온라인 범죄 예방을 위한 실시간 조기 위험 감지 시스템

안진명°[1], 이근배†[1,2]
포항공과대학교 인공지능대학원[1], 포항공과대학교 컴퓨터공학과[2]
{jinmyeong,gblee}@postech.ac.kr

Real-Time Early Risk Detection in Textual Data Streams for Enhanced Online Safety

Jinmyeong An°[1], Geun-Bae Lee†[1,2]
Graduate School of Artificial Intelligence, Pohang University of Science and Technology[1]
Computer Science and Engineering, Pohang University of Science and Technology[2]

요 약

최근 소셜 네트워크 서비스(SNS) 및 모바일 서비스가 증가함에 따라 사용자들은 다양한 종류의 위험에 직면하고 있다. 특히 온라인 그루밍과 온라인 루머 같은 위험은 한 개인의 삶을 완전히 망가뜨릴 수 있을 정도로 심각한 문제로 자리 잡았다. 그러나 많은 경우 이러한 위험들을 판단하는 시점은 사건이 일어난 이후이고, 주로 법적인 증거채택을 위한 위험성 판별이 대다수이다. 따라서 본 논문은 이러한 문제를 사전에 예방하는 것에 초점을 맞추었고, 계속적으로 발생하는 대화와 같은 event를 실시간으로 감지하고, 위험을 사전에 탐지할 수 있는 Real-Time Early Risk Detection(RERD) 문제를 정의하고자 한다. 온라인 그루밍과 루머를 실시간 조기 위험 감지(RERD) 문제로 정의하고 해당 데이터셋과 평가지표를 소개한다. 또한 RERD 문제를 정확하고 신속하게 해결할 수 있는 강화학습 기반 새로운 방법론인 RT-ERD 모델을 소개한다. 해당 방법론은 RERD 문제를 이루고 있는 온라인 그루밍, 루머 도메인에 대한 실험에서 각각 기존의 모델들을 뛰어넘는 state-of-the-art의 성능을 달성하였다.

주제어: 실시간 조기 위험 감지, 강화학습, 문맥 기반 텍스트 분류

1. 서론

온라인 소셜 네트워크의 발전으로 많은 사람들은 익명의 사람들과 다양한 방식으로 상호작용을 하게 되었고 그만큼 악의적인 위험에 노출될 확률이 높아졌다. 예를 들어, 온라인 그루밍과 루머가 그 대표적인 케이스이다. 온라인 그루밍[1]이란 잠재적 성착취자가 온라인에서 피해자들과 감정적인 유대감을 쌓고 그들을 성적인 목적으로 이용하는 범죄를 의미한다. 또 온라인 루머는 한 사람에서 다른 사람으로 어떤 사건이나 공공의 이슈에 대한 그럴듯한 거짓이 퍼지는 현상을 의미한다. 이렇게 온라인 그루밍과 루머와 같이 시간에 따라 변화하는 위험들은 그림 1과 같이 정확성 뿐만 아니라 위험을 판단하는 타이밍이 매우 중요하다. 그러나 많은 경우 기존의 연구들은 신속성에 대한 고려가 부족하다. 예를 들어 기존의 온라인 그루밍[2], 루머 탐지 모델[2]의 경우 특정 발화만 활용하거나, 혹은 대화에 존재하는 모든 발화를 참고함으로써 예방보다는 법적인 증거채택의 목적을 두고 있다. 신속성을 고려한 몇 가지 연구도 존재하지만 rule-based 방법[3]으로 접근하거나, 문맥을 파악하지 못하는 한계점이 있다. 따라서 본 논문은 **실시간 조기 위험 검출 (Real-Time Early Risk Detection; RERD)**이라는 문제를 정의하고, 이 문제를 풀기 위한 새로운 방법론을 제시한다. 본 논문의 기여를 요약하면 다음과 같다.

- RERD 문제를 정의하고, RERD 특징을 공유하는 서로 다른 도메인을 묶어 하나로 정의한다.
- RERD 문제에 대한 데이터셋과 평가지표를 제안하고, 강화학습을 기반으로 정확성과 신속성을 동시에 고려한 새로운 방법론을 제시한다.
- 기존의 베이스라인 모델들과 본 논문의 방법론을 실험을 통해 정량적으로 비교하고, 우리의 시스템이 새로운 state-of-the-art임을 증명한다.

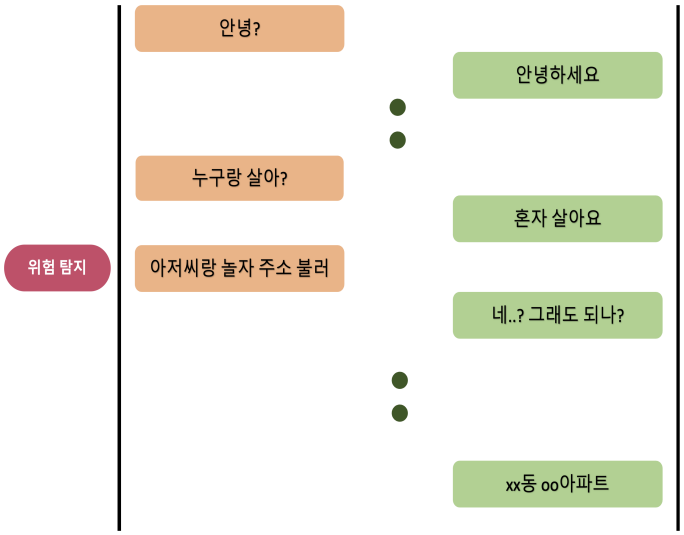

그림 1. 온라인 그루밍 조기 감지 시스템 작동 예시

2. 관련 연구

2.1 정확성을 고려한 위험 검출

온라인 그루밍 문제의 경우 기존의 연구[2]는 특정 발화만을 이용하거나 전체 대화를 참고함으로써 위험을 이진 분류한다. 마찬가지로 온라인 루머[4]의 경우 특정 게시물들을 읽고 위험을 판단한다. 그러나 공통적으로 예방보다는 법적인 증거 채택에 목적을 두고 있다.

2.2 신속성을 고려한 위험 검출

신속성을 고려한 몇 가지 연구도 존재한다. 온라인 그루밍의 경우 신속성을 고려해 PANC[3]라는 벤치마크 데이터셋을 제작했고, 슬라이딩 윈도우 기반 모델을 제안했다. threshold를 이용해 rule-based 방법론으로 조기 검출을 시도한 사례[5]도 있다. 온라인 루머 케이스에서는 조기 루머 검출을 위한 데이터셋을 제작했고 Neural Hawkes Process를 이용한 모델(HEARD)[6]을 제안했다. 온라인 그루밍과 유사하게 미리 정해놓은 threshold[7]를 이용해 조기 검출을 시도하기도 했다.

그러나 크게 3가지 관점에서 기존 연구는 한계점들이 존재한다. 첫째로 문맥을 제대로 고려하지 못하고 있다. 특정 시점의 발화 및 게시물은 이전 문맥을 반영해야 점진적으로 변하는 위험을 감지할 수 있다. 하지만, 기존 연구의 슬라이딩 윈도우 및 단순한 recurrent 모델은 문맥을 적절히 반영할 수 없다. 둘째, 조기 위험 검출을 목적으로한 훈련이 제대로 이뤄지지 않는다. 대부분의 연구에서는 신속성을 고려하여 학습된 모델 없이, 정확성 기반 모델을 가지고 추론단계에서 threshold를 통해 조기 탐지를 시도한다. 마지막으로 조기 위험 검출이라는 공통점을 가진 서로 다른 연구들이 같은 연구주제로 다뤄지지 않았다. 도메인 다르더라도 온라인 그루밍과 루머는 사실 같은 조기 위험 검출 문제로 정의될 수 있지만 기존의 연구는 이들을 다르게 바라보고 있다.

따라서 본 논문은 실시간 조기 위험 검출(RERD)이라는 새로운 문제를 제시하고, 이 문제를 풀기 위한 방법론인 RT-ERD 모델을 소개한다. 그루밍과 루머같이 기존의 서로 다른 문제들을 RERD 문제로 병합하고 문맥을 고려한 강화학습 기반의 RT-ERD 시스템을 제시한다. 그리고 RERD 문제를 위한 평가 지표도 제시하여 기존에 존재하던 모델과 본 논문의 모델을 비교한다.

3. 제안 방법

3.1 문제 정의

조기 위험 감지 문제는[8] 대화나, SNS 리트윗(Retweet)처럼 시간에 따라 변하는 데이터에 대해 실시간으로 위험성을 경고하는 것을 목적으로 한다. 이 때, 모든 대화 및 SNS 게시물들을 보지 않고 신속성과 정확성을 모두 고려하여 위험을 판단한다.

3.1.1 데이터 스트림(Data Stream)

조기 위험 감지는 시간에 따라 순차적으로 주어지는 데이터 스트림을 다룬다. 예를 들어, 대화 C의 경우 시간 순으로 주어지는 발화문장인 $\{utt_1, utt_2, ..., utt_n\}$ 로 정의되고, 게시물 모음 P의 경우 시간순으로 작성된 게시물인 $\{post_1, post_2, ..., post_n\}$ 로 정의된다. 편의상 C와 P를 모두 대화, utt_i와 $post_i$를 발화라고 정의한다. 이 때, 대화를 C라 한다면 전체 데이터셋은 $\mathbf{C} = \{C\}$ 로 표현될 수 있다. 발화 문장의 모음을 S 그리고 각각의 발화 문장을 u_i 라고 한다면 $S = \{u_1, u_2, ..., u_n\}$으로 정의할 수 있다. 그리고 각각의 C는 해당 S가 위험한지 아닌지를 판단하는 레이블이 존재한다. $C = \{S, y\}, y \in \{0, 1\}$. 이 때 $y = 1$을 위험(risk), $y = 0$를 정상(normal)이라고 정의한다.

3.1.2 현재 시점까지의 발화(Prefix)

$S = \{u_1, u_2, ..., u_t, ...\}$ 일 때, 현재의 시점을 t라고 한다면 처음부터 t까지의 모든 발화$S(t)$는 다음과 같이 정의할 수 있다. $S(t) := \{u_1, u_2, ..., u_t\}$ 그리고 이를 prefix라고 정의한다.

3.1.3 실시간 조기 위험 감지 시스템(RT-ERD)

$C \in X_{test}$ 를 만족하는 X_{test} 데이터가 주어졌다고 하자. RT-ERD 시스템은 그림 2처럼 각 대화 C에 대해서 $t = 1, ..., n$가 증가함에 따라 변화하는 $S(t)$를 분류하여 t시점에 위험을 정확하게 탐지하는 것을 목적으로 한다. RT-ERD 시스템이 특정 t 시점에서 위험이라는 결정을 하면 전체 대화 C를 위험이라 생각하고 t시점 이후 발화가 포함된 $S(t+1), S(t+2), ...$ 는 사용하지 않는다.

단, RT-ERD 시스템은 $S(t)$를 시점 t가 마지막 발화 시점인 n이 될 때 까지 정상으로 분류하지 않는다. 그 이유는 언제든 대화 C는 시간이 지남에 따라 위험으로 변할 가능성이 있기 때문이다.

3.2 위험 감지 모듈(Risk Detection Module; RDM)

위험 감지 모듈(RDM)은 AGHMN[9] 모델을 베이스라인으로 하고, 크게 3가지 레이어로 구성된다. 발화 문장 임베딩 레이어, 이전 발화의 문맥 정보를 현재 발화문장에 반영하는 어텐션 레이어, 그리고 발화문장의 위험성을 분류하는 분류 레이어로 구성된다.

3.2.1 문장 임베딩 레이어

대화 C를 이루는 각 발화 $\{u_1, u_2, ..., u_n\}$을 사전 학습된 트랜스포머 인코더를 이용해 $\{e_1, e_2, ..., e_n\}$로 임베딩 한다. 이 때, 트랜스포머 인코더로써 $BERT_{base}$[10]를 사용한다. $< CLS >$ 토큰을 각 발화 u_i에 대한 임베딩 벡터 $e_i \in \mathbb{R}^{d_e}$로 사용하고,

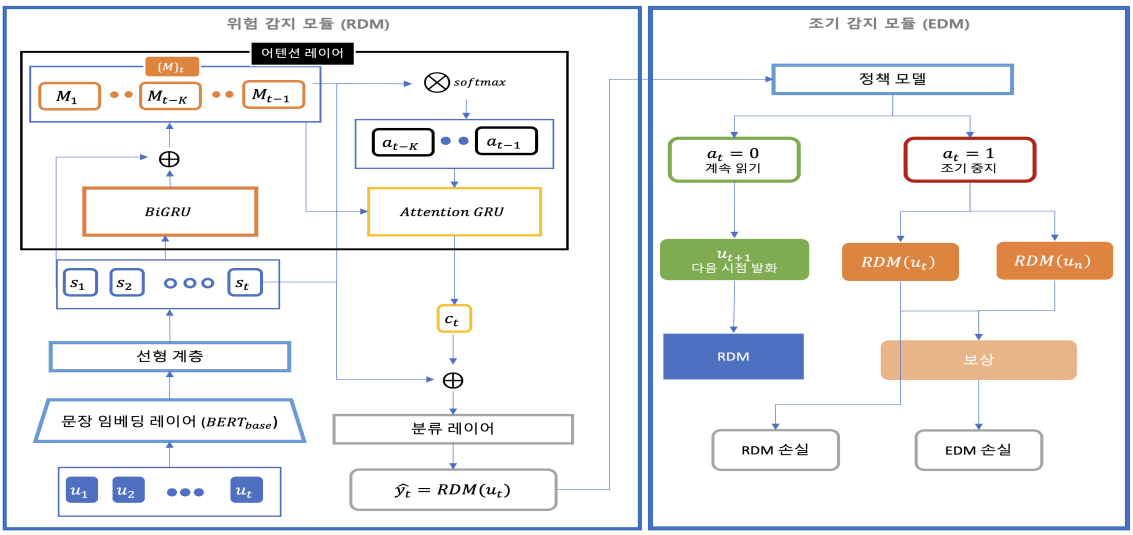

그림 2. RDM과 EDM으로 구성된 실시간 조기 위험 감지 시스템(RT-ERD)

최종적으로 임베딩 벡터를 선형 계층 $W_s \in \mathbb{R}^{d_e \times d_s}$에 통과시켜 문장 임베딩 벡터 $s_i \in \mathbb{R}^{d_s}$를 얻는다. 즉, $s_i = W_s(e_i) + b_s$로 정의할 수 있다.

3.2.2 어텐션 레이어

우선 현재 시점의 발화 문장 이전의 문맥 정보를 저장하기 위해 이전 K개의 문장발화로 이루어진 메모리 $\{M\}_t$를 구축한다. 그림 2와 같이 최근 K개의 문장 임베딩 벡터를 Bi-GRU의 입력으로 넣어 문맥정보를 보존하는 메모리 벡터 $\{M\}_t$를 제작한다.

$$\{M\}_t = \{BiGRU(s_{t-K-1+k}) + s_{t-K-1+k}\}_{k=1}^{K} \quad (1)$$

그리고 현재 시점의 s_t와 이전 발화의 문맥정보인 $M_k \in \{M\}_t$와의 상관관계를 알기 위해 어텐션 가중치 a_k를 구한다.

$$a_k = \{softmax(s_t \cdot M_k)\}_{k=t-K}^{t-1} \quad (2)$$

어텐션 GRU[9]를 활용해 현재 시점의 발화문장을 표현하는 벡터 c_t를 구한다. 어텐션 GRU의 경우 어텐션 가중치 a_k를 업데이트 게이트로 활용하고, 내부 hidden state인 $\tilde{h}_t$를 업데이트 한다.

$$h_k = a_k \circ \tilde{h}_k + (1 - a_k) \circ h_{k-1} \quad (3)$$

$$c_t = AttetionGRU(M_{t-1}, h_{t-2}) \quad (4)$$

3.2.3 분류 레이어

분류 레이어에서는 이전 문맥을 반영한 현재 시점의 발화문장 벡터 c_t와 현재 시점의 발화 문장 임베딩 벡터 s_t의 합을 입력으로 받고 선형 계층과 $softmax$ 함수를 통해 발화문장의 위험성을 계산하고, 크로스 엔트로피 손실 함수를 이용해 RDM를 훈련시킨다.

$$\hat{y} = softmax(W_c(c_t + s_t) + b_c) \quad (5)$$

3.3 조기 감지 모듈(Early Detection Module; EDM)

조기 감지 모듈(EDM)은 어느 시점에서 RDM이 멈춰야 하는지 결정한다. 이 때, 강화학습 방법론 중 REINFORCE 알고리즘[11]을 활용하는데 그 이유는 replay memory[12]를 사용하지 않으므로써 빠른 수렴을 하기 위함이다. REINFORCE 알고리즘은 식 (6)처럼 Monte-Carlo[13] 방법을 통해 구한 total estimated return G_t을 이용해 정책 모델(policy model) π_θ의 파라미터 θ를 업데이트하는 과정이다. 이 때, 정책 모델 π_θ은 현재 상태 s_t와 action a_t를 매핑 시켜준다.

$$\theta \leftarrow \theta + \alpha\gamma^t G_t \nabla_\theta ln\pi_\theta(a_t|s_t) \quad (6)$$

이 때, 본 논문은 REINFORCE 알고리즘의 고질적인 문제인 높은 분산(variance)을 줄이기 위해 baseline value[14]를 활용하고 advantage function $A(s_t, a_t)$을 baseline value로 정한다. advantage function $A(s_t, a_t)$은 state-action value function $Q(s_t, a_t)$에서 state value functions $V(s_t)$를 뺄셈연산으로 얻을 수 있다. 최종적으로 total estimated return G_t와 advantage function $A(s_t, a_t)$를 뺄셈을 하여 식 (8)와 같이 정책 모델 π_θ의 파라미터 θ를 업데이트 한다.

$$A(s_t, a_t) = Q(s_t, a_t) - V(s_t) \quad (7)$$

$$\theta \leftarrow \theta + \alpha\gamma^t(G_t - A(s_t, a_t))\nabla_\theta ln\pi_\theta(a_t|s_t) \quad (8)$$

전체적인 과정은 t시점의 발화 u_t가 정책 모델 π_θ의 입력으로 들어올 때, 정책 모델 π_θ가 $a_t = 0$ 를 출력하면 다음 시점의 발화를 읽고, $a_t = 1$를 출력하면 발화를 읽기 중단하고 현재 시점에서 전체 대화 C가 위험하다고 판단하는 것이다.

3.4 훈련 방법

3.4.1 RDM 훈련

우선 RDM은 크로스 엔트로피 손실 함수를 기반으로 먼저 훈련이 된다. 각 대화 C에 포함된 모든 시점의 발화 $\{u_1, u_2, ..., u_n\}$를 보고 RDM를 먼저 훈련시킨다.

3.4.2 EDM 훈련

EDM은 사전 학습된 RDM의 정확도와 표 1과 같이 보상(reward)을 통해 훈련된다. 각 대화 C에 대해 현재 시점 t의 발화 u_t가 입력으로 주어졌을 때, RDM은 u_t에 대한 hidden state h_t와 위험성에 대한 정확도 p_t를 출력한다. EDM은 h_t와 p_t를 입력으로 받고 정책 모델을 통해 action value a_t를 출력한다. 이 때, EDM은 발화를 계속 읽거나 $a_t = 0$, 계속 읽기를 중단 $a_t = 1$하는 결정을 내린다. 그리고 표 1과 같이 EDM가 결정한 action과 RDM의 정확도 예측을 기반으로 총 4가지 종류의 보상(reward)을 받게 된다. 이 때 $RDM(u_t)$은 t시점에서 RDM이 예측한 정확도이고 $RDM(u_n), n = len(S)$은 모든 발화를 봤을 때, RDM이 예측한 정확도이다

표 1. EDM 보상(Reward) 설계

Action Value	RDM의 예측	보상
a = 1	$RDM(u_t)$ =정상	M
	$RDM(u_t) \neq RDM(u_n)$ =정상	$-P$
	$RDM(u_t) \neq RDM(u_n)$ =위험	$-p$
a = 0 or 1	나머지 경우	$-\epsilon$

t시점까지의 발화문장 u_t, 마지막 시점의 발화문장 u_n에 대한 RDM의 예측 정보가 활용되고 보상 $M > 0$, $P \gg 1$, $p > 0$, $0 < \epsilon \ll 1$이 주어진다. 이러한 보상을 바탕으로 식 (8)에 의해 EDM은 학습된다.

4. 실험 및 결과

4.1 실험 데이터

본 논문은 PAN12, PANC, PHEME 총 3가지의 데이터셋을 사용한다. PAN12[15]는 2012 CLEF 컨퍼런스에서 PAN Lab이 공개한 데이터셋으로 온라인 그루밍 문제를 풀기 위해 제작된 데이터셋이다. 두 명 이상의 인물로 구성된 인터넷 채팅 형식으로 이뤄져 있고, 각 대화는 온라인 그루밍 인지 아닌지 이진 레이블링 되어있다. 전체 데이터셋 중 2.58% 만이 온라인 그루밍 대화이다. 따라서 정상 레이블과 위험 레이블의 비율을 2:1[5]로 조정하였고, 각 대화의 평균 발화문장 수는 63개이다. PANC[3]는 온라인 그루밍 판별 문제에서 정확도 뿐만 아니라, 신속성도 평가하기 위해 제작된 데이터셋이다. PAN12와 다른 점은 온라인 그루밍으로 레이블링 된 평균 발화 2248개의 긴 대화가 있다는 점이다. 이는 충분한 대화시간을 제공하기 때문에 위험을 판단하는 신속성을 평가하기에 용이하다. 마찬가지로 정상과 위험 레이블의 비율을 2:1로 조정했다. PHEME[16]은 SNS 상에 존재하는 루머 감지를 위한 데이터셋으로, 평균 15개의 게시물로 이뤄져 있다.

표 2. 각 실험 데이터의 훈련, 검증, 평가 데이터 개수

데이터셋	훈련 데이터	검증 데이터	평가 데이터
PAN12	2367	261	4589
PANC	5091	630	4563
PHEME	3314	409	420

4.2 비교 모델

PAN12, PANC, PHEME 데이터를 대표하는 state-of-the-art 베이스라인과 본 연구의 모델을 비교한다. 우선 PAN12를 대표하는 모델은 AGHMN 기반 threshold[5]를 이용한 조기 감지 모델이다. AGHMN의 t시점 발화 u_t의 정확도가 특정 threshold가 넘어가면 조기 위험을 감지하는 모델이다. PANC를 대표하는 모델은 슬라이딩 윈도우 기반 BERT이다. 슬라이딩 윈도우 안의 발화 문장들을 이진 분류하고 최근 K개의 슬라이딩 윈도우 중 s(skepticism)개 이상이 위험으로 분류 되었을 때 중단 후 조기 감지를 하는 모델이다. PHEME을 대표하는 모델은 HEARD모델이다. LSTM를 기반으로 발화문장에 대한 위험을 감지하고 Neural Hawkes Process[17]를 이용해 t시점의 u_t가 위험일 때, u_t 이후의 어떠한 발화도 정상 레이블로의 변화가 일어나지 않는 것을 확신할 때, 중단 후 조기 감지를 하는 모델이다.

4.3 실험 세팅

RT-ERD 모델과 BERT, AGHMN+threshold 모델은 학습 시 동일한 하이퍼파라미터를 적용하였다. 학습률(learning rate)는 $1e-4$ 배치 사이즈 (batch size)는 16, 에폭(epoch)은 30이다. 옵티마이저로 Adam[18]을 사용하여 최적화를 했다. HEARD 모델의 경우 학습률(learning rate)은 $2e-4$ 배치 사이즈(batch size)는 16, 에폭(epoch)은 30으로 정했고 옵티마이저로 Adam을 사용했다. AGHMN+threshold 모델의 경우 threshold를 0.8로 정했다. 윈도우 기반 BERT의 경우 윈도우 사이즈를 50으로 정했고, skepticism은 5으로 정했다. RT-ERD 모델과 AGHMN+threshold 모델의 어텐션 레이어를 구성하는 GRU의 hidden 차원은 100으로 구성되어 있고, 메모리 모듈 $\{M\}_t$에 존재하는 최근 발화 문장 수 K는 50으로 설정하였다. EDM을 위한 보상의 경우 $M = (RDM(u_t) \cdot (1 - penalty)) \cdot 2$,

$P = 3$, $p = 1$, $\epsilon = 0$ 로 설정했다. 이 때, *penalty*는 식 (9)로 정의된다.

4.4 평가 방법

정확성과 신속성을 하나의 지표로 평가하기 위해 본 논문은 Latency F1을 사용한다. Latency F1[19]은 위험 검출의 정확성을 대표하는 F1과 신속성을 대표하는 penalty를 곱한 평가 지표이다.

$$penalty(l) := -1 + \frac{2}{1 + exponent(-p \cdot (l-1))} \quad (9)$$

$$speed := 1 - median\{penalty(l) | l \in latency\} \quad (10)$$

$$LatencyF1 := F1 \cdot speed \quad (11)$$

l은 latency로 RT-ERD 시스템이 참고한 발화 문장 수를 의미하며 p는 latency가 증가할 때 얼마나 많은 불이익을 줄지 결정하는 하이퍼파라미터이다. 이러한 penalty를 바탕으로 speed를 정의한다. Speed는 위험 검출의 속도를 정량화 한 지표로 평가 데이터 X_{test}를 이루고 있는 각 대화 중 시스템이 위험으로 조기 탐지한 대화 가운데 실제로 위험으로 분류되는 대화 C를 기준으로 한다. 조건을 만족하는 각각의 C에 대해서 penalty들의 중앙값을 구해 speed를 구한다. 최종적으로 모델이 조기 검출했을 때의 micro F1와 speed를 곱해 Latency F1을 구한다. 따라서 본 논문에서는 Latency F1을 주 평가지표로 사용하고 보조 지표로 micro F1과 speed를 사용한다.

4.5 실험 결과 및 분석

표 3은 각각의 모델들의 총 3가지 데이터셋에 대한 성능을 보여주며 결과를 크게 아래 3가지 측면에서 분석한다.

4.5.1 조기 위험 검출 문제 정의의 정당성

본 연구는 서로 다른 도메인에서 독립적으로 존재하던 베이스라인 모델들을 가지고 서로의 데이터셋에 대한 성능을 측정했다. 그 결과 모든 모델들은 대부분 micro F1과 speed 모두 기존의 베이스라인 성능과 비슷한 0.8 이상의 준수한 성능을 보였다. 이는 각 모델과 데이터셋이 공통의 특징을 지니고 있다는 의미이고, 따라서 각 문제들을 하나의 문제로 정의한 것은 적절하다는 사실을 알 수 있다.

4.5.2 Latency F1

모든 도메인에 대해서 본 연구의 방법론인 RT-ERD 모델이 state-of-the-art의 성능을 보였다. 이로써 위험을 조기 탐지하는데 있어서 강화학습 기반의 정책 모델이 실제로 학습 가능하다는 사실을 알 수 있고, 성능은 기존의 방법론 보다 훨씬 뛰어나다는 것을 증명했다.

4.5.3 정확성과 신속성 간의 상관관계

대체로 micro F1이 커지면, speed가 작아지고 반대의 경우도 마찬가지로 측정된다. 따라서 정확성과 신속성간의 관계는 trade-off 라는 사실을 알 수 있다. 하지만, micro F1이 비슷할 때, speed에서 큰 차이가 나는 경우가 존재한다. 예를 들어 PAN12에서 우리의 모델과 AGHMN+threshold의 결과가 대표적이다. 이 경우를 통해 무조건 더 많은 발화를 본다고 해서 정확성가 오르지는 않는다는 것을 알 수 있고, 신속성과 정확성이 균형을 이루는 최적의 시점이 존재함을 이 결과는 시사하며, 그 시점이 이 연구의 목표임을 알 수 있다.

표 3. 베이스라인 모델과 본 연구 시스템에 대한 평가 결과
BERT*: 본 연구에서 재현한 모델, Latency F1과 speed는 코드가 공개되지 않아 측정할 수 없었음

데이터셋	모델	Latency F1	micro F1	speed
PAN12	RT-ERD	**0.905**	0.966	**0.939**
	AGHMN+th	0.877	0.966	0.908
	BERT*	-	0.943	-
	HEARD	0.793	**0.970**	0.818
PANC	RT-ERD	**0.893**	0.942	**0.948**
	AGHMN+th	0.849	**0.965**	0.880
	BERT[3]	0.840	0.910	0.923
	HEARD	0.795	0.942	0.844
PHEME	RT-ERD	**0.669**	0.800	**0.836**
	AGHMN+th	0.485	0.797	0.609
	BERT*	-	**0.820**	-
	HEARD	0.367	0.757	0.485

5. 결론

본 연구는 실시간 조기 위험 검출 문제(RERD)을 정의하고 RERD의 특성을 가지는 서로 다른 종류의 도메인(온라인 그루밍, 루머)을 하나로 정의했다. 그리고 위험 감지 모듈(RDM)과 조기 감지 모듈(EDM)를 통해 위험 감지에 있어서 신속하면서도 정확성이 보장되는 시스템을 제안했다. 특히, GRU와 attention을 활용해 이전 발화 문맥을 활용하고, 강화학습 방법론 중 REINFORCE 알고리즘을 통해 최적의 조기 감지 타이밍을 찾았다. 또한 기존의 state-of-the-art의 성능을 가지고 있던 베이스라인 모델들과 본 연구의 방법론을 Latency F1이라는 평가지표로 비교했고, 우리의 모델이 도메인을 초월하여 최고의 성능을 만들어낸다는 사실을 확인했다.

감사의 글

본 연구는 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2021-0-00575, 음성 · 텍스트 딥러닝 기반 보이스피싱 예방 기술 개발)과 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음(IITP-2023-2020-0-01789)

참고문헌

[1] S. Wachs, K. D. Wolf, and C.-C. Pan, "Cybergrooming: Risk factors, coping strategies and associations with cyberbullying," *Psicothema*, pp. 628–633, 2012.

[2] P. Bours and H. Kulsrud, "Detection of cyber grooming in online conversation," *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6, 2019.

[3] M. Vogt, U. Leser, and A. Akbik, "Early detection of sexual predators in chats," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4985–4999, 2021.

[4] N. Ruchansky, S. Seo, and Y. Liu, "Csi: A hybrid deep model for fake news detection," *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 797–806, 2017.

[5] S. Bae and G.-B. Lee, "Incremental early text classification system for early risk detection," *Annual Conference on Human and Language Technology*, pp. 91–96, 2021.

[6] F. Zeng and W. Gao, "Early rumor detection using neural hawkes process with a new benchmark dataset," *arXiv preprint arXiv:2306.02597*, 2023.

[7] C. Song, C. Yang, H. Chen, C. Tu, Z. Liu, and M. Sun, "Ced: credible early detection of social media rumors," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 33, No. 8, pp. 3035–3047, 2019.

[8] J. Parapar, P. Martín-Rodilla, D. E. Losada, and F. Crestani, "Overview of erisk 2022: early risk prediction on the internet," *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 233–256, 2022.

[9] W. Jiao, M. Lyu, and I. King, "Real-time emotion recognition via attention gated hierarchical memory network," *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34, No. 05, pp. 8002–8009, 2020.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[11] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," *Advances in neural information processing systems*, Vol. 12, 1999.

[12] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[13] N. Metropolis and S. Ulam, "The monte carlo method," *Journal of the American Statistical Association*, Vol. 44, No. 247, pp. 335–341, 1949. [Online]. Available: http://www.jstor.org/stable/2280232

[14] L. Weaver and N. Tao, "The optimal reward baseline for gradient-based reinforcement learning," *arXiv preprint arXiv:1301.2315*, 2013.

[15] G. Inches and F. Crestani, "Overview of the international sexual predator identification competition at pan-2012." *CLEF (Online working notes/labs/workshop)*, Vol. 30, 2012.

[16] A. Zubiaga, G. Wong Sak Hoi, M. Liakata, and R. Procter, "Pheme dataset of rumours and non-rumours," 2016.

[17] H. Mei and J. M. Eisner, "The neural hawkes process: A neurally self-modulating multivariate point process," *Advances in neural information processing systems*, Vol. 30, 2017.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] F. Sadeque, D. Xu, and S. Bethard, "Measuring the latency of depression detection in social media," *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 495–503, 2018.