정책 특집

AI 챗봇 '이루다' 논란의 이슈 변화와 시사점

Identifying Issue Changes of AI Chatbot 'Iruda' Case and Its Implications

최새솔 (S.S. Choi, saesol.choi@etri.re.kr) 지능화정책연구실 선임연구원
홍아름 (A.R. Hong, areumh@etri.re.kr) 지능화정책연구실 연구원

ABSTRACT

The controversy over Artificial Intelligence (AI) chatbot "Iruda," which suspended its service 20 days after its launch, can be seen as the first case to inform the public of AI ethics issues. Based on this context, this study examines the controversy and social semantic formation of "Iruda" service cases using news topic modeling techniques. 963-news articles were used for the analysis, and the event's duration was analyzed based on major events, such as service start, controversy, and suspension, to understand the progress. From the analyses results, we obtain major keywords and a total of 16 topics (5, 4, 7) from the period. Finally, the implications for the development and utilization of AI services obtained through this controversy were discussed based on the analysis results.

KEYWORDS AI 챗봇, 이루다, AI 윤리, 토픽 모델링, LDA

Ⅰ. 서론

인공지능(이하 AI)이 빠르게 발전함에 따라 다양한 서비스가 출시되고 있다. 챗봇(Chatbot)은 음성이나 문자를 통해 인간과 대화할 수 있는 컴퓨터 시스템으로, AI가 가장 빠르게 도입되고 있는 응용 분야 중 하나이다. 시장 조사 기관에 따르면, 전 세계 챗봇 시장 규모는 2019년 25.7억 달러에서 2024년 94.3억 달러로 연평균 약 30%의 빠른 성장이 전망된다[1].

초기 챗봇 서비스가 규칙(룰) 기반 처리 방식으로 업무 자동화에 주로 적용되었던 것과 달리, 최근의 챗봇 서비스는 자연어처리(NLP) 기반의 끊어짐 없는 대화(Multi-turn)와 특정 임무 수행이 아닌 일상대화형 서비스로 진화하는 것이 특징이다. 본고에서 분석 대상으로 삼은 스캐터랩이 출시한 AI 챗봇 '이루다' 서비스 역시, 스무 살 여대생으로 의인화한 챗봇과의 일상대화를 통해 감정적 충족감을 제공하는 것을 서비스 목표로 했다는 점에서 기존의 챗봇과 결을 달리한다.

사실, AI 서비스가 보일 수 있는 편향성과 불완

* DOI: https://doi.org/10.22648/ETRI.2021.J.360210
* 이 논문은 한국전자통신연구원 연구운영지원사업의 일환으로 수행되었음[21ZR1400, 국가지능화 기술정책 및 표준화 연구].

전성은 '이루다' 서비스 이전에도 많은 지적을 받아왔다. '이루다' 서비스와 유사한 챗봇 서비스에서도 2016년 MS가 공개한 테이(Tay)는 인종차별, 혐오 발언으로 출시 16시간 만에 서비스가 전면 중단되었고[2], 2020년 MS, 아마존, IBM 등 빅테크 기업들은 편향성 문제가 불거지자 자사의 안면 인식 서비스를 보류 또는 개발 중단을 선언한 바 있다[3].

'이루다'의 개발자 또한 출시 전, 어느 정도 논란을 예상했다고는 하나[4], 성적(性的) 대상화, 동성애, 장애인 등에 대한 차별 및 혐오 발언, 개발과정에서의 「개인정보보호법」 위반혐의 등 논란은 예상보다 컸고, 결국 출시 20여 일 만에 서비스를 중단하기에 이른다.

AI 챗봇 '이루다' 논란은 막연하게 여겨지던 AI 윤리와 문제점을 우리 사회에 인식시킨 최초 사례로 볼 수 있다. 이에 본 고는 다양한 논란을 일으키고 서비스를 폐쇄한 AI 챗봇 '이루다'가 남긴 쟁점과 이슈 변화에 대해 뉴스 토픽 모델링 기법을 통해 알아보고, 그 함의를 논의한다.

Ⅱ. 이루다 서비스 개요

AI 챗봇 '이루다'는 국내 스타트업 스캐터랩이 개발한 페이스북 메신저 기반 일상대화 챗봇(Open-domain Conversational AI)이다.

'이루다'는 스무 살 여성 대학생을 페르소나로 하고 있으며, 딥러닝 알고리즘을 이용하여 친근하고 자연스러운 일상대화를 구현한 것이 특징이다.

개발사인 스캐터랩은 2011년 설립된 감정분석 전문 스타트업으로, 연인과 주고받은 메시지를 통해 연애 감정을 파악하는 '연애의 과학', 메신저 기반 감정분석 서비스인 '텍스트앳' 등 대화형 AI에 대한 사업모델을 영위하며 이와 관련한 많은 기술과 노하우를 확보한 것으로 알려져 있다.[1] 스캐터랩은 '이루다'의 개발과정에서 자사가 운영 중인 '텍스트앳'과 '연애의 과학' 등에서 확보한 연인 간 대화 데이터 100억 건을 기초로 '이루다' 학습 모델을 구축하여 자연스럽고 실감 나는 대화 기능을 구현하였다[5].

'이루다'는 2020년 6월 15일에 베타 테스트를 진행하여 1,500명의 베타 테스터와 대화를 나누었으며, 12월 23일부터 정식 서비스를 시작하였다[6]. '이루다'는 초기 자연스럽게 대화를 이어갈 수 있고 사용자의 말투를 따라 하는 등 기존의 임무 수행 중심의 챗봇과는 다른 생동감 있는 대화 서비스로, 출시 3주 만에 약 80만 명의 이용자를 확보하였다[7].

그러나 2020년 12월 30일 온라인 커뮤니티 '아카라이브'에 '이루다'를 성적(性的) 대상화하는 게시물이 올라오는 등 서비스 내 성희롱 및 챗봇의 혐오·차별 발언에 대한 여러 가지 논란이 제기되었다. 논란이 가속화되자 스캐터랩은 챗봇에 대한 성희롱을 예상했고 사전 조처를 했으나, 결과적으로 모든 부적절한 대화를 막지는 못했음을 인정하며, 향후 서비스를 개선하겠다는 공식입장을 2021년 1월 8일에 발표하였다[4]. 하지만 「개인정보보호법」 위반 등으로 논란이 더욱 확대되면서 결국 11일 서비스 중단을 선언하고, 12일 서비스를 잠정 중단하였다[8]. 또한, 1월 15일 스캐터랩은 이루다 서비스 DB 및 딥러닝 대화 모델의 폐기를 선언하였다[7].

Ⅲ. 분석 방법

본 연구는 AI 챗봇 '이루다' 서비스를 둘러싼 이슈를 추적하기 위해서 포털사이트(네이버)의 뉴스 기사를 분석의 대상으로 택하였다. 뉴스 기사는 우리 사회에서 발생하는 사건과 이슈를 담고 있으며

1) http://www.demoday.co.kr/company/스캐터랩 참고

대중들에게 정보를 제공하는 주요 매체로, 특정 주제 또는 분야에 대한 전반적인 동향을 파악하는 데 유용한 자료이다[9]. 따라서 뉴스 기사 분석은 AI 챗봇 '이루다' 서비스의 논란과 이에 따른 사회적 의제 형성과정을 추적하고자 하는 본 연구 목적에 부합하며, 본 연구는 국내 대표적 포털사이트인 네이버의 뉴스 기사를 수집하여 키워드 분석 및 토픽 모델링 기법을 통해 '이루다' 서비스의 논란과 이에 따른 함의를 분석한다.

1. 데이터 수집

분석을 위한 데이터는 2020.06.15.~2021.02.09.(자료수집일 기준) 기간에 네이버 뉴스로 검색(검색키워드: 이루다+챗봇)된 기사 1,553건을 자체 작성한 파이썬 크롤러를 통해 수집하였다. 분석의 용이성을 위해 수집된 기사 중 네이버 뉴스섹션에서 자체 제공하는 기사 963건 만으로 대상을 한정하였다.[2)]

2. 데이터 전처리

데이터 전처리는 다음과 같이 진행하였다. 첫째, 본문 내용이 없는 사진기사나 본문이 짧은 기사(500자 미만), 중복 수집된 기사, 여러 단신에 함께 소개되는 기사[3)]는 정확한 분석에 도움이 되지 못하므로 제외하였다. 이를 통해 963건에서 최종 849건(중복 및 단신 포함 기사 78건, 본문 길이 500자 미만 36건 제외)의 기사만이 분석에 활용되었다. 둘째, 분석에 불필요한 영어 및 특수문자는 모두 제거하였다. 셋째, 빈번히 등장하나 핵심의미 구성에는 불필요한 단어들은 한글 불용어 사전[4)]을 통해 우선 제거하고, 토픽 모델링을 반복하면서 상위 빈도 단어 중 주제 연관성이 떨어지는 단어를 추가 제외하는 방식을 택하였다. 끝으로, 파이썬 KoNLPy 라이브러리 중 mecab 토크나이저(Tokenizer)를 사용하여 1음절 이상의 한글 명사만을 추출하여 분석에 사용하였다.

3. 토픽 모델링

토픽 모델링은 구조화되지 않은 비정형 문서에서 잠재적 정보를 찾는 기법으로, 문서의 의미론적 구조를 탐색하는 데 효과적인 방법의 하나이다[10]. 특히, 단순 키워드 분석과 달리 키워드들의 군집으로 형성되는 문서의 주제를 도출할 수 있다는 장점이 있다[11].

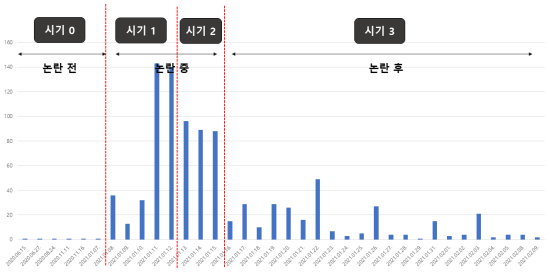
본 분석에서는 토픽 모델링의 대표적인 모델인 LDA(Latent Dirichlet Allocation) 기법을 사용하였다. LDA는 디리클레 분포를 기반으로 문서 내 단어들이 어떤 주제에 포함되는지에 대한 확률을 계산하는 추론방법[10]으로, 데이터 차원축소가 우수하며, 해석 가능한 일관된 주제를 추출하는 데 장점이 있어 텍스트 마이닝에서 널리 사용된다[11].

본 연구는 LDA 분석을 위해서 파이썬 패키지인 겐심(Gensim)을 이용하였는데, 겐심은 다양한 토픽 모델링을 지원하며 대규모 데이터 처리가 가능하다는 장점이 있다[11]. 겐심의 설정값(Hyperpa-

2) 개별 언론사별 뉴스 페이지(HTML) 구성이 모두 다른 관계로, 데이터 분석의 용이성을 위해 네이버가 자체 뉴스 페이지로 편성한 기사(URL이 'https://news.naver.com'로 시작하는 기사)만을 분석에 활용. 전체 기사 1,553건 중 네이버 뉴스 기사에 포함된 건은 963건으로 네이버 기사 비중은 약 62%(963/1553)임

3) 가령, '오늘의 헤드라인 뉴스'와 같이 본문에서 여러 뉴스 내용 중 하나로 소개되는 경우, 관련 없는 내용이 분석에 포함됨으로 해당 기사는 분석에서 제외

4) https://www.ranks.nl/stopwords/korean의 한국어 불용어 목록을 활용



출처 데이터 분석 후 직접 작성

그림 1 일자별 수집된 뉴스 기사 분포

rameter)인 알파(Alpha)와 에타(Eta)값은 기본값으로 설정하였다[12].

토픽의 개수를 결정하는 방법으로 본 연구에서는 토픽 수를 정하는 데 널리 쓰이는 측정방식인 일관성 점수(Coherence score)를 사용하였다. 토픽 개수(2~15개)를 변화하면서 도출한 일관성 점수의 그래프를 구한 다음 최대점의 토픽의 개수로 선정하는 방법이다.

LDA 토픽 모델링 방법은 확률기반 모델로 수행할 때마다 분류결과와 일관성 점수가 달라지므로, 반복 측정을 통해 최적 모델을 탐색하였다.

Ⅳ. 분석 결과

본 연구는 논란의 진행과 이슈 변화 과정을 살피기 위해 서비스의 시작과 논란, 중단 등 주요 이벤트를 기준으로 4개(논란 이전 1개, 논란 이후 3개)의 시기로 구분지어 분석을 진행한다. 이는 시간의 흐름 순으로 분석하여 진행 과정을 더욱 명확히 파악하기 위한 목적이다. 그림 1은 시기 구분과 시기별 네이버 뉴스 분포를 보여주고 있다.

1. 키워드 분석 결과

시기 0은 논란 이전의 시기이다. '이루다' 관련

표 1 시기별 워드 클라우드 및 주요 키워드

시기	워드 클라우드
시기 0: 논란 전 ~ 21.01.07. (6건)	 상위 10개 단어: 가상, 친구, 기술, 인간, 일상, 관계, 메신저, 생각, 출시, 페이스북
시기 1: 논란증폭 · 중단 21.01.08. ~ 21.01.12. (340건)	 상위 10개 단어: 혐오, 차별, 문제, 논란, 연애, 출시, 윤리, 중단, 발언, 성희롱
시기 2: 중단 직후 21.01.13. ~ 21.01.15. (256건)	 상위 10개 단어: 연애, 문제, 논란, 동의, 수집, 이용, 발언, 사회, 윤리, 조사
시기 3: 논란 이후 21.01.16. ~ 21.02.09. (247건)	 상위 10개 단어: 윤리, 기술, 보호, 사람, 기업, 수집, 동의, 유출, 인간, 차별

출처 데이터 분석 후 직접 작성

첫 뉴스('20.06.15. 베타 테스트 시작)가 보도된 시점부터 논란이 본격화되기 직전인 2021년 1월 7일까지로 이 시기는 수집된 뉴스가 6건 밖에 존재하지 않는다. 논란 전까지는 정식 서비스 출시 관련 기사도 없을 만큼, '이루다' 서비스는 큰 주목을 받지 못했다. 이 시기의 상위 키워드로는 가상, 친구, 일상, 관계, 메신저 등 부정적 단어 없이 서비스 특징을 소개하는 용어로 구성되어 있다.

시기 1은 성희롱 논란이 본격화되고, 소수자 차별 및 혐오 발언에 대한 논란, 개인정보 유출 및 서비스 잠정중단 결정에 이르기까지의 기간('21.01.08~'21.01.12.)으로 340건의 가장 많은 뉴스가 보도된 시기이다. 이 시기는 상위 키워드는 혐오, 차별, 문제 등으로 논란이 증폭되는 시기임을 보여준다.

시기 2('21.01.13.~'21.01.15.)는 서비스 중단 직후의 반응과 논란의 확장 양상을 확인할 수 있는 시기로, 이 시기의 상위 키워드로는 연애, 동의, 수집, 이용, 발언 등이 있다.

시기 3은 '21.01.16. 이후로 자료수집일인 '21.02.09.까지의 기간으로 논란이 잦아든 이후의 다양한 의제가 생산된 시기로 판단된다. 이 시기의 상위 키워드로는 윤리, 기술, 보호 등이 있다.

표 1은 시기별 워드 클라우드와 주요 키워드를 보여주고 있다.

2. 토픽 모델링 결과

논란 이후 3개의 시기에 대해서 각각 토픽 모델링을 진행하였다. 시기 0의 경우, 기사 건수가 너무 적고, 연구 목적과 연관성이 적어 분석에서 제외하였다.[5] 반복적인 수행을 통해 최종적으로 시기 1은 5개, 시기 2는 4개, 시기 3은 7개의 총 16개의 토픽을 도출하였다.[6]

본 연구는 기사 하나당 하나의 토픽을 가지는 것으로 가정하여 각 기사를 가장 높은 확률을 가지는 대표토픽 1개와 대응시켰다. 즉, 전체 기사 849건의 뉴스 기사는 각각 16개의 주제 중에 하나로 대응되었다.

다음으로 토픽별 고빈도 출현 키워드와 출현확률이 높은 주요 뉴스 제목 및 뉴스 본문 내용을 바탕으로 토픽명을 부여하였다. 토픽별 토픽명과 주요 키워드, 주요 뉴스 제목 등 분석 결과는 표 2에 제시되어 있다.

가. 시기 1: 논란 증폭 및 중단

시기 1에서는 5개의 토픽이 도출되었다. [토픽1]은 일부 이용자들이 챗봇을 성적(性的) 대상화했다는 논란에 대한 것으로 '성희롱 논란'으로 명명했으며, 관련 기사 건수는 61건(17.9%)이다. [토픽2]는 챗봇의 동성애 등 소수자에 대한 혐오 및 차별 발언 논란에 대한 것으로 '차별 혐오 논쟁'으로 토픽명을 부여하였고, 관련 기사는 81건(23.8%)이었다. [토픽3]은 서비스 개발에 사용된 카카오톡 메시지가 대화상대의 동의 없이 수집되었다는 논란과 관련된 것으로, '카톡 非 동의 수집'으로 명명하였고, 관련 기사는 72건(21.2%)이었다. [토픽4]는 충분히 비식별화되지 않은 카카오톡 메시지 데이터 공유로 인한 개인정보 유출에 대한 개발사의 대응과 해당 논란 관련 설명기사들에 대한 것으로 '개인정보 유출'이라고 명명하였고, 관련 기사는 40건(11.8%)이었다. [토픽5]는 개발사의 해명에도

5) 토픽 모델링 분석 목적이 논란 전개과정과 이후의 사회의 이슈를 파악하는 것이므로, 시기 0의 분석은 제외해도 무방하다고 판단하였음

6) 토픽 개수 선정을 위한 일관성 점수에 대한 절대적 기준은 없으나, 모든 시기의 반복된 분석에서 일관성 점수는 0.45~0.55 수준으로, 이는 수용할만한 것으로 판단하였음[11]

표 2 시기별 토픽 및 주요 키워드

토픽 및 주요 키워드	주요 뉴스 제목	뉴스 건수	비중(%)
시기 1: 21.01.08. ~ 21.01.12. (5일간)		340	100
토픽1 성희롱 논란 성희롱, 출시, 성적, 여성, 논란	• 출시 일주일 만에…'20살 AI 여성' 성희롱이 시작됐다 • AI 여성 캐릭터 '이루다' 출시되자 '성노예 만드는 법' 등장…사회적 논란 • [이효석의 게임인] SF게임 같은 'AI 성희롱'이 현실에 일어났다	61	17.9
토픽2 차별 혐오 논쟁 윤리, 문제, 차별, 혐오, 논란, 중단	• "인권? 듣기 싫은 말만 골라 하네" 누가 AI에게 性차별·혐오 심었나 • "AI가 동성애·장애인 혐오?"…이루다가 불붙인 'AI 윤리' 논쟁 • '이루다' 논란 커지자…AI윤리협회 "서비스 중단하고 개선 촉구"	81	23.8
토픽3 카톡 开동의 수집 수집, 카톡, 동의, 조사, 연인, 유출	• "前남친과 카톡대화, AI가 학습"…제3자 동의 없는 정보수집 논란 • 헤어진 그녀가 생각난다?…정부, 개인정보 유출 논란 'AI 이루다' 조사한다 • "이루다 개발사 직원들, 수집한 연인 카톡 대화 공유"	72	21.2
토픽4 개인정보 유출 알고리즘, 문장, 중단, 필터링, 이름	• 스캐터랩 "이루다 알고리즘, 문장 속 실명 전부 못 걸러냈다" • '이루다' 서비스 중단한 스캐터랩 "알고리즘 전면 개선" • [전문] 이루다 "비식별화 조치에도 문맥 따라 인물 이름 남은 점 사과"	40	11.8
토픽5 서비스 중단 혐오, 차별, 논란, 유출, 중단	• [전문] "혐오발언에 정보유출까지" '이루다' 서비스 논란 속 퇴장 • AI 챗봇 '이루다', 서비스 잠정 중단…"혐오·차별 발언 사과" • '혐오 논란' AI 챗봇 '이루다' 서비스 잠정 중단	86	25.3
시기 2: 21.01.13. ~ 21.01.15. (3일간)		256	100
토픽6 모델/DB 폐기 연애, 동의, 조사, 공유, 모델	• AI '이루다'는 폐기 수순… '연애의 과학' 카톡 100억건은? "전량 폐기 안 해" • 스캐터랩 "이루다 DB·딥러닝 모델 폐기…개인정보 동의 절차 강화할 것" • 이루다 개발사 결국 사과… "이용자 카톡대화, 온라인 공유했다"	140	54.7
토픽7 카카오 근절 원칙 카카오, 증오, 원칙, 근절, 차별	• 카카오, 온라인 게시물 '증오발언 근절 원칙' 마련 • 카카오 "차별·증오발언 강경 대응"…네 가지 원칙 발표 • 인권위 "카카오, 증오발언 대응 원칙 발표 환영"	26	10.2
토픽8 AI 윤리 필요 책임, 윤리, 사회, 기술	• AI 개발 급급해 윤리는 뒷전… 업계 "악해지지 말자" 뒤늦은 고민 • [이루다가 남긴 것⑦] 업계도 "윤리적 AI 필요" vs "인간부터 반성해야" • 방통위 "AI 이용자 차별 우려…사람 중심 정책기반 마련"	77	30.1
토픽9 알페스 논란 알페스, 성대결, 성희롱	• 이루다 후폭풍… 연예인 성희롱 '알페스'·'딥페이크' 처벌 국민청원 빗발쳐 • 性대결로 번진 AI '이루다' 논쟁…남녀 갈려 온라인 성폭력 비판	13	5.1
시기 3: 21.01.16. ~ 21.02.09. (22일간)		247	100
토픽10 피해자 집단소송 소송, 피해자, 유출, 신청, 증거, 보전	• "AI 이루다에 내 개인정보 유출"…집단소송 400여명 참여(종합2보) • AI챗봇 '이루다' 집단소송에 약 400명…'카톡 대화DB' 증거보전신청 • 개인정보 쏙 빼갔네!…이루다 피해자 스캐터랩에 집단소송	66	26.7
토픽11 AI윤리 정립 노력 윤리, 교육, 정책, 기술, 세미나	• 고학수 서울대 교수, "AI윤리 관련 논문 겨우 20~30건, 산학연 연구 필요" • 과기부, 'AI 윤리 정책세미나' 개최…"교육 방안 마련할 것" • 생활과 산업 곳곳 파고든 AI…'모럴 해저드' 대처 방안은	22	8.9
토픽12 AI 개발수칙 마련 보호, 동의, 안심, 처리, 수칙	• [일문일답] "동의 만능주의 없앤다…사전동의 제도 실질화 추진" • '이루다 사건' 재발 않게…인공지능 개인정보보호 준칙 만든다 • '제2의 이루다' 사태 막는다…정부, AI 개인정보보호수칙 3월 발표키로	42	17.0
토픽13 AI 차별규제 마련 윤리, 차별, 혐오, 법규 필요, 기술	• AI 기술보다 뒤처진 법, 차별·혐오 발언 논란 못 막는다 • 이루다 사건 재발 방지 위해서는 '윤리'와 '규제'가 필요 • AI의 소수자 차별 막으려면…법학자 "차별금지법 필요"	24	9.7
토픽14 알페스/젠더 갈등 알페스, 처벌, 아이돌, 착취, 여성	• 단순 팬덤인가 성범죄인가…커지는 '알페스' 논란 • 男 "알페스 생산자 처벌해 달라" vs 女 "우린 더 독한 영상에 당했다"	14	5.7
토픽15 기업관점의 대응 기술, 기업, 윤리, 보험, 인식, 리스크	• [현장에서] "기술에는 양심이 없다", CES의 또 다른 화두 • 한달만에 폐기된 이루다, AI윤리 논쟁 촉발… "신산업 족쇄" 우려도 • '이루다' 혐오발언 논란…"AI 리스크도 보험 들어야"	41	16.6
토픽16 시민단체 고발 인권, 수집, 동의, 보호법, 위반, 조사	• "이루다, 철저하게 조사하라"…민변·진보넷·참여연대, 진정서 제출 • '이루다' 개발사, 항의 폭주에 게시판 폐쇄…"2차 피해는 핑계" • "인공지능 기술 남용 막아야"…시민단체, 챗봇 '이루다' 국가인권위 진정	38	15.4

불구하고 논란이 가라앉지 않자 혐오 및 차별 발언, 개인정보 유출 등 다양한 이슈를 남기며 서비스 중단을 선언한 내용을 다룬 기사들로, '서비스 중단'이라는 토픽명을 부여하였다. [토픽5]는 시기 1에서 가장 높은 비중인 86건(25.3%)의 기사가 포함되었다.

나. 시기 2: 서비스 중단 직후

시기 2는 서비스 중단 직후 3일간의 기간으로 총 4개의 토픽이 도출되었다.

[토픽6]은 서비스 중단 이후에도 비식별 처리가 안 된 일부 학습데이터의 깃허브(git hub) 공유 논란 등이 계속되자 개발사가 '이루다' 서비스 학습 모델 및 데이터의 전량 폐기를 선언한 내용과 그럼에도 불구하고 '이루다' 서비스의 학습데이터의 모태가 된 '연애의 과학' DB는 폐기하지 않는다는 보도 내용과 관련한 것으로 '모델/DB 폐기'로 토픽명을 부여하였고, 관련 기사는 시기 2에서 가장 비중이 높은 140건(54.7%)이었다. [토픽7]은 1월 13일 카카오가 자사 플랫폼 내의 증오(혐오) 발언에 대한 강경 대응과 근절을 위한 4가지 원칙을 발표한 내용을 다루고 있다. '카카오 근절 원칙'으로 명명하였고, 관련 기사는 26건(10.2%)이다. [토픽8]은 '이루다' 서비스 논란 이후 인식된 AI에 대한 책임규명과 윤리 정립에 대한 필요성을 제기하는 내용으로 'AI 윤리 필요'로 명명하였고, 관련 기사는 77건(30.1%)이었다. [토픽9]는 '이루다' 논란이 성 대결 양상으로 확산하는 점을 보여준다. '이루다'를 성적(性的) 대상화 한 일부 남성에 대한 비난이 일자, 미성년 남성 아이돌을 성희롱한 내용이 다수 담긴 팬픽 문화, 일명 '알페스(RPS)'에 대한 처벌 청원이 덩달아 논란이 된 내용을 담고 있다. 이를 '알페스 논란'으로 명명하였고, 관련 기사는 13건(5.1%)이었다.

다. 시기 3: 논란 이후

시기 3은 논란의 정점 이후 한 달여간의 기간으로, 다양한 사회 의제의 형성 내용을 살필 수 있다. 분석 결과 7개의 토픽이 도출되었다. [토픽10]은 개인정보 유출 피해자들의 집단소송 준비 내용에 관한 것으로 '피해자 집단소송'으로 명명하였고, 관련 기사는 66건(26.7%)이다. [토픽11]은 학계, 정부의 AI 윤리교육 및 정책 마련 노력과 활동에 관한 것으로 'AI 윤리 정립 노력'으로 토픽명을 부여하였고, 관련 기사는 22건(8.9%)이었다. [토픽12]는 AI 개발 관점에서 개인정보보호 준칙 마련 및 사전동의 제도개선 등과 관련된 내용으로 'AI 개발수칙 마련'으로 명명하였고, 관련 기사는 42건(17.0%)이었다. [토픽13]은 AI 서비스 내에서의 혐오, 차별 등 알고리즘 편향에 대한 규제 관련 기사를 다룬 것으로 'AI 차별 규제 마련'으로 토픽명을 부여하였고, 관련 기사는 24건(9.7%)이었다. [토픽14] 시기 2에 이어 알페스 처벌 논란 및 남녀 갈등에 관한 기사들로 '알페스/젠더 갈등'으로 명명하였고, 관련 기사는 14건(5.7%)이었다. [토픽15]는 AI 활용에 대한 기업관점의 리스크 인식과 대응에 관한 내용으로 '기업관점의 대응'으로 명명하였고, 관련 기사는 41건(16.6%)이었다. [토픽16]은 지속되는 항의로 개발사가 게시판을 폐쇄했다는 내용과 시민단체들이 철저한 조사를 요청하며 관계부서에 진정서를 제출했다는 내용을 다룬 것으로 토픽명을 '시민단체 고발'로 명명하였고, 관련 기사는 38건(15.4%)이었다.

Ⅴ. 결과 토의

표 3은 '이루다' 서비스가 남긴 쟁점 사항을 시기별로 도출된 토픽 간의 연관성을 고려하여 도식화한 것이다.

표 3 '이루다' 논란의 쟁점/시기별 연관 토픽

구분	AI 윤리/편향성	개인정보 보호	성(性) 대결
시기1	T1. 성희롱 논란 T2. 차별 혐오 논쟁	T3. 카톡 캡동의 수집 T4. 개인정보 유출	
	T5. 서비스 중단		
시기2	T7. 카카오 근절 원칙 T8. AI 윤리필요	T6. 모델/DB 폐기	T9. 알페스 논란
시기3	T11. AI윤리 정립 노력 T13. AI차별규제 마련	T10. 피해자 집단소송 T16. 시민단체 고발	T14. 알페스/젠더갈등
	T12. AI개발수칙 마련 T15. 기업 관점 대응		

출처 저자 작성

'이루다' 서비스를 둘러싼 쟁점은 세 가지로 분류할 수 있는데, 서비스 내 성희롱, 챗봇의 혐오 차별 발언 등과 관련한 AI 윤리와 편향성에 대한 것이 첫 번째 쟁점이고, 서비스 개발과정에서 드러난 「개인정보보호법」 위반이 두 번째 쟁점이다. 끝으로, 알페스 처벌 청원 논란이 보여준 성(性) 대결 양상이 세 번째 쟁점이다.

첫째로, '이루다' 서비스는 사회적으로 통용되는 가치와 상반되는 차별적 발언을 하면서 논란의 중심에 서게 되었다. 일부 온라인 커뮤니티 사이트를 중심으로 '이루다'를 성적 대상으로 놓고 희롱하는 행태가 유행하면서 논란이 야기되었다. 온라인 커뮤니티 '아카라이브', '디시인사이드' 등에서 '이루다'를 성적 대상화하고 이러한 대화를 인증하는 게시물들이 등장하면서, AI 챗봇에 대한 성적 학대·악용에 대한 문제가 제기되었다. 또한 '이루다'는 학습된 데이터들을 바탕으로 장애인, 성소수자, 흑인, 임산부 등의 특정 소수 집단에 대해 혐오와 편견이 담긴 발언들을 하였으며[13], 개발사에서는 이를 제대로 걸러내지 못하면서 논란의 대상이 되었다. 이러한 논란은 서비스 중단(토픽5) 이후에도, 서비스 내 혐오 및 차별 근절 원칙(토픽7), AI 윤리 고려에 대한 필요성(토픽8), 실질적인 AI 윤리 교육과 정책 마련(토픽11), AI 차별규제 및 차별금지법 논의(토픽13), 기업관점에서의 AI 리스크 관리(토픽15) 등 다양한 의제를 생산하였다.

'이루다' 사태의 두 번째 쟁점은 '이루다' 서비스가 「개인정보보호법」을 위반하였다는 논란이다. 스캐터랩은 개발 과정에서 '텍스트앳'과 '연애의 과학' 이용자와 이용자의 연인에게 개인정보 이용·활용 동의를 제대로 받지 않았다는 의혹을 받았다. 또한, 데이터 비식별화(익명화)를 위한 처리가 부족하여 '이루다'와의 대화 과정에서 개인정보가 노출되거나, 제대로 비식별화하지 않은 데이터를 오픈소스 공유 플랫폼 '깃허브'에 공유하기도 하는 등의 여러 가지 문제가 발생하였다[7]. 이러한 문제들로 인해 스캐터랩은 1월 12일부터 「개인정보보호법」 위반혐의로 대한 조사를 받고 있으며, 개발사는 1월 15일 '이루다' 서비스의 DB와 딥러닝 대화 모델을 모두 폐기(토픽6)하겠다고 발표하였다. 발표 이후에도 해당 논란은 피해자 400여 명의 집단소송(토픽10), 시민단체들의 철저한 조사를 촉구하는 진정 제출(토픽16), AI 개발수칙 마련 논의(토픽12) 등으로 이어졌음이 확인되었다.

끝으로, '이루다' 서비스와 직접적 관련이 낮음에도 알페스 처벌 청원을 둘러싼 온라인상에서의 격화된 논쟁(토픽9, 13)은 우리 사회에 잠재된 오래된 성(性) 갈등 구조를 보여주었다고 할 수 있다.

VI. 결론

본 연구에서는 AI 챗봇 '이루다' 서비스를 둘러싼 논란과 쟁점을 탐색하고자 네이버 뉴스에 등재된 뉴스 기사 963건을 수집하여 키워드 분석과 토픽 모델링을 수행하고, 도출된 다양한 토픽 간의 연관성을 고려해 시간의 흐름에 따른 이슈변화의 구조화를 시도하였다.

분석의 결과는 AI 서비스가 편향성 및 윤리 이슈의 제기뿐만 아니라, 개인정보보호 문제로 확대되고 젠더 갈등과 같이 예상치 못한 방향으로도 전개될 수 있음을 보여준다. 또한, 현재의 AI 서비스가 채택하고 있는 딥러닝 방식의 갖는 데이터 의존성의 문제는 학습데이터 확보와 개인정보보호가 끊임없이 상충하게 되는 문제임을 상기시킨다. 아울러, 분석 결과는 AI 서비스의 위법성 논란, 소송 등으로 기업의 혁신 활동이 위축되는 것을 방지하기 위해서는 실질적 AI 개발수칙과 개발자의 윤리교육 등 기업의 리스크 관리를 위한 사전적 대응이 중요해지고 있음을 시사한다.

사실, AI 서비스로 빚어진 윤리적 문제는 새로운 것은 아니다. AI 윤리와 개인정보보호 강화는 최근 IT 업계가 가장 고민하는 화두 중 하나이다. 그러나 문제는 이러한 접근이 다분히 추상적이고 선언적 수준에 그치고 있어, 실제 현장에선 적용할 내용이 부족하다는 점이었다.

이러한 점에서 이번 '이루다' 논란은 AI 윤리이슈가 이제는 선언적인 문제 또는 아직 다가오지 않은 미래 이슈가 아닌 당면한 문제임을 사용자와 개발자 모두가 자각하는 계기가 되었다는 점에서 의미가 있다.

기술의 발전과 함께 AI 서비스는 갈수록 일상생활영역까지 확대될 것이다. AI의 대중화를 피할 수 없다면, 이번 사태를 계기로 "어떻게 AI와 공존해야 하는가?"에 대한 우리 사회의 더욱 심도 있고 다양한 논의가 전개되기를 희망해 본다.

용어해설

토픽 모델링 문서의 집합을 여러 개의 숨겨진 주제(hidden topic)로 분류하는 알고리즘

팬픽 팬 문화의 하나로, 팬이 만들어 낸 연예인 중심의 허구. 팬픽션(fan fiction)의 줄임말

약어 정리

LDA	Latent Dirichlet Allocation
NLP	Natural Language Processing

참고문헌

[1] Markets and Markets, "Chatbot market global forecast to 2024," 2019.

[2] 김현지, "AI 편향·혐오 발언 '이루다'만이 아니다…MS·아마존도 겪었던 문제," 머니투데이, 2021. 1. 15.

[3] 최새솔, "안면인식기술 도입의 사회적 논란과 시사점: 미국사례 중심으로," ETRI 기술정책 브리프, 2020. 12.

[4] 오달란, "AI 챗봇 '이루다' 성희롱 논란에 개발사 "예상한 문제…자정 노력 부탁"," 서울신문, 2021. 1. 8.

[5] 임재우, "AI '이루다' 멈췄지만…성차별·혐오는 인간에게 돌아온다," 한겨레, 2021. 1. 11.

[6] 최광민, "스캐터랩, 세계 최고 수준의 언어능력 보유한 인공지능 '이루다' 정식 출시," 인공지능신문, 2020. 12. 23.

[7] 이효석, "AI 이루다, 논란 일주일 만에 사실상 종료…"중추신경계 폐기"(종합2보)," 연합뉴스, 2021. 1. 15.

[8] 장영은, "AI 알고리즘 편향성 논란, '이루다'가 처음 아니다," 이데일리, 2021. 1. 13.

[9] 권민지, "토픽 모델링 기반 뉴스 기사 분석을 통한 서울시 이슈 도출," 한국방송·미디어공학회 추계학술대회, 2019, pp. 11-13.

[10] 박소현 외, "토픽 모델을 사용한 베스트셀러 서적 단문 의미 분석 연구," 영상문화콘텐츠연구, 제15권, 2018, pp. 101-112.

[11] 하영욱, "ETRI 이슈 리포트, 코로나-19 이후의 비대면 사회 이슈 변화 분석('20년 상반기)," 2020.

[12] 배기웅, 김혜진, "초기 시청자 반응과 드라마 평균 시청률 사이의 관계: 토픽 모델링 측면에서, 방송문화연구," 제31권 제1호, 2019, pp. 103-138.

[13] 이상진, "[인공지능]② 인간의 혐오와 편견이 만든 'AI챗봇 이루다' 논란," 뉴스포스트, 2021. 2. 2.