

생성형 AI의 지상파 방송 뉴스 데이터 이용의 적정 대가는?

변상규
호서대학교 문화영상학부 교수

목차

- 1 서문
- 2 대규모 언어 모델과 뉴스 데이터
- 3 저작권료 산정 모형
- 4 저작권료 산정
- 5 결론 및 시사점

요약

최근 대중화되는 생성형 AI는 딥러닝 기술을 이용하여 학습 단계를 거친다. 그런데 학습 데이터로부터 성능이 큰 영향을 받는다. 주로 인터넷에서 크롤링 등으로 데이터를 확보하는데, 뉴스 등 저작권이 있는 데이터에 대해서도 저작권료 지불을 회피하면서 저작권자와의 갈등이 심화되고 있다. 그러므로 저작권자와 AI 개발사 모두 공생할 수 있는 합리적인 저작권료 산정 필요성이 제기된다.

본고에서는 AI의 학습 및 서비스에 이용되는 뉴스 데이터의 합당한 저작권료를 추정하는 모형을 개발하였다. 학계 및 업계 전문가, 선행연구, 관련 법령 등을 두루 검토하고, 수익접근법과 원가접근법, 1년 단위의 로열티 지급방식을 채택하였다. 그리고 AI 서비스의 이용자 효용과 데이터 제작비용을 추정에 이용하였다.

1. 서문

최근 인공지능(Artificial intelligence, AI)이 출현하고 성능이 급속히 향상되면 서 인류는 경이와 우려 속에서 정보통신 사회의 절정기를 맞고 있다. 최근에 딥러닝(deep learning) 기술을 이용한 생성형(generative) AI가 출현하면서 AI 서비스가 본격적으로 대중화되고 있다. 특히 챗GPT는 출시 2년여 만에 가입자가 5억 명을 넘어선 것으로 알려지면서 상업적으로 크게 성공한 생성형 AI 서비스로 자리잡았다(송영찬, 2025.4.2.).

생성형 AI 모델은 추론을 통해 스스로 예측하여 문장, 그림, 영상, 음악 등 다양한 결과물을 생산한다. 이를 위해 AI는 개발단계에서 자료를 학습하는 과정을 거친다. 대규모 언어모델(Large Language Model, LLM)의 경우 다양한 어문 자료를 학습하여 인간의 언어를 습득한다. AI 개발자들은 인터넷에서 크롤링(crawling)으로 확보한 자료를 학습에 주로 사용하며, 이 과정에서 자료의 복제와 저작이 이루어진다. 그런데 그 중에는 저작권이 있는 자료도 포함되어 있다. 특히 언론사가 생산하는 뉴스 데이터는 AI의 학습에 요긴한 고품질 자료로 인정받고 있으며, 저작권자가 있다. 그러나 AI 개발자들은 공정이용(fair use)에 해당한다고 주장하면서 저작권자에게 이용 허락이나 저작권료 지불을 거부하는 경우가 많다. 그래서 AI 사업자와 저작권자 사이에 갈등이 발생하고 법정 다툼으로까지 진행되기도 한다. 국내에서는 하이퍼클로바 X를 개발 중인 네이버와 지상파 3사가 뉴스 데이터에 대한 저작권료 소송을 벌이고 있다.

그런데 최근 AI 사업자들이 저작권자와 계약을 체결하여 학습 데이터의 이용료를 지불하는 사례도 나타나고 있다. 그리고 국내의 AI 전문가들도 학습 데이터에 대한 저작권료 지급에 대해 큰 이견이 없었다. 그러나 합당한 근거와 수준에 대한 우려가 크다(양현석 등, 2024.12). 그러므로 저작권자와 AI 개발사 모두 공생할 수 있는 합리적이고 객관적인 저작권료를 산정할 필요성이 제기된다.

본고에서는 국내 지상파 3사가 생산한 뉴스 데이터를 대규모 언어모델 생성형 AI가 이용하는 것에 대한 합당한 저작권료 수준을 산정한다. 이 과정에서 학계 및 업계 전문가, 선 행연구, 관련 법령 등을 두루 검토한다.

2. 대규모 언어 모델과 뉴스 데이터

1) 생성형 AI의 학습과 저작권 이슈

생성형 AI는 산출물에 따라 언어·이미지·음성 등으로 구분할 수 있으며, 최근에는 복수 유형의 산출물을 생성하는 다중 모달(multimodal) 모델도 출시되었다. 오픈 AI가 개발한 ‘챗지피티(ChatGPT)’, 구글의 ‘제미나이(Gemini)’, 앤트로픽(Anthropic)의 ‘클로드(Claude)’, 메타(Meta)의 ‘라마(Llama)’, 마이크로소프트의 ‘코파일럿(Copilot)’, ‘미드저니(Midjourney)’, ‘퍼플렉시티(Perplexity)’등이 출시되어 있다. 국내에서 개발된 AI는 SK텔레콤의 ‘에이닷(adot.ai)’, 네이버의 ‘하이퍼클로바 X(HyperCLOVA X)’, ‘뤼튼(Wrtn)’ 등이 있다.

본고에서 중점적으로 다루는 대규모 언어모델 생성형 AI는 주어진 맥락에서 확률이 가장 높은 단어들을 순차적으로 예측하여 연관성 높은 문장을 출력한다(문화체육관광부·한국저작권위원회, 2023.12.27.). 이 때 AI는 각 연결마다 입력되는 신호가 출력값에 미치는 중요도를 의미하는 가중치(weight) 값을 사용하는데, 사전에 어문 데이터를 학습하여 그 값을 결정한다. 그리고 학습을 위해서 데이터의 획득, 수집, 정제, 가공을 거친다(과학기술정보통신부·한국지능정보사회진흥원, 2022.9). 그런데 AI 개발자가 인터넷상의 데이터들을 탐색, 수집할 때 복제가 일어난다. 그리고 이를 입력 데이터로 가공하는 과정에서 다시 복제가 발생하며, 이후 백업 혹은 검증의 목적으로 추가적인 복제가 발생할 수 있다(류시원, 2023.1). 그리고 분석을 위해 데이터를 읽어 들이게 되는데, 이때에도 일시적인 복제가 발생한다. 업계의 전문가는 AI 학습 과정에서 발생하는 복제는 일시적이며, 분석이 끝난 후에 데이터를 삭제하므로 AI 시스템에 학습 데이터가 남아 있을 가능성이 낮다고 주장한다(이철남, 2024.10; 양현석 등, 2024.12). 그러나 생성형 AI가 학습 데이터를 복제하여 출력물을 만드는 현상들이 실제로 보고되고 있다(신용우 등, 2024.9). AI 학습과정에서 발생하는 복제와 이로 인한 결과물의 유사성은 AI 개발자와 저작권자 사이에 벌어지는 법적 다툼에서 주요 쟁점이 되고 있다(권구민·김현석, 2023.2.27.).

2) 뉴스 데이터의 기여도

AI 학습에 많은 데이터가 필요하지만 특히 학습의 효율성을 높이기 위해서는 고품질 데이터가 필수적이다. 언론사가 만드는 뉴스 데이터는 언어모델 AI가 자연스러운 문장을 생성하는 등 언어능력을 높이는데 필요한 고품질 데이터로 인정받는다(양현석 등, 2024.12). 그리고 기자 등 전문가들에 의해 사실을 기반으로 작성되고, 게이트 키핑(gatekeeping) 과정도 거치는 등 신뢰성 높은 정보를 제공한다. 그래서 AI의 환각(hallucination)을 줄이는데 도움이 된다(한국언론진흥재단, 2024; 김선호, 2025.1.3.). 또한 정치·경제·사회·문화·스포츠 등 다양한 주제를 폭넓게 다루므로 AI가 다양한 분야의 어휘를 학습할 수 있으며, AI 모델이 최신성을 갖추는 데 중요한 역할을 한다. 또한 양적으로도 풍부해서 데이터 부족을 해결하는데 기여한다.

생성형 AI의 개발에 뉴스 데이터가 얼마나 이용되는지에 대한 자료는 부족하다. 쟁 GPT3의 경우 학습 데이터 중에서 뉴스 저작물의 비중이 5~10%라는 비공식적인 추정치가 알려져 있다(한국언론진흥재단, 2024). 하이퍼클로바 X는 개발 과정에서 총 5,618억 토큰을 학습에 사용하였는데, 이 중 블로그 자료가 2,736억 토큰, 온라인 카페 자료가 833억 토큰, 뉴스 자료가 738억 토큰, 댓글이 411억 토큰, 지식인 자료가 273억 토큰이 사용된 것으로 네이버 측이 공식적으로 발표한 적이 있다(한국언론진흥재단, 2024). 그러므로 뉴스 데이터는 토큰 측면에서 약 13.1%를 차지한다.

AI의 학습이 끝난 후에도 뉴스 데이터가 AI 서비스에 이용될 수 있다. 최근 출현한 ‘검색증강생성’(Retrieval Augmented Generation, RAG) 모델은 새로운 정보를 외부에서 검색하고 그 내용을 결과물에 통합할 수 있도록 지원한다. 그러므로 학습 데이터의 정보를 최신 뉴스 정보로 보완하여 더 정확하고 풍부한 결과물을 제공한다(한국언론진흥재단, 2024). 이 기술을 이용하면 AI의 결과물이 뉴스 기사의 이용을 대체하여 언론사의 전통적인 뉴스 시장을 잠식할 가능성이 있다(김선호, 2025.1.3.). 이러한 시장 대체효과는 저작권료 분쟁에서 중요하게 고려되는 요소이다.

그러므로 뉴스 저작물은 AI에게 고품질의 학습 데이터임과 동시에, AI 서비스의 신뢰성과 최신성을 향상시키는 데 기여한다. 그러므로 저작권료 산정 범위도 이에 부합하도록 확장되는 것이 바람직할 것으로 판단된다(한국언론진흥재단, 2024; 김창화, 2024.2.29.).

3. 저작권료 산정 모형

AI의 학습에 저작물을 이용하려면 공개된 저작물이어도 저작권자의 동의가 필요하며, 면책 여부와 관계없이 대가를 지급해야 한다는 주장이 있다(김윤명, 2023; 한국언론진흥재단, 2024; 신용우 등, 2024.9). 해외에서는 최근 오픈AI가 AP통신이나 콘텐츠 라이브러리 셀러스톡(Shutterstock) 등으로부터 학습 데이터를 구입하였다(임대준, 2023.7.14.). 국내에서도 AI 전문가들이 데이터에 대한 저작권료 지불의 필요성을 인식하고 있다(양현석 등, 2024.12). 그러므로 합리적인 저작권료 산정 방안을 찾아서 데이터 거래를 활성화하여 AI 개발을 촉진하고 저작권자의 권리 보호도 도모할 시점이다(문화체육관광부·한국저작권위원회, 2023.12.27.).

1) 저작권료 산정 근거

생성형 AI는 다양한 데이터 소스를 학습하기 때문에 기존 디지털 플랫폼과 콘텐츠 제작자 사이의 수익배분 방식을 적용하기가 어렵다(Wang et al, 2024). 그러므로 새로운 방식을 모색할 필요가 있는데, 해외에서는 저작물당 요금을 설정한 사례가 있다(이철남, 2024.10). 어도비(Adobe)는 이미지를 개당 6~16센트, 동영상은 분당 2.62~7.25달러에 구매한 것으로 알려진다. 디파인드 AI(Defined.ai)는 구글, 메타, 애플, 아마존 등에 사진은 1~2달러, 영상은 2~4달러, 긴 영화는 100~300달러, 문자는 1,000자에 1달러를 받았다(박찬, 2024.4.7.). 그리고 일정 기간 사용료를 지급한 사례도 있다. 오픈AI는 월스트리트저널(The Wall Street Journal)에 5년 동안 3,500억 원, 파이낸셜타임즈(Financial Times)와 액셀 스프린저(Axel Springer)에 3년 동안 각 350억 원을 지불할 것으로 조사된다(양현석 등, 2024.12). 커뮤니티 사이트 레딧(Reddit)은 구글에게 연간 6천만 달러를 받고 있고, 애플(Apple)은 언론 및 출판사들과 5천만 달러 규모의 다년 계약을 체결한 것으로 알려진다. 그런데 이 정보들은 과편적이며, 계약의 근거나 내역을 자세히 파악하기가 어렵다. 그리고 금액의 편차가 심하며, 해외 사례이므로 시장 규모가 다른 우리나라에서 참고하기는 어렵다.

한편, 국내 「저작권법」에 따르면 저작권리자는 침해자의 이의액이나 일반적으로 받을 수 있는 통상사용료 상당액을 추정해서 손해배상을 청구할 수 있다(제125조 제1항, 제2

항)¹⁾. 그런데 이익액을 구하기 위해서는 특정 데이터의 기여율을 알아야 하는데, AI의 산출물이 어느 학습 데이터에 얼마나 의존하여 생성되었는지를 정확하게 입증하기가 어렵다 (Linden, 2017; 김윤명, 2023). 그리고 통상사용료는 AI 분야에서 알려진 바가 거의 없다.

학계의 연구 사례를 살펴보면 양현석 등(2024.12)은 전자책 및 전송서비스를 참고하여 4가지 방안을 제안하였다. 즉, 언론진흥재단의 학습용 뉴스 데이터 가격, ‘매일경제신문’의 데이터 거래소 가격, 출판업의 경상로열티 수준, 생성형 AI 개발사업의 예산 규모 등을 기준으로 저작권료를 설정할 것을 제안하였다. 그러나 앞의 세 가지 가격체계는 데이터의 이용 목적이 AI와 전혀 다르며, 시장의 규모도 크게 다르다는 한계가 있다. 마지막 방안은 어문 자료, 이미지, 음향 자료 등 데이터의 형식에 따라 사업마다 비용 차이가 크다. 그리고 개발자가 이미 확보한 자료와 새로 수집해야 하는 자료일 경우, 자체적으로 생산하거나 혹은 외부에서 구매할 경우 등에 따라 차이가 크다. 그리고 자료 확보과정에서의 난이도도 상이하다. 그러므로 하나의 값으로 종합화하기가 매우 어렵다.

Chiu & Chen(2003)은 특허권의 가치를 평가하는 기준과 가중치를 계층분석적 의사결정 방법(Analytic Hierarchy Process, AHP)으로 도출하였다. 그러나 지표들 사이의 가중치를 결정하였을 뿐, 이 결과를 이용하여 저작권료의 규모를 산정하지는 못한다.

Wang et al.(2024)은 AI가 생성한 콘텐츠에서 저작권자의 기여도를 협동 게임 이론(Cooperative Game Theory)으로 측정하였다. 콘텐츠를 생성할 확률의 로그 우도(log-likelihood)를 활용해 모델의 예측 정확도를 효용함수로 삼아서 학습 데이터의 효용과 기여도를 평가하였다. 그리고 샤플리 밸류(shapley value)를 구하여 수익을 배분하였다. 그런데 특정 데이터 세트를 제외하거나 포함시켜 AI를 학습시키는 과정을 반복적으로 시뮬레이션하여야 하는데, 이를 상업용 거대 AI에 적용하기는 시간적, 비용적 측면에서 거의 불가능하다.

1) 제125조(손해배상의 청구) ①저작재산권 그 밖에 이 법에 따라 보호되는 권리(저작인격권 및 실연자의 인격권은 제외한다)를 가진자가 고의 또는 과실로 권리를 침해한 자에 대하여 그 침해행위에 의하여 자기가 받은 손해의 배상을 청구하는 경우에 그 권리를 침해한 자가 그 침해행위에 의하여 이익을 받은 때에는 그 이익의 액을 저작재산권자등이 받은 손해의 액으로 추정한다.

②저작재산권자등이 고의 또는 과실로 그 권리를 침해한 자에게 그 침해행위로 자기가 받은 손해의 배상을 청구하는 경우에 그 권리의 행사로 일반적으로 받을 수 있는 금액에 상응하는 액을 저작재산권자등이 받은 손해의 액으로 하여 그 손해배상을 청구할 수 있다.

한편, 저작권료가 저작권의 객관적인 경제적 가치를 의미하며, 그 가치를 평가하는 방법으로 시장접근법(market approach), 수익접근법(income approach), 원가접근법(cost approach)이 다수의 전문가나 선행연구에서 제안되었다(강신하, 2018; 한국언론진흥재단, 2024; 이규호·남선우, 2024). 시장접근법은 경매 모델을 기반으로 하는데, AI 데이터는 경매시장이 없는 상황이다. 수익접근법은 AI 데이터가 창출할 이익을 평가한다. 원가접근법은 데이터의 생산비용을 기준으로 하므로 직관적이다. 그러나 뉴스는 AI 학습을 위해서 제작된 콘텐츠가 아니므로, 뉴스 제작비 전액을 AI에 부담시킬 수는 없다. 그러므로 비용 중에서 어느 비율만큼을 AI에서 회수하여야 하는지가 또 다른 과제가 된다.

그리고 저작권료 산정에 반영되어야 할 요인들도 구체적으로 제안되었다. 김선호(2025.1.3)와 한국언론진흥재단(2024), 양현석 등(2024.12)에서 전문가들을 대상으로 시행한 토론 및 인터뷰에서 제안된 고려요인들을 정리하면 다음과 같다. 첫째, 데이터의 품질이다. 고품질 데이터가 고품질 서비스로 연결되므로 저작권료가 데이터의 품질을 반영해야 한다. 둘째, 콘텐츠의 최신성이다. 최신 기사는 AI 모델의 최신성 유지에 도움을 주고 이용자의 정보 욕구도 충족시키므로, 1년 이내의 최신 기사에 높은 가격을 부과해야 한다. 셋째, 데이터 이용 범위이다. AI 모델이 데이터를 학습용으로 사용하는지 또는 출시 후에 서비스용으로도 사용하는지에 따라 가격이 달라져야 한다. 넷째, AI 개발자의 규모와 서비스 이용자의 규모를 반영하여야 한다. 다섯째, 계약기간이다. AI가 데이터를 학습한 후에는 해당 데이터가 불필요하지만 그 후에도 AI는 경제적 가치를 창출한다. 그러므로 1회 이용과는 다른 대가 산정 체계가 필요하다. 여섯째, 과학기술정보통신부에서 수행 중인 학습용 데이터 구축지원 사업에서 데이터 구축 예산의 규모를 저작권료에 반영하여야 한다.

그 외에도 AI의 사용 범위가 내부용인지 또는 외부용인지, 뉴스 데이터가 분석기사인지 또는 보도기사인지, AI 서비스가 단일 용도인지 또는 다용도인지, 개발된 AI가 영리적 인지 또는 비영리적인지를 고려해야 한다는 주장이 있다. 그리고 AI의 연산력, 기사 1단위의 원가, 저작물의 잔여 보호기간, 동일한 영역에서 유사한 저작물에 인정되는 사용료 등을 고려하자는 의견도 있다.

한편, 저작권료 지불 방식도 중요한 고려사항인데, 개발자는 정액제를 선호하는데 비해 저작권자는 종량제를 합리적이라고 주장한다.

2) 모형의 프레임

본고에서는 국내 지상파 3사의 뉴스 데이터를 AI 개발자가 이용할 때 지불할 합당한 저작권료 산정 모형을 구축한다. 이 모형의 틀은 전술한 전문가 의견, 선행연구, 법령 등을 고찰하여 만들었다. 우선 주요 고려요인을 검토하면 데이터의 품질은 뉴스 데이터가 AI의 성능 향상에 기여한 내용을 선별하고, 매출액에 대한 기여도를 분석하여 반영한다. 콘텐츠의 최신성은 매년 로열티 방식을 채택하면 반영된다. 기사 전체적으로는 매년마다 거의 일정한 비율의 1년 이하의 최신기사가 기사 풀(pool)에 새로 추가되고 기존의 최신기사는 1년 이상의 기사 풀로 이전된다. 그래서 전체 기사 풀에 대해 매년 로열티를 지불하면 최신성에 대한 보상이 자동으로 이루어진다. 데이터 이용 범위는 학습과 서비스 단계로 나누어 분석하여 반영된다. AI를 개발하는 기업의 규모는 시장에서의 사적 거래에서는 따로 고려하지 않아도 될 것으로 판단된다. 서비스 최종 이용자 규모는 매출액을 기준으로 저작권료를 산정하여 반영된다. 계약기간은 연간 로열티 방식으로 반영된다. 과기정통부의 데이터 구축 지원 사업의 예산은 전술한 이유로 배제하였다.

추가로 AI의 사용 범위가 내부용인지 또는 외부용인지, AI 서비스가 단일 용도인지 또는 다용도인지, AI가 영리성인지 또는 비영리성인지 등 구매자의 상황은 상업적 거래에서는 중요한 고려사항이 아니므로, 모형에서 배제하였다. 분석기사와 보도기사의 경우 AI 학습에 어떤 차별적인 영향을 미칠지를 확신하기가 어렵다. AI의 연산력은 AI 서비스의 성능에 영향을 미칠 것이므로 매출액에 반영될 것이다. 기사 1단위의 원가 기준은 원가접근법에 반영이 가능하다. 저작물의 잔여 보호기간은 연간 로열티 방식으로 반영 가능하며, 동일한 영역에서 유사한 저작물에 인정되는 사용료는 아직 없다.

그러므로 저작권료 산정모형에서는 수익접근법과 원가접근법, 1년 단위의 로열티 지급 방식을 선택한다. 또한 뉴스 데이터의 이용 범위를 AI 학습과 서비스 단계로 구분하여 매출액 증가를 추정한다. 수익접근법에서는 뉴스 데이터로 인해 AI에서 향상되는 기능을 도출하고, 그 기능으로 창출되는 매출액 증가분을 추정한다. 원가접근법에서는 지상파 3사의 뉴스 제작비용의 일정 비율을 AI에 배분한다.

4. 저작권료 산정

1) 수익접근법

뉴스 데이터가 AI 서비스 이용자에게 제공하는 가치를 토대로 매출이 창출된다. 그러므로 데이터의 가치를 추정하기 위해서는 매출의 토대가 되는 이용자의 효용을 추정할 필요가 있다. 여기에 월 이용료 자료를 이용할 수 있는데, 무료 AI 서비스가 많고 이용량과는 무관한 정액요금제가 대부분이다. 그래서 시장의 현시선후자료(revealed preference data)가 충분치 않다. 이 경우에는 유사 재화의 현시선후자료를 이용할 수 있으나, AI 학습용 데이터처럼 혁신적인 재화에 대한 유사 재화를 찾기가 어렵다. 그런데 이용자가 재화에 대한 선호를 평가할 수 있는 경우에는 선호를 직접 물어서 진술선후자료(stated preference data)를 수집하여 분석할 수 있다. AI 서비스의 성능이나 품질은 이용자들이 체감할 수 있으므로, 뉴스 데이터로 인해 개선되는 AI의 기능이나 속성들이 이용자에게 제공하는 효용을 이용자 조사를 통해 추정할 수 있다.

진술선후자료를 이용하여 지불의사액을 추정할 때 컨조인트(conjoint) 분석이 많이 활용된다. 컨조인트 분석은 상품의 여러 속성에 대해 소비자의 평가나 지불의사를 분리하여 추정할 수 있는 간접적인 방법이다. 그러므로 다양한 속성들이 조합된 상품의 효용을 추정할 수 있다. 그래서 컨조인트 방법으로 설문조사 문항을 만들고 선호를 물었다.

저작권료를 분석하기 위해 뉴스 데이터가 기여하는 개선 사항을 학습 단계와 서비스 단계로 구분하였다. 학습단계에서 뉴스 데이터는 언어능력의 고급화에 기여하는 것으로 정의하였다. 구어체 위주의 블로그, 카페에 비해 문어체 위주의 고품질 뉴스 데이터가 AI 언어능력의 고급화를 지원하는데, 설문조사에서는 응답자들의 이해를 돋기 위해 그 효과를 ‘고등학생 수준’에서 ‘대학생 수준’으로의 개선으로 표현하였다. 활용 단계에서는 AI가 뉴스 기사를 검색하고 이를 정리해서 정확하고 최신성 있는 정보나 답변을 제공할 수 있는 기능으로 정의하였다.

그 외에도 그림/사진 인식 및 그리기, 환각, 음성 인식/출력, 동영상 만들기, 월 이용료 등을 추가하여 총 7개의 속성으로 AI 서비스 상품을 구성하였다. 월 이용료는 생성형 AI

서비스의 최저요금제가 대부분 30,000원 수준임을 우선 고려하였고, 무료 요금제 이용자가 아직 많기 때문에 사회적인 체감 요금수준이 그보다 낮을 것으로 가정하여 15,000원을 하한 요금으로 설정하였다. 주효과 직교설계(main effect orthogonal design)를 거쳐 8 개의 상품 카드를 구성하였다. 그리고 각 상품 카드별로 1~10점 사이의 점수를 부여하도록 요구하였다.

[표 1] 생성형 AI 서비스 상품에 대한 선호도를 묻는 컨조인트 문항

Q5-1.

다음은 언어모형 생성형 AI 서비스 상품을 보여주는 카드입니다. 각 상품에서 제공하는 서비스 내용을 살펴보시고, 각 상품에 대해 선호도를 1~10점 사이의 점수를 부여해 주세요. (가장 선호 10점, 가장 비선호 1점, 동점도 가능) (중복은 두 개까지만 제한)

| 상품카드 번호 | 한국어 구사 능력 | 뉴스/보도 검색, 정리, 출처 제공 | 그림/사진 인식 및 그리기 | 활각 | 음성 인식/출력 | 동영상 만들기 | 월이용료 |
|---------|-----------|---------------------|----------------|----|----------|---------|--------|
| 1 | 대학생 | 검색불가 | 가능 | 없음 | 가능 | 가능 | 15,000 |
| 2 | 고등학생 | 검색가능 | 가능 | 있음 | 가능 | 불가능 | 15,000 |
| 3 | 대학생 | 검색가능 | 불가능 | 있음 | 불가능 | 가능 | 15,000 |
| 4 | 고등학생 | 검색불가 | 불가능 | 있음 | 가능 | 가능 | 30,000 |
| 5 | 대학생 | 검색가능 | 불가능 | 없음 | 가능 | 불가능 | 30,000 |
| 6 | 고등학생 | 검색불가 | 불가능 | 없음 | 불가능 | 불가능 | 15,000 |
| 7 | 대학생 | 검색불가 | 가능 | 있음 | 불가능 | 불가능 | 30,000 |
| 8 | 고등학생 | 검색가능 | 가능 | 없음 | 불가능 | 가능 | 30,000 |

진술선호자료를 확보하기 위해 2025년 9월 2~12일에 설문조사를 시행하였다. 전국에 거주하는 16~60세 주민 중에서 생성형 AI 서비스 이용 경험을 가진 응답자 1,300여 명을 대상으로 인터넷을 통해 조사를 진행하였다. 불성실한 응답을 제외하고 총 1,297명의 응답을 분석에 사용하였다. 서열프로빗(ordered probit) 방법으로 분석한 결과 모든 설명변수가 유의수준 1%에서 선호에 대해 유의미한 설명력을 가진 것으로 나타났다. 그러므로 속성의 선별이 잘 이루어진 것으로 판단된다.

분석 결과를 이용하여 속성별로 한계지불의사액을 구한 후, 지상파 3사 뉴스 데이터의 기여율을 추정하여 적용하고 이를 국민경제 차원으로 확장하면 최종적인 저작권료를 산정할 수 있다. 기여율의 추정은 다시 두 단계로 이루어진다. 첫 번째 단계에서는 AI의 언어능력과 최신성 두 기능에 대해 뉴스 데이터가 기여하는 정도를 추정하고, 두 번째 단계에서는 뉴스 데이터에서 지상파 3사의 데이터가 기여하는 정도를 추정하여야 한다. 그런데 기여율을 구하는 데 필요한 자료가 없으므로 합리적인 가정을 추가하였다.

첫 번째 단계에서 하이퍼클로바 X의 학습 데이터 자료를 활용할 수 있겠다. 뉴스는 738억 토큰으로 학습 데이터에서 13.1%를 차지하는데, 나머지 자료는 모두 구어체 위주의 자료이다. 273억 토큰이 사용된 지식인의 일부 정도가 조금 나은 품질일 것으로 생각된다. 논문이나 도서, 연구보고서 등도 고품질의 어문 데이터이지만 이에 대한 내용은 확인되지 않는다. 다만 627억 토큰을 차지하는 기타 부문에 일부가 포함되었을 것으로 가정하였다. 그러나 도서는 디지털화된 비중이 낮고, 논문이나 연구보고서 등은 회원제로 이용되거나 유료화된 경우가 다수 있어서 AI의 접근성이 낮을 것이다. 그래서 기타 부문의 일부와 지식인을 합치면 900억 토큰으로 뉴스 데이터보다 소폭 많지만, 뉴스와 ‘지식인+기타’의 기여도를 50대 50으로 가정하였다. 이는 뉴스 데이터에게는 매우 보수적인 접근이다.

최신성은 주로 최신 정보의 검색일 것이므로, 뉴스 데이터의 기여도가 압도적일 것이다. 논문이나 보고서도 검색 대상에 포함되겠지만, 대중의 이용과는 거리가 있고 접근성도 떨어진다. 그래서 최신성의 90% 만큼 뉴스 데이터가 기여하는 것으로 가정하였다.

2단계에서는 뉴스 데이터에서 지상파 3사의 비중을 추정하여야 한다. 학습 데이터의 비중이나 학습에 반영된 비율을 적용할 수 있으나, 자료를 구할 수가 없다. 그래서 뉴스 매체의 지위를 대체 지표로 고려하였다. 한국언론진흥재단(2023.12.31.)이 설문조사를 수행하고 발표한 언론사의 뉴스 신뢰도 중에서 지상파 3사가 53.0%로 나타났고, 영향력은 55.3%였다.

본 연구에서도 언론사의 이용도를 AHP 방법으로 조사하였다. 언론사를 지상파 3사, 종편 및 보도전문 PP, 종이신문 발행 신문사, 온라인 전용 신문사 등 4개로 나누었다. 그리고 이들의 중요도를 객관적으로 추정하기 위해 AHP 조사 및 분석을 수행하였다. 분석 결

과 지상파 3사의 영향력이 35.47%임을 확인하였다. 두 결과를 종합하면 언론사 중에서 지상파 3사의 비중이 35.47~55.3%였다.

[표 2] 언론사 유형별 영향력을 조사하는 AHP 문항

| | | | | | | | | | | | | | | | | |
|---|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|
| Q8. | | | | | | | | | | | | | | | | |
| 언론사는 크게 방송사와 신문사로 구분할 수 있습니다. 그 중 방송사는 지상파 3사와 유료방송의 언론사로, 신문사는 종이신문을 발행하는 신문사와 온라인 전용 신문사로 나눌 수 있겠습니다. 귀하께서 뉴스/보도 기사를 이용할 때 이용도나 부여하는 신뢰성이 언론사에 따라 차이가 있을 것 입니다. 귀하께서 언론사의 뉴스/보도에 대한 이용도나 신뢰성 등을 종합하여 영향력을 평가해 주십시오. [행 별 1개 선택] | | | | | | | | | | | | | | | | |

| 평가지표 | 중요 <————— 동일 ————— 중요 | | | | | | | | | | | | | 평가지표 | | | |
|---------------|-----------------------|-------|----|-------|----|-------|----|-------|-------|---|---|---|---|------|---|---|---|
| | 절대 중요 | 매우 중요 | 중요 | 약간 중요 | 동일 | 약간 중요 | 중요 | 매우 중요 | 절대 중요 | | | | | | | | |
| 지상파 방송 3사* | ⑨ | ⑧ | ⑦ | ⑥ | ⑤ | ④ | ③ | ② | ① | ② | ③ | ④ | ⑤ | ⑥ | ⑦ | ⑧ | ⑨ |
| 지상파 방송 3사 | ⑨ | ⑧ | ⑦ | ⑥ | ⑤ | ④ | ③ | ② | ① | ② | ③ | ④ | ⑤ | ⑥ | ⑦ | ⑧ | ⑨ |
| 지상파 방송 3사 | ⑨ | ⑧ | ⑦ | ⑥ | ⑤ | ④ | ③ | ② | ① | ② | ③ | ④ | ⑤ | ⑥ | ⑦ | ⑧ | ⑨ |
| 종편, 보도 전문 방송사 | ⑨ | ⑧ | ⑦ | ⑥ | ⑤ | ④ | ③ | ② | ① | ② | ③ | ④ | ⑤ | ⑥ | ⑦ | ⑧ | ⑨ |
| 종편, 보도 전문 방송사 | ⑨ | ⑧ | ⑦ | ⑥ | ⑤ | ④ | ③ | ② | ① | ② | ③ | ④ | ⑤ | ⑥ | ⑦ | ⑧ | ⑨ |
| 종이신문 발행 신문사 | ⑨ | ⑧ | ⑦ | ⑥ | ⑤ | ④ | ③ | ② | ① | ② | ③ | ④ | ⑤ | ⑥ | ⑦ | ⑧ | ⑨ |

* 지상파 방송 3사(KBS, MBC, SBS)

** 종합편성 4사(itbc, TV조선, 채널A, MBN), 보도전문 채널2사(YTN, 연합뉴스TV)

*** 조선일보, 동아일보, 중앙일보, 경향신문, 매일경제, 문화일보, 서울신문, 한겨레신문, 한국일보 등 전국 신문, 부산일보 등 지역 신문, 스포츠 신문 등 전문지, 주간지

**** 이데일리, 뉴스토마토, 이데일리, 데일리안, 노컷뉴스, 뉴데일리 등

기여율을 구한 후 분석 결과를 국민경제 차원으로 확장하여야 한다. 설문조사의 표본을 우리나라 인구통계 현황을 모사하여 추출하였으므로 확장이 가능하다. 그런데 국내 AI 이용자 자료를 구할 수가 없으므로, AI 이용자 100만 명당 연간 매출 기여액을 구하였다. 그 결과 언어능력에 대한 지상파 3사의 기여도는 이용자 100만 명당 연간 166.1~258.9 억 원으로 나타났다. 그리고 최신성에 대한 지상파 3사의 기여도는 이용자 100만 명당

547.3~853.3억 원으로 나타났다. 두 기능의 개선에 대한 지상파 3사 뉴스 데이터의 기여도는 이용자 100만 명당 총 713.4~1,112.2억 원으로 나타난다. 이 금액은 AI 이용자에게 제공한 편익으로, AI 서비스의 가격에 반영되어 매출에 기여한다.

[표 3] 생성형 AI에 대해 지상파 3사 뉴스 데이터가 창출하는 편익

| 기능 | 단계 | 기여율 | 100만 명당(억/연) |
|------|-----|--------|--------------|
| 언어능력 | 1단계 | 50% | 468.3 |
| | | 55.30% | 258.9 |
| | 2단계 | 35.47% | 166.1 |
| 최신성 | 1단계 | 90% | 1,543.0 |
| | | 55.30% | 853.3 |
| | 2단계 | 35.47% | 547.3 |
| 합계 | 하한 | | 1,112.2 |
| | 상한 | | 713.4 |

2) 원가접근법

원가접근법을 적용하기 위해 두 단계를 거쳐야 한다. 첫째, 지상파 3사의 뉴스 제작 비용을 파악해야 한다. 둘째, 뉴스 제작비용의 어느 비율을 AI 사업자에게 분담시킬지를 결정해야 한다. 방송사 입장에서는 뉴스를 유통하는 모든 매체에 제작비용을 배분해야 하겠지만 아직 합의된 기준이 없다.

지상파 3사는 2024년에 총 4,283억 원을 뉴스 제작에 사용한 것으로 나타난다. 여기에 제작비, 인건비, 일상경비, 간접비가 모두 포함되며, KBS는 지역사의 비용까지 포함하였고, MBC와 SBS는 중앙사의 비용만 포함하였다.

우선 매체별 뉴스 소비량을 기준으로 제작비용을 배분할 수 있겠다. 그러나 소비량이 시청률(방송), 클릭 수(포털), 시청 횟수 및 시청 시간(유튜브), 학습·검색·활용도(AI) 등 다양한 형태로 측정되므로, 종합화하기가 어렵다. 차선책으로 뉴스 콘텐츠의 이용도를 기준으로 삼을 수 있겠다. 즉 수용자가 많이 이용하는 뉴스 매체가 이에 비례하여 비용을 분담하는 것이 합당할 것이다.

AHP 분석을 위해 [그림 1]과 같이 TV/라디오 등 방송, 포털, 유튜브 그리고 생성형 AI로 계층 구조를 만들었다. 여기에 뉴스 이용도를 묻는 문항을 만들어서 설문조사에서 제시하였다. Saaty(1983)가 9점 척도가 실제치에 가장 근접한 결과를 나타냄을 실험으로 확인하였으므로, 본 문항에서도 9점 척도를 이용하였다.

[그림 1] 방송 뉴스 콘텐츠의 이용도에 대한 AHP 계층 구조



[표 4] 방송 뉴스 콘텐츠의 이용도를 묻는 AHP 문항

Q7.

방송사가 만든 뉴스의 보도기사는 1)TV 또는 라디오 방송(지상파·유료방송 포함), 2)온라인 포털(유·무선), 3)유튜브, 4)AI 등을 통해 전파됩니다. 최근 생성형 AI가 뉴스를 검색 및 정리하는 서비스를 제공하면서 새로운 뉴스 매체로 떠오르고 있습니다. 구하께서 평소에 방송사에서 만든 뉴스를 이용하는 습관을 생각해 보시고, 방송 뉴스를 접하는 미디어의 이용도를 답변해 주십시오. 표의 각 행마다 양 끝단에 제시된 두 미디어 중에서 방송사 뉴스의 이용도가 더 높은 쪽에 그 차이를 반영하여 표시해 주세요.

| 평가지표 | 중요 <———— 동일 ————— 중요 | | | | | | | | | | | | | 평가지표 | | | | |
|-----------------|----------------------|-------|----|-------|----|-------|----|-------|-------|---|---|---|---|------|---|---|---|-----------------|
| | 절대 중요 | 매우 중요 | 중요 | 약간 중요 | 동일 | 약간 중요 | 중요 | 매우 중요 | 절대 중요 | | | | | | | | | |
| TV/라디오 방송 | ⑨ | ⑧ | ⑦ | ⑥ | ⑤ | ④ | ③ | ② | ① | ② | ③ | ④ | ⑤ | ⑥ | ⑦ | ⑧ | ⑨ | 온라인 포털 (유무선) |
| TV/라디오 방송 | ⑨ | ⑧ | ⑦ | ⑥ | ⑤ | ④ | ③ | ② | ① | ② | ③ | ④ | ⑤ | ⑥ | ⑦ | ⑧ | ⑨ | 유튜브 |
| TV/라디오 방송 | ⑨ | ⑧ | ⑦ | ⑥ | ⑤ | ④ | ③ | ② | ① | ② | ③ | ④ | ⑤ | ⑥ | ⑦ | ⑧ | ⑨ | 생성형 AI |
| 온라인 포털 (유무선) | ⑨ | ⑧ | ⑦ | ⑥ | ⑤ | ④ | ③ | ② | ① | ② | ③ | ④ | ⑤ | ⑥ | ⑦ | ⑧ | ⑨ | 유튜브 |
| 온라인 포털 (유무선) | ⑨ | ⑧ | ⑦ | ⑥ | ⑤ | ④ | ③ | ② | ① | ② | ③ | ④ | ⑤ | ⑥ | ⑦ | ⑧ | ⑨ | 생성형 AI |
| 유튜브 | ⑨ | ⑧ | ⑦ | ⑥ | ⑤ | ④ | ③ | ② | ① | ② | ③ | ④ | ⑤ | ⑥ | ⑦ | ⑧ | ⑨ | 생성형 AI |

조사 결과 일관성 조건 10% 이하를 만족하는 응답이 총 490개로 나타났고, 이들의 응답으로 구성한 쌍대비교 행렬들의 원소를 기하평균하여 종합화했다. 분석 결과 생성형 AI에서의 뉴스 이용도가 20.5%로 나타났으므로, 이를 지상파 3사의 뉴스 제작비용에 적용하면 저작권료가 연간 877.6억 원에 이른다.

5. 결론 및 시사점

세계적으로 다양한 기능을 갖춘 생성형 AI 서비스의 상업화가 진행되고 있으며, 상당한 규모의 이용자를 확보한 AI가 출현하는 등 AI 산업은 경제적으로도 의미를 확보해 나가고 있다. 생성형 AI 개발 과정에서 학습 데이터가 AI의 성능에 중요한 영향을 미치는 것으로 알려져 있으나, AI 개발자들이 저작권자에 대한 보상 없이 대량으로 학습에 사용하고 있어 갈등이 발생하고 있다. 그런데 최근에는 사업자간 협상을 통한 계약도 이루어지기 시작했다. 우리나라에서도 다수의 생성형 AI가 개발되었고, 데이터에 대한 저작권료 갈등도 진행 중이다. 특히 AI 학습의 효율을 높여주는 고품질 데이터로 인정받는 뉴스 데이터에 대한 저작권료를 두고 지상파 3사와 포털사업자인 네이버 측이 법정 공방을 시작하였다. 국내에는 유사한 판결이나 구축된 판례가 거의 없고, 법적인 판단에는 시간이 소요된다. 그러므로 불투명한 국내 데이터 시장의 환경이 AI 산업에 장애가 되어 경쟁력과 성장잠재력을 잠식하고, 저작권자에게도 손해를 입힐 것으로 우려된다. 그러므로 사업자 간 협상을 통한 계약이 활성화되어 저작권료 갈등을 합리적으로 해결하는 관행을 확립할 필요가 있다.

사업자간 계약에서 사업자의 협상력과 합리적인 기준은 항상 충돌한다. 즉, 사업자의 시장 지배력 등에 기반한 협상력의 차이가 협상 결과에 영향을 미치는 것이 일반적이다. 그런데 암묵적으로 또는 공식적으로 널리 인정받는 합리적인 근거나 기준이 있다면, 협상력의 차이에 따른 왜곡을 줄일 수 있을 것이다. 그러므로 합리적인 저작권료 산정모형이 합당한 데이터 거래를 지원함으로써 저작권자의 권리를 보호하고 데이터 거래를 활성화하여 태동 단계에 있는 국내 AI 산업의 효율성을 높이는데 기여할 것이다.

본고에서는 국내 지상파 3사의 뉴스 데이터를 AI 사업자들이 학습과 서비스에 활용하는 데 대한 대가를 객관적인 방법으로 추정하였다. 이를 위해 법령과 해외 법원의 유사 판례, 업계와 학계 등에서 제시한 기준들과 지금 방법 등 주요 고려 사항들을 종합적으로 검토하였고, 그 결과를 모형에 최대한 적용하여 실무적인 활용도를 높이려고 노력하였다. 우선 AI가 뉴스 데이터를 이용하는 상황을 학습과 운용 단계로 구분하였다. 그리고 뉴스 데이터가 AI에 기여하는 바를 언어능력의 향상과 최신성으로 규정하였다. 그리고 수익접근법과 원가접근법, 1년마다 로열티 형식으로 지불하는 방식을 채택하였다. 그런데 시장에서 축적된 데이터가 없으므로, 설문조사를 통해 진술선호자료를 수집하였다. 수익접근법에서는 뉴스 데이터가 기여한 AI의 이용자 효용 증加分을 매출의 대리 지표로 삼았고, 자료를 서열프로빗 모형으로 분석하였다. 원가접근법에서는 방송사가 뉴스를 제작하는 데 소요되는 비용을 구한 후 AHP 방법으로 구한 매체별 뉴스 이용도에 따라 AI에 배분하였다. 그리하여 수익접근법으로는 지상파 3사의 뉴스 데이터가 AI 이용자 100만 명당 713~1,112억 원의 매출에 기여하는 것으로 확인되었다. 원가접근법으로는 연간 878억 원의 저작권료를 분담시킬 수 있을 것으로 확인되었다. 그런데 수익접근법에서는 매출에 대한 기여도를 추정한 것이므로 여기에서 AI 사업자의 비용을 차감하여야 한다. 그리고 정확한 이용자 수에 대한 자료도 적용하여 최종적인 저작권료 규모를 산정하여야 한다. 그러나 원가접근법에서는 데이터 상품에 대한 가격의 개념이므로 추가적인 절차가 불필요하다.

그런데 학습 데이터로 사용된 비중과 학습에 반영된 비율, 이용자의 수, 투입 비용 등의 자료들을 구할 수가 없었다. 이러한 정보의 비대칭성도 협상에 큰 영향을 미친다. 본고에서는 언론에서 보도된 제한적인 내용과 합리적인 가정을 바탕으로 데이터의 부족을 처리하였다. 그러나 AI 사업자가 가진 자료가 가장 정확할 것이다. 그러므로 앞으로 생성형 AI 시장 환경을 정비하면서 이러한 필수 자료들을 AI 사업자들이 공개하도록 추진할 필요가 있겠다. 본고에서 제안한 저작권료 산정 모델과 객관적인 추정 결과가 AI 산업계에서 활용되어, AI 개발자들과 데이터 권리자들이 공정한 논의를 시작하는 축매제가 되기를 기대한다.

참고문헌

- 강신하(2014.9), 「저작권법」, 제2판, 진원사.
- 과학기술정보통신부 · 한국지능정보사회진흥원(2022.9). AI 학습용 데이터 구축 비용산정 가이드.
- 권구민 · 김현석(2023.2.27). 콘텐츠산업의 생성형 AI 활용 이슈와 대응 과제, 한국콘텐츠진흥원.
- 김선호(2025.1.3). 뉴스 미디어의 AI 환경 대응 좌표를 마련하다.
- 김윤명(2023). 데이터 공정이용, 계간 저작권, 제141호, 한국저작권위원회.
- 김창화(2024.2.29). 생성형 AI를 둘러싼 최근 저작권 분쟁 동향과 시사점, Regulatory Law Review.
- 류시원 (2023.1). 저작권법상 텍스트 · 데이터 마이닝(TDM) 면책규정 도입 방향의 검토, 선진상사법을연구, 통권 제101호.
- 문화체육관광부 · 한국저작권위원회(2023.12.27). 생성형 AI 저작권 안내서
- 박찬(2024.4.7.), AI 학습 데이터 시장 급속 확대 단어 1000개 당 1달러, AI타임스.
- 송영천(2025.4.2.). 지브리 열풍 불더니 챗GPT 기입자 5억명 돌파, 한국경제신문
- 신용우 등(2024.9). 인공지능 윤리와 저작권의 규제체계 연구, 국회입법조사처.
- 양현석 등(2024.12). AI 학습에 이용되는 어문저작물의 적절한 대가에 관한 연구, 한국저작권위원회.
- 이규호 · 남선우(2024). 저작권 및 저작물의 가치평가에 대한 연구, 정보법학, 제28권 제1호, 105~145.
- 이철남(2024.10). AI 저작권 법제도 개선방안 연구, 한국저작권위원회.
- 임대준(2023.7.14.), 오픈AI, 저작권 확보 위해 AP통신 · 셋터스톡 계약, AI TIMES.
- 한국언론진흥재단(2023.12.31). 2023 언론수용자 조사.
- 한국언론진흥재단(2024). 2024 AI시대 뉴스저작권 포럼 종합보고서.
- Chiu, Y. & Chen, Y. (2003). Using AHP on Patent Valuation, ISAHP 2005, Honolulu, Hawaii, July 8–10.
- Linden, T.C.G.(2017). Algorithms for journalism: The future of news work. The Journalism of Media Innovations, 4(1), 60~76.
- Saaty, T.L. (1983). Priority setting in complex problems. IEEE Transactions on Engineering Management, 30(3), 140~155.
- Wang, J.T., Deng, Z., Chiba-Okabe, H. Barak, B. & Su, W.J. (2024). An Economic Solution to Copyright Challenges of Generative AI.