중에 감성 분석을 하고 있음
광고의 효율을 높이거나 특정 약품이 사람들에게 실제로 효과가 있는 지를 알아보는 데에도 사용

3. 절차
1) 문서(문장)의 어떤 부분에 의견이 담겨있는 지를 정의(Opinion definition)
2) 첫 번째 단계를 통해 모아진 의견을 요약(Opinion summerization)

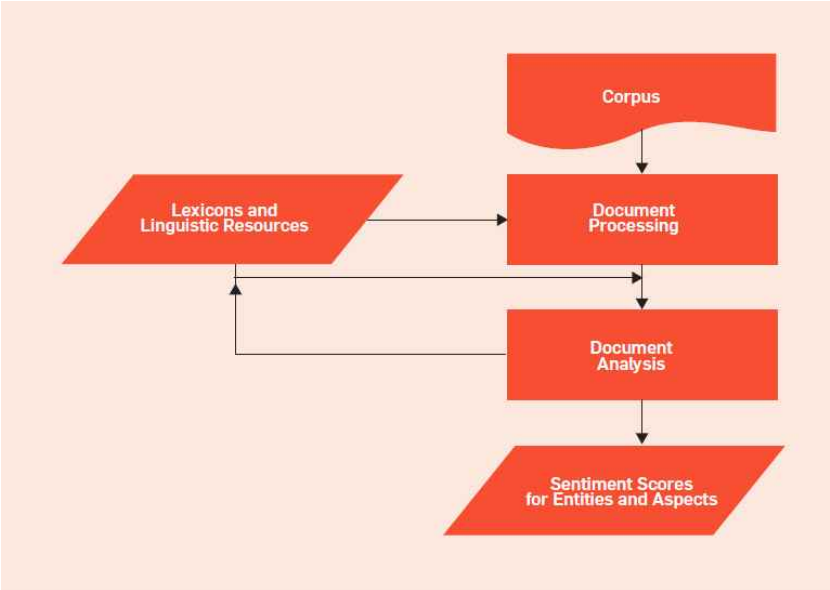
1) Opinion Definition
- 가장 먼저 해야하는 일은 분석에 필요한 4가지 요소를 찾는 것
- 4가지 요소란 분석 대상이 되는 개체(Entity)나 개체의 특성(Aspect/Feature), 개체에 대한 의견에 담겨있는 감성(Sentiment), 의견을 표현하는 주체(Opinion holders), 그리고 발화 시점(Time)
- 우리가 찾고자 하는 의견은 일반적으로 일반 의견(Regular opinions)과 비교 의견(Comparative opinions)의 2가지로 분류 가능
- 전자는 단일한 개체(혹은 개체의 특성)에 대해 의견을 표현한 것으로, “이 제품은 좋다”와 같은 직접적인 의견(Direct opinions)과 “이 제품은 잘 작동한다”와 같은 간접적인 의견(Indirect opinions)두 가지로 분류 가능
- 후자는 2개 이상의 대상을 언급하며 서로에 대한 의견을 나타내는 것

4. 구성 요소
- 비교 의견은 2개 이상의 대상이 언급되므로 분석하기가 상당히 어려움. 따라서 일반 의견에 한정
- 일반 의견은 대상(Target)이 하나이므로 위에서 알아본 4가지 요소(대상, 감성, 주체, 시점)를 다음과 같이 나타낼 수 있음 : $(g_i, so_{ijk}, h_j, t_k)$
- 위 표현에서 gi 는 분석 대상, hi 는 발화 주체, tk 는 발화 시점을 의미. soijk 는 i 에 대해서 k 시점에 j 가 표현한 감성임. 분석 대상이 하나인 경우라도 개체에 관한 서술인지, 개체의 특성에 관하여 서술한 것인지를 구분하기 위해 g 를 e (entity) 와 a (aspect/feature)로 구분하여 나타내기도 함. 이런 경우 문장에서 의견을 정의할 때 아래의 5가지 요소를 분석

대상으로 놓게 됨. $(e_i, a_{jl}, so_{ijkl}, h_j, t_k)$

위 다섯가지 요소를 활용하면 문장에서 의견을 나타내는 부분을 구조화 가능 문장 뿐만 아니라 문서와 같이 긴 텍스트에 내포된 감정을 분류할 때에도 같은 요소를 추출하여 나타냄.

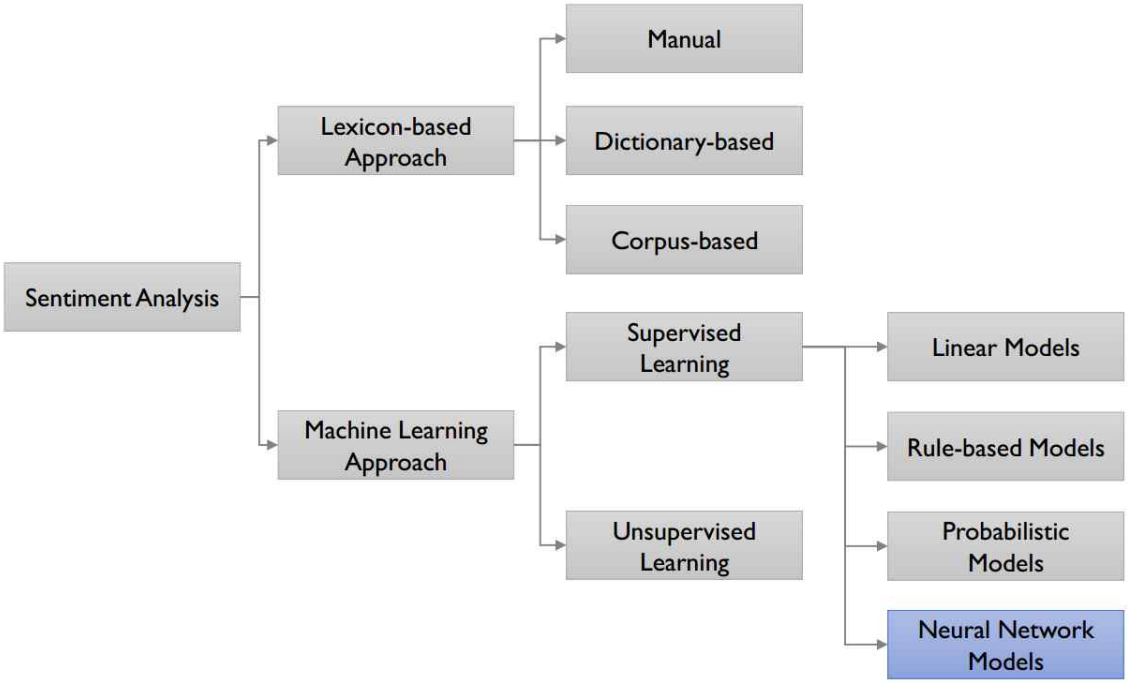
5. Flow chart



- 실제로 감성 분석을 수행할 때에는 요소의 개수를 줄이기 위해 “하나의 문서 내에서는 한 사람이 한 개체에 대해서만 즉시 평가한다.”고 가정
- 분석 대상인 리뷰나 고객 피드백 등은 발화자 자신이 사용한 제품에 대해서 평가하는 것이므로 위 가정을 제법 잘 만족함
- 이 가정을 사용하면 5개의 요소에서 대상(e,a), 발화자(h), 발화시점(t) 등을 무시 가능
- 그렇기 때문에 문서에서 드러나는 감성 표현만 잘 추출해도 분석할 수 있게 되는 것

6. Method

- 어휘 기반(Lexicon-based)의 감성 분석과 머신러닝 기반(ML-based)의 감성 분석으로 분류
- 각각의 방법은 더욱 세부적인 방법들로 나눌 수 있으며 최근에는 BERT와 같은 트랜스포머 계열 신경망 모델이 많이 사용되고 있습니다.



7. 어휘 기반(Lexicon-based)의 감성 분석

- 어휘 기반의 감성 분석은 총 세 가지 세부 방법을 가지고 있음.
- 첫 번째는 모든 단어에 대한 감성 사전을 수동(Manual)으로 구축하는 방법
- 감성 사전이란 문서의 각 단어가 가지는 긍/부정의 정도를 -1(부정) 부터 1(긍정) 사이의 점수로 레이블링한 것. 감성이 들어갈 수 있는 품사인 명사, 형용사, 동사 키워드를 추출한 뒤에 이들에 대한 긍/부정 레이블링을 진행
- 이렇게 구축된 감성 사전을 사용하면 개체에 따른 극성이나 긍/부정을 시각화할 수 있고, 연관어 분석을 통해서 어떤 단어가 해당하는 단어들과 같이 사용되었는지도 알 수 있음.
- 수동으로 감성 사전을 구축하는 방법은 한 번 수행하고나면 적용하기 쉽다는 장점이 있음.
- 하지만 도메인에 따라 사용하는 어휘가 달라지고 긍/부정 점수도 달라질 수 있기 때문에 모든 도메인에 적용할 수 있는 감성 사전을 구축하는 것이 매우 어렵다는 단점도 있음.

8. Dictionary-based

- 첫 번째 방법과의 차이는 매 분석마다 새로운 사전을 쓰는 것이 아니라 기존에 잘 구축되어 있는 외부 사전을 차용한다는 것
- 이 방법은 외부 사전을 가져오기 때문에 수동으로 구축하지 않아도 되며 분석에 적용하기 쉽다는 장점도 있음.
- 하지만 이 방법 역시 도메인 확장성이 없다는 단점을 가지고 있음.

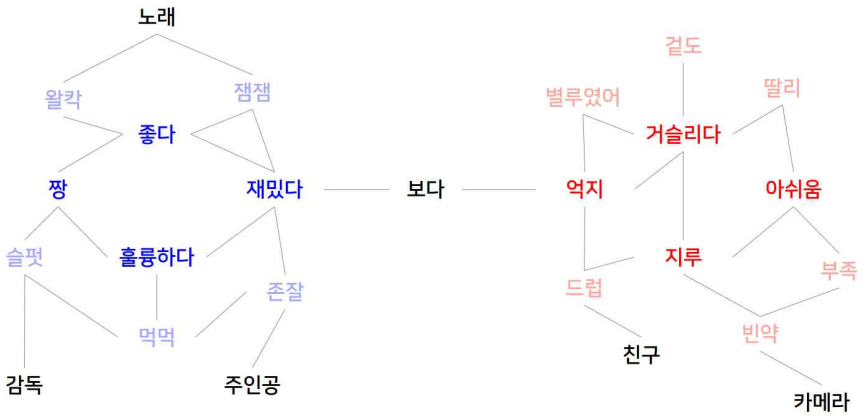
ex) 영화 리뷰 데이터에서 “졸리다” 라는 단어는 부정적인 의미, 침대 상품평 데이터라면 “졸리다” 라는 단어는 긍정적인 의미. 이렇게 도메인에 따라 달라지는 단점이 있기 때문에 만약 차용하려는 사전이 적용하려는 데이터와 다른 도메인으로부터 생성된 것이라면 재고 필요.

9. Corpus-based
마지막은 해당 말뭉치에 맞는 적절한 감성 어휘를 재구축하는 말뭉치 기반(Corpus-based)의 접근 방법입니다. 이 방법은 이전 방법들의 단점인 도메인 의존성을 극복할 수 있지만 좋은 사전 구축을 위해서 많은 데이터(거대한 말뭉치)를 필요로 한다는 조건을 가지고 있습니다. 특정 말뭉치를 분석하는 경우에는 해당 단어가 사용되었을 때 문장의 긍/부정을 t-Test를 통해서 판단합니다.

10. 머신러닝 기반(ML-based)의 감성 분석
머신러닝 기술이 발달하면서 어휘 기반의 감성 분석 방법보다는 머신러닝 모델 기반의 감성 분석이 많이 수행되고 있습니다. 특히 지도 학습(Supervised Learning) 기반의 방법이 많이 시도되고 있습니다.

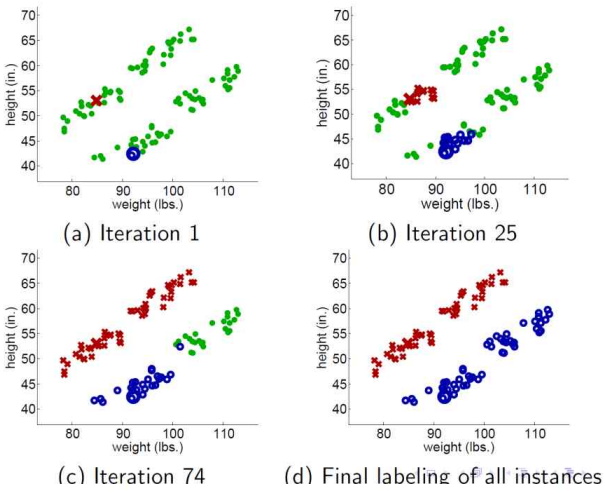
11. Linear
그 중 하나는 회귀 분석을 통해 사전을 구축하는 방식입니다. 일반적으로 리뷰 데이터에는 평점을 부여하는데, 이렇게 평점을 레이블링 된 데이터에 회귀 분석 모델을 적용하여 각 단어에 대한 감성 사전을 구축합니다. 구축한 이후에는 교차 검증을 통해 감성 사전으로서의 타당한 성능을 지니는지를 평가합니다. 검증을 통해 구축한 사전의 성능이 확보된 후에는 새로운 문서에 대한 감성 분석을 수행할 수 있게 됩니다.

12. Semi-supervised
지도 학습과 비지도 학습의 중간인 준지도 학습(Semi-supervised Learning)을 통해 감성 사전을 구축하기도 합니다. 준지도 학습 기반의 방법은 2가지 세부 방법으로 나뉘게 됩니다. 첫 번째는 감성 그래프(Sentiment graph)를 이용하는 방식입니다. 고차원 단어를 저차원으로 임베딩한 후에 거리를 기반으로 각 단어 사이의 네트워크를 구축합니다. 감성 점수가 확실한 수개의 단어를 미리 레이블링(Pre-labeled sentiment words)한 뒤에 공간상의 위치에 따라 나머지 단어의 감성 점수를 측정합니다. 아래는 감성 그래프를 적용했을 때 단어가 긍/부정으로 레이블링 되는 과정을 시각화한 것입니다.



위에서는 “좋다”, “짱”, “재밌다”, “훌륭하다” 등을 긍정으로 “거슬리다”, “억지”, “지루”, “아쉬움” 등을 부정으로, 그리고 “보다”, “감독”, “주인공”, “친구”, “카메라” 와 같이 중립으로 미리 레이블링 했습니다. 그리고 이들과의 관계로부터 나머지 단어의 감정 점수를 매기게 됩니다.

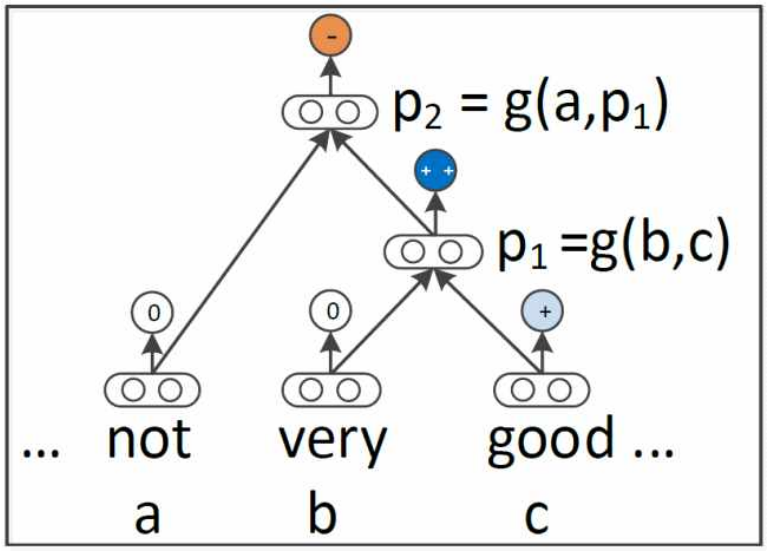
준지도 학습을 통한 두 번째 접근 방식은 자가 학습(Self-training)입니다. 이 방법 역시 감성 그래프 방법에서 했던 것과 같이 특정한 수 개의 단어는 미리 레이블링 되어 있습니다. 이 어휘로만 분류기를 학습한 뒤에 정답이 없는 어휘에 분류기를 적용하여 결과의 신뢰도가 높으면 정답으로 지정한 후에 학습기를 재학습하는 방식입니다. 아래는 자가 학습 방법을 이용한 감성 분석을 사용한 예를 시각화한 것인데 반복 학습 횟수가 늘어날수록 레이블링 되는 단어가 더 많은 것을 볼 수 있습니다.



13. Neural Network

2016년에는 RNTN(Recursive Neural Tensor Network) 방법이 제시되었습니다. 여기서의 R은 Recursive(재귀적인)를 나타내는 단어로 RNN에서 사용되는 Recurrent(순환하는)와는 다른 의미를 가지고 있습니다. RNTN은 두 가지 벤치마크 모델이 있습니다.

첫 번째는 RecursiveNN 입니다. 각 단어를 아래와 같이 구조적으로 나타낸 이후 각 단어를 구(Phrase)로 하나씩 결합하면서 감성이 어떻게 나타내는 지를 계속해서 학습합니다. 논문에서는 긍정과 부정을 25단계로구분하였으며 동일한 구에 댛서는 3명이 평가한 결과물의 평균을 사용하여 레이블링 하였습니다. 여기서 재귀적(Recursive)이라는 수식어를 사용하는 이유는 각 단어 혹은 구마다 동일한 가중치를 적용하기 때문입니다. 아래 이미지는 RecursiveNN이 진행되는 과정을 나타낸 것입니다. 이 때 첫 번째 구 p1의 긍/부정을 판별할 때와 두 번째 구 p2의 긍/부정을 판별할 때 동일한 함수 g 를 사용하는 것을 볼 수 있습니다.

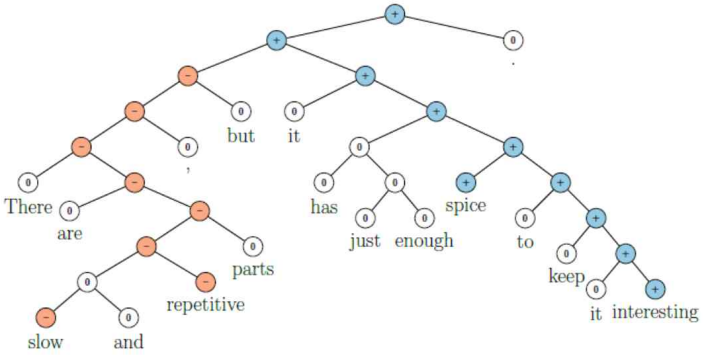


이 과정을 수식으로 나타내면 다음과 같습니다.

yword=softmax(Ws・word)p1=f(W[bc]),p2=f(W[ap1])
위 식에서 활성화 함수인 f 는 일반적으로 tanh 가 사용되며 이 때 가중치인 W 는 변하지 않습니다. 그리고 새로 만들어지는 구를 구성하는 단어 혹은 작은 구를 표현하는 벡터는 Concatenate된 후에 가중치와 내적하게 됩니다.

두 번째 벤치마크 모델은 MV-RNN(Matrix-Vector Recursive Neural Network) 입니다. 이 방법은 더 긴 문장의 문맥을 행렬에 저장하여 기존 RecursiveNN의 한계점을 해결하고자 합니다.

그리고 두 벤치마크 모델을 결합하여 텐서로 쌓아 나타낸 것이 바로 아래의 RNTN입니다. RNTN은 일반적으로 각각의 벤치마크 모델보다 성능이 더 좋습니다. 특히 아래 이미지와 같이 “but”으로 두 문장이 이어져 있거나 복잡한 부정 표현(High-level Negation)이 문장에 존재하는 경우에 기존 모델보다 훨씬 더 좋은 성능을 나타냅니다.


There
are
slow
and
repetitive
parts
,
but
it
has
just
enough
spice
to
keep
it
interesting
.