딥러닝과 감성 분석에 따른 보이스피싱 여부 판별

김원웅*, 강예준**, 김현지***, 양유진*, 오유진*, 이민우*, 임세진**, 서화정***†
*한성대학교 IT융합공학부 (학부생)
**한성대학교 컴퓨터공학부 (학부생)
***한성대학교 IT융합공학부 (대학원생)
***†한성대학교 IT융합공학부 (교수)
dnjsdndeee@gmail.com, etus1211@gmail.com, khj1594012@gmail.com,
yujin.yang34@gmail.com, oyj0922@gmail.com, minunejip@gmail.com,
dlatpwls834@gmail.com, hwajeong84@gmail.com

Determination of voice phishing based on deep learning and sentiment analysis

Won-Woong Kim*, Yea-Jun Kang**, Hyun-Ji Kim***, Yu-Jin Yang*, Yu-Jin Oh*, Min-Woo Lee*, Se-Jin Lim**, Hwa-Jeong Seo***†
*Dept. of IT Convergence Engineering, Han-Sung University (Ungraduated)
**Dept. of Computer Science, Han-Sung University (Ungraduated)
***Dept. of IT Convergence Engineering, Han-Sung University (graduated)
***†Dept. of IT Convergence Engineering, Han-Sung University (Professor)

요 약

본 논문에서는 점차 진화되어가는 보이스피싱 수법에 대하여 딥러닝 기반 네트워크인 DNN(Deep Neural Network)를 통한 보이스피싱 여부 판별할 뿐만 아니라, CNN, Bi-LSTM을 활용한 다양한 관점에서의 감성 분석을 통하여 보이스피싱 조직원의 감성 상태를 파악하여 판별된 결과에 신뢰도를 높여주는 모델을 제안하였다.

1. 서론

최근 보이스피싱 수법이 진화하면서 가족, 지인 또는 공공기관의 관리자 등을 사칭하거나, 정부 긴급 재난 지원금과 같은 사회 현상에 맞춰진 수법들이 나타나고 있다. 또한 초기 보이스 피싱은 특정 지역의 어눌한 말투나 한국의 실정과 맞지 않는 내용 등으로 손쉽게 판별이 가능하였으나 최근의 보이스피싱은 한국인이 참여하거나 전문적인 용어 등이 사용되거나 유출된 개인정보를 바탕으로 범행을 저지름으로써 오히려 2~30대가 가장 큰 피해를 입는 것으로 나타났다[1]. 이와 같이 특정 연령, 상황, 성별에 따라 사기수법에 취약하며 이러한 문제를 개개인이 완벽히 해결하기에는 어려움이 존재한다. 따라서 보이스피싱 여부에 대한 판별을 도울 수 있는 기술이 필요하다.

2015년 7월 28일 서울지방경찰청은 보이스피싱 조직을 검거하며 상황별 범행수법 시나리오 80여 가지를 압수하였다[2]. 따라서 보이스피싱 범죄에는 특정한 패턴이 존재하게 되고, 각 시나리오별 특정 텍스트 및 음성 패턴 또한 존재하게 된다. 본 논문에서는 이를 통하여 딥러닝 모델을 학습시키며, 그 뿐만 아니라 감성에 대한 분석 결과를 추가하여 딥러닝 모델의 일반화 성능을 향상시켜 보다 더 높은 확률로 보이스피싱 여부에 대하여 판별할 수 있는 방법을 제안한다. 또한 감성 분석을 진행할 때에는 텍스트에 대한 감성분석만으로는 오히려 신뢰도가 떨어질 수 있으므로 음성에 대한 감성분석 또한 진행하여 서로의 단점을 보완해주게 되며 결과적으로 모델의 신뢰도를 높이게 된다.

2. 관련 연구

2.1 DNN (Deep Neural Network)

DNN은 신경망 기법 중 하나로써 각 1개의 입력층과 출력층, 그리고 2개 이상의 은닉층을 지닌 신경망을 뜻한다. DNN은 기존의 신경망 기법과 마찬가지로 비선형적인 구조를 가지고 있으며, 각 계층마다 존재하는 뉴런들이 입력값과 출력값을 지니고 있어, 입력값을 이용하여 함수를 계산한 후 그 출력값을 뉴런의 결과값으로 사용한다. 이 때 신경망은

피드-포워드 방식을 지니고 있어 다음 뉴런들은 이전 뉴런들로부터의 가중치를 가지고 있으며 학습을 반복하며 이러한 가중치를 좀 더 나은 방향으로 조정해나가며 모델 스스로 분류 및 회귀 문제들을 해결하기 위하여 사용한다. DNN은 목적에 따라 RNN, CNN으로 분류될 수 있으며 이외에도 LSTM, GRU 등이 있다.

2.1.1 CNN (Convolution Neural Network)

CNN은 딥러닝 모델 중 하나로써 주로 이미지 데이터를 학습하는 데에 사용된다. 이미지 데이터를 학습할 때 Fully Connected Layer를 갖는 MLP(Multi-Layer Perceptron)를 사용한다면 3차원 데이터인 이미지 데이터를 학습할 때 1차원 데이터로 만들어 학습을 하기 때문에 이미지가 갖는 인근 픽셀에 대한 특성을 담지못하며 서로 연관이 없는 픽셀 사이의 값까지 학습을 하게 된다. 또한 신경망이 깊어질 경우 기하급수적으로 증가하는 연산 횟수와 과적합 등과 같은 문제가 발생하여 결과적으로 일반화 성능이 떨어지게 된다. 이때 CNN을 사용하게 된다면, filter라고 하는 가중치 행렬을 움직이며 3차원 이미지의 특성을 지역적으로 추출하므로 이미지 데이터를 학습하는 데에 적합하다.

2.1.2 LSTM (Long Short term Memory)

LSTM은 신경망 기법 중 하나로써 시계열 데이터에 주로 사용되는 신경망인 RNN(Recurrent Neural Network)에 cell state를 추가하며 신경망이 깊어질 경우 과거의 데이터를 잊게되는 기울기 소실 문제를 해결한 모델이다.

LSTM의 Cell state는 Input gate, Forget gate 그리고 Output gate에 의해 가공된 데이터가 이동하게 되는데 이를 이용하여 중요한 데이터와 중요하지 않은 데이터를 판별하여 과거의 중요한 데이터의 정보를 유지하며 보낼 수 있게 된다.

LSTM은 사용되는 목적에 따라서 One-to-many, Many-to-one, Many-to-Many로 나눌 수 있다. One-to-many는 이미지를 보고 이미지에 대하여 설명하는 문장을 생성하는 모델에 사용된다. Many-to-one은 과거의 주가 데이터를 이용하여 주가를 예측하거나 사용되는 텍스트에 따른 감성 분석에 사용된다. Many-to-many는 번역기 등에 사용된다.

2.2 감성 분석

감성 분석이란 텍스트, 이미지 그리고 음성 등과 같은 분석 대상에 들어있는 감성이나 태도 등의 주관적인 정보를 딥러닝 모델을 통하여 분석하는 과정이다. 기존의 연구들은 주로 감정을 happy, fear, disgust, angry, surpirse, neutral, sad와 같이 7가지로 분류하였다[3]. 그리고 이러한 감정들을 크게 긍정/중립/부정으로 나누어 사용한다. 또한 감성 분석을 실제 분야에 접목시키기 위해서는 정량화된 데이터를 습득하는 실시간으로 것이 필수적이고 그에 따라 다양한 기술들이 연구되고 있다[4].

2.2.1 텍스트 감성 분석

텍스트 감성 분석은 텍스트 마이닝 기법을 응용한 기술로, 문장을 특정한 단위로 나눠 미리 구축된 감성 사전에 대입하여 문장에 존재하는 긍정/부정을 판단하는 기술이다. 이때 감성사전의 감성 어휘를 딥러닝의 입력 자질로 사용하게 되었을 때 더 높은 정확도를 내게 된다. 하지만 텍스트 기반의 감성 분석은 같은 텍스트일지라 하더라도 적용되는 도메인에 따라서 긍정과 부정이 달라지는 경우도 있으므로 도메인에 따라 올바른 감성 사전을 적용하는 것이 중요하다. 예를 들어 '슬프다'라는 감정은 일반적으로 부정의 의미를 나타내지만 영화 감상평의 경우에는 긍정의 의미로 사용되기도 한다[5].

2.2.2 음성 감성 분석

일반적으로 음성 신호로부터 사람의 감정을 파악할 수 있는 요소는 톤(Tone), 음성신호의 피치(Pitch), 포만트 주파수(Formant Frequency), 말의 빠르기 등이 있다. 사람은 심리적 변화에 따라 신체적인 변화가 일어나게 되므로 자연스레 톤, 피치, 주파수 등의 변화로 귀결된다[7]. 따라서 본 논문에서는 감정의 변화에 따른 음성의 변화들을 파형과 같은 가시적인 데이터로 변환하여 해당 데이터를 CNN 모델에 입력값으로 사용하여 모델을 학습시킨다.

3. 제안 모델

본 논문에서는 딥러닝 및 각종 감성 분석을 이용한 보이스피싱 여부를 판별하는 모델을 제안한다.

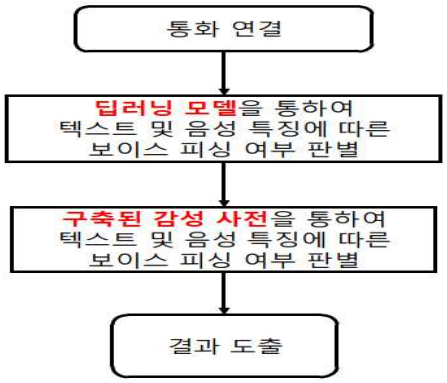

그림 1 보이스피싱 여부 판별 흐름도

그림 1은 전체적인 흐름을 도식화 한 것이다. 먼저 보이스피싱 여부에 대하여 영향을 끼칠 수 있는 요소를 크게 3가지로 분류하였다. 먼저 첫 번째로 보이스피싱 과정에서 빈번히 나타나는 특정 단어가 존재한다는 것이다. 주로 공공기관의 일원을 사칭하며 금융사기에 연루되었다는 내용의 통화를 진행하거나, 받을 수 있는 혜택이 존재한다며 개인정보유출을 유도하는 식의 언어가 사용된다[6]. 두 번째로 보이스피싱을 하는 과정에서 특정한 말의 높낮이(pitch), 톤(tone), 데시벨 등이 존재한다. 보이스피싱과 일반적인 통화의 가장 큰 차이점은, 음성의 높낮이에 대한 편차와 발화 속도이다. 주로 보이스피싱에서 더 높은 발화 속도와 음성의 높낮이가 나타났다. 세 번째로, 통화를 진행하는 동안의 보이스피싱 조직원의 감성 상태이다. 감성 상태를 파악하는 방법으로는 크게 텍스트와 음성으로 분류할 수 있다. 텍스트와 음성에 따라서 상대를 압박하거나 당황을 했을 때 또는 분노 같은 특정 감성 상태를 분류할 수 있다. 예를 들어 당황을 했을 때 말을 더듬는다거나 음성의 pitch가 높아지는 등의 특징이 나타난다. 이러한 요소들을 통하여 보이스피싱 여부에 대하여 판별한 결과값에 가중치를 부여하게 된다.

우선 일정시간 통화가 진행되면, 진행 동안에 텍스트 및 음성적인 특징을 수집한다. 수집된 데이터를 보이스피싱 시나리오를 이용하여 학습된 딥러닝 모델의 입력값으로 사용하여 해당 통화가 보이스피싱인지에 대한 여부를 판별한다. 그 후, 미리 구축된 감성 사전을 이용하여 수집된 텍스트 및 음성적인 특징에서 나타나는 감성에 대하여 정의한다. 정의된 감성에 따라 딥러닝 모델을 통하여 판별한 보이스피싱 확률에 가중치를 더하거나 감소시킨다.

3.1 Neural Network를 통한 보이스피싱 여부 판별

해당 모델에서는 총 2개의 신경망을 사용한다. 먼저 첫 번째로는 통화 내용에 나타나는 단어를 data로 사용하고 해당 통화의 보이스피싱 여부에 대한 판별을 label로 사용하는 신경망이다. 다음으로 음성의 pitch와 발화 속도를 수치화하여 data로 사용하고 보이스피싱 여부를 label로 사용하는 신경망이다. 우선적으로 이와 같은 신경망을 통하여 보이스피싱 여부에 대하여 확률값으로 나타내게 된다. 첫 번째 신경망을 실제 상황에 적용하게 될 때는 통화가 진행되면서 나타난 단어를 입력값으로 사용하게 된다. 또한 두 번째 신경망은 음성의 pitch와 발화 속도를 입력값으로 사용하게 된다. 두 개의 신경망의 출력층에서는 활성함수로써 sigmoid function을 사용하며 여부에 대한 결과값을 확률적으로 나타내게 된다.

표 1 text 기반 신경망

표 2 pitch, 발화 속도 기반 신경망

3.2.1 Bi-LSTM을 통한 텍스트 감성 분석[5]

특정 도메인에 영향을 받지 않는 범용 한국어 감성 사전을 구축한다. 이는 특정 도메인에 영향을 받지 않는 독립적인 감성 어휘로 구성되어 있다. 이 감성사전은 표준국어대사전에 수록된 뜻풀이를 활용하여 감성 어휘를 추출하고 구축한다. 이때, Bi-LSTM(Bidirectional Long-Short Term Memory)을 통하여 뜻풀이 감성 분류 모델을 구축한다. 다음으로 구축된 감성 분류 모델을 사용하여 뜻풀이를 긍정/부정으로 분류한다. 마지막으로 긍정으로 분류된 뜻풀이에서는 긍정에 관련된 감성 어휘를, 부정으로 분류된 뜻풀이에서는 부정에 관련된 감성 어휘를 추출하여 감성 사전을 구축한다. 예를 들어, '좋아하다'라는 문장이 있을 때, 일반적으로 긍정적인 감성을 나타내지만, 부정적인 감성으로 '남의 어리석은 말이나 행동을 비웃거나 빈정거릴 때 하는 말'이라는 뜻풀이가 있다고 했을 때, 해당 뜻풀이가 부정으로 분류되었을 경우 부정 감성 어휘를 추출하

게 된다. 이때 부정 감성 어휘는 '어리석은', '비웃거나', '빈정거릴 때'가 그 어휘이며 해당 어휘들을 추출하여 감성사전을 구축하게 된다. 이러한 구축된 감성사전을 통하여 보이스피싱 조직원이 당황을 하거나 다급한 상황에서의 감성을 나타내게 된다면 보이스피싱 판별 여부에 대한 확률값에 가중치를 더하게 된다.

3.2.2 CNN을 이용한 음성 감성 분석

일반적으로 음성신호로부터 사람의 감성을 인식할 수 있는 요소는 다양하지만 그 중 일반 통화와 보이스피싱에서 차이점을 관찰할 수 있는 감성에 영향을 끼치는 음성적인 특징은 음성신호의 피치(Pitch), 포만트 주파수(Formant Frequency)가 있다. 본 논문에서는 피치와 포만트 주파수에 의해 형성된 다양한 파형을 data로 사용하며 해당 파형에 대한 감성을 label로 사용하여 학습하게 된다. 추후 감성분석 단계에서 플레이어의 음성 파형의 이미지를 입력값으로 받아 해당 이미지에 cnn을 적용 및 특징을 추출하여 파형의 감성을 분류하는 방법을 제안한다. 통화를 진행하면서 나타난 파형에 대해 CNN 모델을 통한 분석을 진행하여 격양되거나 당황했을 때의 음성 신호의 파형이 나타난다면 이 때도 역시 보이스피싱 여부에 대한 확률값에 가중치를 더해주게 된다.

4. 결론

본 논문에서는 점차 진화되어 가는 보이스피싱에 수법에 맞춰 딥러닝 모델의 기술적인 도움을 받아 보이스피싱에 대한 여부를 판별하는 모델을 제안한다. 또한 해당 기능에 대한 일반화 성능을 높여주기 위하여 음성에 대한 특징을 적용시킨 딥러닝 모델이나 감성 분석을 통하여 보이스피싱 여부에 대한 판별값에 신뢰성을 더해주게 된다. 해당 모델은 한 가지의 감성 분석만으로는 오히려 모델의 일반화 성능을 떨어트릴 가능성이 존재하기 때문에 총 두 가지의 감성 분석을 통하여 서로 간의 취약점을 보완해주는 것에 의의를 두었다. 하지만 감성 사전 특성상 도메인의 영향을 많이 받기에 해당 도메인에 적합한 감성사전을 구현해야 한다. 추후에 해당 사항을 고려하여 시스템의 구체적인 구현을 할 계획이다.

5. Acknowledgment

이 논문은 부분적으로 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(No.2018-0-00264, IoT 융합형 블록체인 플랫폼 보안 원천 기술 연구, 50%) 그리고 부분적으로 2021년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2020R1F1A1048478, 50%).

참고문헌

[1] 이범주, et al. "ICT 기술을 적용한2016 년 5 월 18 일 국가 공개 수배 보이스 피싱의 음성 특징 규명." 한국통신학회 학술대회논문집 (2016): 441-442.\
[2] 금융감독원, 금융감독정보 제2015-30호 (통권 843호), 10쪽.
[3] 이자연, et al. "딥러닝 표정 인식을 활용한 실시간 온라인 강의 이해도 분석." 멀티미디어학회논문지 23.12 (2020): 1464-1475.
[4] 이의철, et al. "상반된 감성에 따른 안면 움직임 차이에 대한 분석." 한국콘텐츠학회논문지 15.10 (2015): 1-9.
[5] 박상민, 나철원, 최민성, 이다희, 온병원."Bi-LSTM 기반의 한국어 감성사전 구축 방안." 지능정보연구24.4(2018):219-240.
[6] 이승아. "보이스피싱에 대한 텍스트언어학적 연구." 텍스트언어학 45 (2018): 177-195.
[7] 박창현, et al. "음성으로부터 감성인식 오소 H 분석." (2001).