속성 추출 기반 스미싱 탐지 모델 개발

김진일○ 정지원 유이새 남재창
한동대학교 전산전자공학부
jiniljeil1@gmail.com, idjjm95@gmail.com, yooisae02@gmail.com, jcnam@handong.edu

Smishing Detection Using Attention-based Aspect Extraction

Jinil Kim○, Jiwon Jung, Isae Yoo, Jaechang Nam
School of Computer Science and Electrical Engineering, Handong Global University

요 약

본 연구에서는 스미싱 문자 탐지 모델의 성능을 높이기 위해 속성 용어(Aspect keyword)를 추출하는 ABAE(Attention-based Aspect Extraction) 기법을 적용하였다. ABAE 모델을 통해 스미싱과 관련된 속성 용어를 추출하고 이를 특징(Feature)으로 넣어 5가지 기계학습 모델을 훈련하여 스미싱 문자 탐지 결과를 확인한다. 이 논문에서는 ABAE 모델에서 추출한 속성 용어를 특징으로 사용한 스미싱 탐지 모델과 그렇지 않은 Non-ABAE 스미싱 탐지 모델 간의 유의미한 차이가 있는지 확인하고자 했다. 그 결과 탐지 모델에서 정밀도(Precision)가 근소하게 줄었고 재현율(Recall), F1-Score, AUC, MCC가 향상됐다. ABAE과 Non-ABAE 스미싱 탐지 모델 간 성능 변화가 없다는 것을 귀무가설로 Wilcoxon Signed-Rank Test를 진행했다. 그 결과 ABAE를 활용한 스미싱 탐지 모델이 정밀도를 제외한 지표에서 통계적으로 유의미한 성능 향상을 관찰했다.

1. 서 론

스미싱(Smishing)[1]은 “문자메시지와 피싱의 합성어로 악성 앱 주소가 포함된 휴대폰 문자(SMS)를 대량으로 전송 후 이용자가 악성 앱을 설치하도록 유도하여 금융정보 등을 탈취하는 신종 사기수법”을 칭한다. 스미싱의 대표 유형[1]은 지인 사칭, 택배 사칭, 공공기관 사칭, 사회적 이슈, 기타유형으로 총 6가지 종류로 분류된다.

기존 연구[2]는 스미싱 문자를 분석할 때 단어의 빈도수를 기준으로 한다는 한계점이 있다. 이는 중요하지만 자주 등장하지 않는 단어를 고려하지 않는 경향이 있다. ABAE[3]는 단어를 문맥에 따라서 감지하여 검토하기에 말뭉치에 존재하는 데이터 희소성 문제를 극복한다.

본 연구에서는 ABAE(Attention-based Aspect Extraction)를 통해 속성을 추출함으로써 스미싱 탐지 모델을 향상시키는 것을 목적으로 한다. 사람들이 스미싱을 당하는 이유는 실제 문자 내용과 매우 흡사하다. 또한 친구 사칭 등의 유형은 심신미약 상태로 유도해 정상적인 사고를 못하게 한다. 이를 위해서 스미싱 모델에 대해서 경각심을 불러일으키는 모델이 필요하다. 따라서 본 논문에서는 속성 추출을 이용해 정밀도 차이가 적으면서 재현율이 높은 스미싱 모델을 구현하고자 한다.

2. 관련연구

Jain, A. K., & Gupta, B. B.는 스미싱 문자의 특징 기반으로 스미싱 탐지 모델 연구[2]를 했다. 전화번호, URL, 이메일 등 여러 특징을 기반하여 스미싱을 탐지하는 모델을 구현했다.

He, Ruidan, et al.은 ABAE 모델 연구[3]를 했다. 문장에서 주제어의 역할을 하는 속성 용어를 추출하는 비지도 학습 모델을 제시하였다. ABAE는 기존의 비지도 학습을 통해 속성을 추출한 LDA의 단점을 보완해 높은 Coherence를 보여주었다.

박명현, 최회련, 이홍철은 육아용품 관련 리뷰들의 속성 용어를 추출한 연구[4]를 했다. ABAE 모델을 통해 얻은 속성으로 리뷰를 분류했고 다른 모델보다 ABAE가 좋은 성능을 갖는다는 것을 입증했다.

Mishra, S. and Devpriya S.은 URL의 도메인 비교 및 SMS 분류를 통해 스미싱 탐지 연구[5]를 했다. URL의 도메인 추출 및 키워드로 Google 검색 상위 5개 도메인과 비교해 특징으로 사용하여 스미싱 탐지 모델을 구현했다.

3. 접근 방법

본 연구는 스미싱 모델의 성능 향상을 위해 스미싱에 주로 등장하는 속성 용어를 추출해 스미싱 탐지 모델에 특징으로 활용하는 기법을 제안한다. ABAE 모델은 크게 문장 임베딩, 속성 임베딩, 정규화 및 훈련으로 나누어져 있다.

문장 임베딩 단계에서는 각 단어를 Word2Vec Skip-gram 방식를 통해 e_{w_i}는 d차원 벡터로 변환한다. 이때 문장에서 관계없는 단어들은 (1) 연산을 통해서 제거된다. 계산된 d_i와 계산식 (3)을 통해 Attention 가중치 a_i를

계산한 후 문장에 대한 임베딩 z_s를 구한다.

속성 임베딩 단계에서는 문장 임베딩 단계에서 구한 결과를 이용해 (5) 과정을 통해 $Softmax$ 함수를 사용해서 속성에 대한 확률을 구한다. 이를 (6) 과정을 거치면서 z_s를 r_s로 재구성한다.

정규화 및 훈련 단계에서는 재구성 오류를 줄이기 위해 (7)식을 목표함수로 두고 학습을 진행한다. 이를 위해 부정 표본(negative sampling)으로 m개를 선택한다. 속성 중복을 피하기 위해서 (8) 을 통해 정규화를 진행한다. 이를 (7)식과 더해 최종 비용함수 $L(\theta)$를 구하고 $L(\theta)$를 최소화 하는 방향으로 훈련시킨다.

(1) $d_i = e_{w_i}^T \bullet M \bullet y_s$ (2) $y_s = \frac{1}{n}\sum_{i=1}^{n} e_{w_i}$

(3) $a_i = \frac{exp(d_i)}{\sum_{j=1}^{n} exp(d_j)}$ (4) $z_s = \sum_{i=1}^{n} a_i e_{w_i}$

(5) $p_t = softmax(W \bullet z_s + b)$ (6) $r_s = T^T \bullet p_t$

(7) $J(\theta) = \sum_{s \in D}\sum_{i=1}^{m} max(0, 1 - r_s z_s + r_s n_i)$

(8) $U(\theta) = \left\| T_n \cdot T_n^T - I \right\|$, I는 항등행렬

(9) $L(\theta) = J(\theta) + \lambda U(\theta)$

$E \in \mathbb{R}^{V \times d}$, $T, W \in \mathbb{R}^{K \times d}$, $M \in \mathbb{R}^{d \times d}$, $e_w \in \mathbb{R}^d$, K : 속성 개수

아래 그림1은 He et al(2017)에서 제시한 세 단어가 한 문장일 때 ABAE의 흐름도를 나타냈다.

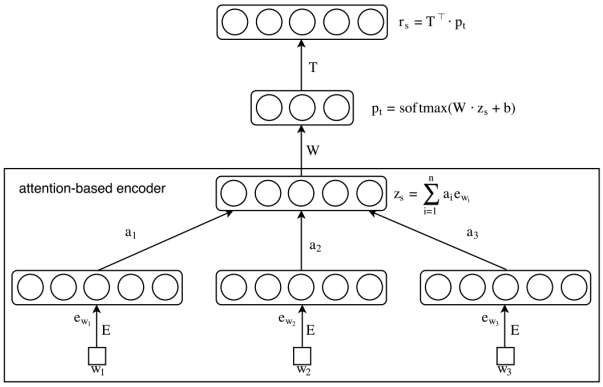

그림 1. An Example of ABAE Structure(He et al., 2017)

4. 실험 설계 및 결과

4.1 ABAE 파라미터 설정

표 1. ABAE Model Hyperparameters

Hyperparameters	Value
K	15
Regularization(λ)	1.0
Number of epochs	200
Optimizer	Adam
learning rate	0.01

ABAE 모델에서 사용된 파라미터는 표1과 같고, 부정 표본(negative sample) 5개를 사용했다.

4.2 실험 데이터 및 전처리 (Preprocessing)

실험 데이터[6]는 영어 문자 데이터로 구성되어있다. 총 문자 개수는 5574개로 일반 문자는 4827개, 스미싱 문자는 747개이다.

ABAE 모델의 경우, 스미싱 대표 단어를 추출하기 위해 스미싱 문자 데이터만을 사용했다. 스미싱 문자는 분기별로 유행하는 경향이 있기에 중복을 제외한 스미싱 문자 653개를 모두 사용했다. 전화번호, URL은 문자 내에서 정규 표현식을 통해 제거했다. 특수 문자와 2글자 이하인 단어 또한 제거했다. Word Tokenizer로 토큰화를 진행했고, 정규화를 위해 소문자로 변환 후 Lemmatizer를 사용하여 표제어로 변환했다. NLTK를 사용하여 영어 불용어를 제거했다. 수학 기호, 스미싱 상징 기호는 스미싱 탐지 모델에서 특징로 활용되어 제거했다.

스미싱 탐지 모델의 경우, 전체 데이터 셋에 중복되는 문장을 제거하여 총 5171개 데이터를 사용했다. 전체 데이터 중 일반 문자는 4518개, 스미싱 문자는 653개로 구성되어있다. 스미싱 탐지 모델은 ABAE 모델과 유사하게 전처리를 진행했다.

4.3 특징 (Feature)

4.3.1. 전화번호
F_1= {1: 문자 내에 전화번호 포함, 0: 포함 X}

4.3.2. URL 존재 유무
F_2= {1: 문자 내 URL 포함, 0: 포함 X}

4.3.3. 문자 길이
F_3= {1: 200 자 초과, 0: 200자 이하}

4.3.4. 수학 기호 (+, %, −, /, ^)
F_4= {1: 수학 기호 포함, 0: 포함 X}

4.3.5. 스미싱 상징 기호
(\$, dollar, ₩, £, pound, money)
F_5= {1: 스미싱 상징 기호 포함, 0: 포함 X}

4.3.6. 주제어
$F_{6\sim20}$= {1: 문장 내 속성 용어 포함, 0: 포함 X}

4.4. ABAE 모델로부터 추출된 주제어

K를 증가시키며 실험 결과간 속성 변화가 가장 적은 결과를 선별했다. 이를 통해 K를 15로 정했다. 스미싱 문자의 이상치를 제외한 평균 단어수는 10.54개 이므로 코사인 유사도 함수를 통해 상위 10개의 속성 용어를 추출했다.

위 과정을 거쳐 나온 속성 용어를 통해서 속성을 경품, 광고, 우승, 행사, 만남으로 분류했다. 표2는 그 결과이다.

표 2. ABAE 결과

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

4.5 실험 결과

표 3. ABAE을 통한 모델 결과

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

표3의 구현 모델로 Random Forest, Logistic Regression, Neural Network, Decision Tree, SVM(Support Vector Machine)를 사용했고 위와 같은 결과를 얻었다. 표3은 10-fold-cross-validation을 10번 진행한 결과의 평균 값이다. 각 평가 지표마다 가장 우수한 값을 볼드처리하였다. 또한 각 지표마다 ABAE 기반 스미싱 탐지 모델과 Non-ABAE 기반 모델에 대해 Wilcoxon Signed-Rank Test을 통해 p-value (0.05)를 나타냈다. 가장 우수한 결과들에 대해 통계적으로 유의미한 차이를 보이는 경우 밑줄로 표시했다.

표3에서 ABAE를 사용한 모델이 그렇지 않은 모델보다 정밀도가 낮아졌다. 하지만, 정밀도를 제외한 재현율, F1-Score, AUC, MCC 지표들은 ABAE를 사용한 모델이 그렇지 않은 모델보다 성능이 좋아진 것을 볼 수 있다.

5. 결론 및 향후 연구

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. 표3에 따르면 정밀도를 제외한 모든 지표가 증가한 것을 볼 수 있다. 이 실험의 목표했던 정밀도의 차이가 가장 적으면서 재현율에서 좋은 성능을 보인 모델은 SVM이며 재현율은 16.0% 증가한 85.9%이다.

한국어 데이터의 제약으로 영어 데이터를 사용했지만, 한국어 데이터를 수집한다면 이를 한국어 특징에 맞게 한국어 데이터를 통해서 구현이 필요하다. 또한 현재 적은 특징을 사용해서 구현했지만 더욱 많은 특징들을 사용한다면 정밀도와 재현율이 크게 향상될 것으로 보인다.

본 연구는 스미싱 키워드가 스미싱 탐지에 미치는 영향을 파악하기 위한 예비 연구이기에 정확한 실험을 위해서는 더 많은 데이터 및 정교한 실험이 필요하다. 스미싱 탐지 모델에서 사용되는 스미싱 문자를 키워드 추출에 모두 사용했다. 향후 연구에서 추가 실험을 통해 더 많은 데이터를 확보하여 ABAE 추출을 위한 데이터와 스미싱 탐지를 위한 데이터를 구분하여 진행할 계획이다.

6. 참고문헌

[1] “스미싱”, KISA 인터넷 보호나라, 2022년 10월 15일 접속, https://www.boho.or.kr/cyber/smishing.do

[2] Jain, A. K., & Gupta, B. B., Feature based approach for detection of smishing messages in the mobile environment. *Journal of Information Technology Research (JITR), 12*(2), 17-35, 2019.

[3] He, Ruidan, et al. "An unsupervised neural attention model for aspect extraction." *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2017.

[4] 박명현, 최회련, 이홍철. 비지도 학습 기반의 한국어 속성 추출에 적합한 전처리 방법 연구: 육아용품 상품평을 대상으로. 대한산업공학회지, 47(1), 56-67, 2021.

[5] Mishra, S. and Devpriya S., “DSmishSMS-A System to Detect Smishing SMS.” *Neural Computing & Applications* (2021): 1 - 18.

[6] Almeida, T.A., Gomez Hidalgo, J.M., Yamakami, A. Contributions to the study of SMS Smish Filtering: New Collection and Results. Proceedings of the 2011 ACM Symposium on Document Engineering (ACM DOCENG'11), Mountain View, CA, USA, 2011.