

Prädiktive Algorithmen zur Persönlichkeitsprognose auf Basis von Social-Media-Daten

Von **Prof. Dr. Ricardo Büttner** (Hochschule Aalen)

Weil Persönlichkeitsmerkmale stabile Verhaltensprädiktoren sind (Costa & McCare, 1992), wäre es nützlich, wenn man Persönlichkeitsmerkmale vollautomatisch mittels prädiktiver Algorithmen prognostizieren könnte. Der Personalbereich könnte dann die Persönlichkeitsinformationen der Bewerber bspw. im Rahmen der Eignungsfeststellung bei der Stellenbesetzung nutzen (Buettner, 2014a). Die Bewerber könnten sich jedoch ebenfalls über zu ihrer Persönlichkeit passende Arbeitgeber informieren und E-Recruiting-Plattformanbieter wären in der Lage, ihre Matchingalgorithmen zu verbessern (Buettner, 2016g).

Der vorliegende Beitrag stellt empirische Ergebnisse automatisierter Prognosen der Big-Five-Persönlichkeitsmerkmale durch maschinelles Lernen auf Basis von Xing-Daten vor und zeigt deren konzeptionelle Anwendungsmöglichkeiten für den Personalbereich.

Digitale Fußabdrücke

Die Persönlichkeit eines Menschen beeinflusst maßgeblich sein Denken, Fühlen und Verhalten. Das zeigt sich in verschiedenen privaten und beruflichen Lebensbereichen. Da das Verhalten auf Social-Media-Plattformen wie Facebook und Twitter digital aufgezeichnet und gespeichert wird, hinterlassen deren Nutzer persönlichkeitspezifische Fußabdrücke (Buettner, 2016e). Es konnte in einer Reihe von Korrelationsanalysen gezeigt werden, dass schwache, aber stabile Zusammenhänge zwischen Persönlichkeitsmerkmalen und den digitalen Fußabdrücken bei der Social-Media-Nutzung bestehen (vgl. Abb. 1). So bestehen stabile Zusammenhänge zwischen spezifischen Social-Media-Nutzungsattributen (Anzahl Kontakte, Anzahl Nachrichten, Anzahl Bilder etc.) und den Big-Five-Persönlichkeitsmerkmalen. Diese Korrelationen wurden kohärent über unterschiedliche Social-Media-Plattformen hinweg gefunden, insb. Facebook, Twitter, LinkedIn, Youtube, Myspace, Renren (Buettner, 2016h).

Prädiktive Algorithmen zur Persönlichkeitsprognose

Prädiktive Algorithmen werden durch maschinelles Lernen gefunden. Hierbei lassen sich Muster in den Daten entdecken

und evaluieren, die durch einfache Verfahren wie Regressionsanalysen nicht auffindbar sind. Zu den häufig verwendeten Standardansätzen maschinellen Lernens zählen hauptsächlich Entscheidungsbäume, künstliche neuronale Netze und Support Vector Machines.

Entscheidungsbäume dienen als geordnete und gerichtete Graphen mit einer Wurzel der Klassifikation von Datenobjekten. Ältere Entscheidungsbaumverfahren können Bäume nur auf Basis diskreter Attribute erzeugen. Neuere CART-Verfahren (classification and regression trees) können zudem reelle Attribute verarbeiten und nutzen eine Reihe von Optimierungsstrategien wie bspw. das Rückschneiden (pruning) von Baumabschnitten mit geringer prädiktiver Kraft. Zur CART-Familie zählt auch das ID3-Verfahren (iterative dichotomiser 3), das entropiebasiert dasjenige Attribut mit dem höchsten Informationsgewinn auswählt, um an der Stelle einen neuen Baumknoten zu generieren.

Künstliche neuronale Netze orientieren sich am Vorbild der Nervenzellverbände und abstrahieren deren Informationsverarbeitung auf einem oder mehreren Computern. Dabei erfolgt eine Simulation der biologischen Vorgänge wie bspw. das Aktivieren von Neuronen oder die Synapsenveränderungen aufgrund von Lernvorgängen. Neben Assoziativspeichern werden häufig mehrschichtige vorwärtsgerichtete künstliche neuronale Netze (azyklische, gerichtete Graphen) zur Mustererkennung oder für prädiktive Modelle eingesetzt.

Support Vector Machines separieren klassenunterschiedliche Objekte durch die Einpassung von Hyperebenen im Vektorraum, in welchem jedes Objekt durch einen Vektor repräsentiert wird. Die Hyperebenen werden so gelegt, dass um die Klassengrenzen herum ein möglichst großer Bereich frei von Objekten bleibt. Support Vector Machines wurden in den letzten Jahren ebenfalls methodisch erweitert. So bieten Kernelfunktionen bspw. die Möglichkeit, auch an sich nicht linear trennbare Objekte zu klassifizieren.

Um die unterschiedlichen Ansätze hinsichtlich ihrer Güte miteinander vergleichen zu können, hat Buettner (2016g) das Caret-Paket in der Statistiksoftware R x64 3.2.2 (R Core Team 2015) genutzt. Das Caret-Paket stellt ca. 200 unterschiedliche Algorithmen zum maschinellen Lernen bereit.

ABSTRACT

Forschungsfrage: Lassen sich Social-Media-Daten zur Prognose von Persönlichkeitsmerkmalen nutzen?

Methodik: Vergleich prädiktiver Algorithmen basierend auf maschinellen Lernverfahren, die einen empirisch erhobenen Datensatz der Social-Media-Plattform Xing auswerten.

Praktische Implikationen: Prädiktive Algorithmen zur Persönlichkeitsprognose lassen sich insbesondere in E-Recruiting-Systemen bei Kosten- und Zeitersparnis einsetzen.

Die Algorithmen wurden mittels eines Xing-Datensatzes evaluiert. Dazu wurden berufsbegleitend Studierende hinsichtlich ihrer Nutzung der Social-Media-Plattform Xing befragt. Von den 760 Teilnehmern nutzten 395 die Xing-Plattform. Von diesen 395 Xing-Nutzern (189 Frauen, 206 Männer) waren 4 Personen unter 21 Jahre, 259 zwischen 21 und 30 Jahre, 92 zwischen 31 und 40 Jahre, 32 zwischen 41 und 50 Jahre, 7 zwischen 51 und 60 Jahre und schließlich eine Person über 60 Jahre alt. Das Nutzungsverhalten wurde mittels der in Abbildung 2 dargestellten

Items erfasst. Die Big-Five-Persönlichkeitsmerkmale wurden mit dem Ten Item Personality Inventory (TIPI) von Gosling et al. (2003) auf einer 5-Punkt-Likert-Skala erhoben.

Im Ergebnis zeigte sich, dass sich Entscheidungsbäume am besten zur Persönlichkeitsprognose eigneten (Buettner 2016g). Unter den verschiedenen Entscheidungsbaumansätzen erzielte das C5.0-Verfahren (Kuhn und Johnson, 2013, S. 394ff.) die besten Ergebnisse (Treffergenauigkeiten: C5.0 68,4%; Random Forest 61,9%; C4.5 J48 54,2 %; vgl. Buettner, 2016c, 2016g). Das

Abb. 1: Positive (+) und negative (-) Zusammenhänge zwischen Social-Media-Nutzungsattributen und den Big-Five-Persönlichkeitsmerkmalen

Indikatorgruppe	Indikator	Offenheit	Gewissenhaftigkeit	Extraversion	Verträglichkeit	Neurotizismus
Statische Profilinformationen	Interessenanzahl	+				
	Gruppenanzahl					+
	Anzahl Profilbilder			+	+	-
Dynamische Profilinformationen	Häufigkeit Profilbildänderungen	+		+		+
	Häufigkeit neuer Bilder	+		+	+	
Nutzungsintensität	Loginhäufigkeit			+	-	
	Nutzungsdauer		-	+		+
	Besuchshäufigkeit anderer Profile		-	+	+	
Kommunikation	Anzahl gesendeter Nachrichten			+	+	-
	Anzahl eigener Kommentare			+	+	+
	Anzahl fremder Kommentare			+	+	
	Anzahl Fauxpas		-		-	
	Anzahl Blogeinträge	+				
	Anzahl emotionaler Inhalte					+
Netzwerk	Kontaktanzahl	+	+	+	+	-
	Weitere Zentralitätsmaße	+			+	

Quelle: in Anlehnung an Buettner 2016h

Abb. 2: Items zur Xing-Nutzung

#	Item	Skala	MW	SD
I ₁	How often do you use Xing?	[1-never..5-daily]	2.96	1.16
I ₂	How often do you use the Xing jobsearch function?	[1-never..5-daily]	1.70	0.87
I ₃	How often do you use the Xing blogging function?	[1-never..5-daily]	1.24	0.51
I ₄	How often do you use the Xing messaging function	[1-never..5-daily]	2.36	0.90
I ₅	How often do you use the Xing event organization function?	[1-never..5-daily]	1.14	0.41
I ₆	How often do you use the Xing event participation function?	[1-never..5-daily]	1.27	0.48
I ₇	How often do you use the Xing advantageous offers function?	[1-never..5-daily]	1.29	0.56
I ₈	Have you filled out your educational background on Xing?	[1-no/2-yes]	1.32	0.47
I ₉	Have you filled out your work experience on Xing?	[1-no/2-yes]	1.95	0.21
I ₁₀	Have you filled out your organizations on Xing?	[1-no/2-yes]	1.68	0.47
I ₁₁	Have you filled out your interests on Xing?	[1-no/2-yes]	1.73	0.44
I ₁₂	Have you filled out your awards on Xing?	[1-no/2-yes]	1.32	0.47
I ₁₃	Have you filled out your language skills on Xing?	[1-no/2-yes]	1.86	0.35
I ₁₄	Have you filled out your haves on Xing?	[1-no/2-yes]	1.73	0.44
I ₁₅	Have you filled out your wants on Xing?	[1-no/2-yes]	1.69	0.46
I ₁₆	Have you filled out your about me information on Xing?	[1-no/2-yes]	1.52	0.50
I ₁₇	Do you have a Xing premium membership?	[1-no/2-yes]	1.26	0.44
I ₁₈	How many Xing contacts do you have?	number	121	160

Quelle: in Anlehnung an Buettner 2016h

C5.0-Verfahren ist der verbesserte Nachfolger des C4.5-Entscheidungsbaumalgorithmus (Quinlan, 1993), der wiederum eine Erweiterung von Quinlans früheren Entropie-basierten ID3-Verfahren ist. C5.0 gehört zu den stärksten Verfahren zur Findung prädiktiver Algorithmen.

Das n=395 Sample wurde in eine Trainingspartition ($n_T=261$) und eine Evaluierungspartition ($n_E=134$) geteilt. Danach erfolgte ein systematischer Vergleich der Algorithmen hinsichtlich ihrer Prognosequalität (Treffergenauigkeit, Sensitivität, Spezifität, Relevanz und Trennfähigkeit). Die zu prognostizierende Variable war das jeweilige Persönlichkeitsmerkmal, das jeweils balanciert auf Basis des Mittelwerts in die Ausprägungen „hoch“ und „niedrig“ gesplittet wurde (Bsp. „Offenheit hoch“ oder „Offenheit niedrig“).

Im Ergebnis zeigte sich, dass das C5.0-Verfahren (Kuhn und Johnson 2013, S. 394ff.) die besten Ergebnisse erzielte. Die Gü-

temaße des prädiktiven C5.0-Entscheidungsbaumes zur Prognose der Big-Five-Persönlichkeitsmerkmale sind in Abbildung 3 dargestellt.

Die Treffergenauigkeit (Vertrauenswahrscheinlichkeit) entspricht dem Anteil aller Objekte, die korrekt klassifiziert wurden. Die Sensitivität (Richtig-Positiv-Rate, Empfindlichkeit, Trefferquote) entspricht dem Anteil der richtigerweise als positiv klassifizierten Objekte an der Gesamtheit der tatsächlich positiven Objekte. Die Spezifität (Richtig-Negativ-Rate) entspricht dem Anteil der richtigerweise als negativ klassifizierten Objekte an der Gesamtheit der in Wirklichkeit negativen Objekte. Die Relevanz (positiver Vorhersagewert, Wirksamkeit, Genauigkeit, positiver prädiktiver Wert) zeigt den Anteil der richtigerweise als positiv klassifizierten Ergebnisse an der Gesamtheit der tatsächlich positiv klassifizierten Ergebnisse. Unter Trennfähigkeit (negativer Vorhersagewert)

versteht man den Anteil der richtigerweise als negativ klassifizierten Ergebnisse an der Gesamtheit der tatsächlich negativ klassifizierten Ergebnisse.

Der prädiktive Nutzen (Erhöhung der Treffergenauigkeit) des gefundenen C5.0-Entscheidungsbaumes zur Persönlichkeitsprognose beträgt zwischen 31.4 und 46.2 Prozent. Das heißt, dass die Xing-Plattform tatsächlich wertvolle Daten zur Persönlichkeitsprognose enthält.

Anwendung im elektronischen Recruiting

Bei der Rekrutierung neuer Mitarbeiterinnen und Mitarbeiter ist es notwendig, eine Prognose für die spätere Passung der Bewerberin bzw. des Bewerbers auf die jeweils vakante Stelle abzugeben. Eine hohe Kandidatenpassung führt zu höherer Arbeitsleistung, höherer Arbeitszufriedenheit, höherem Commitment und längerer Beschäftigungsdauer (Kristof, 1996; Kristof-Brown et al., 2005; Verquer et al., 2003). Zur Prognose der Kandidatenpassung machen sich die Personalmitarbeiter und die Führungskräfte bisher ein Bild durch den Einsatz traditioneller Personalauswahlinstrumente (Lebenslaufanalyse, persönliches Vorstellungsgespräch etc.).

Die Kandidatenpassung setzt sich aus drei Teilbereichen zusammen (Edwards, 1991; Kristof, 1996; Werbel & Johnson, 2001): 1. die Persönlichkeit der Bewerberin bzw. des Bewerbers sollte zur Unternehmenskultur passen (person-organization fit, P-O), 2. die sozialen Fähigkeiten, Rollen und Kommunikationsstile der Bewerberin sollte in die Arbeitsgruppe passen (person-group fit, P-G) und 3. Fähigkeiten, Fertigkeiten und Wissen der Bewerberin sollten auf die konkrete Stelle passen (person-job fit, P-J).

Existierende E-Recruiting-Lösungen und Job-Empfehlungssysteme fokussieren hauptsächlich nur auf das Matching zwischen den Stellenanforderungen und den Bewerberfähigkeiten (P-J fit) und decken damit lediglich einen der insgesamt drei Passungsteilbereiche ab. Genau hier liegt jedoch ein enormes Potenzial für zukünftige E-Recruiting-Lösungen und Job-Empfehlungssysteme (Buettner, 2014b).

Denn, wie oben gezeigt wurde, enthalten Social-Media-Plattformen wie Xing entsprechende Daten, die für Persönlichkeitsprognosen genutzt werden können. Damit ließe sich – bei Vorliegen der Unternehmenskulturmerkmale – der person-organization fit automatisiert abschätzen.

Damit könnte der person-organization fit vollautomatisch evaluiert werden, wodurch Personalmitarbeiter und Führungskräfte Zeit und Kosten sparen.

Weitere Anwendungsmöglichkeiten im Personalbereich

Über die Persönlichkeitsprognose hinaus lassen sich aus großen und komplexen Datenmengen ebenfalls Informationen zur Organisationskultur, den vorhandenen Kommunikationsstilen und Gruppenrollen ableiten (Buettner, 2014a). Kommunikationsstile lassen sich bspw. auf Basis der Häufigkeiten verwendeter Schlüsselwörter mittels Text Mining ableiten. Zudem sind sowohl Kommunikationsstile als auch potenzielle Gruppenrollen Dispositionen von Persönlichkeitsmerkmalen.

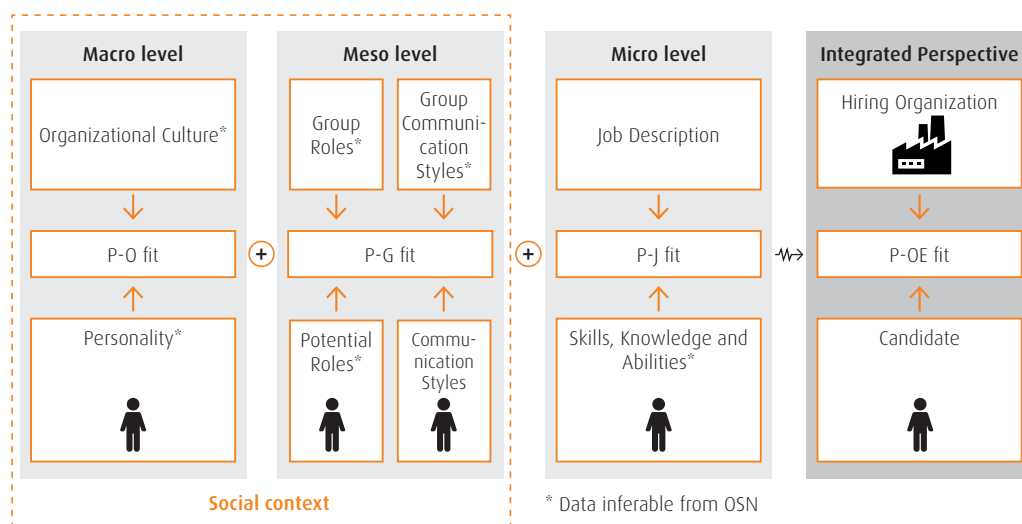
Deshalb könnten über die automatisierte Persönlichkeitsprognose (als ersten Teil des person-organization fits) hinaus auch vollautomatisierte Prognosen zum gesamten person-organization fit und zum person-group fit mittels prädiktiver Algorithmen durchgeführt werden (vgl. Abb. 4).

Abb. 3: Qualitätsmaße des C5.0-Entscheidungsbaumes

	Treffergenauigkeit	Sensitivität	Spezifität	Relevanz	Trennfähigkeit
Offenheit	0,731	0,487	0,832	0,543	0,798
Gewissenhaftigkeit	0,657	0,759	0,548	0,659	0,654
Extraversion	0,672	0,641	0,700	0,661	0,681
Verträglichkeit	0,664	0,609	0,714	0,661	0,667
Neurotizismus	0,694	0,625	0,757	0,702	0,757
Ø	0,684	0,624	0,710	0,645	0,711

Quelle: Buettner 2016g

Abb. 4: Social-Media-E-Recruiting-Framework



Quelle: Buettner, 2014a

Durch die breite Anwendung des Social-Media-E-Recruiting-Frameworks (Buettner, 2014a) könnte die Arbeitszufriedenheit und die Arbeitsleistung der Belegschaft gesteigert werden, weil gut zur Unternehmung passende Kandidaten schneller rekrutiert werden (Kristof, 1996; Kristof-Brown et al., 2005; Verquer et al., 2003).

Das sind insgesamt sehr vielversprechende Entwicklungen innerhalb des Electronic Human Resource Managements an der Schnittstelle zur Wirtschaftsinformatik.

Risiken für Arbeitnehmer

Der zunehmende Einsatz prädiktiver Algorithmen bei verbesserter Mitarbeiterdatenlage birgt jedoch auch Risiken:

In erster Linie bestehen Datenschutzbedenken beim Einsatz prädiktiver Algorithmen zur Persönlichkeitsprognose. Deshalb ist im Bundesdatenschutzgesetz (derzeit noch) geregelt, dass Einzelentscheidungen wie bspw. die Ablehnung einer Bewerberin oder eines Bewerbers nicht ausschließlich auf die Anwendung automatisierter Algorithmen fußen darf, die Persönlichkeitsmerkmale verarbeiten (§ 6a BDSG).

Es besteht die Gefahr, dass Bewerber oder bestehende Mitarbeiter hinsichtlich ihrer Persönlichkeit diskriminiert werden, weil der Arbeitgeber vollautomatisierte Persönlichkeitsanalysen durchführen könnte. Das ist ethisch und rechtlich beden-

lich bzw. verboten, technisch jedoch möglich. So könnten in Zukunft Unternehmen auch schneller die Kandidaten identifizieren, die nicht mehr so gut zu ihnen passen, um sich schneller von ihnen zu trennen und durch besser passenden Mitarbeitern auszutauschen.

Die Unternehmen könnten mit schnellen und kostengünstigen prädiktiven Algorithmen ihre Arbeitnehmer sogar permanent und in Echtzeit hinsichtlich ihrer Persönlichkeit, und darüber hinaus hinsichtlich ihrer Arbeitsleistung/Leistungsfähigkeit, ihres Kommunikationsverhaltens, ihrer politischen Einstellungen oder ihrer Einstellungen dem Arbeitgeber gegenüber durchleuchten („Gläserner Mitarbeiter“). Das würde den unproduktiven Druck auf Arbeitnehmer verstärken.

Zusammenfassung

In dem Beitrag wurde gezeigt, dass Menschen persönlichkeitskorrelierte digitale Fußabdrücke bei ihrer Social-Media-Nutzung hinterlassen, die mittels prädiktiver Algorithmen wiederum zur Persönlichkeitsprognose genutzt werden können. Mithilfe eines Xing-Datensatzes wurde dargestellt, dass insbesondere entscheidungsbaumbasierte Verfahren wie der C5.0-Algorithmus in der Lage sind, substanzielle Prognoseergebnisse zu liefern (prädiktiver Nutzen zwischen 31.4 und 46.2%). Es wurde weiterhin umrissen, wie mit prädiktiven AI-

gorithmen der person-organization fit und der person-group fit geschätzt werden können. Schließlich wurden Arbeitnehmerrisiken durch den Einsatz solcher Algorithmen genannt.

Es bleibt jedoch abzuwarten, inwieweit prädiktive Algorithmen mit etablierten Personalauswahlinstrumenten mithalten können und wenn dies tatsächlich der Fall sein sollte, ob Personalverantwortliche prädiktive Algorithmen tatsächlich einsetzen werden. Wie Biemann und Weckmüller (2012) gezeigt haben, ist „in keiner anderen personalwirtschaftlichen Teilfunktion ... der Unterschied zwischen wissenschaftlichen Erkenntnissen und Einschätzung von Personalverantwortlichen so groß wie bei der Bewertung der Nützlichkeit und Qualität von Personalauswahlinstrumenten“ (S. 49).



PROF. DR. RICARDO BÜTTNER

Professur für Data Science an der Fakultät für Betriebswirtschaft

Hochschule Aalen

E-Mail: ricardo.buettner@hs-aalen.de

www.prof-buettner.com

LITERATURVERZEICHNIS

- Biemann, T./Weckmüller, H. (2012):** Methoden der Personalauswahl: Was nützt? *PERSONALquarterly* 46(1): 46-49.
- Buettner, R. (2014a):** A Framework for Recommender Systems in Online Social Network Recruiting. In: HICSS 2014 Proceedings: 47th Hawaii International Conference on System Sciences (HICSS-47), Januar 6-9, 2014, Big Island, Hawaii, pp. 1415-1424.
- Buettner, R. (2014b):** efficientRecruiting 2.0 - Effizientes Recruiting von Fachkräften im Web 2.0. In BITKOM KnowTech 2014 Proceedings, October 15-16, 2014, Hanau, Germany, pp. 119-127.
- Buettner, R. (2016c):** Mining a user's personality from social media data: A comparison between the random forest and the C4.5 J48 based approach. In VHB 2016 Proceedings: 78. Wissenschaftliche Jahrestagung des Verbandes der Hochschullehrer für Betriebswirtschaft (VHB), May 18-20, 2016, Munich, Germany.
- Buettner, R. (2016e):** Personality as a predictor of business social media usage: An empirical investigation of XING usage patterns. In PACIS 2016 Proceedings: 20th Pacific Asia Conference on Information Systems (PACIS), June 27 - July 1, Chiayi, Taiwan.
- Buettner, R. (2016g):** Innovative Personality-based Digital Services. In PACIS 2016 Proceedings: 20th Pacific Asia Conference on Information Systems (PACIS), June 27 - July 1, Chiayi, Taiwan.
- Buettner, R. (2016h):** Predicting user behavior in electronic markets based on personality-mining in large online social networks: A personality-based product recommender framework. *Electronic Markets: The International Journal on Networked Business*, 2016, pp. 1-19.
- Costa, P. T./McCrae, R. R. (1992):** Revised NEO personality inventory (NEO-PI-R) and the NEO Five-Factor inventory (NEO-FFI): Professional manual. PAR, Odessa, FL, USA.
- Edwards, J. R. (1991):** Person-job fit: A conceptual integration, literature review, and methodological critique. *International review of industrial and organizational psychology*. In: *International review of industrial and organizational psychology*. Cooper CL, Robertson IT (Ed.). Vol. 6. Oxford: John Wiley & Sons, pp. 283-357.
- Gosling, S. D./Rentfrow, P. J./Swann Jr., W. B. (2003):** A very brief measure of the Big-Five personality domains. *Journal of Research in Personality* 37 (6), 504-528.
- Kuhn, M./Johnson, K. (2013):** *Applied Predictive Modeling*. New York: Springer.
- Kristof, A. L. (1996):** Person-organization fit: An integrative review of its conceptualizations, measurement, and implications. *Personnel Psychology* 49 (1): 1-49.
- Kristof-Brown, A. L./Zimmerman, R. D./Johnson, E. C. (2005):** Consequences of individuals' fit at work: A meta-analysis of person-job, person-organization, person-group, and person-supervisor fit. *Personnel Psychology* 58 (2): 281-342.
- Quinlan, J. R. (1993):** *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.
- R Core Team (2015):** *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.
- Verquer, M. L./Beehr, T. A./Wagner, S. H. (2003):** A meta-analysis of relations between person-organization fit and work attitudes. *Journal of Vocational Behavior* 63 (3): 473-489.
- Werbelt, J. D./Johnson, D. J. (2001):** The Use of Person-Group Fit for Employment Selection: A Missing Link in Person-Environment Fit. *Human Resource Management* 40 (3): 227-240.

SUMMARY

Research question: Is it possible to predict personality traits from social media data?

Methodology: Comparison of machine learning based predictive algorithms based on an empirical social media dataset from Xing.

Practical implications: Algorithms for predicting personality traits are useful in eRecruiting systems saving costs and time.