한국언론진흥재단
Korea Press Foundation

KPF 미디어브리프
MediaBrief

2024년 12호

AI로 진화하는 팩트체크: 가능성과 한계

이현우 | 한국언론진흥재단 선임연구위원

- AI 정보 생태계, 알고리즘 역할 축소와 신뢰성 있는 개인화 콘텐츠 부상
- AI 기반 팩트체크는 자연어 처리 기반 기술, 자동 증거 검색 시스템, 이미지 및 비디오 검증 도구, AI-인간 협업 시스템 등 다양한 형태로 진화
- 기술 발전에도 불구, 데이터 의존성 문제, 설명 가능성 부족, 맥락 이해의 한계, 편향성과 오류 가능성 등 다양한 한계 존재
- 저널리즘 분야 활용을 위해서 AI 기술의 발전과 더불어 인간과 AI의 협업이 필수적

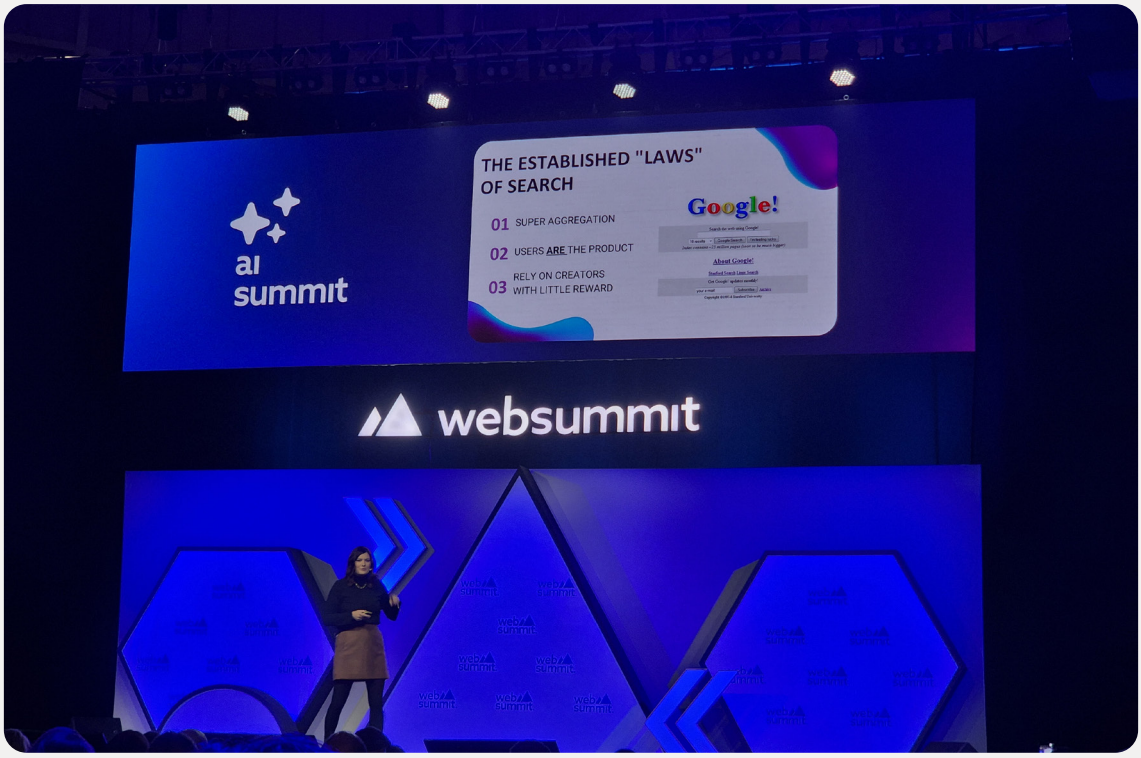


AI 시대가 도래하면서 정보의 생성과 확산 방식에 급격한 변화가 나타나고 있다. 기존에는 대규모 알고리즘이 정보를 주도적으로 제공하며 전반적인 정보 흐름을 통제했지만, 이제는 신뢰성 있는 맞춤형 콘텐츠가 점점 더 중요한 위치를 차지하고 있다. 이러한 상황을 AI 기술기업 CEO 브룩 하틀리(Brooke Hartley)는 다음과 같이 설명하고 있다.

"AI 정보 생태계에서는 알고리즘의 역할이 축소되고 새로운 방식의 정보 제공이 가능해지고 있습니다. 기존에는 검색 엔진과 같은 시스템이 광고 모델에 기반하여 정보를 제공했다면, 앞으로는 알고리즘 의존에서 벗어나 복잡하고 전문적인 질문에 더욱 효율적으로 답할 수 있는 AI 시스템의 가능성이 열리고 있습니다. 앞으로 AI는 개인화된 상호작용을 통해 정보를 제공하게 될 것입니다.

대형 언어 모델은 인간의 사고를 모방하지만, 기본적으로 패턴 매칭 기계일 뿐입니다. 이러한 모델은 고품질의 데이터 없이는 제대로 작동할 수 없으며, 이를 제공하는 콘텐츠 제작자가 핵심 역할을 하게 됩니다. AI가 신뢰받는 기술이 되기 위해서는 정확하고 사실에 기반한 콘텐츠를 생산하는 사람들이 중요하며, 이들을 지원하고 보상하는 구조가 필수적입니다[1]."

그림 웹서밋 2024 'AI Summit' 세션



출처: 한국언론진흥재단. 2024년 11월 14일 웹서밋 AI 세션에서 브룩 하틀리 CEO가 연구 결과를 발표하고 있다.

이처럼 새로운 AI 정보 생태계 환경이 도래하면서 신뢰할 수 있는 콘텐츠와 콘텐츠 검증 기술이 필수 요소로 부상하고 있다. 특히, 허위정보가 사회적 신뢰와 의사결정에 부정적인 영향을 미치는 상황에서 팩트체크는 정보 신뢰도를 높이는 핵심 역할을 담당하고 있다.

AI 기반 팩트체크 기술은 텍스트, 이미지, 비디오 등 다양한 데이터 유형에서 정보를 분석하고 검증할 수 있는 기능을 제공하며, 신속하고 정확한 사실 검증을 가능하게 한다.

본문에서는 자연어 처리 기반 기술, 자동 증거 검색 시스템, 이미지 및 비디오 검증 도구, AI–인간 협업 시스템 등 AI 팩트체크 기술의 주요 유형과 실제 사례를 살펴볼 것이다. 또한, 이러한 기술들이 지닌 한계도 함께 논의하며 AI와 인간이 협력하여 신뢰할 수 있는 정보 생태계를 구축하는 방안을 모색할 것이다.

1) Brooke Hartley, <The dawn of the fact-checking engine>, Websummit 2024 발표 내용

■ AI를 활용한 팩트체크 기술 진화

AI 기반 팩트체크 기술은 여러 유형으로 구분할 수 있다. 각 유형은 텍스트, 이미지, 비디오와 같은 다양한 데이터 유형을 처리하고, 주어진 주장의 진위 여부를 검증하기 위해 특정 기술을 사용한다.

▲ 자연어 처리(NLP)/머신러닝 기반 팩트체크

자연어 처리(NLP) 기반 팩트체크는 텍스트 주장을 분석하고 이를 검증하기 위한 증거를 검색하는 데 사용된다. 주요 기능으로는 주장 감지, 증거 추출, 그리고 증거 기반 판단(증거가 주장을 지지하거나 반박하는지 판단)이 있다.

- ClaimBuster : 정치적 토론이나 연설 중 실시간으로 사실 검증이 필요한 주장을 탐지하는 AI 기반 도구이다. 텍사스 알링턴 대학교 IDIR(Innovative Data Intelligence Research Laboratory)에서 개발한 이 시스템은 머신러닝 알고리즘을 활용하여 문장에서 사실 확인이 필요한 주장을 식별한다.

  ClaimBuster는 수집된 텍스트 데이터를 분석하여 각 문장이 사실 검증의 필요성이 있는지 점수화한다. 이러한 점수화 과정을 통해 우선적으로 검토해야 할 주장을 식별하고, 팩트체커들이 효율적으로 작업할 수 있도록 지원한다. 이 도구는 특히 정치적 토론이나 연설과 같은 실시간 이벤트에서 신속한 사실 검증이 가능하도록 설계되었다.

  ClaimBuster는 API를 통해 외부 개발자나 조직이 자체 시스템에 통합하여 사용할 수 있으며, 이를 통해 다양한 플랫폼에서 자동화된 사실 검증 기능을 구현할 수 있다. 또한, 트위터와 같은 소셜미디어에서 사실 검증이 필요한 주장을 탐지하는 기능도 제공한다[2].

- FEVER(Fact Extraction and VERification) : 대규모 사실 추출 및 검증을 위한 데이터셋으로, 위키백과에서 추출한 문장을 변형하여 생성된 185,445개의 주장으로 구성되어 있다. 이러한 주장은 '지원됨(Supported)', '반박됨(Refuted)', '정보 부족(NotEnoughInfo)' 세 가지 범주로 분류되며, 전자 두 범주의 경우 판단에 필요한 증거 문장도 함께 기록되어 있다[3].

  FEVER는 자연어 처리 분야에서 사실 검증 모델의 개발과 평가를 위한 벤치마크로 활용되며, 연구자들은 이 데이터셋을 통해 텍스트 소스에 대한 주장 검증 시스템의 성능을 향상시키고 있다[4].

2) https://idir.uta.edu/claimbuster/
3) https://aclanthology.org/N18-1074/
4) https://paperswithcode.com/dataset/

▲ 자동 증거 검색 시스템

데이터베이스나 온라인 소스에서 주장을 뒷받침하거나 반박할 수 있는 관련 증거를 자동으로 검색한다.

- Factmata : 데이터베이스에서 증거를 찾는 데 머신러닝과 의미론적 분석을 활용하는 시스템이다. 온라인상의 유해하거나 신뢰할 수 없는 콘텐츠를 자동으로 감지하고 분류하기 위해 자연어 처리(NLP) 기술을 활용하는 AI 기반 플랫폼이다. 이 플랫폼은 24개의 전문 커뮤니티와 2,000명 이상의 전문가(기자와 연구자 포함)로 구성된 네트워크를 통해 지속적으로 학습 데이터를 생성하고 알고리즘의 정확성을 향상시킨다.
  AI는 검사하는 모든 콘텐츠에 신뢰 등급을 지정하여 사기성 콘텐츠에 대한 분류 시스템 역할을 한다. 신뢰 등급은 유해함(toxicity), 정치적 편향(political bias), 클릭베이트(clickbait) 등 9가지로 구분되어 품질, 안전성 및 신뢰성에 대한 정보를 제공한다. 현재 이 기술은 개발 초기 단계에 있다[5].

- Google Fact Check Explorer : 전 세계의 팩트체크 결과를 한곳에서 검색하고 확인할 수 있는 도구로, 언론인, 연구자, 일반 사용자들이 사실과 허위정보를 구분하는 데 도움을 준다. 이 도구는 독립적인 팩트체크 기관들이 조사한 내용을 모아 제공하며, 사용자는 특정 주제나 인물에 대한 팩트체크를 손쉽게 검색할 수 있다.
  최근에는 이미지 검색 기능이 추가되어, 사용자가 이미지를 업로드하거나 링크를 입력하면 해당 이미지에 대한 팩트체크 결과를 확인할 수 있게 되었다. 이는 이미지 기반의 허위정보를 식별하는 데 큰 도움이 된다. 또한, Fact Check Explorer는 ClaimReview 마크업을 활용하여 팩트체크된 콘텐츠를 식별하고 표시한다. 이를 통해 사용자는 신뢰할 수 있는 팩트체크 정보를 더욱 쉽게 접할 수 있다[6].

▲ 구조화된 데이터베이스(지식 그래프)

지식 그래프 및 의미론적 매칭 기술은 구조화된 데이터베이스에 저장된 사실과 주장을 비교한다. 주로 금융, 의료, 통계와 같은 고정된 정보를 검증하는 데 유용하다. 최근에는 검색증강생성(RAG) 모델과도 결합되어 다양한 분야로 확장되는 양상을 보인다.

- IBM Watson : 금융 및 의료 분야에서 지식 그래프를 활용하여 사실 검증을 수행한다. 의료 분야에서 Watson for Oncology는 자연어 처리와 기계 학습을 활용하여 최신 의학 증거를 기반으로 치료 권장사항을 제공한다. 이 시스템은 환자의 의료 기록, 최신 연구 결과 등을 분석하여 개인화된 치료 옵션을 제시하며, 종양학자들이 최적의 치료 결정을 내리는 데 도움을 준다[7].
  이러한 지식 그래프 기반 사실 검증 시스템은 구조화된 데이터베이스를 활용하여 정확한 정보를 제공하지만, 구축과 유지에 상당한 자원과 노력이 필요하다. 또한, 지식 그래프의 품질과 범위는 시스템의 성능에 직접적인 영향을 미치므로, 지속적인 업데이트와 관리가 필수적이다[8].

5) https://www.nesta.org.uk/feature/ai-and-collective-intelligence-case-studies/factmata/
6) https://newsinitiative.withgoogle.com/ko-kr/resources/trainings/google-fact-check-tools/
7) https://link.springer.com/article/10.1007/s00146-020-00945-9
8) https://research.ibm.com/blog/knowledge-graph-ai

▲ 이미지 및 비디오 검증 시스템

이미지 및 비디오 검증 시스템은 시각적 콘텐츠에서 변조나 메타데이터 이상 여부를 분석한다.

- InVID : 소셜 미디어에서 확산되는 뉴스 영상의 신뢰성을 평가하고 검증하기 위해 개발된 도구로, 특히 기자와 연구자들이 사용자 생성 콘텐츠의 진위 여부를 판단하는 데 도움을 준다.
  먼저, YouTube, Facebook, Twitter 등에서 동영상을 프레임 단위로 분할하여 키프레임을 추출하고, 이를 분석하거나 역이미지 검색에 활용할 수 있다. 역이미지 검색은 추출된 키프레임을 대상으로 Google, Baidu, Yandex 등 검색 엔진을 통해 이미지의 원 출처나 유사한 콘텐츠를 탐색할 수 있는 기능이다.
  또한, 이미지나 동영상 파일의 메타데이터를 분석하여 촬영 위치, 시간, 사용된 장치 정보 등을 확인함으로써 콘텐츠의 진위를 판단하는 데 도움을 준다[9].

- 딥페이크 탐지 AI : 영상에서 얼굴 움직임이나 조명 불일치와 같은 미묘한 차이를 분석하여 조작된 콘텐츠를 식별하는 기술이다. 이러한 시스템은 얼굴의 자연스러운 움직임과 조명 패턴을 학습하여, 실제 영상과 합성 영상 간의 차이를 감지한다.
  딥페이크 영상에서는 자연스럽지 않은 눈 깜빡임 패턴이나 감정 표현의 불일치가 자주 발견된다. 이러한 눈 깜빡임의 불규칙성을 딥페이크 식별의 주요 단서로 활용할 수 있다. 또한, 조명 반사와 그림자에서도 불일치가 나타날 수 있다. 특히, 눈동자에 반사되는 빛의 패턴이 일관되지 않거나 부자연스러운 경우, 합성 영상임을 판별하는 중요한 근거가 된다[10].
  딥페이크 얼굴은 감정 표현에서도 부자연스러움을 보일 수 있다. 예를 들어, 입은 웃고 있지만 눈은 감정을 반영하지 않는 경우가 이에 해당한다. 이러한 미묘한 불일치는 딥페이크 탐지의 유용한 단서로 활용된다[11].

▲ AI-인간 팩트체커 협업 시스템

혼합형 시스템은 AI의 자동화 기능과 인간의 검증 작업을 결합한다. AI는 반복적인 작업(예: 증거 검색)을 처리하고, 인간은 복잡한 주장을 분석한다.

- FactStream : 자동화된 주장 감지와 인간 편집 과정을 통합한 도구로, 정치 토론 등의 실시간 팩트체크에 활용된다. FactStream은 듀크대학교의 리포터스 랩(Reporters' Lab)에서 개발한 실시간 팩트체크 애플리케이션으로, 주요 정치 이벤트 동안 신뢰할 수 있는 팩트체킹 기관의 검증된 정보를 사용자에게 즉시 제공하는 것을 목표로 한다.
  FactStream은 PolitiFact, FactCheck.org, The Washington Post와 협력하여 최신 팩트체크를 통합하여 제공하며, 사용자는 일일 스트림을 통해 최신 팩트체크를 확인하고, 특정 주제나 인물에 대한 검색 기능을 활

9) https://www.invid-project.eu/tools-and-services/invid-verification-plugin/
10) https://en.wikipedia.org/wiki/Deepfake
11) https://link.springer.com/chapter/10.1007/978-3-031-43153-1_24

용할 수 있다. 또한, 중요한 팩트체크가 게시되면 알림을 통해 사용자에게 즉시 전달한다[12].

- Logically : AI와 인간 전문성을 결합해 허위정보와 잘못된 정보를 해결하는 기술 회사이다. Logically의 플랫폼은 AI를 활용해 주장들을 필터링하고, 기존 검증된 데이터베이스와 비교해 신뢰 점수를 부여한다. 이후 인간 팩트체커들이 해당 주장을 평가하며 최종적으로 정확성을 판단한다.
  Logically Intelligence(LI) 서비스는 수천 개의 웹사이트와 소셜미디어 데이터를 수집해 위험 콘텐츠를 식별하고 이를 유형별 그룹으로 정리한다. 이는 정부 및 비정부기구(NGO)가 온라인 허위정보를 식별하고 대응하도록 돕는 것을 목표로 한다[13].

- Debunk.org : AI 알고리즘을 통해 온라인 뉴스 기사를 자동으로 스캔하고 허위정보를 식별하며 대응하는 조직이다. 이 플랫폼은 26개 언어로 2,500개 이상의 웹 도메인을 모니터링하며, 매일 약 30,000개의 기사를 처리한다.
  AI 알고리즘은 키워드와 600개 이상의 프로파간다 및 허위정보 내러티브를 검색하며, '엘프'로 불리는 자원봉사자 팀이 위험성 플래그가 지정된 콘텐츠를 수동으로 검토해 위협 수준을 결정한다. 매달 15,000개 이상의 콘텐츠가 수동으로 검토된다. 이러한 AI와 인간 전문성의 결합은 허위정보를 신속하게 식별하고 대응할 수 있게 한다[14].

■ AI 기반 팩트체크의 한계 : 인간과 AI의 협업 필요

AI 기반 팩트체크 기술은 지속적으로 발전하고 있지만 여전히 여러 한계를 가지고 있다. 이러한 한계는 데이터 품질, 기술적 제약, 그리고 인간의 판단력을 대체할 수 없는 본질적인 문제에서 비롯된다[15].

▲ 데이터 의존성 문제

AI 기반 팩트체크 시스템은 대규모 데이터셋에 의존해 학습하지만, 대부분의 데이터셋이 제한적인 크기이거나 특정 분야(도메인)로 국한된다. 이로 인해 시스템이 새로운 유형의 정보나 다른 문화적 맥락을 이해하는 데 어려움을 겪는다. 특히 한국어와 같은 소수 언어나 자원이 불충분한 환경에서는 이러한 데이터 부족이 더욱 두드러진다.

12) https://www.factstream.co/

13) https://en.wikipedia.org/wiki/Logically_%28company%29

14) https://en.wikipedia.org/wiki/Debunk.org

15) Hanselowski, A., Stab, C., Schulz, C., Li, Z., & Gurevych, I. (2019). A Richly Annotated Corpus for Different Tasks in Automated Fact-Checking. https://consensus.app/papers/richly-annotated-corpus-different-tasks-automated-hanselowski/05101dadcaa15ab88690c73fb992f277/
Vallayil, M., Nand, P., Yan, W. Q., & Allende-Cid, H. (2023). Explainability of Automated Fact Verification Systems: A Comprehensive Review. https://consensus.app/papers/explainability-automated-fact-verification-systems-vallayil/da43ae4884bf5dbdb9eedbf9f3a5746c/
Hrckova, A., Móro, R., Srba, I., Simko, J., & Bieliková, M. (2022). Automated, not Automatic: Needs and Practices in European Fact-checking Organizations as a basis for Designing Human-centered AI Systems. https://consensus.app/papers/automated-automatic-needs-practices-european-hrckova/12bd5d9f63c85c8d89c8b6445ce22ccf/

▲ 설명 가능성 부족

AI 시스템의 판단 근거가 충분히 투명하지 않기 때문에, 사용자들이 결과를 신뢰하기 어렵다. 특히 자동화된 팩트체크 시스템이 제공하는 결과가 올바르더라도, 그 과정을 납득하지 못하면 실제 활용도는 낮아진다. 이를 해결하기 위해 설명 가능한 AI(XAI)의 필요성이 대두되고 있지만, 현재의 시스템은 여전히 신뢰성이 부족하다.

▲ 맥락 이해의 한계

AI는 주장의 맥락적 의미를 이해하거나 미묘한 뉘앙스를 파악하는 데 한계가 있다. 예를 들어, 문화적 차이, 은유적 표현, 그리고 특정 도메인 지식이 요구되는 경우 AI는 부정확한 결과를 도출하는 경우가 많다. 이러한 한계는 특히 정교한 언어적 해석이 필요한 정치적 발언이나 복잡한 과학적 주장에서 두드러진다.

▲ 편향성과 오류 가능성

AI는 학습 데이터의 편향을 그대로 반영할 수 있으며, 잘못된 정보를 학습할 경우 이를 사실로 간주해 잘못된 결과를 생성할 수 있다. 심지어 잘못된 정보를 사실처럼 보이게 만들 수 있는 '환각(hallucination)' 문제도 존재한다. 이는 특히 높은 신뢰성을 요구하는 저널리즘 분야에서 치명적일 수 있다.

AI 기술의 발전에도 불구하고 팩트체킹 과정에 대한 인간의 개입 필요성은 여전히 크다. 완전히 자동화된 팩트체크 시스템은 현실적으로 불가능하며, 인간 검증자가 AI의 결과를 분석하고 최종 결정을 내리는 데 중요한 역할을 한다. 저널리즘 분야에서 AI 단독으로 팩트체크를 수행하는 것은 기술적, 윤리적으로 바람직하지 않다.

결론적으로, AI 기반 팩트체크 기술은 정보 검증의 속도와 효율성을 높이는 데 기여하고 있지만, 데이터 품질, 맥락 이해, 편향성 문제 등 여러 한계를 극복해야 한다. 이러한 문제를 해결하기 위해서는 AI 기술의 발전뿐만 아니라 인간과 AI의 협업이 필수적이다.

KPF 2024년 12호
미디어브리프

발행인 김효재
편집인 남정호
기　획 한국언론진흥재단 미디어연구센터
발행일 2024년 12월 13일

한국언론진흥재단 미디어연구센터
04520 서울특별시 중구 세종대로 124 프레스센터빌딩 13층
전화 (02) 2001-7755 팩스 (02) 2001-7740
www.kpf.or.kr

편집 (주)나눔커뮤니케이션
04034 서울특별시 마포구 잔다리로7길 16 교평빌딩 304호
전화 (02) 333-7136 팩스 (02) 333-7146

ⓒ한국언론진흥재단 미디어연구센터 2024

〈KPF미디어브리프〉는 한국언론진흥재단의 정부광고 수수료로 제작되었습니다.
자세한 내용은 QR코드를 통해 확인해 주세요.