Generative KI

Trainingsdaten in Form bringen

[28.08.2023] Guter Input ergibt guten Output – diese einfache Regel gilt auch für die Daten, mit denen KI-Anwendungen trainiert werden. Je besser Daten aufbereitet sind, desto effizienter ist die Entwicklung und desto sicherer die spätere KI-Lösung.

Anwendungen, die Daten mithilfe Künstlicher Intelligenz (KI) oder maschinellem Lernen (ML) verarbeiten, werden derzeit breit diskutiert. Die Debatte konzentriert sich dabei vor allem auf ethische und sicherheitsrelevante Aspekte und damit auf Fragen des richtigen Einsatzes von (generativer) KI. Die Frage nach der Qualität solcher Anwendungen, die wiederum von den Daten abhängt, mit denen die Algorithmen trainiert werden, wird in der öffentlichen Diskussion vernachlässigt. Zu diesem Schluss kommt das Schweizer Data-Intelligence-Unternehmen Aparavi. Aparavi schätzt, dass bis zu 80 Prozent der Daten, die für das KI-Training infrage kommen, unstrukturiert sind. In diesen Beständen verbergen sich nicht nur veraltete Dokumente oder risikobehaftete Daten, sondern auch wichtige und wertvolle Informationen. Solche Datenbestände müssten schon vorab klassifiziert und bereinigt werden.

Sensible Daten aussieben

Eine saubere Data Collection ist für eine sinnvolle, effektive Entwicklung von KI-Apps essenziell. Denn die Qualität des Outputs bei der KI-Entwicklung hängt zwangsläufig von der Qualität des Inputs ab – je gepflegter die Trainingsdaten, desto höher der Anwendungsnutzen. Ideal sind „transparente, klassifizierte, strukturierte und priorisierte Daten und Metadaten“, so Aparavi, die auch von Dubletten bereinigt sein sollten. Ebenso wichtig sei es, kritische sensible Daten und Dokumente, die aus rechtlichen Gründen nicht verwendet werden dürften, auszusieben. Dazu gehören beispielsweise personenbezogene Daten oder Inhalte, die vor dem Stichtag einer Änderung rechtlich relevanter Vorgaben datieren. Um Verfälschungen, aber auch Risiken und Strafzahlungen zu vermeiden, müssten diese identifiziert und aus Datenbeständen entfernt werden, noch bevor die Datenbestände für das Training generativer KI-Anwendungen herangezogen werden.

Entwicklungszeiten abkürzen

Die Nutzung künstlich erzeugter Datensätze, so genannter synthetischer Daten, nimmt zu. Die Fachleute von Aparavi sehen dies als Indikator der Unzufriedenheit von Data Scientists mit dem vorhandenen echten Datenmaterial. Dennoch seien synthetische Daten kein vollwertiger Ersatz für das Training von KI-Anwendungen: Mit Originaldaten könnten Algorithmen und Anwendungen deutlich schneller und effizienter entwickelt werden als mit simulierten Datensätzen.
KI-Entwicklung ist per se ein iterativer Prozess mit hohem Ressourcenbedarf – und verursacht folglich hohe Kosten. Ein schlechter Dateninput verlängert die Entwicklungszeiten und erhöht die Kosten zusätzlich. Ein sauberer, auf relevante, sinnvolle Daten kondensierter Datenbestand kann die Anwendungsentwicklung hingegen beschleunigen und damit auch den finanziellen Aufwand reduzieren. „Clean and Lean Data spielen bei der Entwicklung von KI- und ML-Apps eine überragende Rolle“, sagt der Aparavi-CEO Adrian Knapp. Ob eine KI-Anwendung erfolgreich wird, entscheide sich an den Daten, die sozusagen das Futter für die zu trainierenden Algorithmen darstellen.

(sib)

Stichwörter: Panorama, Aparavi, KI, künstliche Intelligenz