Predicting Cognitive and Motivational Style from German Text using Multilingual Transformer Architectures

Henning Schäfer **Ahmad Idrissi-Yaghir** **Andreas Schimanowski**
Michael Raphael Bujotzek **Hendrik Damm** **Jannis Nagel** **Christoph M. Friedrich***
Department of Computer Science
University of Applied Sciences and Arts Dortmund (FHDO)
44227 Dortmund, NRW, Germany

Abstract

This document describes participation results of team *FHDO* for the first task of the GermEval 2020 competition on Classification and Regression of Cognitive and Motivational style from German texts, which was split into two subtasks. The first subtask was to create a ranking based on predicted high school grades and intelligence quotients (IQ) from freely associated texts of images and questions. Based on the predicted data, a ranking system was evaluated. The second task was the classification of an Operant Motive Test (OMT). The goal was to label the textual answers to images with one of five motives on the level of a psychologist. This work used Bidirectional Encoder Representations from Transformers (BERT) with the pre-trained *Digitale Bibliothek Münchener Digitalisierungszentrum* (DBMDZ) German model.

The best models achieved a Pearson correlation coefficient of 0.3154 for subtask 1 and an F_1-Macro score of 70.40 % for subtask 2. In the competition, the submitted model for subtask 2 achieved the best results (1st place) in all individual categories (motives, levels, motives and levels).

1 Introduction

GermEval 2020 consists of 4 different tasks. This work covers the first task (Johannßen et al., 2020), which is about long-term behavior and development of students and is split into two different subtasks.

The first subtask is to predict the measures of cognitive and motivational style. To achieve this, the z-standardized high school grades and intelligence quotients were summed and globally ranked (Johannßen et al., 2019). This ranking is artificial and it should be seen as an explorative task, yet without any real-world application, because predicting academic development is highly controversial as reflected in (Johannßen et al., 2020). The ranking was created based on actual high school grades (mathematics, German, and English grade) and intelligence quotients (IQ) for logic (logic IQ) and language (language IQ). The samples are written in German and provided by NORDAKADEMIE (NORDAKADEMIE, 2018) who are performing annually aptitude college application tests since 2011. The dataset for this subtask contains answers from 2,595 participants and includes 77,850 answers.

The second subtask is to label freely associated texts from participants on the level of a psychologist. The dataset was obtained by asking more than 14,600 volunteers to describe 15 provided images with implicit motives (OMT) (Scheffer, 2004). The OMT defines a few pre-defined questions for each image, such as "*How does the person feel?*". The answers were then classified by researchers of the University of Trier. The labeling consists of 5 different classes ("M": power, "A": affiliation, "L": achievement, "F": freedom, "0": zero) and 6 psychometric levels.

2 Previous Work

2.1 Subtask 1

For the prediction of the rank in *Subtask 1: Regression of artificially ranked cognitive and motivational style*, no prior work exists. The ranking is artificial and there is no real-world setting for applicants. The idea is to find an alternative to the Numerus Clausus (NC) to predict academic development (Zimmerhofer and Trost, 2008). NC is a selection mechanism used for college acceptance determination in Germany, applied given too many

* Corresponding author
christoph.friedrich@fh-dortmund.de

applicants for a specific course. The criticism on NC is, that applicants are measured only by a single grade. This does not reflect the potential intellectual ability of applicants. As a consequence, there are institutions looking for alternative measures.

2.2 Subtask 2

The problem based on *Subtask 2: Classification of the Operant Motive Test (OMT)* has already been discussed by Johannßen et al. (2019), who used a logistic model tree (LMT) and paired it with a broadly utilized psychometrical language analysis tool called Linguistic Inquiry and Word Count (LIWC) (Tausczik and Pennebaker, 2010).

Johannßen et al. (2019) were able to achieve a score of F_1 = 81 % on a different dataset, that did not include the motive class "F". The goal of the work was to automate the motive classification by a machine learning model. For pre-processing purposes, Johannßen et al. (2019) removed spam from the record. Identical sequences of letters, empty answers, and random collections of symbols were removed as well. Additionally, other languages besides German and texts with encoding problems were deleted.

Besides having little to no prior work specifically on the two subtasks, there have been related tasks in the psycholinguistic domain in the past, such as the categorization of social media customer feedback into sentiment classes (Hövelmann and Friedrich, 2017) and early detection of depression based on written text sequences (Trotzek et al., 2020).

3 Datasets

The datasets of the two subtasks were split into training, development, and test data, and are different in content. In both cases, the first 80 % of the dataset has been used for training and the remaining 20 % were equally split into development and test data. The following paragraph provides a short description of the datasets for the respective subtasks. For a detailed description, see (Johannßen et al., 2020).

3.1 Subtask 1

The datatset contains the image and answer number, the universally unique identifier (UUID), and the Motive Index (MIX) text. It also contains the German, English, and mathematics grade, the language IQ and logic IQ (z-standardized) as numeric values. For example, a MIX text looks like: "*Sie fühlt sich besorgt und ist verantwortungsbewußt.*" which translates to "*She feels concerned and responsible.*". Furthermore, the dataset contains the students rank. These informations are connected by a student identification (ID). The test dataset (base for the submission) is a separate file containing the students' MIX texts.

Answers consist of an average of 15 terms per document. In addition, the shortest answer consists of 3 terms and the longest of 42 terms. The standard deviation is 8 terms. See (Johannßen et al., 2020) for a detailed overview of the average grades, average IQ scores and, the standard deviation of this elements.

3.2 Subtask 2

The dataset contains the student answers of the Operant Motive Test (OMT), an ID, and the corresponding motive and level.

Answers are generally presented in German (e.g., "*sie führt das gespräch.überlegen.sie führt die situation.hänschen muss in zimmer und die kassette heut abend fällt aus.*" which translates to "*she conducts the conversation.superior.she performs the situation.small hans must be in room and the cassette tonight will be cancelled.*"). However, some of the answers are given in English (e.g., "*to give help the other, he decides to give good advice.confident and responsible.he has the capacity to help.happily for both of them.*") or French (e.g., "*Etre écoutée, elle communique avec un groupe de personnes.supérieure.elle est seule à devoir communiquer à plusieurs autres personnes.*" which translates to "*To be listened to, she communicates with a group of superiors. she is alone in having to communicate with several other people.*"). This should be considered when working on the task. Explorative data analysis revealed that some texts contain spelling and grammar errors, which should be considered as well.

The shortest answer contains 4 terms and the longest answer 79 terms. The average length of answers is 22 terms. The standard deviation is 12 terms.

Table 1 shows that the motives (rows) and levels (columns) are unbalanced, which adds more complexity to this task. The statistics shown in Table 1 are compiled based on the training dataset with 167,200 labelled text records.

	Σ	0	A	F	L	M
Σ	100 %	4.55 %	16.83 %	17.59 %	19.63 %	41.02 %
0	4.6 %	4.55 %	0.01 %	0.00 %	0.00 %	0.01 %
1	9.9 %	0.00 %	1.70 %	1.06 %	1.43 %	5.67 %
2	20.8 %	0.00 %	5.73 %	3.33 %	7.69 %	4.11 %
3	13.6 %	0.00 %	0.81 %	2.57 %	3.76 %	6.46 %
4	30.7 %	0.00 %	4.51 %	5.42 %	4.51 %	16.25 %
5	20.4 %	0.00 %	4.07 %	5.57 %	2.24 %	8.52 %

Table 1: Class distribution for subtask 2, based on training data. "M": power, "A": affiliation, "L": achievement, "F": freedom, "0": zero

4 Pre-processing

Students' answers provided for the first subtask are directly vectorized with term frequency-inverse document frequency (TF-IDF) (Spärck Jones, 1972) without any further pre-processing.

For subtask 2, 13 documents were removed since they were not labeled. The motive level in the record on line 11549 in the training dataset with the value of "4^" was assumed to be a typing error and therefore corrected to "4". Further data exploration revealed a small portion of the documents in English and French. To distinguish which documents were given in these languages, a FastText (Joulin et al., 2017) pre-trained model[1] was utilized. This resulted in finding 303 (0.18 %) French and 158 (0.09 %) English documents. Only predictions with a probability $\geq$ 0.75 were taken into account. These documents have been translated into German, using the MarianMT (Junczys-Dowmunt et al., 2018) Helsinki-NLP/opus-mt models provided in HuggingFace (Wolf et al., 2019). After the translation a combination of two spellcheckers[2,3] was used to correct the spelling mistakes in the provided documents. For the traditional systems, the documents were vectorized using TF-IDF. These vectors were combined with the LIWC features that were determined using the German version of the LIWC analysis tool. Furthermore, around 15 % of the 223,220 features were selected performing χ^2-selection (Liu and Setiono, 1995). Other pre-processing techniques, like removing punctuation or replacing German umlauts ("ä", "Ä", "ö", "Ö", "ü" and "Ü") and ligatures (e.g., "ß") were briefly tested and showed no improvements.

[1] https://fasttext.cc/blog/2017/10/02/blog-post.html, last access: 2020-06-08

[2] https://github.com/barrust/pyspellchecker, last access: 2020-06-08

[3] https://github.com/mammothb/symspellpy, last access: 2020-06-08

5 Methods

5.1 Subtask 1

For subtask 1 several models were tested, including Random Forest (RF) (Breiman, 2001), Linear Regression (LR), Neural Networks (NN - Multi Layer Perceptron) (Rumelhart et al., 1986), BERT (Devlin et al., 2018), Extreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016), and a Support Vector Regressor (SVR) – a specific form of a Support Vector Machine (SVM) (Cortes and Vapnik, 1995). The trained models from BERT, XGBoost, and SVR provided the best results (shown in Table 2). All models use TF-IDF vectorized texts. Each model was developed for all given targets (English grade, logic IQ, etc.), because none target resulted in a best correlation with the given students' text. Experimenting with the parameters of the SVR model has shown that choosing a low complexity parameter of ϵ resulted in a remarkable correlation improvement. That way, the German grade as a target performed better. Tests with changing other SVR parameters did not show notable improvements.

The SVR model provided overall the best result for the development dataset with a Pearson correlation coefficient of 0.3459 on the development set and 0.3154 on the independent test set. It was parameterized with $\gamma = 0.001$, the complexity parameters $C = 1$, and $\epsilon = 0.2$ for the English grade target. Since no model provided an unambiguous result for a specific target, it can be assumed that no single target offers an optimal prediction and a combination should be used. The final results on the independent test set are shown in Table 3.

After the competition, the three submitted models were trained using the sum of the available targets. Except for the BERT-model, improved results were achieved compared to the single target models. The XGBoost model even reached a better result than the winning submission of subtask 1 as shown in the second column of Table 3.

5.2 Subtask 2

Subtask 2 has been tested with a variety of classification models, listed in Table 4. As a baseline, a linear Support Vector Classifier (SVC) was trained on the vectorized texts with their combined motive level labels. Another classical approach was Logistic Regression on the TF-IDF vectors which outperformed the previous model. The FastText classifier was also tested and trained on the pre-

Method	m_grade	e_grade	g_grade	lang_iq	logic_iq
NN	**0.2821**	0.1674	0.1945	0.1197	0.1929
RF	**0.2187**	0.1972	0.2110	0.1080	0.0424
XGBoost	**0.2975**	0.2450	0.2926	0.1975	0.1311
SVR	0.2492	**0.3459**	0.3152	0.0255	0.1641
SVR low ϵ	0.2504	0.3120	**0.3423**	0.2035	0.1741
LR	**0.2726**	0.2065	0.1697	0.1876	0.2411
BERT	**0.3236**	0.3146	0.3058	0.1260	0.0848

Table 2: Results for subtask 1 on development set. This task uses the Pearson r correlation coefficient as a metric. Each column denotes different targets ("m_grad": maths grade, "e_grade": English grade, "g_grade": German grade, "lang_ig": language IQ.)

Method	Pearson r (submitted)	Pearson r (ex post)
FHDO_BERT_DBMDZ_uncased	0.2533	0.2208
FHDO_SVR_TF_IDF	**0.3154**	0.3427
FHDO_XGB_TF_IDF	0.2841	0.3939

Table 3: Results for subtask 1 on test set. The ex post results were obtained using the sum of the targets instead of one specific target.

processed text. In another experiment two models were trained separately, however, the approach did not perform better than the combined labels. Transformers (Vaswani et al., 2017) based models were considered to further improve results.

The BERT architecture has proven to be exceptionally effective in many downstream natural language processing (NLP) tasks, therefore it was selected for the first tests. However, this model was not trained from scratch due to computational expense. Instead, pre-trained models on German text and multilingual models were adopted. The first German model was published by the German company Deepset AI.[4] It was trained from scratch on the German Wikipedia dump, court decisions, and news articles. Yet, only a cased model was published. Cased and uncased models were afterwards published by DBMDZ,[5] trained on German Wikipedia dump, European Union (EU) bookshop corpus, Open Subtitles, and Web Crawls. In addition, cased multilingual models were tested, provided by Devlin et al. (2018) and trained on 104 languages including German. All models use the BERT base model architecture which consists of 12 transformer blocks, 12 attention heads, and 110 million parameters. It was pre-trained in two phases: (1) "masked language modeling", and (2) "next sentence prediction". In the first phase, the model predicts a percentage of random "masked" words from a sentence. In the second phase it predicts if the second sentence is the actual next sentence of the first sentence. From all tested models, the DBMDZ uncased model performed best on the development set. The cased DBMDZ model was then used in an ensemble together with the cased Deepset AI model, which turned out to be the second best model.

For the final results, the uncased DBMDZ model was able to achieve the highest score out of all participant submissions, as shown in Table 4. The hyperparamter used for training these models can be seen in Table 5.

In a different approach, documents have been translated to English using the MarianMT pretrained model, to be able to use the comprehensive English-based BERT models. The model provides overall good translation quality in both directions and was also able to capture and successfully translate wrong spelled words. BERT base and the Generalized Autoregressive Pre-training for Language Understanding (XLNet) (Yang et al., 2019) model with relative position encoding features performed well, using the translated corpus. Another tested model was a Robustly Optimized BERT Pre-training Approach (RoBERTa) (Liu et al., 2019), which is considered to be more robust because of the larger training data. Finally, Text Encoders As Discriminators Rather Than Generators (ELECTRA) (Clark et al., 2020) was tested, where a generator is trained to perform Masked Language Modeling (MLM) prior to predicting whether each token in the input was replaced by a generator sample or not, using the discriminator.

Eventually, all models based on translation could not compete with the German corpus-based finetuning.

It is worth noting that while inspecting the confusion matrices of a few models, it was recognized that the models can not fully distinguish between the motive "M" and "F". This Problem should be further explored in future works as it might result in better models.

6 Conclusion

The results show that using transformer-based NN architectures is appropriate for the classification of OMT. Despite having samples in other lan-

[4] https://deepset.ai/german-bert, last access: 2020-06-08

[5] https://github.com/dbmdz/berts, last access: 2020-06-08

Method	Precision % (dev)	Recall % (dev)	$F_{1\text{macro}}$ % (dev)	$F_{1\text{macro}}$ % (test)
FHDO__DBMDZ_uncased *	70.99	69.77	**70.67**	**70.40**
FHDO__BERT_ensemble_cased *	71.37	69.97	70.59	70.16
FHDO__DBMDZ_cased *	69.60	69.60	69.56	69.77
german-DBMDZ-uncased	70.61	70.42	70.42	70.13
german-BERT-cased *	70.11	69.80	69.89	69.93
german-DBMDZ-cased	69.92	69.80	69.80	70.00
RoBERTa-large	67.55	69.43	69.44	69.52
multilingual-BERT-cased	68.22	68.08	68.12	-
Electra-large	67.82	68.00	67.84	66.98
bert-base-cased	67.23	67.41	67.47	68.37
xlnet-large	67.43	67.24	67.56	69.00
FastText	67.07	66.08	66.48	66.86
Logistic Regression	67.20	65.74	66.30	66.87
LinearSVC	66.88	65.76	66.03	66.77

Table 4: Results for subtask 2 on development and test set. (*) denotes models where all documents have been spellchecked and English as well as French documents have been translated to German. The first three systems were used for the submission. The other results on the test set were determined after the competition.

Hyperparameter	Value
epochs	4
max sequence length	128
learning rate	6e-4
optimizer	LAMB (You et al., 2020)

Table 5: Hyperparameter used for the BERT models

guages e.g., French and English as well as samples with wrong whitespacing, and incorrect word spelling within the dataset, using extensive preprocessing did not improve the performance remarkably. It can be assumed that this happens due to the large dataset and robust transformer architecture. Nevertheless, the multilingual answers can be explained with the fact that some students do not speak German, e.g., exchange students.

While achieving good results with English-based BERT models on the translated corpus, models pre-trained specifically on German datasets were superior in all categories. For future work, different translation models will be considered to examine the predictions based on English BERT models. In a different approach, a single model was trained to predict the motive. Additional models were trained on each motive-level combination to predict the level. These results did not compete with the default approach, where motives and levels were merged into classes. This indicates a relation between motive and level of the OMT and should be further investigated in future work.

References

Leo Breiman. 2001. Random Forests. *Machine Learning*, 45(1):5 – 32.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. *CoRR*, abs/2003.10555.

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. *CoRR*, abs/1810.04805.

L. Hövelmann and C. M. Friedrich. 2017. Fasttext and Gradient Boosted Trees at GermEval-2017 on Relevance Classification and Document-level Polarity. *Proceedings of the German Society for Computational Linguistics and Language Technology (GSCL) Workshop: GermEval Shared Task on Aspect-based Sentiment in Social Media Customer Feedback, Berlin, Germany, September 12, 2017*, pages 30–35.

Dirk Johannßen, Chris Biemann, Steffen Remus, Timo Baumann, and David J. Scheffer. 2020. Germeval 2020 task 1 on the classification and regression of cognitive and emotional style from text: Companion paper. In *Proceedings of the 5th SwissText & 16th KONVENS Joint Conference 2020*, volume 2624 of CEUR Workshop Proccedings, Zurich, Switzerland (held online due to COVID19 pandemic).

Dirk Johannßen, Chris Biemann, and David Scheffer. 2019. Reviving a psychometric measure: Classification and prediction of the operant motive test. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 121–125, Minneapolis, Minnesota. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Huan Liu and Rudy Setiono. 1995. Chi2: feature selection and discretization of numeric attributes. In *Proceedings of 7th International Conference on Tools with Artificial Intelligence, ICTAI '95*, pages 388–391, Herndon, VA, USA. IEEE Computer Society.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

NORDAKADEMIE. 2018. Assessment center an der nordakademie. *Campus Forum Nr. 66/Juni 2018*, page 8.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning Representations by Back-propagating Errors. *Nature*, 323(6088):533–536.

David Scheffer. 2004. *Implizite Motive: Entwicklung, Struktur und Messung*. Hogrefe Verlag.

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

M. Trotzek, S. Koitka, and C. M. Friedrich. 2020. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*, 32(3):588–601.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: NeurIPS 2019*, pages 5754–5764, Vancouver, BC, Canada.

Yang You, Jing Li, Sashank J. Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2020. Large batch optimization for deep learning: Training BERT in 76 minutes. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Alexander Zimmerhofer and Günter Trost. 2008. *Auswahl- und Feststellungsverfahren in Deutschland - Vergangenheit, Gegenwart und Zukunft*, volume 1. Hogrefe Verlag.