감정분석 방법에는 두가지 방법이 있다

1. 감정이 labeling 되어있는 데이터를 통해서 감정분석을 하는 방법

2. 기존에 만들어진 lexicon을 통해서 감정분석을 하는 방법

한글 감정분석에 사용되는 lexicon은 서울대에서 만든 lexicon이 존재한다.

감정분석이 Data에 의존하는지 알아보자.

감정에 대해서 라벨링 된 텍스트 데이터를 수집하는데 문제가 생긴다. 온라인상에 많은 수의텍스트 데이터가 존재하지만 Labeling된 텍스트 데이터를 찾기는 쉽지 않다. 또한 실제로 수집한 텍스트데이터에 라벨링을 하는 작업의 경우에도 많은 시간과 비용이 소모된다. 따라서 온라인상에서 쉽게 얻을 수 있는 영화 한줄평 데이터를 사용하였다. 영화 한줄평의 경우 영화에 대한 평점과 함께 댓글을 남기므로 평점데이터를 점수구간별로 나누어 긍정/부정의 라벨링을 할 수 있다.

Train data : 네이버 평점 데이터

Test data : 트위터 크롤링을 통해 텍스트 데이터 수집 후 수작업 라벨링

1. 서울대 lexicon을 사용하여서 감정분석

2. 네이버평점을 사용한 lexicon을 사용한 감정분석

테스트 데이터를 작성하는데

1, 목적

텍스트 데이터를 활용한 감정분석은 기계학습 기반의 지도학습을 통한 방법과, 감성어 사전 기반의 방법 2가지 방법이 존재한다. 그중 감성어 사전 기반의 분석의 경우 기존에 만들어진 감성어 사전을 이용하는데 한글 텍스트 분석의 경우 서울대에서 만든 Korean Sentiment Lexicon을 주로 사용하여서 분석한다.

그러나 텍스트 데이터가 domain에 dependent하다면 해당 domain의 특성을 반영해 줄 수 있는 ML기반의 지도학습이 Korean Sentiment Lexicon보다 더 좋은 감성분석 결과를 도출 해 낼 수 있을 것이라고 예상된다. 따라서 Korean Sentiment Lexicon을 사용한 감성분석 결과와, ML기반의 감성분석 결과를 비교하여서, 예측 정확도를 평가지표로 사용하여 어떠한 분석이 비교우위를 가지는지 비교해 보고자 한다

2. 수집방법

온라인상에서 감성정보가 라벨링 되어있는 텍스트 데이터를 확보하기는 쉽지 않다. 그러나 ‘네이버 영화’의 한줄평 데이터의 경우 영화에 대한 평점과 한줄평이라는 텍스트 데이터가 함께 제공되고 있다. 따라서 평점데이터를 구간별로 나누어 긍정/부정적인 감성으로 라벨링하고 이를 train데이터로 사용하여서 ML기반의 감성분석을 수행해 보도록 하겠다.

‘네이버 영화’의 텍스트 데이터의 경우 request함수를 사용하여 수집이 가능하므로 request함수를 사용하여서 영화의 제목과 장르, 사용자별 평점과 한줄평 데이터를 수집하겠다.

감성분석 모델링의 분석 결과를 test하기 위해서는 twitter의 api를 이용해 특정 영화에 대한 텍스트 데이터를 수집하고 그 중에서 분명하게 감정이 드러나는 텍스트를 추출하고, 수작업으로 라벨링 하여서 test셋으로 만들고 예측 정확도를 비교하는데 사용하겠다.

Twitter의 경우 api를 사용하여서 텍스트 데이터를 추출하는데, 일간 최대 100개, 최장 일주일간의 데이터 까지만 수집할 수 있는 한계가 존재한다. 따라서 특정 영화의 제목으로 검색하여서 추출되는 최대 700개의 트윗을 수집하겠다. 수집되는 트윗중에서 중복되는 트윗과 광고를 제거하는 전처리를 수행하고 수작업으로 라벨링을 하여서 test data로 활용하겠다.

3. 분석방법

Lexicon을 사용하여서 test데이터의 감성을 예측해보고 실제 라벨링과의 비교를 통해서 예측 정확도를 계산하고, ML기반의 분석을 통해 train데이터를 사용하여서 모델링을 진행하겠다. 또한 train데이터에서 해당 장르의 데이터만을 사용하여서 즉 domain에 더 dependent한 모델을 만들어서 분석을 진행하여 보겠다.