Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.35, NO.6, Dec. 2025 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2025.35.6.1611

LLM 기반 마약 은어 키워드 탐지 시스템*

김 민 석,[1†] 구 형 준[2‡]
[1,2]성균관대학교 (대학원생, 교수)

LLM-Based Drug Term Detection in Korean Messenger Conversations*

Minseok Kim,[1†] Hyungjoon Koo[2‡]
[1,2]Sungkyunkwan University (Graduate Student, Professor)

요 약

디지털 소통이 일상화되면서 온라인 마약 거래가 심각한 사회 문제로 부상하고 있다. 본 연구는 온라인 대화에서 한국어 기반의 마약을 일컫는 (알려지지 않은) 은어나 변형어를 자동으로 탐지하는 LLM (Large Language Model) 기반 탐지 시스템을 제안한다. 기존의 간단한 키워드 매칭 방식이나 텍스트를 벡터공간에서 의미와 문맥 정보를 파악하는 Word2Vec 기반 단어 임베딩 기술은 지속적으로 진화하는 은어와 의도적인 변형에 대응하기 어려운 한계점이 있다. 본 연구는 단어 빈도-역문서 빈도 (TF-IDF: Term Frequency-Inverse Document Frequency) 기반으로 통계적 가중치를 통해 자동으로 변형어를 생성하고, 이를 이용해 LLM 기반의 대규모 학습 데이터셋을 구축한다. 또한 슬라이딩 윈도우 기반으로 문맥을 인식하는 아키텍처와 이중 손실 함수를 활용한 메시지 수준의 어텐션 학습 모델을 이용한 마약 은어 키워드 탐지 시스템을 제안한다. KLUE/RoBERTa와 KLUE/BERT 모델을 활용한 실험 결과, 제안 시스템은 0.9816의 정확도와 0.9763의 재현율을 달성하였다.

ABSTRACT

We propose an LLM-based system that automatically detects (unknown) Korean slang and its variations referring to drugs in online conversations. Traditional approaches, such as simple keyword matching or Word2Vec-based word embedding for capturing semantic and contextual information in a vector space, have limitations in coping with continuously evolving slang and intentional word alterations. In this work, we generate linguistic variations using statistical weighting based on Term Frequency - Inverse Document Frequency (TF-IDF), constructing an LLM-based large-scale training dataset. Besides, we introduce a drug-referring slang detection system that employs a sliding-window-based contextual recognition architecture and a message-level attention learning model trained with a dual-loss function. Experimental results using the KLUE/RoBERTa and KLUE/BERT models demonstrate that the proposed system achieves an accuracy of 0.9816 and a recall of 0.9763.

Keywords: Drug Slang Detection, Social Media Mining, Large Language Model

I. 서 론

디지털 커뮤니케이션이 일상화되면서 온라인 플랫폼을 통한 불법 마약 거래가 심각한 사회 문제로 대두되고 있다. 특히 메신저 앱, SNS, 온라인 포럼 등은 익명성과 접근 용이성으로 인해 마약 거래의 주요 경로로 악용되고 있다 [1]. 거래 당사자들은 수사기관의 탐지를 회피하기 위해 직접적인 마약명 대신 '아이스', '얼음', '차가운 술' 등의 은어를 사용하며, 이러한 은어는 지속적으로 진화하고 변형되는 추세다 [2].

전통적인 키워드 기반 탐지 시스템은 이러한 동적

Received(10. 15. 2025), Modified(11. 18. 2025), Accepted(11. 21. 2025)
* 본 논문은 2025년도 정부(과학기술정보통신부) 정보통신기획평가원 No.RS-2024-00398745, 디지털 환경에서의 증거인멸행위 증명 및 대응기술 개발 지원으로 수행한 연구임.
† 주저자, for8821@g.skku.edu
‡ 교신저자, kevin.koo@g.skku.edu(Corresponding author)

특성에 대응하기 어려우며, 특히 한국어 환경에서는 언어적 특성으로 인한 추가적인 어려움이 존재한다. 한국어는 교착어적 특성으로 인한 복잡한 어미 변화, 띄어쓰기의 비일관성, 그리고 한글-영어-숫자의 혼용 등으로 인해 은어 탐지가 더욱 복잡해진다 [3]. 예를 들어, 마약류 '필로폰'을 뜻하는 '아이스'라는 은어는 'ㅇice', 'a이ㅇs', '아01스' 등 수십 가지 형태로 변형되어 사용하고 있으며, 이는 단순한 정규표현식 기반 패턴 매칭으로는 새로운 변형은 탐지가 불가능하다 [11].

온라인 마약 거래 탐지의 중요성은 단순히 불법 거래 차단을 넘어서 사회적 안전망 구축에 있다. 특히 청소년들이 손쉽게 접근할 수 있는 메신저 플랫폼에서의 마약 거래는 사회적 파급효과가 크다 [4]. 기존 연구들이 대부분 영어나 중국어에 집중되어 있는 반면 [5, 6], 한국어의 고유한 언어적 특성을 고려한 탐지 시스템에 대한 연구는 제한적이다 [11]. 또한, 은어의 지속적인 진화와 의도적 난독화 전략으로 인해 정적인 규칙 기반 접근법의 한계가 명확해지고 있다. 새로운 은어가 등장할 때마다 수동으로 규칙을 업데이트하는 것은 비효율적이며, 창의적인 변형 패턴에 대한 대응력이 부족하다. 이러한 문제를 해결하기 위해서는 문맥을 이해하고 패턴을 학습할 수 있는 지능형 시스템이 필요하다.

본 연구의 목적은 한국어 메신저 대화에서 마약 관련 은어를 효과적으로 탐지할 수 있는 대규모 언어모델(Large Language Model, LLM) 기반 시스템을 개발하는 것이다. 구체적으로는 한국어 환경에 특화된 은어 변형 패턴을 체계적으로 분석하고, 라벨링된 데이터 (labeled data) 부족 문제를 해결하기 위한 자동 데이터 증강 방법론을 제안한다. 또한 메신저 대화의 문맥적 특성을 고려한 슬라이딩 윈도우 기반 탐지 아키텍처를 설계하고, 실용적 운영 환경에서의 적용 가능성을 검증한다.

본 연구는 다음과 같은 측면에서 학술적·실용적 기여를 제공한다. 기술적으로는 TF-IDF [22] 기반 자동 변형어 생성을 통한 학습 데이터 확장 방법론을 제안하고, 슬라이딩 윈도우와 어텐션 메커니즘을 결합한 문맥 인식 탐지 아키텍처를 개발한다. 또한 윈도우 수준 분류와 메시지 수준 위치 탐지를 동시에 수행하는 이중 손실 함수를 도입한다. 본 연구는 최초로 한국어 은어를 체계적으로 탐지하는 방안을 제안하여 후속 연구의 기반을 제공하며, 언어 특화적 접근법의 중요성을 실증한다. 사회적으로는 온라인 마약 거래 차단을 통한 사회 안전 향상에 기여하고, 법 집행 기관과 플랫폼 운영자를 위한 실용적 도구를 제공하며, 청소년 보호 및 디지털 범죄 예방에 활용될 수 있는 기술적 기반을 마련할 수 있다. 제안된 시스템은 KLUE/RoBERTa와 KLUE/BERT 모델을 활용한 실험에서 0.9816의 정확도와 0.9763의 재현율을 달성하였다.

II. 기존 연구

2.1 키워드 기반 접근 방식

초기 마약 은어 탐지 연구는 주로 정규표현식과 사전 기반 매칭을 활용했다. 최민재 등은 Facebook, Instagram, KakaoTalk, Twitter, Telegram 등 5개 SNS 플랫폼에서 한국어 마약 은어 패턴을 분석하고 변형 규칙을 체계화했다. [15]. 이들은 한국어 환경에서 나타나는 은어 변형의 주요 패턴을 다음과 같이 분류했다.

첫째, 초성/중성/종성 치환 패턴이다. 한글의 자모를 유사한 모양의 영문자나 숫자로 대체하는 방식으로, 'ㅇ'을 '0'으로, 'ㅣ'를 '1'이나 'l'로 치환하는 등의 패턴이 관찰되었다. 이러한 치환은 시각적 유사성을 활용하여 사람은 읽을 수 있지만 기계적 탐지를 어렵게 만든다.

둘째, 영어/숫자 혼용 패턴이다. '아이스'를 'ice', '아1스', 'ㅇice' 등으로 표기하는 방식으로, 한글과 영어, 숫자를 혼합하여 사용한다. 이는 단순한 키워드 매칭을 무력화시키는 효과적인 회피 전략이다.

셋째, 특수문자 삽입 패턴이다. 단어 중간에 의미 없는 특수문자를 삽입하여 '아이스', '아_이_스' 등으로 표기하는 방식이다. 이는 언어모델 내 내장된 토크나이저의 정상적인 작동을 방해하여 탐지를 회피한다 [11].

그러나 이러한 규칙 기반 접근법은 수동으로 구축된 규칙의 확장성 한계를 드러낸다. 새로운 변형 패턴이 나타날 때마다 수동적으로 규칙을 추가해야 하며, 규칙 간의 충돌이나 우선순위 문제도 직면한다. 또한 이 방법은 문맥을 고려하지 않기 때문에 일반 단어와 은어를 구분하지 못하는 높은 오탐률 문제를 야기한다 [17].

2.2 임베딩 기반 접근 방식

Holbrook 등은 Reddit 데이터에서 Word2Vec을 활용하여 오피오이드 (opioid) 관련 은어를 자동으로 발견하는 시스템을 제안했다. [4] 이들의 접근법은 핵심 마약 용어 ('opioid', 'fentanyl', 'heroin')를 시드(seed)로 사용하여 벡터 공간 (vector space)에서 가까운 단어들을 은어 후보로 추출한다. 이 방법은 'blues'가 'fentanyl'과 유사한 문맥에서 사용됨을 포착하여 은어 관계를 발견했다. 예를 들어, "got some blues today"와 "got some fent today"가 유사한 문맥에서 나타나면, 두 단어의 임베딩이 가까워져 'blues'를 fentanyl의 은어로 식별한다. Word2Vec의 또 다른 장점은 언어의 의미적 관계를 포착할 수 있다는 점이다. 'heroin' - 'drug' + 'medication' = 'methadone'과 같은 유추 관계를 학습하여, 은어 간의 복잡한 관계를 모델링할 수 있다.

그러나 이 접근법은 여러 한계를 보인다. 첫째, 단어 수준 유사도만 고려하여 문맥적 의미 파악에 실패한다. 예를 들어, '얼음'이 날씨 관련 대화에서 사용될 때와 마약 거래 문맥에서 사용될 때를 구분할 수 없다. 둘째, 철자 변형이나 의도적 난독화에 취약하다. 'ice'와 '○ice', 'i-c-e' 등은 완전히 다른 단어로 처리되어 은어 관계를 포착할 수 없다. 셋째, 학습 데이터에 나타나지 않은 새로운 은어는 탐지가 불가능하다.

2.3 딥러닝 기반 접근 방식

최근 연구는 BERT [8]등 사전학습된 LLM을 활용하여 문맥을 고려한 탐지를 시도한다. 이러한 모델은 Transformer [21]를 기반으로 하여 단어 간의 복잡한 의존 관계를 포착할 수 있다.

Dong 등은 중국어 소셜미디어에서 BERT 기반 마약 관련 게시물 분류기를 개발했다 [18]. 이들의 시스템은 게시물 전체를 입력으로 받아 마약 관련 여부를 판별하며, 0.8575의 F1 점수 (F1 score, F1)를 달성했다. BERT의 양방향 문맥 이해 능력을 활용하여 은어가 사용되는 미묘한 문맥적 단서를 포착할 수 있었다.

Song 등은 원격 지도 학습 (distant supervision)과 탈어휘화 (delexicalization) 전략을 결합한 JEDIS를 제안하였다 [19]. 이들은 Reddit Drug·Silk Road Forum 말뭉치에서 코카인·엑스터시 등 46개 시드 용어를 활용해 라벨링된 긍정·부정 샘플을 구축하고, negative sampling을 통해 부정 샘플의 비율을 조절하였다. 이후 대상 단어를 마스킹한 채 문맥만으로 은어 여부를 판단하는 context-only 모듈과, 마스크된 단어의 문맥적 임베딩을 활용하는 word attribute 모듈을 각각 BERT 기반으로 학습한 뒤 앙상블하여 최종 예측을 수행하였다. 실제 평가에서 Reddit Drug과 Silk Road Forum 데이터셋에서 각각 0.6419, 0.6320의 F1 점수를 기록하였다.

그러나 대부분의 연구가 영어나 중국어에 집중되어 있으며, 한국어의 언어적 특성을 고려한 연구는 부족한 실정이다. 한국어는 교착어로서 조사와 어미 변화가 복잡하고, 띄어쓰기가 일정하지 않으며, 한글, 영어 및 한자를 혼용하는 특성이 있어 별도의 접근 방식이 필요하다.

III. 문제 정의

3.1 문제 정의

본 연구에서 다루는 문제는 주어진 메신저 대화에서 마약 관련 은어를 포함하는 메시지를 식별하고자 한다. 즉, 메신저 대화 $C=\{m_1, m_2, \cdots, m_n\}$가 주어졌을 때, 각 메시지 m_i에 대해 이진 레이블 $y_i \in \{0,1\}$을 예측하는 문제로 정의한다. 여기서 $y_i = 1$은 마약 관련 은어를 포함한 메시지다.

3.2 챌린지

마약 은어 탐지 시스템 개발에는 여러 기술적 도전 과제가 존재한다.

동적 은어 진화. 은어는 사법 집행 기관의 탐지를 피하기 위해 지속적으로 진화한다. 사용자들은 새로운 은어를 계속 생성하고, 기존에 활용되던 은어도 끊임없이 변형시킨다. 예를 들어, '아이스'라는 은어가 널리 알려지면 '얼음', '차가운 술', '크리스탈' 등의 새로운 은어가 등장하고, 각각이 다시 다양한 변형을 거친다. 이러한 동적 특성은 정적인 사전 기반 접근법을 무력화시킨다.

다의성. 많은 은어가 일상적인 단어에서 차용되었

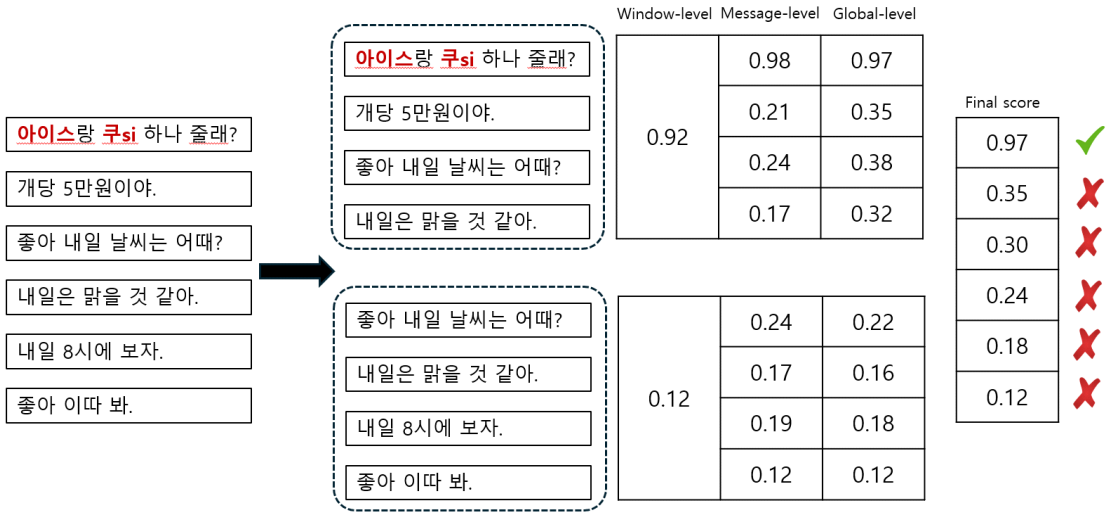

Fig. 1. Overview of our proposed model. To detect sentences containing the slang referring to a drug, the system divides a sentence into several sliding windows and computes a likelihood based on window-level predictions (e.g., 0.92) and message-level attention weights (e.g., 0.97). The final detection score is obtained by averaging the scores across overlapping windows, then compared the result against an adaptive threshold (e.g., 0.36 + 0.29 = 0.65) to determine whether the sentence contains the slang.

기 때문에 대화 내 문맥 없이는 구분이 불가능하다. '얼음 좀 사와'라는 메시지가 실제로 얼음을 구매하는 것인지, 메스암페타민을 거래하는 것인지는 전후 문맥을 통해서만 판단할 수 있다. 이는 단순한 키워드 매칭이나 단어 수준 분석의 한계를 보여준다.

의도적 난독화. 은어 사용자들은 탐지를 회피하기 위해 창의적인 변형을 시도한다. 단순한 철자 변형을 넘어서 이모티콘 삽입 ('아😂이😂스'), 초성 분리 ('ㅇㅏㅇㅣㅅㅡ'), 숫자 치환 ('아01스') 등 예측하기 어려운 패턴을 사용한다. 이러한 변형은 규칙으로 모두 정의하기 불가능하며, 일반화 능력이 필요하다.

데이터 불균형. 실제 메신저 대화에서 마약 관련 은어가 포함된 메시지는 극히 일부이다. 이로 인해 모델이 모든 메시지를 정상으로 분류하는 편향을 학습할 위험이 있다. 또한 은어의 민감한 특성상 레이블된 데이터를 대량으로 수집하기 어려운 현실적 제약도 있다.

한국어 특성으로 인한 복잡성. 한국어는 교착어로서 조사와 어미 변화가 복잡하고, 띄어쓰기가 일정하지 않다. '아이스있어?', '아이스를원해', '아이스필요해' 등 다양한 형태로 표현될 수 있으며, 각각에 대해 변형이 적용되면 경우의 수가 기하급수적으로 증가한다. 또한 한글, 영어, 숫자, 특수문자를 혼용하는 특성은 토큰화 (tokenization)와 정규화를 어렵게 만든다.

IV. 방법론

4.1 데이터셋 생성 방식

학습 데이터 부족은 은어 탐지 시스템 개발의 주요 장애물이다. 실제 은어가 포함된 대화를 대량으로 수집하는 것은 법적, 윤리적 제약이 있으며, 수동으로 레이블링하는 것은 비용이 많이 든다. 이를 해결하기 위해 본 연구는 자동 변형어 생성 기법을 제안한다.

자동 변형어 생성. Fig. 2에서 보여지듯이, 제안하는 자동 변형어 생성 방법은 일반적인 대화를 마약 은어가 포함된 대화로 변환하는 것이다. 핵심 아이디어는 대화에서 중요한 단어를 무작위 단어로 대체하면, 그 단어가 문맥상 특별한 의미를 갖는 것처럼 보

Table 1. Original Korean drug slang terms and their four generated obfuscated variants (Type 1 - Type 4), using the evasion pattern technique from [15], for constructing the evaluation test set.

Original	Type 1 variants	Type 2 variants	Type 3 variants	Type 4 variants
아이스	a이스	아1스	아이s	ai스
차가운 술	cha가운 술	차가un 술	차가운s술	차gaun 술
작대기	jak대기	작dae기	작대gi	작😂대😂기
공포의 백색가루	gong포의 백색가루	공포i의 백색ga루	공포의 백색가ru	공포😂의 백😂색garu

Algorithm 1: Dataset Generation

Input: Corpus D, Korean morpheme V
Output: Augmented dataset D'
foreach *conversation* $c \in D$ **do**
  $texts \leftarrow \text{ExtractMessages}(c)$;
  $nouns \leftarrow \text{KoreanNounExtractor}(texts)$;
  tfidf_matrix $\leftarrow \text{ComputeTFIDF}(nouns)$;
  tfidf_scores $\leftarrow \text{SumColumns}(\text{tfidf\_matrix})$;
  top_nouns $\leftarrow \text{SelectTopK}(\text{tfidf\_scores}, k = 3)$;
  replacement_nouns $\leftarrow \text{RandomSample}(V, k = 3)$;
  for $i = 1$ **to** $|top\_nouns|$ **do**
    foreach *message* m *in* c **do**
      if $Contains(m, top\_nouns[i])$ **then**
        $m \leftarrow \text{Replace}(m, \text{top\_nouns}[i], \text{replacement\_nouns}[i])$;
        $\text{SetLabel}(m, \text{positive} = \text{True})$;
    $D' \leftarrow D' \cup \{c\}$;
return D';

Fig. 2. Dataset generatation algorithm. We augment the training dataset by replacing top TF-IDF [23] nouns with random nouns to synthesize slang-like terms, labeling them as positive.

인다는 관찰에 기반한다. 이 알고리즘의 핵심은 TF-IDF [23]를 사용하여 대화에서 중요한 단어를 식별하는 것이다. TF-IDF가 높은 단어는 해당 대화에서 높은 중요도를 가질 가능성이 크며, 이를 다른 단어로 대체하면 자연스럽게 은어처럼 보이는 효과가 있다. 예를 들어, "오늘 날씨 좋네. 커피 마시러 갈까?"라는 대화에서 '커피'가 높은 TF-IDF 점수를 받았다면, 이를 '와이축'(형태소 사전에서 무작위 선택된 명사)으로 대체하여 "오늘 날씨 좋네. 와이축 마시러 갈까?"를 생성한다. 이렇게 생성된 문장에서 '와이축'은 문맥상 특별한 의미를 갖는 은어처럼 해석될 수 있다.

대화 데이터셋 생성. 모델의 실제 은어 탐지 능력을 평가하기 위해 실제 마약 은어와 그 변형어로 구성된 테스트셋을 구축하였다. 구글 검색을 통해 수집한 42개의 실제 마약 은어를 기반으로, 각 은어에 대해 4가지 유형의 변형어를 생성하여 총 210개의 은어 변형어 데이터셋을 완성하였다. 변형어 생성 규칙은 최민재 등의 연구에서 제시된 실제 온라인 환경에서 관찰되는 회피 패턴 [15]을 반영하도록 설계되었으며, Gemini-2.5-flash [23] 언어 모델을 활용하여 단어 내 자연성을 고려한 변형을 수행하였다. Table 1에서 보여지듯이, Type 1 변형은 첫 번째 글자를 시각적으로 유사한 영문자로 치환하는 방식으로, '아'을 영문 'a'로 대체하여 '아이스'를 'a이스'로 변형한다. Type 2 변형은 단어 중간 글자 중 하나를 영문자로 변환하는 방식으로 '아이스'를 '아i스'로 변형하며, Type 3 변형은 마지막 글자를 영문자로 대체하여 '아이s'와 같이 변환한다. Type 4 변형은 변형의 다양성을 극대화하기 위해 앞선 변형 기법들을 조합한 복합 변형 방식 (예: ai스)이다. 이러한 변형들은 한글 자모를 영문자나 숫자로 치환하거나, 한글-영어-숫자를 혼용하거나, 특수문자를 삽입하는 등 실제로 사용되는 회피 전략을 포괄하여 시각적 유사성을 활용해 인간은 읽을 수 있지만 기계적 탐지를 어렵게 만드는 특성을 구현하였다.

이렇게 생성된 은어와 변형어를 실제 대화 코퍼스에 삽입하여 테스트셋을 구성했다. 각 은어는 적절한 문맥에 삽입되어 실제 사용 환경을 시뮬레이션했다. 예를 들어, "얼음 좀 구해줄 수 있어?"와 같은 문장에서 '얼음'을 다양한 변형으로 대체하여 모델이 각 변형을 올바르게 탐지할 수 있는지 평가했다.

4.2 모델 구조 (Model Architecture)

제안하는 모델 아키텍처는 대화의 문맥을 효과적으로 포착하면서도 개별 메시지의 은어 포함 여부를 정확히 판별할 수 있도록 설계되었다. 핵심 구성 요소는 슬라이딩 윈도우 기반 문맥 인코딩, 어텐션 기반 특징 추출, 분류기 및 이중 손실 함수이다.

슬라이딩 윈도우 기반 문맥 인코딩. 메신저 대화는 연속적인 메시지의 시퀀스 (sequence)로 구성되

며, 각 메시지의 의미는 전후 문맥에 크게 의존한다. 예를 들어, "내일 몇시에 볼래?"라는 메시지는 그 자체로는 의미가 모호하지만, 앞에 "아이스 하나 갖다 줘."라는 메시지가 있다면 마약 거래 문맥일 가능성이 높아진다. 이러한 문맥 의존성을 모델링하기 위해 본 연구에서는 슬라이딩 윈도우 접근법을 채택했다. 구체적으로, 연속된 l개의 메시지 (예: $l=10$)를 하나의 입력 단위 (window)로 고정하고, 스트라이드 (stride, 예: $s=5$) 단위로 이동하면서 다음 윈도우를 순차적으로 생성한다. 윈도우의 집합 W은 수식 (1)과 같이 정의한다.

$$W=\left\{w_i:[m_{(i-1)s+1},m_{(i-1)s+2},\cdots,m_{(i-1)s+l}]\right\}_{i=1}^{\lfloor (n-l+1)/s \rfloor} \tag{1}$$

여기서 n은 전체 메시지 수이다.

마약 은어 분류기. 마약 은어 분류기는 사전학습된 KLUE/RoBERTa 모델을 통과한 문맥 표현을 정규화와 드롭아웃이 포함된 다중 레이어 퍼셉트론 (MLP)을 활용하여 로짓을 계산한 뒤, 소프트맥스를 적용해 윈도우 수준 이상 확률 p_{w_i}를 산출한다. 또한, 각 윈도우 내에서 개별 메시지의 기여도를 평가하기 위해 메시지 수준 어텐션 분포 $\alpha^{w_i}=[\alpha_1^{w_i},\alpha_2^{w_i},\ldots\alpha_l^{w_i}]$를 계산한다. 각 메시지 m_i의 전역 이상도 점수 s_i는 해당 메시지가 중첩된 $|W_i|$개 윈도우에서의 기여도와 윈도우 위험도를 β 비율로 혼합하여 계산한다:

$$s_i=\frac{1}{|W_i|}\sum_{w\in W_i}(\beta\alpha_i^w+(1-\beta)p_w) \tag{2}$$

결론적으로, 대화 내 $\{s_i\}$의 평균 $\bar{s}$와 표준편차 σ_s로 동적 임계치 $T=\bar{s}+\sigma_s$를 설정하고, $s_i>T$를 만족하는 메시지를 이상 메시지로 분류한다.

구체적 계산 예시. Fig. 1의 예시를 통해 계산 과정을 설명한다. 첫 번째 윈도우는 "아이스랑 쿠si 하나 줄래?", "개당 5만원이야.", "좋아 내일 날씨는 어때?", "내일은 탑을 것 같아."의 4개 메시지를 포함하며, 윈도우 수준 점수 $p_{w_1}=0.92$를 얻었다. 이 윈도우 내에서 "아이스랑 쿠si 하나 줄래?" 메시지는 어텐션 점수 $\alpha_1^{w_1}=0.97$을 받았다. 이 메시지는 두 윈도우에 중첩되어 나타나므로, 식 (2)에 따라 최종 점수 $s_1=0.97$을 얻었다. 전체 대화에서 각 메시지의 최종 점수는 0.97, 0.35, 0.30, 0.24, 0.18, 0.12이다. 이들의 평균 $\bar{s}=0.36$, 표준편차 $\sigma_s=0.29$를 계산하면 임계치 $T=\bar{s}+\sigma_s=0.65$가 된다. 해당 메시지는 $s_1>T$이므로 은어 포함 메시지로 분류된다.

4.3 이중 손실 함수 정의

제안하는 이중 손실 함수 (Dual loss function)는 윈도우 수준의 분류 정확도와 메시지 수준의 위치 정확도를 동시에 최적화한다. 이는 단순히 윈도우에 은어가 있는지 판별하는 것을 넘어서, 어떤 메시지가 은어를 포함하는지 정확히 식별할 수 있게 한다.

윈도우 수준 분류 손실. 본 연구에서는 클래스 불균형 문제를 해결하기 위해 Focal Loss [7]를 채택했다. Focal Loss는 다수를 차지하고 있는 클래스(대부분의 정상 메시지)에 대한 손실을 줄이고 소수 클래스 (은어 포함)에 집중하도록 설계했다:

$$L_{window}=-w_t(1-p_t)^\gamma\log(p_t) \tag{3}$$

여기서 p_t는 정답 클래스에 대한 모델의 예측 확률이다. w_t는 클래스별 가중치로, 은어 포함 클래스에 더 높은 가중치를 부여한다. γ는 focusing 파라미터로, 본 연구에서는 2.0을 사용했다.

메시지 수준 어텐션 감독 손실. 윈도우 안에서 실제 은어가 포함된 메시지에 높은 어텐션이 할당되도록 명시적인 감독 (supervision)을 추가한다.실제 은어 포함 여부를 나타내는 원-핫 벡터를 $y^{w_i}=[y_1^m,y_2^m,\ldots y_l^m]$라 하면, 다음과 같은 교차 엔트로피 (Cross Entropy) 형태로 손실을 계산한다:

$$L_{attention}=-\sum_{j=1}^{l}y_j^m\log(\alpha_j) \tag{4}$$

이 손실은 "어떤 메시지가 윈도우 내에서 은어를 담고 있는가"라는 위치 정보를 어텐션이 반영하도록 강

제한다. 결과적으로 모델은 단순히 윈도우 전체가 은어와 관련 있다고 판단하는 수준을 넘어, 구체적으로 어느 메시지가 그 판단의 근거인지를 더 잘 구분하게 된다.

최종 손실 함수. 두 구성 요소를 가중 결합하여 최종 학습 목표를 정의한다:

$$L_{total} = \lambda L_{window} + (1-\lambda) L_{attention} \quad (5)$$

본 연구에서는 $\lambda = 0.7$을 사용하여 윈도우 수준 분류에 더 높은 가중치를 부여했다. 이는 전체적인 탐지 성능을 우선시하면서도 위치 정확도를 향상시키는 균형을 제공한다.

V. 구 현

본 연구를 위한 소프트웨어 스택은 PyTorch 1.13.0을 기반으로 구축했다. Transformers 라이브러리 4.25.1을 사용하여 사전학습 모델을 로드하고 파인튜닝 (fine-tuning)하였으며, 한국어 형태소 분석을 위해 KoNLPy 0.6.0과 Kkma 형태소 분석기를 사용했다. 데이터 처리와 분석을 위해 pandas 1.5.2, numpy 1.23.5, scikit-learn 1.2.0을 활용했다.

모델 구현 세부 사항. 본 연구는 한국어 LLM인 KLUE/RoBERTa 기반 3개 모델 (small, base, large)와 KLUE/BERT- base 모델, 그리고 DistilBERT (monologg/distilkobert) [28] 및 Electra (monologg/koelectra-base-v3-discriminator) [29] 모델을 실험했다. 추가적으로, 상용 언어모델인 GPT-4o 및 Gemini 모델 또한 실험에 활용하였다. 각 모델에 대해 동일한 아키텍처를 적용했으며, 특수 토큰 추가로 인한 임베딩 크기 조정을 진행하였다. 슬라이딩 윈도우 구현에서는 효율성을 위해 배치 처리를 최적화했다. 동일한 대화에서 생성된 여러 윈도우를 하나의 배치로 묶어 처리하고, 패딩을 최소화하기 위해 동적 배치 구성을 사용했다. 옵티마이저로는 AdamW를, 스케줄러로는 PyTorch에 내장된 get_linear_schedule_with_warmup 함수를 사용하였다.

하이퍼파라미터. 학습은 다음과 같은 하이퍼파라미터로 수행되었다. 배치 크기는 16으로 설정했으며, 경사 누적 (gradient accumulation)을 2 스텝마다 적용하였다. 학습률은 2e-5로 시작하여 선형 감쇠 (linear decay)를 적용했다. 가중치 감소 (Weight decay)는 0.01을 사용했다. 총 20 epoch 동안 학습했으며, 검증 데이터에 대한 손실 (validation loss)가 5 epoch 동안 개선되지 않으면 early stopping을 적용했다. 또한, gradient clipping을 1.0으로 설정하여 학습 안정성을 보장했다. 손실함수의 하이퍼파라미터는 $\gamma = 2.0$, $\alpha = 0.25$ 및 $\lambda = 0.7$로 구성하였고, 분류기는 $\beta = 0.8$을 활용하였다.

VI. 평 가

본 연구의 모든 실험은 Intel(R) Core(TM) i9-10900X CPU @ 3.70GHz, 128GB RAM, Quadro RTX 8000 GPU로 구성된 64-bit Ubuntu 20.04 시스템에서 수행되었다.

6.1 실험환경 구축

데이터셋. 본 연구는 국립국어원 메신저 말뭉치 [16]를 기본 데이터셋으로 사용했다. 이 데이터셋은 카카오톡 및 라인 메신저 내 사용자들의 약 3,800개의 대화로 구성되며, 총 691,000개의 메시지를 포함한다. 각 대화는 평균 180개의 메시지로 이루어져 있으며, 일상적인 주제부터 업무 관련 대화까지 다양한 내용을 포함한다.
데이터 전처리 과정에서 다음과 같은 단계를 거쳤다. 먼저, 빈 메시지나 단순 이모티콘만 포함된 메시지를 제거했다. 둘째, 지나치게 긴 메시지 (512 토큰 초과)는 문장 단위로 분할했다. 셋째, 개인정보로 의심되는 패턴 (전화번호, 주소 등)을 마스킹 처리했다.
자동 변형어 생성을 통해 3,500개의 학습용 대화를 생성했다. 각 원본 대화에서 평균 3개의 변형 대화를 생성하여 데이터셋 크기를 확장했다. 테스트셋은 42개의 실제 은어와 각 4개의 변형을 380개의 대화에 삽입하여 총 1,900개의 테스트 샘플을 구성했다.

평가지표. 모델 평가는 전통적인 분류 성능 지표와 불균형 데이터 특성을 반영한 판별 능력 평가를 통해 수행되었다. 첫째, 정확도 (accuracy), 정밀도 (precision), 재현율 (recall), F1 점수 (F1)를 계산하여 모델의 예측 정확성과 마약 은어카 포함된

Table 2. Performance comparison of different language models on a drug-relavant slang detection task. Note that all models are pretrained on KLUE [17] and fine-tuned on our proposed algorithm.

Model	Method	Accuracy	Precision	Recall	F1	AUC
BERT-base	baseline	0.5123	0.5234	0.4987	0.5108	0.5156
	proposed	**0.9744**	**0.9798**	**0.9704**	**0.9751**	**0.9878**
RoBERTa-small	baseline	0.5087	0.5198	0.4923	0.5058	0.5134
	proposed	**0.9743**	**0.9806**	**0.9692**	**0.9749**	**0.9880**
RoBERTa-base	baseline	0.5156	0.5267	0.5034	0.5148	0.5189
	proposed	**0.9754**	**0.9807**	**0.9710**	**0.9758**	**0.9917**
RoBERTa-large	baseline	0.5234	0.5345	0.5123	0.5233	0.5278
	proposed	**0.9816**	**0.9879**	**0.9763**	**0.9821**	**0.9946**
DistilBERT-base	baseline	0.4981	0.4993	0.4951	0.4972	0.4992
	proposed	**0.9801**	**0.9824**	**0.9781**	**0.9802**	**0.9833**
Electra-base	baseline	0.5313	0.5458	0.5321	0.5388	0.5412
	proposed	**0.9724**	**0.9743**	**0.9652**	**0.9697**	**0.9703**
GPT-4o	baseline	0.6013	0.6234	0.5876	0.6049	0.6178
Gemini	baseline	0.5734	0.5923	0.5534	0.5723	0.5889

대화 클래스 및 포함되지 않은 클래스 간 성능 평가를 진행하였으며, AUC-ROC (i.e., AUC)를 측정하여 모델의 분류 성능을 종합적으로 평가하였다.

6.2 실험 결과

주요 실험 결과. 제안한 시스템의 성능을 다양한 측면에서 평가했다. Table 2와 같이 RoBERTa-large 모델이 가장 우수한 성능을 보였으며, 0.9816의 정확도, 0.9879의 정밀도, 0.9763의 재현율을 달성했다. 또한, 모델 크기에 따른 성능 차이를 분석한 결과, large 모델이 base 모델 대비 약 0.06의 성능 향상을 보였다. 이는 모델의 파라미터가 클수록 복잡한 은어 패턴을 더 잘 학습할 수 있음을 시사한다. 그러나 small 모델도 0.9742의 정확도를 달성하여, 리소스가 제한된 환경에서도 실용적인 성능을 제공할 수 있음을 확인했다. BERT와 RoBERTa 아키텍처를 비교한 결과, RoBERTa가 전반적으로 더 나은 성능을 보였다. 이는 RoBERTa의 동적 마스킹으로 인한 학습이 은어 탐지 태스크에 더 적합함을 나타낸다.

제거 연구 (Ablation study). 슬라이딩 윈도우의 크기와 스트라이드가 성능에 미치는 영향을 체계적으로 분석하기 위해 RoBERTa-small을 5 epoch 동안 학습시켰다. Table 3에서 보여지듯이, 윈도우 크기를 6에서 10까지 늘리자 정확도는 0.9384에서 0.9557로, F1 점수는 0.9225에서 0.9571로 꾸준히 향상되었다. 윈도우 크기 6에서는 전후 문맥이 부족하여 일반 대화와 마약류 은어를 포함한 대화를 구분하기 어려움을 보여준다.

Table 3. Impact of a sliding window size on detection performance of RoBERTa-small language model with pretraining KLUE [17] and fine-tuning our proposed algorithm.

Window size	Accuracy	F1
6	0.9384	0.9225
8	0.9476	0.9436
10	**0.9557**	**0.9571**

추가적으로, 스트라이드의 변화에 따른 성능 평가도 진행하였다. Table 4에서 보여지듯이, 스트라이드 3은 학습 시간 대비 성능 향상이 크지 않았으며, 스트라이드 5는 정확도 0.9557, F1 0.9571로 스트라이드 4 (정확도 0.9549, F1 0.9529) 대비 소폭 우수한 성능을 보이면서도 중복 윈도우 처리를 줄여 연산 효율성까지 확보하여 성능과 효율성의 최적 균형을 달성하였다.

Table 4. Impact of a stride on detection performance of RoBERTa-small language model with pretraining KLUE [17] and fine-tuning our proposed algorithm.

Stride	Accuracy	F1
3	0.9535	0.9518
4	0.9549	0.9529
5	**0.9557**	**0.9571**

6.3 제안 접근 방식 평가

실제 운영 환경에서의 적용 가능성을 검증하기 위해 학습된 모델의 추론 처리량과 메모리 요구량을 CPU 및 GPU 환경에서 정량적으로 측정하였다. 해당 실험에서는 Roberta-small 모델을 기준으로 측정하였다.

추론 처리량. 트레이닝 데이터셋에서 생성한 슬라이딩 윈도우 (131,792개)를 대상으로 실험한 결과, CPU 환경에서는 초당 평균 9.4개의 윈도우를 처리할 수 있었다. 반면 GPU 환경에서는 초당 61.4개경 대비 약 6배 이상의 처리 성능 향상을 제공하였다. 이러한 처리 속도는 실시간 또는 준실시간 대화 모니터링 시스템의 요구 조건을 충족시킬 수 있는 수준으로 해석할 수 있다.

메모리 사용량. CPU 및 GPU 환경에서 메모리 사용량을 트레이닝 데이터셋에서 측정하였다. CPU 환경에서는 프로세스가 주기억장치 (RAM)을 점유하고 있는 크기인 RSS (Resident Set Size)를 측정한 결과, 프로세스의 RSS가 약 1.39GB로 나타났으며, 대화량이 증가하더라도 메모리 사용이 급격히 상승하지 않는 모습을 보였다. 또한, GPU 환경에서의 실제 사용량 (allocated memory)는 636MB였고 예약된 메모리 (reserved memory)는 654MB였다. 종합하면, 본 모델은 GPU를 활용한 배치 또는 스트리밍 추론 환경에서 실시간성 요구를 만족하면서도 메모리 자원 측면에서 경량성을 유지하므로, 실제 운영 시스템에 통합할 수 있는 실용적인 성능 프로파일을 갖추고 있는 것으로 판단된다.

6.4 일반화 능력 평가

본 절에서는 6.2절에서 제안한 TF-IDF 기반 변형어 데이터 증강과 슬라이딩 윈도우/어텐션 구조를 적용하여 학습된 모델(이하 제안 모델)의 일반화 능력을 평가한다. 구체적으로, 학습에 사용되지 않은 메신저 도메인인 AI-Hub 일상 대화 코퍼스를 대상으로 학습할 수 있음을 검증한다.

6.2절과 동일하게 국립국어원 메신저 말뭉치를 학습 데이터로 사용하여 제안 모델을 학습한 후, 추가 파인튜닝 없이 AI-Hub에서 제공하는 주제별 텍스트 일상 대화 데이터셋 [27]에 직접 적용하였다. 이때 AI-Hub 데이터 중 카카오톡을 제외한 4개 메신저(페이스북, 인스타그램, 밴드, 네이트온)의 대화 19,809건을 추가 평가 데이터셋으로 활용하였다. 해당 데이터셋은 플랫폼, 사용자 구성, 대화 주제 등 여러 측면에서 국립국어원 메신저 말뭉치와 상이하므로, 학습에 사용되지 않은(out-of-domain) 메신저 환경으로의 일반화 능력을 검증하기에 적합하다.

Table 5에서 AI-Hub의 주제별 텍스트 일상 대화 데이터셋에서도 제안 모델은 모든 언어 모델에서 baseline 대비 일관된 성능 향상을 보인다. 예를 들어 DistilBERT [28]의 경우 baseline 대비 정확도가 0.4796에서 0.9792로, F1 점수가 0.4974에서 0.9804로 향상되었으며, ELECTRA [29] 모델 또한 정확도 0.9712, F1 0.9682의 성능을 달성하여 비교적 경량 모델에서도 제안 기법이 다른 메신저 도메인으로 잘 학습할 수 있음을 보여준다. RoBERTa-large 모델은 AI-Hub 데이터셋에서도 가장 높은 정확도 0.9819를 기록하여, 대규모 파라미터를 가진 모델이 도메인 이동(domain shift) 상황에서 은어 및 변형어 패턴을 상대적으로 더 잘 일반화할 수 있음을 시사한다.

추가적으로, 모델의 오류 양상을 정량적으로 파악하기 위해 국립국어원 메신저 말뭉치 및 AI-Hub 내 주제별 텍스트 일상 대화 데이터셋를 기준으로 오탐율(false positive rate, FPR)과 미탐율(false negative rate, FNR)을 산출하였다. 제안된 기법으로 파인튜닝 된 Roberta-large 모델의 국립국어원 메신저 말뭉치 데이터셋의 FPR은 0.0321, FNR은 0.0478 수준으로 나타났으며, 모델별·데이터셋별 세부 수치는 Table 6에 정리하였다. 이러한 정량적 지표는 7.3절에서 제시하는 오탐 및 미탐 사례 분석과 함께 해석함으로써, 모델이 어떤 유형의 발화에서 특히 취약한지를 보다 구체적으로 파악할 수 있다.

Table 5. Evaluation of varying models trained on the NIKL messenger corpus [16] (3,500 conversations from KakaoTalk, LINE; Section 6.2) on the out-of-domain AI-Hub messenger dataset [27] (19,809 conversations from Facebook, Instagram, Band, and NateOn).

Model	Method	Accuracy	Precision	Recall	F1	AUC
BERT-base	baseline	0.5105	0.5192	0.5015	0.5096	0.5122
	proposed	**0.9739**	**0.9755**	**0.9712**	**0.9733**	**0.9850**
RoBERTa-small	baseline	0.5091	0.5201	0.4910	0.5051	0.5140
	proposed	**0.9740**	**0.9810**	**0.9685**	**0.9746**	**0.9875**
RoBERTa-base	baseline	0.5160	0.5280	0.5011	0.5142	0.5190
	proposed	**0.9760**	**0.9815**	**0.9720**	**0.9767**	**0.9920**
RoBERTa-large	baseline	0.5240	0.5305	0.5100	0.5200	0.5280
	proposed	**0.9819**	**0.9830**	**0.9780**	**0.9805**	**0.9950**
DistilBERT-base	baseline	0.4976	0.4981	0.4967	0.4974	0.4982
	proposed	**0.9792**	**0.9814**	**0.9794**	**0.9804**	**0.9827**
Electra-base	baseline	0.5327	0.5491	0.5347	0.5418	0.5425
	proposed	**0.9712**	**0.9739**	**0.9625**	**0.9682**	**0.9699**

Table 6. False positive rate (FPR) and false negative rate (FNR) of the RoBERTa-large based proposed model on NIKL messenger corpus [16] and AI-Hub messenger [27] dataset.

Dataset	FPR	FNR
NIKL	0.0321	0.0478
AI-Hub	0.0284	0.0523

VII. 사례 연구

7.1 자연스러운 대화 내 은닉된 대화 탐지

본 연구에서 제안한 모델의 주요 강점 중 하나는 일반적인 대화 속에 교묘하게 숨겨진 마약 관련 은어를 효과적으로 탐지한다는 점이다. Fig. 3에서 보여지듯이, 모델은 프로젝트 진척도 및 회의자료 공유 등이 포함된 일상 대화 속 "어제 쿠si 잘 받았어?"라는 은어가 포함된 대화를 효과적으로 구분할 수 있었다. 이는 모델이 단순한 키워드 매칭을 넘어서 문맥적 단서와 대화 패턴을 종합적으로 고려하여 판단을 내리고 있음을 시사한다.

7.2 다인원 대화 확장성

실제 메신저 환경에서는 일대일 대화뿐만 아니라 다수가 참여하는 단체 대화방도 빈번하게 사용된다. 이러한 환경에서의 모델 성능을 평가하기 위해 4명의 화자가 참여하는 단체 대화 시나리오를 구성하여 실험을 진행했다.

각 대화에서 화자별 조합 ($\binom{4}{2}$)을 고려하여 100개의 테스트 케이스를 Gemini [23]을 활용해 생성했다. 실험 결과, 다인원 대화에서도 모델은 0.9234의 정확도를 유지하며 안정적인 성능을 보였다. 이는 화자가 증가하고 대화의 복잡도가 높아져도 모델이 은어 탐지 능력을 유지한다는 것을 보여준다.

다만, 화자 수가 증가할수록 대화의 맥락이 분산되고 암시적 표현이 증가하는 경향이 있어, 일대일 대화 대비 약 5% 정도의 성능 하락이 관찰되었다. 이는 향후 멀티턴 대화 및 화자 특성을 고려한 모델링이 필요하다는 것을 시사한다.

7.3 오탐 및 미탐 사례 연구

국립국어원 메신저 말뭉치와 AI-Hub 일상 대화 데이터셋을 대상으로 산출한 정량적 지표에서, 제안 기법으로 파인튜닝된 RoBERTa-large 모델의 FPR은 0.0321, FNR은 0.0478 수준으로 나타났다(Table 6). 이는 전체 대화 중 약 3% 수준의 정상 메시지가 오탐지(False Positive)되고, 약 5% 수준의 마약 은어 포함 메시지가 미탐지(False Negative)되는 경향이 있음을 의미한다. 본 절에서는 이러한 정량적 지표를 뒷받침하기 위해, 대표적인

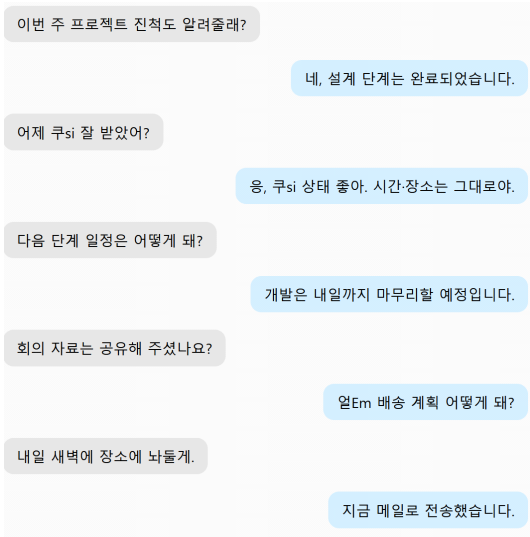

Fig. 3. Example of the drug-indicating slang detection in a everyday conversation. Our system successfully identifies drug-related slang (e.g., "어제 쿠si 잘 받았어?") that is subtly embedded within otherwise ordinary workplace dialogues about project updates and meeting materials.

오탐지 및 미탐지 사례를 유형별로 분석하고 모델 구조와의 연관성을 논의한다.

False Positive 사례 분석. False positive의 주요 원인은 일상적인 단어가 특수한 문맥에서 사용될 때 발생했다. 예를 들어, "오늘 너무 덥다…", "차가운 맥주 한잔 하자"에서 '차가운'이 은어로 오분류되는 경우가 있었다. 이는 모델이 '차가운'과 관련된 은어 패턴('차가운 술' 등)을 과도하게 일반화했기 때문이다. 또 다른 False Positive 패턴은 숫자와 영어가 혼용된 일반적인 표현이었다. "내일 meeting 3시에 하는 거 맞지?" "meeting 3시에 하자"와 같은 문장에서 영어-한글 혼용 패턴이 은어 변형과 유사하여 오탐지되는 경우가 있었다. 이는 변형 패턴의 특징이 일상적인 코드 스위칭과 겹치는 문제를 보여준다.

False Negative 사례 분석. False negative는 주로 학습 데이터에 없는 새로운 은어나 극도로 창의적인 변형에서 발생했다. 예를 들어, "어제 그 크리스탈 물건 괜찮더라", "다음에는 두 개만 챙겨줘"와 같은 완전히 새로운 은어나 최신 유행어를 사용한 경우 탐지에 실패했다. 이는 모델의 일반화 능력의 한계를 보여준다. 문맥이 매우 암시적인 경우도 False negative의 원인이었다. "그거 어제 거 또 있어?", "응", "그럼 내일 그때 보자"과 같은 극도로 간결한 대화에서는 은어가 명시적으로 나타나지 않아 탐지가 어려웠다. 이는 문맥 정보만으로는 한계가 있음을 시사한다.

VIII. 한계점 및 향후 연구

8.1 한계점

스트라이드 기반 탐지. 고정된 윈도우 크기와 스트라이드로 인해 은어가 윈도우 경계에 걸쳐 나타날 경우 탐지 성능이 저하될 수 있다. 예를 들어, "아이스" / "있어?"와 같이 핵심 은어와 문맥 단서가 서로 다른 윈도우에 분리될 경우, 각 윈도우만으로는 은어 사용을 판단하기 어렵다. 현재 시스템은 스트라이드를 통한 중첩으로 이를 부분적으로 해결하지만, 완벽한 해결책은 아니다. 또한 대화의 길이가 매우 길어질 경우, 중요한 문맥 정보가 윈도우 범위를 벗어나 손실될 수 있다. 이는 특히 은어가 대화 초반에 언급되고 실제 거래는 후반에 이루어지는 경우와 같은 장거리 의존성을 포착하지 못하는 문제로 이어진다.

단일 도메인 의존성. 본 연구는 메신저 대화 데이터에 특화되어 학습되었기 때문에, 다른 형태의 텍스트에서는 성능이 보장되지 않는다. SNS 게시물, 댓글, 포럼 글 등은 메신저 대화와는 다른 언어적 특성을 가진다. 예를 들어, 트위터는 글자 수 제한으로 인해 더 압축적인 표현을 사용하고, 인스타그램은 해시태그를 통한 은어 사용이 빈번하다. 이러한 플랫폼별 특성을 현재 모델이 충분히 반영하지 못할 가능성이 있다. 또한 음성 기반 메신저나 이미지/비디오를 포함한 멀티모달 커뮤니케이션에서의 은어 사용은 전혀 다른 접근이 필요하다.

8.2 향후 연구 방향

동적 윈도우 크기 조정. 동적 윈도우 크기 조정은 고정된 윈도우의 한계를 극복할 수 있는 유망한 방향이다. 대화의 특성에 따라 윈도우 크기를 적응적으로 조절하는 메커니즘을 개발할 필요가 있다. 예를 들어, 짧고 빈번한 메시지 교환이 있는 부분에서는

작은 윈도우를, 긴 메시지가 이어지는 부분에서는 큰 윈도우를 사용하는 방식이다. 이를 위해 강화학습(reinforcement learning)이나 메타러닝 (meta-learning) 기법을 활용하여 최적의 윈도우 크기를 자동으로 결정하는 에이전트를 학습시킬 수 있다. 또한, 계층적 어텐션 (hierarchical attention) 메커니즘을 도입하여 로컬 문맥과 글로벌 문맥을 동시에 고려하는 구조도 고려할 수 있다.

변형어 학습 효과의 정량적 분석. TF-IDF 기반으로 선정한 핵심 용어에 다양한 변형어를 부여하여 학습에 반영한 것이 성능에 얼마나 기여하는지에 대해서는, 본 논문에서 일부 결과 (Table 2 및 Table 5)에만 제시되어 있으며 체계적인 전·후 비교 실험은 제한적이다. 향후 연구에서는 동일한 모델과 동일한 데이터 조건에서 TF-IDF 기반 변형어 학습을 적용한 경우와 적용하지 않은 경우를 직접 비교하여, 변형어 학습 반영 전/후의 성능 차이를 정량적으로 분석할 필요가 있다. 이를 통해 변형어 생성 비율, TF-IDF 임계값 등 하이퍼파라미터가 실제 탐지 성능과 일반화 성능에 미치는 영향을 보다 체계적으로 규명할 수 있을 것이다.

주기적 재학습 및 능동학습 기반 동적 업데이트. 새로운 은어 및 극단적 변형에 대한 취약성을 완화하기 위해서는, 모델을 한 번 학습한 뒤 고정하는 정적인 방식이 아니라 주기적인 재학습과 능동 학습(active learning)을 통한 동적 업데이트 전략이 필요하다. 예를 들어, 실제 운영 환경에서 수집되는 대화 중 모델의 예측 확률이 애매한 샘플, 반복적으로 오탐·미탐이 발생하는 샘플을 우선적으로 선별하여 전문가 (예: 수사기관·모니터링 인력)의 라벨링을 받은 뒤, 주기적으로 모델을 재학습하는 형태를 구성할 수 있다. 이러한 능동학습 기반 파이프라인을 도입하면, 사람이 일일이 모든 대화를 검토하지 않더라도 라벨링 효율을 극대화하면서 최신 은어 패턴을 빠르게 학습시킬 수 있다.

다국어 및 멀티모달 확장. 다국어 및 멀티모달 확장은 현실적인 커뮤니케이션 환경을 반영하기 위해 필요하다. Multilingual BERT [24]와 같은 다국어 모델을 활용하여 언어 간 은어 패턴을 학습할 수 있다. 또한, 이미지나 비디오에 포함된 은어(예: 마약 사진에 붙는 특정 이모티콘)를 탐지하기 위한 멀티모달 모델 개발도 중요한 연구 방향이다. 음성 메시지의 경우, speech-to-text 변환 후 탐지하는 것뿐만 아니라 음성 특징 자체에서 은어 사용 패턴을 학습하는 end-to-end 접근도 고려할 수 있다.

설명 가능한 AI. 설명 가능한 AI (Explainable AI) 기술 통합은 시스템의 신뢰성과 법적 활용도를 높이기 위해 중요하다. LIME [25], SHAP [26] 등의 사후 설명 기법을 적용하여 모델의 예측 근거를 제시할 수 있다. 또한, 어텐션 메커니즘을 더욱 정교하게 설계하여 어떤 단어나 문맥이 은어 판단에 기여했는지 시각화할 수 있다. 또한, 대조적 설명 (Counterfactual explanation)을 생성하여 "만약 이 단어가 없었다면 결과가 어떻게 바뀌었을까?"와 같은 질문에 답할 수 있는 시스템도 개발할 수 있다.

IX. 결 론

본 연구는 한국어 메신저 대화에서 마약 관련 은어를 효과적으로 탐지하는 LLM 기반 시스템을 제안했다. 본 연구는 기존 접근법의 한계인 정적 패턴 의존성, 문맥 이해 부족, 새로운 은어 대응 불가 등의 문제를 해결하기 위해 혁신적인 방법론을 개발했다.

제안 시스템의 핵심 기여는 다음과 같다. 첫째, TF-IDF 기반 자동 데이터 증강 방법을 통해 레이블 데이터 부족 문제를 해결했다. 이는 일반 대화를 은어 포함 대화로 자연스럽게 변환하여 대규모 학습 데이터를 생성할 수 있게 했다. 둘째, 슬라이딩 윈도우 기반 문맥 인코딩을 통해 대화의 맥락을 효과적으로 포착하는 데 성공하였다. 셋째, 이중 손실 함수를 통한 메시지 수준 어텐션 학습으로 은어의 정확한 위치를 식별할 수 있게 했다.

실험 결과는 제안 방법의 효과성을 명확히 보여준다. KLUE/RoBERTa-large 모델은 0.9816의 정확도와 0.9763의 재현율을 달성하였다. 추가로, 실용성 평가에서는 GPU 환경에서 초당 약 61.4개의 윈도우를 처리할 수 있어 준실시간 모니터링이 가능하며, 메모리 부담도 비교적 낮아 실제 운영에 적합한 성능 프로파일을 보였다.

본 연구의 의의는 기술적 성과를 넘어서 사회적 영향에도 있다. 마약 은어 탐지 시스템은 온라인 마약 거래를 차단하고 청소년을 보호하는 데 기여할 수 있다. 법 집행 기관은 이 시스템을 활용하여 수사 효

율성을 높일 수 있으며, 플랫폼 운영자는 유해 콘텐츠를 사전에 차단할 수 있다. 또한, 본 연구에서 개발한 방법론은 마약 은어뿐만 아니라 다른 종류의 유해 언어 탐지에도 응용될 수 있다.

그러나 본 시스템은 여전히 개선의 여지가 있다. 고정된 윈도우 크기, 단일 도메인 의존성 등의 한계는 향후 연구를 통해 해결되어야 한다. 향후 연구 방향으로는 동적 윈도우 조정, 다국어 멀티모달 확장, 설명 가능한 AI 통합 등을 제시했다. 이러한 연구가 진행된다면, 더욱 실용적이고 신뢰할 수 있는 은어 탐지 시스템을 구축할 수 있을 것이다.

결론적으로, 본 연구는 한국어 환경에서 마약 은어를 탐지하는 최초의 체계적인 연구로서, 기술적 혁신과 실용적 가치를 동시에 제공한다. 제안된 방법론과 시스템은 디지털 시대의 새로운 형태의 범죄에 대응하는 중요한 도구가 될 것으로 기대된다.

References

[1] A. Sarker, K. O'Connor, R. Ginn, M. Scotch, K. Smith, D. Malone, and G. Gonzalez, "Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from Twitter," Drug Safety, vol. 39, no. 3, pp. 231-240, Mar. 2016.

[2] H. Hu, N. Phan, S. A. Chun, J. Geller, H. Vo, X. Ye, R. Jin, K. Ding, D. Kenne, and D. Dou, "An insight analysis and detection of drug-abuse risk behavior on Twitter with self-taught deep learning," Computational Social Networks, vol. 6, no. 10, Nov. 2019.

[3] S. S. Simpson, N. Adams, C. M. Brugman, and T. J. Conners, "Detecting novel and emerging drug terms using natural language processing: A social media corpus study," JMIR Public Health and Surveillance, vol. 4, no. 1, Jan. 2018.

[4] E. Holbrook, B. Wiskur, and Z. Nagykaldi, "Discovering opioid slang on social media: A Word2Vec approach with Reddit data," Drug and Alcohol Dependence Reports, vol. 13, 100302, Dec. 2024.

[5] P. M. Lavanya and E. Sasikala, "Auto capture on drug text detection in social media through NLP from the heterogeneous data," Measurement: Sensors, Vol.24, 100550, Dec. 2022.

[6] M. A. Al-Garadi, Y.-C. Yang, H. Cai, Y. Ruan, K. O'Connor, G. Gonzalez-Hernandez, J. Perrone, and A. Sarker, "Text classification models for the automatic detection of nonmedical prescription medication use from social media," BMC Medical Informatics and Decision Making, vol. 21, no. 27, Jan. 2021.

[7] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proceedings of the IEEE International Conference on Computer Vision, Oct. 2017, pp. 2980-2988.

[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, June 2019, pp. 4171-4186.

[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, Jul. 2019.

[10] Y. Kim, J.H. Kim, J. M. Lee, M. J. Jang, Y. J. Yum, S. Kim, U. Shin, Y.-M. Kim, H. J. Joo, and S. Song, "A pre-trained BERT for Korean medical natural language processing," Scientific Reports, vol. 12, 13847, Aug. 2022.

[11] S. Kim, T. Kang, T. K. Chung, Y. Choi, Y. Hong, K. Jung, and H. Lee, "Automatic extraction of comprehensive drug safety information from adverse drug event narratives in the Korea Adverse Event Reporting System using natural language processing techniques," Drug Safety, vol. 46, pp. 781-795, Aug. 2023.

[12] J. Tassone, P. Yan, M. Simpson, C. Mendhe, V. Mago, and S. Choudhury, "Utilizing deep learning and graph mining to identify drug use on Twitter data," BMC Medical Informatics and Decision Making, vol. 20, no. 304, Dec. 2020.

[13] National Institute of Justice, "Taking on the Dark Web: Law Enforcement Experts ID Investigative Needs," https://nij.ojp.gov/topics/articles/taking-dark-web-law-enforcement-experts-id-investigative-needs#1-0, Accessed: Nov, 24, 2025

[14] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234-1240, Feb. 2020.

[15] M. Choi, H. Lee, J. Kim, and C. Kim, "A Study on Keyword Collection System for Tracking Drug Crimes on Social Media," in Proceedings of the Conference on Information Security and Cryptography-Summer 2023, 31(2), pp. 209-212, Jul. 2023.

[16] National Institute of Korean Language, "Messenger Corpus v2.0," https://kli.korean.go.kr/corpus/main/requestMain.do#none, Accessed: Nov. 24, 2025

[17] S. Park, J. Moon, S. Kim, W. I. Cho, J. Han, J. Park, C. Song, J. Kim, Y. Song, T. Oh, J. Lee, J. Oh, S. Lyu, Y. Jeong, I. Lee, S. Seo, D. Lee, H. Kim, M. Lee, S. Jang, S. Do, S. Kim, K. Lim, J. Lee, K. Park, J. Shin, S. Kim, L. Park, A. Oh, J.W. Ha, K. Cho, "KLUE: Korean Language Understanding Evaluation," in Proceedings of the 35th Conference on Neural Information Processing Systems, Dec. 2021, pp. 3583-3595.

[18] F. Dong, W. Guo, J. Liu, T. A. Patterson, H. Hong. "BERT-based language model for accurate drug adverse event extraction from social media: implementation, evaluation, and contributions to pharmacovigilance practices." Front Public Health. Apr. 2024.

[19] M. Song, E. Jang, J. Kim, and S. Shin. "Covering Cracks in Content Moderation: Delexicalized Distant Supervision for Illicit Drug Jargon Detection". In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (KDD '25), Jul. 2025. pp. 1265-1276 .

[20] F. Mirić, "Criminological, criminalistic and linguistic aspects of drug addicts' slang," Science and Society, vol. 5, no. 2, pp. 489 - 495, 2020

[21] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000-6010, Dec. 2017.

[22] Juan Ramos, "Using TF-IDF to determine word relevance in document queries," in Proceedings of the First Instructional Conference on Machine Learning. pp. 29 - 48., Jan. 2003,

[23] Gemini Team, "Gemini 2.5: Pushing the Frontier with Advanced Reason-

ing, Multimodality, Long Context, and Next Generation Agentic Capabilities.", Jul. 2025
〔24〕 Telmo Pires, Eva Schlinger, and Dan Garrette. "How Multilingual is Multilingual BERT?", In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4996 - 5001. Jul. 2019.
〔25〕 Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should I trust you?" Explaining the predictions of any classifier.","" In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135-1144. Aug. 2016.
〔26〕 Mosca, Edoardo, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. "SHAP-based explanation methods: a review for NLP interpretability.," In Proceedings of the 29th international conference on computational linguistics, pp. 4593-4603. Oct. 2022.
〔27〕 AI-Hub, "Topic-based Text Daily Conversation Dataset," https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=543, Accessed: Nov. 24, 2025
〔28〕 Sanh, V. et al., "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.," In proceedings of 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing, Oct. 2019
〔29〕 Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators.," In International Conference on Learning Representations (ICLR). Sep. 2020.

〈저자소개〉

김 민 석 (Minseok Kim) 학생회원
2024년 8월 : 성균관대학교 소프트웨어학과 학사
2024년 9월 ~ 현재 : 성균관대학교 소프트웨어학과 석사과정
〈관심분야〉 정보보호, 인공지능, 인공지능 보안

구 형 준 (Hyungjoon Koo) 종신회원
2010년 2월 : 고려대학교 정보보호학과 석사
2019년 5월 : 스토니 브룩 뉴욕주립대 컴퓨터과학과 박사
2020년 12월 : 조지아텍 박사 후 연구원
2021년 3월 ~ 현재 : 성균관대학교 소프트웨어학과 부교수
〈관심분야〉 소프트웨어/시스템 보안, 인공지능을 활용한 정보보안