Dacon 14회
금융문자 분석
경진대회

DA501

목차

1. 데이터 설명

KB금융그룹 및 한국인터넷진흥원에서 제공한 정상&스미싱 문자 학습데이터 295,945개와 테스트데이터 1,626개

	id	year_month	text	smishing
0	0	2017-01	XXX은행성산XXX팀장입니다.행복한주말되세요	0
1	1	2017-01	오늘도많이웃으시는하루시작하세요XXX은행 진월동VIP라운지 XXX올림	0
2	2	2017-01	안녕하십니까 고객님. XXX은행입니다.금일 납부하셔야 할 금액은 153600원 입니...	0
3	4	2017-01	XXX 고객님안녕하세요XXX은행 XXX지점입니다지난 한 해 동안 저희 XXX지점에 ...	0
4	5	2017-01	1월은 새로움이 가득XXX입니다.올 한해 더 많이행복한 한해되시길바랍니다	0

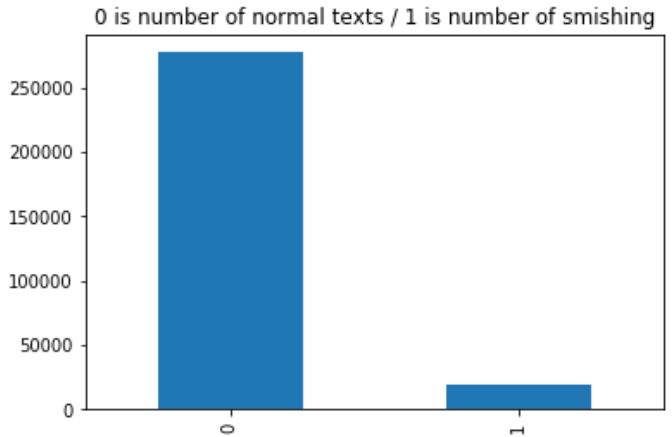
year_month
- 고객이 문자를 전송 받은 연도와 월
- 2017년 1월부터 2018년 12월까지 총 24개월의 데이터 존재

text
- 고객이 전송 받은 문자의 내용
- 개인정보 보호를 위해 은행명과 고객명 등은 XXX로 처리되어 있음

smishing
- 해당 문자의 스미싱 여부
- 0: 스미싱 아님
- 1: 스미싱

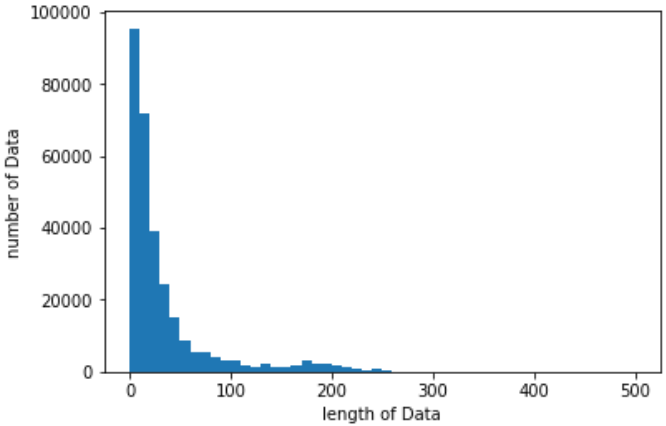
2. EDA - 문자 개수 분포

Train 데이터의 문자 분포



정상 문자에 비해 스미싱 문자의 개수가 현저히 적음

토큰화된 문자의 길이별 개수

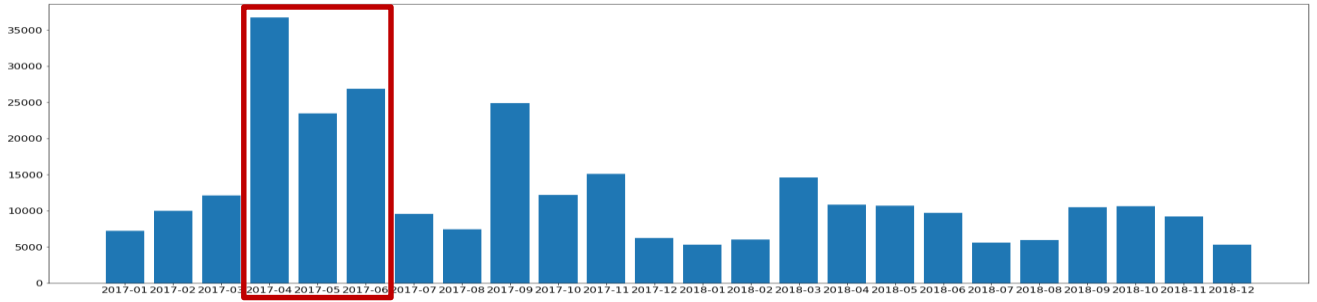


문자의 최대 길이: 499, 문자의 평균 길이: 34.04

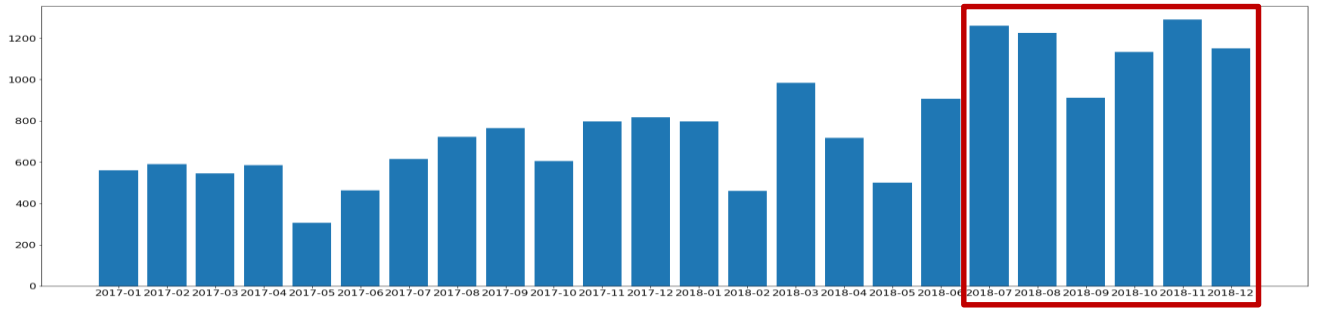
2. EDA - 시계열 분포

전체 문자의 월별 분포

전체 문자의 수는 2017년 4월~6월에 증가하는 반면,
스미싱 문자의 수는 2018년에 더 증가하는 경향이 있음.



스미싱 문자의 월별 분포



2. EDA - 워드클라우드

스미싱 문자와 정상 문자에서 쓰이는 단어의 차이를 분석하기 위해 빈도 수 및 TF-IDF 기준으로 워드클라우드를 그림

정상 문자

[빈도 수 기반]

[TF-IDF 기반]

- '은행'에 직접 방문하라는 '내점', '고객'을 '전담'하는 직원들의 '감사' 문자가 많았음
- 예상 외로, 사이트 URL이나 어플 가입을 유도하는 링크가 일반 문자에서 보다 더 나타남
- 서비스 및 상품 가입 유도보다 고객 관리 의도로 세법개정에 관한 내용인 '공제', '세액'을 공지하는 정보성 문자가 많음

스미싱 문자

[빈도 수 기반]

[TF-IDF 기반]

- '금리', '한도', '대출', '등급'과 같이 서비스 및 상품 설명과 관련된 단어가 스미싱 문자에 자주 나타남
- 스미싱 문자가 타겟팅하는 저신용자를 유혹하는 단어가 많이 나타남('부채', '자체등급', '채무통합', '조회기록', '전환신용등급' 등)
- '문자예약', '신청방법문자', '오픈채팅', '문자상담' 등 비공식적인 통로로 소통하려는 시도가 보임

3. 데이터 전처리

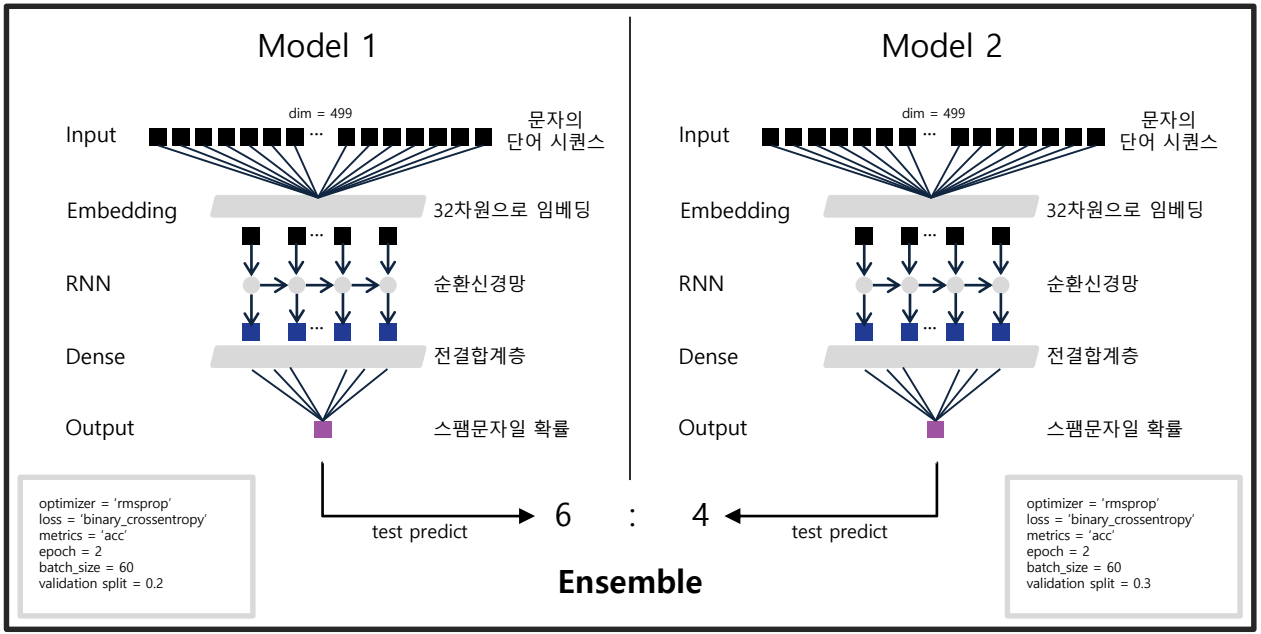
DACON
DATA TO VALUE

데이터 전처리 과정

1. Komoran 형태소 분석기를 사용하여 train/test data의 명사만 추출
2. 형태소 분석된 train/test data를 행 기준으로 concat
3. Concat된 결과의 result(문자) 컬럼을 X_data로, smishing 컬럼을 y_data로 저장
4. Tokenizer를 사용하여 297571개의 row를 가진 X_data의 각 행에 토큰화 수행
5. texts_to_sequences 함수를 사용하여 토큰화된 각 행을 인덱스로 변환하여 저장
6. {단어:인덱스} 딕셔너리를 word_to_index로 저장
7. (word_to_index의 길이+1)을 vocab_size로 저장
8. pad_sequences 함수를 사용하여 문자 최대 길이 499로 각 데이터의 길이를 패딩
9. Concat된 train/test data를 슬라이싱하여 X_train, y_train, X_test, y_test로 나눔

4. 모델링

- Sequence 모델인 **RNN**(Recurrent Neural Network)을 활용
- 두 모델을 **앙상블**한 구조



5. 결과

‘RNN 앙상블’ 모델의 장점

1. LSTM이나 다른 pre-training model(ELMO, BERT 등)보다 빠른 학습 및 예측
   - 알고리즘을 현업에서 사용하기 위해서는 변화하는 스미싱 추세를 빠르게 학습하여 예측할 수 있어야함

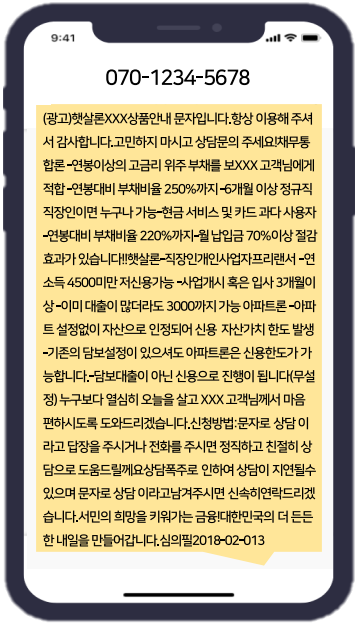
2. 앙상블을 통한 예측의 안정성 및 정확성 향상
   - 단일 모델보다 예측의 안정성과 정확성을 높이기 위해 2개의 RNN 모델을 앙상블하여 사용함

기업 측면 기대효과	개인 측면 기대효과
금융기관으로부터 전송된 메시지에 대한 안전성 확보, 스팸문자 식별 및 스미싱 예방에도 큰 기여	본인도 모르는 결제, 금융 정보 유출 등의 심각한 피해 예방

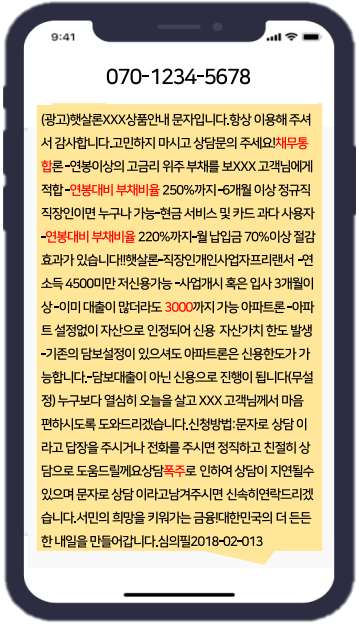
“ 스미싱 문자 탐지는 기업과 소비자로부터 안전한 금융 활동을 보장함 ”

5. 결과 - 응용예시

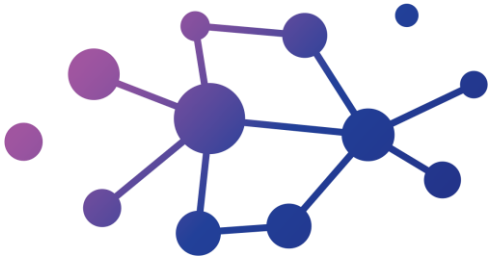
1. 스미싱 의심 문자 수신



2. 스미싱 관련 키워드 탐지



3. 스미싱일 확률 및 이유 안내



THANK YOU