JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 28 No. 8, pp. 21-30, August 2023
https://doi.org/10.9708/jksci.2023.28.08.021

# A Study on Deep Learning Model for Discrimination of Illegal Financial Advertisements on the Internet

Kil-Sang Yoo*, Jin-Hee Jang*, Seong-Ju Kim*, Kwang-Yong Gim**

*Ph.d student, Graduate School of IT Policy and Management, Soongsil University, Seoul, Korea
*Ph.d student, Graduate School of IT Policy and Management, Soongsil University, Seoul, Korea
*Ph.d student, Graduate School of IT Policy and Management, Soongsil University, Seoul, Korea
**Professor, Dept. of Business, Soongsil University, Seoul, Korea

## [Abstract]

The study proposes a model that utilizes Python-based deep learning text classification techniques to detect the legality of illegal financial advertising posts on the internet. These posts aim to promote unlawful financial activities, including the trading of bank accounts, credit card fraud, cashing out through mobile payments, and the sale of personal credit information. Despite the efforts of financial regulatory authorities, the prevalence of illegal financial activities persists. By applying this proposed model, the intention is to aid in identifying and detecting illicit content in internet-based illegal financial advertisining, thus contributing to the ongoing efforts to combat such activities. The study utilizes convolutional neural networks(CNN) and recurrent neural networks(RNN, LSTM, GRU), which are commonly used text classification techniques. The raw data for the model is based on manually confirmed regulatory judgments. By adjusting the hyperparameters of the Korean natural language processing and deep learning models, the study has achieved an optimized model with the best performance. This research holds significant meaning as it presents a deep learning model for discerning internet illegal financial advertising, which has not been previously explored. Additionally, with an accuracy range of 91.3% to 93.4% in a deep learning model, there is a hopeful anticipation for the practical application of this model in the task of detecting illicit financial advertisements, ultimately contributing to the eradication of such unlawful financial advertisements.

▸ **Key words**: Illegal Financial Advertisements, Deep Learning, CNN, RNN, LSTM, GRU

• First Author: Kil-Sang Yoo, Co-Author: Jin-Hee Jang, Seong-Ju Kim, Corresponding Author: Kwang-Yong Gim
*Kil-Sang Yoo (rks912@naver.com), Graduate School of IT Policy and Management, Soongsil University
*Jin-Hee Jang (equite1@naver.com), Graduate School of IT Policy and Management, Soongsil University
*Seong-Ju Kim (megnaspirit@naver.com), Graduate School of IT Policy and Management, Soongsil University
**Kwang-Yong Gim (gygim@ssu.ac.kr), Dept. of Business, Soongsil University
• Received: 2023. 07. 11, Revised: 2023. 08. 17, Accepted: 2023. 08. 18.


http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

[요 약]

인터넷 불법금융광고는 인터넷 카페, 블로그 등을 통해 통장매매, 신용카드·휴대폰결제현금화 및 개인신용정보매매 등 불법금융행위를 목적으로 한다. 금융감독당국의 노력에도 불구하고 불법금융행위는 줄어들지 않고 있다. 본 연구는 인터넷 불법금융광고 게시글에 파이썬 딥러닝 기반 텍스트 분류기법을 적용해 불법여부를 탐지하는 모델을 제안한다. 텍스트 분류기법으로 주로 사용되는 합성곱 신경망(CNN: Convolutional Neural Network), 순환 신경망(RNN: Recurrent Neural Network), 장단기 메모리(LSTM: Long-Short Term Memory) 및 게이트 순환 유닛(GRU: Gated Recurrent Unit)을 활용한다. 그동안 수작업으로 심사한 불법확인 결과를 기초 데이터로 이용한다. 한국어 자연어처리와 딥러닝 모델의 하이퍼파라미터 조절을 통해 최적의 성능을 보이는 모델을 완성하였다. 본 연구는 그동안 이뤄지지 않았던 인터넷 불법금융광고 판별을 위한 딥러닝 모델을 제시하였다는데 큰 의미가 있다. 또한 딥러닝 모델에서 91.3~93.4% 수준의 정확도를 보임으로써 불법금융광고 탐지에 딥러닝 모델을 실제 적용하여 불법금융광고 근절에 기여할 수 있기를 기대해 본다.

▸ **주제어**: 불법금융광고, 딥러닝, 합성곱 신경망, 순환 신경망, 장단기 메모리, 게이트 순환 유닛

## I. Introduction

불법금융행위는 금융생활 안전성과 금융시장 질서를 훼손하는 불법행위이다. 정부가 불법금융을 척결하기 위한 노력을 하고 있지만, 불법금융행위 중 보이스피싱 피해건수는 2018년부터 2021년까지 매년 3만건을 유지하고 있다[1]. 불법금융행위는 피해자에게 금전적으로 정신적으로 극심한 고통을 안기며 커다란 사회적 비용을 초래하게 한다[2]. 이런 불법금융행위의 미끼가 사회관계망서비스(SNS: Social Network Service) 등 인터넷에 게시되는 불법금융광고이다. 인터넷 불법금융광고는 인터넷 카페, 블로그 및 인터넷 커뮤니티 등에서 불법대부 상담을 유도하거나 통장매매 및 개인신용정보매매 등을 목적으로 한다[3]. 인터넷 불법금융광고를 통해 불법금융업자에게 넘어간 각종 정보는 보이스피싱 등 불법사금융에 악용되어 자금편취 등의 추가 피해를 유발한다. 인터넷 불법금융광고는 코로나 바이러스 감염증(COVID-19: Corona virus Disease)으로 대면 활동이 위축되면서 더욱 기승을 부렸다. 금융감독원(FSS: Financial Supervisory Service)이 인터넷 불법금융광고를 적발하여 방송통신심의위원회(이하 '방심위')에 게시글 삭제 등 조치의뢰건수가 2020년 10,641건에서 2021년 16,092건으로 큰 폭으로 증가하였다[4].

금융감독원은 Fig. 1과 같이 인터넷 불법금융광고를 수집하여 불법여부를 확인한 후 조치의뢰하는 대응체계를 갖추고 있다. 금융감독원은 시민감시단의 제보, 일반인들의 제보 및 한국인터넷진흥원(KISA: Korea Internet & Security Agency)에서 접수한 제보로 인터넷 불법금융광고 의심 데이터를 수집한다. 또한 키워드 기반으로 인터넷 게시글을 수집(스크래핑)하는 감시시스템을 통해서도 인터넷 불법금융광고 의심 데이터를 수집한다. 금융감독원은 동 게시글이 불법금융광고인지 아닌지 불법여부를 확인하여 불법으로 판단되는 건에 대해서는 방심위에 게시글 삭제 등의 조치를 의뢰한다[4].

금융감독원은 불법금융광고로 의심되는 수많은 인터넷 불법금융광고에 대한 불법여부를 전담직원들이 수작업으로 일일이 확인하는 방식으로 심사하고 있다. 결국 제보건 중심으로 심사업무가 이뤄지고 있다. 따라서 적시 조치에 어려움이 있으며, 직원의 주관적 판단에 의존하여 심사가 이뤄지는 근본적인 한계를 가지고 있다.

이에 본 연구에서는 딥러닝을 활용하여 인터넷 불법금융광고 의심 게시물 중 불법금융광고를 보다 빠르고 일관되게 탐지할 수 있는 모델을 제안한다. 인터넷 불법금융광고 의심 게시물인 텍스트에 딥러닝 기반 텍스트 분류 기법을 적용한다. 금융감독원이 방심위에 조치의뢰한 인터넷 게시글 1,780건을 인터넷 불법금융광고로 레이블링하고,


Fig. 1. The FSS's System for responding to Illegal Financial Advertisements[4]

금융감독원이 불법여부 확인 결과 인터넷 불법금융광고가 아닌 것으로 분류한 게시글 중 일부인 5,338건을 인터넷 불법금융광고가 아닌 것으로 레이블링하여 총 7,118건을 기초 데이터로 활용하였다. 입력 데이터가 텍스트이므로 한국어 자연어 처리 및 딥러닝 모델을 기반으로 인터넷 불법금융광고 여부를 판별한다. 텍스트 분류의 대표적인 딥로닝 모델인 합성곱 신경망(CNN: Convolutional Neural Network), 순환 신경망(RNN: Recurrent Neural Network), 장단기 메모리(LSTM: Long-Short Term Memory) 및 게이트 순환 유닛(GRU: Gated Recurrent Unit) 기법을 적용한다. 해당 딥러닝 모델들을 파이썬으로 구현하고, 하이퍼파라미터를 조절하면서 학습과 테스트를 반복하여 모델별 성능분석을 실시하고, 성능이 우수한 최적의 딥러닝 모델을 제시한다.

본 논문의 구성은 다음과 같다. 제2장에서는 관련 연구를 살펴보고, 제3장에서는 연구 모형을 제시하고, 제4장과 제5장에서는 각각 실험과정과 실험결과를 보여주며, 마지막으로 제6장에서 결론을 내린다.

## II. Related Works

코로나바이러스감염증으로 인한 시장불안, 비대면거래 확대, 정보기술을 활용한 범죄 수법의 발달 등으로 전체 범죄의 감소에도 불구하고 불법금융행위가 개입될 여지가 많은 사이버범죄는 Fig. 2와 같이 지속적으로 증가하는 추세이다[5]. 대표적인 금융사기인 보이스피싱 피해금액은 2019년 6,720억원에 최대를 기록한 후 2022년에 1,451억원으로 점차 줄어들고 있으나, 보이스피싱으로 집계되지 않는 가상화폐 투자유인 등 신종 사기피해가 발생하고 있다[6]. 또한 보험사기 적발금액은 2021년에 9,434억원이었으나 2022년에는 1조 818억원을 기록하여 지속적으로 증가하고 있다[7]. 적발되지 아니한 것까지 포함하면 실제 불법금융으로 인한 피해금액은 훨씬 클 것이다. 영국의 금융사기와 사이버범죄 통합관리센터인 사기대응반(Action Fraud)은 2020년부터 1년간 접수된 사기범죄로 인한 손실액은 약 23.5억 파운드이며, 그중 사이버범죄 손실액은 9.6백만 파운드로 분석하였다[8]. 미국 연방수사국(FBI: Federal Bureau of Investigation) 산하 인터넷범죄신고센터(IC3: Internet Crime Complaint Center)는 2022년 기준 인터넷 범죄 피해액은 약 103억 달러이며 그중 정부 사칭 등 전화사기 피해액은 10.5억 달러로 추정하였다[9].

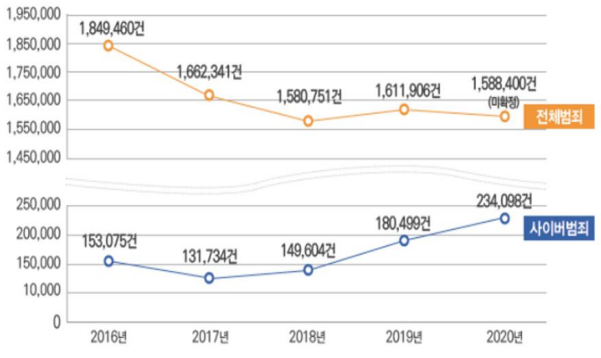


Fig. 2. Overall Crime Trends in the Last 5 Years[5]

인터넷 불법금융광고는 누구나 접근 가능한 인터넷 공간의 특성을 이용한다. 인터넷의 다양한 매체와 커뮤니티에 불법금융광고 게시글을 노출시켜 불법금융행위에 빠져들게 사람들을 유인한다. 불법금융광고의 특성상 Fig. 3과 같이 대부분 텍스트로 이뤄진다. 이미지 형태도 있으나 광고 내용은 텍스트 형태이다. 텍스트를 벡터화하는 자연어처리와 벡터화된 텍스트를 딥러닝 모델로 학습하는 과정을 거쳐 인터넷 불법금융광고 인지 아닌지 이진분류 할 수 있다.

인터넷 게시글에 대한 불법여부를 판별하는 한국어 텍스트 분류로는 불법도박 사이트 및 불법도박 게시글, 마약 등 불법의약품판매 게시글 등을 탐지하는 연구가 있다. 불법도박 의심 사이트의 소스코드를 스크래핑하여 프로그램 예약어 등을 제외한 후 텍스트만을 토큰화하여 미리 정의한 불법도박 광고 키워드와 비교[10]하거나 머신러닝 방법으로 불법도박 사이트인지 여부를 구별하는 연구가 있다[11]. 불법 스포츠 도박을 광고하는 문구가 삽입된 게시글을 사전 키워드별로 빈도별 가중치를 정의하여 불법 스포츠 도박 광고를 차단하는 모델에 관한 연구도 있다[12]. 인터넷 댓글에 대하여 감성분석 및 서포트벡터머신(SVM: Supprot Vector Machine) 기법을 적용하여 악성 댓글을 탐지[13]하거나, 머신러닝과 인공신경망 기법으로 마약매매 게시물을 탐지하는 연구가 있다[14].


Fig. 3. An Example of Illegal Financial Advertisement[4]

텍스트 분류기법으로는 딥러닝 모델이 전통 머신러닝 모델보다 우수한 성능을 보이고 있는 것으로 알려져 있다. 뉴스 데이터의 텍스트 감정분류 실험결과 머신러닝보다 딥러닝 CNN-LSTM 융합모델의 정확도가 우수하였다[15]. 소셜 네트워크 플랫폼에서 중국어 텍스트 분류에 대하여도 CNN-LSTM 융합한 딥러닝 모델이 더 우수한 결과를 가져 왔다[16]. 영화 리뷰에 대한 텍스트 데이터에 대한 감성분류 연구에 있어서도 딥러닝 모델이 우수하였으며 딥러닝 모델중 CNN-LSTM 융합 모델이 가장 우수한 결과를 가져 왔다[17]. 영화리뷰(MR: Movie Reviews, SST-1, SST-2: Stanford Sentiment Treebank) 및 고객리뷰(CR: Customer Reviews)에 대한 긍정·부정 분류, 문장에 대한 주관·객관 분류, 질문 데이터 셋에 대한 사람, 위치, 숫자 정보 등 질문유형을 분류하는 여러 가지 데이터셋에 대한 벤치마크에서 CNN 모델이 탁월한 결과를 얻어 다양한 분류 작업에 활용될 수 있음을 보여주었다[18]. 딥러닝 모델을 한국어 상품평 텍스트의 감성분석에 적용할 때 영어와 한국어의 언어적인 차이로 부딪히게 되는 기본적인 이슈들에 대하여 실증적으로 살펴보며, 모든 품사를 고려한 형태소 임베딩과 CNN를 적용한 결과 감성분석의 분류 정확도가 향상되는 결과를 제시하기도 하였다[19].

각국의 금융감독기관들은 디지털 감독 혁신에 정책적 우선순위를 두고 인공지능, 머신러닝 및 빅데이터 기술을 활용한 섭테크(SupTech: Supervisory Technology) 전략을 도입하고 있다. 우리나라 금융감독원도 도입 초기 단계이긴 하지만 보이스피싱 및 불법채권추심 등에 인공지능 분석을 적극 활용하고자 하고 있다[20]. 또한 이와 관련된 학술연구도 있다. 수사관들의 보이스피싱 사건정보를 자연어 처리 및 딥러닝 모델 개발을 통해 실제 사건 수사자료에 적용하여 보이스피싱 사건을 자동 인식·추출하는 연구가 있다[2]. 불법 채권추심행위를 점검하기 위하여 불법행위를 판별하는 규칙기반 검출과 SVM 등 머신러닝을 결합한 불법채권추심 분류모델에 대한 연구도 있다[21]. 금융상품 불완전판매를 적발하기 위해 상담원과 고객간의 상담내용을 텍스트로 변환(STT: Speech To Text)한 후 해당 텍스트에 딥러닝 기반 텍스트 분류 기법을 적용한 연구도 있다[22]. 금융감독분야 중 시세조종, 내부자거래, 불공정거래 행위 적발 등 시장감시는 물론 금융사기 예방·적발 및 인터넷 불법금융광고 판별 등 소비자보호와 관련하여 인공지능 기반의 섭테크를 확대 적용·활용할 필요가 있다[20]. 인공지능을 활용한 섭테크 관련 텍스트 분류에 대한 연구는 다양하게 존재함에도 불구하고 인터넷 불법금융광고를 주제로 한 자연어처리, 머신러닝 및 딥러닝 등 섭테크를 활용한 학술연구는 찾을 수가 없었다. 레이블링된 기초 데이터를 구하기 어려웠기 때문으로 판단된다. 본 연구는 인터넷 불법금융광고 의심 게시물 중 불법금융광고를 탐지하기 위한 딥러닝을 적용한 텍스트 분류모델을 제안하였고, 그 중 CNN 모델의 분류 정확도가 더 높다는 것을 보여 주었다.

## III. Proposed Model

인터넷 게시글 등 한국어 텍스트 분류에 관한 선행 연구를 살펴본 결과 텍스트를 토큰화하여 머신러닝과 인공신경망 기법으로 분석하였으며, CNN 및 LSTM 등 딥러닝 모델의 성능이 우수함을 확인하였다. 특히 보이스피싱, 불법 채권추심 및 금융상품 불완전판매 등 금융감독 분야에 섭테크를 적용한 연구에서는 인공신경망 기법을 적용하고 있었다.

본 연구는 Fig. 4와 같이 인터넷 불법금융광고 감시시스템을 통해 수집한 인터넷 불법금융광고 의심 게시글을 활용한다. 해당 게시글 중 금융감독원 직원의 불법여부 확인을 거쳐 불법금융광고 여부를 판별한 정보를 기초 데이터로 이용한다. 제보로 입수된 데이터는 대부분 텍스트를 포함한 이미지이기 때문에 이번 연구에서는 활용

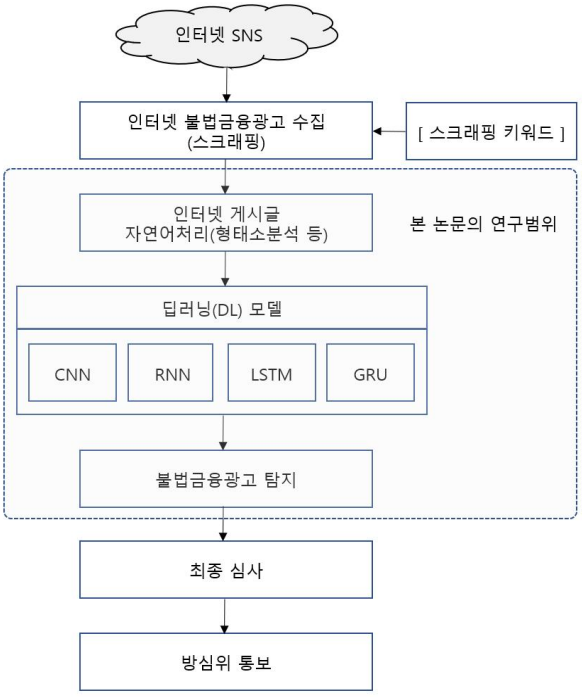


Fig. 4. Proposed Model

하지 않았다. 텍스트 토큰화 및 임베딩 벡터화 등 자연어 처리를 하여 딥러닝 학습과 테스트를 위한 데이터셋을 갖춘다. 이후 CNN, RNN, LSTM 및 GRU 딥러닝 모델을 구현하여 학습과 테스트를 반복적으로 수행한다. 패딩 길이, 임베딩 사이즈, 활성화 함수 및 옵티마이저 등 딥러닝 모델의 하이퍼파라미터를 조절하면서 평가하여 최고의 성능을 보이는 딥러닝 모델을 완성한다.

## IV. Experimental Process

### 1. Data set

2019년부터 2022년까지 약 4년간 인터넷 불법금융광고 관련 키워드로 웹 스크래핑하여 수집한 데이터 중 불법여부 확인을 거친 심사결과를 기초 데이터로 사용한다. 기초 데이터로 불법금융광고 내용과 심사의견 등을 추출한다. 불법금융광고 내용은 인터넷 게시글로 한국어 텍스트이다. 심사결과는 불법여부 및 조치의뢰 결과를 확인할 수 있는 값이다. 해당 심사결과 값을 기준으로 인터넷 불법금융광고 여부를 판단하는 레이블 값을 '1' 또는 '0'의 값으로 부여한다. 이는 금융감독원 직원이 불법여부를 심사하여 방심위에 조치의뢰 여부를 결정한 값으로 전문가가 분류한 결과이기에 정답으로 정의했다. 기본적인 데이터 정제를 위해 불법금융광고 내용에서 불필요한 문자를 삭제하기 위해 한글과 숫자, 공백 외의 문자를 빈값(Null) 처리하였다. 'ㅋ', 'ㅎ' 등의 의성어와 같은 특이한 문자들을 빈값 처리하는 과정도 거쳤다. 기초 데이터셋은 인터넷 불법금융광고가 1,780건(25.0%), 불법이 아닌 일반금융광고가 5,338건(75.0%)으로 총 7,118건이다. 불법이 아닌 일반금융광고는 불법금융광고보다 훨씬 많은 비중을 차지하므로 본 연구에서는 불법금융광고의 3배 수준으로 샘플링하였다.

### 2. Split of training and test data

기초 데이터셋의 70%를 학습데이터로, 나머지 30%를 테스트데이터로 사용하기로 하고 데이터를 분리한다. 학습데이터와 테스트데이터를 생성한 후 랜덤하게 순서를 재배치했다. 학습데이터는 불법금융광고 1,246건과 일반금융광고 3,736건으로 총 4,982건을 사용한다. 테스트데이터는 나머지 불법금융광고 534건과 일반금융광고 1,602건으로 총 2,136건을 사용한다.

### 3. Text tokenization

한글 텍스트의 토큰화 방법은 다양하나 한글 형태소 분석기를 활용한 단어 토큰화(Word Tokenization)를 실시한다. 단어 토큰화는 형태소 분석을 기준으로 이뤄지며 여러 품사를 태깅하기도 한다. 본 연구에서는 다양한 한국어 토크나이저 중 Okt 형태소 분석기를 사용하여 형태소 토큰화(Morpheme Tokenization)를 수행한다. 형태소 토큰화를 하면서 함께 불용어(Stopword)를 제외 처리하기로 한다. 텍스트 마이닝을 위한 한국어 범용 불용어 목록 293개[23]에 해당 목록에 없는 '을', '를', '이', '가' 등의 조사 15개를 추가하여 308개의 불용어를 정의하였다. 학습 및 테스트 데이터를 토큰화하면서 해당 불용어는 토큰에서 제외시킨다. Fig. 5는 토큰화된 학습 및 테스트 테이터 중 레이블이 '1'인 불법금융광고의 형태소 36,075건에 대한 워드 클라우드(Word Cloud)이다.


Fig. 5. Word Cloud

### 4. Integer encoding and padding

학습데이터 4,982건을 빈도수 기준으로 단어집합(Vocabulary)을 생성한다. 단어집합의 전체 크기는 32,524건이다. 이중 등장 빈도가 2번 이하인 단어를 희귀단어로 제외한다. 희귀단어는 15,675건으로 단어집합에서 차지하는 비율은 48.19%이다. 하지만, 전체 등장빈도에서 희귀단어의 등장빈도 비중은 0.82%로 낮아 단어집합에서 희귀단어를 삭제한다. 단어집합의 크기는 0번 패딩을 고려하여 16,850건(= 32,524 - 15,675 + 1)으로 정의한다. 해당 단어집합의 크기를 기준으로 데이터에 대한 정수 인코딩을 다시 수행하여 희귀단어를 제외시켰다.

학습데이터에 대한 정수 인코딩 결과 최대 길이는 4,116이었으며, 평균 길이는 480.15이다. Fig. 6은 정수 인코딩한 학습데이터의 길이 분포도이다. 길이가 1,800 이하인 데이터의 비율은 98.77% 수준이다. 패딩 길이를 1,800으로 정하여 학습데이터와 테스트데이터를 패딩 길이에 맞춰 동일하게 패딩하였다. 일반적으로 패딩 길이보다 긴 데이터는 잘리게 되며, 패딩 길이보다 짧은 데이터의 경우 뒷 부분 빈 공간에 '0'을 채우는 트레일링 패딩(Trailing Padding)을 하게 된다. 이후 패딩길이를 1,500(97.15%) 및 1,200(91.75%) 등으로 조절하면서 딥러닝 모델의 성능 변화를 확인한다.

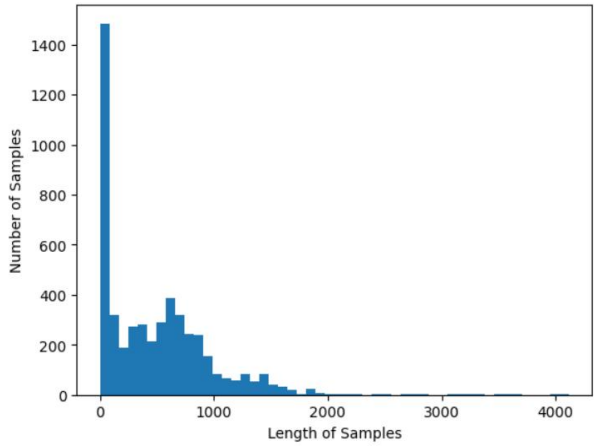


Fig. 6. The Length Distribution of the Integer-Encoded Training Data

## 5. Embedding

본 연구에서는 정보손실을 줄이고 데이터를 효과적으로 나타낼 수 있는 밀집벡터(Dense Vector)로 표현하는 워드 임베딩(Word Embedding)을 실시한다. 이는 데이터의 차원수를 줄이기도 하지만 단어 벡터간에 거리나 유사도를 측정할 수 있다. 의미가 유사한 단어는 벡터공간에 가깝게 반대의 경우는 멀게 배치하는 방식이다. 본 연구에서는 한국어 텍스트를 정수 인코딩을 하였기에 케라스(Keras)에서 제공하는 기본 도구인 Embedding() 함수를 사용하였다. 임베딩 벡터의 차원수는 256로 설정하였다. 이후 128 및 64 등으로 조절하면서 딥러닝 모델의 성능 변화를 확인하였다.

## 6. Deep learning model

딥러닝은 머신러닝의 일부분으로 인공신경망의 층을 연속적으로 쌓아 올려 모델을 생성한다. 입력층과 출력층 사이에 여러 은닉층을 포함한 심층신경망(DNN: Deep Neural Network)을 이용한다. 본 연구에서는 텍스트 분류를 위한 대표적인 딥러닝 모델인 1차원 CNN (Conv1D), RNN, LSTM, 양방향 LSTM(Bidirectional LSTM) 및 GRU 모델을 구현한다. 각 모델의 입력층과 출력층 사이의 은닉층(Hidden Layer)를 1개 또는 3개로 구현하여 학습과 테스트를 반복하였다.

1차원 CNN의 경우 합성곱 연산에 사용되는 필터 커널 사이즈를 5 또는 3으로 하면서 출력 필터수를 128 또는 64개로 조절한다. RNN, LSTM, Bidirectional LSTM 및 GRU의 경우 은닉층의 유닛수를 128 또는 64개로 조절하였다. 은닉층을 3개층으로 구현할 경우에는 유닛수를 128, 64, 32 또는 64, 32, 16과 같이 절반씩 줄여나가는 방식으로 설정하였다. 딥러닝 활성화 함수(Activation)도 렐루(Relu), 하이퍼볼릭 탄젠트(Tanh), 시그모이드(Sigmoid) 및 소프트맥스(Softmax) 등으로 조절하였다. 또한 과적합을 방지하기 위한 드롭아웃 비율(Dropout Rate)도 0.5에서 0.2 사이값으로 설정을 변경하면서 딥러닝 모델의 성능 변화를 확인하였다.

각 딥러닝 모델을 학습할 경우 적용하는 옵티마이저는 RMSprop(Root Mean Square Propagation), 아담(Adam), SGD(Stochastic Gradient Descent) 및 아다그라드(Adagrad: Adaptive Gradient) 등으로 변경하며 학습률(Learning Rate), 베타(Beta), 로우(Rho) 및 모멘텀(Momentum) 등의 값에 변화를 주면서 해당 모델들을 학습 및 테스트를 반복 수행한다. 테스트 데이터의 손실이 증가하면 과적합 징후이므로 손실이 5회 증가하면 학

Table 1. Hyperparameter setting values

| Hyperparameters | | Setting values |
|---|---|---|
| related to the model architecture | input length of embedding layer | 1800, 1,500 or 1,200 |
| | output dim of embedding layer | 256, 128 or 64 |
| | number(size) of hidden layers | 1(128 or 64)<br>3(128-64-32 or 64-32-16) |
| | activation | relu, tanh, sigmoid or softmax, etc. |
| | dropout rate | 0.5, 0.3 or 0.2 |
| related to the training algorithm | optimizer | RMSprop, Adam, SGD or Adagrad, etc. |
| | learning rate | 0.01, 0.05 or 0.001 |
| | beta_1 | 0.3 |
| | beta_2 | 0.5 |
| | rho | 0.9 or 0.6 |
| | momentum | 0.9, 0.7 or 0.6 |
| | weight_decay | 0.001 or 0.0 |
| | epoch | 40 or 20 |
| | batch_size | 60, 30 or 20 |
| | validation_split | 0.3 or 0.2 |

습을 조기 종료(Early Stopping)하도록 한다. 테스트 데이터의 정확도가 이전보다 좋을 경우에는 해당 모델을 저장하도록 한다. 본 연구 모델의 문제 해결은 이진 분류이므로 손실(Loss) 함수를 'Binary_CrossEntropy'로 설정한다. 학습과 테스트 중 모델의 성능을 모니터링하는 평가 지표(Metrics)로는 정확도인 'Acc'를 지정한다. 또한 에포크(Epoch)를 20으로 설정하고 필요시 40으로 늘려 실험한다. 배치사이즈(Batch Size)는 60, 30 및 20으로 조절하면서 테스트 데이터 사용비율(Validation Split)을 0.3 또는 0.2로 변화를 주면서 딥러닝 모델의 성능을 확인하였다. Table 1은 각 딥러닝 모델의 하이퍼파라미터를 튜닝할 때 설정했던 주요 값들이다. 실행시 마다 성능변화 폭을 최소화하기 위해 시드값을 고정하였다.

### 7. Experimental evaluation

본 연구는 인터넷 게시글이 불법금융광고에 해당되는지 아닌지를 판별하는 이진 분류의 문제이다. 이진 분류의 성능을 평가하는 지표는 일반적으로 Table 2와 같이 혼동행렬(Confusion Matrix)을 사용한다. 분류 모델이 예측한 결과값(Predicted Value)과 실제 결과값(Observed Value)을 비교하여 얼마나 맞추었는지 틀렸는지를 나타내는 행렬이다. 이 혼동행렬을 기반으로 Table 3과 같이 정확도(Accuracy), 정밀도(Precision), 재현율(Recall) 및 F1 스코어(F1 Score) 등을 계산하여 측정한다.

Table 2. Confusion Matrix

| Predicted value | Observed value | |
|---|---|---|
| | illegal (1) | Normal (0) |
| illegal (1) | TP(True Positive) | FP(False Positive) |
| Normal (0) | FN(False Negative) | TN(True Negative) |

Table 3. Performance Evaluation Metrics

| Items | Calculation formula |
|---|---|
| Accuracy | (TP + TN) / (TP + TN + FP + FN) |
| Precision | TP / (TP + FP) |
| Recall | TP / (TP + FN) |
| F1 Score | 2×(Precision×Recall) / (Precision + Recall) |

### 8. Experimental environment

본 연구의 실험 환경은 Table 4와 같다. 실험 컴퓨팅 엔진은 구글 코랩(Colab)을 이용하였다. 하이퍼파라미터 튜닝 등 딥러닝 모델의 성능 확인 작업은 구글 코랩 프로 플러스(Colab Pro+) 환경에서 하였다. 모델 구현 등 실험을 위해 사용한 프로그램 언어는 파이썬이며, 딥러닝 라이브러리는 텐서플로우(Tensorflow), 코앤엘파이(KoNLPy) 및 케라스(Karas) 등을 사용하였다.

Table 4. Experiment environment

| Items | Values |
|---|---|
| Compute Engine | Google Colab Pro+ |
| Language | Phython 3.10.11 |
| Deep learning library | Tensorflow 2.12.0<br>KoNLPy 0.6.0<br>Keras 2.12.0<br>Pandas 1.5.3<br>Numpy 1.22.4<br>Matplotlib 3.7.1 |

## V. Experimental Results

본 연구의 딥러닝 모델로 1차원 CNN, RNN, LSTM, 양방향 LSTM 및 GRU 모델을 구현한 후 하이퍼파라미터를 변경하면서 학습과 테스트를 반복한 결과 Table 5와 같이 1차원 CNN과 양방향 LSTM 모델에서 우수한 정확도를 보였다. 각 모델에 대해서는 5회 이상 반복 실행하여 최고 및 최저 정확도를 확인하였다. 테스트 데이터에 대한 성능 평가결과 1차원 CNN 모델이 0.9237 ~ 0.9340로 가장 좋은 정확도를 보였다. 다음으로는 양방향 LSTM 모델이 0.9190 ~ 0.9251의 정확도를 보였다. 각 모델별로 최고의 정확도를 보인 임베딩 층은 자연어 처리한 16,850개의 단어집합을 입력 시퀀스 길이를 1,800개로 128 차원으로 임베딩하는 조건으로 동일하였다. 임베딩 차원이 256 또는 128일 경우 정확도 영향은 크지 않았다. 딥러닝 모델의 구조 측면에서 3개층으로 은닉층을 구성한 모델보다 Fig. 7과 같이 1개의 은닉층으로 구성한 단순한 모델이 오히려 더 좋은 정확도를 보였다. 은닉층의 활성화 함수를 'Relu' 또는 'Tanh'로, 출력층의 활성화 함수는 'Sigmoid'로 설정한 경우가 정확도가 높았다. 학습 알고리즘인 옵티마이저로는 'RMSprop' 또

```
Model: "sequential_8"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_6 (Embedding)     (None, None, 128)         2156800

 conv1d (Conv1D)             (None, None, 64)          41024

 dropout_3 (Dropout)         (None, None, 64)          0

 dense_4 (Dense)             (None, None, 1)           65

=================================================================
Total params: 2,197,889
Trainable params: 2,197,889
Non-trainable params: 0
```

Fig. 7. 1D CNN Model Summary

Table 5. Experiment Result

| Items | | | 1D CNN (Conv1D) | RNN (SimpleRNN) | LSTM | Bidirectional LSTM | GRU |
|---|---|---|---|---|---|---|---|
| hyper-parameter related to the model architecture | embedding layer | | input_dim (size of vocabulary) : 16,850<br>input_length (length of input sequences) : 1,800<br>output_dim (dimension of the dense embedding) : 128 | | | | |
| | hidden layer | number(size) | 1(128) | 1(128) | 1(128) | 1(128) | 1(128) | 1(128) |
| | | activation | relu | relu | relu | tanh | tanh | tanh |
| | | dropout rate | 0.3 | - | - | - | 0.3 | - |
| | output(dense) layer | activation | sigmoid | sigmoid | sigmoid | sigmoid | sigmoid | sigmoid |
| hyper-parameter related to the training algorithm | optimizer | | RMSprop | Adam | Adam | RMSprop | Adam | Adam |
| | learning rate | | 0.001 | 0.01 | - | 0.001 | 0.01 | - |
| | beta_1 (only Adam) | | X | 0.3 | - | - | 0.3 | - |
| | beta_2 (only Adam) | | X | 0.5 | - | - | 0.5 | - |
| | rho (only RMSprop) | | 0.9 | X | X | 0.9 | X | X |
| | momentum (only RMSprop) | | 0.2 | X | X | 0.0 | X | X |
| | weight_decay | | - | - | - | - | - | - |
| | epoch | | 20 | 20 | 20 | 20 | 20 | 20 |
| | batch_size | | 20 | 20 | 60 | 60 | 30 | 60 |
| | validation_split | | 0.3 | 0.3 | 0.3 | 0.3 | 0.2 | 0.3 |
| experiment result | val_accuracy | highest | 0.94783 | 0.95117 | 0.94247 | 0.93311 | 0.94985 | 0.92910 |
| | | lowest | 0.94515 | 0.94515 | 0.93244 | 0.92843 | 0.94283 | 0.92843 |
| | test_accuracy | highest | **0.9340** | **0.9316** | **0.9204** | **0.9185** | **0.9251** | **0.9199** |
| | | lowest | 0.9260 | 0.9237 | 0.9190 | 0.9134 | 0.9190 | 0.9167 |
| | | Gap | 0.0080 | 0.0079 | 0.0014 | 0.0051 | 0.0061 | 0.0032 |

는 'Adam'으로 설정 한 경우 정확도가 좋았다. 은닉층의 드롭아웃 비율을 0.3으로 세팅하고, RMSprop 옵티마이저를 이용하면서 학습률, 로우 및 모멘텀 값을 각각 0.001, 0.9, 0.2로 설정한 1차원 CNN 모델에서 최고의 정확도를 보였다. 해당 1차원 CNN 모델은 학습 과정에서 11번째 에포크때 조기 종료되었다. Fig. 8은 에포크 진행에 따른 학습 데이터 및 테스트 데이터의 1차원 CNN 모델의 정확도 'Acc'값과 손실 'Loss'값의 변화 추이이다.

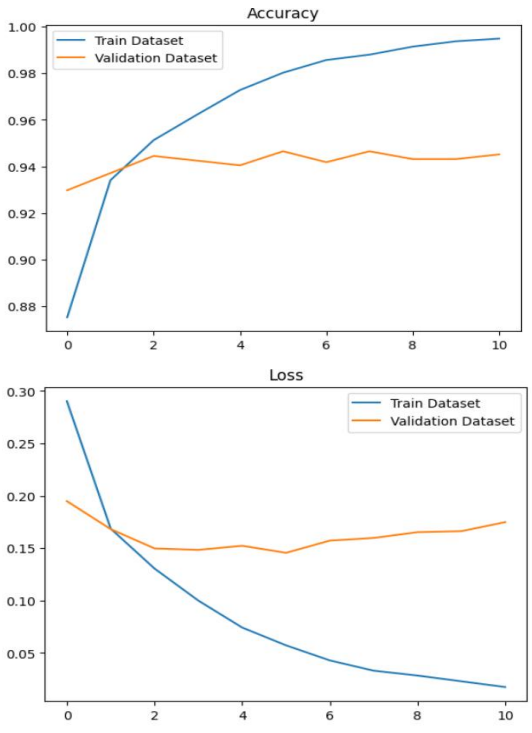


Fig. 8. Accuracy and Loss Curves of the 1D CNN Model

최고의 정확도를 보인 1차원 CNN 모델에 테스트 데이터를 적용하여 성능 평가지표를 확인한다. Table 6은 학습된 1차원 CNN 모델에 테스트 데이터 2,136건을 적용한 결과인 혼동행렬이다. 이를 기반으로 Table 7과 같이 정확도(0.9340), 정밀도(0.8673), 재현율(0.8689) 및 F1 스코어(0.8681)를 확인한다. 테스트 데이터 2,136건 중 불법금융광고 464건과 일반금융광고 1,531건을 제대로 분류하여 93.40% 수준의 정확도를 보였다. 불법금융광고로 예측한 535건 중 464건을 정확히 예측하여 86.73%의 정밀도를 보였다. 불법금융광고 534건 중 464건을 정확히 판별하여 86.89%의 재현율을 보였다.

Table 6. Confusion Matrix of 1D CNN Model

| Predicted value | Observed value (Number of cases) | | Sum |
|---|---|---|---|
| | illegal (1) | Normal (0) | |
| illegal (1) | 464 | 71 | 535 |
| Normal (0) | 70 | 1,531 | 1,601 |
| Sum | 534 | 1,602 | 2,136 |

Table 7. Result of classification evaluation index

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| 1D CNN | 0.9340 | 0.8673 | 0.8689 | 0.8681 |

## VI. Conclusions

본 연구는 그동안 연구되지 않았던 인터넷 게시물에 대한 불법금융광고 여부를 판별하는 딥러닝 모델을 제시하였다는 점에서 큰 의미가 있다. 또한 1차원 CNN 딥러닝 모델의 하이퍼파라미터 조절을 통해 최고 0.9340의 정확도를 보임을 검증하였다. 특히 본 연구에서 구현하였던 1차원 CNN, RNN, LSTM, 양방향 LSTM 및 GRU의 모든 딥러닝 모델에서 0.9134 ~ 0.9340의 정확도를 보임을 확인하였다. 이는 해당 딥러닝 모델을 실제 적용할 수 있음을 보여준다. 금융감독분야에 섭테크(SupTech)를 적용하는 좋은 사례가 될 수 있을 것이다.

한편 제보 건은 인터넷 게시글이 대부분 이미지 또는 인터넷 주소(URL)로 수집되어 본 연구에 활용하지 못한 한계가 있었다. 기초 데이터가 많지 않았음에도 본 연구는 금융감독 분야에 최신 기술을 적용하는 섭테크의 일환으로 인터넷 불법금융광고 탐지 업무에 딥러닝을 실무적 적용 가능성을 확인하는 성과를 거두었다.

향후 제보된 인터넷 주소(URL) 등을 이용하여 제보를 접수함과 동시에 텍스트를 확보하는 등의 방법으로 대량의 학습 데이터를 확보하여 추가 연구를 하고자 한다. 특히 SNS 카테고리별로 딥러닝 모델을 각각 구현하는 등 성능 향상 방안과 이미지로 된 불법금융광고 탐지 영역으로 연구 범위를 추가·확장하고자 한다.

## REFERENCES


[1] Korean National Police Agency, "Monthly Status Report on Voice Phishing Case," Public Data Portal [Online], *Availability: https://www.data.go.kr/data/15099013/fileData.do*

[2] H. J. Kim, “An Implementation of Natural Language Processing and Deep Learning in Phone Scam Investigation,” Journal of Korean Criminological Association, Vol. 16, No. 1, pp. 123-141, 2022. DOI: 10.29095/JKCA.16.1.6

[3] Korea Ministry of Government Legislation, “The Types and Regulations of Illegal Financial Advertising,” The Easy Legal Information Service [Online], *Availability: https://easylaw.go.kr/CSP/CnpClsMain.laf?csmSeq=901&ccfNo=1&cciNo=1&cnpClsNo=1*

[4] Financial Supervisory Service, “The Crackdown, Collection, and Actions taken against Illegal Financial Advertisements in 2021,” Press Releases [Online], *Availability: https://www.fss.or.kr/fss/bbs/B0000188/view.do?nttId=56214&menuNo=200218&cl1Cd=&sdate=&edate=&searchCnd=1&searchWrd=%EB%B6%88%EB%B2%95%EA%B8%88%EC%9C%B5%EA%B4%91%EA%B3%A0&pageIndex=1*

[5] Korean National Police Agency, “Analysis Report on Cybercrime Trends in 2020,” Notifications/News [Online], *Availability: https://www.police.go.kr/user/bbs/BD_selectBbs.do?q_bbsCode=1001&q_bbscttSn=20210610100256713*

[6] Financial Supervisory Service, “Voice Phishing Incidents and Key Characteristics in 2022,” Press Releases [Online], *Availability: https://www.fss.or.kr/fss/bbs/B0000188/view.do?nttId=127319&menuNo=200218*

[7] Financial Supervisory Service, “Insurance Fraud Detection Status and Future Plans in 2022,” Press Releases [Online], *Availability: https://www.fss.or.kr/fss/bbs/B0000188/view.do?nttId=58396&menuNo=200218&cl1Cd=&sdate=&edate=&searchCnd=1&searchWrd=%EB%B3%B4%ED%97%98%EC%82%AC%EA%B8%B0&pageIndex=1*

[8] The UK’s Action Fraud, “Fraud Crime Trends” & “Cyber Crime Trends,” Fraud and Cyber Crime Statistics [Online], *Availability: https://www.actionfraud.police.uk/fraud-stats*

[9] The USA’s Internet Crime Complaint Center(IC3), “Internet Crime Report 2022”, IC3’s Annual Report [Online], *Availability: https://www.ic3.gov/Home/AnnualReports*

[10] K. S. Lee, J. H. Lee, and H. M. Cho, “Keyword Combination based Classification for Illegal Gambling Websites,” Korea Computer Congress 2021, pp. 1194-1196, 2021.

[11] C. W. Song, and H. C. Ahn, “Development of an Intelligent Illegal Gambling Site Detection Model Based on Tag2Vec,” Journal of Intelligence and Information Systems, Vol. 28, No. 4, pp. 211-227. 2022. DOI: 10.13088/jiis.2022.28.4.211

[12] J. A. Kim, and G. B. Lee, “An Effective Method for Blocking Illegal Sports Gambling Ads on Social Media,” Journal of the Korea Society of Computer and Information, Vol. 24, No. 12, pp. 201-207, 2019. DOI: 10.9708/jksci.2019.24.12.201

[13] J. U. Hong, S. H. Kim, J. W. Park, and J. H. Choi, “A Malicious Comments Detection Technique on the Internet using Sentiment Analysis and SVM,” Journal of the Korea Institute of Information and Communication Engineering, Vol. 20, No. 2, pp. 260-267. 2016. DOI: 10.6109/jkiice.2016.20.2.260

[14] J. H. Park, S. Y. Cho, J. H. Lee, H. T. Lim, and Y. G. Cheong, “Detection of Illicit Drug Selling Post on an Online Community Using Phoneme Separation and Machine Learning Algorithm,” Korea Computer Congress 2020, pp. 368-370, 2020.

[15] S. Y. Shin, K. S. Shin, and H. C. Lee, “Text Classification Using LSTM-CNN,” Journal of Information and Communication Convergence Engineering, Vol. 23, No. 2, pp. 692-694. 2019.

[16] Y. Chen, T. Juan, and H. K. Jung, “Text Classification on Social Network Platforms Based on Deep Learning Models,” Journal of Information and Communication Convergence Engineering, Vol. 21, No. 1, pp. 9-16, 2023. DOI: 10.56977/jicce.2023.21.1.9

[17] H. Y. Park, and K. J. Kim, “Sentiment Analysis of Movie Review Using Integrated CNN-LSTM Model,” Journal of Intelligence and Information Systems, Vol. 25, No. 4, pp. 141-154, 2019. DOI: 10.13088/jiis.2019.25.4.141

[18] Y. Kim, "Convolutional Neural Network for Sentence Classification," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746-1751, 2014. DOI: 10.3115/v1/D14-1181

[19] H. J. Park, M. C. Song, and K. S. Shin, “Sentiment Analysis of Korean Reviews Using CNN - Focusing on Morpheme Embedding,” Journal of Intelligence and Information Systems, Vol. 24, No. 2, pp. 59-83, 2018. DOI: 10.13088/jiis.2018.24.2.059

[20] Y. K. Huh, “Legal Implications of Financial Supervision with Artificial Intelligence,” The Korean Journal of Securities Law, Vol. 23, No. 1, pp. 221-250, 2022. DOI: 10.17785/kjsl.2022.23.1.221

[21] T. H. Kim, J. I. Lim, “A Classification Model for Illegal Debt Collection Using Rule and Machine Learning Based Methods,” Journal of the Korea Society of Computer and Information, Vol. 26, No. 4, pp. 93-103. 2021. DOI: 10.9708/jksci.2021.26.04.0

[22] J. H. Kim, and J. I. Won, “Discrimination Model On Misselling of Financial Products Using Deep Learning,” KIISE Transactions on Computing Practices, Vol. 25, No. 6, pp. 294-302, 2019. DOI: 10.5626/KTCP.2019.25.6.294

[23] X. H. LU, and J. Jin, “A Study on the Lists of Common Korean Stopwords for Text Mining,” Korean Language Research Circle, Vol. 63, No. 13, pp. 1-15, 2022. DOI: 10.16876/klrc2022..63.13

## Authors

Kil-Sang Yoo received the M.S. degree in Information Security from Korea University, Seoul, Korea in 2012. Kil-Sang Yoo is currently a Ph.D. student in the Graduate School of IT Policy and Management at Soongsil University, Seoul, Korea. He has been in the Financial Supervisory Service since 1993. He is interested in Electronic Finance, Virtual Assets and Artificial Intelligence(Machine Learning and Deep Learning)

Jin-Hee Jang received the M.S. degree in electrical and electronic engineering from Hanyang University, Seoul, Korea in 2007. Jin-Hee Jang is currently a Ph.D. student in the Graduate School of IT Policy and Management at Soongsil University, Seoul, Korea. He is currently working as a Business in Information Security Industry. He is interested in Blockchain, Artificial Intelligence, Information Security.

Seong-Ju Kim received the M.S. degree in Information Technology Policy & Management from Soongsil University, Seoul, Korea in 2022. Seong-Ju Kim is currently a Ph.D. student in the Graduate School of IT Policy and Management at Soongsil University, Seoul, Korea. He is currently working as a Software Engineer in Metaverse Industry. He is interested in Metaverse, Computer Vision, Artificial Intelligence, IoT.

Kwang-Yong Gim received the M.S. degree from the Graduate School of Business at Korea University, Seoul, Korea, in 1983. He earned his Ph.D. degree in Business from Mississippi State University, USA, in 1993. Dr, Gim is currently a Professor in the Department of Business at Soongsil University, Seoul, Koera. He is interested in Software Testing, Quality Assurance, Management Information systems (MIS), Information Security and Affective Computing.