GermEval 2020 Task 1 on the Classification and Regression of Cognitive and Motivational Style from Text: Companion Paper

Dirk Johannßen[1,2], Chris Biemann[1], Steffen Remus[1], Timo Baumann[1], and David Scheffer[2]

[1]MIN Faculty, Dept. of Computer Science, Universität Hamburg, 22527 Hamburg, Germany
[2]Faculty of Economics, NORDAKADEMIE, 25337 Elmshorn, Germany
http://lt.informatik.uni-hamburg.de/
{*johannssen,biemann,remus,baumann*}*@informatik.uni-hamburg.de*
david.scheffer@nordakademie.de

Abstract

This paper describes the tasks, databases, baseline systems, and summarizes submissions and results for the GermEval 2020 Shared Task 1 on the Classification and Regression of Cognitive and Motivational Style from Text. This shared task is divided into two subtasks, a regression task, and a classification task. Subtask 1 asks participants to reproduce a ranking of students based on average aptitude indicators such as different high school grades and different IQ scores. The second subtask aims to classify so-called implicit motives, which are projective testing procedures that can reveal unconscious desires. Besides five implicit motives, the target labels of Subtask 2 also contain one of six levels that describe the type of self-regulation when acting out a motive, which makes this task a multiclass-classification with 30 target labels. 3 participants submitted multiple systems. Subtask 1 was solved (best $r = .3701$) mainly with non-neural systems and statistical language representations, submissions for Subtask 2 utilized neural approaches and word embeddings (best macro $F_1 = 70.40$). Not only were the tasks solvable, analyses by the participants even showed connections to the implicit psychometrics theory and behavioral observations made by psychologists.

1 Introduction

Despite the growing interest in NLP and its methods since 2015 (Manning, 2015), application fields of NLP in combination with psychometrics are rather sparse (Johannßen and Biemann, 2018). Aptitude diagnostics can be one of those application fields. To foster research on this particular application domain, we present the *GermEval-2020 Task 1 on the Classification and Regression of Cognitive and Motivational Style from Text*[1] [2] [3]. The task contains two subtasks. For Subtask 1, participants are asked to reproduce a ranking of students based on different high school grades and intelligence quotient (IQ) scores solemnly from implicit motive texts. For Subtask 2, participants are asked to classify each motive text into one of 30 classes as a combination of one of five implicit motives and one of six levels. Quantitative details on participation are displayed in Table 1.

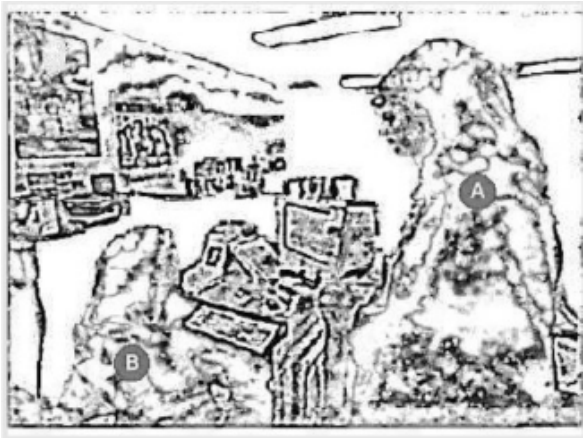

Figure 1: One example of an image to be interpreted by participants utilized for the motive index (MIX).

The validity of high school grades as a predictor of academic development is controversial (Hell et al., 2007; Schleithoff, 2015; Sarges and Scheffer, 2008). Researchers have found indications that linguistic features such as function words used in a prospective student's writing perform better

[1]GermEval is a series of shared task evaluation campaigns that focus on Natural Language Processing for the German language. The workshop is held as a joint Conference SwissText & KONVENS 2020 in Zürich.

[2]https://www.inf.uni-hamburg.de/en/inst/ab/lt/resources/data/germeval-2020-cognitive-motive.html

[3]The data and annotations were provided by Nicola Baumann (Universität Trier) and Gudula Ritz (Impart GmbH).

in predicting academic development (Pennebaker et al., 2014) than other methods such as GPA values.

During an aptitude test at a rather small university of applied sciences NORDAKADEMIE in Germany with roughly 500 students enrolling each year, participants are asked to write freely associated texts to provided questions, regarding shown images. Psychologists can identify so-called implicit motives from those expressions. Implicit motives are unconscious desires, which are measurable by operant methods (Gawronski and De Houwer, 2014; McClelland et al., 1989). Psychometrics are metrics, which can be utilized for assessing psychological phenomena. Operant methods, in turn, are psychometrics, which are collected by having participants write free texts (Johannßen et al., 2019). Those motives are said to be predictors of behavior and long-term development from those expressions (McClelland, 1988; Scheffer, 2004; Schultheiss, 2008).

From a small sample of an aptitude test collected at a college in Germany, the classification and regression of cognitive and motivational style from a German text can be investigated. Such an approach would extend the sole text classification and could reveal insightful psychological traits.

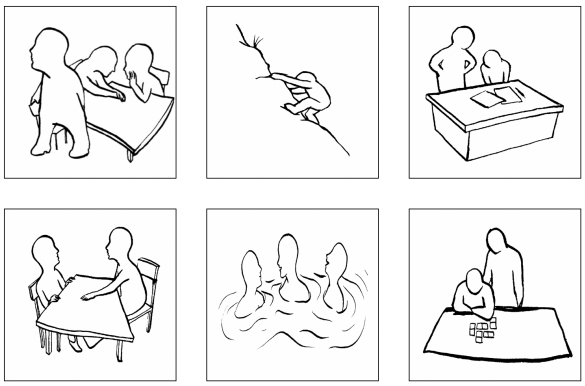

Figure 2: Some examples of images to be interpreted by participants utilized for the operant motive test (OMT) with A being the so-called affiliation motive and M being the power motive, two out of the five motives besides L for achievement, F for freedom and 0 for the zero / unassigned motive.

The Operant Motive Test (OMT, displayed in Figure 2) or the Motive Index (MIX, displayed in Figure 1) are tests that employ operant methods. For those tests, participants are required to use introspection and assess their psychological attributes unconsciously. Psychologists label these textual answers with one of five motives (M - power, A - affiliation, L - achievement, F - freedom, 0 - zero), and corresponding levels (0 to 5, with 0 being the zero level). Those levels describe the type of self-regulation when acting out a motive. For both, motives and levels, a zero is assigned, if no clear motive or level can be identified. The first level is the ability to self-regulate positive affect, the second level is the sensitivity for positive incentives, the third level is the ability to self-regulate negative affect, the fourth level is the sensitivity for negative incentives and the fifth level is the passive coping of fears (Scheffer and Kuhl, 2013).

There are findings for implicit motives being indicators for behavioral long-term developments. Scheffer (2004) found a weak, but significant correlation of $r = .2$ between high-school grades and the achievement motive. McClelland (1989) could show that if an achievement is highly visible to peers, a higher power motive creates a flow situation. The development of managers has been researched by McClelland and Boyatiz (1982): even after 18 years, managers with a higher achievement motive moved up higher in the company's hierarchy. By analyzing documents, speeches, messages or media commentaries, and other text resources. Winter (2007) measured implicit motives through content analysis of government statements, speeches, and diplomatic documents and showed that war situations and political crises were connected with higher levels of the power motive, whilst peace times were rather connected with the achievement motive. Semenova and Winter (2020) analyzed Russian presidents and found a high level of achievement motives in general, except for the third term of Vladimir Putin's office when international frictions grew stronger.

For our task, we provide extensive amounts of textual data from both, the OMT and MIX, paired with IQ and high school grades and labels.

The task is to predict measures of cognitive and motivational style solemnly based on text. For this, z-standardized high school grades and IQ scores of college applicants are summed and globally 'ranked'. This rank is utterly artificial, as no applicant in a real-world-setting is ordered in such fashion but rather there is a certain threshold over the whole of the hour-long aptitude test with multiple different test parts, that may not be undergone

by applicants.

2 Prior and related Work

In the prior work on this task (Johannßen and Biemann, 2019; Johannßen et al., 2019), the authors have performed classification of a reduced set of implicit motives, in which the so-called freedom motive (F) was not present and in which the levels were not part of the target labels either. Thus prior classification tasks related to the Subtask 2 are not directly comparable with this shared task. Johannßen et al. (2019) first utilized a logistic model tree (LMT) and hand-crafted features (e.g.
spelling mistakes, type-token ratio, part-of-speech (POS) tags) paired with a broadly utilized psychometrical language analysis tool called Linguistic Inquiry and Word Count (LIWC, pronounced 'Luke'). The authors were able to achieve a score of $F_1 = .81$, approaching the pairwise annotator intra-class correlation coefficient of $r = .85$ (for the four target classes M, A, L and 0). The LMT approach was not the most innovative but offered a chance of investigating algorithmic decisions made, as the structure is easier interpretable than methods used for deep learning (i.e. explainable AI is a widely open research problem (Rai, 2019)).

Later that year, Johannßen et al (2019) deepened their approach by employing deep learning to a similar classification task. The combination of an LSTM paired with an attention mechanism allowed the authors to investigate algorithmic decisions made, even though this approach was not to be confused with a true explanation. Thus, the authors also investigated correlations between classified motives and subsequent academic success in the form of college grades and found weak connections of the achievement motive with better college grades.

3 Aptitute test and college

Since 2011, the private university of applied sciences NORDAKADEMIE performs an aptitude college application test.

Zimmerhofer and Trost (2008, p. 32ff.) describe the developments of the German Higher Education Act. A so-called Numerus Clausus (NC) Act from 1976 and 1977 ruled that colleges in Germany with a significant amount of applications have to employ a form of selection mechanism. For most colleges, NC was the threshold for many applicants. Even though this value is more complex, it roughly can be understood as a GPA threshold. Since this second Higher Education Act, colleges are also free to employ alternate selection forms, as long as they are scientifically sound, transparent, and commonly accepted in Germany (Tschentscher, 1977).

Even though Hell (2007, p. 46) found the correlation coefficient of high school grades of r = .517 to be the most applicable measure for academic suitability, criticism emerged as well. The authors criticized the measure of grades by just one single institution (i.e. a high school) does not reflect upon the complexity of such a widely questioned concept of intellectual ability. Schleithoff (2015, p. 6) researched the high school grade development of different German federal states on the issue of grade inflation in Germany and found evidence, that supports this claim. Furthermore, in most parts of Germany, the participation grade makes up 60% of the overall given grade and thus is highly subjective.

Since operant motives are said to be less prone to subjectivity, the NORDAKADEMIE decided to employ an assessment center (AC) for research purposes and a closely related aptitude test for the application procedure (Gragert et al., 2018). Rather than filtering the best applicants, the NORDAKADEMIE aims with the test for finding and protecting applicants that they suspect to not match the necessary skills required at the college[4]. Thus, every part of the aptitude test is skill-oriented.

Furthermore, this test contains multiple other parts, e.g. math- and an English test, Kahnemann scores, IQ scores, a visual questionnaire, knowledge questions to the applied major or the implicit motives, the MIX.

4 Ethical considerations

Even though parts of this test are questionable and are currently under discussion, no single part of this test leads to an application being rejected. Only when a significant amount of those test parts are well below a threshold, applicants may not enter the second stage of the application process, which is applying at a private company due to the integrated study program the college offers. Roughly 10 percent of all applicants get rejected based on their aptitude test results. Furthermore, every applicant has the option to decline the data

[4] https://idw-online.de/de/news492748

	Task A	Task B motives	task B levels	task B motives + levels
# Teams	2	3	3	3
# Submissions	6	7	7	7
Best Team	TueOslo	FH Dortmund	FH Dortmund	FH Dortmund
Best pearson r	.3701	-	-	-
Best Macro-F_1	-	70.46	66.50	70.40
Impr. over baseline	.1769	5.52	6.92	5.95

Table 1: Quantitative details of submissions.

to be utilized for research purposes and still can apply to study at the NORDAKADEMIE. All anonymized data instances emerged from college applicants that consented for the data to be utilized in this type of research setting and have the opportunity to see any stored data or to have their personal data deleted at any given moment (e.g. sex, age, the field of study).

Any research performed on this aptitude test or the annually conducted assessment center (AC) at the NORDAKADEMIE is under the premise of researching methods of supporting personnel decision-makers, but never to create fully automated, stand-alone filters (Binckebanck, 2019). First of all, since models might always be flawed and could inherit biases, it would be highly unethical. Secondly, the German law prohibits the use of any – technical or non-technical – decision or filter system, which can not be fully and transparently be explained. Aptitude diagnostics in Germany are legally highly regulated.

The most debated upon the part of the aptitude test is the IQ. Intelligence in psychology is understood as results measured by an intelligence test (and thus not the intelligence of individuals itself). Furthermore, intelligence is always a product of both, genes and the environment. Even though there are hints that the IQ does not measure intellectual ability but rather cognitive and motivational style (DeYoung, 2011), it is defined and broadly understood as such.

Mainly companies in Europe employ IQ tests for selecting capable applicants. In the United Kingdom, roughly 69 percent of all companies utilize IQ. In Germany, the estimate is 13 percent (Nachtwei and Schermuly, 2009).

Since IQ tests only measure the performance in certain tasks that rather ask for skill in certain areas (logics, language, problem-solving) than cognitive performance, such intelligence tests should rather be called comprehension tests. Due to unequal environmental circumstances and measurements in non-representative groups, minorities can be discriminated by a biased (Rushton and Jensen, 2005). One result of research on the connection between implicit motives and intelligence testing could help to improve early development and guided support.

It is this bias, which leads to unequal opportunities especially in countries where there is a rich diversity among the population. Intelligence testing has had a dark history. Eugenics during the great wars e.g. in the US by sterilizing citizens (Lombardo, 2010) or in Germany during the Third Reich are some of the most gruesome parts of history.

But even in modern days, the IQ is misused. Recently, IQ scores have been used in the US to determine which death row inmate shall be executed and which might be spared. Since IQ scores show a too large variance, the Supreme Court has ruled against this definite threshold of 70 (Roberts, 2014). However, (Sanger, 2015) has researched an even more present practice of 'racial adjustment', adjusting the IQ of minorities upwards to take countermeasures on the racial bias in IQ testing, resulting in death row inmates, which originally were below the 70 points threshold, to be executed.

There is an ethical necessity to carefully view, understand and research the way intelligence testing is conducted and how those scores are – if at all – correlated with what we understand as 'intelligence', as they might be mere cognitive and motivational styles. Further valuable research can be conducted to investigate connections between other personality tests such as implicit motives with intelligence or comprehension tests. Racial biases are measurable, variances are large and many critics state that IQ scores reflect upon skill or cognitive and motivational style rather than real

intelligence as it is broadly understood.

It is important to note, however, that this task is not about automating IQ predictions from text or to research the IQ, but to conduct basic research on the possibility to predict psychological traits by text with a focus on implicit motives.

A more detailed ethical evaluation of this task and a recommended read has been formulated by Johannßen et al. (2020).

5 Data

5.1 NORDAKADEMIE Aptitude Data Set

Since 2011, the private university of applied sciences NORDAKADEMIE performs an aptitude college application test, where participants state their high school performance, perform an IQ test, and the implicit psychometrical test MIX. The MIX measures so-called implicit or operant motives by having participants answer questions to those images like the one displayed below such as "who is the main person and what is important for that person?" and "what is that person feeling?". Furthermore, those participants answer the question of what motivated them to apply for the NORDAKADEMIE.

The data consists of a unique ID per entry, one ID per participant, of the applicants' major and high school grades as well as IQ scores with one textual expression attached to each entry. High school grades and IQ scores are z-standardized for privacy protection.

The data is obtained from 2,595 participants, who produced 77,850 unique MIX answers and have agreed to the use of their anonymized data for research purposes.

The shortest textual answers consist of 3 words, the longest of 42 and on average there are roughly 15 words per textual answer with a standard deviation of 8 words. The (for illustrative purposes not z-standardized) average grades and IQ scores are displayed in Table 2.

The *IQ language* measures the use of language and intuition such as the comprehension of proverbs. The *IQ logic* tests the relations of objects and an intuitive understanding of mainly verbalized truth systems. The averaged IQ includes IQ language and logic as well as further IQ tests (i.e. language, logic, calculus, technology, and memorization).

Metric	score	standard deviation
German grade	9.4 points	1.84
English grade	9.5 points	2.15
Math grade	10.1 points	2.2
IQ language	66.8 points	19.0
IQ logic	72.6 points	15.6
IQ averaged	77 points	14.1

Table 2: Average scores and standard deviations of data for Subtask 1.

5.2 Operant Motive Test (OMT)

The available data set has been collected and hand-labeled by researchers of the University of Trier. More than 14,600 volunteers participated in answering questions to 15 provided images such as displayed in the figure below.

The pairwise annotator intraclass correlation was $r = .85$ on the Winter scale (Winter, 1994).

The length of the answers ranges from 4 to 79 words with a mean length of 22 words and a standard deviation of roughly 12 words. Table 3 shows the class distribution of the motives, the levels, and all the combinations. The number of motives in the available data is unbalanced with power (M) being by far the most frequent with 54.5%. The combined class of M4 is by far more frequent than e.g. the combination F_1. This makes this task more difficult, as unbalanced data sets tend to lead to overfitting. Those percentages were measured on the training set, containing a subset of 167,200 labeled text instances.

		Motives					
		Σ	**0**	**A**	**F**	**L**	**M**
	Σ	**100%**	**4.55%**	**16.83%**	**17.59%**	**19.63%**	**41.02%**
Levels	**0**	**4.6%**	4.55	.01	.00	.00	.01
	1	**9.9%**	.00	1.70	1.06	1.43	5.67
	2	**20.8%**	.00	5.73	3.33	7.69	4.11
	3	**13.6%**	.00	.81	2.57	3.76	6.46
	4	**30.7%**	.00	4.51	5.42	4.51	16.25
	5	**20.4%**	.00	4.07	5.57	2.24	8.52

Table 3: An overview of the Subtask 2 classes distributions (percentages). Values were rounded.

6 Task definitions

The *Shared Task on Classification and Regression of Cognitive and Motivational Style from Text* consists of two subtasks, described below. Participants could participate in any of them, may use external data and/or utilize the other data respec-

tively for training, as well as perform e.g. multi-task or transfer learning. Both tasks are closely related to the main research objective: implicit motives (see Section 1). For this first subtask, MIX texts are the basis for classifying cognitive and motivational style. For the second subtask, the OMT can be classified into main motives and levels.

6.1 Subtask 1: Regression of artificially ranked cognitive and motivational style

This task had yet never been researched and was open: It was neither certain, whether this task can be achieved, nor how well this might be possible before this task.

The goal of this subtask is to reproduce the artificial 'ranking' of students. Systems are evaluated by the Pearson correlation coefficient between system and gold ranking. An exemplary illustration can be found in Section 5. We are especially interested in the analysis of possible connections between text and cognitive and motivational style, which would enhance later submission beyond the mere score reproduction abilities of a submitted system.

A z-standardized example was provided with with a unique ID (consisting of studentID_imageNo_questionNo), a student ID, an image number, an answer number, the German grade points, the English grade points, the math grade points, the language IQ score, the math IQ score, and the average IQ score (all z-standardized). The data is delivered as displayed in Tables 4, 5, and 6.

The data is delivered in two files, one containing participant data, the other containing sample data, each being connected by a student ID. The rank in the sample data reflects the averaged performance relative to all instances within the collection (i.e. within train / test / dev), which is to be reproduced for the task.

The training data set contains 80% of all available data, which is 62,280 expressions and the development and test sets contain roughly 10% each, which are 7,800 expressions for the development set and 7,770 expressions for the test set (this split has been chosen in order to preserve the order and completeness of the 30 answers per participant).

For the final results, participants of this shared task were provided with a MIX_text only and were asked to reproduce the ranking of each student relative to all students in a collection (i.e. within the test set).

Field	Value
student_ID	1034-875791
image_no	2
answer_no	2
UUID	1034-875791_2_2
MIX_text	Die Person fühl sich eingebunden in die Unterhatung. [The person feels involved in the conversation.]

Table 4: Subtask 1 data file 1

Field	Value
student_ID	1034-875791
german_grade	-.086519991198202
english_grade	.3747985587188588
math_grade	.511555970796778
lang_iq	-.010173719700624
logic_iq	-.136867076187825

Table 5: Subtask 1 data file 2

Field	Value
student_ID	1034-875791
rank	15

Table 6: Subtask 1 data file 3

System submissions were evaluated on the Pearson rank correlation coefficient.

6.2 Subtask 2: Classification of the Operant Motive Test (OMT)

For this task, we provided the participants with a large dataset of labeled textual data, which emerged from an operant motive test (described in Section 1). The training data set contains 80% of all available data (167,200 instances) and the development and test sets contain 10% each (20,900 instances). The data is delivered as displayed in Tables 7, and 8.

On this task, submissions are evaluated with the macro-averaged F_1-score.

7 Systems

While 31 teams were registered on CodaLab, only 3 teams submitted systems for the official evaluation. In this section, we will describe the systems as well as the organizer's baseline, which was sur-

Field	Value
UUID	6221323283933528M10
text	Sie wird ausgeschimpft, will jedoch das Gesicht bewahren. [She gets scolded, but wants to save face.]

Table 7: Subtask 2 data file 1

Field	Value
UUID	6221323283933528M10
motive	F
level	5

Table 8: Subtask 2 data file 2

passed by most submissions. All results are displayed in Table 9.

7.1 Organizer's baseline systems

For both tasks, the organizers chose rather simple approaches that utilize support vector machines (SVM) paired with frequency-inverse document frequency (tf-idf) document representations.

SVMs are a class of statistical machine-learning algorithms that aim to map data to a higher dimensional feature space that best linearly separates target classes with the largest margin between them, which normally would not be separable linearly (this is called the *kernel trick*) and were first created by Cortes and Vapnk (1995). Tf-idf is a statistical evaluation of how important words are for documents and was first used by Luhn (1957).

7.1.1 Subtask 1

For Subtask 1, a Support Vector Regressor (SVR) was utilized. This statistical method tries to find an ideal line that best fits provided training data and thus examines a relationship between two continuous variables. Text is represented via tf-idf and a simple count vectorizer, which tokenizes text and builds vocabulary.

The SVR system achieved a Pearson ρ of .32, which is quite a big signal for data sources produced by human behavior. As there were 260 values to be ranked, we determined a T-value of 5.33 with a degree of freedom of 259, leading to a p-value of 2.096e-07. This means, that the result is highly significant and the null hypothesis can be rejected.

7.1.2 Subtask 2

As for the classification task, a linear support vector classifier (SVC) was chosen. 30 (combined motive-level labels) binary SVCs one-vs-all classifiers were trained. The data was centered and C (regularization) was set to the default 1.0 and the chosen loss is the *squared hinge*. It is useful for binary decision or when it is not of importance how certain a classifier is. The loss is either 0 or increases quadratically with the error. The system reached a macro F_1 score of 64.45 on the motive + labels classification task.

7.2 Submitted Systems

This section will provide a rough overview of the submitted systems, chosen word representations, some outstanding parameter choices, and some of the most interesting findings. For more details, it is recommended to read the resp. publications. Some details can be found in Table 9.

We notice two different approaches from the teams, especially from Subtask 1 to Subtask 2: i) statistical and non-neural word representations and systems and ii) neural approaches and word embeddings.

The team from Tübingen (Çöltekin, 2020) was very successful on the first subtask by using linear models with statistical n-gram features, exceeding the baseline by .1778 points and the second-placed team FH Dortmund by .0547 points. The authors note in their discussion, that, even though neural approaches nowadays offer broad applicability on all sorts of tasks, for the proposed regression task, their linear approach with n-gram features was sufficient. Even if the authors did not reach the first place on the second subtask with their self-designated *simple* linear and statistical approach, they still surpassed the organizer's baseline system on the second task by 3.36 percent points. Their results showed, that there is a signal in the implicit texts is sufficient for re-creating the ranking above chance.

The team Idiap (Villatoro-Tello et al., 2020) reached the second place for every type of Subtask 2 goal with a *Simple Transformer*, approach, which utilizes the attention mechanism without any recurrent units. Words were represented with pre-trained BERT (Devlin et al., 2018) embeddings. Since the attention mechanism offers the chance of investigating algorithmic decisions made, the authors plan for future work to investi-

Team	Classifier Approach	Task	Resp. score	Text Features
Tübingen (Çöltekin, 2020)	Subtask 1	Linear single 2	.3701	n-grams
FH Dortmund (Schäfer et al., 2020)	Subtask 1	SVR	.3154	tf-idf
Baseline	Subtask 1	SVR	.1923	tf-idf
FH Dortmund (Schäfer et al., 2020)	Subtask 2 motives	BERT ensemble cased	70.46	BERT
Idiap (Villatoro-Tello et al., 2020)	Subtask 2 motives	SimpleTransOut BERT LATEST	69.63	BERT
Tübingen (Çöltekin, 2020)	Subtask 2 motives	SVM adaptive	68.04	n-grams
Baseline	Subtask 2 motives	SVC	64.94	tf-idf
FH Dortmund (Schäfer et al., 2020)	Subtask 2 levels	BERT ensemble cased	66.50	BERT
Idiap (Villatoro-Tello et al., 2020)	Subtask 2 levels	SimpleTransOut BERT LATEST	65.32	BERT
Tübingen (Çöltekin, 2020)	Subtask 2 levels	linear-single2	63.35	n-grams
Baseline	Subtask 2 levels	SVC	59.85	tf-idf
FH Dortmund (Schäfer et al., 2020)	Subtask 2 motives + levels	DBMDZ uncased	70.40	BERT
Idiap (Villatoro-Tello et al., 2020)	Subtask 2 motives + levels	SimpleTransOut BERT LATEST	69.97	BERT
Tübingen (Çöltekin, 2020)	Subtask 2 motives + levels	SVM adaptive	67.81	n-grams
Baseline	Subtask 2 levels + motives	SVC	64.45	tf-idf

Table 9: Overview of the submitted approaches. Only the best submitted systems per team and task were considered. The entries are grouped by the type of task and displayed in descending order. DBMDZ stands for *Digitale Bibliothek Münchener Digitalisierungszentrum* and is a pre-trained German BERT model. SimpleTransOut stands for the Simple Transformer library from pypi.org.

gate those, possibly better understanding the OMT and the underlying patterns. During their presentations at the GermEval20 Task 1 session, the authors displayed tokens, which acquired high attention mass and concluded, that firstly, function words were more influential than content words, secondly, the so-called freedom motive was harder to distinguish from power than e.g. the achievement motives and that finally, negations were influential for classifying the power motive with level 4.

Lastly, the team from the FH Dortmund (Schäfer et al., 2020) utilized BERT word representations, exceeding the baseline-system of the motives + levels approach (30 target classes) by 5.95 percent points and the second-placed team Idiap by .61 percent points. The team experimented with different pre-processing steps but found, that they did not greatly influence the performance of their system, despite the data being mixed with different languages and some noise. For their approaches to solving Subtask 2, the authors experimented with different word representations, namely fasttext (Bojanowski et al., 2017) and BERT. Interestingly, the authors state that it was more useful for solving Subtask 2 to predict all 30 classes with a single model, than to train two classifiers for motives and levels respectively and to combine the predictions.

8 Discussion

The Organizer's SVM tf-idf systems have shown, that solutions of both subtasks above chance are possible. Subtask 2 with its implicit motives and levels appears to be a bit more trivial, as a macro score of $F_1 = 64.45$ is already strong, considering that the 30 target classes are unevenly distributed.

The submitted systems of the shared task participants revealed some interesting findings, which could be impactful for the implicit motive theory and their practical assessment.

Team Thübingen (Çöltekin, 2020) could recreate the Subtask 1 ranking above chance, even though there were no available manual labels. Since the impacts of identified implicit labels functioned as interim steps for behavioral pre-

dictions before (Johannßen and Biemann, 2019), those findings indicate the psychological validity of this implicit psychometric.

Team FH Dortmund (Schäfer et al., 2020) observed that for Subtask 2, excessive pre-processing did not make much of a difference. This, paired with an already strong but simple SVM tf-idf baseline system, indicates that language modeling already could be sufficient for classifying implicit motives and levels. If that were the case, the most impactful utterances per target class should be investigated and compared to the implicit motive theory.

Furthermore, the team found the direct prediction of 30 target classes of the motive + levels combination to be more sufficient than two models separately. Since the operant motive (OMT) theory states, that motives and levels are disjunct or orthogonal and thus not directly connected, those findings indicate incorrectness of this psychological empirical assumption. According to the provided data, this indicates that the OMT theory has to be investigated in terms of a connection between motives and levels. If that holds, it would be a very novel procedure, revising an empirical psychological theory based on NLP experiments and findings.

Lastly, some of the findings by the participants, have shown strong connections to behavioral research made on behalf of the implicit psychometrics theory. Winter (2007) identified so-called activity inhibition (AI) as good behavioral predictors for war and crisis situations by analyzing political speeches. AI is being described as negations in combination with the power motive. This connection between the power motive and negations was also observed by team Idiap (Villatoro-Tello et al., 2020) and thus reproduces earlier findings in other settings. Those findings could foster implicit psychometrics theory and thus advance aptitude diagnostics, which is the very reason for conduction such shared tasks.

References

Lars Binckebanck. 2019. Digitale Unterstützung für Personaler – Mitarbeitende finden mithilfe von Künstlicher Intelligenz. Published: [Online]. Available: https://www.nordakademie.de/news/digitale-unterstuetzung-fuer-personaler-mitarbeitende-finden-mithilfe-von-kuenstlicher.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Colin G. DeYoung. 2011. Intelligence and personality. In *The Cambridge handbook of intelligence*, Cambridge handbooks in psychology, pages 711–737. Cambridge University Press, New York, NY, US.

Bertram Gawronski and Jan De Houwer. 2014. Implicit measures in social and personality psychology. *Handbook of research methods in social and personality psychology*, 2:283–310.

Elisabeth Gragert, Jörg Meier, Christoph Fülscher, and NORDAKADEMIE. 2018. Campus Forum. Technical Report 66, NORDAKADEMIE, Elmshorn, Germany.

Benedikt Hell, Sabrina Trapmann, and Heinz Schuler. 2007. Eine Metaanalyse der Validität von fachspezifischen Studierfähigkeitstests im deutschsprachigen Raum. *Empirische Pädagogik*, 21(3):251–270.

Dirk Johannßen and Chris Biemann. 2018. Between the Lines: Machine Learning for Prediction of Psychological Traits - A Survey. In *Machine Learning and Knowledge Extraction - Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 International Cross-Domain Conference, CD-MAKE 2018, Hamburg, Germany, August 27-30, 2018, Proceedings*, volume 11015 of *Lecture Notes in Computer Science*, pages 192–211. Springer.

Dirk Johannßen and Chris Biemann. 2019. Neural classification with attention assessment of the implicit-association test OMT and prediction of subsequent academic success. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers*, pages 68–78, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Dirk Johannßen, Chris Biemann, and David Scheffer. 2019. Reviving a psychometric measure: Classification of the Operant Motive Test. In *Proceedings of the Sixth Annual Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*, pages 121–125, Minneapolis, MN, USA.

Dirk Johannßen, Chris Biemann, and David Scheffer. 2020. Ethical considerations of the GermEval20 Task 1. IQ assessment with natural language processing: Forbidden research or gain of knowledge? In *Proceedings of the GermEval 2020 Task 1 Workshop in conjunction with the 5th SwissText & 16th KONVENS Joint Conference 2020*, pages 30–44,

Zurich, Switzerland (online). German Society for Computational Linguistics & Language Technology.

Paul Lombardo. 2010. Three Generations, No Imbeciles: Eugenics, the Supreme Court, and Buck v. Bell. *Three Generations, No Imbeciles: Eugenics, the Supreme Court, and Buck v. Bell*, 101:1–365.

Hans P. Luhn. 1957. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM J. Res. Dev.*, 1(4):309–317.

Christopher D. Manning. 2015. Computational Linguistics and Deep Learning. *Computational Linguistics*, 41:699–705.

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

David C. McClelland and Richard E. Boyatzis. 1982. Leadership motive pattern and long-term success in management. *Journal of Applied Psychology*, 67(6):737–743.

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

Jens Nachtwei and Carsten C. Schermuly. 2009. Acht Mythen über Eignungstests. *Harvard Business Manager*, (04/2009):6–10.

James W. Pennebaker, Cindy K. Chung, Joey Frazee, Gary M. Lavergne, and David I. Beaver. 2014. When Small Words Foretell Academic Success: The Case of College Admissions Essays. *PLOS ONE*, 9(12):e115844.

Arun Rai. 2019. Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science*, 48:137–141.

John Roberts. 2014. Freddie Lee Hall, Petitioner v. Florida. URL: https://www.apa.org/about/offices/ogc/amicus/hall.

John P. Rushton and Arthur R. Jensen. 2005. Thirty years of research on race differences in cognitive ability. *Psychology, Public Policy, and Law*, 11(2):235–294.

Robert M. Sanger. 2015. IQ, Intelligence Tests, 'Ethnic Adjustments' and Atkins. SSRN Scholarly Paper ID 2706800, Social Science Research Network, Rochester, NY, USA.

Werner Sarges and David Scheffer. 2008. *Innovative Ansätze für die Eignungsdiagnostik*. Hogrefe Verlag, Göttingen, Germany.

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

David Scheffer and Julius Kuhl. 2013. *Auswertungsmanual für den Operanten Multi-Motiv-Test OMT*. sonderpunkt Verlag, Münster, Germany.

Fabian Schleithoff. 2015. Noteninflation im deutschen Schulsystem — Macht das Abitur hochschulreif? *ORDO — Jahrbuch für die Ordnung von Wirtschaft und Gesellschaft*, 66:3–26.

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

Henning Schäfer, Ahmad Idrissi-Yaghir, Andreas Schimanowski, Michael R. Bujotzek, Hendrik Damm, Jannis Nagel, and Christoph M. Friedrich. 2020. Predicting Cognitive and Motivational Style from German Text using Multilingual Transformer Architectures. In *Proceedings of the GermEval 2020 Task 1 Workshop in conjunction with the 5th SwissText & 16th KONVENS Joint Conference 2020*, pages 17–22, Zurich, Switzerland (online). German Society for Computational Linguistics & Language Technology.

Elena Semenova and David Winter. 2020. A Motivational Analysis of Russian Presidents, 1994–2018. *Political Psychology*.

Alex Tschentscher. 1977. Numerus clausus II.

Esaú Villatoro-Tello, Shantipriya Parida, Sajit Kumar, Petr Motlicek, and Qingran Zhan. 2020. Idiap & UAM participation at GermEval 2020: Classification and Regression of Cognitive and Motivational Style from Text. In *Proceedings of the GermEval 2020 Task 1 Workshop in conjunction with the 5th SwissText & 16th KONVENS Joint Conference 2020*, pages 11–16, Zurich, Switzerland (online). German Society for Computational Linguistics & Language Technology.

David Winter. 2007. The Role of Motivation, Responsibility, and Integrative Complexity in Crisis Escalation: Comparative Studies of War and Peace Crises. *Journal of personality and social psychology*, 92:920–37.

Alexander Zimmerhofer and Günter Trost. 2008. Auswahl- und Feststellungsverfahren in Deutschland - Vergangeheit, Gegenwart und Zukunft. In *Studierendenauswahl und Studienentscheidung*, 1., aufl. edition, pages 32 – 42. Hogrefe Verlag, Germany.

Çağrı Çöltekin. 2020. Predicting Educational Achievement Using Linear Models. In *Proceedings of the GermEval 2020 Task 1 Workshop in conjunction with the 5th SwissText & 16th KONVENS Joint Conference 2020*, pages 23–29, Zurich, Switzerland (online). German Society for Computational Linguistics & Language Technology.