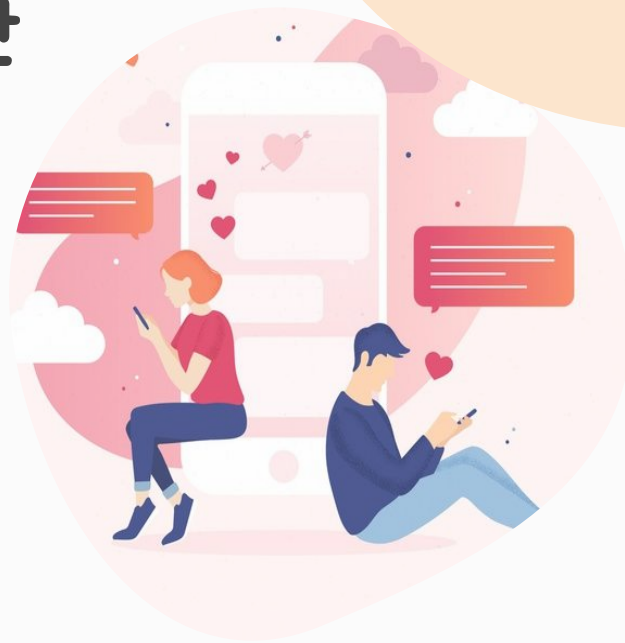


# 자연어 처리 및 딥러닝 모델을 통한 데이팅 앱 유저 리뷰 감성분석

AI 07기 임현민



## 목차

01

문제 정의

02

데이터 소개 및 가설 설정

03

EDA 및 데이터 전처리

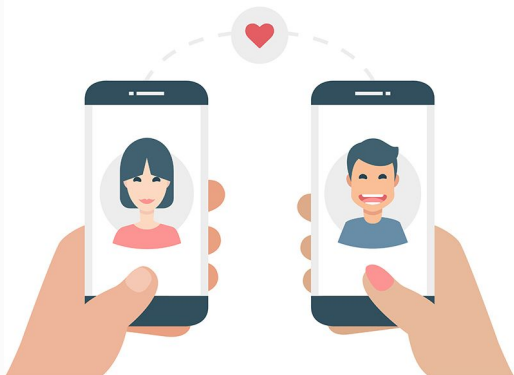
04

모델 비교

05

결과 및 한계점

## 문제 정의



### 01 데이트 앱 시장 급성장

- 전세계 3조원 규모 (2020)
- 국내 830억원 규모 (2020)

### 02 경계심은 여전히 높은 상황

- 범죄, 불건전한 목적 앱 악용
- 불필요한 과금 유도

### 03 앱 유저 리뷰 감성분석

- 리뷰 텍스트 데이터를 이용한 자연어 처리 및 딥러닝 모델로 서비스에 대한 감성이 어떨지 파악하고자 한다.






## 데이터 소개

### 데이팅 앱 '범블(BUMBLE)' 구글 플레이스토어 리뷰 데이터






- 데이터 수집 방법
  - 케글(Kaggle) CSV 파일 다운로드
- 데이터 세부 정보
  - 총 105,954 개
  - Content -> 앱 리뷰 텍스트
  - Score -> 앱 평가 점수



#### 소비자 지출기준

	Tinder
	Bumble App
	Pairs
	Badoo
	MeetMe

#### 평균 월간 사용자

	Tinder
	Badoo
	Bumble App
	happn
	Grindr

## 가설 설정

### Model A

Score 3 -> 긍정에 포함

Score	Label
1	부정
2	
3	긍정
4	
5	

이진분류

VS

### Model B

Score 3 -> 중립으로 분리

Score	Label
1	부정
2	
3	중립
4	긍정
5	

다중분류

### 가설

레이블을 세분화한  
모델 B의 정확도가 더  
높을 것이다.

# 데이터 전처리

## 전처리 1

- **결측값 & 중복값 제거**  
(105,954 → 86,923)

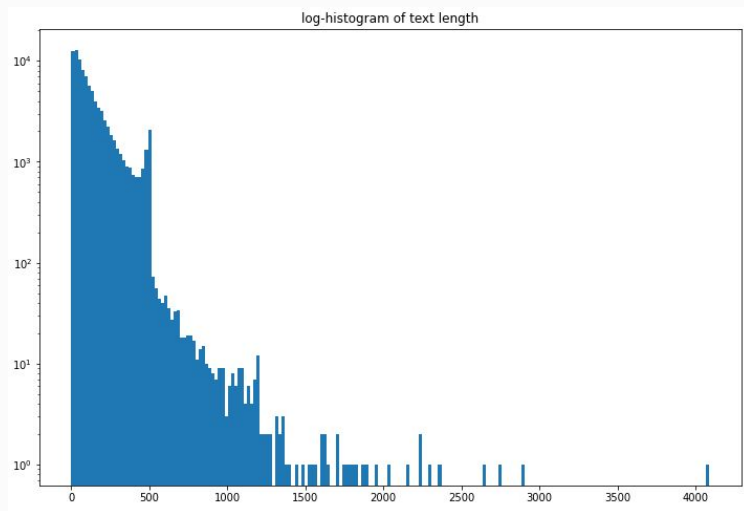
## 전처리 2

- **축약형 단어 확장**  
(I'm / didn't 등)  
-> Contractions
- **구두점, 특수문자, 이모지 제거**  
-> re.sub()
- **소문자 통일**  
-> lower()

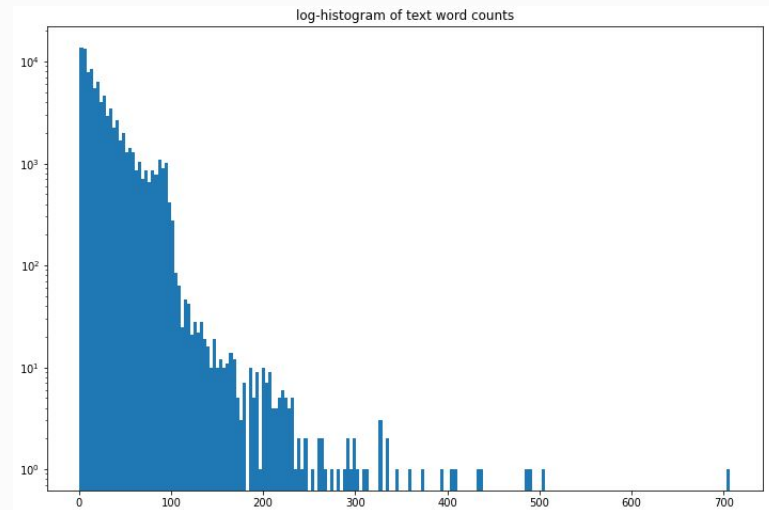
## 전처리 3

- **불용어 제거**  
-> spaCy
- **표제어 처리**  
-> spaCy

## EDA - 텍스트 데이터 분포

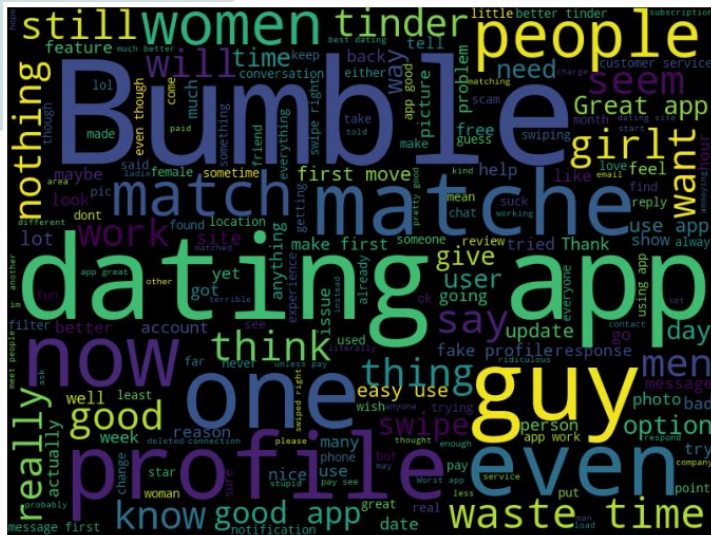


- 텍스트 길이 최대값: 4088
- 텍스트 길이 최소값: 1
- 텍스트 길이 중간값: 90
- 텍스트 길이 평균: 137

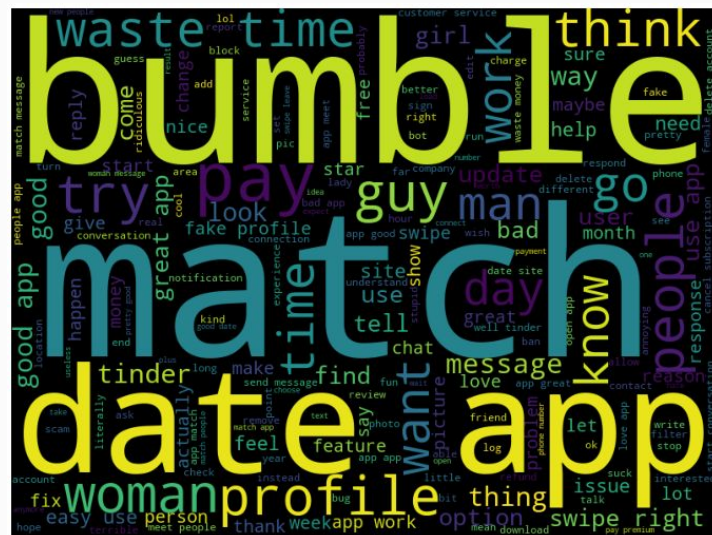


- 텍스트 하나당 단어 갯수 최대값: 707
- 텍스트 하나당 단어 갯수 최소값: 1
- 텍스트 하나당 단어 갯수 중간값: 17
- 텍스트 하나당 단어 갯수 평균: 26

## EDA - 워드 클라우드 비교



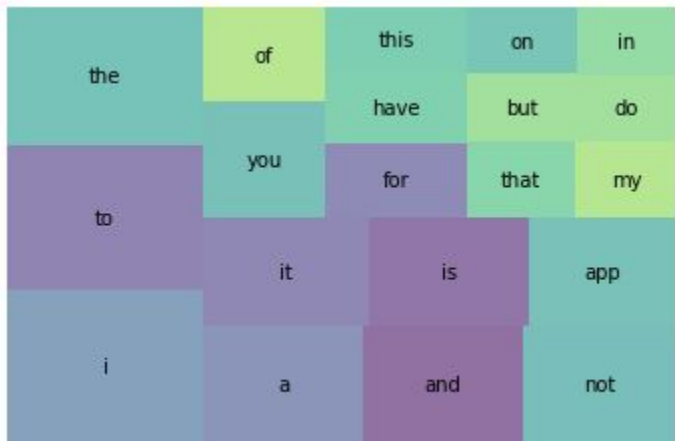
## 전처리 전



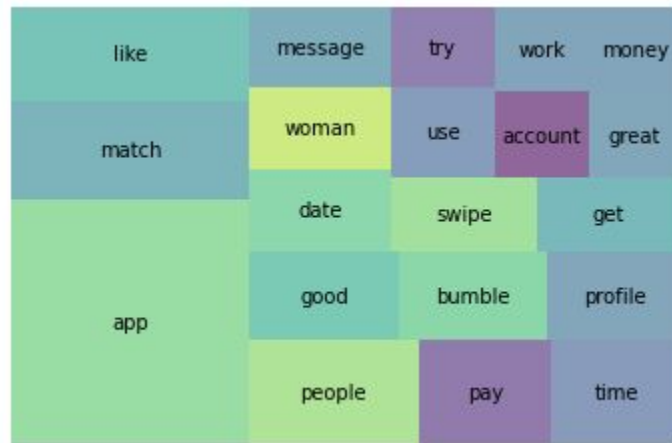
## 전처리 후



## EDA - Top20 단어 비교



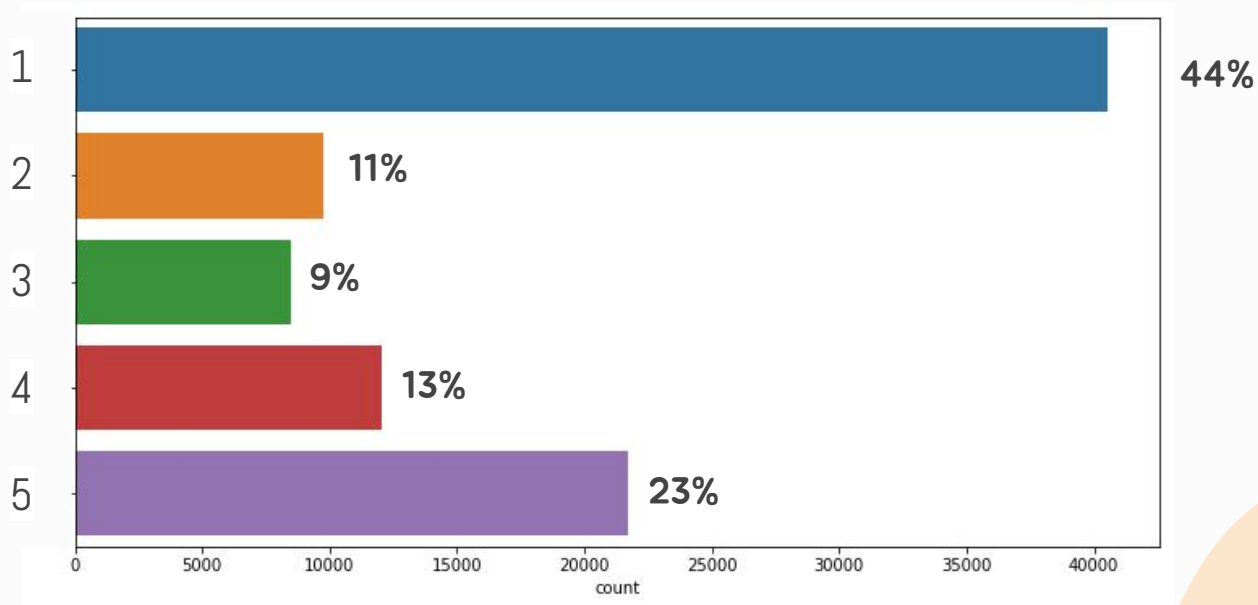
전처리 전



전처리 후

## EDA - 타겟 레이블 분포

SCORE

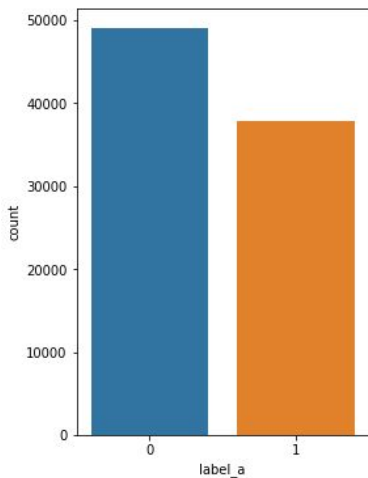


# EDA - 타겟 레이블 분포

Model A

Score 3 -> 긍정에 포함

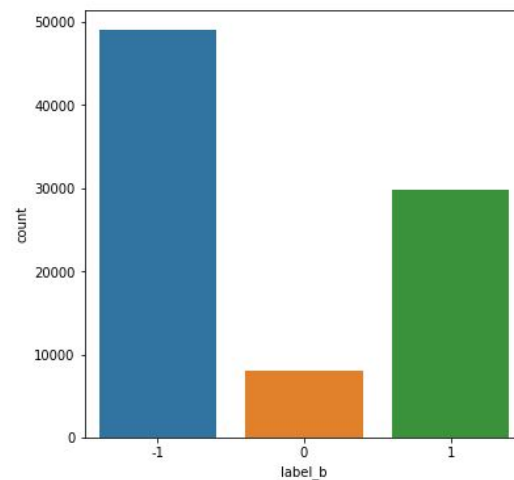
Score	Label
1	부정 (0)
2	
3	긍정 (1)
4	
5	



Model B

Score 3 -> 중립으로 분리

Score	Label
1	부정 (0)
2	
3	중립 (1)
4	긍정 (2)
5	




# 모델 학습



## LSTM 모델

- 텍스트: 단어의 순서가 중요한 대표적인 시퀀스 데이터
- RNN 은 과거 정보를 기억하고 이에 맞추어 새로운 샘플을 처리할 수 있기 때문에 시퀀스 데이터를 다룰 때 장점을 갖지만, 단어의 길이가 길어지면 앞쪽에 입력된 단어의 의미가 사라지게 되는 문제가 있다.
- LSTM 은 문장이 길어지더라도, 앞 단어를 더 오래 기억할 수 있기 때문에 기존 RNN 의 단점을 보완한다.

## 모델 학습

	Model A 	Model B
학습 데이터셋 정확도	0.8503	0.8085
검증 데이터셋 정확도	0.8251	0.7954
테스트 데이터셋 정확도	0.8332	0.7997
Chance level	0.5	0.333

## 가설

레이블을 세분화한 모델 B의 정확도가 더 높을 것이다. -> (X)

## 결과

### 한계점 및 추후 보완점

- 이모지에 대한 전처리 미흡
  - 이모지를 데이터로서 어떤식으로 변환하여 사용할 수 있을지 고민 필요
- 다양한 하이퍼파라미터 튜닝 부족
  - KerasTuner 활용 세밀한 튜닝 필요
  - 모델 성능 높이기

The background features abstract organic shapes in a light orange color on the left and a light teal color on the right, set against a white background. The word "Thanks!" is centered in a dark grey serif font.

Thanks!