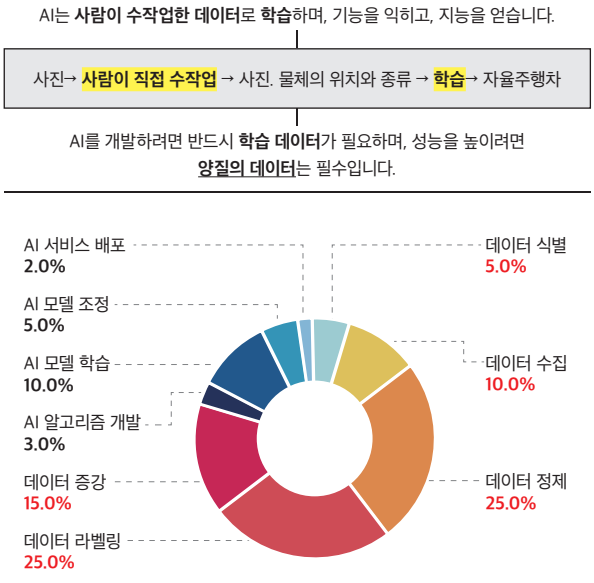
#2 AI 개발, 학습데이터의 중요성

글. **김세엽** 셀렉트스타 대표

업종, 규모, 기술력에 상관없이 AI 산업을 관통하는 단 하나의 키워드는 '학습데이터의 중요성'이다. 2021년 3월 Deeplearning.AI의 앤드류 응(Andrew Ng)교수가 학습데이터의 중요성을 강조한 이래, 데이터 중심의 인공지능(Data-Centric AI) 개발은 업계 표준으로 자리 잡았다. AI 성능을 높이는 데는 알고리즘과 코드 개선보다, 양질의 학습데이터를 구하는 작업이 훨씬 효과적이라는 맥락이다. 본 기고문에서는 AI 학습데이터 구축의 핵심 기술과 사례를 소개한다.

그림 1. 전체 프로젝트 기간의 약 80%가 학습데이터 구축 (데이터 수집,정제,가공 등)에 활용



크라우드 소싱 플랫폼의 장점

데이터 구축 산업은 본질적으로 매우 노동집약적이다. 수십만 건의 원천데이터를 수집하고 가공하는 과정에서 상당한 노동력이 투입되기 때문이다. 셀렉트스타는 이 점에 착안해 크라우드 소싱 플랫폼 캐시미션을 런칭했다. 셀렉트스타가 고객사로부터 수주한 데이터 생산 프로젝트를 캐시미션 플랫폼에 등록하면, 크라우드 작업자는 원하는 시간만큼 데이터 가공 업무에 참여하고 적절한 보상을 받아 갈 수 있다.

기업 입장에서는 작업자를 상시 고용 및 유지하지 않고 데이터 구축 업무 전반을 용역으로 할 수 있기에 많은 비용과 시간을 절감할 수 있다. 인공지능 학습데이터 시장은 이러한 수요를 바탕으로 빠르게 성장하고 있다. 특허청 자료에 따르면 인공지능 학습데이터 시장 규모는 22년 기준 약 2억 1천만 달러로, 연평균 성장률은 23.14%로 예상된다(YV Intelligence Report '21).

셀렉트스타는 2018년 설립 이후 200개 이상 기업과 협업해 1억 3천만 건이 넘는 데이터를 생산했으며, 2022년 10월 기준 크라우드 소싱 플랫폼 캐시미션에는 약 24.5만 명의 작업자가 등록돼 있다.

데이터 품질 관리 노하우 및 솔루션

데이터 구축 서비스의 경쟁력은 납품 데이터의 품질에서 온다. 불특정 다수의 노동력을 활용하는 크라우드 소싱 과정에서 일관된 작업 수준을 유지하기 어렵기 때문이다.

셀렉트스타는 데이터의 일관성을 확보하기 위해 회사 설립

그림 2. 크라우드 소싱 플랫폼 '캐시미션' 알고리즘

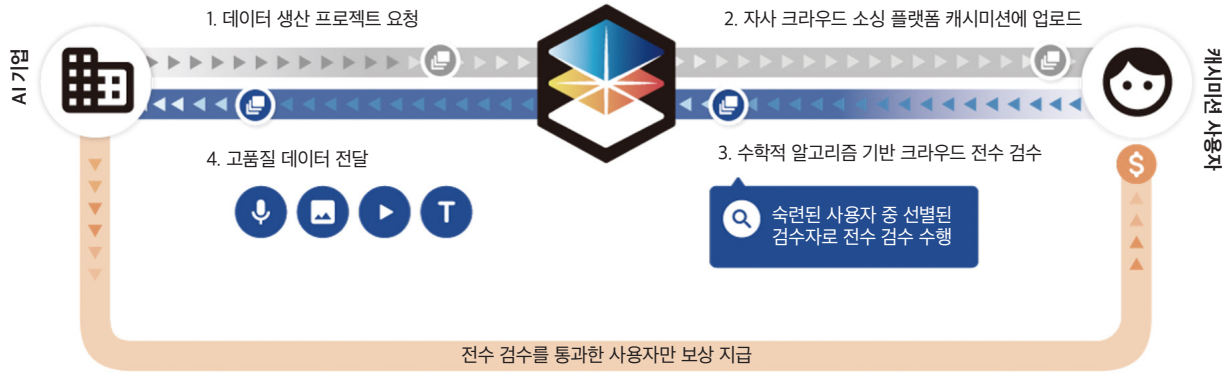
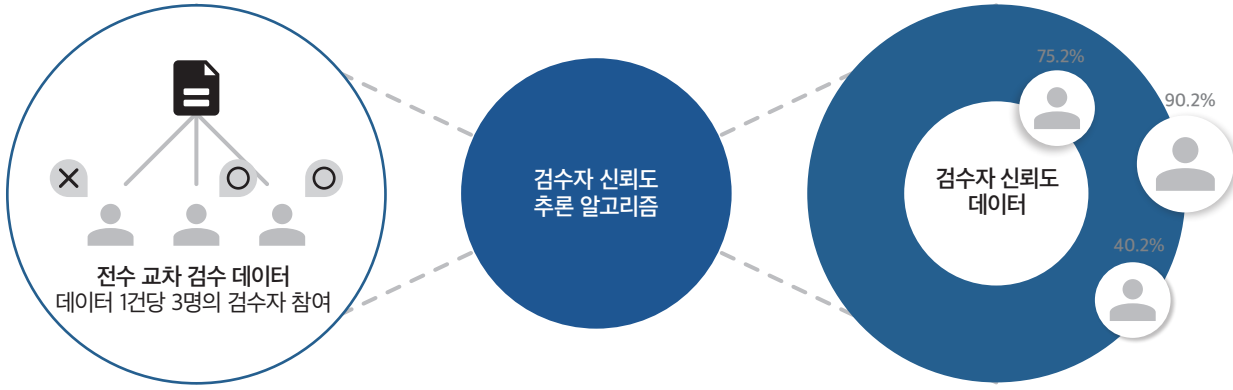


그림 3. 작업자 신뢰도 정보 반영 수학적 알고리즘 (특허 제 10-2333644호)



초기부터 국내 유일 전문 가이드팀을 운영하고 있다. 교육 자료, 작업 상세 기준, 작업 참여를 위한 테스트 문항 등을 문서화하는 조직이다. 크라우드 워커는 작업 가이드라인 숙지를 확인하는 테스트를 통과해야만 작업에 참여할 수 있으며, 작업 중에는 정답이 정해진 '함정 문제'를 풀며 어뷰징 여부를 검증받는다. 이 밖에도 셀렉트스타는 유저별 작업 현황 실시간 모니터링, 우수 작업자 집단 선별 및 관리 등 데이터 품질에 기여할 수 있는 모든 작업을 분업하며 관련 노하우를 축적하고 있다.

한편 데이터를 수집하는 과정에서는 고객사의 AI 개발 방향성이 수정되거나 필요한 데이터의 특성이 바뀌는 일이 비일비재하다. 이러한 과정에서 일반 대중이 아닌 특정 분야 전문성을 갖춘 인력을 필요로 하거나, 굉장히 높은 수준의 정보 보안을 요구하는 프로젝트도 곧잘 발생한다. 데이터 구축 서비스 기업은 이를 위한 전문 인력 섭외 및 고용, 작업 인력 인하우스 파견 및 관리 역량 또한 갖추고 있어야 한다. 이밖에도 프로젝트 착수부터 최종 데이터 납품까지 모델 성능에 영향을 줄 수 있는 희소한 엣지케이스를 고려하고, 데이터 구축 솔루션을 제안하는 컨설팅 업무 또한 데이터 구축 서비스 플랫폼의 역할이다.

관련 기술 특허

셀렉트스타는 데이터 품질 관리를 위한 10건의 기술 특허를 보유하고 있다. 특허 목록에는 반자동 레이블링, 수학적 알고리즘 등 기술적인 솔루션뿐 아니라, 플랫폼 운영 노하우로 개발한 유저 작업 인터페이스도 포함돼 있다.

그림 4. 반자동 Bounding BOX 라벨링 특허

특허 제 10-2176458호

데이터 라벨링을 위한 바운딩 박스 그리기 방법 및 장치

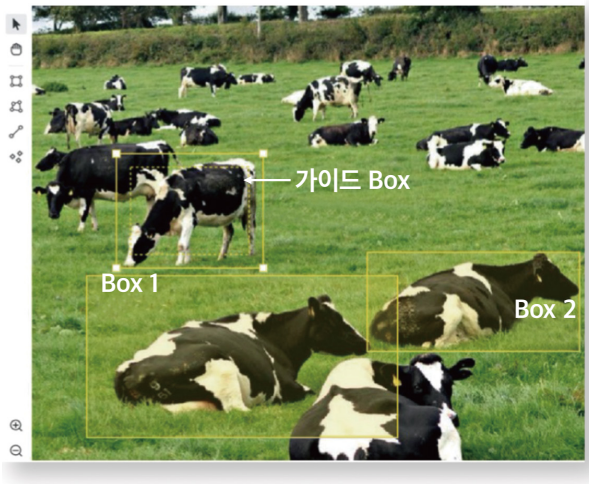


그림 5. 크라우드 작업자 교육 자료. 혐오 표현 지정

무슨 화근을 불러올지 아무도 모른다.

[아프리카 남성, 난민 승소 확정]

카테고리 : 집단_민족_인종_그 외

→ 혐오 대상(아프리카 남성)이 문장에 명시되지 않았지만 추측할 수 있습니다. 구분선에 라벨을 그리고 카테고리를 선택합니다.

작업자 신뢰도 정보 반영 수학적 알고리즘, 특허 제10-2333644호

본 특허는 크라우드 작업자의 작업 능력을 수치화하여, 이를 검수 결과에 반영하는 방법이다. 예를 들어 하나의 데이터에 대해 검사자 2명이 '통과', 1명이 '불통'을 제출할 경우, 단순 다수결이 아니라 작업자의 신뢰도 데이터를 의사 결정에 반영하는 알고리즘이다. 이때 크라우드 작업자는 작업 신뢰도를 측정하는 별도의 작업을 수행할 필요가 없다. KAIST 수리과학 박사진이 개발한 수학적 알고리즘을 통해, 일정 분량 이상 작업을 제출하는 시점에 비교 검사로 작업 수행 능력을 자동 예측한다. 검수자 신뢰도를 추론하는 수학적 알고리즘은 글로벌 Top 경쟁사에서도 찾아볼 수 없는 셀렉트스타만의 원천 기술이다. 셀렉트스타는 데이터납품 전 해당 알고리즘을 활용한 최종 검수로 데이터 퀄리티를 보증한다.

반자동 Bounding BOX 라벨링, 특허 제10-2176458호

본 특허는 이미지 내 물체 주변 Bounding BOX를 그리는 유저 인터페이스(UI)에 관한 내용이다. 통상 라벨링 작업에선 라벨 대상과 Bounding BOX 사이의 간격이 좁을수록 양질의 데이터가 생성된다. 하지만 크라우드 작업자마다 Bounding BOX와 객체 경계 사이의 간격 기준이 다를 수 있다. 예를 들어 객체 '소'를 대상으로 라벨을 그릴 때, BOX 1에서는 소와 박스 사이의 간격이 널널한 반면 BOX 2에서는 그 간격이 좁은 문제가 발생한다.

셀렉트스타에서는 라벨 작업자마다 발생하는 작업 편차를 줄이기 위해, 설정된 값에 따라 반자동으로 보조 Bounding BOX를 생성해 주는 특허를 보유하고 있다. 이제 작업자는 객체 경계를 두 Bounding BOX 사이에 위치함으로써, 보다 정확하고 일관된 라벨링 작업이 가능하다. 셀렉트스타는 데이터 가공/검수에도 동일한 기준을 적용해 일관된 Bounding BOX 데이터를 납품한다.

한국어 혐오 표현 데이터 셋 KOLD

셀렉트스타의 데이터 구축 역량을 보여주는 사례로 KOLD(Korea Offensive Language Dataset)을 소개한다. KOLD는 KAIST 연구진과 셀렉트스타가 협업하여 구축한 한국어 혐오 표현 데이터 셋으로, 네이버 뉴스와 유튜브 플랫폼 등에서 수집한 4만 429개의 혐오 댓글과 그에 대한 주석으로 구성돼 있다. 수천 명의 크라우드 작업자들이 제공된 댓글에서 혐오 키워드를 발췌하고, 일정한 기준에 따라 인종 젠더 지역 등으로 혐오 대상을 분류했다.

데이터 구축의 핵심은 '혐오에 대한 일관된 기준을 적용하기 어렵다'였다. 당연한 얘기지만 어떠한 표현이 혐오 표현인지는 수학적으로 정의 내리기 어렵다. 혐오에 대한 정의는 일정한 사회적 가치를 공유하는 집단 속에서 공시적으로 결정되기 때문이다. 따라서 혐오 표현 분류 작업에는 소수의 숙련자보다 크라

그림 6. 크라우드 작업자 교육 자료. 혐오 대상 카테고리 분류

대분류	소분류	대분류	소분류
종교	이슬람교	기타	나이
	기독교		신체적 특성
	천주교		장애
	불교		질환
	그외		사회 경제적 조건
정치	진보		그외
	보수		
	그외		

우드 소싱으로 구성된 작업자 집단이 효과적일 수 있다.

크라우드 작업자들에게 주어진 문장 속 혐오 표현에 해당하는 부분에 대한 '라벨링 작업'과 혐오 대상에 대한 '분류 작업'을 함께 요청했다. 수집한 데이터 전량에 대해 최소 3회 교차 검수를 진행하면서, 혐오에 대한 공중(公衆)의 주관을 일관되게 반영할 수 있었다. 작업자가 욕설과 혐오에 노출될 수 있기에 만 19세 이상의 성인만 모집했고, 연구기관 요청에 따라 모든 작업자로부터 '정보보안서약서'에 동의받았다.

KOLD는 대부분 영어로 구성된 자연어 처리(NLP, Natural Language Processing) 분야 학습데이터 시장에서, 소수 언어인 한국어 혐오 표현을 수집 · 가공한 양질의 정제 데이터 셋이다. KAIST 연구진에 따르면, 연구 결과 한국어 혐오 표현의 대상 그룹 분포가 기존 영어 데이터 세트와 크게 다르다는 사실이 발견됐다고 한다. KOLD를 한국어 처리 모델 학습에 활용하면 인공지능 번역, 독해, 챗봇 성능을 효과적으로 높일 수 있다.

KOLD 데이터 셋과 관련 논문은 오는 12월 아랍에미리트 아부다비에서 자연어 처리(NLP, Natural Language Processing) 분야 세계 최고 권위 국제 학술대회 'EMNLP'에서 공식 발표될 예정이다.

맺음말

절대다수의 기업들이 시시각각 변하는 AI 개발 방향성을 학습데이터에 반영하는 데 어려움을 겪는다. 셀렉트스타는 축적한 기술력과 노하우로, 데이터 산업 현장 일선의 문제를 해결하기 위한 데이터 플랫폼 서비스(SaaS)를 준비 중이다. AI 생애 주기(ML Lifecycle)에 걸쳐 양질의 학습데이터를 제공할 수 있다.

저자소개

김세엽 셀렉트스타 대표는 KAIST 전기전자공학부 재학 중 인공지능 학습데이터 구축 분야 창업을 결심했다. 2018년 KAIST 창업경진대회 E5에서 우승을 차지하며 동반 창업자들과 함께 크라우드 소싱 플랫폼 '캐시미션'을 개발했다. 2021년에는 美 포브스가 선정한 '2021년 아시아 30세 이하 리더'에 선정됐다. 셀렉트스타는 2022년 8월 기준 시리즈A 익스텐션 라운드를 종료하며 누적 투자액 134억 원을 달성했다.