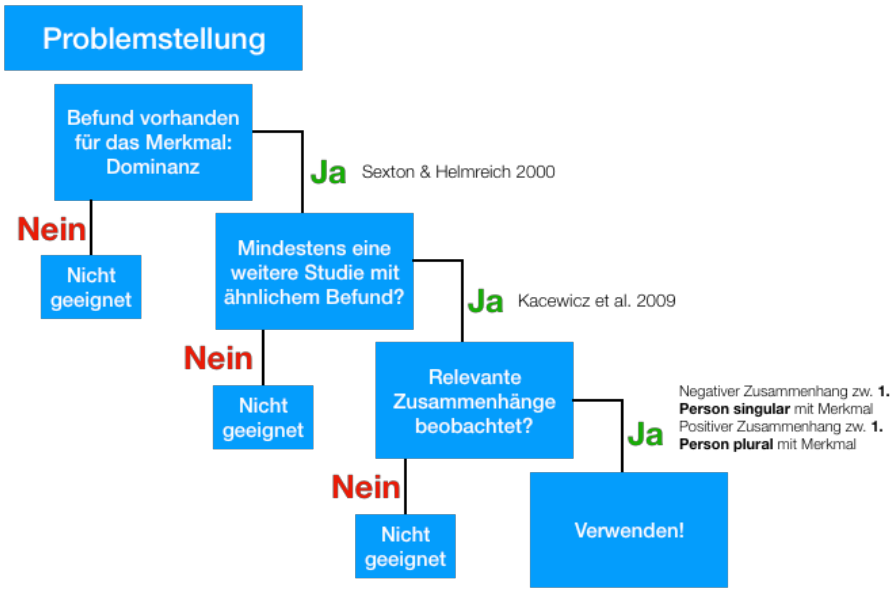
Menschen zuverlässig verstehen

Wie 100 Worte die höchsten Qualitätsstandards in der Textanalyse erreicht

Von der Wissenschaft zur Anwendung – Umsetzung wissenschaftlicher Befunde zur Entwicklung einer regelbasierten Textanalyse

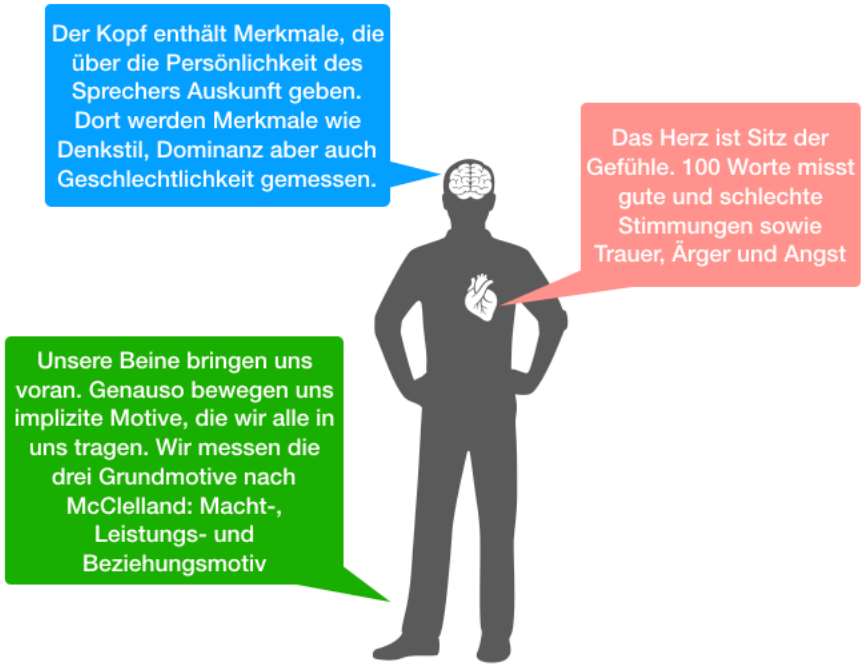
Die Häufigkeit, mit der wir bestimmte Worte und Wortkombinationen benutzen, verrät etwas über die Persönlichkeit. Diesen Zusammenhang, den viele auch intuitiv vermuten würden, konnte in vielen wissenschaftlichen Studien gezeigt und repliziert werden (Tausczik et al. 2010 für einen Überblick). Bei der Entwicklung der 100 Worte Analyse wurden zahlreiche psychologische Studien berücksichtigt an denen Forscher aus verschiedenen Disziplinen beteiligt waren. In ihren Studien haben die Forscher Hypothesen aufgestellt, wie bestimmte Sprachelemente mit der Persönlichkeit von Menschen in Verbindung stehen könnten. Diese prüften die Forscher dann mit experimentellen Methoden und konnten ihre Hypothesen so belegen. Die 100 Worte Textanalyse basiert auf diesen Befunden und wir entwickelten daraus eine regelbasierte Anwendung. Bevor wissenschaftliche

Befunde in die 100 Worte Textanalyse eingeflossen sind, fand jedoch eine kritische Prüfung statt. Das war nötig, denn nur so konnte eine hohe Qualität sichergestellt werden. Befunde wurden für unsere Anwendung ausgewählt, wenn sie einen relevanten Beitrag zur Erklärung eines Persönlichkeitsmerkmals leisteten (=statistisch signifikanter[2] Zusammenhang zwischen Sprachelement und Persönlichkeitsmerkmal mit einem Korrelationskoeffizienten min. größer gleich .20) und sich in mindestens zwei Studien nachweisen ließen. Für das Merkmal Dominanz ist das Vorgehen beispielhaft dargestellt.



Sind diese Zusammenhänge gefunden, überträgt ein Team bestehend aus Psychologen, Data-Scientists und Programmierern diese in Code-Sprache, sodass sie im 100 Worte Core nutzbar werden. Eine Auflistung aller aus Texten gemessener Merkmale findet sich im Anhang (siehe *Überblick Studien*).

Die 100 Worte Sprach-Kategorien umfassen verschiedene psychologische Merkmale. Insgesamt beinhaltet unsere Textanalyse 38 Kategorien, die sich in drei Gruppen aufteilen: Merkmale der Persönlichkeit, Merkmale der Stimmungen und Emotionen und Merkmale der Motive und Bedürfnisse. Diese Merkmale lassen sich mithilfe einer Mensch-Metapher eingängig darstellen. Eine Darstellung aller Kategorien findet sich im Anhang (siehe *Überblick 100 Worte Kategorien*).



Eigene Arbeiten

100 Worte gibt sich aber mit der bloßen Übertragung von bestehenden Forschungsbefunden nicht zufrieden. Unser Anspruch ist es, die Qualität unserer Textanalyse selbst nachzuweisen. Um das zu erreichen, führen wir Studien durch und vergleichen unsere Analyse mit anderen Verfahren sowie mit Wettbewerbern. Aus dieser Forschungstätigkeit sind bereits vielversprechende Resultate entstanden. Wir verglichen dabei die 100 Worte Stimmungskategorien mit drei gängigen Sentiment-Corpi für die deutsche Sprache: SentiWS, German Polarity Clues und Zürich Lexicon. Wir ließen die vier Wörterbücher in der Vorhersage von Stimmungen in rund 8000 deutschen Tweets gegeneinander antreten. Die 100 Worte Stimmungskategorien erzielte dabei eine vier bis 15 Prozent höhere Genauigkeit (=F-Score[1]) in der Vorhersage der Stimmung gegenüber den konkurrienden Wörterbüchern.

Wir evaluierten aber nicht nur unsere Stimmungs- sondern auch unsere Motivkategorien. Dazu verwendeten wir einen öffentlich zugänglichen Datensatz von Schultheiss, der 2009, im Rahmen der Studie *„Activity inhibition: a predictor of lateralized brain function during stress*" entstand: http://www.psych2.phil.uni-erlangen.de/%7Eoschult/humanlab/publications.htm

Die Details der Studie sollen hier nicht Thema sein. Interessierte können sie hier einsehen: https://www.100worte.de/blog/page/27/wie-gut-sagen-100worte-unsere-motive-vorher-unsere-studie-zur-validierung/

In diesem Paper solle es stattdessen darum gehen, wie wir methodisch vorgehen und wie wir das im Vergleich mit anderen tun, beispielhaft an unserer Motiv-Studie dargestellt.

	Wettbewerb	100 Worte
Ziel	Validierung der Big-Five Kategorien	Validierung der Motiv-Kategorien
Label	**30** Fremdeinschätzung **unerfahrener** Bewerter von DAX 30 CEOs hinsichtlich Big Five basierend auf „beruflichen und privaten Informationen“. Verwendet aus Mai et al., 2015.	**600** Textproben von **100** Versuchspersonen, die von **zwei erfahrenen** Bewertern mithilfe eines seit Jahrzehnten verwendeten Manuals hinsichtlich Motiven eingeschätzt wurden. Verwendet aus Schultheiss et al., 2009.
Label: Gütemaße[3]	Objektivität: **nicht angegeben** Reliabilität: **nicht angegeben** Validität: **nicht angegeben**	Objektivität: **gegeben** Reliabilität: **mittel** Prädikative Validität: **hoch**
Label: Verfügbarkeit	Daten **nicht** verfügbar	Daten **frei** verfügbar: http://www.psych2.phil.uni-erlangen.de/%7Eoschult/humanlab/publications.htm
Label: Interessens-konflikt	**möglich**, da Autoren im Board von Wettbewerb sitzen	**ausgeschlossen**, da Autoren unabhängig von 100 Worte
Datengrundlage	112 subjektiv ausgewählte Interviews, Pressekonferenzen, YouTube-Beiträge von Dax 30 CEOs	600 Textproben von 100 Versuchspersonen
Datengrund-lage: Interessens-konflikt	**möglich**, da von Wettbewerb selbst erstellt	**ausgeschlossen**, da Autoren unabhängig von 100 Worte
Ergebnisse: konvergente Validität	**r= -0.06 - 0.16** (keine Angabe von Signifikanz)	**r= 0.21 - 0.52** (Signifikanz auf dem 0.001-Niveau)
Replikation	nicht vorhanden	vorhanden: Schultheiss, 2013
Ableitung	von Wettbewerber einerseits: „Die ausführliche Prüfung der [konvergenten Validität] zeigt, dass diese sowohl aufgrund der signifikanten Mittelwertsunterschiede auf 5%-Niveau als auch aufgrund der niedrigen Korrelation abgelehnt werden muss.“ andererseits: „Aufgrund der Übereinstimmung der Ausprägungstendenzen in Bezug auf die Big- Five-Dimensionen von … mit dem aktuellen Forschungsstand wird davon ausgegangen, dass die automatisierte Sprachanalyse ein geeignetes Instrument zur Persönlichkeitsbeurteilung ist.“	von Schultheiss 2013 (basierend auf seinen Befunden): „Assessment of implicit motives with a word-count approach yields scores that converge with content-coded motive measures, that predict well-documented validation criteria of implicit motive measures, and that respond sensitively to experimental arousal of motivation.“

Dargestellt ist das unterschiedliche Vorgehen bei zwei Validierungsstudien, die von einem Wettbewerber (linke Spalte) und von 100 Worte (rechte Spalte) durchgeführt wurden. Ziel beider Studien war es, die konvergente Validität der Verfahren zu erheben. Der gemachte Vergleich eignet sich gut um aufzuzeigen, mit welchen Standards 100 Worte arbeitet.

Zunächst ist es 100 Worte wichtig, in keinem Interessenskonflikt zu stehen oder diese zu benennen, wenn sie vorliegen. Interessenskonflikte bestehen immer dann, wenn selbst erzielte Forschungsbefunde als Argument für den Vertrieb des eigenen Produkts dienen. Ein solcher Interessenskonflikt besteht sowohl bei unserem Wettbewerber als auch bei 100 Worte. Damit unsere Aussagen aber trotzdem einen wissenschaftlichen Wert haben, achtet 100 Worte darauf, dass die Studien an sich frei von verfälschenden Einflüssen von Stakeholdern sind. Das erreichte 100 Worte, indem der zur Validierung genutzte Datensatz von anderen – nicht mit 100 Worte in Beziehung stehenden – Wissenschaftlern erhoben wurde und er die Anforderungen an statistische Güte erfüllt. Beides, Unabhängigkeit in der Erstellung des Datensatzes als auch statistische Güte, ist bei unserem Wettbewerb nicht gegeben (siehe Tabelle Zeilen 1 bis 4). Weiterhin ist es 100 Worte wichtig, Befunde nur dann als relevant zu deklarieren, wenn sie statistische Signifikanz aufweisen (also keine Zufallsbefunde sind) und der beobachtete Effekt relevant ist. Ein gängiges Maß für den Zusammenhang zwischen Merkmalen ist die Korrelation. Die Korrelation kann Werte zwischen 1 (für einen perfekten positiven Zusammenhang), -1 (für einen perfekten negativen Zusammenhang) und 0 (wenn es keinen Zusammenhang gibt) annehmen. Weil aber Persönlichkeitsmerkmale oftmals von anderen Einflüssen verdeckt sind, werden nur selten Korrelationen über 0,4 gefunden. In der psychologischen Forschung gelten Werte ab 0,2 als relevant. Gemäß dieser Systematik bewertet 100 Worte die gemachten Befunde. Auf die oben vorgestellte Studie übertragen, ist der gemachte Befund sowohl statistisch hoch signifikant als auch relevant, da die Effektmaße über der geforderten Grenze von 0,2 liegen. Beides trifft für unseren Wettbewerb nicht zu. Deswegen kann hier auch nicht von einem im statistischen Sinne validen Verfahren gesprochen werden. Das hat Konsequenzen für die Empfehlungen, die sich aus den Befunden ableiten. Für 100 Worte käme es nicht in Frage, fragwürdige Befunde als nützlich zu erklären. Hier unterscheiden wir uns ebenfalls von unserem Wettbewerb (Tabelle Zeile *Ableitung*). Schließlich ist es 100 Worte wichtig, dass Befunde keine Glückstreffer sind, sondern tatsächlich systematische Zusammenhänge darstellen. Deswegen versuchen wir stets unsere Befunde in die bestehende Forschungslage einzubetten. Besonders hilfreich ist es dabei, wenn auch andere Forscher mit ähnlichen Methoden zu ähnlichen Ergebnissen kommen. Für die vorliegende Studie ist das bei 100 Worte der Fall (Tabelle Zeile *Replikation*). Unser Wettbewerb kann jedoch keine Replikation vorweisen.

Schließlich will 100 Worte zuverlässige Anwendungen entwickeln. Es geht also auch um die Reliabilität[4]. Hierunter ist die Zuverlässigkeit, also die Stabilität der Messwerte über die Zeit zu verstehen. Die Beurteilung der Güte von Textanalysen ist aber nicht so einfach, wie es zunächst vielleicht scheint. Für Fragebogen-basierte Tests ist das Vorgehen klar und vergleichsweise simpel: Im Test werden eine Reihe von Fragen zur Ausprägung eines bestimmten Persönlichkeitsmerkmals

gestellt wie z. B. Gewissenhaftigkeit. Eine Person beantwortet diese Fragen, indem sie entweder „Ja" oder „Nein" ankreuzt oder das Maß der Zustimmung oder Ablehnung zu einer Aussage im Test angibt. Anschließend werden diese Fragen miteinander korreliert und man erhält eine Aussage über die Zuverlässigkeit bzw. interne Konsistenz des Fragebogens. Je stärker die Antworten korrelieren, desto mehr messen die Fragen das Gleiche. Damit ist die interne Konsistenz und damit die Reliabilität des Tests belegt. Das gleiche Vorgehen könnte man auch zur Einschätzung von Wörterbüchern verwenden. Doch ist Sprache komplexer als die Antwort auf eine Frage in einem Persönlichkeitstest. Wenn eine Person etwas sagt, wird sie es nicht im nächsten oder den darauf folgenden Sätzen wieder sagen, denn normalerweise wird etwas gesagt und dann zum nächsten Thema übergegangen. Es wäre seltsam, wenn sich eine Person ständig wiederholt. Aber genau das, nämlich die Wiederholung, ist ein grundlegendes Konstruktionsmittel von Fragebogen-Verfahren: Unterschiedlich formulierte Fragen versuchen das gleiche Konstrukt (z. B. Gewissenhaftigkeit) zu erfassen. Da die Messung von Sprache „im Feld" erfolgt, ist die Wiederholung unwahrscheinlich. Aufgrund dieser Tatsache haben Textanalyse-Verfahren die Schwierigkeit mit „komplexen Antworten" umzugehen.

Oftmals wird eine hohe interne Konsistenz als Voraussetzung für Validität angesehen. Verschiedene Wissenschaftler (z. B. Reuman, 1984) haben darauf hingewiesen, dass – zumindest in der Messung von Motiven in Sprache – geringe interne Konsistenz keinen Einfluss auf die Konstruktvalidität aufweist. Ein weiterer Wissenschaftler (= Atkinson, 1981) verwies darauf, dass die Annahmen der klassischen Testtheorie (nach Gulliksen, 1950) für die Messung von Motiven in Sprache unangemessen seien. Aufgrund der geschilderten Besonderheiten hält Pennebaker (2015) fest, dass die Zuverlässigkeitskoeffizienten der natürlichen Sprache niedriger sind als die, die sonst bei psychologischen Tests gemessen werden. 100 Worte teilt die Auffassung von Pennebaker.

Aber wie genau sieht das Vorgehen zur Messung und Berechnung der internen Konsistenz der einzelnen Wörterbücher, die 100 Worte verwendet, aus? Die 100 Worte Wörterbücher sind nach psychologischen Merkmalen gruppiert. So gibt es ein Wörterbuch für die Emotion Ärger, eines für das Motiv Leistung, usw. Wir nehmen an, dass mehrere Worte eines Wörterbuchs verwendet werden, wenn über das übergeordnete psychologische Merkmal gesprochen wird. Wenn eine Person z. B. traurig ist, wird sie verschiedene Worte verwenden, die Trauer ausdrücken. Diese Worte haben also eine höhere Wahrscheinlichkeit, gemeinsam aufzutauchen, als Worte, die zu einem ganz anderen Merkmal gehören (z. B. Freude). Diese gemeinsame Auftretenswahrscheinlichkeit kann gemessen werden und gibt die interne Konsistenz unserer Wörterbücher an. Wir erreichen Werte für die interne Konsistenz zwischen .16 und .89. Damit liegen wir in etwa bei den Werten des Language Inquiry and Word Count (LIWC, 2008) für die englische Version. Dort werden Werte zwischen .18 und .93 berichtet. Für die 100 Worte Motivkategorien werden interne Konsistenzen zwischen .48 und .53 im Datensatz von O. Schultheiss (2009) erzielt.

Abschließend kann zum Thema Reliabilität von Textanalyse-Verfahren Folgendes festgehalten werden: Fragebogen-Verfahren erreichen höhere interne Konsistenz-Werte. Diese führen aber nicht zwangsläufig zu validieren Aussagen. So zeigten sich textbasierte Verfahren zur Messung von Motiven in der Vorhersage von Berufserfolgen prädiktiv valide – sogar 16 Jahre nach Messung. Allerdings zeigten sich im selben Zeitraum keine Zusammenhänge mit fragebogen-basierten Motivmessungen (z. B. McClelland & Boyatzis, 1982, Jenkins, 1994 oder Collins, Hanges & Locke, 2004). Erklärt werden kann dieser scheinbar widersprüchliche Zusammenhang zwischen mittelmäßiger Reliabilität (=Homogenität) und hoher prädiktiver Validität von Textverfahren durch die Unmittelbarkeit von Sprache. Sie stellt eine unverfälschte, direkte Verhaltensprobe einer Person dar. Fragebögen erfordern stattdessen eine bewusste Auseinandersetzung, da sie kein bereits vorhandenes Verhalten erfassen. Dadurch sind per se latente (unbewusste) Inhalte nicht zugänglich oder durch Vorstellungen davon (=Introspektion) verfälscht. Gerade latente Variablen sind aber langfristig wirksam und deren Berücksichtigung ist unverzichtbar in der Vorhersage von Verhalten.

Literatur

- Atkinson, J. W. (1981). Studying personality in the context of an advanced motivational psychology. American Psychologist, 36, 171-128.
- Collins, C. J., Hanges, P. J., & Locke, E. A. (2004). The Relationship of Achievement Motivation to Entrepreneurial Behavior: A Meta-Analysis. *Human Performance, 17*(1), 95-117.
- Gulliksen, H. (1950). Theory ofmental tests. New York: Wiley.
- Jenkins, S. R. (1994). Need for power and women's careers over 14 years: Structural power, job satisfaction, and motive change. *Journal of Personality and Social Psychology, 66*(1), 155-165.
- McClelland, D. C., & Boyatzis, R. E. (1982). Leadership motive pattern and long-term success in management. *Journal of Applied Psychology, 67*(6), 737-743.
- Schultheiss, O. C., Riebel, K., and Jones, N. M. (2009). Activity inhibition: a predictor of lateralized brain function during stress. Neuropsychology 23, 392–404.

- Tausczik, Y.R., & Pennebaker, J.W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology, 29, 24–54.
- Veroff, J., Reuman, D., u. Feld, S. (1984). Motives in American men and women across the adult life span. Developmental Psychology, 20, 1142-1158.

Anhang

Überblick Studien

	100 Worte Validierung	Newman et al., 2008	Schultheiss, 2013	Pennebaker et al., 2014	Kacewicz et al., 2013
Stimmung	**0,6**				
Geschlecht	**0,3**	0,3			
Motiv Führung	**0,4**		0,2		
Motiv Leistung	**0,7**		0,3		
Motiv Beziehung	**0,6**		0,3		
Denkstil	Studie in Arbeit			0,2	
Dominanz	Studie in Arbeit				0,2

Überblick 100 Worte Kategorien

	Kategorie bedeutet…	Kategorie beinhaltet…
Stimmungen und Emotionen	Stimmungen und Emotionen geben Auskunft über die generelle Stimmung sowie die aktuelle Befindlichkeit. Indikative Worte für gute bzw. schlechte Stimmung sind z. B.	Gute Stimmung Schlechte Stimmung Angst

	Kategorie bedeutet…	Kategorie beinhaltet…
	„freudig“, „sorgenfrei“, „gefällt“; „unzufrieden“, „gelangweilt“, „ignorant“.	Ärger Trauer
Bedürfnisse und Motive	Menschen haben unterschiedliche Motive und Bedürfnisse, die ihr Handeln und Denken (damit auch das Sprechen) leiten. Besonders sympathisch und bezogen wirkt das Beziehungsmotiv in der Sprache weil hier Worte der Verbundenheit (z. B. "gemeinsam", "zusammen", "miteinander") verwendet werden.	Führung: Macht vs. Ohnmacht Leistung: Erfolg vs. Misserfolg Beziehung: Gemeinschaft vs. Einsamkeit
Charakteristika	Hierunter sind verschiedene Kategorien zusammengefasst, die die Persönlichkeit beschreiben.	Denkstil: Analytisch vs. Spontan Dominanz: dominant vs. nachgiebig Authentizität: formell vs. frei-heraus Weiblichkeit / Männlichkeit
Regulatorischer Fokus	Der regulatorische Fokus beschreibt zwei grundlegende motivationale Orientierungen. Menschen mit einem Verlustfokus ist es wichtig, Fehler und Gefahren zu vermeiden. Bei einem Gewinnfokus steht dagegen das Wachstum und der Gewinn im Vordergrund.	Reward Risk
Zeitliche Perspektive	Menschen haben unterschiedliche zeitliche Orientierungen in der Sprache. Manche verwenden eher eine zukunftsorientierte Sprache, andere eine vergangenheitsorientierte.	Vergangenheit Zukunft
Sprachliche Position (aktiv vs. passiv)	Menschen, die aktiv formulieren, leiten das Gespräch. Menschen, die passiv formulieren, sind Empfänger von Anweisungen. Sie werden geleitet. Im Umgang mit Kunden empfiehlt sich eine aktive Sprache.	Agent: aktive Position Patient: passive Position

	Kategorie bedeutet…	Kategorie beinhaltet…
Sprachliche Genauigkeit	Sprache unterscheidet sich in konkrete und unkonkrete Sprache. Ist eine Sprache unkonkret, lässt sie Raum für Unklarheit und Interpretation. Im Kundenkontakt ist eine konkrete Sprache anzustreben.	DAV: Descriptive Action Verbs IAV: Interpretative Action Verbs SV: State Verbs Adj: Adjectives
Sicherheit/Unsicherheit	Wenn sich Menschen einer Sache sicher sind, drücken sie das auch in absoluten Worten (z. B. "Immer", "auf jeden Fall"). Andernfalls verwenden sie Worte der Unsicherheit (z. B. "Möglicherweise").	Tentative Worte Absolute Worte
Persönliche Sprache	Sprache kann persönlich oder unpersönlich wirken. Bei einer persönlichen Ansprache werden Menschen direkt angesprochen, indem sie miteinander, übereinander oder über sich selbst sprechen.	Personalpronomen Ich Du Wir
Funktionsworte	Diese Kategorien haben inhaltliche Bedeutung, geben der Sprache aber Struktur. Sie sind wichtig, um Denkstile, Dominanz oder Authentizität in der Sprache zu messen.	Personalpronomen Artikel Präpositionen Pronomen Adverbien Hilfsverben
Inhaltskategorien	Diese Kategorie umfasst Worte, die einen bestimmten Inhalt ausdrücken.	Finanzen

Begriffsdefinition

[1] Bei der statistischen Analyse der binären Klassifizierung ist der F1-Score (auch F-Score oder F-Messung) ein Maß für die Genauigkeit eines Tests. Es berücksichtigt sowohl die Genauigkeit p als auch die Trefferquote r des Tests: p ist die Anzahl der korrekten positiven Ergebnisse dividiert durch die Anzahl aller positiven Ergebnisse, die vom Klassifikator zurückgegeben werden, r ist die Anzahl der korrekten positiven Ergebnisse dividiert durch die Anzahl aller relevanten Proben (alle Proben, die als positiv identifiziert werden sollten). Der F1-Wert ist der harmonische Mittelwert aus Genauigkeit und Trefferquote, wobei ein F1-Wert seinen besten Wert bei 1 (perfekte Genauigkeit und Trefferquote) und seinen schlechtesten bei 0 erreicht.

[2] Statistisch signifikant wird das Ergebnis eines statistischen Tests genannt, wenn Stichprobendaten so stark von einer vorher festgelegten Annahme (der Nullhypothese) abweichen, dass diese Annahme nach einer vorher festgelegten Regel verworfen wird. Sinnvollerweise wird bei der Festlegung dieser kritischen Schwelle bedacht, welche Konsequenzen der Fall hätte, dass irrtümlich angenommen wird, ein beobachteter Unterschied sei nur zufällig. Hält man diese Folgen eher für gravierend, so wird man hier eher ein niedriges Niveau als ein höheres wählen, beispielsweise lieber 1 % als 5 %, oder aber 0,1 % für die maximal zulässige Irrtumswahrscheinlichkeit festlegen. Diese Wahrscheinlichkeit wird als Signifikanzniveau α bezeichnet. So bedeutet $\alpha=0{,}05$: Falls die Nullhypothese richtig ist, darf die Wahrscheinlichkeit dafür, dass sie fälschlich abgelehnt wird (Fehler 1. Art), nicht mehr als 5 % betragen. Entsprechend beträgt dann die Wahrscheinlichkeit, eine richtige Nullhypothese aufgrund des statistischen Tests nicht abzulehnen, $1-\alpha=0{,}95$, sprich mindestens 95%.

[3] Jede wissenschaftliche Messmethode muss bestimmten Gütekriterien (im Sinne von Qualitätskriterien) genügen. Objektivität und Zuverlässigkeit sind Forderungen für nahezu alle Messungen. Im engeren Sinne sind diese Kriterien für psychologische Tests bzw. allgemeiner psychodiagnostische Verfahren verfeinert worden – sie sind Spezifikationen allgemeingültiger Gütekriterien für wissenschaftliche Erkenntnismethoden. Als Hauptgütekriterien gelten (jede nachfolgende Stufe ist nur nach Erfüllung der vorhergehenden zu erreichen):

- Objektivität: Sind die Ergebnisse unabhängig von Einflüssen der Untersucher oder der Untersuchungssituation bei Durchführung, Auswertung und Interpretation zustande gekommen?
- [4]Reliabilität: Wird das Merkmal zuverlässig gemessen oder ist die Messung in zu großem Ausmaß mit Messfehlern behaftet?
- Validität: Misst das Verfahren tatsächlich das gewünschte Merkmal? Ist die Verwendbarkeit des Verfahrens für eine diagnostische Entscheidung gegeben?