FAHSE, Christian & LUTZ, Tim
Landau

Codieren mit KI - Herausforderungen und Chancen

Codieren von Texten und Transkripten ist fundamental für empirische didaktische Forschung aber auch ressourcenintensiv. Was liegt näher, als hierfür Künstliche Intelligenz einzusetzen? Aber ist solch eine Auswertungsmethode auch didaktisch vertretbar angesichts der grundsätzlich statistischen Natur von KI? (Stichworte overfitting, stochastic parrot (Bender et al. 2021)) Wie ein valides Vorgehen aussehen kann und welche neuen Chancen KI bietet, wird hier an einem konkreten Beispiel skizziert.

Studie zur Argumentationskompetenz

Auf dem Weg zu einer Sonde (Fahse 2017) zur Messung der Kompetenz K1 (Argumentation) auf Lerngruppenebene wurde mit Qualitativer Inhaltsanalyse ein theoretisches Modell entwickelt, das drei Argumentationstypen unterscheidet: Reichhaltige (R), pseudofachliche (P) und apodiktische (A) Argumentation (Fahse 2017, siehe dort).

Dieser Artikel berichtet von Daten eines einzelnen schriftlich erhobenen Items bei Lernenden ab Jahrgangstufe 7 bis hin zu Abitur und Studium (Was ist das Ergebnis der Aufgabe 7:0? Begründe Deine Meinung so, dass jemand, der die Antwort nicht kennt, es versteht.). Die Herausforderung für die Codierung ist hier hinreichend komplex, da z. B. die mathematische Korrektheit der Antwort kein Kriterium für die Argumentationskompetenz sein soll, um nicht zwei potenziell voneinander unabhängige latente Fähigkeiten zu vermischen. Der Typ R umfasst Begründungsstrategien, die grundsätzlich inhaltlich erfolgreich sein könnten, wie z. B. das Nicht-Gelingen eines Verteilungsprozesses als Grundvorstellung zur Division oder das Scheitern der Umkehraufgabe als algebraische Grundvorstellung. Der Typ P versucht, eine mathematisch-inhaltliche Begründung zu imitieren (z. B. durch eine Phantasieregel "7:0=0, denn alle Aufgaben mit 0 ergeben null."). Diese imitierende Vorgehensweise entspricht Lernstrategien, die beim Erwerb von Fremdsprachen (Stichwort induktive Grammatik) durchaus zielführend sind. Der apodiktische Typ A begründet auf einer anderen als der inhaltlichen Ebene, nämlich weshalb die Quelle verlässlich ist für die angeführte Regel "es kommt immer 0 heraus, weil wir das in der Grundschule so gelernt haben". Es liegen ca. 1400 Texte vor, von denen ca. 400 codiert sind. Die "händischen" Codierungen auf Basis eines sehr ausführlichen und in Teilen operationalisierten Codiermanuals wurden mindestens zu zweit durchgeführt mit nachfolgenden Konsensgesprächen.

In: P. Ebers, F. Rösken, B. Barzel, A. Büchter, F. Schacht & P. Scherer (Hrsg.),
Beiträge zum Mathematikunterricht 2024.
57. Jahrestagung der Gesellschaft für Didaktik der Mathematik. WTM.
https://doi.org/10.37626/GA9783959872782.0

KI-Einsatz - grundsätzliches Vorgehen

KI wird hier eingesetzt, um eine Einteilung von Texten in 4 Klassen vorzunehmen: R, P, A und S (Sonstige). Zunächst wird ein Language Model (LM) aus dem Textkorpus erstellt. Dieses Modell weist jedem Wort einen mehrdimensionalen Vektor zu. Es gibt Standardmodelle des Deutschen mit ca. 50 Dimensionen (nnlm-de-dim50, Bengio et al. 2003). Die Standardmodelle schnitten aber hier im Vergleich schlechter ab als das von uns an den spezifischen Daten selbst trainierte Language Model (LM).

Mit dem LM werden alle Texte in Tokens (jedes Wort ein Vektor) übersetzt und ein erster Teil (z. B. 50 %) trainiert in dieser Gestalt ein neuronales Netz. Hier hat man Freiheiten bei der Festlegung der Netz-Geometrie: Anzahl der Neuronen, Verknüpfung der Ebenen (meist dense layer), Verknüpfung zum Kontext (N-Gram). Um overfitting zu vermeiden, sind hohe Neuronenzahlen und ungewöhnliche Geometrien eher hinderlich für eine Verallgemeinerung des Modells. Ein systematischer bias ist durch Anpassungen der Geometrie nicht zu erwarten, da die Modelle ohnehin mit dem Evaluationsdatenset (s. u.) bewertet werden. Die erste Ebene eines Netzes besteht aus Zellen, die die Tokens eines jeden Textes aufnehmen. Die letzte Ebene des Netzes hat vier Zellen mit Zahlen zwischen 0 und 1, die grob als Wahrscheinlichkeiten, besser als Maß für das Vorliegen eines bestimmten Typs (R, P, A oder S) angesehen werden können. Mit einem Teil der Trainingsdaten werden die Kantengewichte des Netzes, die diese vier Endwerte steuern, so lange angepasst, bis das Netz möglichst oft dem korrekten Typ eines jeden Textes den höchsten Endwert zuweist. Wenn die Zellenanzahl zu groß ist, kann man die Trefferquote auf nahezu 100% steigern, ohne dass irgendeine allgemeine Struktur durch die KI erfasst würde. Dieses sogenannte overfitting wird dadurch vermieden, dass man einerseits das Netz schlank hält und andererseits seine Qualität an einem weiteren Teil (z. B. 25%) des Datensets, den Evaluationsdaten, misst. Als Cross-Check für die Auswertung bietet sich an, dieses Verfahren mehrfach mit verschiedenen Aufteilungen in Trainings- und Evaluationsdaten durchzuführen und zu vergleichen. Auf diese Weise werden mehrere Modelle untersucht und mittels ihrer Performanz auf den Evaluationsdatensätzen selektiert bis hin zum finalen Modell. Die Erkennungsraten bei Anwendung auf einen dritten "Testdatensatz" bewerten die Anwendung des fertigen Modells auf neue Daten. Dieser Testdatensatz (z. B. 25%) wird von Daten gespeist, die explizit weder zum Training noch beim Vergleich der Modelle eingesetzt wurden. Insbesondere die Erkennungsraten beim Testdatensatz dienen zum Bericht der Modellgüte (vgl. Xu & Goodacre 2018).

KI-Einsatz in der Studie

Die Bedeutung des aufwendigen preprocessing für die Modellgüte kann leicht unterschätzt werden. Zunächst werden Beschreibungen von nicht-lexikalen Elementen (Skizzen, Formeln), sowie die Rechtschreibung und Satzzeichen vereinheitlicht. Anschließend entfernt man sogenannte stopwords (An & Chen 2005). Stopwords sind vorher festgelegte Wörter, deren Verwendung keinen wesentlichen Beitrag für die inhaltliche Erfassung des Textes und somit für die Codierung der Eingabe leisten. Die Identifizierung relevanter stopwords muss jeweils domänenspezifisch vorgenommen werden (Sarica & Luo 2021). Danach werden ausgewählte Wortgruppen, z. B. Verben, auf Wortstämme reduziert und mit diesen ein Bigramm-LM unsupervised trainiert. Hierbei ist zu beachten, dass für das LM nicht nur die codierten, sondern bereits alle transkribierten Texte verwendet werden können. Die Kombination von unsupervised trainierten Modellen der transkribierten Texte mit supervised trainierten Modellen der bereits codierten Texte hat sich in unserer Studie bewährt und ist ein Schlüssel für den KI-Einsatz, denn dieser ist ja gerade dafür gedacht, mittels vergleichsweise weniger von Hand codierter Texte viele transkribierte Texte zu klassifizieren. Das anschließend mit den lediglich 400 codierten Texten entwickelte Klassifikationsmodell wies Erkennungsraten von über 80% auf. Da die Studie noch nicht abgeschlossen ist, können hier noch keine genaueren Angaben gemacht werden.

Diskussion

Ist KI grundsätzlich zur Codierung geeignet? Unser Beispiel spricht dafür. Natürlich führt die KI nicht zu exakt den gleichen Ergebnissen wie die händische Codierung. Differenzen zur händischen Codierung entstehen nicht nur aus Unvollkommenheiten der trainierten KI, sondern auch durch menschliche Fehleinschätzungen, die in die Trainingsdaten eingingen. Die Konsenscodierungsschleifen entstehen ja gerade durch die Graubereiche, bei denen die Codierenden eine Klassenzuweisung angesichts von Pro- und Contra-Argumenten abwägen müssen. Es wäre denkbar, dass KI dazu Hinweise geben könnte, welche Zuweisung am besten zu den anderen Strukturen passt - der KI könnte in dieser Hinsicht eine Aufgabe bei der Theoriebildung zukommen. Dies bedarf aber unbedingt der Transparenz und mehr noch der kritischen Kontrolle, inwieweit das Instrument (hier KI) die epistemischen Prozesse beeinflusst (Weiss 2019, S. 9f).

Ist es nützlich und sinnvoll, KI für Codierungen einzusetzen? Die trotz wenig codierten Texten gute Performanz der KI in unserer Beispielstudie spricht dafür, dass sich Zeit und damit menschliche Ressourcen einsparen lassen. Die "Tuchfühlung" mit dem Material geht dabei nicht verloren, da eine erhebliche Menge von Trainingsdaten ohnehin von Hand codiert werden muss.

Und mehr noch: Die Nachhaltigkeit der Ergebnisse qualitativer Textanalysen könnte durch die Erstellung von ML-Modellen (bei geeigneter Datensatzgröße) zur Textklassifizierung deutlich gesteigert werden. Denn diese Modelle könnten als Produkt wissenschaftlicher Textveröffentlichungen für andere Forschungsvorhaben bereitgestellt werden und so einen Anknüpfungspunkt für Folgeforschung bieten.

Kann KI Scheinobjektivität vorgaukeln? Das Modell ist wenig in menschlich fassbaren Kategorien zu durchschauen (bis auf Methoden der XAI), wenn auch formal zu verstehen und objektiv anzuwenden. Bei der Suche nach Antwort sollte bedacht sein, dass auch in klassischen Auswertungen unhinterfragt statistische Verfahren z. T. unsinnig angewandt werden (Problematik der p-Werte, statistische Verfahren, deren Sinn aus der vordigitalen Zeit stammt, publication bias, vgl. Open Science Collaboration, 2015). Auch Codiermanuale und Daten werden häufig nicht veröffentlicht. Eine KI-gestützte Auswertung könnte hierzu ein willkommenes Gegengewicht liefern.

Literatur

An, J. & Chen, Y.-P. (2005). Keyword extraction for text categorization. In *Proceedings of the 2005 International Conference on Active Media Technology, IEEE*, 556-561.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Conference on Fairness, Accountability, and Transparency (FAccT '21)*. https://doi.org/10.1145/3442188.3445922

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin. C. (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research 3*, 1137-1155. https://www.kaggle.com/models/google/nnlm

Fahse, C. (2017). Issues of a Quasi-Longitudinal Study on Different Types of Argumentation in the Context of Division by Zero. In T. Dooley & G. Gueudet (Hrsg.), *Proceedings of the Tenth Congress of the European Society for Research in Mathematics Education (CERME10)*, 147-154.

Open Science Collaboration (2015). Estimating the Reproducibility of Psychological Science. *Science 349*(6251), 943–943. DOI: 10.1126/science.aac4716

Sarica S. & Luo, J. (2021). Stopwords in technical language processing. *PLoS ONE 16*(8). https://doi.org/10.1371/journal.pone.0254937

Weiss, Y. (2019). Gedanken zur Digitalisierung des Mathematikunterrichts aus der Sicht des Werkzeugbegriffs. *International Cultural-historical Human Science, 15*(1), 53-74.

Xu, Y. & Goodacre, R. (2018). On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *Journal of Analysis and Testing 2*, 249–262.