KUBIG NLP CONTEST

★★★★★

뉴스 제목을 통한 다중 감성 분석

텍스트 다중 분류

13기 김현지 , 13기 박주영

팀 소개

뉴스 인사이드 아웃

13기 김현지 13기 박주영

차례

VOL.000
Contents

★★★★★

NLP Project

뉴스 제목을 통한 다중 감성 분석 프로젝트

뉴스 제목을 통한 다중 감성 분석

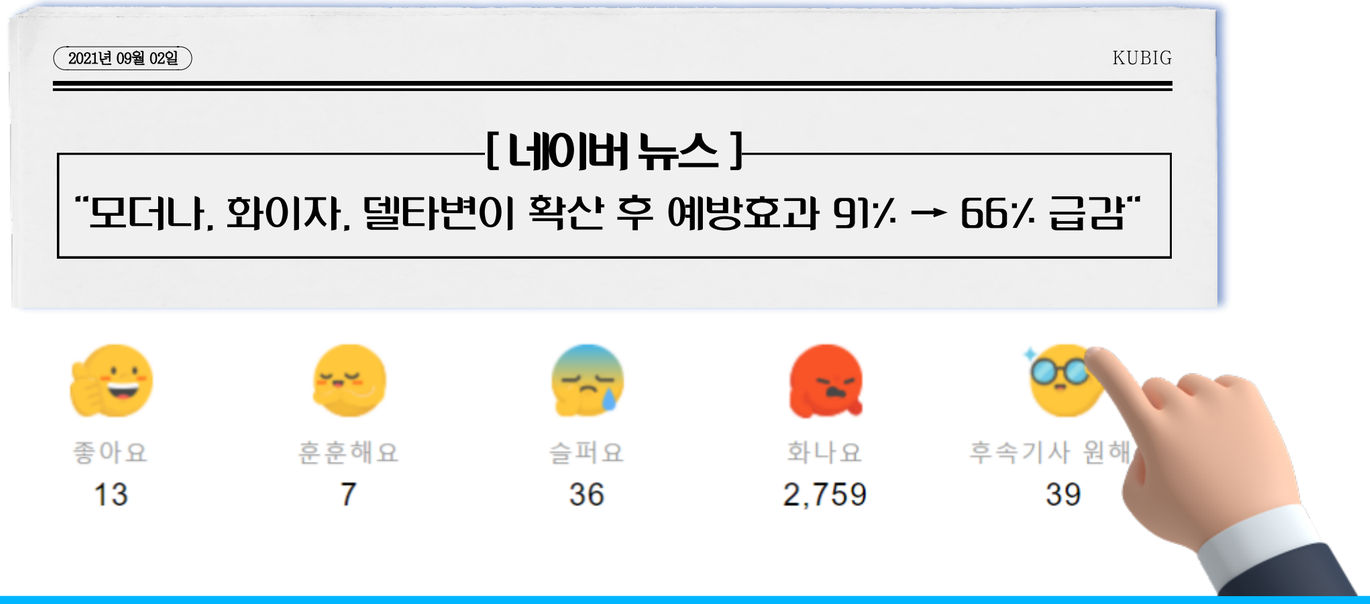
★★★★★

VOL.001 분석 주제

분석 주제

뉴스 제목을 보고, 독자들의 반응 예측해보기

네이버 뉴스 독자들은 해당 뉴스 기사에 대해 '좋아요', '훈훈해요', '슬퍼요', '화나요', '후속기사 원해요'의 총 5가지의 감정 표현을 할 수 있다.
본 프로젝트에서는 뉴스 제목을 통해 이러한 독자들의 감정 반응을 예측해보고자 한다.



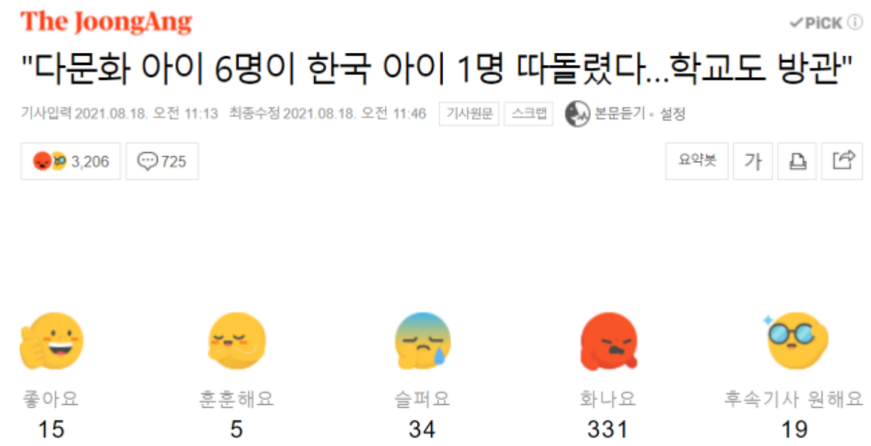
데이터 수집

★★★★★

VOL.002 데이터 수집

다양한 분야의 뉴스의 제목과 반응별 공감수 크롤링

네이버 뉴스 크롤링



뉴스 제목
각 뉴스별 헤드라인 제목 수집

반응별 공감수
5개 반응에 대한 공감수 수집

다양한 분야의 뉴스 데이터
코로나, 경제, 환경, 생활 등

데이터 수집

★★★★★

VOL.002 데이터 수집

수집된 데이터 형태 (예시)

기사명	좋아요	훈훈해요	슬퍼요	화나요	후속기사 원해요
통계청 2분기 가계동향조사 결과 발표	2	0	5	6	15
김영론 지사, 여성경제인 간담회.."전남 행복시대 동참을"	6	2	1	0	8
전경련"한국 갈등지수 최악..정치,경제,사회 모두 심각"	4	2	10	17	4

데이터 정제

결측치 포함 데이터 제거
공감수 총합이 1개 미만인 데이터 제거

라벨링

각 기사 제목에 대응하는 감정을 1개씩 라벨링

데이터 셋 완성

판다스 데이터 프레임 형태
기사명 / 레이블

데이터 수집

★★★★★

VOL.002 데이터 수집

라벨링

1. 각 기사에 표시된 공감 반응에 대해 1st Max, 2nd Max 확인
2. '좋아요-훈훈해요'에 대해 라벨링을 '훈훈해요'로 부여
   '화나요-슬퍼요'에 대해 '슬퍼요' 라벨링 부여
3. '후속 기사 원해요'는 표본이 적어 그냥 제거 + '감정'만을 확인

기사명	1st Max	2nd Max
실종 33년 만에 가족과 눈물의 상봉 40대 여성	좋아요	훈훈해요
수척해진 김정은에 눈물 젖는다던 北...다이어트 만화 됐다	화나요	좋아요
"건강했던 남편 백신 맞고 급성 백혈병 진단 후 숨져" 아내 올린 눈물의 靑청원	화나요	슬퍼요
'윤희숙 사퇴'에 분노한 野...눈물 흘린 이준석 "야만적 연좌제"	화나요	슬퍼요
의원직도 내려놓겠다는 윤희숙, 여당은 비판 세례 "기만자 눈물의 사퇴쇼"	화나요	좋아요

```
df_like = df[(df['1st Max'] == '좋아요')]
df_like['label'] = np.where(df_like['2nd Max'] == '훈훈해요', '훈훈해요', '좋아요')

df_angry = df[(df['1st Max'] == '화나요')]
df_angry['label'] = np.where(df_angry['2nd Max'] == '슬퍼요', '슬퍼요', '화나요')

df_warm = df[(df['1st Max'] == '훈훈해요')]
df_warm['label'] = '훈훈해요'

df_sad = df[(df['1st Max'] == '슬퍼요')]
df_sad['label'] = '슬퍼요'

df_sad = df[(df['1st Max'] == '슬퍼요')]
df_sad['label'] = '슬퍼요'
```

데이터 수집

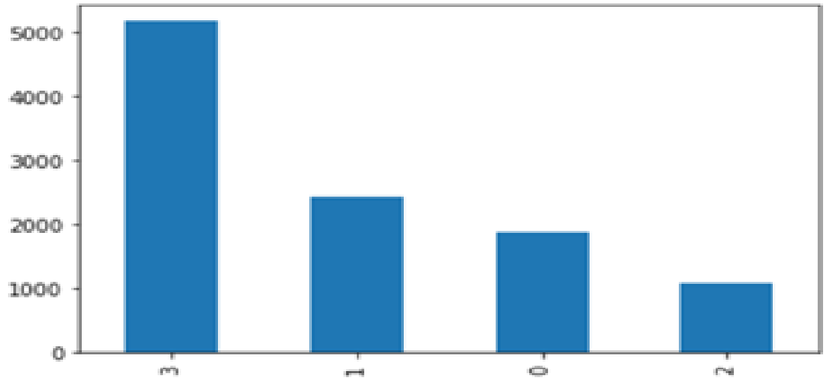
★★★★★

VOL.002 데이터 수집

최종 데이터 셋 형태 (예시)

기사명	레이블
통계청 2분기 가계동향조사 결과 발표	1
김영론 지사, 여성경제인 간담회.."전남 행복시대 동참을"	0
전경련"한국 갈등지수 최악..정치,경제,사회 모두 심각"	3

레이블 비율



데이터 전처리

★★★★★

VOL.003 데이터 전처리

01	**데이터 셋 분리**	Train / Test 데이터셋으로 분리
02	**불용어 제거**	정규표현식을 사용해 한글을 제외한 문자 모두 제거
03	**토큰화**	KONLPY의 Okt를 사용해 형태소 추출을 통한 토큰화 진행

각 감정 별 빈도가 높은 단어

좋아요: '원금', '흥행', '청원', '국민', '재난', '금리', '유튜브'

훈훈해요: '친환경', '채용','흥행', '유튜브', '코로나', '후원', '감동'

슬퍼요: '청원', '눈물', '사망', '국민', '코로나', '접종', '백신'

화나요: '원금', '청원', '국민', '대출', '백신', '재난', '코로나'

04	**패딩**	해당 단어 리스트를 정수 인코딩으로 벡터화 시킨 후, 문장의 길이를 25로 패딩 진행

모델링

★★★★★

VOL.004 모델링

자료의 불균형 문제 해결

01

불균형 처리를 하지 않은 데이터

minor한 클래스에 특별한 관심을 두는 게 아닌 경우 불균형 상태를 그냥 두는 것이 나을 수도 있다

02

SMOTE

oversampling을 통해 각 class의 샘플의 개수를 같게 만드는 방법

03

class weight

loss에 가중치를 두어 각 클래스가 loss에 미치는 영향을 동등하게 만들어 불균형 해소

04

focal loss

다중 클래스 분류 문제의 역전파 때 가중치를 업데이트하는 과정에서 loss의 계산에 현재까지의 클래스 별 정확도를 고려한 가중치를 부여하여 즉 분류하기 어려운 데이터에 대해 집중을 하게 되어 전반적인 모델의 정확도를 높이는 방법

모델링

★★★★★

VOL.004 모델링

6가지 모델을 사용해 예측을 진행한 후 **성능을 비교** 해본다.

1

LSTM

2

KoBERT

3

Machine Learning

Random Forest / SVM /
Naive Bayes / Logistic Regresson

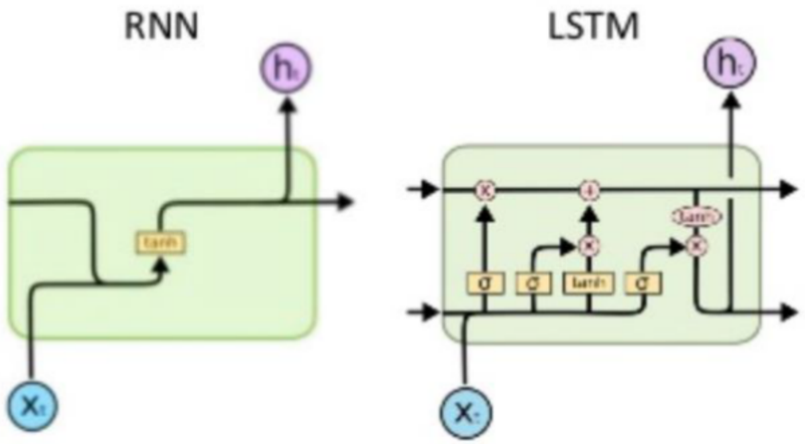
모델 1 | LSTM

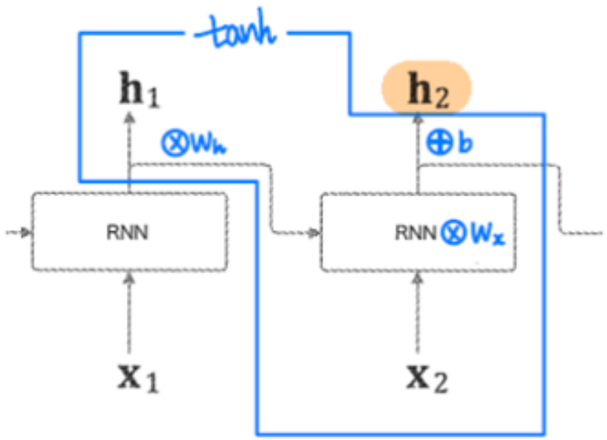
★★★★★

VOL.004 모델링

LSTM 순환하는 경로를 구현한 딥러닝 알고리즘인 RNN에 대해 미분 사라짐 문제를 해결하기 위해 제안된 방법

- 뉴런들이 자기 자신과 연결(전 시점의 출력이 현 시점의 입력)되기 때문에 과거의 정보를 기억할 수 있음

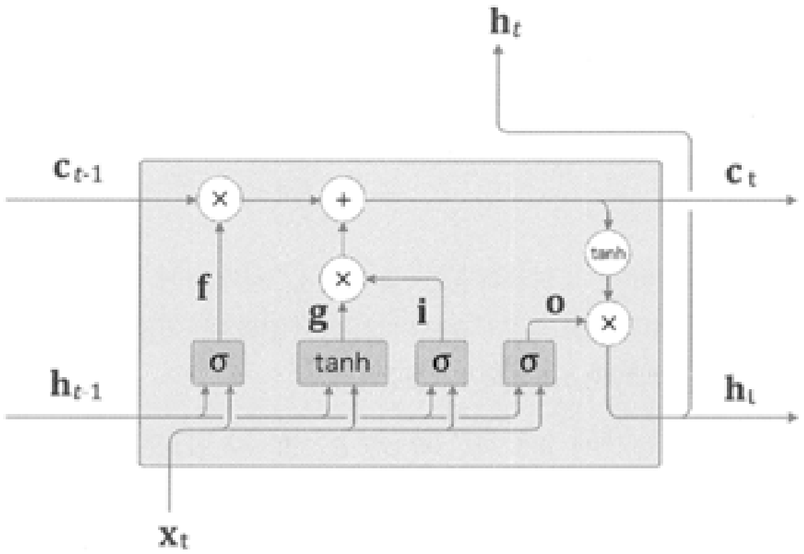




모델 1 | LSTM

★★★★★

VOL.004 모델링



LSTM의 구조

1) 메모리 셀(c)

과거의 필요한 정보를 간직한다. 이를 바탕으로 은닉상태를 계산한다.

2) output 게이트(o)

기억 셀의 원소들이 다음 시각의 은닉상태에 얼마나 중요한가를 반영 (x)

3) forget 게이트(f)

무엇을 잊어야 하나를 반영 (x)

4) 새로운 기억 셀(g)

새로 기억해야 할 정보를 추가 (+)

각 게이트들은 전 시점의 은닉 상태와 현 시점의 입력 데이터를 받아 열림 상태를 반영하기 위해 sigmoid 함수 계산을 통해 출력

모델 1 | LSTM

★★★★★

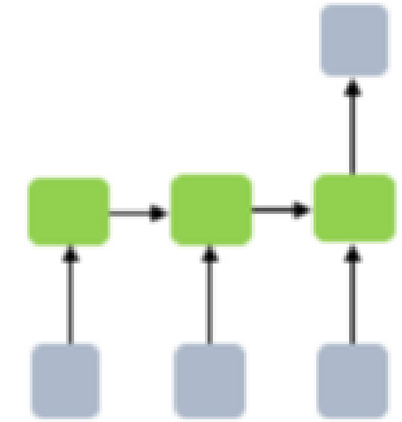
VOL.004 모델링

LSTM과 Classification

Many-to-one 문제: 모든 시점(time step)에 대해서 입력을 받지만 최종 시점의 셀만이 은닉 상태를 출력해 적절한 활성화 함수를 사용해 정답을 선택

다중 클래스 분류

1) 출력층의 활성화 함수로 소프트맥스 함수를, 손실 함수로 categorical_crossentropy를 사용
2) 클래스가 N개라면 출력층에 해당되는 밀집층(dense layer)의 크기는 N



다 대 일(many-to-one)

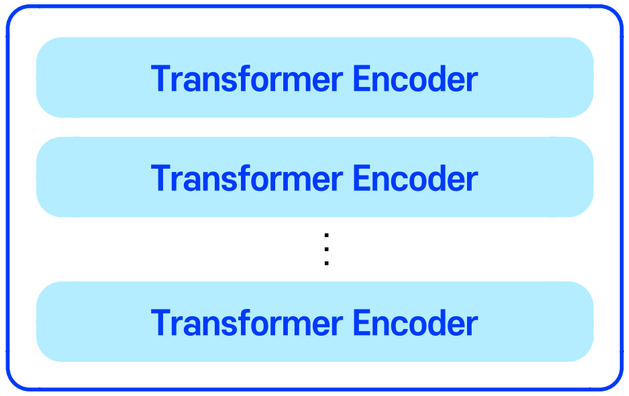
모델 2 | KoBERT

★★★★★

VOL.004 모델링

BERT 트랜스포머를 이용해 구현되었으며위키피디아(25억개)와 BookCorpus(8억개)와 같은 레이블이 없는 텍스트 데이터로 사전 훈련된 언어 모델
KoBERT: BERT 모델에서 한국어 데이터를 추가로 학습시킨 모델로 한국어 위키의5백만개의 문장과 54백만개의 단어로 학습되었다.

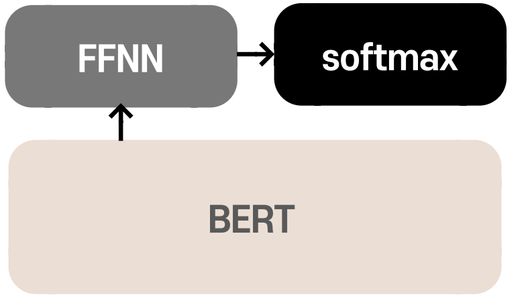
기본 구조



트랜스포머의 인코더를 쌓아올린 구조

모델 사용

예) 스팸 메일 분류기



사전 훈련된 BERT를 가지고 다른 Task에서의 추가 훈련과 함께
하이퍼파라미터를 재조정하여 내가 원하는 Task를 수행할 수 있다.

모델 2 | KoBERT

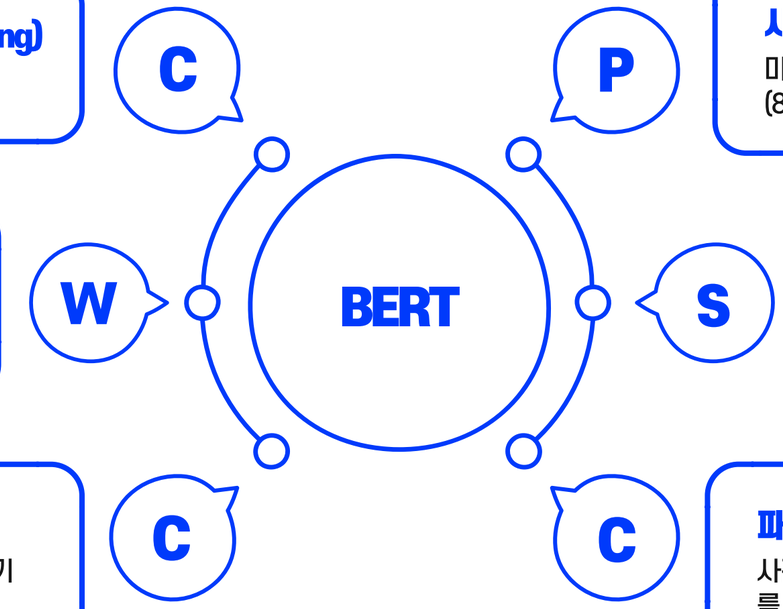
★★★★★

VOL.004 모델링

문맥을 반영한 임베딩(Contextual Embedding)
문맥을 더욱 자연스럽게 파악할 수 있다.

서브워드 토크나이저: WordPiece
단순히 띄어쓰기로 토큰을 나누는 것보다 효과적으로 토큰을 구분한다.

포지션 임베딩(Position Embedding)
각 토큰의 위치를 알려주는 임베딩. 입력을 한 번에 받기 때문에 각 토큰의 위치를 기억할 필요가 있다.



사전 훈련(Pre-training)
마스크드 언어 모델, 다음 문장 예측 방법으로 BookCorpus (8억 단어)와 위키피디아(25억 단어)로 학습되었다.

세그먼트 임베딩(Segment Embedding)
두 개의 문장을 입력받기도 하는데, 이때 두 개의 문장을 구분하기 위한 임베딩 층이 존재한다.

파인 튜닝(Fine-tuning)
사전 학습된 BERT에 우리가 풀고자 하는 태스크의 데이터를 추가로 학습 시킨다.

모델 2 | KoBERT

★★★★★

VOL.004 모델링

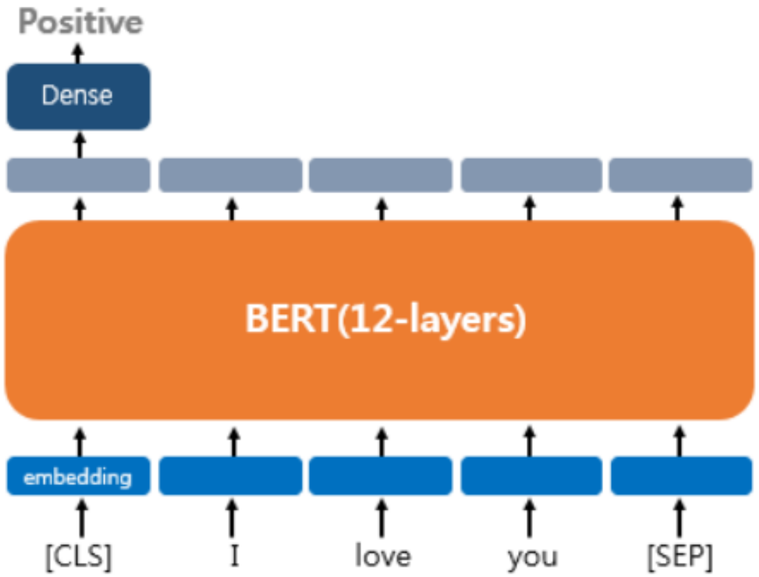
BERT를 Fine-tuninig 하기

Text Classification

파인 튜닝 단계:

사전 학습된 BERT에 우리가 풀고자 하는 태스크의 데이터를 추가로 학습 시켜서 테스트하는 단계. 실질적으로 태스크에 BERT를 사용하는 단계에 해당한다.

문서의 시작에 [CLS] 라는 토큰을 입력한다. 텍스트 분류 문제를 풀기 위해서 [CLS] 토큰의 위치의 출력층에서 밀집층(Dense layer) 또는 같은 이름으로는 완전 연결층(fully-connected layer)이라고 불리는 층들을 추가하여 분류에 대한 예측을 하게 된다.



Baseline | 머신러닝 모델

★★★★★

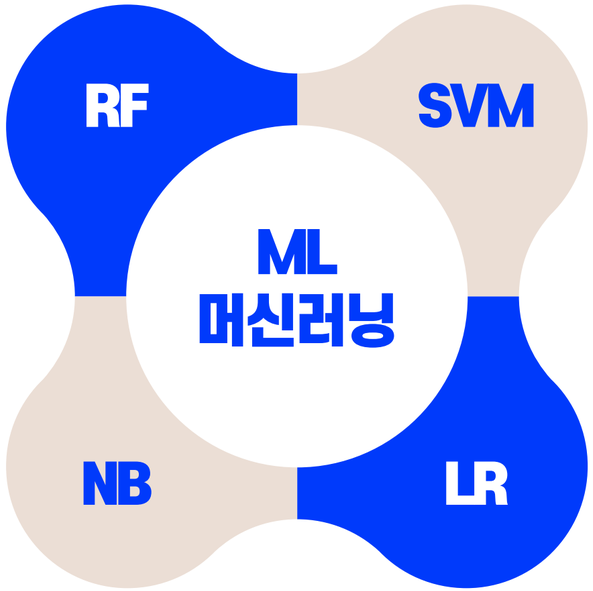
VOL.004 모델링

Random Forest

Decision Tree classifier를 바탕으로
구한 예측치들에 대한 앙상블 기법

Naive Bayes

특성들 사이의 독립을 가정하는 베이즈
정리를 적용한 확률 분류기의 일종



SVM

각 그룹의 관측치 중 가장 가까운 거리를 바탕으로 생성된 분류선 중
그 밴드가 가장 두껍도록 boundary(green)를 결정해 그룹을 분류하는
방법 → 범주형 변수를 다루므로 SVC (OneVsRest Classifier)로 다중
분류 확장 가능

Logistic Regression

귀를 사용하여 데이터가 어떤 범주에 속할 확률을 0에서 1 사이의 값으로
예측하고 그 확률에 따라 가능성이 더 높은 범주에 속하는 것으로 분류해
주는 지도 학습 알고리즘

결과 비교

★★★★★

VOL.005 결과

LSTM **KoBERT** **ML** (Logistic Regressor)

	LSTM	KoBERT	RF	SVM	NB	LR
기본	0.6302	0.6588	0.6277	0.5516	0.6148	0.6446
class weight	0.5988	0.6313	0.6296	0.5531	0.5987	0.6175

결과

★★★★★

VOL.005 결과

```
end = 1
while end == 1 :
    sentence = input("기사 제목을 입력하세요 : ")
    if sentence == "끝" :
        break
    predict(sentence)
    print("₩n")
```

... 기사 제목을 입력하세요 :

결과

★★★★★

VOL.005 결과

기사 제목을 입력하세요 : "경비실 에어컨 전기료 내라" 논란에 경비원 "주민들은 몰라"
>> 이 기사에 대한 주된 반응은 화나요 로 예상됩니다.

기사 제목을 입력하세요 : 의정부 수락산서 60대 등산객 벌에 쏘여 사망
>> 이 기사에 대한 주된 반응은 슬퍼요 로 예상됩니다.

결과

★★★★★

VOL.005 결과

기사 제목을 입력하세요 : 홍진경, 2억 4천만 원 쾌척! 열일. 행보 속 돋보이는 선행
>> 이 기사에 대한 주된 반응은 훈훈해요 로 예상됩니다.

기사 제목을 입력하세요 : K팝 4세대 주도권을 선점하라.. 쑥쑥 크는 차세대 아이돌
>> 이 기사에 대한 주된 반응은 좋아요 로 예상됩니다.

기사 제목을 입력하세요 :

발표를 마칩니다

감사합니다

뉴스 인사이드 아웃 | 13기 김현지, 13기 박주영