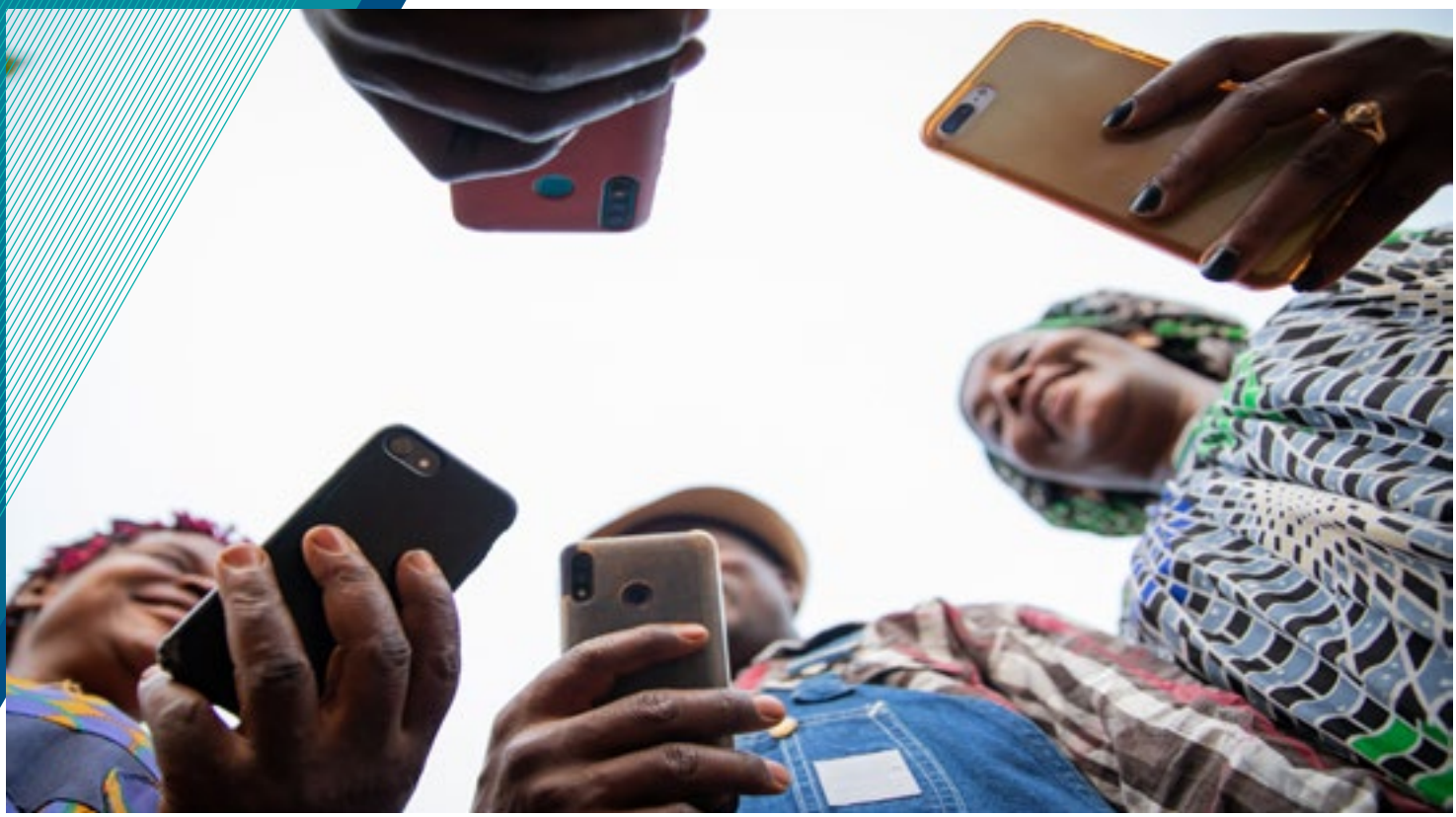
Nr. 521 / Januar 2024
Analysen & Argumente

KONRAD ADENAUER STIFTUNG



KI in Subsahara-Afrika

Ein Afrofuturismus ohne afrikanische Sprachen?

Jan-Ole Voß

- Im Bereich der Entwicklung von *large language models* (LLMs) haben sowohl Subsahara-Afrika als auch Europa einen schweren Stand. Gemeinsame Gründe sind unter anderem die Diversität der gesprochenen Sprachen und ein Mangel an verfügbaren Daten für das Training von LLMs.
- Erste Entwicklungen von LLMs in regional dominierenden Sprachen setzen den Fokus auf jene Sprachen, die die größten Märkte umfassen. Die sprachliche Diversität im digitalen Raum könnte so sinken.
- Um die Verbreitung von KI-Anwendungen zu fördern, ist eine verstärkte Kooperation mit Datenerhebungsinstitutionen empfehlenswert, ebenso sollte ein Werben für liberale Datenschutzregulierungen bei politischen Partnern in Betracht gezogen werden.
- Innovative, effizienz- und partizipationssteigernde KI-Lösungen existieren bereits und bieten ein großes Potenzial für menschliche sowie wirtschaftliche Entwicklung in Subsahara-Afrika, bedürfen aber guter Rahmenbedingungen.

Inhaltsverzeichnis

KI, Sprache und Daten

Sogenannte *large language models* (LLMs) sind darauf ausgelegt, Sprache zu erkennen, zu verstehen und wiederzugeben. Ein mit entsprechenden Daten trainiertes LLM kann Sprache anhand von Vorhersagen generieren, bekanntestes Beispiel für LLMs ist GPT-3.5 von OpenAi. Die öffentlich zugängliche Demo ChatGPT löste sowohl Euphorie als auch Sorge aus. Schon heute bieten derartige KI-Modelle enorme Chancen, ihre fortwährende Entwicklung verspricht weitreichende Effizienzsteigerungen und erleichterte Zugänge zu beispielsweise medizinischen Diagnosen. Um diese Fähigkeiten zu erlangen, müssen jedoch zuerst riesige Datenmengen zugeführt werden. Diese Notwendigkeit der Verfügbarkeit großer Datenmengen impliziert jedoch Schwierigkeiten für Bereiche, in denen wenige Daten verfügbar sind und solche, bei denen Daten zwar vorhanden, aber nicht digitalisiert sind.

Der Großteil der KI-Entwicklung fand bisher in den USA und China statt, beide Standorte haben einige Wettbewerbsvorteile gegenüber Entwicklern aus der EU oder anderen Regionen. Neben massiven Investitionen von Seiten der Tech-Konzerne, Finanzinvestoren und im Falle Chinas auch die Unterstützung durch die chinesische Zentralregierung, ist hier besonders die zurückhaltende Regulierung von Daten zu nennen. Nutzbare digitalisierte Daten für das Training von LLMs sind in beiden Staaten in großer Menge verfügbar. Besonders Entwickler in den USA haben Zugriff auf große internationale Datensätze von Millionen Nutzern US-amerikanischer Internetdienste. Ein weiterer Vorteil: Circa 55 Prozent des Internets sind auf Englisch, daraus ergibt sich ein deutlich größerer Pool an verfügbaren Daten. Auch chinesische Unternehmen haben Zugriff auf enorme Datenmengen, vor allem wenn es um den nationalen Markt geht. Besonders zu nennen sind hier Daten der Regierung, die Unternehmen für die Entwicklung von fortschrittlicher Gesichtserkennung zur Verfügung gestellt werden.[1] Im Vergleich ist der europäische Raum ein kompliziertes Feld für Entwickler, starke Datenschutzregulierungen, sprachliche Fragmentierung und eine träge Investitionsumgebung erschweren die Entwicklung. Eine Armut an digitalisierten Daten herrscht in Europa nicht, nur ist ihre Nutzbarkeit stark eingeschränkt, beispielsweise durch Effekte der DSGVO.[2] [3]

Subsahara-Afrika

Ein Blick auf den afrikanischen Kontinent offenbart eine noch größere sprachliche Diversität als in Europa. Etwa ein Drittel aller weltweit gesprochenen Sprachen stammt aus Subsahara-Afrika. Schätzungen beziffern die Gesamtzahl an afrikanischen Sprachen auf 1.000 bis 2.000.[4] Insgesamt existieren zurzeit 75 afrikanische Sprachen, die jeweils von einer Million oder mehr Menschen gesprochen werden. Zusätzlich zu dieser komplexen linguistischen Geografie kommt eine oft nicht vollständige Dokumentierung dieser Sprachen. Viele der bis zu

2.000 Sprachen existieren nur in mündlicher Tradition, es gibt häufig keine standardisierten schriftlichen Überlieferungen, geschweige denn digitalisierte Datensätze. Als digitalisierte Ressourcen sind im Grunde sämtliche Spracherzeugnisse zu verstehen, dazu zählen nicht ausschließlich Romane, wissenschaftliche Artikel oder Zeitungen, sondern auch Chatverläufe, E-Mails und Audioaufnahmen. Zusätzlich zu einer Fülle an digitalisierten Daten ist eine gewisse Standardisierung der Sprache wichtig, also eine klare Rechtschreibung und Grammatik, auf die sich LLMs beziehen können. Bei Sprachen mit relativ begrenzten Datensätzen oder unklaren Strukturen spricht man im Kontext von LLMs von *low-resource languages*. Als *high-resource languages* gelten Sprachen, die in großem Ausmaß als digitalisierte Ressource verfügbar sind, allen voran Englisch, Mandarin aber auch zu einem gewissen Grad Deutsch. Aus diesen Gründen wird von vielen Seiten befürchtet, dass zahlreiche Sprachen im Laufe der Zeit mehr und mehr aus dem digitalen Raum verdrängt und durch sogenannte *high-resource languages* ersetzt werden.[5]

Ein eklatanter Datenmangel existiert nicht nur im Bereich der Sprachen, auch statistisch repräsentative Daten zu verschiedenen sozioökonomischen Kategorien des gesellschaftlichen Zusammenlebens sind in vielen afrikanischen Ländern unvollständig vorhanden und oft von Schwierigkeiten bei den Erhebungsprozessen geplagt.[6] Derartige repräsentative native Daten aus spezifischen afrikanischen Kontexten sind indes unabdingbar, um akkurate KI-Modelle zu entwickeln. KIs auf Basis importierter Daten aus den USA oder China beinhalten die Spezifika der neuen soziokulturellen Umgebung nicht und drohen so falsche Vorannahmen, im Englischen *bias* genannt, zu vermitteln.[7] Daher ist eine höhere Qualität sowie größere Quantität an repräsentativen Daten zu allen Nischen afrikanischer Märkte und Gesellschaften notwendig.[8] Diese müssen nicht direkt für eine KI-Entwicklung von Grund auf genutzt werden, oft ist es auch möglich, neue Daten einzuspeisen und vorhandene Systeme so an die neuen Anforderungen anzupassen.

Neben der sprachlichen Fragmentierung und dem Problem des Datenmangels sind unter anderem auch mangelnde Weiterbildungschancen für KI-Interessierte, eine gering ausgeprägte digitale Infrastruktur und fehlende nationale KI-Strategien Herausforderungen für die Entwicklung und Nutzung von KI auf dem Kontinent.[9] Das Meistern dieser Herausforderung würde sich allerdings als überaus lohnend erweisen. Die Chancen, die KI auch in infrastrukturell schlecht erschlossenen Regionen eröffnen kann, umfassen nahezu sämtliche Aspekte des menschlichen Zusammenlebens.

Die Chancen, die KI auch in infrastrukturell schlecht erschlossenen Regionen eröffnen kann, umfassen nahezu sämtliche Aspekte des menschlichen Zusammenlebens.

Silicon Savannah

Führend im Tech-Sektor Subsahara-Afrikas ist die sogenannte Silicon Savannah Nairobis. Ihren Beinamen verdiente sich die Stadt durch den großen kommerziellen Erfolg von mPesa, einem Bezahldienst auf Basis von Mobiltelefonen. Gemeinhin ist Nairobi seither als Tech-Hub Subsahara-Afrikas bekannt. Kenia ist außerdem neben Uganda, Tansania und Ruanda eines der vier Länder, in denen Swahili Amtssprache ist. Swahili ist die weitverbreitetste afrikanische Sprache und wird heute von bis zu 200 Millionen Menschen gesprochen.[10]

In der Silicon Savannah werden währenddessen die ersten Schritte für die Zukunft von Swahili im Bereich von KI-Anwendungen getan. Im Oktober 2023 startete UlizaLlama ein auf Basis von Metas Llama2 entwickeltes LLM. Die Besonderheit ist, dass UlizaLlama als Free-to-Use- und Open-Source-Software entwickelt wurde, hierdurch stellt der Entwickler Jacaranda Health, eine kenianische NGO, eine Grundlage für andere Entwickler zur Verfügung. UlizaLlama ist darauf ausgelegt, auf den eigenen Servern der Entwickler zu laufen

und bietet damit die Chance, sensible Daten am Standort zu behalten. Jacaranda Health plant die auf Swahili trainierte KI in der medizinischen Beratung von Müttern und werdenden Müttern einzusetzen.[11] Mit einer Verteilung von aktuell 0,2 Ärzten auf 1.000 Einwohnern fehlt vielen Kenianerinnen der Zugang zu jeglicher ärztlicher Beratung, vergleichsweise kommen in Deutschland auf 1.000 Einwohner 4,4 Ärzte.[12] Besonders ländliche Regionen Kenias sind betroffen, hier ist der Mangel noch eklatanter. Durch die vergleichsweise gute Mobilfunkabdeckung und die breite Verfügbarkeit von Mobiltelefonen in Kenia kann UlizaLlama auch in entlegeneren Gegenden oder ärmeren Gesellschaftsschichten Beratungsleistungen erbringen. Der Mehrwert für die menschliche Entwicklung durch eine derartige KI-gestützte Beratungsleistung ist evident. Auch wirtschaftlich sind Gewinne erkennbar, eine frühere und zugänglichere Beratung durch gute KI kann spätere gesundheitliche Komplikationen reduzieren und so Folgekosten senken. Der Einsatz von KI in anderen Bereichen weist ähnliche Vorteile auf. Wichtig ist hierbei die sprachliche Zugänglichkeit der Anwendungen, häufig wird auf automatische Übersetzungen zurückgegriffen. Eine derartige teilweise Umgehung des Problems der *low-resource languages* durch das Zwischenschalten von automatischen Übersetzungssoftwares führt aber zu weiteren Problemen. LLMs werden durch automatische Übersetzungen oft unpräziser in ihren Formulierungen, kontextspezifische Botschaften gehen so verloren.[13] Daher sind Entwicklungen wie UlizaLlama überaus wichtig für eine praxisorientierte und nutzerfreundliche Ausbreitung neuer KI-Lösungen in Subsahara-Afrika. Für viele afrikanische Sprachen sind häufig noch keine verlässlichen Übersetzungsprogramme verfügbar, ihrer Entwicklung müsste oft eine formelle Bestandsaufnahme der Sprache vorhergehen. Dabei müsste nicht nur syntax-, sondern insbesondere auch kontextbezogene Semantik erfasst werden. Der Kostenaufwand, um diese Vorarbeit für die 75 verbreitetsten afrikanischen Sprachen zu leisten, wäre enorm. Daher ist aus rein wirtschaftlichen Gründen eine Fokussierung auf einige wenige weitverbreitete Sprachen zu erwarten. Dies zeigt sich auch dadurch, dass Swahili aktuell die einzige afrikanische Sprache ist, auf die Bard, Googles Inhouse-KI, trainiert wird.[14] Verschiedene Autoren sehen hier besonders die verschiedenen Regierungen in der Bringschuld, die Voraussetzungen zu verbessern und die Datenverfügbarkeit zu stärken.[15] [16]

LLMs werden durch automatische Übersetzungen oft unpräziser in ihren Formulierungen, kontextspezifische Botschaften gehen so verloren.

Aus rein wirtschaftlichen Gründen ist eine Fokussierung auf einige wenige weitverbreitete Sprachen zu erwarten.

Ausblick

Aus den Perspektiven des Datenmangels und der sprachlichen Diversität weist die Situation Subsahara-Afrikas einige Parallelen zu aktuellen offenen Fragen in Europa auf. Allerdings kommen in Subsahara-Afrika noch zahlreiche, hier nur am Rande erwähnte Probleme hinzu. Sowohl Europa als auch Subsahara-Afrika sollten sich auf Wesentliches konzentrieren, um nicht abgehängt zu werden. Ein Blick auf die zahlreichen Sprachen Subsahara-Afrikas, ihre lokalen Unterschiede sowie die oft komplizierte Datenlage lässt vermuten, dass viele Sprachen nicht die nötige finanzielle Aufmerksamkeit von Investoren bekommen werden. Daher werden nicht alle der dutzenden gesprochenen Sprachen Kenias oder der 1.000 bis 2.000 Sprachen des Kontinents die gleiche Rolle spielen können. Dies gilt ebenso für Minderheitensprachen in Europa. Investitionen werden sich auf die Sprachen mit dem größten potenziellen Markt fokussieren. Das heißt allerdings nicht, dass afrikanische Nationen sich keinerlei sprachliche Autonomie im digitalen Raum bewahren können. Wie das Beispiel UlizaLlama zeigt, bieten Sprachen wie Swahili schon heute die Möglichkeit, massentaugliche LLM-Projekte zu entwickeln. Ein nationaler oder sogar regionaler linguistischer Konsens, wie er mit Swahili in Ostafrika vorhanden ist, kann die Attraktivität für zeitnahe Entwicklungen im LLM-Bereich stärken. Derartige Projekte werden, sofern sie erfolgreich sind, auch die lokalen sprachlichen Unterschiede reduzieren und so zur Vereinheitlichung beitragen.

Um solche Entwicklungen zu ermöglichen, ist die Verfügbarkeit von großen Mengen qualitativ hochwertiger Daten essenziell, dies können linguistische Ressourcen oder statistische Daten zu Wirtschaft, Gesellschaft oder Politik sein. Wenn Subsahara-Afrika die möglichen Fortschritte im Bereich KI für ein weiteres Leap-Frogging nutzen will, braucht es Maßnahmen, die die Grundlagen für KI-Anwendung und -Entwicklung erleichtern. Das könnten neben klassischer digitaler Infrastruktur (Breitbandanschlüsse, Rechenzentren etc.) auch die Stärkung nationaler Statistikbehörden sein. Beispielsweise könnte der Zugang zu Haushalts- und Finanzdaten transparenter gestaltet werden. Regelmäßige umfassende Haushaltsumfragen, gekoppelt mit detaillierten und transparenten Finanzdaten aus der Verwaltung könnten neue Potenziale für den Einsatz von KI, besonders im Good-Governance-Bereich, eröffnen. Klare, liberale und entwicklerfreundliche Datenschutzregulierungen sind ebenfalls bedeutend. Einen stärkeren Schutz von individuellen Datenrechten nachzusteuern, ist immer möglich. Durch zu restriktive Auflagen unmöglich gemachte Innovation nachträglich zu „verordnen“ ist hingegen unmöglich.

Wenn Subsahara-Afrika die möglichen Fortschritte im Bereich KI für ein weiteres Leap-Frogging nutzen will, braucht es Maßnahmen, die die Grundlagen für KI-Anwendung und -Entwicklung erleichtern.

Implikationen

Für die deutsche Politik und Entwicklungszusammenarbeit wäre eine Kooperation mit afrikanischen Partnern in den Bereichen Datenregulierung und Datenerhebung interessant. Beide Bereiche sind geeignet, die Grundvoraussetzung für KI-Entwicklung und Anwendung zu verbessern und so neue Märkte zu erschließen. Eine Zusammenarbeit mit Datenerhebungsinstituten, wie zum Beispiel der NGO Afrobarometer, wäre denkbar. Ebenso wäre ein Werben für eine liberale und transparente Datenpolitik bei politischen Partnern anzustreben. Hierdurch könnte die EU dazu beitragen, in dem wichtigen wirtschaftlichen Feld der Datenpolitik ein Gegengewicht zu China zu schaffen. Rahmenbedingungen könnten so auf breiter Ebene verbessert werden. Eine transparentere Datenpolitik vonseiten der verschiedenen Regierungen hätte auch den Nebeneffekt, dass sie Korruption und Verschwendung sichtbarer machen würde, ohne dabei diese sensiblen Themen direkt anzusprechen. Bei allen Maßnahmen sollte auch stets die digitale Stärkung der Peripherie eine Rolle spielen, damit nicht nur die Realitäten des Zentrums in den Daten abgebildet werden.

1 Beraja, Martin et al. (2021). "Data-Intensive Innovation and the State: Evidence From AI Firms in China", in: NBER Working Paper Series, Nr. 27723.

2 Martin, Nicholas et al. (2019). "How Data Protection Regulation Affects Startup Innovation", in: Information Systems Frontiers, https://doi.org/10.1007/s10796-019-09974-2 (zuletzt abgerufen am 5.1.2024).

3 Wallace, Nick. Castro, Daniel (2018). "The Impact of the EU's New Data Protection Regulation on AI". https://datainnovation.org/2018/03/the-impact-of-the-eus-new-data-protection-regulation-on-ai/ (zuletzt abgerufen am 14.12.2023).

4 Harvard University, The African Language Program at Harvard. "Introduction to African Languages". https://alp.fas.harvard.edu/introduction-african-languages#:~:text=With%20anywhere%20between%201000%20and,more%20than%20one%20million%20speakers (zuletzt abgerufen am 11.12.2023).

5 Pereira, David (2023). "How can NLP impact the future of minority languages", in: Towards Data Science, https://towardsdatascience.com/how-can-nlp-impact-the-future-of-minority-languages-555b0fc80bd0 (zuletzt abgerufen am 18.12.2023).

6 Seidler, Valentin et al. (2023). "Subnational Variations in the Quality of Population Health Data: A Geospatial Analysis of Household Surveys in Africa". https://ssrn.com/abstract=4508419 (zuletzt abgerufen am 11.12.2023).

7 Chinganya, Oliver (2023). "The Future of AI in Statistics in Africa: Is the Continent Ready?". https://www.isi-web.org/article/blogs/future-ai-statistics-africa-continent-ready (zuletzt abgerufen am 11.12.2023).

8 Ade-Ibijola, Abejide. Okonkwo, Chinedu (2023). "Artificial Intelligence in Africa: Emerging Challenges", in: Responsible AI in Africa, Challenges and Opportunities. Hrsg. Damian Okaibedi Eke et al., S. 101–17, S. 107–8.

9 Ebd., S. 105.

10 Harvard University, The African Language Program at Harvard. "Introduction to African Languages".

11 Jacaranda Health (2023). "Jacaranda launches first-in-kind Swahili Large Language Model". https://jacarandahealth.org/jacaranda-launches-first-in-kind-swahili-large-language-model/ (zuletzt abgerufen am 11.12.2023).

12 Weltbank. "Physicians (per 1,000 people)". https://data.worldbank.org/indicator/SH.MED.PHYS.ZS (zuletzt abgerufen am 11.12.2023).

13 Nicholas, Gabriel. Bhatia, A. (2023). "Lost in Translation: Large Language Models in Non-English Content Analysis", S. 26. https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/ (zuletzt abgerufen am 11.12.2023).

14 Oginga, Bernard (2023). "Kiswahili at the center of digital advancement, Artificial Intelligence". https://www.un.org/africarenewal/magazine/july-2023/kiswahili-center-digital-advancement-artificial-intelligence (zuletzt abgerufen am 11.12.2023).

15 Ade-Ibijola, Abejide. Okonkwo, Chinedu (2023). "Artificial Intelligence in Africa: Emerging Challenges", S. 113.

16 Chinganya, Oliver (2023). "The Future of AI in Statistics in Africa: Is the Continent Ready?".

Impressum

Der Autor

Jan-Ole Voß ist Trainee im Auslandsbüro der Stiftung in Nairobi. Er studierte unter anderem an der Kobe Universität in Japan und an der Zhejiang Universität in China.

Konrad Adenauer-Stiftung e. V.

Jan-Ole Voß
Subsahara-Afrika
Europäische und Internationale Zusammenarbeit
jan-ole.voss@kas.de

Postanschrift: Konrad-Adenauer-Stiftung e. V., 10907 Berlin

Herausgeberin: Konrad-Adenauer-Stiftung e. V., 2024, Berlin
Gestaltung: yellow too, Pasiek Horntrich GbR
Satz: Janine Höhle, Konrad-Adenauer-Stiftung e. V.
Hergestellt mit finanzieller Unterstützung der Bundesrepublik Deutschland.

ISBN 978-3-98574-209-7