By: Nir Kshetri

Kshetri, Nir (2021). "Data labeling for the artificial intelligence industry: Economic impacts in developing countries" IEEE IT Professional, 23(2), 96-99. https://doi.org/10.1109/MITP.2020.2967905

Abstract:

Artificial intelligence (AI) applications, although at a nascent stage of development, are already having considerable economic and social impacts in developing countries.1 An equally fascinating and important aspect of the development of the global AI industry is that a high proportion of jobs that require relatively low skills in this industry are being performed by the developing countries. The multibillion-dollar data labeling industry represents an interesting example to illustrate this trend.

Keywords: Data models | Labeling | Artificial intelligence | Developing countries

Article:

Artificial intelligence (AI) applications, although at a nascent stage of development, are already having considerable economic and social impacts in developing countries.[1] An equally fascinating and important aspect of the development of the global AI industry is that a high proportion of jobs that require relatively low skills in this industry are being performed by the developing countries. The multibillion-dollar data labeling industry represents an interesting example to illustrate this trend.

According to the professional services firm PwC, AI's contribution to the global economy will reach about US$16 trillion by 2030 (https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html). Accurately labeling data for training machine learning (ML) models is integral to the creation of this value. The global market for data labeling, also known as content labeling or data annotation, is expected to reach US$5 billion by 2023.[2]

Data labeling activities, however, are extremely time and labor intensive. Developing countries are in a position to take advantage of the opportunities provided by the global AI industry. This phenomenon has thus created a whole new industry of data labeling in developing countries, which is described as “a new type of blue-collar industry” (https://www.forbes.com/sites/korihale/2019/05/28/google-microsoft-banking-on-africas-ai-labeling-workforce/#34b0fd0d541c). Data labelers are referred to as the blue-collar workers of the AI age (https://towardsdatascience.com/the-invisible-workers-of-the-ai-era-c83735481ba)**.**

DATA LABELING IN DEVELOPING COUNTRIES POWERING THE GLOBAL AI INDUSTRY

Data need to be cleaned, categorized into appropriate groups, and labeled so that AI algorithms can recognize patterns.[3] For instance, for ML algorithms to accurately recognize a car, the algorithms may need to be trained with a large number of car pictures (https://www.vice.com/en_us/article/7xyabb/china-ai-dominance-relies-on-young-data-labelers). In the most common form of AI–the supervised learning—the algorithms need to be fed with millions of pretagged examples of car pictures until they correctly identify the pictures.[2] These activities need a large amount of human labor. For instance, one hour of video data related to autonomous driving may need as much as 800 man-hours of data labeling work (https://www.axios.com/ai-data-labeling-billion-dollar-market-409704bc-e63c-4af0-b0d0-44424abcd561.html). Estimates suggest that data labeling activities account for as much as 80% of the time needed to build AI systems.[4]

Labeling data for some AI apps may need high levels of skills and knowledge. For instance, to develop an AI app to detect cancer on images from a CT scan, experienced radiologists may have to train the algorithms (https://towardsdatascience.com/the-invisible-workers-of-the-ai-era-c83735481ba). However, there are many tasks that computers lack the capability to perform as well as ordinary human beings. Even less-skilled workers can easily be trained to perform such tasks. Most data labeling trainings can be completed in a short period of time. For instance, in order to learn their tasks, data labelers at iMerit typically take a one-week online training course via video calls with U.S.-based trainers.[4]

Table 1. Some Examples of Data Labeling Companies Operating in Developing Economies.

Data labeling company	**Operations and workforce**	**Profiles of clients**
iMerit	Based in India and the U.S. 2500 in data labeling centers in India such as Ranchi, Shillong, Vizag, and Kolkata (https://tinyurl.com/y465cqmq).	90% are U.S.-based (https://tinyurl.com/yyzn8vjd).
Samasource	Offices in San Francisco, New York, the Hague, Kenya, and Uganda. Global staff of 2900: East Africa's largest AI and data annotation employer (https://tinyurl.com/wqanmja).	Include 25% of the Fortune 50 companies including major automakers and U.S.-based technology companies such as Google, Microsoft, and IBM (https://tinyurl.com/qnwlgxo).
MBH	300,000 in China's backward provinces (https://tinyurl.com/yeb7qtbb).	Mainly Chinese companies such as the Beijing-based video-sharing social networking platform TikTok.
Playment	Based in India and the U.S. More than 300 000 crowdsourced data labelers (https://tinyurl.com/y2dqm3c3).	Over 100 customers in more than 12 countries (https://tinyurl.com/upp4wsu). Some include Samsung, Didi Chuxing Technology, Alibaba, Drive.ai, and Continental AG. Most clients are in the autonomous vehicle industry (https://tinyurl.com/y2dqm3c3).

Table 1 provides some examples of data labeling companies operating in developing economies. These firms mainly serve foreign clients. Chinese data labeling firms such as MBH, on the other hand, mainly support the domestic AI industry.

In China, which has gained global prominence in recent years in the AI field, research and development activities to support the growth of the AI industry are conducted in wealthy cities such as Beijing, Shanghai, Hangzhou, and Shenzhen. Most of the data labeling works, on the other hand, are performed in the country's disadvantaged regions such as smaller towns and rural areas in Shandong, Henan, Hebei, and Shanxi provinces in the North.[5] In particular, Henan Province is the epitome of an economy that has experienced an AI-led transformation. The Province's Zhengzhou city has been known for the manufacturing plants of Taiwanese electronics company Foxconn. It was estimated that Foxconn employed about 350 000 people and produced about half of the world's iPhones in Zhengzhou in 2016 (https://www.businessinsider.com/apple-iphone-factory-foxconn-china-photos-tour-2018-5).

The growth in data labeling also reflects the influences of a number of trends in the manufacturing sector. First, manufacturing activities have become increasingly automatized. For instance, between 2012 and 2016, Foxconn was reported to deploy tens of thousands of robots, which replaced more than 400 000 jobs. The company's plan includes full automation of 30% of production activities by 2020 (https://www.scmp.com/economy/china-economy/article/2185993/man-vs-machine-chinas-workforce-starting-feel-strain-threat).

Second, manufacturing jobs in China are declining due to a decline in the country's economic growth and, hence, demand for products, increasing costs, and competition from other economies (https://www.nytimes.com/2016/07/23/business/international/china-jobs-donald-trump.html). For instance, the Boston Consulting Group's study conducted in 2015 found that manufacturing costs in some of China's major manufacturing hubs were almost at the same levels as in the U.S. Unsurprisingly, many western companies have been moving and relocating manufacturing activities into other developing countries in Asia as well as to the U.S., Canada, and Mexico (https://www.nytimes.com/2016/07/23/business/international/china-jobs-donald-trump.html). For instance, factory workers in Bangladesh and Vietnam can be hired for less than a quarter and a half, respectively, of the salary of a Chinese worker.

Finally, there has been a generational shift in preference for work. An increasing number of Chinese millennials do not like dull, boring, and tedious factory jobs (https://www.latimes.com/world/la-fg-china-millennials-jobs-20190512-story.html).

The upshot of the above is a decline in industrial manufacturing and exports in China. For instance, Henan's mobile phone exports reduced by 23.7% in January 2019 compared to a year earlier (https://www.scmp.com/economy/china-economy/article/2188162/foxconn-tale-slashed-salaries-disappearing-benefits-and-mass). China's economically backward regions are embarking on ambitious plans to develop the data labeling industry. For instance, North China's landlocked province Shanxi, which is among the poorest provinces in China, plans to attract more than 100 data labeling companies and train more than 10 000 workers by 2022. The province's goal is to have a RMB 5 billion (US$726 million) industry by 2025 (https://kr-asia.com/data-labeling-jobs-are-coming-to-underdeveloped-regions-in-china-but-can-they-stay). Workers who were employed in assembly lines and construction sites before are thus finding new jobs in the Chinese data labeling industry.[6]

Partly as a response to the decline in manufacturing activities, in recent years, data labeling companies are mushrooming in the towns and villages of backward regions such as the Henan province.[7] In the Pingdingshan village of the province, some large data labeling projects are reported to employ tens of thousands of people.[8]

ECONOMIC AND SOCIAL IMPACTS OF THE DATA LABELING INDUSTRY

Finding high quality AI talents such as ML engineers has been big challenge for companies in developing economies (https://factordaily.com/india-ai-talent-gap-google-microsoft/*)*. For instance, India is estimated to have only about 50-75 AI researchers (https://via.news/asia/artificial-intelligence-development-india/). According to India's skill assessment company Aspiring Minds' Annual Employability Survey 2019, 80% of Indian engineers were considered to be unfit for a job. The survey also found that only 2.5% of them had skills required by the AI industry (https://www.businesstoday.in/current/corporate/indian-engineers-tech-jobs-survey-80-per-cent-of-indian-engineers-not-fit-for-jobs-says-survey/story/330869.html). India has also faced a severe shortage of qualified faculty members to teach AI courses in its universities.[9]

On the other hand, there is an abundant supply of low-skill and low-wage labors in India and other developing countries. For instance, Indian high schools graduated 20 million students in 2017.[10] There are, however, only a few job opportunities available to absorb these graduates. As an example, in a Mumbai city police's job advertisement for 1137 constable positions with US$357/month salary, which required only a high school education, over 200 000 people applied. Many of the candidates had been trained in highly skilled jobs such as doctors, lawyers, and engineers.[11]

In light of the high rates of unemployment and underemployment in developing countries, the economic impacts of the newly emerging data labeling industry are undoubtedly important. However, controversies have arisen regarding ethical and fair practices of data labeling firms.

Some data labeling firms have been accused of paying low wages to their workers. In this way, organizations engaged in this industry are allegedly facilitating a "new kind of slavery in the digital era" (https://www.weforum.org/agenda/2019/08/ai-low-skilled-workers/).

Although some data labeling firms have claimed to produce positive social effects, such claims cannot be easily verified. There are virtually no regulations that govern the working conditions of data labelers. Industry standards related to ethical sourcing are weak. There is no standard for reporting data-labeling firms' social impacts. There is also the lack of validation of such impacts by third parties. There is little hard evidence to counter critics' concerns that these firms claiming that they engage in impact sourcing is nothing more than marketing gimmicks or "impact-washing."[3]

Some data labeling firms such as CloudFactory, DDD, iMerit, and Samasource are members of the Global Impact Sourcing Coalition (GISC), which was founded in 2016. Among other measures, the GISC has established an "impact sourcing standard." The standard defines minimum requirements that the GISC members are to satisfy. It has also provided voluntary best practices for employment. The GISC requires its members' performance on criteria such as nondiscrimination and equal pay to be assessed every two years.[3]

When it comes to promoting ethical and socially responsible behaviors, however, the GISC at best is of questionable effectiveness and value. For instance, the violators face no penalty or sanction if they do not pass the assessment. They do not lose the GISC membership. the GISC members also differ significantly in the terms of the information they publish. Samasource's impact audits report includes indicators such as workforce demographics and the number of people lifted from poverty. The data labeling firm DDD's reports contain information about employees' earnings and increase in lifetime income. CloudFactory had not published social impact report since 2015. As of 2019, iMerit had not published any such reports.[3] The U.S.-based provider of data labeling and annotation services Alegion, which is not a GISC member, has outlined broad targets that it seeks to achieve but lacks specific metrics.[3]

Another challenge that arises here is that unlike fair-trade goods that are directly sold to consumers, data labeling firms provide their services to businesses. These firms thus face little public pressure to be honest, and it is ineffective to pressure them to engage in ethical practices.[3]

SUMMARY

The rapidly growing global AI industry has created demands for highly skilled jobs such as ML engineers and data scientists, as well as less-skilled works such as those of data labelers. In particular, most AI systems heavily rely on human-powered data labeling activities. Developing countries provide very large and seemingly bottomless sources and resources to support these activities to boost the global AI industry. Data labelers in these countries are playing a key role in curating the data that powers AI systems.

A notable feature of the data labeling industry is that it does not favor locations with specific cultural contexts. The data labeling market thus is characterized by a low entry barrier for most developing countries. Whereas the outsourcing of call center jobs gravitated to countries with sizable English-speaking populations such as India and the Philippines, English skill is less of a factor in data labeling jobs. Digital literacy is sufficient to participate in most data labeling tasks such as image classification.

Among developing countries, China has emerged as a key global AI player. The country's wealthy regions and big cities lack attractiveness for the development of data-labeling services industry. Such services in the country are thus mostly performed in the poorer and backward regions, which are providing economic incentives for data labeling firms.

What is not clear is whether data labeling firms in developing countries are operating in more or less ethical ways compared to other foreign firms operating in these countries. Some critics have claimed that this industry has features that are akin to slavery. While data labeling firms claim to engage in activities that have positive social impacts, it is not easy to assess the validity of such claims. Data labeling companies have their own definitions as to what are ethical and fair practices. Moreover, the definitions vary widely across them. Thus, we may not be able to take self-reported information provided by data labeling firms as proof positive that these firms are creating more positive social impacts in developing countries compared to other foreign firms. In many cases, the problem of assessing such claims is made more complex by the fact that they do not publish any information or fail to provide relevant information on such impacts.

REFERENCES

1. N. Kshetri, “Artificial intelligence in developing countries,” *IEEE IT Pro.*, to be published, doi: 10.1109/MITP.2019.2951851.
2. “Data-labelling startups want to help improve corporate AI,” *The Economist*, Oct. 17,2019. [Online]. Available: https://www.economist.com/business/2019/10/17/data-labelling-startups-want-to-help-improve-corporate-ai
3. K. Kaye, “These companies claim to provide “Fair-trade” data work. Do they?” *MIT Technol. Rev.*, Aug. 7,2019. [Online]. Available: https://www.technologyreview.com/s/614070/cloudfactory-ddd-samasource-imerit-impact-sourcing-companies-for-data-annotation/
4. C. Metz, “A.I. is learning from humans. Many humans,” *The New York Times*, Aug. 16,2019. [Online]. Available: https://www.nytimes.com/2019/08/16/technology/ai-humans.html?auth=linked-google
5. S. Dai, “AI promises jobs revolution but first it needs old-fashioned manual labour—from China,” *South China Morning Post*, Oct. 8,2018. [Online]. Available: https://www.scmp.com/tech/article/2166655/ai-promises-jobs-revolution-first-it-needs-old-fashioned-manual-labour-china
6. L. Yuan, “How cheap labor drives China's A.I. ambition,” *The New York Times*, Nov. 25,2018.
7. H. Wu, “China Is achieving AI dominance by relying on young blue-collar workers,” *Vice*, Dec. 21,2018. [Online]. Available: https://www.vice.com/en_us/article/7xyabb/china-ai-dominance-relies-on-young-data-labelers
8. C. Cadell, “Faces for cookware: Data collection industry flourishes as China pursues AI ambitions,” *Reuters*, Jun. 28,2019. [Online]. Available: https://www.reuters.com/article/us-china-ai-data-insight/faces-for-cookware-data-collection-industry-flourishes-as-china-pursues-ai-ambitions-idUSKCN1TS3EA
9. S. Ravi and P. Nagaraj, “Harnessing the future of AI in India,” *Brookings Inst.*, Oct. 18,2018. [Online]. Available: https://www.brookings.edu/research/harnessing-the-future-of-ai-in-india/

10. K. Sheehy, “High school grads in China, India are better prepared for college,” *U.S. News.*, Aug. 27,2012. [Online]. Available: https://www.usnews.com/education/blogs/high-school-notes/2012/08/27/high-school-grads-in-china-india-are-better-prepared-for-college

11. N. Bagri, “India is trapping its young people,” *Foreign Policy*, May 14,2019. [Online]. Available: https://foreignpolicy.com/2019/05/14/india-is-trapping-its-young-people/

Nir Kshetri is currently a Professor of management with the Bryan School of Business and Economics, University of North Carolina at Greensboro. Contact him at nbkshetr@uncg.edu.