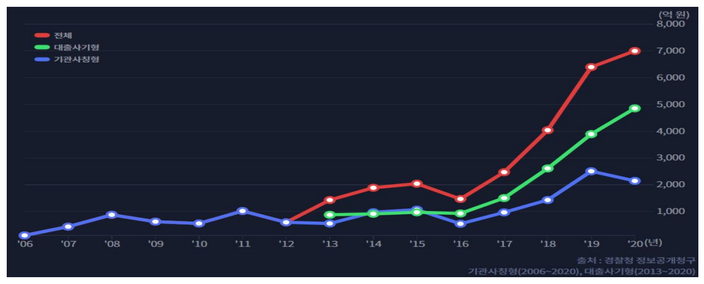
속성 추출 기반 스미싱 탐지 앱 개발

2022년 가을학기
캡스톤축제
[캡스톤 디자인]

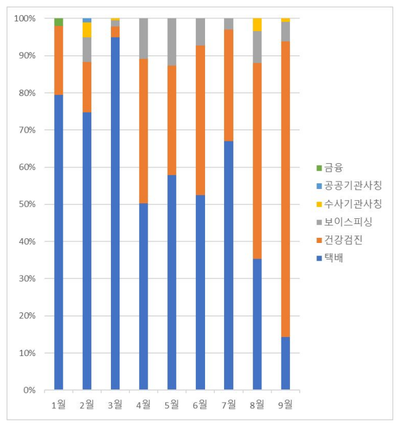
참여기업체: AI 굿윌보이스 **지도교수님**: 남재창 **팀원**: 유이새, 정지원, 김진일

1. 필요성 및 문제 정의

- **필요성**
  - 금융사기는 매년 증가하고 있는 추세
  - 상황과 시기에 따른 **스미싱 키워드** 변화



출처 : 경찰청, 금융감독원 정보공개청구



출처 : 이스트시큐리티 2022 월별 스미싱 키워드

- **문제 정의**

Problem Statement

- ABAE(Attention Based Aspect Extraction)을 이용한 스미싱(SMS, MMS 피싱)탐지 어플 개발

Constraints

- 안드로이드 기반 어플
- 데이터가 개인정보를 포함하고 있으므로 수집이 어려움

Objectives

- ABAE를 이용한 스미싱 키워드 추출
- 문자 텍스트, 전화번호, URL, 키워드 등을 기준으로 피싱 여부를 판단
- 고령층에서도 쉽게 사용할 수 있는 UI

Main Functions

- 피싱 문자 탐지 및 유형 분석
  - 전체 문자 분석
  - 실시간 문자 분석: 문자가 오면 Broadcast Receiver에서 인텐트를 수신한다.
  - ABAE로 추출된 키워드들과 스미싱 관련 특징들에 의해 데이터를 가공한다. 이후 머신러닝 모델을 이용해 문자의 스미싱 확률을 계산한다.
- 그래프를 이용한 시각화
- 휴대폰 위험도 분석

2. 기존 연구/제품 비교 분석

- **Existing Methods**
  - He, Ruidan, et al. (2019)은 속성 용어를 추출하는 비지도 학습 ABAE 모델을 제시하였고, LDA의 단점을 보완해 높은 Coherence를 보여주었다.
  - Mishra, S. and Devpriya S. (2021)은 URL의 도메인 비교 및 SMS 분류를 통해 스미싱 탐지 모델을 구현했다.

- **Differences**

	피싱캣S	라인 백신	닥터스미싱	스마트피싱보호
대표 어플	피싱캣S	라인 백신	닥터스미싱	피싱제로
차별점	인공지능(ABAE모델)을 탑재, **(선제적 대응 가능 모델)** **위험문자단계별 경고 시스템** **분석 · 예측**이 뛰어나 사용성 높음 **선제적 대응**	스미싱 범죄에 대해 뒤늦은 후행성 대처		스미싱 예측율이 낮아 사용성이 떨어짐
	축적된 스미싱 기록과 비교 분석하기 때문에 변형된 유형의 범죄에 취약			

- **Advantages**
  - 키워드 변화에 민감하게 반응하여 탐지 정확도를 높인다.
  - 시기별로 자주 등장하는 유형에 맞춤으로 탐지가 가능하다.

3. 핵심 내용 요약

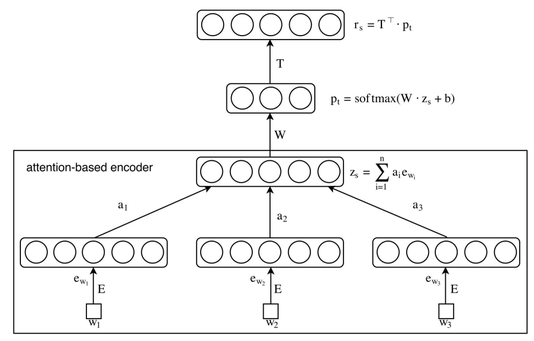
- **Key Technologies**

Smishing Aspect Extraction

1. 문장 임베딩
   각 단어를 Word2Vec를 통해 d차원으로 변환한 후 non-aspect 단어를 걸러내고, Attention 가중치를 계산하여 문장 임베딩 z_s를 구함
2. 속성 임베딩
   문장 임베딩 단계에서 구한 결과를 이용해 Softmax 함수를 사용해서 속성에 대한 확률을 계산 후 r_s로 재구성
3. 정규화 및 훈련 단계
   비용함수 최소화 하는 방향으로 훈련을 진행

Features

1. 전화번호 존재 유무:
   F_1= {1: 전화번호 포함, 0: 포함 X}
2. URL 존재 유무:
   F_2= {1:URL 포함, 0: 포함 X}
3. 메시지 길이
   F_3= {1: 200자 초과, 0: 200자 이하}
4. 수학 기호 (+, %, -, /, ^)
   F_4: {1:수학 기호 포함, 0: 포함 X}
5. 스미싱 기호 ($, dollar, ₩, £, pound, money)
   F_5= {1: 스미싱 기호 포함, 0: 포함 X}
6. 스미싱 주제어:
   F_6~F_{20}= {1: 주제별 스미싱 단어 포함, 0: 포함 X}

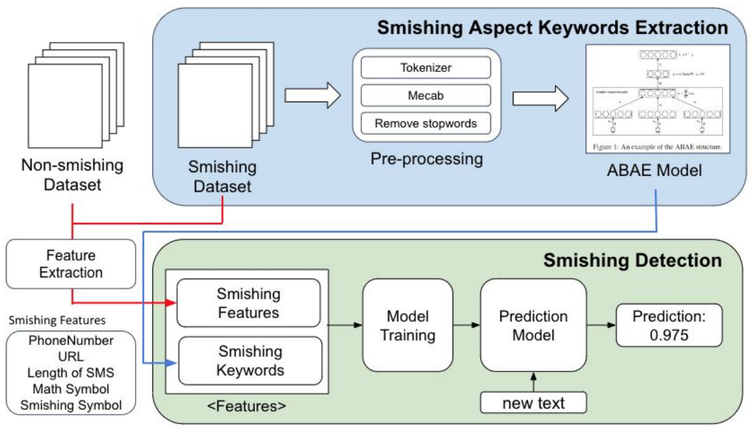


ABAE (Attention Based Aspect Extraction) 모델

출처: He, Ruidan, et al. "An unsupervised neural attention model for aspect extraction.". 2017.

속성 (Inferred Aspect)	속성 용어 (Aspect Keywords)
경품	top, gift, claim, text, welcome, chat, week, txt, info, hear
광고	mobile, year, send, unsub, girl, top, contact, service, msg, dont
광고	min, txt, pls, text, digital, top, box, dvd, claim, first
경품	hear, mob, collection, text, cost, gift, box, top, show, del
우승	summer, pobox, box, rply, txt, girl, message, time, award, dvd
우승	award, hear, win, network, welcome, important, club, babe, end, box
우승	top, text, like, box, award, welcome, txt, important, summer, double
광고	part, help, info, digital, please, pobox, want, real, access, important
광고	like, top, claim, contact, girl, mobile, may, first, love, time
우승	want, win, top, text, call, time, babe, week, send, cost
만남	real, girl, next, help, voucher, redeem, text, first, top, mobile
광고	tone, important, access, mob, txt, half, flirt, real, click, babe
행사	msg, like, txt, pls, top, contact, send, buy, text, discount
광고	chat, landline, half, top, send, msg, network, welcome, del, news
경품	pobox, first, help, text, private, girl, choose, please, gift, network

- **System Design**

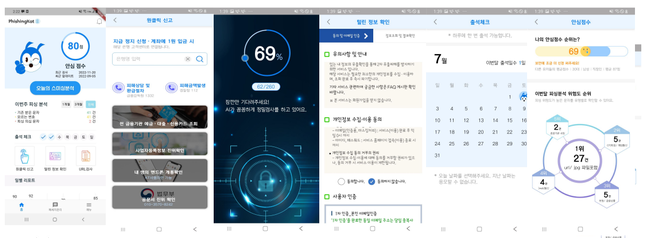


- **Specifications**

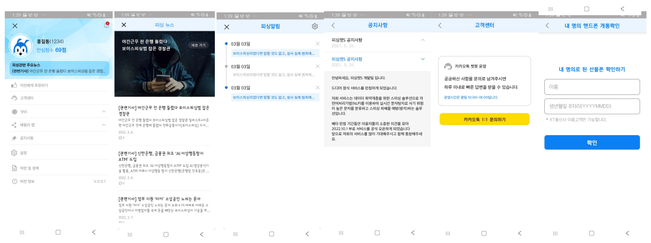
android Flutter Dart
Java PyTorch python

4. 실험 결과/평가

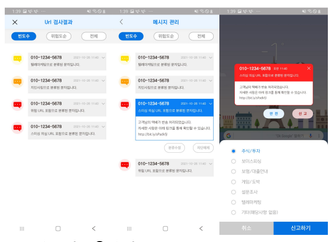
- **Implementations**
  - 스미싱 예방 및 탐지 어플 제작
    - Home



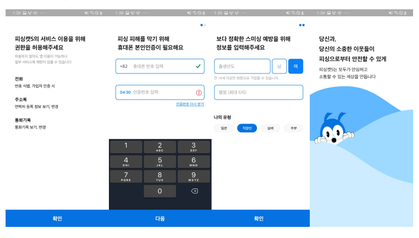
    - Menu



    - Smishing Message Management



    - Login & Join



- **Evaluation**

model 비교		Random Forest	Logistic	Neural Network	Decision Tree	SVM
Precision	ABAE	0.952	0.951	0.924	0.950	0.963
	Non ABAE	0.966	**0.968**	0.939	0.966	**0.968**
	P-value	0.002	0.002	0.006	0.002	0.002
Recall	ABAE	0.832	0.843	0.856	0.847	**0.859**
	Non ABAE	0.700	0.700	0.684	0.700	0.699
	P-value	0.002	0.002	0.002	0.002	0.002
F1-Score	ABAE	0.888	0.893	0.879	0.894	**0.908**
	Non ABAE	0.810	0.811	0.775	0.810	0.810
	P-value	0.002	0.002	0.002	0.002	0.002
AUC	ABAE	**0.982**	**0.982**	0.949	0.961	0.949
	Non ABAE	0.918	0.917	0.917	0.918	0.915
	P-value	0.002	0.002	0.002	0.002	0.002
MCC	ABAE	0.876	0.881	0.890	0.882	**0.898**
	Non ABAE	0.801	0.803	0.801	0.801	0.802
	P-value	0.002	0.002	0.002	0.002	0.002

ABAE 기반 스미싱 탐지 모델과 Non-ABAE 기반 모델에 대한 Wilcoxon Signed-Rank Test 결과

- 정밀도를 제외한 재현율, F1-Score, AUC, MCC 지표들은 ABAE를 사용한 모델이 그렇지 않은 모델보다 성능이 향상되었다.
- 스미싱에서 자주 등장하는 단어는 스미싱 탐지 모델 성능 향상에 유의미한 영향을 미친다는 것을 관찰했다.
- 정밀도는 다소 감소했지만, 재현율의 성능 향상 폭이 커, 속성 용어를 갖고 있는 문장은 스미싱 관련 문자일 가능성이 높다는 결과를 보여 준다.

- **성과**
  - 한국소프트웨어종합학술대회 (KSC 2022) 논문 게재 결정
  - 한동 창업경진대회 (RPM) 동상