



네이버 영화 리뷰 다중 감정 분류기 MR2SentiNet

페르소나 시스템

2019.11.26

진명훈 김호현

목차 index



Ch01 문제 정의

- 프로젝트 / 개발 개요
- 요구사항 분석
- 감정 범주 정의

Ch02 관련 연구

- 극성 분류
- Topic Modeling
- Language Model

Ch03 활용 데이터

- NSMC
- Crawled Data
- Senti. Word Dictionary
- Sejong-corpus
- Simple EDA



Ch04 데이터셋 구축

- 띄어쓰기, 오타자 전처리
- 데이터 라벨링 – 감정 사전 매핑
- 데이터 라벨링 – JST
- Tokenizing
- BERT Tokenizing

Ch05 Modeling

- Process
- 분류 모델
- KorBERT
- Ensemble
- Optimizer

Ch06 결론 및 향후 방향성

- 결론
- 향후 방향성

Appendix.

Reference.

Ch01 문제 정의

- 프로젝트 / 개발 개요
- 요구사항 분석
- 감정 범주 정의

01. 문제 정의

프로젝트 / 개발 개요

프로젝트 명

- 희로애락 다중 감정 분류 시스템 구축

프로젝트 기간







- '19.09.10 ~ '19.11.22

과업배경 및 목표

- IITP, KSA 주관 국비 사업 프로젝트
- 산업 특화형 인공지능 인재 배양을 위한 현장형 프로그램
- 협력 기업 연계 프로젝트 실습을 통해 실제 적용 역량 배양

수행조직 및 일정

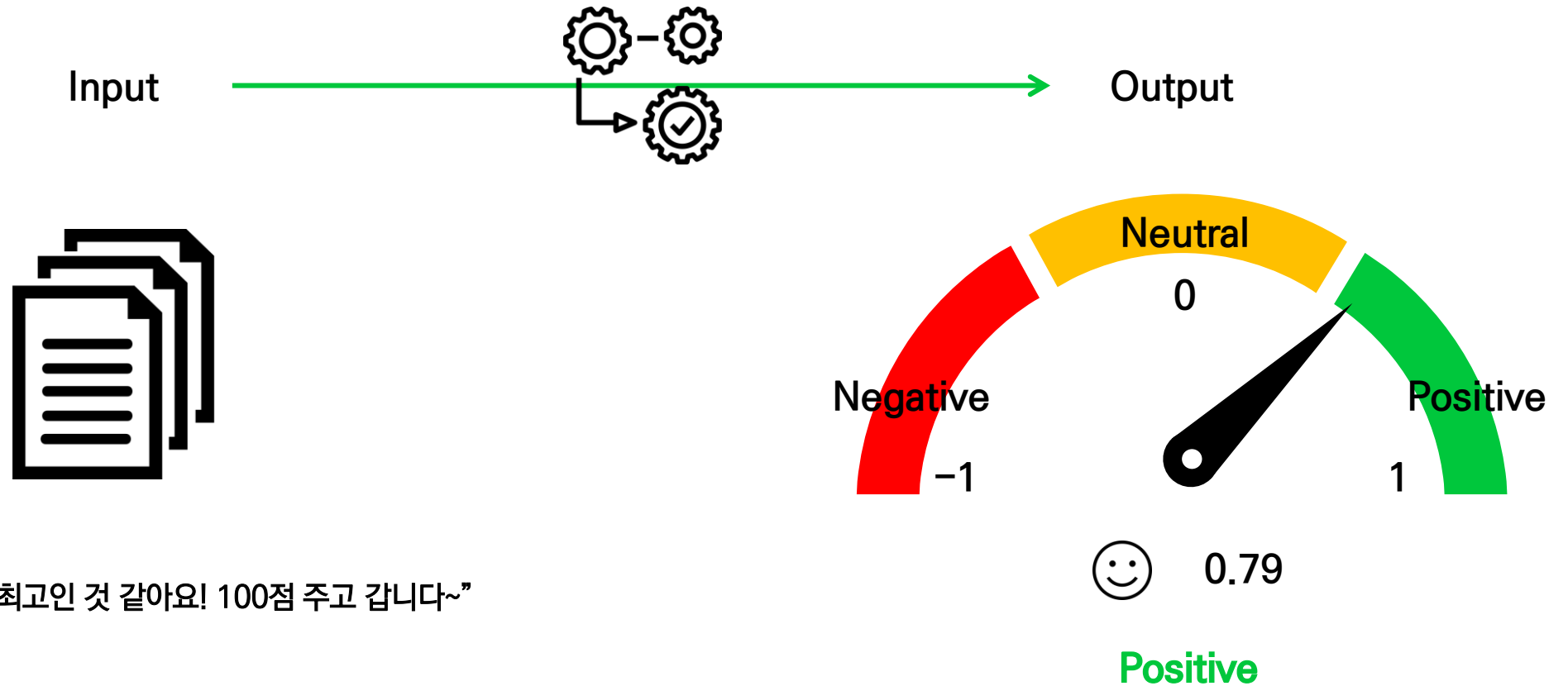
- 팀장: 진명훈
- 팀원: 김호현

	2019년		
	9월	10월	11월
관련 연구 탐색			
DB 및 감정 사전 구축			
데이터 라벨링(JST, 사전)			
분류 모델 구축			
성능 향상 및 보고서 작성			
모듈화 및 결과 발표			

01. 문제 정의

프로젝트 / 개발 개요

〈 기존의 감정 분류 〉



01. 문제 정의

프로젝트 / 개발 개요

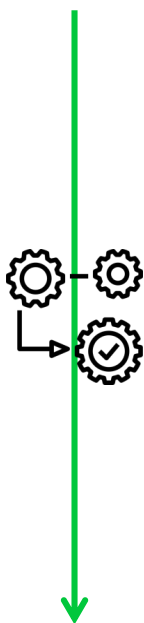
〈 히로애락 다중 감정 분류 시스템 구축 〉



요구 사항

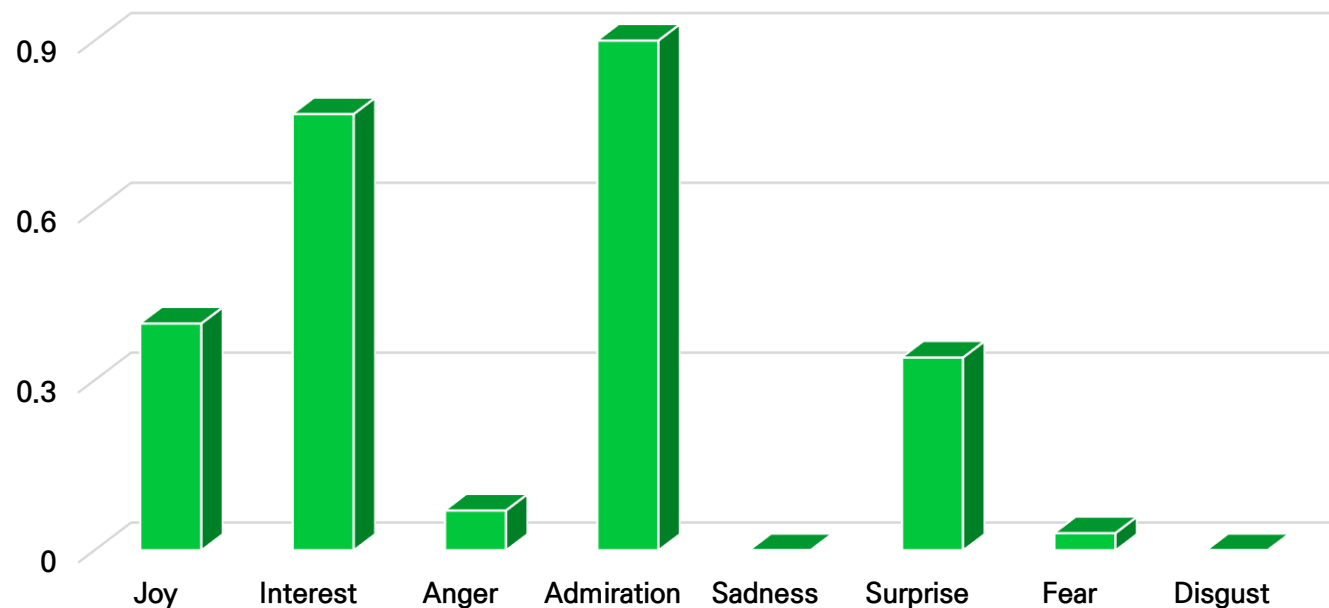
- 텍스트가 Input으로 들어오면 **특정 감정일 확률**을 Output으로 받는 모형 구축

Input:



Output:

“페르소나 시스템은 최고인 것 같아요! 100점 주고 갑니다~”



01. 문제 정의

요구사항 분석



- 앞선 요구 사항에 대해 다음의 6가지 해결 과제를 선정
- 3개월 간 다음의 과제를 수행하기 위한 process를 수립 및 실시

“

2D-SLAP

”



Differentiation

기존 연구와의 차별성은?



Data

어떤 Domain의 데이터로 분석할 것인가?



Sentiment

감정의 범주에 대한 정의를 어떻게 내릴 것인가?



Labeling

Label이 없는 데이터로 어떻게 학습할 것인가?



Algorithms

확률 값을 얻기 위해 어떤 방식을 사용할 것인가?



Preprocessing

Text 데이터 전처리 및 임베딩을 어떻게 실시할 것인가?(전체)

01. 문제 정의

감정 범주 정의

- 러셀의 감정, 동 서양에서 고전적인 감정에 대해 정의한 것을 소개한 후
- 우리가 택할 감정은 범주이고 극성이 아닌 카테고리별 이라는 것을 강조

01. 문제 정의

감정 범주 정의

Ch02 관련 연구

- 극성 분류
- Topic Modeling
- Language Model

02. 관련 연구

02. 관련 연구

02. 관련 연구

02. 관련 연구

Ch03 활용 데이터

- NSMC
- Crawled Data
- Senti. Word Dictionary
- Sejong-corpus
- Simple EDA

03. 활용 데이터

03. 활용 데이터

03. 활용 데이터

ksenticnet

5,465개 단어
4차원의 sentic value
각 단어 별 2개의 감정 매핑
총 8개의 감정으로 구성 (admiration, anger, disgust, fear, interest, joy, sadness, surprise)
극성에 대한 자료도 존재 (positive vs negative, [-1, 1])
각 단어 별 유사한 단어도 기입되어 있음

sentiwordknu

428개 단어
감정정도의 평균과 분산이 존재
각 단어 별 1개의 감정 매핑
총 11개의 감정으로 구성 (혐오, 슬픔, 기쁨, 중성, 흥미, 분노, 지루함, 놀람, 공포, 기타, 통증)
극성에 대한 자료 X
각 단어 별 유사한 단어도 기입 X

03. 활용 데이터

Ch04 데이터셋 구축

- 띄어쓰기, 오타자 전처리
- 데이터 라벨링 – 감정 사전 매핑
- 데이터 라벨링 – JST
- Tokenizing
- BERT Tokenizing

04. 데이터셋 구축

띄어쓰기, 오타자 전처리

최초의 시도 : soynlp의 maxtokenize
Sejong corpus try
But 잘 안됨...

Soyspacing, 핑퐁의 라이브러리 활용하여 띄어쓰기 선 전처리 실시

04. 데이터셋 구축

04. 데이터셋 구축

04. 데이터셋 구축

04. 데이터셋 구축

04. 데이터셋 구축

Ch05 Modeling

- Process
- 분류 모델
- KorBERT

05. Modeling

05. Modeling

05. Modeling

05. Modeling

05. Modeling

05. Modeling

05. Modeling

05. Modeling

05. Modeling

05. Modeling

Ch06 결론 및 향후 방향성

- 결론
- 향후 방향성

06. 결론 및 향후 방향성

06. 결론 및 향후 방향성

“

여러분의 6개월은
어떠셨나요?

”

감사합니다.

Appendix.

Appendix.

Reference.
