https://doi.org/10.7840/kics.2024.49.4.489

Transformer 기반 LLM의 학습을 이용한 112 허위오인신고분류·예측모델 개발

정 재 훈°, 박 현 호*

Development of a False Alarm Classification and Prediction Model Using Transformer-Based Large Language Model

Jae-hoon Jeong°, Hyunho Park*

요 약

112 긴급신고 시스템은 국민 안전을 위한 경찰의 최전선으로, 신속한 출동 및 사건처리가 무엇보다도 중요하다. 허위오인신고는 경찰력이 낭비될뿐 아니라 진정으로 도움을 필요로 하는 상황에 대응하기가 어려워진다는 점에서 그 문제가 크다. 최근들어 증가하고 있는 허위오인신고에 대응하기 위하여 본 연구자는 딥러닝을 기반으로한 112 허위오인신고 분류·예측모델을 제안한다. 본 모델은 112상황실 접수요원이 요약한 신고내용 텍스트를 입력받아 해당 신고의 허위오인여부를 결정하게 된다. 악의적 허위신고와 오인신고중 후자는 신고접수자 입장에서 표면적으로 알아채기가 불가능에 가깝다. 이로 인해 연구자는 허위오인신고 전체 데이터와 악의적인 허위신고만을 담은 데이터를 나누어 실험하였다. 트랜스포머 구조를 가진 BERT, KoBERT,ELECTRA, KoELECTRA, RoBERTa 5가지 모델에 대하여 동일한 하이퍼파라미터로 모델 훈련을 진행했다. 본 연구는 자연어처리와 LLM을 이용하여 경찰의 실제 치안업무에 대해 문제해결적 접근을 수행했다는 점에 의의가 있다. 본 연구의 결과가 악의적인 허위신고를 빠르게 식별하고 경찰이 의사결정을 지원하는데에 도움이 될 것으로 기대한다.

키워드 : 트랜스포머, 자연어처리, 112, 허위신고, 이진분류
Key Words : Transformer, NLP, 112, False Alarm, Binary Classification

ABSTRACT

The 112 emergency reporting system is the police's front line for public safety, and rapid dispatch and incident handling are of the utmost importance. False reports are problematic in that they not only waste police power, but also make it difficult to respond to situations that need help. In order to respond to the increase in false reports, this researcher proposes a 112 false report classification and prediction model based on deep learning. This model receives the report text summarized by the 112 situation room receptionist and determines whether the report is false or misidentified. Mistaken reports are almost impossible to detect on the surface from the point of view of the person who filed the report. Because of this, the researcher conducted an experiment dividing data containing all false reports and data containing only malicious false reports. Model training was performed with the same hyperparameters for five models with transformer structures: BERT,

※ 이 논문은 2024년도 정부(경찰청)의 재원으로 지원받아 수행된 연구결과임 [내역사업명: 112 긴급출동 의사결정 지원 시스템/ 연구개발과제번호: PR08-03-000-21]
°• First and Corresponding Author : Koraen National Police University, 20210042@police.ac.kr, 학생회원
* Electronics and Telecommunications Research Institute, hyunhopark@etri.re.kr, 정회원
논문번호 : 202311-145-A-RN, Received November 18, 2023; Revised December 13, 2023; Accepted December 19, 2023

KoBERT, ELECTRA, and RoBERTa. This study is significant in that it took a problem-solving approach to the police's actual security work using natural language processing and LLM. It is expected that the results of this study will help identify malicious false reports and support police decision-making.

Ⅰ. 서 론

경찰의 112 신고서비스는 개인적 사회적 법익을 수호하고 치안을 유지하는 경찰작용의 핵심이다. 112 신고체계를 통해 경찰은 신속한 출동과 조치로 당면한 위험 또는 장해를 제거할 수 있다. 이는 이미 발생한 범죄행위를 사후적으로 조사하는 수사경찰의 업무와 명백히 구분된다. 112 현장출동 경찰관은 빠르고 효율적으로 움직여야만 한다. 신속한 대응은 현장 경찰관뿐 아니라 전체적인 치안상황을 조망하며 지령을 내리는 112 상황실의 몫이기도 하다.

112 신고접수에서 현장출동까지의 정보흐름은 상황실에게 이상적이지 않다. 신고대상 현장의 정보가 이미 신고자를 거쳐 한 번 왜곡되기 때문이다. 신고라는 시스템 자체의 한계이다. 주어진 정보는 현장의 당사자 혹은 제3자인 시민의 신고뿐이다.

사회 전체적으로 신고의식이 높아져 112 신고가 늘어나는 것은 긍정적인 현상이다. 그러나 상당수의 시민들이 112 긴급신고를 '경찰 호출벨'로서 사용하고 있는 실정이다. 경찰청 통계자료에 따르면 출동이 필요없는 단순상담등 신고가 전체 신고의 43.8%를 차지하고 있다. 경찰청은 비출동 민원상담신고번호인 182에 대해서 지속적으로 홍보하고 있으나 그 효과는 미미하다.

경찰이 제공하는 치안서비스는 공공재이다. 누구나 제한없이 이용할 수 있는 비배제적 서비스이다. 동시에, 일정 이상의 수요가 존재할 경우 일부는 해당 서비스를 이용할 수 없는 경합적 성격을 지닌다[1]. 경찰은 경합적인 치안서비스를 적재적소에 제공하기 위하여 그림 1[2]에서 볼 수 있듯이 112 신고에 대한 차별적 대응을 하고 있다. 긴급도에 따라 신고대응코드를 5단계로 나누어 긴급한 사건을 우선하여 대응한다.

비출동신고인 CODE 4의 경우 불필요한 현장출동까지 이어지는 사건이 적다. 신고접수 요원이 담당기능으로 이관 및 이첩할 수 있기 때문이다. 그러나 허위신고 및 오인신고의 경우는 다르다. 마치 실제로 긴급한 사건인 것과 같이 접수된다. 이는 전술한 바와 같이 신고시스템이 가진 한계에서 기인한다. 지난 5년간 112 신고시스템에 접수된 허위오인신고는 약 148만 건으로, 1년에 대략 29만건 이상이 허위오인신고로 접수된다.

경합적 서비스인 112 긴급신고에 있어 허위오인신고는 치명적이다. 악의적이거나 부주의한 신고로 인해 경찰력이 낭비된다. 경찰력의 낭비뿐 아니라 경찰이 긴급한 도움이 필요한 상황에 출동할 수도 없게 된다. 특히 악의적 허위신고의 경우 신고자들이 대체로 중대한 사건을 지어내기 때문에 문제가 더욱 크다.

허위신고 관련 통계가 작성된 이래로 112에 대한 허위오인신고는 매년 줄어가는 추세였다. 그러나 최근에 들어 다시금 허위오인신고가 증가하고 있다. 2023년 7월 25일에는 서울 송파구 롯데월드타워 100층에 폭발물이 설치되었다는 신고가 접수되었다. 수십명의 경찰, 소방인력이 건물을 수색했음에도 폭탄은 발견되지 않았다.

같은해 7월 9일에는 한 여성이 남자친구에게 성폭행당했다고 112에 허위신고 하였다. 허위신고의 이유는 남자친구가 연락을 받지 않아서 화가 났기 때문으로 밝혀졌다. 위 건 접수로 인해 경찰차 3대와 경찰관들을 동원해 피해자 구조에 나섰으나 결국 헛걸음하였다.

경찰은 신속한 출동시스템을 구축하고 효율적인 현장대응을 하기 위해서 경찰력의 낭비를 최대한 막아야 한다. 이에 본 논문은 자연어처리와 LLM을 이용하여 112 허위오인신고를 필터링하는 시스템을 제안한다. 제안하는 시스템은 112 신고가 접수되고, 하위

구분	개선 전		개선 후	분류 기준	출동목표시간
긴급	코드1	➡	코드0	코드1 중 이동범죄, 강력범죄 현행범 등의 경우(선지령 및 제반출동요소 공조출동) 예) 남자가 여자를 강제로 차에 태워 갔다, 여자가 비명을 지른 후 끊긴 신고	최단 시간 내
			코드1	생명·신체에 대한 위험이 임박, 진행 중, 직후인 경우 또는 현행범인인 경우 예) 모르는 사람이 현관문을 열려고 한다, 주차된 차문을 열어보고 다닌다	최단 시간 내
비긴급	코드2	➡	코드2	생명·신체에 대한 잠재적 위험이 있는 경우 또는 범죄예방 등을 위해 필요한 경우 예) 엄엄이 끊났는데 손님이 깨워도 일어나지 않는다, 집에 와보니 도둑이 들었는지 집이 난리다	긴급 신고 지장 없는 범위 내 가급적 신속 출동
			코드3	즉각적인 현장조치는 불필요하나 수사, 전문 상담 등이 필요한 경우 예) 언제인지 모르지만 금반지가 없어졌다, 며칠 전에 폭행을 당해 병원치료중이다	당일 근무시간 내
비출동	코드3	➡	코드4	긴급성이 없는 민원·상담 신고	타기관 인계

그림 1. 현행 112 신고분류 체계
Fig. 1. Current 112 Report Classification System

기능으로 지령이 하달되는 과정에서 해당 신고가 허위오인신고인지 여부를 판단한다. 본 연구자는 LLM을 파인튜닝하고, 이를 기반으로 이진분류 모델을 구축할 것이다. 이어서, 사용하는 언어모델과 데이터 전처리 기법에 따른 실험환경을 구축하여 어떤 모델이 가장 우수한 성능을 보이는지 비교해보고자 한다.

Ⅱ. 관련 선행연구

본 논문에서 제시하는 시스템과 관련된 선행연구는 크게 (1) 112 신고시스템 개선 연구, (2) 112 거짓신고 대처방안 연구로 나누어볼 수 있다. 112 신고데이터를 빅데이터 및 자연어처리 방법론으로 접근한 연구들도 다수 존재했다.

논문 [3]은 긴급, 비긴급, 비출동으로 나누어 사건을 5단계 세분화한 현행 112 신고분류체계에 대하여 실증분석하였다. 연구는 1년간 서울경찰청에서 처리한 112 신고사건을 기반으로 수행되었다. 저자는 긴급신고(코드0, 1)에 대하여 약 40%이상이 비긴급신고에 해당하는 등 긴급코드가 지침으로서의 기능을 하지 못함을 지적하였다. 긴급신고중 사실은 비긴급사건이었던 사례에서 허위/오인신고가 5.9%로 상당한 비율을 차지하였다.

논문 [4]에서는 국민의 관점을 기반으로 112 신고상황에 대한 긴급코드를 재편하는 연구를 진행하였다. 신고분류중 인적피해와 공공안전에 관련된 신고에 대해서 국민이 경찰보다 더 민감한 반응을 보였다. 경찰의 경우 물적 피해와 불분명한 신고에 대해서 더 높은 긴급도를 지정하였다. 이어진 현장경찰관 인터뷰는 112 신고에 있어 긴급성코드가 지침으로서의 역할을 하고 있지 못함을 보였다.

112 거짓신고 대처방안 관련 논문인 [5]는 신고자의 기존 신고이력등을 축적하여 거짓신고 평가에 참고할 것을 제안했다. 위 연구는 거짓신고가 의심될시 처벌고지를 명확화하여 차후 신고분류에 이용해야 함을 제언한다. 논문 [6]은 실제신고와 거짓신고의 통화중 음성 특징정보의 차이를 연구하였다.

112 신고데이터를 빅데이터, 자연어처리 및 딥러닝 등 방법론으로 접근한 연구도 존재한다. 논문 [7]은 112 신고접수자료에 대하여 텍스트마이닝 분석을 진행하였다. 저자는 신고유형(절도, 폭력, 기타형사법, 주취자 등)별로 워드클라우드를 생성하여 키워드를 시각화하였다.

논문 [8]은 112 신고내용과 종결시 사건코드를 이용하여 신고내용 텍스트를 벡터화하고 군집분석하였다. 위 연구는 형성된 군집들 간 거리를 통해 서로 다른군집으로 분류되었음에도 유사한 양상을 보이는 사건들의 관계를 규명하였다.

이외에도 해외의 경찰 등 긴급신고시스템에서의 오남용 문제를 다룬 문헌도 존재한다. 논문 [9]는 캐나다 퀘벡 주의 거짓 긴급신고의 녹취를 담화분석하였다. 위 분석을 통해 거짓신고의 경우 신고접수자의 추가적인 정보요청이나 질문에 제대로 대답하지 못하는 경우가 많음을 지적했다. 연구자는 이러한 특징에 착안하여 거짓신고 판별을 위한 의사결정나무 방법론을 제시한다.

논문 [10]에서는 긴급신고의 특성을 기반으로 기존의 이진분류가 아닌 삼진분류방식을 제안한다. 긴급신고는 허위신고로 판단되더라도 실제로 긴급한 상황일 경우를 놓치게 되면 큰 문제가 발생한다. 이를 보완하기 위해 허위신고로 분류된 사건들에 대하여 다시 Reliability Veryfying을 통해 일반신고로 분류되도록 하는 방법을 제시했다.

Ⅲ. 연구 대상 데이터

3.1 데이터 개요

본 연구의 분석 대상 데이터는 비식별화된 112 신고접수 및 현장대응 데이터로, 신고내역 총 999,995건을 포함하고 있다. 데이터의 상세는 아래 표 1과 같다.

112 신고접수 데이터 중 대응코드의 경우는 그림 1의 각 코드에 해당하며 비긴급인 CODE 4는 제거하였다. 접수사건 종별은 신고가 접수될 당시 112상황실에서 해당사건의 유형을 태깅한 것이다. 연구 데이터에 존재하는 접수사건은 총 49종이다. 종결종류는 신고출동 후 현장조치의 유형에 따른 대분류이며 총 7가지이다. 종결종류 세부구분은 각 대분류에 대한 세부구분으로, 현장종결-현장조치, 검거-체포등을 예시로 들 수 있다.

표 2에는 연구 대상 데이터의 텍스트 컬럼과 그 예시가 있다. 신고접수 내용은 112 상황실 접수요원이 신고자와 전화통화하며 정리한 텍스트이다. 위 텍스트에는 신고상황의 핵심과 통화중 발화내용이 요약되어 있다. 뿐만 아니라 112상황실에서 현장경찰관에게 보내는 요청사항도 포함하고 있다. '출동 시 신고자 통화 요망', '허위신고시 처벌 요망'등을 예로 들 수 있다. 현장처리 내용은 112 신고출동한 현장경찰관이 신고조치 이후 보고하는 내용을 담고 있다.

표 1. 연구 대상 데이터 상세(범주형)
Table 1. Detailed Description of Data (Categorical)

No.	Variable	Variable Varient	Freq.	Percentage	Cumula-tive Per.
1.	대응코드	C0	7,397	0.73	0.73
		C1	298,437	29.85	30.58
		C2	620,725	62.07	92.65
		C3	73,436	7.35	100.00
2.	접수사건 종별	기타형사범	151,820	15.18	15.18
		보호조치	150,391	15.03	30.21
		시비	89,437	8.94	39.15
		교통사고	69,017	6.90	46.05
		폭력	62,161	6.21	52.56
		…	…	…	…
3.	종결종류	현장종결	556,851	55.69	55.69
		비출동종결	183,711	18.37	74.06
		미처리	89,678	8.97	83.03
		계속조사	55,182	5.52	88.55
		검거	44,929	4.50	93.05
		인계종결	34,975	3.49	96.54
		허위오인	34,669	3.46	100.00
4.	종결종류 세부구분	현장조치	504,435	50.44	50.44
		동일	178,979	17.90	68.34
		타부서인계	48,719	4.87	72.21
		불발견	37,762	3.78	75.99
		상담안내	35,608	3.56	79.55
		…	…	…	…

표 2. 연구 대상 데이터 상세(텍스트)
Table 2. Detailed Description of Data (Texts)

No.	Variable	Examples
1.	신고접수 내용	1. 가게 입구/노숙자가 누워있다고 2. 싸워서 시끄럽다며 // 3. 차 대 보행자 교통사고/ …
2.	현장처리 내용	1. 귀가조치 함 2. 친구들끼리 대화중이었던 것… 3. 교통사고접수하여 가해자 및… …

3.2 데이터 분석

본 연구자는 연구주제인 '허위오인신고 분류·예측 시스템'의 구현을 위한 실험에 앞서, 연구 대상 데이터를 통하여 허위오인신고의 현황을 대략적으로 파악하고자 한다.

앞선 데이터 개요에서 볼 수 있다시피 112 신고 종결종류에는 '허위오인'이라는 항목이 존재한다. 그러나 모든 허위오인신고가 해당 항목으로 분류되지는 않는다. 예를 들어 악의적인 허위신고로 인해 해당 신고자가 경범죄처벌법으로 즉결심판의 대상이 되거나 또는 공무집행방해로 체포되는 경우 각각 검거-즉결심판, 검거-체포 등으로 분류된다. 종결종류는 다르게 태깅되었으나 현장처리 내용에만 해당 신고가 허위/오인이었음을 기록하는 경우도 존재한다. 허위오인신고에 대하여 처벌하지 않고 경고조치만 취하는 경우를 예로 들 수 있다. 따라서 본 연구는 종결분류가 허위오인이 아니지만 실제로는 허위오인에 해당하는 경우 또한 허위오인으로 분류하였다. 아래 표 3은 이러한 데이터들의 예시를 보여준다.

실질 허위오인신고를 파악하기 위하여 전체 신고건 999,995건중 허위오인 분류에 해당하는 34,669건을 추출하였다. 이후 그중에서 비상벨 오작동에 해당하는 18,681건을 제외하였다. 마지막으로 표3에서와 같이 종결종류가 다르지만 실질적으로 허위신고에 해당하는 경우들을 합쳐 실질 허위오인신고를 정리하였다. 실질 허위오인 신고는 총 999,995건중 18,735건이다. 비율로는 1.87%으로 2021년 기준 허위오인신고율(2022, 경찰청) 1.60%와 유사하다.

다음으로는 허위오인신고로 판단된 신고건들이 가지는 특징적 분포를 확인한다. 표 4는 긴급코드별 전체신고분포와 허위오인신고분포, 그리고 허위신고율을 나타낸 것이다.

표 4에 따르면, 코드0은 전체 신고데이터의 0.73%에 불과하나 허위오인 데이터중에서 4.17%를 차지한다. 전체 코드0 신고중 허위오인신고의 비율은 10.56%이다. 긴급신고로 분류되는 코드1의 경우도 마찬가지이다. 코드1은 전체 신고데이터의 29.85%를 차지하나, 허위오인신고 데이터중 약 절반이 코드1에 해당한다. 허위오인 데이터의 긴급코드별 분포가 전체 모집단에 비해 높은 긴급코드에 몰려있음을 알 수 있다. 이는 허위오인신고의 상당수가 긴급사건에 해당하

표 3. 분류와 다르게 허위신고에 해당하는 경우
Table 3. Examples of False Reporting(not classified)

No	신고접수 내용	현장처리 내용	종결종류
1.	같이있는 사람이 흉기소지했다며/ 칼아닌 다른물건/ 그냥오라고만함	경범죄처벌법위반 허위신고로 현행범체포하여 형사계인계함	검거-체포
2.	술집에 들어가려 여자주인이 못들어오게 밀어서 입술에 상처났다며//	돈을 빌리러가서 업주가 나가라하자 폭행당했다 거짓신고한 것이다. 허위신고 즉결심판청구	검거-측결심판
3.	술먹고 사람을 때렸다며?/ 전화끊김	술에 취해 기분이 우울해서 신고했다하여 허위신고로 처벌될 수 있음을 강력계도	현장종결-현장조치

표 4. 긴급코드별 분포(전체/허위오인)
Table 4. Distribution by Emergency Code

No.	긴급코드	All	False report	False rate
1.	CODE0	7,397 (0.73%)	781 (4.17%)	10.56%
2.	CODE1	298,437 (29.85%)	9299 (49.63%)	3.12%
3.	CODE2	620,725 (62.07%)	8087 (43.17%)	1.30%
4.	CODE3	73,436 (7.35%)	568 (3.03%)	0.77%

여 신속한 출동을 요구함을 보여준다. 이러한 현상은 경찰력의 심각한 낭비와 비효율을 초래한다.

49종의 접수사건분류에 대한 허위오인신고율을 계산한 결과는 표 5, 6과 같다. 상하위 각 6개의 접수사건분류만을 표시하였다.

표 5에 따르면 허위오인율 상위 6개 사건분류는 각 납치감금, 살인, 강도, 아동학대, 절도, 가정폭력으로 강력범죄가 다수이다. 특히 상위 3개 사건분류의 허위오인율이 20% 이상으로, 전체평균을 크게 웃돈다. 상위 5, 6위에 해당하는 절도와 가정폭력의 경우 전체신고에서 차지하는 비율도 적지 않다. 표에는 드러나지 않으나 신고사례중 약 15%를 차지하고 있는 기타형사범 사건에 대하여도, 오인신고율은 3.1%로 전체평균을 상회한다.

표 6에 따르면 하위 6개분류는 주취자, 보호조치, 교통사고, 교통불편, 소음, 행패소란이다. 이들중 절반인 3개분류(주취자, 보호조치, 행패소란)가 주취자 관련 사건분류이며, 2개분류(교통사고, 교통불편)는 교통관련 사건분류이다.

표 5. 접수사건분류별 분포(허위오인율 상위)
Table 5. Distribution by Incident Classification(Top 6)

No.	접수사건분류	All	False report	False rate
1.	납치감금	780 (0.07%)	25 4(1.36%)	32.56%
2.	아동학대	875 (0.01%)	216 (1.15%)	24.69%
3.	살인	167 (0.01%)	38 (0.20%)	22.75%
4.	강도	148 (0.09%)	27 (0.14%)	18.24%
5.	절도	29,434 (2.94%)	3223 (17.20%)	10.95%
6.	가정폭력	22,383 (2.24%)	2065 (11.02%)	9.23%

표 6. 접수사건분류별 분포(허위오인율 하위)
Table 6. Distribution by Incident Classification(Low6)

No.	접수사건분류	All	False report	False rate
1.	주취자	15,444 (1.54%)	29 (0.15%)	0.187%
2.	보호조치	150,391 (15.04%)	283 (1.51%)	0.188%
3.	교통사고	69,017 (6.90%)	143 (0.76%)	0.207%
4.	교통불편	36,126 (3.61%)	79 (0.42%)	0.219%
5.	소음	49,643 (4.96%)	109 (0.58%)	0.220%
6.	행패소란	49,710 (4.97%)	158 (0.84%)	0.318%

허위오인율 하위 6개 사건분류들에 대해서는 전체사건에서 해당분류가 차지하는 비율이 상위집단에 비해 크다. 또한 주취자나 교통관련등, 신고자가 직접 현장을 확인하기에 용이하다. 반면 표5의 상위분류들은 직접 현장을 바라보거나 명확하게 판단하기 어렵다. 아동학대와 가정폭력의 경우는 옆집에서 아이가 울고 있거나 가족간의 다툼 소리를 듣고 신고하는 경우가 대부분이다. 상위분류중 절도의 경우는 보통 절도 현행범의 목격보다도 본인의 물건이 없어졌다는 신고가 다수이다. 이 경우에도 허위오인율이 높을 여지가 있다.

허위오인율 상하위분류의 특징들은 실제로 단어 빈도분석을 통해서도 확인된다. 연구대상 데이터의 신고접수내용을 토큰화하여워드클라우드를 생성한 결과는 그림 2, 그림 3과 같다. 그림 2의 경우는 전체 신고접수내용에 대한 워드클라우드이다.

전체 신고내용에서는 술, 손님, 취한, 쓰러져 등 주취자 관련 어휘가 두드러지게 나타난다. 더불어 차, 택시, 사고등 교통사고와 관련한 키워드도 등장한다. 이는 허위오인율 하위 신고분류의 특징을 나타내고 있음을 의미한다. 실제로 표 6의 여섯 가지 사건분류는 전체신고중 약 37%로 상당 부분을 차지하고 있다.

그림 3의 허위오인 신고접수내용에서는 전체내용과 대조되는 부분이 있다. 소리, 없어졌다와 같은 어휘가 강조되어 있다. 이는 목전에서 범죄행위 등이 일어나기보다는 간접적으로 상황을 인지하여 신고하는 경우가 많음을 의미한다. 뒤이어 '같다'라는 키워드 또한 특징적이다. 그림 2와 그림 3에서 공통으로 강조된 소리, 술, 연락, 집과 달리 '같다'와 '없어졌다'는 허위오인 신고에서 자주 등장한다. 신고자의 추측성

그림 2. 전체 신고접수내용 워드클라우드
Fig. 2. Word cloud of overall reports

그림 3. 허위오인 신고접수내용 워드클라우드
Fig. 3. Word cloud of false·misco -nfirmed reports

어미 사용이 신고내용의 허위오인여부 판단의 힌트가 될 수 있다.

다만, 강조의 정도는 다르나 주요 키워드라고 볼 수 있는 소리, 술, 집, 연락 등의 어휘는 전체신고에서도 상당부분 확인되었다. 허위오인신고 판별에 있어 위 어휘들은 유의미한 구분자가 되기 어려울 것이다.

Ⅳ. 허위오인신고 분류 모델 개발

4.1 Transformer 구조

본 연구는 Transformer 구조 기반의 LLM을 미세조정(fine-tuning)하여 112 허위오인신고 분류 작업에 이용한다.

Transformer는 Seq2Seq 모델의 단점을 보완하는 Attention Mechanism을 핵심으로 한다. 기존에는 문장의 맥락을 학습하기 위해 LSTM 등 RNN이 자연어처리 영역에서 주로 사용되었다. 그러나 병렬처리가 불가능하고, Gradient Vanishing 문제가 발생하는 등의 한계가 있었다.

Transformer 구조는 Multi-Head Attention layer를 이용해 Attention Mechanism을 구현한다. Transformer의 인코더-디코더 구조는 다음 그림 4와 같다[11].

여기서 각 Attention Layer과 순전파 신경망 Layer를 거칠때에, Residual Learning을 같이 진행하고 있다(Add&Norm). 어텐션을 수행하고 나온 값과 Residual connenction을 통해 바로 입력된 값을 동시에 처리하는 구조이다. 이를 통해 전체 신경망은 기존 정보를 입력받으면서 잔여된 부분만 학습하므로 수렴 속도가 빠르다.

그림 4[11]의 Multi-Head Attention을 살펴보면 세 개의 input이 Attention Layer에 입력되는 것을 볼 수 있다. Multi-Head Attention의 3가지 input에 따른 정보의 흐름은 다음 표 7과 같다.

Transformer 구조는 이어지는 토큰들을 하나의 Sequence로 입력받지 않고 병렬적으로 각 토큰에 대해 Attention을 진행한다. 이후 Positional Embedding을 통해 각 토큰의 위치정보를 보존한다. 이를 통해 효율적인 병렬처리가 가능하다.

이는 보다 긴 데이터를 처리하는데에도 효과적이다. 기존의 RNN은 입력 데이터간에 떨어져 있는 거리에 비례해 계산의 복잡도가 증가하며 전술한 Gradient Vanishing등의 문제가 발생한다. 그러나 Transformer는 모든 위치에 대해 동일하게 병렬적으로 Attention등의 절차를 수행하게 된다. 결과적으로 Transformer가 기존의 RNN, LSTM등 방법론에 비해 처리할 수 있는 데이터의 길이가 크다.

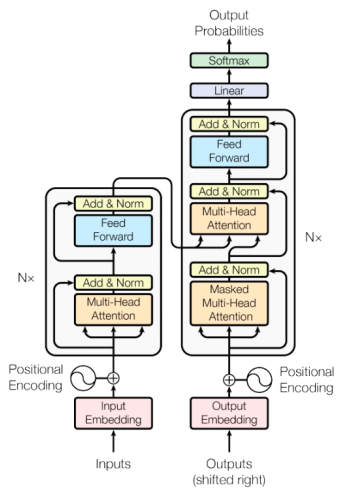

그림 4. Transformer 구조 도식
Fig. 4. Schematic of Transformer structure

표 7. Transformer의 Multi-Head Attention 작동
Table 7. Transformer's Multi-Head Attention operation flow

1	Query, Key, Value는 동일한 input embedding에 대한 복사본. 각 embedding 행렬을 Q, K, V라 하고 행렬K를 전치하여 Query와 Key를 multiplication. $Matmul. = QK^T$
2	Q가 $m \times n$행렬이라면 곱해진 행렬의 값에 대하여 scaling 수행. $sc = \frac{QK^T}{\sqrt{d}}$
3	scale 이후 SoftMax Layer에 통과시켜 Attention Score를 표준화. $Softmax\left(\frac{QK^T}{\sqrt{d}}\right)$
4	Softmax 취한 값을 행렬 V와 내적 $Softmax\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V$
5	최종 어텐션은 다음 수식과 같음 $atn(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_K}}\right)V$

4.2 분류 모델 개요

이 절에서는 본 연구에서 개발하고자 하는 모델의 구조에 대해서 이야기할 것이다. 본 연구에서 제시하는 모델은 Python의 머신러닝 라이브러리인 Pytorch를 이용하여 개발되었다. 개발 과정은 데이터 전처리 및 변환(토큰화), 하이퍼파라미터 지정, 모델 구축, 학습 순으로 이루어졌다. '112 긴급신고 허위오인신고 분류 모델'은 '신고접수 내용' 텍스트를 입력받아, 해당 신고가 허위오인신고인지 여부를 출력한다.

앞선 데이터 분석에서 볼 수 있다시피 '신고접수 내용'은 112 신고 접수요원이 신고자와의 통화내용을 정리한 텍스트이다. 위 텍스트에는 신고상황의 핵심과 통화중 발화내용이 요약되어 있다. 이 텍스트를 모델의 입력 텍스트로 사용한다. 표 2와 표 3에서도 볼 수 있듯이 '신고접수 내용' 텍스트는 완결된 문장이 많지 않다. 또한 '/'와 같은 특수문자로 각 문장이 구분되어 있다는 특징이 있다. 그러나 해당 텍스트에 대한 추가적인 전처리는 수행하지 않았다. 본 모델 구축에서 사용할 LLM들이 '/'와 같은 특수문자 또한 유의미한 구분자로서 이용할 수 있기 때문이다.

모델의 출력층에서는 1 또는 0을 반환하도록 한다. 이는 각각 긍정/부정을 의미한다. 즉, 허위오인여부 긍정인경우 1을, 허위오인여부 부정인경우 0을 반환하도록 할 것이다. 이는 다음 절에서 더 자세히 설명할 것이다.

모델의 입출력 다음으로, 본 연구에서 사용할 학습 방식에 대해 설명하고자 한다. 본 연구에서 제시하는 모델은 다양한 자연어처리 작업에 이용할 수 있는 LLM을 미세조정(fine-tuning)하여 허위오인신고 분류 task를 수행한다. 미세조정이란, 사전 학습된 모델의 파라미터를 조정하여 새로운 작업에 대한 특성을 학습하는 과정을 의미한다. 본 연구에서 제시하는 모델은 LLM의 layer들 (입출력층, 임베딩층, Transformer)의 구조를 변경하지는 않는다. 대신 train 과정에서 모델의 파라미터(가중치와 bias)를 수정하는 방식으로 미세조정이 이루어진다.

본 연구에서 사용하고자 하는 Transformer 구조 기반 LLM은 BERT-base, KoBERT, ELECTRA, KoELECTRA, RoBERTa 다섯가지이다. base 모델과 구조는 같으나 타겟언어가 다른 두 가지(KoBERT, KoELECTRA)를 제외한 base 모델들의 차이점에 대하여 설명하고자 한다. 세 가지 base 모델들(BERT-base, ELECTRA, RoBERTa)은 많은 차이가 있으나, 본 연구에서 제시하는 모델의 개발에 유의할 것이라 판단되는 학습과정의 차이점을 다룰 것이다.

BERT는 마스크 언어 모델링(MLM) 방식으로 사전학습된다. 이는 주어진 input 문장중 15%의 토큰을 무작위 마스킹하여 학습하는 방법이다. RoBERT도 동일한 방법으로 학습하나, 정적 마스크 대신에 동적 마스크를 이용한다. 기존 BERT에서는 한 문장에 대해 적용한 마스킹을 여러 epoch동안 동일하게 사용했으나, RoBERTa에서는 한 문장을 여러개 복사하여 서로다른 random한 mask를 적용하는 것이다. 이를 통해 더욱 효율적인 학습이 가능하다[12,13].

ELECTRA는 앞선 두 모델과 다르게 replaced token detection을 통해 사전학습한다. 이는 마스킹할 대상이 될 토큰들을 다른 토큰으로 바꾼 뒤, 이것이 원래의 토큰인지 교체한 토큰인지를 판별하는 방식이다. BERT와 RoBERTa의 경우 마스킹된 15%의 토큰 예측을 중심으로 모델 학습이 이루어진다. 반면에 ELECTRA는 모든 토큰을 대상으로 학습이 이루어진다. 이는 적은 데이터로도 BERT와 동일한, 혹은 그 이상의 성능을 보일 수 있음을 함의한다[14].

4.3 데이터 전처리 및 변환(토큰화)

허위오인신고 분류 모델 구축을 위해 사용할 데이터는 112 사건접수 및 현장처리 데이터 총 999,995건이다. 이들 중 실질 허위오인 신고로 분류된 18,735건이 타겟으로 사용되었다. 본 연구의 목적은 주어진 '신고접수 내용' 텍스트에 대하여 해당 신고가 허위오인신고인지 여부를 예측하는 이진분류(Binary

Classification)모델 구축이다. 따라서 반대되는 클래스의 데이터가 요구된다.

전체 데이터중 실질 허위오인에 해당하는 데이터의 비율은 약 1.87%이다. 분류 클래스 간의 불균형이 매우 심하기 때문에, 허위오인과 일반신고 데이터간 균형이 요구된다. 오버샘플링기법의 이용은 정보의 손실없이 더 높은 성능을 취할 수 있다는 장점을 가진다. 다만 허위오인 데이터의 엄밀성과 정확성이 흐려질 수 있고, 오버피팅의 가능성이 증가할 수 있기 때문에 본 연구자는 언더샘플링 기법을 사용하였다.

Random undersampling 기법으로 실질 허위오인에 해당하지 않는 일반신고 데이터에 대하여 모집단의 각 범주(긴급코드, 접수사건분류, 종결사건분류 등)별 비율을 반영하여 총 19,000건의 일반신고 데이터를 선별하였다. 각 클래스에 대하여, 허위오인신고는 class변수 값을 1로, 일반신고의 경우에는 0으로 지정하여 타겟 데이터를 구축하였다. 이로써 허위오인신고 분류모델 학습에 사용하기 위한 데이터는 112 신고접수 및 현장처리 데이터 총 37,735(18,735/19,000)건이다. 필터링된 112 긴급신고 37,735건 각각에 대한 '신고접수 내용' 텍스트를 모델 학습에 사용할 것이다. 허위오인신고 분류모델 학습에 사용할 데이터의 형태는 다음 그림 5와 같다.

모델 학습에 사용할 '신고접수 내용' 데이터는 112 접수 요원이 신고자와의 통화내용을 요약한 텍스트이다. 이를 모델에 입력하기 위해, Transformer 구조를 가진 각 언어모델에 해당하는 Tokenizer로 토큰화한다. 이 과정에서 512개의 토큰을 기준으로 문장을 잘라내고 padding하였다. 패딩의 기준이 되는 토큰수 512는 BERT Tokenizer의 최대 입력 크기이다.

악의적인 허위신고의 경우 신고접수 과정에서 112 상황실 요원이 어느정도 파악할 수 있다. 매뉴얼 상에도 허위신고가 의심된다면 접수요원이 허위신고시 처벌가능함을 고지하도록 되어있고, 실제로 고지하였음을 신고접수 내용 텍스트에 적시해놓기도 한다. 뿐만 아니라 대부분의 허위신고는 접수요원의 추가적 질문에 제대로 대답하지 못하는 등의 특이사항이 발생하는 바, 이러한 의심점과 특징들이 전부 신고접수 텍스트에 반영된다. 따라서 악의적인 허위신고의 경우 연구대상 데이터를 통해서 충분히 분류 및 예측이 가능하다.

하지만 신고자가 인지한 현장이 실제와 차이가 있는 경우는 그렇지 못하다. 이러한 오인신고의 상황에서 신고자는 자신이 인지한 범죄상황을 수사기관에 보고하게 된다. 술을 마신 연인이 스킨십하는 모습을 강제추행으로 보거나, 옆집에서 아이가 우는 것을 듣고 아동학대로 신고하는 것 등을 예시로 들 수 있다. 이러한 경우, 112 상황실 접수요원을 비롯하여 신고자조차도 해당사실을 진실로 받아들인다. 결국 신고내용 접수 텍스트만으로 분류모델이 오인신고를 완벽히 구분하기에는 한계가 있다.

본 연구에서는 보다 엄밀한 분석과 모델의 성능 확인을 위하여 새로운 데이터셋을 구축한다. 본 연구자는 출동한 현장경찰관이 신고처리 후 기록한 '현장처리 내용'을 근거로 18,735개의 허위오인신고들 중에서 악의적인 허위신고를 분류했다. 오인신고와 악의적 허위신고는 '현장처리 내용'에서 구별되기 때문이다. 악의적인 허위신고는 현장처리 내용에 다음과 같은 경우를 포함한다. 첫째로, 현장에서 허위신고로 신고자에 대하여 즉결심판 청구하거나 현행범체포한 경우를 포함하고 있다. 둘째로는 신고자가 도망치거나, 현장에 부재했던 경우를 포함한다. 마지막으로는 처벌하지 않고 경고로 넘어가거나, 처리중 수사과로 인계된 경우 등을 비롯한 기타분류이다. 연구자가 '현장처리 내용'을 바탕으로 판단하여 분류한 악의적인 허위신고 건에 대한 '신고접수 내용'을 악위적인 허위신고

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 37735 entries, 0 to 37734
Data columns (total 2 columns):

#	Column	Non-Null Count	Dtype
0	신고접수 내용	37,735 non-null	object
1	레이블	37,735 non-null	object

그림 5. 허위오인신고 분류모델 학습 데이터
Fig. 5. False report classification model training data structure

표 8. 악의적 허위신고(허위거짓신고) 예시
Table 8. Examples of malicious false report

No	신고접수 내용	현장처리 내용
1.	차를 세워 놓고 밥을 먹고 온 사이/차가 없어졌다고/차키뺐음	허위신고로 ○○○(**세) 즉결심판 청구함 즉결심판번호: 제 ***7-******5번
2.	누가 엉덩이와 가슴을 만지고 도주했다며/ 주취 목소리/ 발생일자 구체적인 내용에 답변하지 않음/	신고자 ○○○(**세)으로 자기 애인집이라고 하며 혼자 있으면서 횡설수설하여 강력계도 후 현장정리
3.	노래방에 비상문으로 들어가 달라며 / 도우미가 흰색옷을 입고 들어갔다며/	노래방에는 손님이 한명도 없는 것으로 현장종결.

분류모델의 학습데이터로 이용할 것이다.

전체 허위오인신고 데이터 18,735건 중에서 악의적인 허위신고에 해당하는 4,765건을 분류하였다. 이에 일반신고 데이터를 이전의 방법과 동일하게 4,800건 추출하여 악의적 허위신고 분류 모델 학습을 위한 총 9,565(4,765/4,800)건을 마련하였다. 다음 표 8은 허위오인신고 데이터중 악의적 허위신고로 분류된 데이터의 예시이다. 또한 악의적 허위신고 분류모델 학습에 사용할 데이터의 형태는 다음 그림 6과 같다.

두 데이터셋에 대한 전처리 흐름은 그림 7과 같다.

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9565 entries, 0 to 9564
Data columns (total 2 columns):

#	Column	Non-Null Count	Dtype
0	신고접수 내용	9,565 non-null	object
1	레이블	9,565 non-null	object

그림 6. 악의적 허위신고 분류모델 학습 데이터
Fig. 6. Malicious false report classification model training data structure

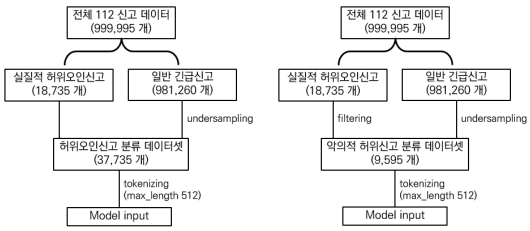

그림 7. 전처리 과정 흐름도
Fig. 7. Preprocessing process flow chart

4.4 하이퍼파라미터 지정

앞선 절에서 구축한 학습 데이터셋을 각각 허위오인신고(37,735건), 허위거짓신고(9,565건) 데이터셋으로 명명한다. 각각의 데이터셋으로 학습된 모델 또한 동일하다. Transformer 구조 기반의 LLM을 이용하여 실험이 진행되었다. BERT-base, KoBERT, ELECTRA, KoELECTRA, RoBERTa 모델이 대상이며, 여타 Hyperparameter는 다음과 같이 동일하게 학습을 수행했다. 두 종류의 데이터셋은 모두 학습과정에서 Train : Validation : Test 8 : 1 : 1로 분할되었다. 이때, 분할은 python scikit-learn 라이브러리의 train_test_split()함수를 이용하였다.

손실함수로는 이진분류 모델에 사용하는 Binary Cross Entropy를 채택했다. 학습률은 0.00001(1e-5), 최적화 함수는 AdamW이다. Batchsize의 경우 32로 고정하였다. Dropout은 p=0.1로 지정하였다. 실험결과의 재현을 위하여 randomstate를 임의로 지정하였다.

Tokenizer는 각 모델에 해당하는 사전학습된 Tokenizer를 불러와 이용하였다.

4.5 모델 구축

앞서 언급한 실험대상 5가지 모델에 대하여, 모델의 아키텍처를 설명하고자 한다. 본 연구에서 제시하는 모델의 구축환경은 Python 환경을 기반으로 한다. Python의 머신러닝 라이브러리인 Pytorch를 이용하였다. 또한 사전학습된 LLM을 불러오기 위하여 Hugging Face를 이용하였다.

Transformer 구조를 가지는 LLM들을 Hugging Face를 이용하여 불러온 후, Classification Task에 맞게 모델의 출력부에 분류기 Layer를 추가하였다. 모델의 핵심구조인 Transformer 구조와 임베딩층에 대해서는 변경을 가하지 않았다. 모델별로 최종 Layer로 추가된 분류기 구조는 다음 그림 7과 같다. 본 연구에서 제시하는 모델은 이진분류 Task를 수행하므로 분류기의 최종 Output에 대하여 out_features를 2로 지정하였다. KoBERT는 그림 8의 BERT 구조와 같으며, KoELECTRA 또한 ELECTRA 구조와 동일하다.

```
# ELECTRA-Base
(classifier): ElectraClassificationHead(
    (dense): Linear(in_features=256, out_features=256, bias=True)
    (dropout): Dropout(p=0.1, inplace=False)
    (out_proj): Linear(in_features=256, out_features=2, bias=True)
)

# RoBERTa-Base
(classifier): RobertaClassificationHead(
    (dense): Linear(in_features=256, out_features=256, bias=True)
    (dropout): Dropout(p=0.1, inplace=False)
    (out_proj): Linear(in_features=256, out_features=2, bias=True)
)

# BERT-Base
(dropout): Dropout(p=0.1, inplace=False)
(classifier): Linear(in_features=768, out_features=2, bias=True)
```

그림 8. 각 모델별 분류기 Layer 구조
Fig. 8. Classifier layer structure for each model

4.6 모델 학습 및 결과 분석

표 9는 허위오인신고 task에 대한 모델별 성능이다. 실험 결과, 상기한 오인신고의 한계점이 드러났다. 모델의 종류와 에포크 수에 무관하게 Binary classification Task임에도 정확도와 F1 Score가 상당히 낮았다. 그럼에도 한국어 특화 모델과 베이스 모델 사이에는 성능의 차이가 어느정도 존재했다. 다섯가지 모델 중에서 가장 나은 성능을 보인 모델은 KoELECTRA이었다.

표 9. 허위오인신고 분류 모델 실험 결과
Table 9. False misreport classification model experiment results

Model	Metric	Epoch		
		30	50	100
BERT-base	Acc.	0.51	0.53	0.54
	F1	0.50	0.50	0.51
KoBERT	Acc.	0.53	0.54	0.56
	F1	0.53	0.52	0.52
ELECTRA	Acc.	0.53	0.55	0.58
	F1	0.52	0.55	0.58
KoELECTRA	Acc.	0.55	0.57	0.63
	F1	0.54	0.57	0.62
RoBERTa	Acc.	0.50	0.52	0.53
	F1	0.50	0.51	0.53

표 9의 결과는 허위오인신고 중에서도 특히 오인신고가 모델 학습에 영향을 준 것으로 추정된다. 오인신고의 경우 신고내용 접수 텍스트가 일반신고와 차이가 적기 때문이다. 데이터 분석의 과정에서 살펴본 차이는 모델의 오인신고 패턴 학습에 큰 영향을 주지 못했다. 사건분류에 대한 고려 없이 신고접수내용 텍스트만을 input으로 받았기 때문에 오인신고의 한계가 더욱 강조되었다고 볼 수 있다.

오인신고에 대하여 자연어처리를 통한 분류가 어려울 것이라는 가설은 표 9와 위 표 10과의 비교로도 확인된다.

표 9의 허위오인 분류 모델과 다르게 전체적으로 평가 metric이 높게 나타났다. 이는 허위거짓신고 분류 task가 허위오인신고 분류보다 수월함을 의미한다. 표 10에서도 베이스 모델보다는 한국어 특화 모델이 좋은 성능을 보였다. 모델간의 비교에서도 또한 표9와 같이 BERT 계열보다는 ELECTRA 계열의 모델이 더 높은 metric을 가졌다. 표 10에 따르면 KoELECTRA 모델의 100에포크에서 보이는 F1 Score가 0.84로 나타난다. 그러나 Epoch 100의 학습과정에서도 여전히 Validation loss가 감소하는 추세를 보였다. 이는 추가적인 학습이 해당 모델의 성능을 더욱 향상시킬 수 있음을 시사한다.

표 10. 허위거짓신고 분류 모델 실험 결과
Table 10. Malicious false report classification model experiment results

Model	Metric	Epoch		
		30	50	100
BERT-base	Acc.	0.71	0.72	0.74
	F1	0.72	0.73	0.75
KoBERT	Acc.	0.75	0.76	0.78
	F1	0.74	0.74	0.79
ELECTRA	Acc.	0.71	0.74	0.76
	F1	0.70	0.75	0.75
KoELECTRA	Acc.	0.78	0.80	0.84
	F1	0.79	0.80	0.84
RoBERTa	Acc.	0.71	0.72	0.72
	F1	0.70	0.71	0.71

실험환경에서 가장 준수한 성능을 보인 허위거짓신고 데이터셋에 대한 KoELECTRA 모델의 Test 결과(Epoch 100)에 따른 Confusion matrix는 그림 9와 같다.

본 모델을 학습하는 과정에서 허위거짓신고인 경우 1, 보통신고인 경우를 0으로 라벨링하였다. 따라서 거짓신고인 False가 Positive value이고, 보통신고인 usual이 Negative value에 해당한다. TP = 434, FN = 64, FP = 69, TN = 377이다. Positive에 대하여 Precision은 434/503 = 0.86이고, Recall은 434/498 = 0.87이다.

다음 그림 10은 허위오인신고 데이터셋에 대한 KoELECTRA 모델(Epoch 100)의 confusion matrix이다. 그림 9의 혼동행렬과 비교한다면 오인신고 데이터가 모델의 학습에 있어 어떤 영향을 미치는지 파악할 수 있다.

그림 10은 그림 9의 혼동행렬에 비하여 False Positive의 비율이 유의하게 높은 것을 확인할 수 있다. TP = 1275, FN = 565, FP = 768, TN = 1023이다. 그림 9와 10의 모델간 차이점은 '학습데이터에 오인신고가 포함되었는지 여부'이다. 전체적으로 Accuracy와 F1 Score가 낮은 점도 주목할만 하지만, FP가 유의하게 높은 점을 짚고 넘어가야 한다. 그림 9의 Precision은 1275/2043 = 0.62이며, Recall은

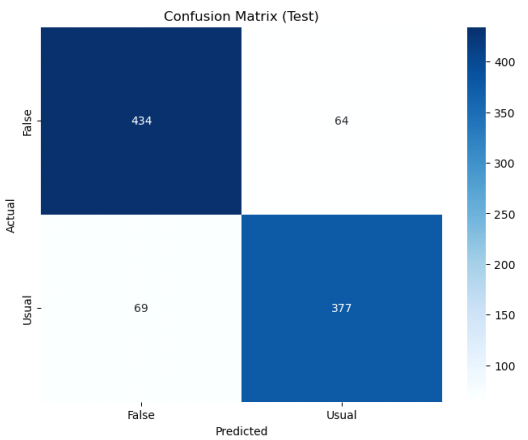

그림 9. 허위거짓신고 KoELECTRA모델의 혼동행렬
Fig. 9. Confusion matrix of malicious false report KoELECTRA model

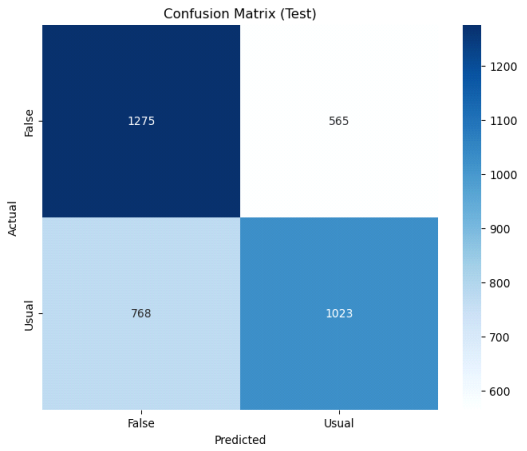

그림 10. 허위오인신고 KoELECTRA모델의 혼동행렬
Fig. 10. Confusion matrix of KoELECTRA model for false misreporting

1275/1840 = 0.69이다. 실제로 허위오인신고인 데이터 중에서 허위오인으로 예측하는 경우는 나름의 정확도를 보인다. 허나, 모델이 허위오인으로 예측한 데이터중에서는 실제로는 허위오인에 해당하지 않은 데이터가 더욱 많다. 앞선 데이터 분석 과정에서 드러난 오인신고의 특징이 영향을 미친 것이다.

오인신고는 대체로 직접 범죄현장을 마주하기보다는 간접적으로 인지하고 신고하는 경우가 많다. 이에 따라 신고접수 텍스트에서도 '소리', '~같다'등의 표현이 두드러진다. 그림 10의 혼동행렬을 보면, 모델이 이러한 패턴이 드러난 신고들을 대부분 허위오인신고로 분류한 것으로 보인다. 즉, '간접적으로 범죄상황을 인지하고 신고하였으나, 이것이 실제로 범죄였던 경우'를 허위오인으로 분류한 것이다.

실제로 직접 보고 신고하는 주취자 및 교통사고 신고와는 다르게 간접적으로 파악한 아동학대 등 신고는 그 불확실성이 높다. 직접 목격하기보다도 소리나 정황을 통해서 간접적으로 신고하는 경우가 상당히 많다. 이들 신고 모두를 '해당 패턴이 나타난다는' 이유만으로 허위오인신고로 분류한다면 큰 문제가 발생한다. 진정으로 출동이 필요한 상황에 경찰이 출동하지 못하기 때문이다. 이러한 점에서 본 모델은 정확도뿐만 아니라 낮은 FP가 무엇보다도 중요하다.

본 연구의 실험환경에서 구축한 두 가지의 데이터셋에 따라 분류 모델의 성능 지표인 Accuracy와 F1 score를 확인했다. 두 데이터셋 사이에 상당한 성능지표 차이가 있었다. 그 이유는 전술한 바와 같이 오인신고의 특성때문으로 보인다. 표 9와 표 10에 따르면, 같은 데이터셋을 동일한 하이퍼파라미터 조건 하에서 학습하였음에도 성능에 상당한 차이를 보였다. 같은 조건임에도 모델별로 상이한 성능을 보인 점에 대해서 짚고 넘어가고자 한다.

표 10과 같이 허위거짓신고 데이터셋에 대하여는 모델별 성능이 KoELECTRA > KoBERT > ELECTRA > BERT > RoBERTa순으로 나타난다. 허위오인신고 데이터셋에 대한 학습결과인 표9에 따르면 모델별 성능이 KoELECTRA > ELECTRA > KoBERT > RoBERTa > BERT순으로 나타난다. 위 순서에서 주의깊게 바라볼 부분은 크게 두가지로 나뉜다.

첫 번째는 BERT와 RoBERTa를 기반으로 하는 모델들에 비하여 ELECTRA 기반의 모델이 높은 성능을 보였다는 점이다. BERT와 RoBERTa는 4.2절에서 서술한 바와 같이 마스크 언어 모델링(MLM) 방식으로 학습하게 된다. 이는 전체 토큰의 15%만을 학습에 활용한다는 특징이 있다. 이에 반해 ELECTRA는 전체 입력 토큰을 학습에 활용할 수 있는 replaced token detection을 활용한다. 이 점에서 ELECTRA는 본 연구에 활용한 112 긴급신고 데이터셋과 같이 비교적 적은 데이터로도 BERT에 버금가는 성능을 보일 수 있었다고 사료된다. 특히나 허위거짓신고(9,565 개) 데이터셋의 경우 ELECTRA의 이러한 강점이 주요했을 것으로 보인다.

두 번째는 대부분의 조건에서 BERT보다 높은 성능을 보이는 RoBERTa가 본 연구의 실험조건에서는 BERT와 비슷하거나 낮은 성능을 보였다는 점이다. 이에 대해서는 학습에 이용한 데이터 크기가 영향을 미친 것으로 분석된다. 본 연구의 실험환경에서 이용한 두 데이터셋중 상대적으로 많은 허위오인신고 데이터셋(37,735 개)에서는 BERT와 RoBERTa가 비슷한 성능을 보였다. 그러나 비교적 적은 허위거짓신고 데이터셋(9,565 개)에 대하여는 오히려 RoBERTa가 BERT보다 낮은 성능을 보였다.

RoBERTa는 BERT에 비해 상당히 복잡한 구조를 가지고 있다. 또한 학습에도 더 많은 자원을 사용하는 등 효과적이나 동시에 효율적이지는 못한 특성을 가지고 있다. 더 많은 파라미터를 가진 RoBERTa가 상대적으로 여타 모델들에 비하여 오버피팅되어 일반화의 측면에서 낮은 성능을 보인 것으로 추측된다.

실제로 RoBERTa의 학습과정에서 각 Epoch별로 확인한 Validation set의 Accuracy와 F1 score는 여타 모델들과 크게 차이나지 않았다. 그럼에도 Test set에 대하여 평가 metric이 낮게 나타났다. 적은 데이터셋

에 대한 실험환경이 RoBERTa의 장점을 펼치기에는 적절하지 않았던 것으로 보인다.

4.7 소결 및 제언

그동안 경찰의 치안업무와 관련된 domain에 있어서 AI를 접목하기 위한 시도는 꾸준히 이루어져 왔다. 그러나 대체로 실질적인 권력관계에서의 의사결정보다는 단순한 업무지원의 목적으로 실험된다는 한계가 있었다. 이는 경찰의 직무특성과도 연결되어 있다. 경찰은 국민의 신체 생명 재산을 보호한다. 그러나 동시에 경찰작용의 대상자가 되는 국민-경찰책임자-에게는 다소간 침익적인 권력작용이 주로 이루어진다. 경찰의 잘못된 직무수행은 국민의 기본권을 심각하게 제한할 수 있다는 점이 문제된다.

본 장에서는 허위오인신고와 허위거짓신고 분류모델에 대한 모델학습이 이루어졌다. 악의적인 허위신고의 경우는 신고접수내용 텍스트만으로 분류작업에서 상당히 높은 성능을 보였다. 추가적인 학습을 통하여 경찰의 업무영역에서도 충분히 '의사결정자' AI모델을 운용할 수 있을 것으로 보인다.

그러나 상술한 바와 같이 AI가 제 기능을 다하더라도 실질적으로 경찰력의 낭비를 막기에는 어렵다. 경찰의 존재이유 및 직무와 연관지어 생각해보아야만 한다. AI의 임의적 결정으로 긴급신고의 출동여부를 결정하는 것은 다소간 도발적인 제안이다. 이를 위해서는 치안서비스의 수요자인 국민의 공감대 형성이 선행되어야 한다. 허나 현행체제에서도 위와같은 허위거짓신고 분류 모델은 다음과 같은 영역에서 충분히 역할을 다할 수 있다.

위 모델을 통하여 허위신고를 분류하였다면 해당 신고자에 대한 즉결심판 혹은 현행범체포 등 후속조치 프로세스를 앞당길 수 있다. 더불어 AI가 악의적인 허위신고로 판단했다는 점까지 현장출동 경찰관에게 전달된다면 이를 통해 현장적 대응과 의사결정에도 상당한 도움을 줄 수 있다.

V. 결 론

관련 통계가 작성된 이래로 112 허위오인신고는 꾸준히 감소했다. 그러나, 지난 몇 년 사이 다시금 이러한 신고가 급격히 증가하는 추세이다. 경찰의 긴급신고시스템을 대상으로 한 허위신고 연구는 여러 방면으로 진행되고 있다. 그러나 허위신고 문제를 해결하기 위한 정책적 제언과 시스템 개선방안등을 제시하였을 뿐, 자연어처리 기술과 LLM 학습을 통한 분류/예측모델에 대한 연구는 미진한 실정이다. 112 신고접수 데이터에 대하여 텍스트마이닝 등 군집분석을 수행한 사례는 존재하였으나 이 또한 특정 문제에 대한 해결적 접근보다는 탐색적 분석에 가까웠다.

이에 따라 본 연구자는 112 신고접수 데이터를 이용하여 해당 신고가 허위오인신고인지를 분류하는 텍스트 이진분류 모델과 시스템을 제안한다. 연구에서 제시한 모델은 112상황실 접수요원이 신고내용을 요약한 신고접수 내용 텍스트를 입력받는다. 텍스트를 보고 해당 신고가 허위오인신고인지 일반적인 신고인지를 구분하는 이진분류 task를 수행한다.

그러나, 오인신고의 경우 표면적으로는 일반적인 신고와 크게 구분되지 않는다. 허위오인신고 전체에 대한 데이터와 악의적 허위신고만을 담은 데이터셋을 따로 나누어 모델 훈련이 진행되었다. 실험을 통해 transformer 구조 기반의 각 LLM별로 어느정도의 성능을 보이는지 파악하였다. 이어서 데이터 구성에 따른 성능차이를 파악하여 오인신고의 특성이 AI의 학습에 어떤 영향을 미치는지를 규명하였다.

실험 결과 base 언어모델보다는 한국어에 특화된 모델이 나은 성능을 보였다. 모델별로 살펴보자면 ELECTRA 기반의 모델이 BERT 기반의 모델보다 높은 평가 metric을 가졌다. 데이터의 구성으로는 악의적인 허위신고만을 학습시킨 허위거짓신고 모델이 높은 정확도를 보였다. 다양한 모델 중에서 가장 높은 성능을 보인 모델은 허위거짓신고 분류 task의 KoELECTRA모델이다. 모델의 정확도는 0.84, F1 Score 0.84로 나타났다.

본 연구는 112 신고데이터를 바탕으로 하여 허위오인신고와 관련된 EDA를 통해 허위오인신고의 살폈다. 더불어 허위오인신고 분류를 위한 실질적인 LLM 학습과 metric 측정까지 나아갔다는 점에서 기존의 연구와 차별성을 보인다. 경찰의 치안 domain에서 자연어처리와 LLM을 이용하여 특정 문제에 대한 해결적 접근을 처음으로 시도해보았다는 것에 의의가 있다.

그러나 본 연구에서 제안하는 시스템은 한계가 존재한다. 첫째로 긴급코드, 접수사건분류등 모델학습에 참고할 수 있는 여타 변수를 배제하고 텍스트만을 학습하였다. 둘째로는 각 모델별로 최대 100에포크로 제한된 학습을 거쳤다. 추가적인 학습을 한다면 성능이 더욱 높아질 여지가 있다. 마지막으로는 허위오인과 일반신고 클래스의 불균형 문제이다. 이를 해결하기 위해 본 연구에서는 언더샘플링을 이용하였으나, SMOTE 등 기법을 이용한다면 데이터를 보다 효율적으로 사용할 수 있다.

경찰의 직무특성에 따르면, AI가 판단하는 대로 출동여부를 결정짓는 것에는 상당한 무리가 있다. 그러나 악의적인 허위신고 예측에 대하여는 해당 신고자에 대한 신속한 조치가 가능하고, 현장 경찰관들의 의사결정에 도움이 될 수 있다. 이로써 본 연구에서 제안하는 시스템이 현재로서는 실제적 치안사무에 '지원자'로서의 역할을 충분히 수행할 수 있을 것으로 기대한다.

References

[1] E. A. Blackstone, A. J. Buck, and S. Hakim, "Evaluation of alternative policies to combat false emergency calls," *Evaluation and Program Planning*, vol. 28, p. 234, 2005. (https://doi.org/10.1016/j.evalprogplan.2004.09.004)

[2] Korean National Police Agency, Police improve 112 reporting dispatch system*(2016),* Retrieved Sep., 24, 2023, from https://www.korea.kr/briefing/pressReleaseView.do?newsId=156119657

[3] S. H. Roh and J. T. Cho, "An empirical study on the problems and improvement of 112 emergency call system in Korea," *The Police Sci. J.*, vol. 11, no. 4, pp. 9-38, 2016. (https://doi.org/10.16961/polips.2016.11.4.9)

[4] J. E. Lee and H. Choi, "A study on the 112 emergency call code system reflecting the public's perspective," *The J. Police Policies*, vol. 37, no. 2, pp. 5-38, 2023. (https://doi.org/10.35147/knpsi.2023.37.2.005)

[5] J. Y. Hwang, "A study on the real condition and eradication plan of the false report through 112 call service," *The J. Police Policies*, vol. 35, no. 1, pp. 7-37, 2021. (https://doi.org/10.35147/knpsi.2021.35.1.007)

[6] D. S. Woo and J. H. Byeun, "A study on the identification of the false report through 112 call service : Through the analysis of voice difference between the false and the actual report," *The J. Police Sci.*, vol. 15, no. 4, pp. 117-140, 2015. (http://doi.org/10.22816/polsci.2015.15.4.005)

[7] J. Y. Jung, "Big data analysis on 112 report data : Focusing on the eda technique," *Korean Secur. J.*, vol. 66, pp. 71-92, 2021. (http://doi.org/10.36623/KSSR.2021.66.4)

[8] S. E. Hong, Y. J. Kim, K. H. Jang, and J. S. Bang, "Exploring the similarity between incident types based on text mining and cluster analysis through vectorizing 112 report contents," *The J. Police Sci.*, vol. 20, no. 3, pp. 63-86, 2020. (http://doi.org/10.22816/polsci.2020.20.3.003)

[9] M. Laforest, "The false report during an emergency call: Using discourse analysis to detect deceit," *Int. Assoc. Forensic Ling. Tenth Biennial Conf.,* pp. 139-152, Birmingham, UK, Jul. 2012.

[10] M. D. Firoozjaei, J. W. Park, and H. S. Kim, "Detecting false emergency requests using callers' reporting behaviors and locations," *Int. Conf. Advanced Inf. Netw. and Appl. Wkshps.*, pp. 139-152, Crans-Montana, Switzerland, 2016.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in NIPS*, pp. 5998-6008, 2017. (https://doi.org/10.48550/arXiv.1706.03762)

[12] 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. (https://doi.org/10.48550/arXiv.1810.04805)

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019. (https://doi.org/10.48550/arXiv.1907.11692)

[14] K. Clark, M. T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," *arXiv preprint arXiv:2003.10555*, 2020. (https://doi.org/10.48550/arXiv.2003.10555)

정 재 훈 (Jae-hoon Jeong)

2021년 3월~현재 : 경찰대학 행정학과

<관심분야> 딥러닝, 인공지능, 자연어처리

[ORCID:0009-0008-2768-3100]

박 현 호 (Hyunho Park)

2005년 8월 : 경북대학교 전자전기컴퓨터학부 학사

2007년 8월 : 광주과학기술원 정보기전공 학부 석사

2014년 8월 : 과학기술연합대학원대학교 광대역네트워크공학 박사

2014년 9월~현재 : 한국전자통신연구원 선임연구원

<관심분야> 딥러닝 기반 공공안전 기술개발, 자연어 처리

[ORCID:0000-0002-9943-1497]