Reviving a psychometric measure: Classification and prediction of the Operant Motive Test

Dirk Johannßen
MIN Faculty,
Dept. of Computer Science
Universität Hamburg
& Nordakademie

Chris Biemann
MIN Faculty,
Dept. of Computer Science
Universität Hamburg
22527 Hamnburg, Germany

David Scheffer
Dept. of Economics
Nordakademie
25337 Elmshorn, Germany

http://lt.informatik.uni-hamburg.de/
{biemann, johannssen}@informatik.uni-hamburg.de
{david.scheffer, dirk.johannssen}@nordakademie.de

Abstract

Implicit motives allow for the characterization of behavior, subsequent success and long-term development. While this has been operationalized in the operant motive test, research on motives has declined mainly due to labor-intensive and costly human annotation. In this study, we analyze over 200,000 labeled data items from 40,000 participants and utilize them for engineering features for training a logistic model tree machine learning model. It captures manually assigned motives well with an F-score of 80%, coming close to the pairwise annotator intraclass correlation coefficient of r = .85. In addition, we found a significant correlation of r = .2 between subsequent academic success and data automatically labeled with our model in an extrinsic evaluation.

1 Introduction

In psychology, texts have been analyzed for so-called motives since the 1930s Schultheiss and Brunstein (2010a). Implicit motives are unconscious motives, which are measurable by operant methods. Operant methods, in turn, are psychometrics, which are captured by having participants write free texts, i.e. participants are asked ambiguous questions or are shown faint images, which they describe or interpret. Classically, motives are labeled manually in these descriptions for further analysis (Schultheiss, 2008). Knowledge of operant motives facilitate clinical research on e.g. traumas, as conducted by Weindl and Lueger-Schuster (2016). According to Schultheiss (2008), there are three main motives of the operant system: i) affiliation (hereafter referred to as A), which is a desire for establishing positive relationships, ii) achievement (hereafter referred to as L), described as the capacity of mastering challenges and gaining satisfaction from such and iii) power (hereafter referred to as M), which is the desire to have an impact on one's fellows. Originally, psychological motives were measured with projective techniques, such as the thematic apperception test (TAT, (Murray, 1943)) or with questionnaires (Schüler et al., 2015). During the TAT, participants were shown between 8 and 30 colorless images in two sessions and were asked to tell stories for each of the 10 images per sessions, which took about 20-30 minutes. Besides this time consumption, the TAT showed variable objectivity, thus an acceptable inter-rater agreement could not be achieved. Motives can be also measured by questionnaires, which helps to achieve objectivity but measure something different, i.e. explicit motives. The hypothesis of those independent motivational systems (explicit, implicit) was proposed and shown by McClelland et al. (1989). Implicit motives are aroused by affective incentives that promise direct emotional rewards, whilst explicit motives are aroused by rational incentives, which include social expectations (Schüler et al., 2015).

Even though it is possible to predict the hierarchical development of managers, subsequent academic success and preferred clothing brands (as reviewed in Section 3), research on motives has declined mainly due to labor-intensive and costly human annotation by well-trained psychologists. In this work, we examine how far processing with natural language processing (NLP) techniques can automatize the assignment of operant motives. We evaluate our approach intrinsically as well as extrinsically for the prediction of subsequent academic success as reflected in grades of final student's bachelor's theses.

As far as we are aware, this is the first work that uses the OMT for training a machine learning algorithm in order to classify yet unlabeled data and investigate measurable connections between operant motives and subsequent academic success.

2 The OMT and MIX

The operant motive test (OMT) was originally developed by Kuhl and Scheffer (1999). Different to the TAT by Murray (1943), for measuring motives with the OMT, participants are shown sketched scenarios with multiple persons in underspecified situations, such as displayed in Figure 1.

The OMT has the two main advantages, that participants are asked to state very short answers in contrast to whole stories of the TAT and that the OMT introduces additional *levels* of affective valence to the three main motives ranging from 1 to 5, allowing psychologists to differentiate affects of participants even further. Level 1 stands for self-regulating, 2 for incentive-driven, 3 for self-driven, 4 for active avoidance and 5 for passive avoidance.

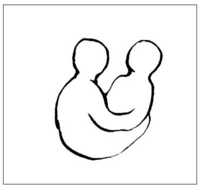

Figure 1: Sketched scenery for participant to answer four (OMT) questions on the narration and involved emotions (Kuhl and Scheffer, 1999)

A so-called zero-motive or zero-level (annotated as 0 for both, the motives and levels) are labeled if no clear motive or level can be identified, resulting in 4 X 6 possible target classes (0, A, M, L with levels 0 to 5). Even though cases are rare, it is possible to assign a level other than 0 with a 0 motive, i.e. no motive could be identified since motives and levels are orthogonal classifications.

A closely related psychometric test is the so-called Motive Index (MIX), developed by Scheffer and Kuhl (2006). The MIX is measured similarly to the OMT with slightly altered questions for an even faster assessment, making the MIX suitable for shortened aptitude diagnostics.

3 Related Work

McClelland and Boyatzis (1982) showed during an assessment center study that managers with a highly developed power motive were significantly more likely to reach higher hierarchy levels within 18 years. Weindl and Lueger-Schuster (2016) utilized the OMT for clinically investigating survivors of childhood abuse in foster care settings, finding connections between certain motive level constellations and symptoms of abuse. Schmidt and Frieze (1997) utilized the motive model of McClelland and Boyatzis (1982) on 142 college students and concluded that a stronger power motive occurrence mediated product involvement such as expensive cars or interview clothing, whilst affiliation was associated with purchasing gift cards. Schultheiss and Brunstein (2010b) analyzed CEO speeches and were able to predict individual and collective behavior of company members or companies. Schüler et al. (2015) compared and related three different motive measures, namely the picture story exercise (PSE, (Schultheiss and Pang, 2007)), the OMT and the multi-motive grid (Sokolowski et al., 2000), and showed that the measures differ in their scoring system and thus show little overlap, indicating them being unexchangeable. It is controversial whether the achievement motive is connected with academic success: Scheffer (2004) was able to predict grades with a significant correlation of r = .2, attributed to the intrinsic desire for excellence, whilst McClelland (1988) found that the power motive is rather correlated with academic success if grades are exposed to peers due to the desire to impress fellows.

Those studies show the validity and promising predictive power of the OMT, which can be utilized for aptitude diagnostics of different fields. In terms of the bachelor thesis grades, which are perceptible by peers, the predictability by the power motive can be hypothesized.

4 Data

Data has been collected by having 40,000 anonymized participants textually associate images in German such as the one in Figure 1 on the two questions i) Who is the main person and what is important for that person? ii) How does that person feel? The participants gave 220,859 answers on 15 different images. After filtering (cf. Section 5.1), we retain 209,716 text instances.

Each answer was labeled manually with the motives 0, A, L or M and a level ranging from 0 to 5. The annotators were psychologists, trained by the OMT manual by Kuhl and Scheffer (1999). The inter-annotator agreement with previously coded motives using the Winter scale (Winter, 1994) reached as high as 97% and 95% for the two annotators after the manual training. The pairwise intraclass correlation coefficient is an often utilized agreement measure, developed by Shrout and Fleiss (1979). This coefficient was measured to be .85 on average for the three motives (Schüler et al., 2015), thus showing the difficulty to standardize the labeling process.

The class distributions of motives and levels displayed in Table 1 show that the power motive (M) is with 59% nearly three times as frequent as the second largest class of achievement (L) with 19%. Furthermore, levels 4 and 5 together represent more than half of all level-labeled instances.

In addition to the roughly 220,000 labeled OMT text data instances, a small dataset of related but unlabeled MIX texts from 105 participants is available, which come with the additional information of the bachelor thesis grades of the anonymized participants. We will use this dataset for the extrinsic evaluation below.

5 Methodology

The main goal of this work is the automatization of the motive classification by training a machine learning model. Another goal will be the first and basic validation of the trained model by classifying the yet unlabeled 105 additional texts and hypothesizing a correlation between the achievement or the power motive with the bachelor thesis grades.

5.1 Pre-processing

We pre-processed the data by first removing spam, which mostly contained the same letters repeated, empty answers or a random variation of symbols. Also, we removed entries in different languages other than German. Lastly, texts with encoding problems were either resolved or removed. After this pre-processing, the whole dataset consisted of 209,716 texts. The distribution of filtered questions is uneven.

	0	**1**	**2**	**3**	**4**	**5**	Σ
0	7,921	0	2	1	2	6	7,932
A	11	2,888	9,581	1,361	7,617	6,822	28,280
L	6	2,455	12,697	6,405	7,542	3,742	32,847
M	25	11,338	12,353	15,248	36,103	23,610	98,677
Σ	7,963	16,681	34,633	23,015	51,264	34,180	167,736

Table 1: The OMT's training classes distribution after filtering and removing a held-out test and development set (10% each).

5.2 Feature engineering

For engineering features, the texts mostly were tokenized and processed per token. Engineered features were the type-token-ratio, the ratio of spelling mistakes and frequencies between 3 and 10 appearances.

Further features are LIWC and language model perplexities. The psychometric dictionary and software *language inquiry and word count* (LIWC) was developed by Pennebaker et al. (1999) and later transferred to German by Wolf et al. (2008). The German LIWC allowed for 96 categories to be assigned to each token, ranging from rather syntactic features such as personal pronouns to rather psychometric values such as familiarity, negativity or fear.

Part-of-speech (POS) tags were assigned to each token and thereafter counted and normalized to form a token ratio. We trained a POS tagger via the natural language toolkit (NLTK) on the TIGER corpus, assembled by Brants et al. (2004) and utilizing the STTS tagset, containing 54 individual POS tags.

We trained a bigram language model for each class and incorporated Good-Turing smoothing for calculating the perplexity. During training, we tuned parameters (e.g. which smoothing to use) via development set and tested the model with a held-out test set of 20,990 instances. The perplexity of a model q is:

$$2^{-\frac{1}{N}\sum_{i=1}^{N}\log_2 q(x_i)}$$

with p being an unknown probability distribution, $x_1, x_2, \ldots x_N$ being the sequence (i.e. the sentence) drawn from p and q being the probability model.

5.3 Model training

Even though deep learning has shown to be powerful, it often comes with a cost of losing transparency, which is crucial for our task, in which we seek to better understand the connection between psychology and language. Therefore we utilized different classical machine learning algorithms such as Naïve Bayes, LMT or regression and found the logistic model tree (LMT) implementation of Landwehr et al. (2005) to be the best-performing one amongst the tested. A LMT is a decision tree, which performs logistic regressions at its leaves. The root differentiates the language model's perplexities (A, M, and L) and thereafter performs the logistic regressions based on further features.

A qualitative post-hoc analysis by psychologists has resulted in an agreement with the model's predictions, except for too many assigned 0 labels and motives.

6 Results

Based on the correlation-based *Feature Subset Selection* by Hall (2000), the most influential features are the LIWC categories *I, Anger, Communication, Friends, Down, Motion, Occup, Achieve* and *TV*, as well as the perplexities of the language models affiliation (A), performance (L) and power (M) and attributive possessive pronoun (PPOSAT) POS tag frequency.

When classifying unlabeled OMT related texts of 105 anonymized participants, counting the motive predictions and analyzing a possible connection with the bachelor thesis grade and said counts, a weak but significant Pearson correlation coefficient of r = .2 could be found between the power motive and the thesis grade value (shown in Figure 2), whilst the achievement motive did not show any correlation. A wordlist-based model, which consists of 415 affiliation, 512 achievement, and 572 power words showed an insignificant correlation of r = .07 with an F-score of 61.07%.

Actual \ Predicted	0	A	L	M	Σ
0	338	92	163	427	1,020
A	51	2,667	105	708	3,531
L	115	66	3,151	804	4,136
M	209	573	556	10,965	12,303
Σ	713	3,398	3,975	12,904	20,990

Table 2: The confusion matrix of the motive classification task (without the levels) on the test set (10% of available data) with filtered values.

The confusion matrices in Table 2 illustrate the model's performance for each class. The model scores an F1 score of 65.4% for classifying the levels and 80.1% for classifying the motives.

An error analysis revealed that misclassified instances contain more words on average (24.2 versus 21.04). Also, misclassifications contain four times the amount of fillers (e.g. you know, like, i mean, Pennebaker et al. (1999)). Those instances are focused on plural personal pronouns twice as often and show a higher amount of answer particle. Moreover, misclassified instances contain 50% more often religious expressions, metaphors, and topics of sadness. Most of the misclassified instances show high perplexity scores of either one motive, are written in all caps and contain one-word sentences. When referring to the OMT manual Kuhl and Scheffer (1999) used for training psychologists on that labeling task, it is controversial whether all caps words should be viewed as a feature in itself and whether single word sentences qualify for being labeled different than 0, hence the OMT asks participants for stories rather than keywords. The annotators seem to have developed an intuition besides the OMT manual, as reflected in their high intraclass correlation coefficients.

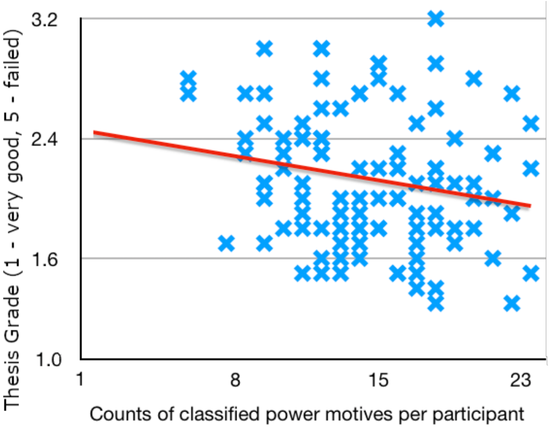

Figure 2: Correlation of r = -0.20 between LMT classifier predicted counts of power motive answers and the bachelor thesis grades. The German grading system ranges from 1.0 (very good) to 5.0 (failed).

7 Conclusion

The psychometric OMT is hampered by costly and labor-intensive manual annotation. Automatization is possible by utilizing the proposed model for motive and level classification. The annotators have had an average intraclass correlation coefficient of .85, whilst the overall F-score has reached 80.1%, clearly exceeding F = 61.07% of the wordlist-based model. Even though both measures are not directly comparable, the respectable F-scores suggest that the feature-engineered machine learning model is approaching human-like performance. Interestingly, the most influential features relate to the OMT theory. Lastly, a first theory validation has resulted in a significant r = .2 correlation between the predicted power motive and bachelor thesis grades. Furthermore, often better performing neural approaches should be considered for future work.

References

Sabine Brants, Stephaie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2004. The tiger treebank. *Journal of Language and Computation*, 2:597–620.

Mark Andrew Hall. 2000. *Correlation-Based Feature Selection for Machine Learning*. dissertation, University of Auckland, New Zealand.

Julius Kuhl and David Scheffer. 1999. *Der operante Multi-Motiv-Test (OMT): Manual [The operant multi-motive-test (OMT): Manual]*. Impart, Osnabrück, Germany: University of Osnabrück.

Niels Landwehr, Mark Andrew Hall, and Eibe Frank. 2005. Logistic Model Trees. *Machine Learning*, 59(1):161–205.

David Clarance McClelland. 1988. *Human Motivation*. Cambridge University Press.

David Clarance McClelland and Richard Boyatzis. 1982. Leadership Motive Pattern and Long-Term Success in Management. *Journal of Applied Psychology*, 67:737–743.

David Clarance McClelland, Richard Koestner, and Joel Weinberger. 1989. How do self-attributed and implicit motives differ? *Psychological Review*, 96(4):690–702.

Henry Alexander Murray. 1943. *Thematic apperception test*. Thematic apperception test. Harvard University Press, Cambridge, MA, US.

James Pennebaker, Martha Eileen Francis, and Roger John Booth. 1999. Linguistic inquiry and word count (LIWC). *Software manual*. http://liwc.wpengine.com (visited: 2019-01-17).

David Scheffer. 2004. *Implizite Motive: Entwicklung, Struktur und Messung [Implicit Motives: Development, Structure and Measurement]*, 1st edition. Hogrefe Verlag, Göttingen, Germany.

David Scheffer and Julius Kuhl. 2006. *Erfolgreich motivieren: Mitarbeiterpersönlichkeit und Motivationstechniken [Motivate Successfully: Employer Personality and Motivational Techniques]*, 1st edition. Hogrefe Verlag, Göttingen, Germany.

Laura Schmidt and Irene Hanson Frieze. 1997. A mediational model of power, affiliation and achievement motives and product involvement. *Journal of Business and Psychology*, 11(4):425–446.

Oliver Schultheiss. 2008. Implicit motives. *Handbook of personality: Theory and research*, pages 603–633.

Oliver Schultheiss and Joachim Brunstein. 2010a. *Implicit Motives*. Oxford University Press, Oxford, New York.

Oliver Schultheiss and Joachim Brunstein. 2010b. *Implicit Motives*. Oxford University Press, Oxford, New York.

Oliver Schultheiss and Joyce Pang. 2007. Measuring implicit motives. In *Handbook of research methods in personality psychology*, pages 322–344, New York, NY, US. Guilford Press.

Julia Schüler, Veronika Brandstätter, Mirko Wegner, and Nicola Baumann. 2015. Testing the convergent and discriminant validity of three implicit motive measures: PSE, OMT, and MMG. *Motivation and Emotion*, 39(6):839–857.

Patrick Shrout and Joseph Fleiss. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.

Kurt Sokolowski, Heinz-Dieter Schmalt, Thomas Langens, and Rosa Maria Puca. 2000. Assessing Achievement, Affiliation, and Power Motives All at Once: The Multi-Motive Grid (MMG). *Journal of Personality Assessment*, 74(1):126–145.

Dina Weindl and Brigitte Lueger-Schuster. 2016. Institutional Abuse (IA) and Implicit Motives of Power, Affiliation, and Achievement - an Alternative Perspective on Trauma-Related Psychological Responses. In *ISTSS International Society for Traumatic Stress Studies 32nd Annual Meeting*, Dallas, Texas, USA.

David Winter. 1994. *Manual for scoring motive imagery in running text*. Dept. of Psychology, University of Michigan (unpublished).

Markus Wolf, Andrea Horn, Matthias Mehl, Severin Haug, James Pennebaker, and Hans Kordy. 2008. Computergestützte quantitative Textanalyse: Äquivalenz und Robustheit der deutschen Version des Linguistic Inquiry and Word Count. *Diagnostica*, 54:85–98.