Automatische Motiv-Vorhersage: Was kann ein Sprachmodell leisten?

Aswathy Velutharambath, M.Sc. Computational Linguistics & Daniel Spitzer, M.Sc. Psych.

März 2020

Hintergrund

Die Vorhersage von psychologischen Merkmalen aus Sprache mithilfe von automatischen Verfahren ist ein recht neues Betätigungsfeld in verschiedenen Wissenschaften wie Psychologie, Natural Language Processing (Emotion Detection) und Linguistik (Soziallinguistik). In den genannten Richtungen werden verschiedene Verfahren verwendet, um sich dem Ziel - der Messung von Persönlichkeit aus Sprache - anzunähern. Aus Psychologie und Linguistik sind regelbasierte eigenschaftsbasierte Verfahren bekannt, die Wortlisten, die mit spezifischen, mit einem Persönlichkeitsaspekt assoziierten Wörtern, gefüllt sind, verwenden. Solch ein Verfahren setzt auch 100 Worte ein und reichert dieses um die Verfahren des NLP (=Natural Language Processing) an um z. B. über die Bedeutung von Homonymen (=Worten mit mehreren Bedeutungen) Auskunft zu erhalten[1]. Die Kombination von regelbasierten Ansätzen mit maschinellen Verfahren hat das Potenzial, die Vorteile beider Verfahren zu nutzen. Regelbasierte Verfahren liefern nachvollziehbare Ergebnisse: Warum Text in einer gegebenen Art und Weise eingeschätzt wird, kann immer nachvollzogen werden. Maschinelle Verfahren sind dagegen häufig genauer, da sie für den spezifischen Sprachanlass und die Vorhersage eines Kriteriums trainiert wurden. Dafür sind die allermeisten maschinellen Verfahren aber Black-Boxes und es kann im Nachhinein nicht mehr nachvollzogen werden, wie ein Ergebnis zustande kam. Daher eignen sich maschinelle Verfahren auch nicht in der Diagnostik. Berechtigterweise wird in der Diagnostik gefordert, dass Ergebnisse nachvollziehbar sind. Nicht nachvollziehbare Ergebnisse bergen das Risiko einer fehlerhaften Aussage: Zwar weißt der Test auf das Vorhandensein eines Merkmals hin,

[1] Für eine genaueren Einblick in die 100 Worte Analyse, folgen Sie diesem Link: https://cdn2.hubspot.net/hubfs/6866529/Manual-Entwicklung-einer-KI-gestuetzten-Textanalyse_Daniel_Spitzer.pdf

doch muss nicht zwangsläufig ein kausaler Zusammenhang zwischen Merkmal und beobachtetem Objekt bestehen. Zwar können solche fehlerhaften Aussagen auch von regelbasierten Verfahren gemacht werden, allerdings lassen sie sich dort überprüfen.
In der vorliegenden Arbeit haben wir versucht, beide Verfahren miteinander zu kombinieren um ein relevantes Persönlichkeitsmerkmal aus Sprache mit einer möglichst hohen Präzision vorherzusagen. Konkret untersuchten wir die Vorhersagbarkeit von impliziten Motiven aus Texten mithilfe des Sprachmodells BERT (Bidrectional Encoder Representations from Transformers); Devlin, Chang, Lee, & Toutanova, 2018).

Implizite Motive

Motive sind interne Handlungsanleiter. Sie veranlassen einen Menschen, in eine bestimmte Richtung zu denken und sich auf eine bestimmte Weise zu verhalten. Sie drücken ein Streben aus: „Ich möchte gerne … sein“ oder „Ich wäre gerne …“. Generell werden drei Grundmotive unterschieden. Es sind die Motive nach Macht, Status und Führung, nach Leistung und Weiterentwicklung und nach harmonischen Beziehungen zu anderen Menschen.
Diese Motive wurden von dem Sozialpsychologen David McClelland formuliert und in der Psychologie beforscht. Die Erforschung und Bestimmung von Motiven macht Sinn, da sich mit ihnen wichtige Dinge wie z. B. den beruflichen Erfolg vorhersagen lassen. Bisher wurden Motive aber mit einem Verfahren erfasst, das nur mit viel Aufwand durchgeführt und ausgewertet werden kann. Texte, die von Personen geschrieben werden und über die implizite Motivstruktur einer Person Auskunft geben, müssen in mühevoller Kleinarbeit von erfahrenen Psychologen untersucht werden. Eine zuverlässige maschinelle Unterstützung wäre daher eine echte Innovation.

Vorgehen

Da zur Vorhersage der impliziten Motive in dieser Arbeit das Sprachmodell BERT (Bidrectional Encoder Representations from Transformers) verwendet wird, soll es im Folgenden vorgestellt werden. BERT ist ein sehr mächtiges Modell für natürliche Sprache. Es ist darauf trainiert, natürlichen Sprachgebrauch zu erkennen und erlangt dadurch ein hohes Maß an Sprachverständnis. Auf Grundlage dieses allgemeinen Modells kann man durch weiteres Training speziellere Aufgaben lösen, die Sprachverständnis benötigen. Das Sprachmodell wird beispielsweise dafür eingesetzt, den Kontext von Sprache zu beachten und somit u. a. die Bedeutungen von Begriffen besser zuordnen zu können. Nach unserem Kenntnisstand stellt BERT aktuell das am weitesten entwickelte Modell für eben diese Aufgabe dar.
Bisher verwendeten wir BERT zur Unterstützung bei der Interpretation von Worten, die mehrere Bedeutungen haben. In den Sätzen „Stell mir bitte die Leiter an die Wand“ und

„Die Leiter sind gerade zu Tisch“, wird „Leiter“ mit unterschiedlichen Bedeutungen verwendet. Für Menschen ist die Bedeutung der beiden „Leiter“ einfach zu erfassen. Für Maschinen stellt es allerdings eine schwierige Aufgabe dar.
Um die Bedeutung ambiger Worte zu erfassen, bedurfte es eines Verfahrens, das den Kontext der in Frage stehenden Wörter berücksichtigt. Hierbei hilft uns unser Sprachmodell BERT. Als besonders effizient und treffsicher erweist sich BERT z. B. bei der Vorhersage von Formulierungen und ganzen Sätzen (Next Sentence Prediction). In verschiedenen Benchmark-Tests erzielt BERT bei dieser Aufgabe eine Genauigkeit von 98% (z. B. Devlin, Chang, Lee, & Toutanova, 2018).
In der vorliegenden Arbeit verwendeten wir BERT aber für eine andere Aufgabe, nämlich zur Klassifizierung von Sätzen hinsichtlich der darin angesprochenen Motive. Für diese Aufgabe stand uns der mehr als 66.000 Sätze umfassende Datensatz von Schönbrodt, Hagemeyer, Brandstätter, Czikmantori, Gröpel et al (2020 in press) zur Verfügung.
Jedem Satz in dieser Sammlung wurde eine Information angehängt, die Auskunft über die Motivausprägung gibt. Diese umfassen die drei Grundmotive need for Affiliation (=Beziehungs-/Anschlussmotiv), need for Power (Führungs-/Machtmotiv) und need for Achievement (Leistungsmotiv). Kodiert wurde eine 1, wenn eines der Motive vorhanden war und 0, wenn eines der Motive nicht vorhanden war. Es ergab sich folglich eine binäre Kodierung. Diese Kodierung wurde jeden einzelnen Satz in der Sammlung von auf diese Aufgabe trainierten Psychologen vorgenommen. Erhoben wurden diese Daten mithilfe der sechs Bilder umfassende „Bilder-Geschichten Übung“ (Picture-Story-Exercise, PSE; McClelland, Koestner, & Weinberger, 1989), der in weiten Teilen dem Thematischen-Auffassungs-Test entspricht, jedoch im Umfang etwas reduziert ist. Die Teilnehmer wurden instruiert, sich für jedes der sechs mehrdeutigen Bilder eine Geschichte zu überlegen und aufzuschreiben.
Den Datensatz teilten wir, wie üblich bei der Vorhersage von Merkmalen, in einen Trainings und einen Testdatensatz auf. Das wiederholten wir mehrmals, indem wir Trainings und Test-Datensatz immer wieder zufällig unterschiedlich wählten. Und jedes Mal ließen wir unser Sprachmodell vorhersagen, welches Motiv in einem gegebenen Satz zum Ausdruck kommt. Um entscheiden zu können, wie gut unser Sprachmodell die Motive richtig klassifizieren konnte, erhoben wir die sogenannte Accuracy, die den prozentualen Anteil aller richtigen Entscheidungen zu allen Entscheidungen angibt und andere Gütemaße.

Ergebnisse

Wir verstehen die vorliegende Arbeit als Klassifizierungsaufgabe einer binären Variable (Merkmal vorhergesagt / Merkmal nicht vorhergesagt) zu einer anderen, ebenfalls binären Variable (Merkmal tatsächlich vorhanden / Merkmal tatsächlich nicht vorhanden). Zu beachten ist, dass wir diese Klassifizierung auf Satzebene durchführten. Eine Aggregation der Daten auf Text oder Personenebene fand nicht statt.

Tabelle 1 gibt Aufschluss über die erzielten Kenngrößen Accuracy und F-Score.

Tabelle 1. *Accuracy, F-Score und Matthews Korrelationskoeffizient (MCC) in der Vorhersage der drei Motive mit BERT*

	Accuracy	**F-Score**	**MCC**
n Macht	0.83	0.84	.67
n Leistung	0.85	0.85	.70
n Beziehung	0.85	0.85	.71

Neben den genannten Gütemaßen berechneten wir auch den Matthews Korrelationskoeffizienten (MCC; Matthews, 1975) für die drei Klassifizierungen. Dieser gilt ebenso als Maß für die Qualität einer binäre Klassifikationsaufgabe.
Die Ergebnisse zeigen, dass das Sprachmodell BERT in der Lage ist, die Motive vorherzusagen. Wie die Qualität dieser Vorhersage einzuschätzen ist, soll im folgenden Abschnitt besprochen werden.

Konklusion

Um die erzielten Ergebnisse einschätzen zu können, ist es nötig, sie mit anderen Ergebnissen und Benchmarks zu vergleichen. In der vorliegenden Arbeit gingen wir der Frage nach, wie gut wir ein Merkmal zuordnen können. Dabei betrachteten wir verschiedene Maßzahlen. Die für unsere Ziele wohl am besten geeignete, ist die Korrektklassifikationsrate (englisch: *Accuracy*). Sie gibt den Anteil aller Entscheidungen an, die richtig getroffen wurden und wird durch die Anzahl aller getroffener Entscheidungen geteilt. Die Accuracy ist also ein Quotient und liegt zwischen 0 und 1. In unseren Klassifikationsaufgaben erreichten wir eine Korrektklassifikationsrate zwischen 0,84 (für die Vorhersage von need for Power) und 0,85 (für die Vorhersage von need for Achievement und need for Affiliation). Des weiteren berechneten wir das sog. F-Maß. Es ist ein kombiniertes Maß aus der Genauigkeit und der Trefferquote der Klassifikation. Die Betrachtung des F-Maßes ermöglicht die Einschätzung der „Klassifizierungskraft“ bezüglich der beiden Maße Genauigkeit und Trefferquote. Aufgrund der Berechnung des F-Maßes sprechen hohe Werte für eine hohe Ähnlichkeit von Genauigkeit und Trefferquote, was für unsere Klassifikationsaufgabe anzustreben ist. Aufgrund der erzielten Ergebnisse ist unser Test als ausgeglichen zu bewerten, da es eine Ausgewogenheit zwischen Genauigkeit und Trefferquote aufweist.
Weiterhin berechneten wir den Matthews Korrelationskoeffizienten (MCC; Matthews, 1975). Im Vergleich zu den bereits berichteten Maßen, beachtet der MCC alle vier Felder (richtig Positive, richtig Negative, falsch Positive und falsch Negative) der Konfusions-

Matrix und gilt daher als das Maß mit der höchsten Informationsdichte. Laut Chicco und Jurman (2020) sollte insbesondere der Matthews Korrelationskoeffizient zur Bewertung der Güte von Klassifizierungsaufgaben herangezogen werden. Nach der Einschätzung von Powers (2011) lassen sich die Werte des MCC mit Werten der Pearson-Korrelation vergleichen. Daher erscheint es auch sinnvoll, Cohen's (1988) Empfehlung zur Interpretation der Stärke des korrelativen Zusammenhangs anzuwenden. Laut Cohen gilt ein Zusammenhang über r=.5 als hoch.
Wie gut ist der hier verwendete Klassifikator aber tatsächlich? Diese Frage ist abschließend schwer zu beantworten. Eine wichtige Vergleichsgröße ist die Interrater-Reliabilität. Dieser Wert wird gewöhnlich angegeben um bewerten zu können, wie hoch die Übereinstimmung zwischen Personen ist, die beide den selben Text bewerten. In der händischen Auswertung der Picture-Story-Exercise dient diese Größe als Gütemaß für die Fähigkeit der Annotatoren, Motive in Texten zu erkennen und richtig zuzuordnen. Diese müssen eine gewisse Vergleichbarkeit untereinander erzielen, da sonst keine reliable Messung der Motive möglich ist. Um die Interrater-Übereinstimmung zu bestimmen, wird die Pearson-Korrelation zwischen der Kodierungen zweier Annotatoren berechnet. Die höchsten Interrater-Reliabilität, die für zwei Personen berichtet werden, liegen zwischen r=.75 und r=.85 (vergleich Schultheiss, 2013). Die in dieser Arbeit erzielten Werte sind nicht direkt miteinander vergleichbar. Am ehesten kann der MCC als Vergleichsgröße herangezogen werden. Dieser liegt in dieser Arbeit zwischen r=.68 und r=.71 und damit leicht unter dem unteren Ende des Kontinuums der weiter oben berichteten Interrater-Reliabilität. Damit kovariiert die hier vorgestellte Klassifizierung mit der menschlichen Annotierung annähernd so hoch, wie die Klassifizierung von menschlichen Annotatoren untereinander.
Einschränkend ist festzuhalten, dass die Leistung unserer Klassifizierung nach oben beschränkt ist. Das beste Ergebnis, dass unsere Klassifizierung hätte erzielen können, wäre eine perfekte Vorhersage der menschlichen Kodierung. Die menschliche Kodierung wiederum ist aber nicht vollkommen sondern enthält eine Konfundierungen mit unbekannten Varianzanteilen, worauf die nicht eindeutige Korrelation zwischen menschlichen Annotatoren hinweist. Es ist also fraglich, ob die hier vorgestellte Klassifizierung noch bessere Ergebnisse erzielt hätte, wenn das Merkmal der Klassifizierung, eine höhere Reliabilität aufgewiesen hätte.

Literatur

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics, 21(1), 6.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure, 405(2), 442-451.

Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

Schönbrodt, F. D., Hagemeyer, B., Brandstätter, V., Czikmantori, T., Gröpel, P., Hennecke, M., ... & Kopp, P. M. (2020). Measuring implicit motives with the picture story exercise (PSE): Databases of expert-coded german stories, pictures, and updated picture norms. Journal of Personality Assessment.

Schultheiss, O. C. (2013). Are implicit motives revealed in mere words? Testing the marker-word hypothesis with computer-based text analysis. Frontiers in psychology, 4, 748.