Emotion Classification in Movie Reviews

한글 영화 리뷰의 감정 분류

김선웅

Result

http://dovvvv.tk/

강추, 꿀잼 영화입니다!

Prediction

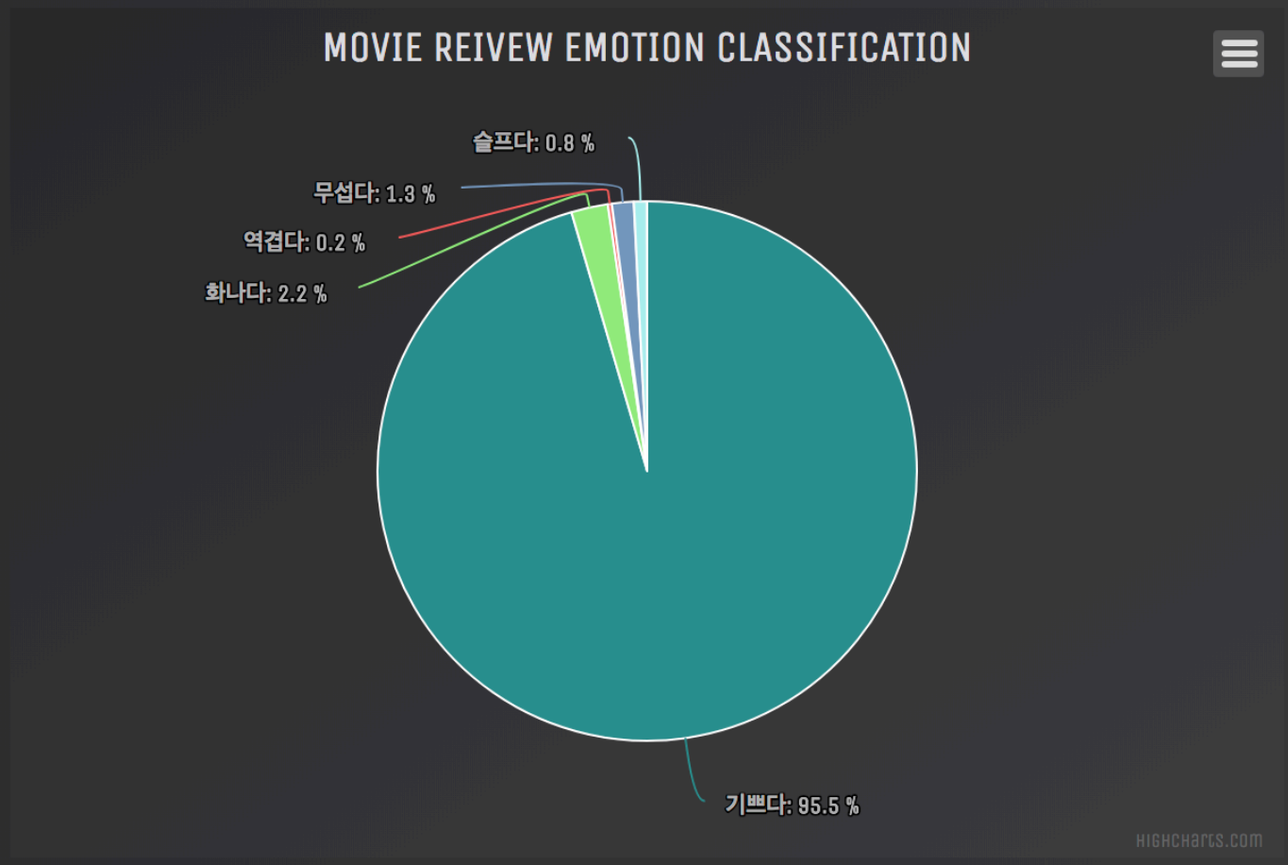
기쁘다 : 95.48%

화나다 : 2.19%

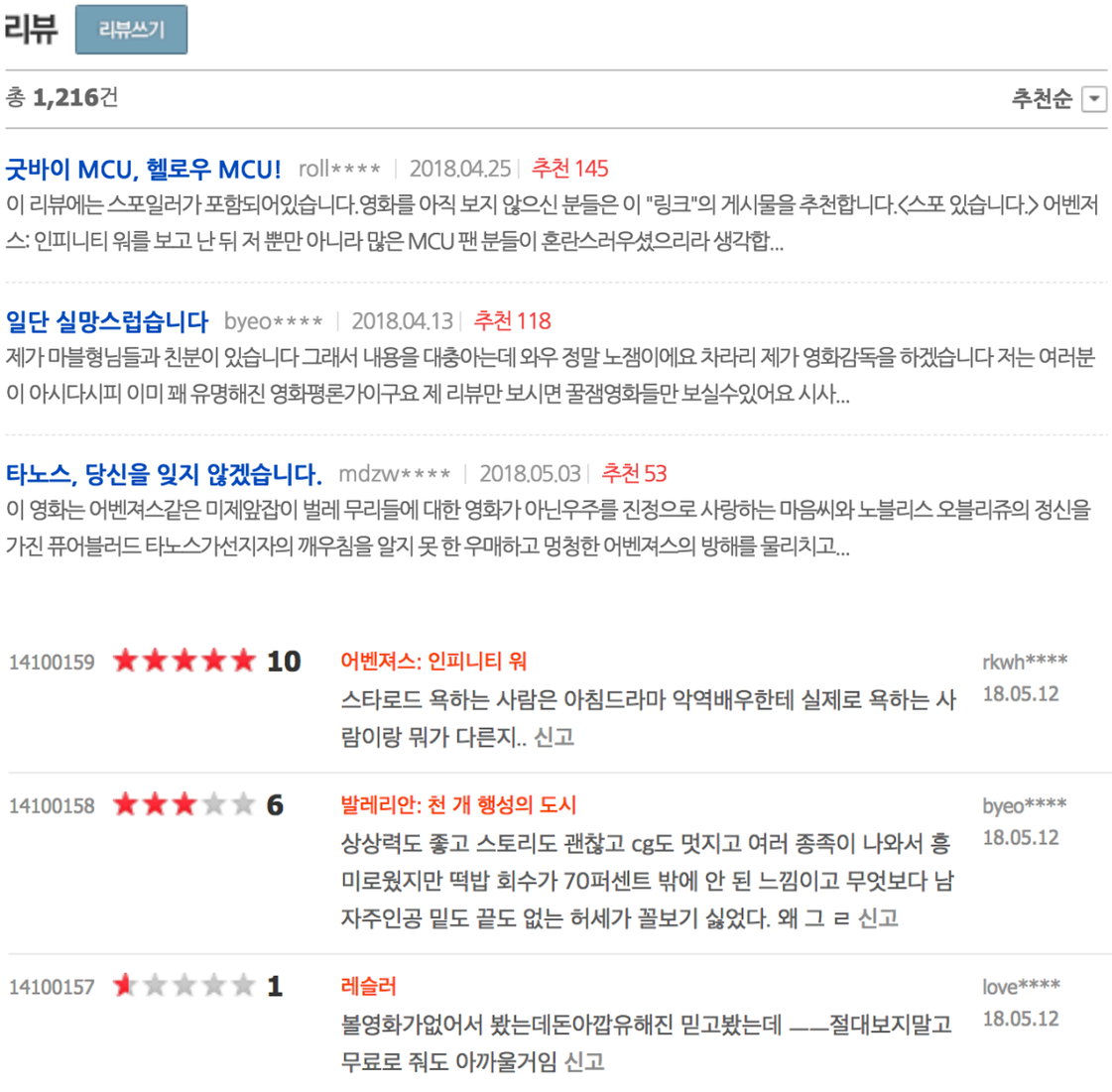
역겹다 : 0.24%

무섭다 : 1.30%

슬프다 : 0.79%



Why Emotion Classification?



리뷰 리뷰쓰기

총 1,216건 추천순

굿바이 MCU, 헬로우 MCU! roll**** | 2018.04.25 | 추천 145
이 리뷰에는 스포일러가 포함되어있습니다.영화를 아직 보지 않으신 분들은 이 "링크"의 게시물을 추천합니다.<스포 있습니다.> 어벤저스: 인피니티 워를 보고 난 뒤 저 뿐만 아니라 많은 MCU 팬 분들이 혼란스러우셨으리라 생각합...

일단 실망스럽습니다 byeo**** | 2018.04.13 | 추천 118
제가 마블형님들과 친분이 있습니다 그래서 내용을 대충아는데 와우 정말 노잼이에요 차라리 제가 영화감독을 하겠습니다 저는 여러분이 아시다시피 이미 꽤 유명해진 영화평론가이구요 제 리뷰만 보시면 꿀잼영화들만 보실수있어요 시사...

타노스, 당신을 잊지 않겠습니다. mdzw**** | 2018.05.03 | 추천 53
이 영화는 어벤져스같은 미제앞잡이 벌레 무리들에 대한 영화가 아닌우주를 진정으로 사랑하는 마음씨와 노블리스 오블리쥬의 정신을 가진 퓨어블러드 타노스가선지자의 깨우침을 알지 못 한 우매하고 멍청한 어벤져스의 방해를 물리치고...

14100159	★★★★★ 10	어벤져스: 인피니티 워	스타로드 욕하는 사람은 아침드라마 악역배우한테 실제로 욕하는 사람이랑 뭐가 다른지.. 신고	rkwh**** 18.05.12
14100158	★★★☆☆ 6	발레리안: 천 개 행성의 도시	상상력도 좋고 스토리도 괜찮고 cg도 멋지고 여러 종족이 나와서 흥미로웠지만 떡밥 회수가 70퍼센트 밖에 안 된 느낌이고 무엇보다 남자주인공 밑도 끝도 없는 허세가 꼴보기 싫었다. 왜 그 ㄹ 신고	byeo**** 18.05.12
14100157	★☆☆☆☆ 1	레슬러	볼영화가없어서 봤는데돈아깝유해진 믿고봤는데 ——절대보지말고 무료로 줘도 아까울거임 신고	love**** 18.05.12

Problems

수많은 리뷰... But,
같은 영화에도 천차만별인 감상평
전체적인 리뷰의 내용을 한눈에 파악하기는 쉽지 않음

개인적으로 평점보다는 리뷰 자체를 더 보는 편

Approach

리뷰에서 좀 더 개별적 정보를 얻어보자
→ 먼저 각 리뷰의 감정을 파악해보자

네이버 영화 https://movie.naver.com/movie/point/af/list.nhn

Problems

학습에 사용할(레이블이 태그된) 데이터 부재
직접 레이블 태깅

Corpus 구축 (유의어 검색): Word2Vec
레이블 태깅 & 노이즈 제거 (직접)

Limitation: 1) 수집한 리뷰로만 말뭉치 구축
2) 토크나이징/POS tagging 시 신조어에 취약

총 809개의 Corpus 구축

기쁘다	화나다	역겹다	슬프다	무섭다
174	196	179	154	106

기쁘다/Adjective, 0

감동/Noun, 0
감사/Noun, 0
고맙다/Adjective, 0
괜찮다/Adjective, 0
굿/Noun, 0
귀엽다/Adjective, 0
귀요미/Noun, 0
기쁘다/Adjective, 0
기쁨/Noun, 0
꿀잼/Noun, 0
달콤/Noun, 0

데이터 셋 구축

데이터 수집

네이버 영화 리뷰
252,847건

nsmc + crawling

MySQL

Corpus 구축

Tokenizing/
POS tagging

KoNLPy (Twitter)

레이블

기쁘다/화나다
역겹다/슬프다/무섭다

Most similar words

Word
Embedding

Word2Vec

Corpus

레이블 태깅

태그된 데이터
165,810건

기쁘다 73,614
화나다 40,501
역겹다 20,390
슬프다 20,701
무섭다 10,604

* 공포영화의 수가 적어
'무섭다' 데이터 수가 적음

Naver Sentiment Movie Corpus https://github.com/e9t/nsmc

데이터 전처리

KoNLPy의 twitter 클래스를 이용해 tokenizing/pos tagging

```
def tokenize(doc):
    return ['/'.join(t) for t in twitter.pos(doc, norm=True, stem=True)]


tokenize('강추, 꿀잼 영화입니다!')
>>>
['강추/Noun',
 ',/Punctuation',
 '꿀잼/Noun',
 '영화/Noun',
 '이다/Adjective',
 '!/Punctuation']
```

모델링

Tf-Idf / Multinomial Naive Bayes
: 외부 데이터에 가장 안정적인 성능을 보임

```
clf = Pipeline([
    ('vect', TfidfVectorizer(min_df=10, ngram_range=(1, 3))),
    ('clf', MultinomialNB(alpha=0.001)),
])

model = clf.fit(X_train, y_train)
model
```

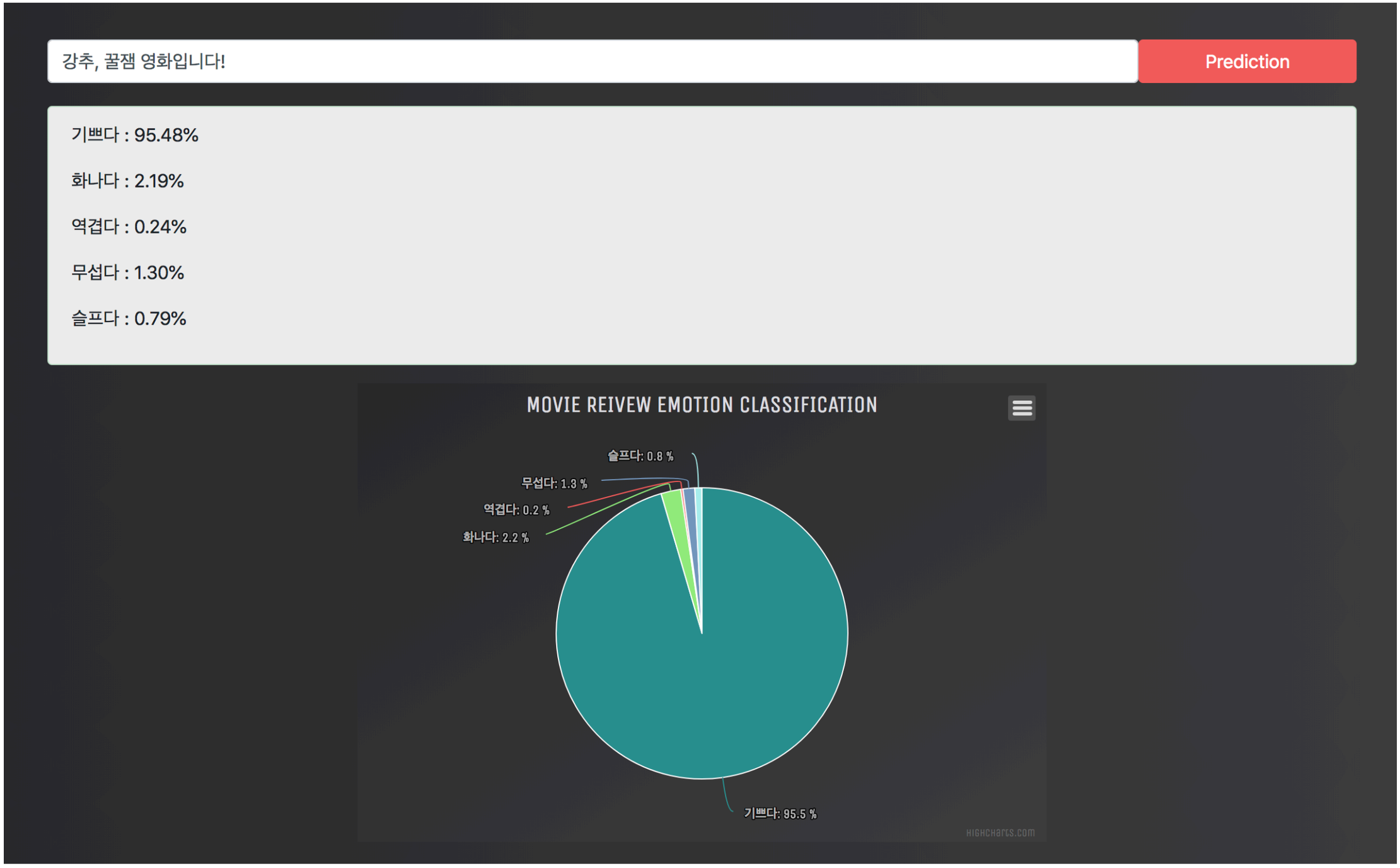
Performance

Accuracy	Recall	F1
0.77	0.77	0.78

Limitation: 5가지 감정으로 분류하기 어려운 리뷰는 학습
데이터 분포 비율대로 예측 (판단을 하지 못함)

웹 어플리케이션

Flask / Bootstrap / AWS
리뷰를 입력하면 해당 리뷰의 감정 판단
http://dovvvv.tk/



Limitations / Lessons

- **전처리**
  - OOV 문제 → 신조어 등에 강한 tokenizer 사용 (e.g. soynlp)
  - 오타/띄어쓰기 대처 전무 → 전처리 과정에서 처리 필요

- **모델링**
  - 5가지로 분류하기엔 미묘한 감정이 많은 한국어
    → 더 많은 레이블 이용 필요 (논문 등에서는 7~11개 이용)
  - 학습 데이터 구축 과정에서 5가지 감정으로 필터링
    → 실제 데이터와 맞지 않는 부분 발생

Thank you

감사합니다