# 내용 기반 보이스피싱 탐지 시스템

손성환*, 김동규**, 강승식*
*국민대학교 컴퓨터공학과
**국민대학교 빅데이터경영통계
e-mail : ssh121@kookmin.ac.kr, kdg9511@naver.com,
sskang@kookmin.ac.kr

# Content based Voice Phishing Detection System

Sung-Hwan Son*, Dong-Kyu Kim**, Seung-Shik Kang*
*Dept of Computer Science, Kookmin University
**Dept of Bigdata Analytics Business Statistics, Kookmin University

**요 약**

실제 통화에서 발생하는 발화를 실시간으로 분석하여, 보이스피싱 여부를 탐지하는 시스템을 구현해보고자 한다. 발화는 다양한 어휘들로 구성되어 있고 해당 어휘들의 빈도 정도, 위치 등의 정보를 토대로 내용을 분석하는 것이 가능하다. 따라서 보이스피싱이 실제 금융 및 대출 상담이나 경찰 수사와 같은 내용을 기반으로 진행되며, 본래의 목적 달성을 위한 시나리오가 존재하는 것에 착안하여, 해당 시나리오를 구성하는 어휘 및 구문(Multi-word expression)의 발생 정도를 기계 학습으로 분석, 탐지하는 방법을 제안한다.

## 1. 서 론

보이스피싱은 법적용어로 '전기통신금융사기'로 분류되며, 스마트폰, 일반전화, PC 등의 통신매체를 이용한 금융사기를 의미한다[1]. 금융감독원의 자료에 따르면, 한국 보이스피싱의 시작은 2006년 6월으로 추정되며, 2018년 상반기까지 총 16만 건 발생하였고, 피해액 규모는 약 1조 5천억원이다. 2019년 대한민국의 과학기술연구개발 예산이 약 2조 8천억원인 것을 감안하면, 이는 엄청난 수치이다. 2015년부터 2017년까지 점차 감소하는 추세를 보이던 보이스피싱은 2018년에 이르러 다시 증가 추세로 돌아섰다. 2017년 상반기 대비 2018년 상반기 피해액은 73.7%(762억원)의 증가율을 보였다. 같은 기간 피해자 수는 56.4%(7,573명)증가율을 나타냈으며, 피해 건수도 40.6%(8,945건)으로 증가했다. 이는 매일 116명의 피해자가 10억원 정도의 피해를 당하고 있음을 의미한다. 전 연령대에서 규모의 차이는 있으나 20-30대에서 약 425억원, 40-50대에서 약 996억원, 60대 이상에서 약 350억원으로 피해가 발생했으며, 유형별로 대출빙자형의 피해금액이 70.7%, 정부기관 등 사칭형 피해금액이 29.3%를 차지했다.

보이스피싱의 수법도 날로 교묘해지고 있다. 다양한 상황에 따른 변수와 보이스피싱 예방 차원 제도를 회피에 대한 시나리오를 작성하고, 다수의 사기범이 한 팀으로 역할을 분담하여 진행한다. 구분하기 어려운 가짜 사이트를 접속하게 하거나, 악성 어플리케이션을 보안용으로 속여 설치하게 하는 경우도 있다.

따라서 매우 정교한 시나리오를 작성해야 하는 사기범들의 입장에서는 다양한 상황에서 나오는 변수를 줄이고, 보이스피싱 예방 차원 제도를 피하기 위한 시나리오와 사기범이 원하는 정보 및 돈을 탈취하기 위한 시나리오 등이 필수불가결하다. 따라서 우리는 해당 시나리오에서 실제 금융권, 정부기관과의 통화에 있어서 존재할 수 없는 대사를 추려내고, 이를 토대로 대상의 의도를 추론하여 해당 통화가 보이스피싱인지 아닌지 탐지하는 시스템을 제안한다.

## 2. 관련 연구

기존의 보이스피싱을 판단하기 위한 방법[2]으로는 Blacklist를 구축하는 방법, GMM(Gaussian mixture model)을 통한 음성 분석 방법이 있다. Blacklist는 말 그대로 실제로 탐지된 보이스피싱 범죄자의 번호 또는 신호 리스트를 분석, 유사도에 따라 탐지하는 방법이다. 실제로 SMS를 매개로 하는 사기 수법에서 악성 어플리케이션을 다운받도록 유도하는 경우 다운받는 어플리케이션에 대해 등록된 어플리케이션인지 검사를 진행하여 등록되지 않은 경우 해당 번호를 리스트에 추가하는 식으로 Blacklist를 구축하는 방법이 있다[3]. GMM을 활용한 방법에는 통화하는 대상의 음성을 분석하여 거짓과 진실을 판단하여 분석, 탐지하는 방법이 존재한다[4].

보이스피싱 이외에도 다양한 피싱 방식이 있는데, SMS나 메일을 매개로 하는 피싱이 대표적이다. 이와 같은 피싱에는 대해서는 Blacklist, Whitelist, Content based와 같은 탐지 방법론으로 나눠질 수 있다. Whitelist는 Blacklist와는 반대로 검증된 목록 이외의 접근에 대하여 차단하는 것을 의미한다. 마지막으로 Content based는 글의 내용을 기반으로 분석하여 사기 여부를 판단하는 방법

이다. Content based를 기반으로 한 방법론으로는 피싱 데이터에서 수집한 어휘를 바탕으로 어휘 사전 및 해당 어휘들로 구축된 묶음 사전을 구축, 사전들과의 매칭 정도를 Fuzzy rule을 기반으로 정의하여 스팸 메일인 확률을 판단하는 방법이 있다[5]. 그리고 마찬가지로 피싱 데이터의 어휘를 가지고 K-means와 Fuzzy support vector machine을 이용, 스팸 메일 여부를 판단 방식이 있다[6].

본 논문에서는 보이스피싱 발화 데이터에서 사기범들이 원하는 방향으로 유도하기 위한 시나리오가 있다는 점에 주목하여, 해당 시나리오를 구성하는 문장, 구문, 어휘를 특성으로 탐지하는 내용 기반 방법론을 제안한다.

## 3. 실험

### 3.1 데이터

보이스피싱 발화 데이터, 양성 데이터(Positive data)는 금감원의 '보이스피싱 지킴이'에서 추출한 신고 사례 데이터이다. 해당 데이터는 크게 '금융권 사칭형', '정부 기관 사칭형'으로 나눠질 수 있으며, 총 315개이다.

학습을 위한 대조 데이터(Negative data)는 개인 정보의 문제로 실제 대출 상담이나 경찰 및 금감원의 통화 발화 데이터를 사용하지 못했다. 따라서 네이버 지식인의 경제 부분의 답변 데이터 36,299개를 추출하였다.

이후 각 텍스트를 형태소 단위로 분할하고, 명사, 동사, 형용사, 부사 토큰만 품사 정보를 추가하였다. 그리고 단편적인 '서울', '중앙'과 같은 어휘보다는 '서울 중앙 지검'과 같이 연결된 어휘가 더 의미있는 경우를 배제하지 않기 위하여 유니그램부터 트라이그램까지의 토큰을 추가했다.

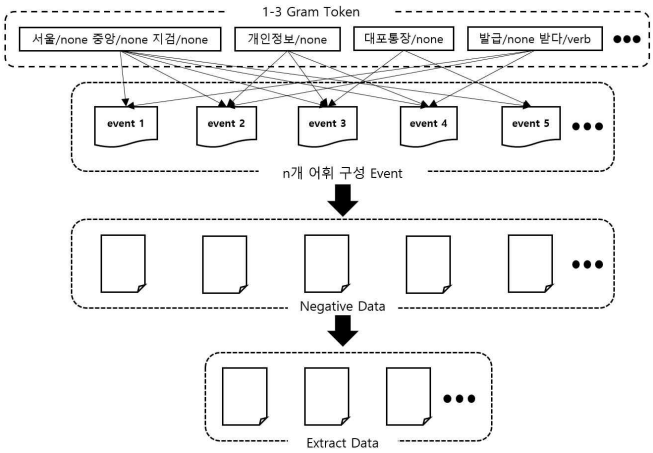


그림 1. Event group을 활용한 유사 데이터 추출 구조

대체적으로 보이스피싱 데이터와 실제 상담 및 수사 내용은 상당 부분 유사하므로, 수집한 대조 데이터에서 양성 데이터와 유사한 데이터를 추려내는 작업을 진행했다. 그림1은 유사 데이터 추출 방식으로 양성 데이터 토큰들의 TF-IDF 상위 어휘 n개의 조합 그룹(Event group)이 음성 데이터에 존재하는 빈도수의 정도로 유사도를 판단하여 추출하는 방법을 가시화한 것이다.

### 3.2 학습

데이터가 부족한 관계로 딥러닝 기법을 사용하기보다 기계 학습 모델을 활용하였다. 특성은 양성 데이터와 대조 데이터에서의 TF-IDF 값의 상위 어휘 500개에 속하는 어휘의 TF-IDF 값을 사용하였고, 상위 어휘 500개에 해당하는 특성 벡터를 사용하여 문서 벡터를 정의하였다.

$$Odds\ ratio = \frac{\Theta}{1-\Theta}$$
$$z = logit(Odds\ ratio) = \log\left(\frac{\Theta}{1-\Theta}\right)$$
$$logistic(z) = \Theta(z) = \frac{1}{1+\exp(-z)} \quad (1)$$

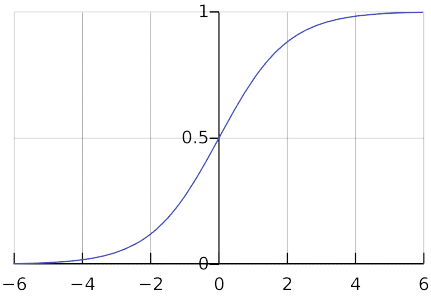


그림 2. 로지스틱 시그모이드 함수 그래프

구축한 문서 벡터를 토대로 기계 학습 모델 중 하나인 로지스틱 회귀 모델(Logistic regression model)을 사용하여 보이스피싱일 확률을 계산, 분류하였다. 로지스틱 회귀는 범주형 데이터를 특정 범주에 속할 확률로 나타낼 수 있도록 구현한 모델로 승산비(Odds ratio)를 기반으로 구현된 수식1을 이용한다. $\Theta$는 1이 될 확률 즉, 보이스피싱일 확률로 볼 수 있고, $\Theta(z)$는 승산비에 로그를 취한 것의 역함수이다. 이에 따른 수식1, 그림2는 로지스틱 함수로 불리며 시그모이드 함수 중 하나이다.

## 4. 실험 결과

이용할 수 있는 데이터가 부족한 관계로 대조 데이터의 개수를 늘려가면서 실험을 진행했다. 유사 대조 데이터를 추출하기 위한 상위 어휘 그룹을 구성하는 어휘 개수를 조절해 가면서 유사도가 높은 데이터 위주로 대조 데이터를 구성했으며, 학습 데이터와 테스트 데이터는 8:2의 비율로 진행하였다.

표 1. 학습데이터 비율에 따른 정확도

| 그룹 어휘 개수/ 대조 데이터 개수 | 정확도 |
|---|---|
| 28/315 | 0.99726 |
| 25/841 | 0.99119 |
| 15/1506 | 0.99178 |

표1과 같이 정확도는 전부 0.99 이상 평가되었지만, 사실상 과적합이나 너무 분류되기 쉬운 작업으로 판단 될

수 있다.

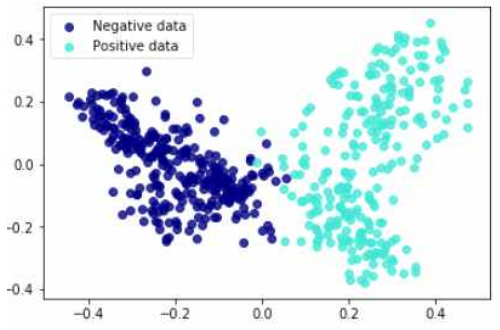


그림 3. 주성분분석을 통한 시각화

그림3은 테스트 데이터의 주성분을 2개로 주성분 분석(PCA)을 진행한 결과이다. 양성 데이터와 대조 데이터가 단순한 선형 결정 경계선으로 분류가 가능한 것을 확인할 수 있다. 따라서 과적합보다 분류되기 쉬운 작업으로 인한 높은 정확도의 결과라고 추측할 수 있다.

## 5. 결론

결과에서 실제 보이스피싱이 아닌 금융 상담 및 수사 발화는 보이스피싱 발화와 매우 유사한 것에 비하여, 사용된 대조 데이터는 보이스피싱 발화 데이터와 유사도가 떨어지는 문제를 확인할 수 있다. 게다가 보이스피싱이 발화 데이터이고 실제 상담 및 수사 발화의 틀을 가지고 있기 때문에 그 자체에 대한 특성도 무시할 수 없다.

결론적으로 더 유사한 대조 데이터를 수집해야 한다. 그리고 정확한 대조군은 보이스피싱의 여집합 즉, 보이스피싱이 아닌 전체 통화 발화이기 때문에 양성 데이터에서만 특성을 추출하여 문서 벡터를 만드는 것이 일반화에 더 좋을 것으로 판단된다. 마지막으로 토큰들을 Word2Vec으로 벡터화시킨 것을 군집화하여 각 군집의 어휘가 얼마나 매칭되는지의 정도를 특성으로 문서 벡터를 구축하고, 군집화를 품사 정보에 따라 개별적으로 진행함으로써 False positive 문제가 발생하는 것을 완화시킴과 동시에 보이스피싱 특정 시나리오의 구문을 탐지하여 분석하는 것이 가능할 것으로 기대한다.

## 참고문헌

[1] 조호대. “보이스피싱 발생 및 대응방안”, 『한국콘텐츠학회논문지』 제12권 제7호, pp.176-182, 2012.

[2] Cik Feresa Mohd Foozy, Rabiah Ahmad, Mohd Faizal Abdollah. “Phishing Detection Taxonomy for Mobile Device”, IJCSI. Vol.10 No.3, pp.338-344, January 2013.

[3] 김정훈, 고준영, 이근호. “빅데이터 기반의 융합 보이스피싱을 이용한 사회공학적 공격 기법과 대응방안”, 『한국융합학회논문지』 제6권 제1호, pp.85-91, 2015.

[4] J. H. Chang, K. H. Lee. “Voice Phishing Detection Technique based on Minimum Classification Error Method Incorporating Codec Parameters”, IET Signal Process. Vol.4 No.5, pp.502-509, 2010.

[5] Shengnan Wang, Xiaoyong Zhang, Yijun Cheng, Fu Jiang, Wentao Yu, Jun Peng. “A Fast Content-based Spam Filtering Algorithm with Fuzzy-SVM and K-means”, IEEE International Conference on Big Data and Smart Computing. January, pp.301-307, 2018.

[6] Zhiqiang Zhang, Aihua Zhang. “A Novel Strategy for Fault Diagnosis of Analog Circuit Online based Modified Kernel Fuzzy C-means”, IEEE International Conference on Industrial Technology, pp.938 - 943, March 2016.