Automatisierte Textanalyse bei der Personalauswahl – Potenziale und Grenzen

KATHARINA SIMBECK
Prof. Dr., Professorin für Betriebswirtschaftslehre und Controlling an der Hochschule für Technik und Wirtschaft Berlin

FINN FOLKERTS
Wiss. Mitarbeiter an der Hochschule für Technik und Wirtschaft Berlin

SHIRIN RIAZY
Wiss. Mitarbeiterin an der Hochschule für Technik und Wirtschaft Berlin

Unternehmen können heute mithilfe von Algorithmen eine Vorauswahl aus eingegangenen Bewerbungen treffen. Dabei setzen sie auch automatisierte Verfahren der Textanalyse durch künstliche Intelligenz ein. Die dahinterliegenden Routinen sind jedoch nicht nachvollziehbar und können unerwünscht diskriminierend wirken. Im Beitrag werden Ergebnisse eines Projekts vorgestellt, in dem die Voreingenommenheit deutschsprachiger Analysesysteme untersucht wurde.

Automatisierte Textanalyse durch künstliche Intelligenz

Textanalyse-Verfahren, die auf künstlicher Intelligenz (KI) basieren, werden bereits bei der automatisierten Vorauswahl von Bewerbungen eingesetzt (vgl. KPMG AG 2018). Hierfür bietet z. B. die deutsche Firma Precire einen vollständigen Workflow an, der neben der automatisierten Textanalyse des Lebenslaufs auch ein Telefoninterview mit einer Sprach-KI beinhaltet, die ein psychologisches Profil erstellt. Das Unternehmen Digital Minds aus Finnland wählt mit einem KI-gestützten System Bewerber/-innen nach einer Textanalyse ihrer E-Mail-Postfächer aus. Selbst Karrierenetzwerke wie LinkedIn und XING nutzen automatisierte Textanalysen, um ihren Mitgliedern potenziell interessante Arbeitgeber bzw. Arbeitnehmer/-innen vorzuschlagen.

Sentiment-Analysen stellen eine Unterkategorie der Textanalyse-Verfahren dar, sie ermitteln (automatisiert) die Stimmung eines Textes. In der Regel unterscheiden sie dabei positive und negative Stimmungen oder ordnen Sätze einer konkreten Emotion, z. B. Traurigkeit, zu. Diese Zuordnung erfolgt anhand zuvor gelernter Trainingsdaten, die bereits nach Emotionen (z. B. Wut oder Freude) klassifiziert wurden. Im Ergebnis kann das System für jeden Satz eine Stimmungslage und deren Intensität ermitteln (vgl. Liu 2012). Dem Satz *»Sie hat ihre Aufgaben stets zu unserer vollsten Zufriedenheit erfüllt.«* könnte mit einer Wahrscheinlichkeit von 98,4 Prozent eine positive Stimmungslage zugeordnet werden.

Künstliche Intelligenz lernt auch Vorurteile

KI-Systeme lernen anhand von Trainingsdaten. Somit ist die Qualität eines KI-Systems von der Qualität dieser Trainingsdaten abhängig. Die verwendeten Trainingstexte spiegeln in der Regel die in der Gesellschaft vorhandenen Vorurteile und Stereotype wider, die Ergebnisse des Systems vermitteln dann eine Scheinobjektivität (vgl. Osoba/Welser 2017).

Dies zeigten Caliskan/Bryson/Narayanan (2017) in einem Experiment. Sie trainierten ein System durch maschinelles Lernen an Internettexten (Textkorpus »Common Crawl«). Sie nutzten dabei den verbreiteten Ansatz der »word embeddings«: Treten Wörter besonders häufig im selben Satz auf, assoziiert das System einen Zusammenhang. Das so trainierte System replizierte klassische Stereotype gegenüber Frauen oder afroamerikanischen Vornamen. Es assoziierte Frauen z. B. mit Familie und Kunst, während Männer eher mit karrierebezogenen Themen und Mathematik verknüpft wurden.

Zu einem ähnlichen Ergebnis kommen Kiritchenko/Mohammad (2018) bei einem Test von über 200 Sentiment-Analysesystemen. Sie erstellten einen englischsprachigen Korpus mit Testsätzen, die Emotionen aus den Kategorien Freude, Angst, Wut oder Traurigkeit beinhalteten. In diesen Testsätzen wurde das Subjekt in Geschlecht und Namen (unterschiedlicher Herkunft) variiert. Bei mehr als 75 Prozent der untersuchten Systeme wurden die Emotio-

nen eines Satzes in Abhängigkeit vom Subjekt unterschiedlich bewertet.

Die Erkenntnis, dass KI-basierte Textanalyseverfahren nicht objektiv sind, ist für das Personalmanagement aus mehreren Gründen relevant: Zum einen werden die im Personalwesen eingesetzten Systeme mit den gleichen Methoden des maschinellen Lernens trainiert. Sprachmodelle und Algorithmen können im Personalwesen eingesetzt werden, auch wenn sie nicht speziell für die diskriminierungsfreie Anwendung entwickelt wurden. Es widerspricht aber den Erwartungen der Nutzer/-innen von solchen automatisierten Systemen, wenn diese nicht konsistent Gleiches gleich bewerten. Zum anderen ist die Bewertung von Texten durch die Systeme aus Nutzersicht intransparent, das System ist eine Blackbox. Einflussfaktoren auf die Klassifizierung sind teilweise sogar technisch nicht nachvollziehbar.

Untersuchung eines deutschsprachigen Sentiment-Analysesystems

Aufgrund der großen Verfügbarkeit englischsprachiger Trainingsdaten für Sentiment-Analysesysteme ist anzunehmen, dass deutschsprachige Systeme eine tendenziell schlechtere Qualität haben (vgl. Narr/Hulfenhaus/Albayrak 2012). Fraglich ist, ob deutschsprachige Sentiment-Analysesysteme ebenfalls gesellschaftliche Stereotype reproduzieren. Dazu wurde im Rahmen eines Projekts an der Hochschule für Technik und Wirtschaft (HTW) Berlin (vgl. Infokasten) ein Experiment analog zu dem von Kiritchenko/Mohammad (2018) repliziert.

Im ersten Schritt wurden Mustersätze mit jedem Vornamen aus Tabelle 1 und jedem Adjektiv aus Tabelle 2 kombiniert (z. B. *»Die Situation macht Leon wütend.«*). Daraus entstand ein Korpus von 612 deutschen Sätzen. Diese wurden danach von dem Sentiment-Analyse-System als positiv, negativ oder neutral mit einer entsprechenden Wahrscheinlichkeit bewertet.* Die Herangehensweise könnte damit der eines Anbieters ähneln, welcher E-Mail-Postfächer von Bewerberinnen und Bewerbern vollständig analysiert.

Tendenziell war das System in der Lage, Emotionen richtig zu klassifizieren. Die Abweichungen in der Klassifizierung von Sätzen mit Subjekten männlichen bzw. weiblichen Geschlechts respektive Vornamen (deutscher oder türkischer/arabischer Herkunft) wurden auf statistisch signifikante Unterschiede (t-test) untersucht. Vom System wiederholt falsch klassifizierte Sätze wurden von der Auswertung ausgeschlossen.

Tabelle 1
Im Experiment verwendete Vornamen

Deutsche Vornamen		Türkische/arabische Vornamen	
männlich	weiblich	männlich	weiblich
Ben	Anna	Ali	Elif
Leon	Emma	Can	Hiranur
Noah	Hannah	Tarek	Zeynep

Tabelle 2
Im Experiment verwendete Adjektive

Angst	Freude	Traurigkeit	Wut
ängstlich	glücklich	trist	wütend
unruhig	fröhlich	traurig	zornig
fürchterlich	großartig	deprimierend	nervig
schrecklich	wundervoll	herzzerreißend	lästig

Das Projekt »Diskriminiert durch Künstliche Intelligenz« (DiKI)

Das Projekt untersucht, ob der Einsatz von künstlicher Intelligenz im Personalmanagement zu Diskriminierung beitragen oder diese reduzieren kann. Hierzu wurde in einem ersten Schritt die Voreingenommenheit des Systems (im Machine Learning als »bias« bezeichnet) in Bezug auf Geschlecht und Vornamen untersucht.

Die Untersuchung wurde mit dem Sentiment-Analyse-System von ParallelDots (www.paralleldots.com) durchgeführt und analysiert 612 deutsche Sätze.

Ein umfangreicheres Experiment mit weiteren Systemen ist bereits geplant.

Das Projekt wird gefördert durch die Hans-Böckler-Stiftung.

Nähere Informationen: http://iug.htw-berlin.de/?page_id=92

»Ali ist wütender als Leon«

In der Abbildung sind die Ergebnisse zur Herkunft des Vornamens dargestellt. Jeder Punkt repräsentiert einen Satz und zeigt die berechnete Wahrscheinlichkeit für die prognostizierte Ausprägung einer Emotion an. Die Raute gibt den Mittelwert für jede Punktwolke an.

Das Ergebnis der Untersuchung zeigt beim Vergleich der Sätze von Personen mit deutschen und türkischen oder arabischen Vornamen deutliche Unterschiede in den Bewertungen. Sätze von Personen mit ausländischen Vornamen assoziiert das System mit signifikant höherer Wahrscheinlichkeit mit Wut, Angst und Traurigkeit. Zwischen weiblichen und männlichen Subjekten gab es meist keinen

* Durchgeführt vom cand. M.Sc. (Wirtschaftsinformatik) Laser Karasahin.

Abbildung

Bewertung der Emotionen nach Herkunft des Vornamens

n = 362

signifikanten Unterschied, nur in den Kategorien Angst und Traurigkeit tendierte das System dazu, Frauen höhere Werte zuzuschreiben.

Aufschlüsse zu den Bewertungsgrundlagen sind gefragt

Deutschsprachige Sätze, die sich nur in der ethnischen Zugehörigkeit des Subjekts unterscheiden, werden vom beispielhaft geprüften Sentiment-Analysesystem unterschiedlich bewertet.

Systeme wie das von uns getestete sind für die Nutzer/-innen sog. Blackbox-Systeme, die Gründe für die unterschiedlichen Bewertungen sind intransparent. Es ist unklar, wie und womit das System trainiert wurde. Eine mögliche Ursache könnte z. B. sein, dass die Trainingsdaten überdurchschnittlich viele Sätze mit türkischen oder arabischen Namen in einem negativen Kontext enthalten. Eine andere Erklärung könnte sein, dass seltenere Vornamen allgemein zu anderen Ergebnissen führen. Hierfür wäre es interessant, in den weiterführenden Experimenten auch seltene deutsche Vornamen zu testen, wie z. B. Adalbert oder Edeltraut.

Der Einsatz von automatisierten Textanalysesystemen im Personalmanagement birgt Diskriminierungspotenziale, weil die verwendeten proprietären Algorithmen und Trainingsdaten innerhalb der Blackbox-Systeme schwer zu erkennen sind. ◀

Literatur

Caliskan, A.; Bryson, J. J.; Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. In: Science 356 (2017), S. 183–186 – DOI: 10.1126/science.aal4230

Kiritchenko, S.; Mohammad, S. M.: Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. 2018 – URL: http://arxiv.org/pdf/1805.04508v1 (Stand: 20.03.2019)

KPMG AG (Hrsg.): Wertschöpfung neu gedacht. 2018 – URL: https://hub.kpmg.de/hubfs/LandingPages-PDF/KPMG_Studie_Artif_Intelligence_April_2017_BF_SEC.pdf (Stand: 20.03.2019)

Liu, B.: Sentiment analysis and opinion mining (Synthesis lectures on human language technologies, lecture #16). San Rafael, Calif. 2012

Narr, S.; Hulfenhaus, M.; Albayrak, S.: Language-independent twitter sentiment analysis. Tagungsbeitrag Knowledge discovery and machine learning (KDML) LWA Dortmund 2012 – URL: www.dai-labor.de/fileadmin/files/publications/narr-twittersentiment-KDML-LWA-2012.pdf (Stand: 20.03.2019)

Osoba, O.; Welser, W.: An intelligence in our image. The risks of bias and errors in artificial intelligence (Research report, RR-1744-RC). Santa Monica, Calif. 2017