

Wie lernt eigentlich eine Künstliche Intelligenz und warum sprechen alle von Daten, Daten und Daten?

Der Artikel erklärt, wie Künstliche Intelligenz (KI) mit vielen Beispielen (Daten) lernt, um bestimmte Aufgaben zu lösen. Die EU fordert nun, dass Unternehmen offenlegen, welche Daten sie für das KI-Training verwenden. Das stellt Anbieter wie Meta und OpenAI vor Herausforderungen. Wichtig für gute KI-Ergebnisse sind passende und genaue Daten sowie Fachwissen, zum Beispiel beim Ermitteln von Baupreisen.



Markus Nussbaum

Die Bauherrenhilfe stellt allen Mitgliedern und Partnern einen „KI-Beauftragten“ zur Seite! Markus Nussbaum hat die HTL für Hochbau sowie Bauingenieurwesen studiert und bereits eine jahrelange Berufserfahrung hinter sich gebracht. Er ist zudem zertifizierter KI-Manager.

Kontakt:
m.nussbaum@bauherrenhilfe.at

Im letzten KI-Beitrag ging es um die Einführung der neuen EU-Verordnung (AI Act), die hauptsächlich darauf abzielt, Künstliche Intelligenz im Hinblick auf „Transparenz“ zu regulieren. Doch welche Auswirkungen hat diese Verordnung, insbesondere auf die Bauwirtschaft? Hier gleich die enttäuschende Antwort: Das lässt sich nicht eindeutig sagen. Die KI-Verordnung regelt nicht alle Aspekte rund um das Thema KI. Es kommen zum Beispiel auch Bestimmungen aus dem Urheberrecht, dem Datenschutzrecht und dem Produkthaftungsrecht ins Spiel.

Herausforderungen für internationale KI-Anbieter

Der Tatsache, dass der AI Act Künstliche Intelligenz innerhalb der EU regulieren soll, steht gegenüber, dass die derzeit meistgenutzten KI-Modelle aus Regionen außerhalb der EU (konkret aus den USA) stammen. Im Rahmen der KI-Verordnung müssen KI-Entwickler unter anderem die Trainingsdaten der verwendeten KI-Modelle transparent offenlegen (vorausgesetzt, die Verordnung gilt für die jeweilige KI, die in der EU eingesetzt wird). Die Reaktion der US-amerikanischen Tech-Giganten auf die EU-Regulierung dürfte kaum positiv ausfallen. So plant Meta (früher Facebook) nach aktuellem Stand, die neueste Version seines Sprachmodells „LLaMA“ nicht in der EU zur Verfügung zu stellen. Andererseits könnte dies die Entwicklung von „AI made in Europe“ fördern. Ausgerechnet OpenAI (Entwickler von ChatGPT) ist doch sehr verschlossen, was die verwendeten Trainingsdaten angeht. Was hat es mit den diesen Daten auf sich?

Was sind eigentlich Daten?

Lassen wir die rechtlichen Aspekte beiseite und tauchen tiefer in das Thema Daten und KI-Training ein. In unserer Welt gibt es nahezu unendlich viele Informationen, die aus den unterschiedlichsten Quellen stammen

können (z. B. durch Messen und Beobachten). Wir versuchen stets, diese Informationen zu speichern und weiterzugeben. Durch diesen Prozess entstehen „Daten“. Ein Beispiel dafür ist ein Datensatz in Form einer Tabelle mit Spalten und Zeilen, der Informationen über etwas enthält (z. B. eine einfache Excel-Tabelle).

Strukturierte und unstrukturierte Daten

Wenn man etwas beobachtet, kann es verschiedene Merkmale aufweisen. Diese Merkmale (auch Features genannt) werden in den Spalten der Tabelle dargestellt. Die jeweiligen Beobachtungen entsprechen den Zeilen. Der Schnittpunkt eines bestimmten Merkmals mit einer bestimmten Beobachtung wird als Datenpunkt bezeichnet. Daten in diesem Format sind „strukturierte“ Daten (geordnet). Viele andere Daten sind jedoch nicht strukturiert (z. B. Bilder, Videos, Audios und Dokumente). In diesem Fall sprechen wir von „unstrukturierten“ Daten. Die Daten selbst bestehen entweder aus Zahlen (numerische Daten) oder Wörtern (kategorische Daten). Diese unterschiedlich strukturierten Daten werden dem jeweiligen ML-Algorithmus zum Training zur Verfügung gestellt. Das Maschinelle Lernen (ML) kennt folgende Lernarten: Überwachtes Lernen, unüberwachtes Lernen und verstärkendes Lernen. Wichtig dabei: Wir unterscheiden zwischen dem KI-Training (das Ergebnis ist ein KI-Modell) und der Nutzung des KI-Modells (das „fertige“ KI-Modell trifft z. B. Vorhersagen). In diesem Beitrag wird das „überwachte Lernen“ thematisiert.

Wie lernt KI durch überwachtes Lernen?

Beim überwachten Lernen soll die KI lernen, aus einem Input (Feature) den entsprechenden Output (Label) zu bestimmen. Im KI-Training werden dem ML-Lernalgorithmus Trainingsdaten mit Features (Eingaben) gegeben, die bereits einen bekannten Output

(Ausgabe oder Label) enthalten. Ein Beispiel dafür wären Bilder von Katzen und Hunden (Input), die korrekt als solche „gelabelt“ (bekannter Output) sind. Während des Trainings erkennt der Lernalgorithmus Muster und Zusammenhänge zwischen dem Input und dem bekannten Output und erstellt daraus ein Vorhersagemodell. Wenn die KI während des Trainings ein Bild fälschlicherweise als „Katze“ klassifiziert, obwohl die korrekte Lösung „Hund“ lautet, kann der Lernalgorithmus das Vorhersagemodell entsprechend anpassen (die Lösung ist der KI im Training also bekannt). Das Ziel ist, dass die KI bei einem Bild von einem Hund mit hoher Wahrscheinlichkeit einen Hund und keine Katze erkennt. Das trainierte KI-Modell soll schließlich in der Lage sein, auf einen „neuen“ Input (ein neues Bild eines anderen Hundes) den richtigen Output vorherzusagen (nämlich, dass es sich um einen Hund handelt).

Modelle und Methoden des überwachten Lernens

Lernmodelle des überwachten Lernens, bei denen ein Label (z. B. Katze/Hund) als Output erzeugt wird, werden als Klassifikationsmodelle bezeichnet. Das Gegenstück dazu ist das sogenannte Regressionsmodell, bei dem eine Zahl (z. B. ein Angebotspreis) der Output ist. Für beide Fälle gibt es verschiedene Modellverfahren (z. B. lineare Regression, logistische Regression, Entscheidungsbäume, Random Forests und Künstliche Neuronale Netze), die im Grunde, wie oben beschrieben, stets dasselbe Ziel verfolgen.

Bedeutung der Datenqualität für KI-Ergebnisse

Noch wichtiger als die Wahl des Modells ist jedoch die Qualität (und nicht nur die Quantität) der Trainingsdaten. Wenn beispielhaft ein KI-Modell für die Angebotserstellung entwickelt wird, ist die Auswahl passender Daten für das KI-Training entscheidend. Bei der Ermittlung der Einheitspreise sind Informationen zu den Leistungsansätzen von zentraler Bedeutung. Wenn diese Eingangsvariablen (aus Erfahrungswerten oder Fachliteratur) im KI-Training fehlen oder unzureichend sind, kann keine hochwertige Beziehung zwischen den Leistungsansätzen und dem finalen Einheitspreis hergestellt

werden. Daher sind auch fundierte Fachkenntnisse, in diesem Fall zur Baupreisermittlung, für die Datenaufbereitung und das KI-Training unerlässlich. Es gilt: Mist rein = Mist raus.