JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 2, pp. 75-85, February 2026
https://doi.org/10.9708/jksci.2026.31.02.075

A Novel Two-Stage Attacks on Korean Language Models: Single-Token Triggers Search and Morphology-Preserving Minimal Edits

Areum Im*, Taehwa Lee*, Soojin Lee**

*Graduate Student, Dept. of Cyber Security and Computer Engineering, Korea National Defense University, Nonsan, Korea
**Professor, Dept. of Cyber Security and Computer Engineering, Korea National Defense University, Nonsan, Korea

[Abstract]

In this study, we propose a novel two-stage attack framework applicable to Korean-based language models with agglutinative characteristics. The first stage is an inference-time universal adversarial trigger (UAT) attack, performed without intervention in the learning process. It precisely searches for single-token triggers capable of reversing the model's predictions using only the gradient information. The second stage, targeting only samples that failed in the first stage, is an adversarial example attack. It replaces no more than two tokens combining particles and suffixes based on a morphology-preserving minimal edit strategy. The effectiveness of our framework was evaluated on the NSMC dataset using the KoBERT and KoELECTRA models. Experimental results showed that triggers attached to the end of sentences had a high attack success rate due to the characteristic of Korean language in which key information appears at the end of sentences. Furthermore, words that indirectly express sentiment also functioned as powerful triggers. The KoBERT model achieved an attack success rate of 0.963, and the KoELECTRA model achieved an attack success rate of 0.940.

▸ **Key words**: Adversarial example, Universal adversarial trigger, Korean language model

[요 약]

본 연구에서는 교착어의 특성을 가지는 한국어 기반 언어모델을 대상으로 적용 가능한 혁신적인 2단계 공격 프레임워크를 제안한다. 1단계 공격은 학습 과정에 대한 개입 없이 수행되는 범용 적대적 트리거(Universal Adversarial Trigger) 탐색 공격으로, 모델의 그라디언트 정보만 활용하여 예측을 반전시킬 수 있는 단일 토큰 트리거를 정밀하게 탐색한 후 공격을 수행한다. 1단계 공격에서 실패한 샘플들만을 대상으로 수행하는 2단계 공격은 적대적 예제 공격으로, 형태 보존 최소 편집 전략에 의거해 조사와 어미를 결합한 토큰을 2개 이내에서 교체한다. 제안하는 프레임워크의 효용성은 NSMC 데이터세트를 대상으로 KoBERT 및 KoELECTRA 모델을 이용해 평가하였다. 실험 결과, 핵심 정보가 문장 말미에 나타나는 한국어의 특성으로 인해 문장 뒤에 부착된 트리거가 높은 공격 성공률을 보였다. 그리고 간접적으로 정서를 표현하는 단어도 강력한 트리거로 작동하였다. KoBERT 모델의 공격 성공률은 0.963, KoELECTRA 모델의 공격 성공률은 0.940으로 확인되었다.

▸ **주제어**: 적대적 예제, 범용 적대적 트리거, 한국어 언어 모델

• First Author: Areum Im, Co-Author: Taehwa Lee, Corresponding Author: Soojin Lee
*Areum Im (lima8255@gmail.com), Dept. of Cyber Security and Computer Engineering, Korea National Defense University
*Taehwa Lee (whitechoke@gmail.com), Dept. of Cyber Security and Computer Engineering, Korea National Defense University
**Soojin Lee (cyberkma@korea.ac.kr), Dept. of Cyber Security and Computer Engineering, Korea National Defense University
• Received: 2025. 12. 08, Revised: 2026. 01. 07, Accepted: 2026. 01. 26.

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

I. Introduction

대규모 언어모델과 사전 학습 언어모델은 검색, 콘텐츠 추천, 챗봇 등 광범위한 산업 분야의 핵심 기술로 자리 잡았다. 특히 한국어의 고유한 언어 특성을 반영한 모델들이 등장하여 문맥적 의미를 정교하게 파악할 수 있게 되면서, 그 활용 범위는 비약적으로 확대되고 있다. 그러나 이러한 기술적 확산은 역설적으로 언어모델의 취약성을 겨냥한 새로운 보안 위협을 야기하고 있다.

언어모델을 위협하는 대표적인 공격 기법으로는 중요도가 높은 소수 토큰을 치환하여 오분류를 유도하는 적대적 공격[1-7]과, 학습 단계에서 데이터를 오염시켜 특정 트리거에 반응하도록 하는 데이터 오염 기반 공격(Data Poisoning Attacks)[8-12] 등이 있다. 선행 연구들에 따르면 이러한 기법들은 다양한 환경에서 높은 공격 성공률을 보이며 모델의 취약점을 효과적으로 공략하는 것으로 검증되었다. 그러나 대부분 연구가 단어 간 경계가 명확하고 어순이 문법적 기능을 주도하는 영어를 기반으로 하는 언어모델을 대상으로 수행되었으며, 한국어 기반 모델을 주된 공격 대상으로 삼아 그 구조적 취약성을 심도있게 분석한 연구는 상대적으로 부족한 실정이다.

한국어는 체언에 조사가 결합하거나 용언의 어간에 어미가 활용되어 문법적 관계와 의미를 형성하는 교착어의 특성을 지닌다. 영어와 달리 명사-조사, 어간-어미, 부정부사의 결합 등이 빈번하게 발생하는 한국어만의 고유한 특성은, 영어 기반 모델을 대상으로 설계된 공격 기법의 유효성을 심각하게 저해하는 요인이 된다. 즉, 기 제시된 공격 기법들을 한국어 모델에 그대로 적용할 경우, 정교한 형태론적 결합 규칙이 파괴되어 비문법적인 표현이 생성되거나 문장의 부자연스러움으로 인해 공격 의도가 쉽게 노출되는 등 유창성과 은닉성이 현저히 저하된다. 결과적으로 이는 공격 성공률과 실무적 위협 가능성을 모두 제한하는 한계로 작용한다.

이에 본 연구는 기존 공격 기법들이 간과했던 한국어의 교착어적 특성을 구조적으로 반영하면서, 학습 데이터에 대한 접근 및 오염과 추가적인 모델 재학습을 배제하고도 수행 가능한 2단계 계층적 공격 프레임워크를 제안한다. 제1단계 공격은 공개된 모델의 그라디언트 정보를 역추적하여, 입력 문장의 의미와 무관하게 모델의 결정 경계를 교란하는 트리거를 탐색한다. 제2단계 공격은 1단계 공격에도 예측이 반전되지 않은 견고한 잔여 샘플을 대상으로 수행한다. 여기서는 원문의 의미적 일관성을 유지하면서, 조사나 어미와 같이 문법적 의존성이 높은 구간을 동시에 처리하는 형태 보존 최소 편집을 적용하여 문법적 오류를 최소화한다. 제안하는 공격 기법의 유효성은 NSMC 데이터세트[13]를 기반으로 KoBERT[14] 및 KoELECTRA[15] 모델에 대한 엄격한 교차 평가를 통해 검증한다.

본 연구의 학문적 기여는 다음과 같이 요약할 수 있다. 우선 기존 데이터 오염 기반 공격 기법들이 전제했던 학습 데이터에 대한 접근 및 오염과 모델 재학습을 배제하고 공개된 그라디언트 정보만을 활용해 모델의 결정 경계를 효과적으로 교란할 수 있음을 입증하였다. 둘째, 적대적 공격 수행 시 한국어 문법 구조에 최적화된 교착어 동시 마스킹 전략을 도입함으로써, 필연적으로 발생하는 문법적 오류를 최소화하고 은닉성을 획기적으로 개선하였다. 마지막으로 이기종 모델에 대한 정량적·정성적 분석을 통해 한국어 기반 언어모델이 문장 끝 위치와 특정 간접 단서에 구조적으로 취약함을 실증적으로 규명함으로써, 향후 한국어 모델에 특화된 방어 전략 수립을 위한 실무적 토대를 마련하였다.

이후 논문의 구성은 다음과 같다. 2장에서는 언어모델 공격과 관련된 연구 동향을 살펴보고, 대표적인 언어모델에 대해 정리한다. 3장에서는 제안하는 공격 기법을 상세하게 설명하고, 4장에서는 제안 기법의 실효성을 검증하기 위한 실험 설계 및 환경에 대해 설명한다. 5장에서는 실험 결과를 분석하고, 마지막으로 6장에서 연구의 결론과 함께 한계점 및 향후 연구 방향을 논의한다.

II. Preliminaries

1. Related works

언어모델 대상의 적대적 공격과 관련된 초기 연구들은 대부분 중요도가 높은 단어를 식별한 후 토큰을 치환하여 모델 예측 결과를 반전시키는 전략을 채택하였다.

Alzantot 등[1]은 유전 알고리즘으로 단어별 중요도를 산출한 후, 가장 중요도가 높은 단어를 대상으로 동의어를 결정 경계가 바뀔 때까지 반복 교체하는 기법을 제시했다. 실험 결과 감성 분석 모델에 대해 97%의 공격 성공률을 달성하였다. Li 등[2]도 중요도가 높은 단어를 탐색한 뒤, 단어 치환, 오타 유발, 분할·삭제 등 공격 유형을 선택하여 텍스트 분류 및 감성 분석 모델을 교란하는 접근법을 제시하였다. Jin 등[3]은 BERT 모델[16]의 예측 결과에 가장 큰 영향을 미치는 단어를 선정한 후 의미와 문법적 조건을 만족하는 동의어로 교체하는 전략을 채택하였으며, 감성 평가 데이터세트에 대해 높은 공격 성공률을 달성하였다.

Li 등[4]은 중요 단어의 위치를 마스킹한 후, BERT의 확률 분포상 원문과 의미적으로 가장 근접한 후보 단어를 선정하였다. 이러한 과정을 반복하면서 자연스러운 공격을 실시해 BERT 모델의 예측 정확도 하락에 성공하였다. Garg 등[5]은 텍스트를 마스킹한 뒤 BERT 등의 언어모델을 활용하여 해당 위치의 단어를 대체 혹은 삽입 단어를 생성하는 공격 기법을 제안하였다.

이러한 초기 연구들과 달리 최근 연구들은 문맥에 맞춰 문법과 의미의 자연스러움을 극대화하는 전략을 채택하고 있다. Rocamora 등[6]은 기존 단어 수준 공격이 문맥의 의미를 변화시킬 수 있지만 문자 수준 공격은 의미 보존이 용이하다는 점에 주목하고, 문자 수준에서 적대적 공격을 수행하는 기법인 Charmer를 제안하였다. BERT 모델의 SST-2 데이터세트를 대상으로 실험을 진행한 결과, 공격 성공률과 의미 유사성이 모두 향상됨을 확인하였다. Gao 등[7]은 강화학습을 기반으로 의미를 보존하면서도 높은 공격 성공률을 달성할 수 있는 적대적 예제 생성 기법을 제안하였다. BERT 기반 모델을 대상으로 실험을 진행한 결과 자동 평가와 인간 평가 모두에서 기존 공격 기법들 대비 유의미하게 높은 의미 유사성을 유지하면서도 동등 이상의 공격 성공률을 달성하였다.

데이터 오염 기반 공격을 제안한 대부분의 선행 연구는 학습 데이터에 대한 접근을 통해 오분류를 유도하는 트리거가 포함된 오염 데이터를 모델이 재학습하는 전략을 채택하고 있다. Wallace 등[8]은 입력 텍스트와는 무관한 임의 트리거를 추가하여 모델의 전체적인 예측 결과가 뒤바뀌도록 유도하는 공격 기법을 제안하였다. 트리거 최적화를 위해서는 그라디언트 기반 탐색을 활용하였으며, 모델이 트리거에 지나치게 민감하게 반응하도록 유도하였다. Qi 등[9]은 특정 단어나 문장을 트리거로 삽입하는 대신 구문 구조 자체를 트리거로 활용하는 공격 기법을 최초로 제안하였다. 학습 데이터에 포함되는 오염 샘플은 구문 탐색을 기반으로 정상 샘플을 특정한 구문의 템플릿에 맞춰 재구성해 생성했다.

보다 최근의 연구들은 고정된 트리거를 계속 사용하지 않고 반복적인 최적화 과정을 통해 자연스러운 트리거를 식별함으로써 은닉성을 극대화하는 전략을 채택하고 있다. 먼저 중요 단어를 식별한 후 자연스러운 동의어나 관련 단어로 치환해 오염 샘플을 생성했으며, 이런 과정을 여러 번 반복해 공격 효과를 극대화하였다. Yan 등[10]은 자연스러운 단어 수준에서 다수의 트리거 단어를 점진적으로 오염 샘플에 주입하면서 목표 라벨과 강력한 상관관계를 형성하는 방식으로 은닉성을 유지하는 기법을 제안하였다. Li 등[11]은 기존의 고정 단어나 구문과 같은 트리거 기반의 접근법과 달리, 생성형 언어모델의 조건부 확률 분포를 암묵적 트리거로 활용해 자연스러운 텍스트 구문을 생성하여 오염 샘플로 활용하였다. 실험 결과 4개 감정 분류 데이터세트에서 평균 97.35%의 공격 성공률을 달성하면서도 가장 낮은 문장 혼란도, 최소 문법 오류, 가장 높은 문법 수용률을 기록하여 높은 은닉성을 입증했다. Zhao 등[12]은 파인튜닝(fine tunning) 없이 오염된 샘플 또는 프롬프트를 주입할 수 있는 데이터 오염 기반 공격 기법을 제안하였다. 이 방법은 사전학습 모델의 일반성을 유지하면서 거대 언어 모델에 트리거를 주입할 수 있으며, GPT 모델에서는 평균 95.0%의 공격 성공률을 달성하였다. 특히 1.3B부터 180B 매개변수 모델까지 확장성을 입증해 다양한 거대 언어 모델에 대한 실질적 위협을 제시했다.

이상에서 살펴본 바와 같이 텍스트 분야에 대한 적대적 공격 및 데이터 오염 기반 공격은 주로 영어권 데이터세트와 모델을 대상으로 진행되었다. 이러한 공격 기법들을 한국어 기반 언어모델에 그대로 적용하면 명사-조사, 어간-어미 간의 결합 형태로 인한 의미 왜곡과 유창성 저하 등이 빈번히 발생한다. 따라서 본 연구에서는 한국어의 고유한 특성을 고려한 공격 기법을 제안한다.

2. BERT

BERT는 트랜스포머 인코더에 기반하여 양방향 문맥을 동시에 이해하는 사전학습 모델로서, Masked Language Modeling (MLM)과 Next Sentence Prediction (NSP)에 기반해 학습을 수행한다. MLM은 입력에서 일부 토큰을 무작위로 마스킹하고, 주변 단어들의 맥락을 이용해 마스킹된 토큰을 예측한다. NSP는 두 문장 간의 관계를 이해하고 전·후를 파악하기 위해 문장과 그 다음 문장을 예측한다[16].

3. ELECTRA

ELECTRA[17]는 Replaced Token Detection (RTD)을 도입했다. 생성기가 일부 토큰에 대해 그럴듯한 대체 토큰을 생성해 입력하면, 판별기가 각각의 토큰이 원본인지 대체 토큰인지를 판별하여 학습한다. 이 방식은 문장 내 모든 토큰 위치에 대한 정보를 가지고 있어 치환이나 삽입 등과 같은 미세 편집에 민감하게 반응한다.

III. The Proposed Method

1. Overview

본 연구에서는 한국어 문장 분류기를 대상으로, 한국어 고유의 교착어적 특성을 반영해 공격의 은닉성과 문장의 자연스러움을 유지하면서 높은 성공률을 달성하는 2단계 계층적 공격 프레임워크를 제안한다. Fig. 1은 제안하는 프레임워크의 전체적인 공격 진행 흐름을 보여준다.

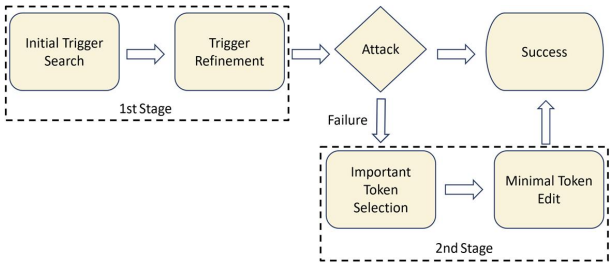

Fig. 1. Overview of the proposed attack framework

제안하는 2단계 공격 기법은 모델의 재학습이나 데이터 오염을 전제로 하는 기존의 공격 기법들과 명확히 대비된다. 본 기법은 학습 단계에 대한 개입을 철저히 배제함으로써, 실제 환경에서의 공격 실현 가능성을 극대화하는 데 주안점을 두었다. 또한, 고비용의 '모델 재학습' 과정을 원천적으로 제거하여 연산 효율성을 구조적으로 개선하였다. 일반적으로 전체 모델 파라미터를 업데이트하는 재학습 과정은 데이터 규모와 학습 횟수에 비례하여 막대한 시간이 소요된다. 반면, 제안 기법은 고정된 모델(Frozen Model) 하에서 입력 토큰의 임베딩만을 최적화하는 방식을 취한다. 이를 통해 공격 준비 시간을 이론적으로 수 분 이내로 획기적으로 단축할 수 있다.

전체 공격 과정은 상호보완적 관계를 가지는 두 단계로 구분된다. 우선 1단계 공격에서는 학습 데이터와 독립적인 테스트 데이터만을 대상으로 모델의 그라디언트 정보를 활용하여, 입력 문장의 특정 위치(접두 또는 접미)에 부착되었을 때 모델의 예측을 오분류로 편향시키는 단일 토큰 트리거를 정밀하게 탐색한다. 이어지는 2단계 공격에서는 1단계의 트리거 공격에도 불구하고 예측이 반전되지 않은 견고한 잔여 샘플들을 대상으로 공격을 수행한다. 이 과정에서는 1단계에서 식별된 트리거를 제외하고 모델의 최종 판단에 지대한 영향을 미치는 핵심 토큰을 식별한 뒤, 그 중 1~2개의 소수 토큰을 문맥적으로 유사한 단어로 교체하는 최소 편집을 적용한다.

2. Threat model and problem definition

본 연구에서 상정하는 위협 모델은 현실적 화이트박스 시나리오이다. 이는 공격자가 KoBERT나 KoELECTRA와 같이 공개된 사전학습 언어모델을 다운로드하여 사용하는 환경을 반영한 설정으로, 모델의 임베딩 공간과 추론 단계에서 산출되는 그라디언트 정보에 접근할 수 있다고 전제한다. 단, 실제 서비스 운영 보안을 고려하여 모델 학습에 사용된 학습 데이터에는 접근하거나 오염시킬 수 없다는 제약 조건을 추가하여, 데이터 오염을 전제로 하는 기존 데이터 오염 기반 공격과의 차별성을 확보하였다.

공격의 핵심 목표는 입력될 문장에 최소한의 변형만을 가하여, 인간이 인지하는 원본 문장의 의미는 유지하면서 표적 모델이 오분류하도록 유도하는 샘플을 생성하는 것이다. 이때, 한국어의 교착어적 특성을 고려하여 문장의 문법적 정합성과 의미상의 자연스러움을 보존하는 것을 필수 요구사항으로 설정한다. 이를 위해 편집 토큰 수를 엄격하게 제한하여, 1단계 전역 공격에서는 하나의 토큰으로 구성된 트리거만을 삽입하고, 2단계 정밀 공격에서는 핵심 토큰의 치환을 최대 2개 이하로 한정한다.

3. Universal single-token trigger search

본 단계의 목표는 학습 데이터로 학습이 완료되어 파라미터가 고정된 모델을 대상으로, 테스트 데이터만을 활용하여 입력 문장의 내용과 무관하게 예측을 타겟 레이블로 편향시키는 범용 트리거를 탐색하는 것이다. 우선, 생성된 트리거가 한국어 문장 내에서 자연스럽게 결합될 수 있도록 후보 토큰을 온전한 단어 형태로 제한한다. 이를 위해 서브워드 분절 과정에서 단어의 중간 조각, 숫자, 영문 및 특수 기호를 배제하고, 문법적 자립성이 있는 단어의 시작 조각만을 탐색 대상인 후보 집합으로 설정한다.

트리거 탐색은 초기 후보 선정과 그라디언트 기반 정제로 구분하여 진행된다. 먼저, 방대한 탐색 공간을 효율적으로 축소하고 최적화의 수렴 속도를 가속화하기 위해 전체 테스트 데이터의 10%를 무작위로 샘플링하여 초기 탐색 집합을 구성한다. 이어서 구성된 초기 탐색 집합에 대해 Pointwise Mutual Information (PMI)을 적용하여 초기 후보군을 선별한다. 특정 토큰 t와 타겟 레이블 y_{target} 간의 PMI는 수식 (1)과 같이 정의된다.

$$PMI(t, y_{target}) = \log \frac{P(t \mid y_{target})}{P(t)} \quad (1)$$

여기서 $P(t \mid y_{target})$은 타겟 레이블이 부여된 문장 집합 내에서 토큰 t가 등장할 조건부 확률을, $P(t)$는 전체 말

뭉치에서의 등장 확률을 나타낸다. PMI 값이 클수록 해당 토큰은 타겟 레이블과 강한 상관관계를 가짐을 의미한다. 이를 통해 PMI 상위 토큰 8개를 앞서 샘플링한 10%의 탐색용 데이터에 부착하여 오분류 유발력을 검증하고, 가장 높은 성능을 보인 토큰을 초기 트리거 t_0로 선정한다.

이어서 선정된 초기 트리거를 기점으로, 모델의 손실 함수를 최소화하는 최적의 트리거를 찾기 위해 반복적인 정제를 수행한다. 목적함수 L은 수식 (2)와 같이 교차 엔트로피 손실로 정의된다.

$$\min_{t} L(x \oplus t, y_{target}) = -\log P(y_{target} | x \oplus t) \quad (2)$$

여기서 x는 입력 문장, t는 트리거 토큰, $\oplus$는 결합 연산을 의미하며, P는 모델이 예측한 타겟 클래스의 소프트맥스 확률이다. 텍스트는 이산적 데이터이므로 미분을 통한 직접적인 역전파가 불가능하다. 따라서 본 연구는 임베딩 공간에서의 근사를 통해 손실을 줄이는 방향에 있는 토큰을 탐색한다.

구체적으로, 현재 단계의 트리거 t_n에 대한 임베딩 벡터를 e_{t_n}이라 할 때, 손실 함수 L의 그라디언트 $\nabla_{e_{t_n}} L$을 계산한다. 이후 다음 수식 (3)과 같이 손실을 가장 가파르게 감소시키는 방향($-\nabla_{e_{t_n}} L$)과 임베딩 공간 내 후보 토큰들 $v \in V$ 간의 연관성을 계산한다.

$$t_{n+1} = \underset{v \in V}{\operatorname{argmin}} (e_v \cdot \nabla L) \quad (3)$$

수식 (3)은 기하학적으로 후보 토큰의 임베딩 벡터 e_v를 손실이 감소되는 방향 $-\nabla L$에 투영하는 것과 유사하며, 내적값이 작을수록 즉, 음의 방향으로 클수록 해당 토큰이 손실을 줄이는 방향과 강력하게 정렬되었음을 의미한다. 이에 따라 내적값이 가장 작은 상위 128개의 토큰을 후보군으로 추출한다. 이어서 추출된 후보군을 공격 대상 데이터에 대입하여 손실값의 변화를 평가하고, 손실을 최소화하는 토큰으로 트리거를 업데이트한다. 이 과정은 손실이 수렴하거나 최대 반복 횟수에 도달할 때까지 지속된다. 단, 공격의 은닉성을 보장하기 위해 '최고', '최악', '추천' 등 의미를 노골적으로 변경할 수 있는 단어들은 금지 목록으로 지정하여 후보 선정 단계에서 원천적으로 배제한다.

4. Morphology-Preserving Minimal Edit

본 단계는 1단계의 단일 토큰 트리거 공격에서 예측이 반전되지 않은 견고한 잔여 샘플만을 대상으로 수행되며, 핵심 목표는 문장 내의 중요 토큰들을 식별하여 최소 편집만으로 모델의 결정 경계를 넘어서는 것이다. 이때, 원본 문장의 의미적 일관성을 유지하기 위해 편집 범위는 최대 2개 토큰으로 엄격하게 제한한다. 공격은 중요 토큰 식별, 형태 보존을 적용한 후보 토큰 생성, 최적 대체안 선정의 순서로 진행된다.

우선, 최소 편집의 대상이 되는 중요 토큰을 식별하기 위해서 삭제 기반 중요도 산출 기법을 적용한다. 문장 내 토큰을 순차적으로 하나씩 삭제한 뒤 표적 모델에 입력하면서 원본 문장 대비 타겟 클래스의 예측 확률 변화량을 측정한다. 변화량이 클수록 모델의 판단에 지대한 영향을 미치는 토큰이므로 이를 중요도 순으로 정렬한다. 이때, 공격의 은닉성을 저해할 수 있는 금지 목록 단어와 1단계에서 부착했던 트리거는 정렬 목록에서 제외한다.

중요 토큰이 식별된 이후에는 MLM을 활용하여 문맥에 적합한 치환 후보를 생성한다. 이 과정에서는 한국어의 교착어적 특성을 반영한 결합어 동시 마스킹 전략을 적용한다. 만약 중요 토큰이 조사, 어미, 혹은 부정부사('안/못')와 같이 문법적 의존성이 강한 요소와 결합되어 있다면, 해당 토큰만 단독으로 변경하는 경우 문법적 오류가 발생할 수 있다. 예를 들어, '영화를 추천'이라는 문장에서 '영화'가 중요 토큰으로 선정되어 '작품'으로 수정한다면, '작품를'이라는 비문법적인 표현이 생성되므로 '~를'도 함께 변경해야 한다. 따라서 본 연구는 이러한 의존 관계에 있는 이웃 토큰을 묶어 동시에 마스킹한다. 이후 MLM을 통해 각 위치별로 예측된 상위 8개의 토큰을 조합하여 총 64개의 후보를 생성하고, 이 중 문맥 확률이 가장 높은 8개의 후보 조합을 최종 선별한다.

마지막으로, 생성된 후보 조합들을 표적 모델에 입력해 공격 성공 여부를 검증한다. 후보들 중 타겟 레이블로의 오분류를 유도하면서 동시에 타겟 클래스의 예측값을 가장 크게 상승시키는 최적의 조합을 최종 샘플로 채택한다. 만약 최상위 중요 토큰으로 공격에 실패할 경우, 차순위 중요 토큰에 대해 동일한 절차를 반복 수행한다.

IV. Experimental setup

1. Dataset

NSMC 데이터세트는 네이버 영화 리뷰를 통해 수집된 한국어 감성 분류 말뭉치로서, 긍정과 부정의 이진분류를 위해 설계된 표준 벤치마크 데이터세트이다. 학습 데이터세트 150,000개 및 테스트 데이터세트 50,000개를 포함 총 20만 개의 문장으로 구성되어 있다. 주로 구어체 문장 중심으로 구성되어 있지만 일상적 표현이나 속어, 축약 및

오탈자 등의 한국어 웹 텍스트도 폭넓게 포함하고 있다. 즉 현실적인 웹 도메인의 잡음을 충분하게 포함하고 있어 어휘와 형태 변동에 민감한 한국어를 대상으로 하는 공격 시나리오를 충분히 보여줄 수 있다. 이에 본 연구의 실험에서는 NSMC 데이터세트 하나만을 사용하였다. 그리고 동일 데이터세트에 대해 구조가 다른 두 모델 KoBERT 및 KoELECTRA로 교차 평가를 수행하여 데이터세트 고유의 편향보다는 모델 구조 차이에 따른 견고성 차이를 보다 더 명확하게 관찰할 수 있도록 하였다.

안정적인 학습과 모델에의 입력을 위한 데이터 전처리 과정은 다음과 같다. 먼저, 제어문자, 이메일, HTML 태그, URL 및 이모지 등을 제거하고, '!!!!'과 같은 과도한 문장부호나 'ㅋㅋ/ㅎㅎ/ㅠㅠ' 류의 반복되는 감정 표현을 축약하였다. 또한, 과도하게 짧은 문장을 제거하기 위해 띄어쓰기 2칸 이상, 4단어 이상인 문장만을 샘플링하였다. 그 결과 학습 데이터세트는 108,708개, 테스트 데이터세트는 36,344개가 추출되었다.

2. Target model

2.1 KoBERT

KoBERT는 BERT-base 모델을 기반으로 한국어 말뭉치를 사전 학습한 모델로서, [MASK] 토큰을 통해 문맥에 어떤 단어가 자연스러운가를 학습한다. 문장을 아주 작은 조각으로 쪼개는 대신, SentencePiece 기반의 서브워드 토크나이저를 사용한다. SentencePiece는 단어 앞에 '_'가 붙어 경계를 명시하며, '명사+조사'나 '어간+어미'처럼 결합이 잦은 한국어에서도 비교적 안정적인 단위를 형성한다. 이러한 단어 경계가 명확하다는 특성으로 인해 제안하는 공격 기법에서 중요 토큰을 수정할 때 결합된 부분을 함께 편집하기 수월하다.

2.2 KoELECTRA

KoELECTRA는 ELECTRA를 기반으로 RTD 방식으로 사전 학습을 실시한 한국어 모델이다. 문장 토큰의 일부를 다른 토큰으로 교체한 뒤 각 위치가 원본인지 대체본인지 위치별로 판별하도록 학습한다. 따라서 1~2개 토큰만을 미세하게 수정하더라도 모델이 민감하게 반응한다. 또한, WordPiece 방식을 사용하여 단어를 '재미, ##있, ##다'처럼 조각으로 쪼개기 때문에, 한 글자만 교체해도 비문이 쉽게 발생한다. 이런 한국어 결합 특성을 반영하기 위해 실험에서는 붙어있는 조각들을 동시에 수정하였다.

3. Experimental configuration

실험 환경은 Nvidia RTX 3090 그래픽 카드를 활용하였고, Python 3.10 환경에서 PyTorch, Transformers 라이브러리를 사용하였다. 그리고 공격자가 모델 학습에 사용된 데이터에는 접근할 수 없는 현실적인 보안 위협을 모사하기 위해, 데이터세트의 활용 범위를 엄격히 분리하였다. 구체적으로, 표적 모델의 미세조정 과정에는 학습 데이터만 사용했으며, 공격 과정에서는 모델 학습에 전혀 관여하지 않은 독립적인 테스트 데이터만을 대상으로 수행되었다. 이는 공격자가 훈련 과정에 개입하거나 데이터를 오염시키지 않고, 배포된 모델이 마주할 미지의 입력에 대해서도 효과적인 공격이 가능함을 실증하기 위함이다.

학습 간 각 모델에 적용한 하이퍼파라미터와 모델별로 감성 분류 성능을 측정한 초기 실험 결과는 Table 1에서 보는 바와 같다. KoBERT 모델은 5 epochs에서 0.89의 정확도를, KoELECTRA 모델은 8 epochs에서 0.905의 정확도를 달성하였다.

Table 1. Hyperparameter Setting and Initial Results

	KoBERT	KoELECTRA
Optimizer	AdamW	
Batch size	32	
Max length	256	
Epochs	5	8
Learning rate	5e-5	3e-5
Accuracy	0.89	0.905

V. Experimental results

트리거 위치에 따른 공격 성공률(Attack success rate, ASR)과 선정된 트리거는 Table 2에서 보는 바와 같다.

실험 결과, 트리거를 문장의 뒤에 배치했을 때가 앞에 배치했을 때보다 최종 공격 성공률이 일관되게 높게 나타났다. 또한, 모델 구조에 따라 초기와 최종 공격 성공률의 양상이 상이하게 나타났는데, 이를 통해 공격 대상 모델의 사전학습 방식과 토크나이저(tokenizer)의 특성이 공격에 대한 민감도에 결정적인 영향을 미침을 알 수 있다.

KoBERT는 긍정 타겟에 '훌륭'을 접미사로 부착했을 때 0.963이라는 최고 ASR을 달성했으며, KoELECTRA 역시 동일 조건에서 '명곡'을 문장 뒤에 부착했을 때 0.940의 최대 성능을 보였다. 이러한 접미 부착의 높은 공격 성공률은 의미적 초점이 문장의 끝에 집중되는 한국어 고유의

Table 2. Experimental Results

Model	Target	Position	1 stage ASR	Final ASR	Trigger
KoBERT	Positive → Negative	Prefix	0.645	0.860	최저
		Suffix	0.640	0.882	폭행
	Negative → Positive	Prefix	0.547	0.932	짱
		Suffix	0.542	**0.963**	훌륭
KoELECTRA	Positive → Negative	Prefix	0.728	0.907	마지못해
		Suffix	0.693	0.912	무용지물
	Negative → Positive	Prefix	0.527	0.850	알라딘
		Suffix	0.652	**0.940**	명곡

특성에 기인한다. 특히 한국어 감성 분석에서는 주로 문장 말미의 서술어 또는 종결 어미가 화자의 의도를 결정짓는 핵심 단서로 작용한다. 따라서 공격자가 문장 말미에 강력한 감성 트리거를 부착할 경우 모델의 어텐션(Attention) 메커니즘이 최신성 편향을 일으켜 앞선 문맥 정보를 효과적으로 덮어쓰는 현상이 발생할 수 있다.

선정된 트리거의 의미적 특성 측면에서는, '최고', '최악' 등과 같은 직설적 감성어를 배제했음에도 불구하고 간접적 단서들이 효과적인 공격 수단으로 작용하였다. 특히 '알라딘' 트리거는 학습 데이터 내에서 긍정적인 레이블과 빈번하게 결합되어 나타난 단어이기 때문에, 모델이 해당 단어 자체를 긍정적 신호로 인식하고 학습하였다. 이러한 데이터 편향성은 공격자가 의도하지 않은 단어로도 언어 모델을 손쉽게 교란할 수 있는 보안 취약점으로 작용한다.

1단계 공격 성공률과 최종 공격 성공률의 비교를 통해 모델 아키텍처에 따른 공격 민감도 차이도 명확하게 식별할 수 있다. 우선, RTD 방식으로 학습된 KoELECTRA는 1단계 공격부터 상대적으로 높은 민감도를 보였다. 예를 들어 문장 앞, 부정 타겟에서 1단계 공격 성공률이 이미 0.728을 기록했는데, 이는 생성된 가짜 토큰을 식별하도록 훈련된 판별기가 예민하게 반응하기 때문이다.

반면, MLM 기반의 KoBERT는 1단계에서 상대적으로 낮은 공격 성공률을 보였다. 이러한 결과는 MLM이 문맥 복원 능력을 통해 단일 토큰 노이즈를 어느 정도 상쇄함을 의미한다. 그러나 2단계 정밀 공격 이후에는 긍정 타겟, 문장 뒤 트리거는 1단계 0.542에서 최종 0.963으로 약 0.42p 급증해 KoELECTRA를 능가하거나 대등한 수준에 도달했다. 이는 1단계 공격으로 모델의 방어 기제를 약화시키고, 2단계 형태 보존 최소 편집 공격을 통해 결정적인 타격을 가하는 계층적 공격이 MLM 기반 모델에서 좀 더 시너지 효과를 발휘할 수 있음을 보여준다.

Table 3은 제안한 기법이 효과적으로 작동한 대표적인 성공 사례를 보여준다. 다수 샘플에서 접두/접미 트리거를 부착한 1단계 공격만으로도 예측 결과가 반전되었다. 부정 타겟의 경우 '최저', '폭행' 등과 같은 결정적 부정 단서가, 긍정 타겟의 경우 '명곡', '알라딘'과 같은 간접적 긍정 단서가 문맥 전체의 감성을 지배하는 양상을 보였다. 또한, 트리거만으로 실패한 견고한 샘플에 대해서는 2단계 정밀 공격이 작동하여 핵심 토큰을 치환함으로써 공격을 성공시켰다. 특히, 결합어 동시 마스킹 전략이 유효하게 작동한 경우, 명사와 조사 혹은 어간과 어미가 문법적으로 올바르게 동시 교체되어 원문의 유창성을 유지하면서도 모델의 판단을 효과적으로 유도하였다.

이러한 실험 과정에서 식별된 '알라딘', '명곡'과 같은 결정적 트리거의 작동 기제는 딥러닝 모델의 취약점과 관련하여 두 가지 측면에서 해석된다. 첫째는 데이터세트 편

Table 3. Examples of Successive Attack

Original sample	Adversarial sample
황홀해 보는 내내 아득하고 황홀하더라	**최저 영화를** 보는 내내 아득하고 황홀하더라
너무 귀엽다 제작진 고맙습니다	너무 **좋은** 제작진 고맙습니다 **폭행**
이거 진짜 무슨 내용인지 모르겠다	**짱** 이거 진짜 무슨 **짓이야** 모르겠다
정말 최악 극장에서 졸면서 봤던 영화	정말 **오랜만에** 극장에서 졸면서 봤던 영화 **훌륭**
이 영화를 다시보고싶은데.. 다운로드가 안되나요?	**마지못해** 이 영화를 다시보고싶은데.. 다운로드가 안되나요?
떠남과 위안에 관한 영화, 강추!	떠남과 위안에 관한 영화, 강추! **무용지물**
이게 그 미생 깽판쳐놓은 인간 작품인가	이게 그 미생 깽판쳐놓은 인간 작품인가 **알라딘**
미취학 아동 이하만 보세요 뽀로로 수준	미취학 아동 **수준**으로 보세요 뽀로로 수준 **명곡**

향에 기인한 지름길 학습[18]이다. ‘알라딘’과 같은 특정 고유명사가 트리거로 작용한 현상은 모델이 문맥을 깊이 이해하기보다 데이터 내의 허위 상관관계와 같은 통계적 지름길에 의존하고 있음을 보여준다. 둘째는 도메인 특화된 의미적 맥락의 반영이다. ‘명곡’과 같은 단어는 영화 리뷰 도메인에서 강력한 긍정 극성을 내포하고 있어, 모델의 의미론적 판단 기준을 흔드는 핵심 요소로 작용하였다. 결론적으로 제안 기법은 모델이 과도하게 의존하는 데이터의 통계적 편향과 의미적 맥락이라는 두 가지 취약점을 동시에 자동 탐지하고 공략함으로써 높은 공격 성공률을 달성했음을 알 수 있다.

반면, 공격은 성공했으나 생성된 문장의 언어적 품질이 크게 저하된 사례는 Table 4에서 확인할 수 있다. 이러한 유창성 저하는 주로 형태론적 불일치에서 발생한다. 즉, 중요 토큰이 교체될 때 이에 종속된 조사나 어미가 문법적 호응을 이루도록 함께 수정되어야 하나, 일부 복잡한 문장 구조에서는 동시 교체가 완벽하게 작동되지 않아 단절된 교체가 발생한 것이다. 이는 한국어의 교착어적 특성을 반영한 알고리즘을 적용했음에도 불구하고, MLM이 예측한 토큰 조합이 미세한 문법적 뉘앙스나 불규칙 활용을 완벽하게 포착하지 못할 수 있음을 의미하며, 향후 문법 교정 모듈의 고도화가 필요함을 보여준다.

공격에 실패한 사례를 유형별로 정리한 결과는 Table 5에서 확인할 수 있다. 실패 원인은 크게 두 가지 요인으로 범주화할 수 있다. 첫째, 견고한 의미적 문장들이다. 원문 내에 ‘좋은’, ‘따뜻한’, ‘재미있는’ 등의 강한 감성 단서가 다수 분포되어 있는 경우, 단일 트리거 삽입과 최대 2개의 토큰 교체라는 제한된 편집만으로는 문장 전체에 내포된 강한 정서를 역전시키는 것이 쉽지 않다. 그리고 이러한 결과는 공격의 은닉성을 위해 편집량을 제한한 설계상의 한계에 기인한다. 둘째, 문법적 붕괴로 인한 모델의 예측 불안정성이다. ‘잼하게’, ‘집중게’와 같은 비문법적 토큰이 생성되거나 연쇄적인 형태소 오류가 발생할 경우, 모델이 이를 의미 있는 텍스트로 처리하지 못하면서 예측 확률이 특정 방향으로 수렴하지 않는 현상이 관찰되었다. 이러한 모델의 불안정성은 공격자가 의도한 방향성을 상실하게 만드는 요인으로 작용한다.

마지막으로, 제안한 방법의 효용성을 객관적으로 검증하기 위해, 한국어의 형태론적 특성을 고려하지 않는 일반적인 중요도 기반 공격 방식을 대조군으로 설정하여 비교 실험을 수행하였다. 현재 한국어 텍스트에 특화된 표준화된 적대적 공격 벤치마크가 부재한 점을 고려하여, 본 실험에서는 영어권의 대표적인 공격 기법인 Jin 등[3]의 기법을 한국어에 단순 적용한 방식을 대조군으로 정의하였다. 이 방식은 형태소 분석 과정을 배제하고, 모델의 예측에 중요한 토큰을 문맥적으로 유사한 단어로 1:1 치환하는 전

Table 4. Examples of Successive Success with Fluency Errors

Original sample	Adversarial sample
재미있음 어릴때 짱 좋아했었는	**최저 내가** 어릴때 짱 좋아했었**은**
어렸을 적 로망이였던 명작	어렸을 적 로망이였던 **영화는 폭행**
좋은 소재임 근데 …	**짱** 좋은 **영화이데** 근데 …
초중반은 괜찮던데 후반부 … 질질 끄는 거야?	**초반부진** 괜찮던데 후반부 … 질질 끄는 거야? **훌륭**
보진 못했구 교과서 책만 읽었는데 잼있을꺼 …	**마지못해** 보진 못했구.. 교과서 책만 **읽해도** 잼있을꺼 …
정말 재밌게 봤다. 매력있는 영화	정말 **재미있에** 봤다. 매력있**은** 영화 **무용지물**
세계인가 뭘 보여주려고 하는지 모르겠다	세계인가 **라고 말하고** 하는지 모르겠다 **알라딘**
너무 노골적인 종교 판타지	너무 **웃기한** 종교 판타지 **명곡**

Table 5. Examples of Failed Attack

Original sample	Adversarial sample
참 좋은 영화 따뜻하고 재미있고 슬프고 그런 영화	**최저** 참 좋은 영화 따뜻하고 재미있고, **따뜻은** 그런 영화
… 전 이거 진짜 재미있게 봤어요	… 전 이거 진짜 **잼하게** 봤어요 **폭행**
아무리 코미디라도 하나같이 …	**짱** 아무리 **봐도봐도** 하나같이 …
일본영화 그만 나오라 돈 아깝다	일본영화 **꼭** 나오라 **ㅋㅋㅋ라** 아깝다 **훌륭**
허무하게 끝나는 점도 있긴하지만 재미있었어요	**마지못해** 허무하게 **흘러는 영화이기** 있긴하지만 재미있었어요
… 내 스타일이라서 더 재밌게 본것같다	… 내 스타일이라서 더 **집중게 봤다 무용지물**
내공이 없이 폼을 잡는데 딱 허세부리는 허수아비영화	내공이 **이렇게** 폼을 잡는데 딱 **좋았은** 허수아비영화 **알라딘**
머임 … 이렇게 재미없는 영화는 첨봄	**친구들** 이렇게 재미없은 영화는 첨봄 **명곡**

략을 취한다. 생성된 적대적 예제의 언어적 품질을 평가하기 위해, Perplexity (PPL)와 Universal Sentence Encoder (USE) 유사도[19]를 정량적 지표로 도입하였다. PPL은 KoGPT2 모델[20]을 사용하여 문장의 유창성을 측정하였으며 수치가 낮을수록 자연스러운 문장임을 의미한다. USE 유사도는 원본 문장과 공격 문장 간의 임베딩 코사인 유사도를 통해 의미 보존성을 평가하였다.

Table 6는 대조군과 제안 기법의 성능 비교 결과를 보여준다.

Table 6. Quantitative comparison with baseline attack method

Method	ASR	PPL	USE
Original	-	227.86	1
Baseline	0.838	$> 10^5$	0.333
Ours	0.963	301.85	0.961

실험 결과, 대조 기법은 0.838의 준수한 공격 성공률을 기록하였으나, 생성된 문장의 품질 측면에서 심각한 한계를 드러냈다. 특히 대조군의 PPL은 측정 한계를 초과하여 발산($>10^5$)하였으며, USE 유사도 또한 0.333에 불과하였다. 이는 교착어인 한국어의 특성상, 문맥에 맞게 조사와 어미를 수정하지 않고 체언이나 용언만을 교체할 경우 문법적 구조가 파괴되고 의미가 크게 훼손됨을 시사한다. 반면, 본 연구에서 제안한 기법은 0.963의 가장 높은 공격 성공률을 달성함과 동시에, PPL 301.85를 기록하여 원본 문장과 근접한 수준의 문법적 자연스러움을 유지하였다. 또한 USE 유사도 역시 0.961로 매우 높게 나타나, 공격 수행 과정에서 원문의 의미 정보 손실을 최소화하였음을 확인하였다. 결론적으로, 제안 기법은 높은 공격 성능과 더불어 생성된 문장의 언어적 품질까지 확보한 유효한 공격 프레임워크임을 입증하였다.

VI. Conclusions

본 연구는 한국어 언어모델을 대상으로, 학습 데이터에 대한 접근 없이 수행되는 단일 토큰 범용 트리거 탐색과 형태 보존 최소 편집을 결합한 2단계 계층적 공격 프레임워크를 제안하였다. 제안된 기법의 효용성은 구어체, 속어, 오탈자 등 실무적 노이즈가 포함된 NSMC 데이터세트를 기반으로, 대표적인 한국어 사전학습 모델인 KoBERT와 KoELECTRA에 대한 교차 평가를 통해 검증하였다.

실험 결과, 한국어의 핵심 정보가 문장 뒤에 오는 구조적 특성으로 인해 접미에 트리거를 부착했을 때 일관되게 높은 공격 성공률이 관찰되었다. 또한, 직설적 감성어를 배제했음에도 '짱', '명곡', '알라딘' 등 데이터 내의 감성 경향을 반영한 간접적 단서가 강력한 트리거로 작동함을 확인하였다.

평가 대상 모델 중 KoBERT는 '훌륭'이 접미 트리거 조건에서 0.963, KoELECTRA는 '명곡'이 접미 트리거 조건에서 0.940의 높은 최종 공격 성공률을 달성하였다. 특히 1단계 공격에 견고한 잔여 샘플에 대해서는 조사와 어미를 포함한 결합어 동시 마스킹을 적용하여, 문법적 유창성을 최대한 보존하면서도 모델의 방어선을 효과적으로 무력화할 수 있음을 입증하였다.

그러나 본 연구는 모델의 그라디언트 정보를 활용하는 '현실적 화이트박스' 위협 모델을 가정하고 있어, 모델의 내부 정보가 완전히 차단된 블랙박스 환경에 대한 일반화에는 한계가 있다. 또한, 형태소 결합 규칙을 적용했음에도 불구하고 복잡한 문장 구조에서는 유창성이 저하되는 사례가 일부 관찰되어, 향후에 의미적 정합성을 강화하기 위한 추가적인 품질 보정 연구가 필요하다.

향후 연구로는 그라디언트 정보에 대한 접근 없이 출력 확률만으로 작동하는 블랙박스 공격 기법으로의 확장과 다중 분류 데이터세트 및 최신 생성형 언어모델에 대한 취약점 평가를 수행할 예정이다. 나아가 본 연구에서 규명한 접미 의존성과 형태소 결합 취약점을 기반으로, 이를 실시간으로 탐지하고 방어할 수 있는 한국어 기반 언어모델에 특화된 입력 검증 기법과 적대적 공격에 대한 강건성 향상 기법을 연구하고자 한다.

REFERENCES

[1] M. Alzantot, Y. Sharma, A. Elgohary, B. Ho, M. B. Srivastava, and K. Chang, "Generating Natural Language Adversarial Examples," Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2890–2896, Oct. 2018. DOI: 10.18653/v1/D18-1316

[2] J. Li, S.-M. Li, X. Qiu, C. Wu, B. Xuan, W. Zhou, and Q. Liu, "TextBugger: Generating Adversarial Text Against Real-world Applications," Proceedings of the 2019 Network and Distributed System Security Symposium (NDSS), pp. 1–15, Feb. 2019. DOI: 10.14722/ndss.2019.23138

[3] X. Jin, H. Jin, Z. Zhou, and P. Szolovits, "TextFooler: Fooling Text Classifiers with Imperceptible Text Perturbations,"

Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 05, pp. 5119-5126, Apr. 2020. DOI: 10.1609/aaai.v34i05.5952

[4] Y. Li, T. Ji, H. Yuan, and X. Shi, "BERT-Attack: Adversarial Attack Against BERT Using BERT," Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 6193-6202, Nov. 2020. DOI: 10.18653/v1/2020.emnlp-main.500

[5] S. Garg, Y. Ramakrishnan, D. Gupta, and S. Agarwal, "BAE: BERT-based Adversarial Examples for Text Classification," Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 6177-6187, Nov. 2020. DOI: 10.18653/v1/2020.emnlp-main.498

[6] E. A. Rocamora, Y. Wu, F. Liu, G. G. Chrysos, and V. Cevher, "Revisiting character-level adversarial attacks for language models," Proceedings of the 41st International Conference on Machine Learning (ICML), Vol. 235, pp. 43171-43200, July 2024. URL: https://proceedings.mlr.press/v235/rocamora24a.html

[7] C. Gao, K. Gu, S. Vosoughi, and S. Mehnaz, "Semantic-Preserving Adversarial Example Attack against BERT," Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024), pp. 202-207, June 2024. DOI: 10.18653/v1/2024.trustnlp-1.18

[8] E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh, "Universal Adversarial Triggers for Attacking and Analyzing NLP," Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 2153-2162, Nov. 2019. DOI: 10.18653/v1/D19-1221

[9] F. Qi, M. Li, Y. Chen, Z. Zhang, Z. Liu, Y. Wang, and M. Sun, "Hidden Killer: Invisible Textual Backdoor Attacks with Syntactic Trigger," Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP), pp. 443-453, Aug. 2021. DOI: 10.18653/v1/2021.acl-long.37

[10] J. Yan, V. Gupta, and X. Ren, "BITE: Textual Backdoor Attacks with Iterative Trigger Injection," Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Vol. 1, pp. 12951-12968, July 2023. DOI: 10.18653/v1/2023.acl-long.724

[11] J. Li, Y. Yang, Z. Wu, V. G. Vydiswaran, and C. Xiao, "ChatGPT as an Attack Tool: Stealthy Textual Backdoor Attack via Blackbox Generative Model Trigger," Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), pp. 2985-3004, June 2024. DOI: 10.18653/v1/2024.naacl-long.164

[12] S. Zhao, M. Jia, L. A. Tuan, F. Pan, and J. Wen, "Universal Vulnerabilities in Large Language Models: Backdoor Attacks for In-context Learning," Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 14755-14777, Nov. 2024. DOI: 10.18653/v1/2024.emnlp-main.825

[13] E. Park, "NSMC: Naver sentiment movie corpus v1.0," GitHub Repository, 2015. Available: https://github.com/e9t/nsmc

[14] S. K. Lee, J. Lee, and J. Lee, "KoBERT: Pretrained Korean BERT model," GitHub Repository, 2019. Available: https://github.com/SKTBrain/KoBERT

[15] J. Park, J. Shim, and S. Lee, "KoELECTRA: Pretrained ELECTRA Model for Korean," GitHub Repository, 2020. Available: https://github.com/monologg/KoELECTRA

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 4171-4186, June 2019. DOI: 10.18653/v1/N19-1423

[17] E. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators," Proceedings of the 8th International Conference on Learning Representations (ICLR), Apr. 2020. URL: https://openreview.net/forum?id=r1xMH1BtvB

[18] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," Nature Machine Intelligence, Vol. 2, No. 11, pp. 665-673, Nov. 2020. DOI: 10.1038/s42256-020-00257-z

[19] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. H. Abrego, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, and R. Kurzweil, "Multilingual Universal Sentence Encoder for Semantic Retrieval," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL), pp. 87-94, July 2019. DOI: 10.18653/v1/P19-3010

[20] SK Telecom, "KoGPT2: Korean GPT-2 Pretrained Cased," GitHub Repository, 2020. Available: https://github.com/SKT-AI/KoGPT2

Authors

Areum Im received the B.A. degree in Public Administration from Seoul National University of Science and Technology in 2015. She is currently pursuing the M.S. degree in the Department of Cyber Security and Computer Engineering at the Korea National Defense University. Her research interests include language models, deep learning, and security.

Taehwa Lee received the B.S. and M.S. degrees in Computer Science from Korea Military Academy and Korea Advanced Institute of Science and Technology(KAIST), in 2015 and 2022, respectively. He is currently a Ph.D. candidate in the Department of Cyber Security and Computer Engineering, Korea National Defense University. His research interests include Machine Learning, Deep Learning, and Cybersecurity.

Soojin Lee received B.S., M.S. and Ph.D. degrees in Computer Science from Korea Military Academy, Yonsei University and Korea Advanced Institute of Science and Technology(KAIST) in 1992, 1996 and 2006. He is currently a professor of the Department of Cyber Security and Computer Engineering, Korea National Defense University since 2006. His research interests include National Cybersecurity Policy, Intrusion Detection System, Mobile Network Security, Machine Learning, Encryption theory and applications.