

자연어 처리 및 딥러닝을 이용한 고객 컴플레인 주제 분류

AIB 07기 임현민

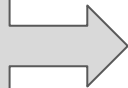
Contents

- 1 주제 선정/데이터 소개
- 2 이전 프로젝트 보완점
- 3 EDA 및 텍스트 전처리
- 4 모델
- 5 결과 및 한계점



고객 컴플레인에 대해
주제를 자동 분류할 수 있을까?

TEXT
DATA



자연어처리

딥러닝

A

B

C

D



미국 금융 서비스에 대한 지역별 컴플레인 데이터

✓ 데이터 수집 방법:

- CFPB (미국 소비자 금융 보호국) 웹사이트에서 CSV 파일 다운로드

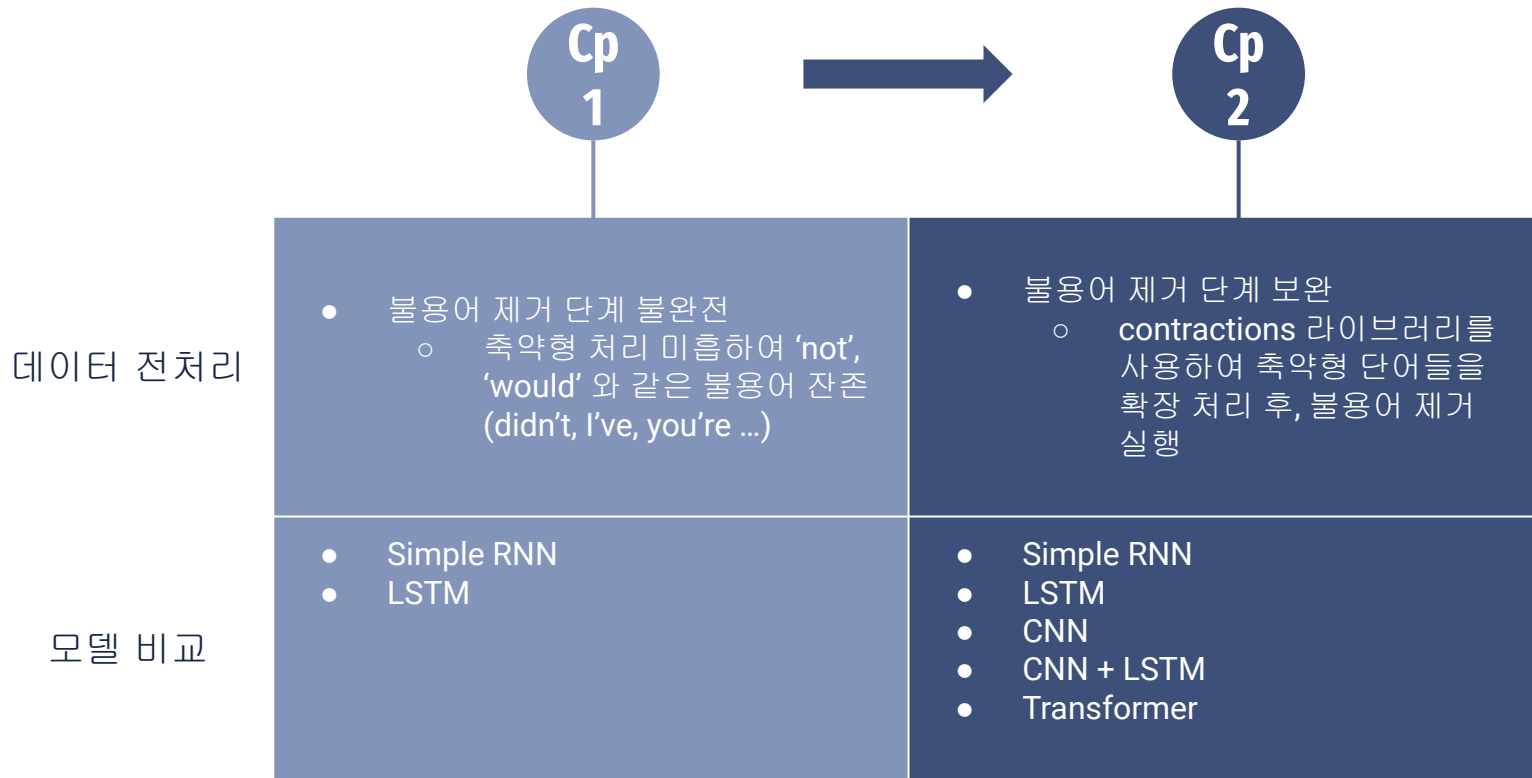
✓ 데이터 정보:

- 2021/01/01 ~ 2021/12/31
- 캘리포니아
- 총 25,126개 텍스트 데이터

가설

**고객 컴플레인 데이터에 대해
자연어 처리 및 딥러닝을 이용하여 주제를 분류할 수 있다.
(정확도 90%)**

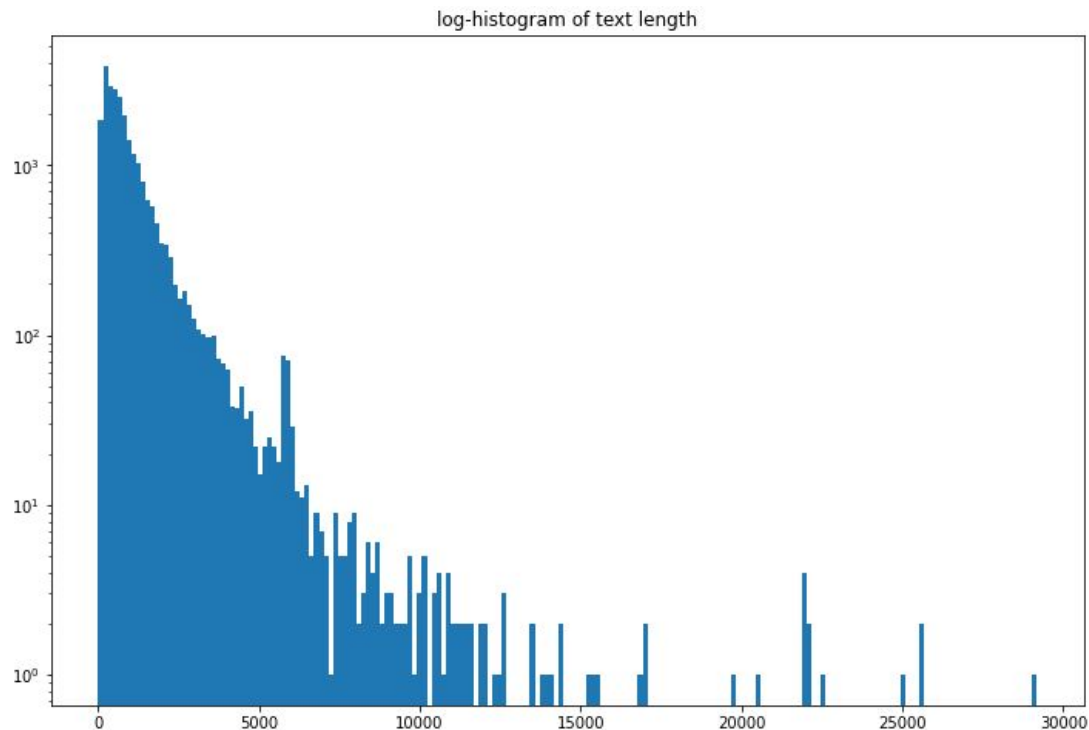
이전 프로젝트 대비 보완점



Project Pipeline

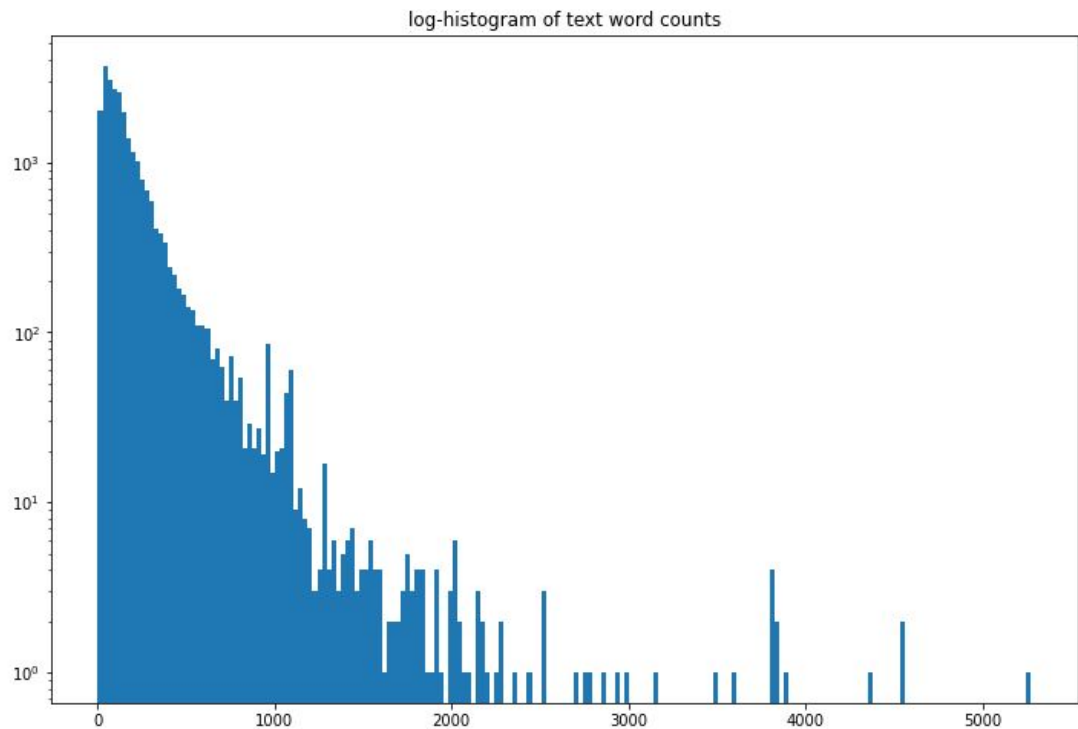


텍스트 길이 분포



- 텍스트 길이 최대값: **29174**
- 텍스트 길이 최소값: **9**
- 텍스트 길이 중간값: **660**
- 텍스트 길이 평균: **1018**
- 텍스트 길이 표준편차: **1297**

텍스트 내 단어 개수 분포



- 텍스트 길이 최대값: **5271**
- 텍스트 길이 최소값: **2**
- 텍스트 길이 중간값: **119**
- 텍스트 길이 평균: **183**
- 텍스트 길이 표준편차: **232**

Word Cloud



- XXXX
-> 개인정보
- credit, account 등
-> 금융관련
- Bank, Reported Identity
-> 대소문자 혼용

...

And now we are in the month of **XX/XX/XXXX** and still nothing from them. That means Experian has NOT complied with the Fair Credit Reporting Act law and continued to report inaccurate information on my credit report. REMOVE THESE FOLLOWING ACCOUNTS COMPLETELY FROM MY CREDIT REPORT : (**XXXX XXXX XXXX XXXX, XXXX XXXX, XXXX XXXX XXXX XXXX XXXX XXXX XXXX XXXX : XX/XX/XXXX** Reference # : **XXXX, XXXX XXXX XXXX XXXX** Filed/Reported : **XX/XX/XXXX** Reference # : **XXXX**)

...

Im a loyal and happy Comenity customer, Ive read Comenity has made adjustments for people affected by Covid and I hope that Comenity can remove the **XX/XX/XXXX**, late payments off BOTH my accounts, given that Im a **XXXX XXXX**. \n\nAccounts information : Account information number 1 : Account name : COMENITY BANK/EXPRESS Account number : **XXXX** Balance : **{\$83.00}**

...

텍스트 특징 확인


99%


대문자


77%

연속된 X문자


39%

\n 문자


96%

마침표 (.)


69%

쉼표 (,)


7%

물음표 (?)

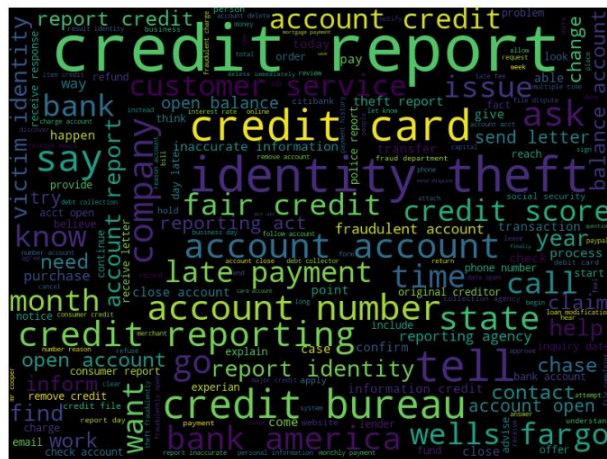
Project Pipeline



Word Cloud 비교



전처리 전

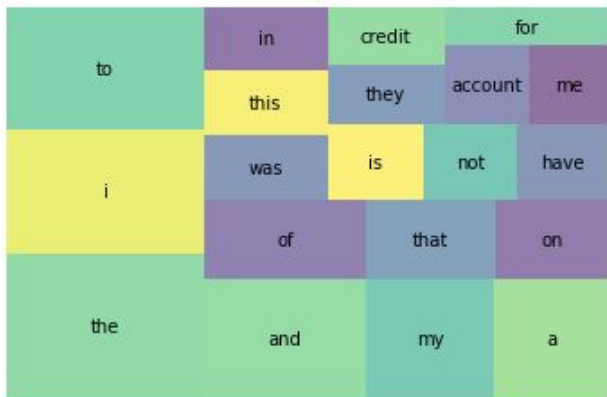


전처리 후
(cp 1)

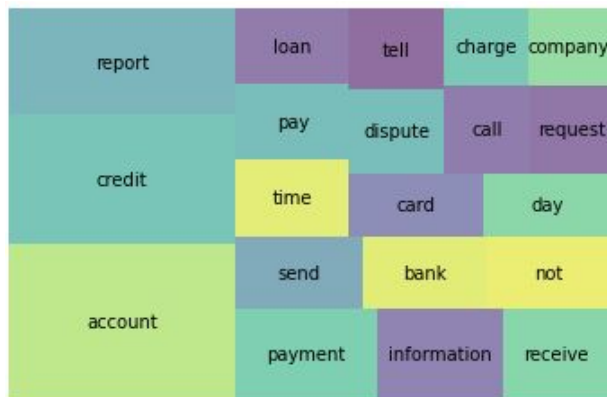


전처리 후
(cp 2)

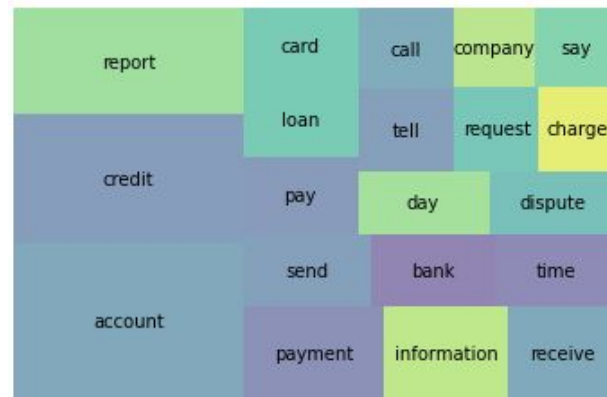
빈도높은 단어 TOP 20 비교(Squarify)



전처리 전



전처리 후
(cp 1)



전처리 후
(cp 2)

타겟 레이블 분포

신용 리포트, 신용 복구 서비스

채권 추심

카드

주택저당대출

예금 계좌

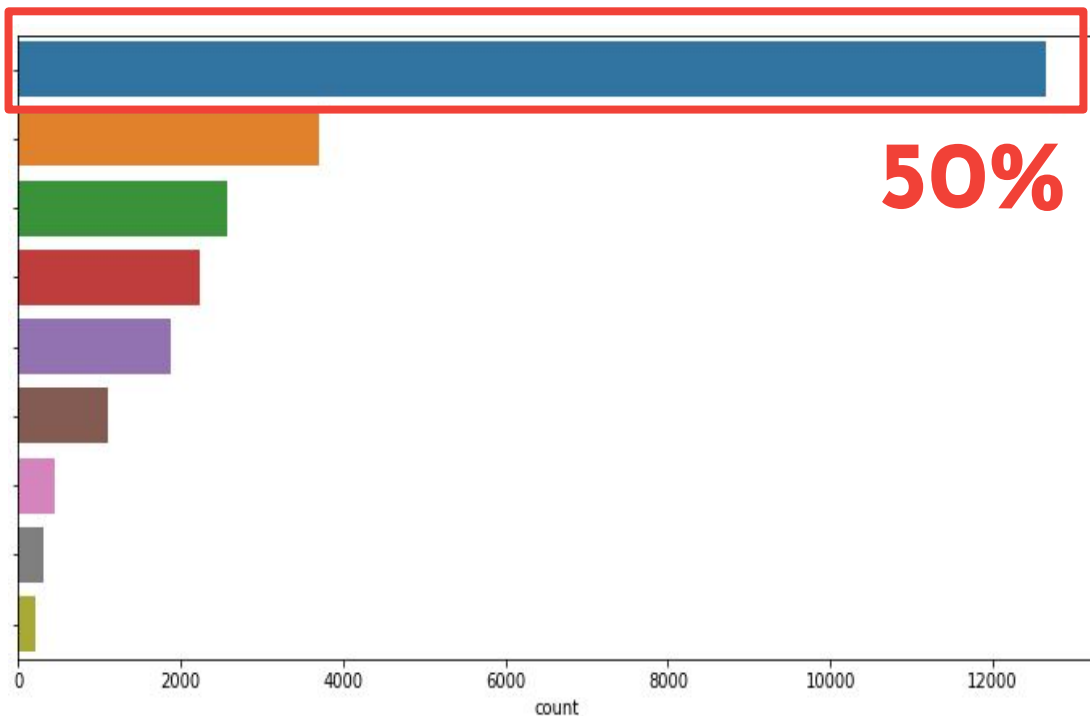
송금, 가상화폐/통화서비스

자동차 대출, 임대

개인 대출

학자금 대출

50%



레이블 인코딩

0	예금 계좌
1	카드
2	신용 리포트, 신용 복구서비스
3	채권 추심
4	송금, 가상화폐/통화서비스
5	주택저당대출
6	개인 대출
7	학자금 대출
8	차량 대출, 임대

모델 비교

SimpleRNN	단어의 입력 순서를 중요하게 반영하며 학습한다.
LSTM	문장의 길이가 길어지더라도, 앞 단어를 더 오래 기억할 수 있다.
CNN	문장의 지역 정보를 보존하고 반영하며 학습한다.
CNN + LSTM	문장의 지역 정보를 반영하며 특징을 추출하는 CNN 과 감지 능력이 탁월한 LSTM 을 결합한 모델.
Transformer	RNN 이 없는 인코더-디코더 구조를 갖고, 병렬연산이 가능하기때문에 학습이 빠르고 성능이 좋다.

모델 성능 비교

Simple RNN	LSTM	CNN	CNN + LSTM	Transformer
0.7197	0.8137	0.8396	0.8134	0.8276

결과

테스트 데이터 평가

Cp1 Cp2
80% -> 84%

(기준모델 71%)

한계점 및 추후 발전 과제

- BERT 모델에 대한 스터디 진행후, 모델 구축하여 기존 모델들과 비교