Idiap & UAM participation at GermEval 2020: Classification and Regression of Cognitive and Motivational Style from Text

Esaú Villatoro-Tello[1,2], **Shantipriya Parida**[2], **Sajit Kumar**[3],
Petr Motlicek[2] **and Qingran Zhan**[2]

[1]Universidad Autónoma Metropolitana, Unidad Cuajimalpa, Mexico City, Mexico.
`evillatoro@correo.cua.uam.mx`
[2]Idiap Research Institute, Rue Marconi 19, 1920 Martigny, Switzerland.
`firstname.lastname@idiap.ch`
[3]Centre of Excellence in AI, Indian Institute of Technology, Kharagpur, West Bengal, India.
`kumar.sajit.sk@gmail.com`

Abstract

In this paper, we describe the participation of the Idiap Research Institute at GermEval 2020 shared task on the Classification and Regression of Cognitive and Motivational style from Text, specifically on subtask 2, Classification of the Operant Motive Test (OMT). Generally speaking, GermEval 2020 aims at encouraging the Natural Language Understanding (NLU) research community in proposing novel methodologies for assessing the connection between freely written texts and its cognitive and motivational styles. For evaluating this task, organizers provided a large dataset containing textual descriptions, in German language, generated by more than 14,000 participants. Our participation aims at evaluating the impact of advanced language representation, e.g., Bert, XLM, and DistilBERT in combination with some traditional machine learning algorithms. Our best configuration was able to obtain an F1 macro of 69.8% on the test partition, which represents a relative improvement of 7.4% in comparison to the proposed baseline.

1 Introduction

The idea that language use reveals information about personality has long circulated in the social and medical sciences. The ways people use words convey a great deal of information about themselves (Pennebaker et al., 2003). Psycholinguistics theory has shown the presence of linguistic indicators that could be important for determining aptitudes and academic development in subjects (Pennebaker et al., 2014), however, many of these research has focused on the analysis of self-reports or essays.

In contrast, implicit motives, indicators used by psychologies during aptitude diagnosis, are not readily accessible features to the conscious mind and, therefore, not assessable using self-reports of personal needs (Gawronski and De Houwer, 2014). Instead, implicit motives are primarily assessed using indirect measures that rely on projective techniques that instruct individuals to produce imaginative stories based on ambiguous picture stimuli that depict people in different situations. Such stimuli influence the content of the individual's fantasy and are projected onto the characters of the stories which the individual writes about from these pictures (Johannßen and Biemann, 2018, 2019; Johannßen et al., 2019). Consequently, this motivational response emerges through the contents of the written imaginative material and can be coded for its motive imagery using standardized and validated content coding systems.

The most frequently used measures of implicit motives are Picture Story Exercise (PSE) (McClelland et al., 1989), Thematic Apperception Test (TAT) (Murray, 1943), Multi-Motive Grid (MMG) (Sokolowski et al., 2000), and Operant Motive Test (OMT) (Kuhl and Scheffer, 1999; Denzinger and Brandstätter, 2018). Generally speaking, these tests are based on the operant methods, i.e., participants are asked ambiguous questions or are shown simple images, which they have to describe. Specifically, during the OMT test, subjects are shown sketched scenarios with multiple persons in non-specified situations, which required to use introspection and assess their psychological attributes unconsciously. Psychologists label these textual answers with one of five motives, namely M-power, A-affiliation, L-achievement, F-freedom, and 0-zero. And, each motive is associated with its corresponding level (from 0 to 5).

Accordingly, the “GermEval 2020 Task on the Classification and Regression of Cognitive and Emotional Style from Text”,[1] shared subtask 2, proposes an exploratory task on the Classification of the Operant Motive Test (OMT). The challenge consists of automatically processing pieces of text, generated by undergraduate students during an OMT test, and to correctly detect subjects corresponding motive/level combination.

To address the OMT task, we evaluate the impact of deep learning architectures such as Transformers (Wolf et al., 2019), namely Bert (Devlin et al., 2019), XLM (Conneau and Lample, 2019), DistilBert (Sanh et al., 2019). We compare its performance against traditional classification methods, e.g., fully connected neural networks. We compared the efficiency of these recent methodologies and compare them under different configuration parameters. Our results indicate that performing a fine-tuning of Bert is possible to obtain a 7.4% relative improvement in comparison to the proposed baseline, and the 2nd place overall during GermEval 2020 edition.

The rest of the paper is organized as follows. Section 2 describes the dataset and provides some statistics. The details of our methodology are provided in Section 3. Performed experiments and obtained results are shown in Section 3.2. Finally, we share the conclusion of our work in Section 5.

2 Dataset

To perform our experiments, we employed the dataset available in the GermEval 2020 shared task on the “Classification and Regression of Cognitive and Motivational style from the text”, described in Johannßen et al. (2020). The provided data, in German language, has been collected from more 14,600 subjects that participated in the OMT test. Each answer was manually labeled with the motives (0, A, L, M, F) and the levels (from 0 to 5), resulting in a 30 class classification problem. This annotation was performed by an expert psychologist, trained by the OMT manual as described in (Kuhl and Scheffer, 1999). The distribution of the dataset is: 167,200 for training (*train*), 20,900 for development (*dev*), and 20,900 for testing (*test*).[2]

	Training	
	Average (σ)	*Total*
Tokens	20.27 (±12.08)	3,389,945
Vocabulary	18.07 (±9.82)	267,620
LR	0.92 (± 0.08)	0.08
	Development	
	Average (σ)	*Total*
Tokens	20.38 (±12.17)	425,880
Vocabulary	18.17 (±9.94)	55,606
LR	0.92 (± 0.08)	0.13
	Test	
	Average (σ)	*Total*
Tokens	20.24 (±12.01)	423,018
Vocabulary	18.05 (±9.76)	55,592
LR	0.92 (±0.08)	0.13

Table 1: Statistics of the OMT dataset in terms of number of tokens, vocabulary size and lexical richness. The minimum length of the texts is 1 token, while the maximum length is 99, 90, and 96 tokens for *train*, *dev*, and *test* partitions respectively. In all partitions, the 75% of the data has a length of 27 tokens.

Table 1 shows some statistics of the GermEval 2020 dataset, for *train*, *dev*, and *test* partitions. We compute the average number of tokens, vocabulary, and lexical richness of each text in the dataset. Lexical richness (LR), also known as “type/token ratio” is a value that indicates how the terms from the vocabulary are used within a text. LR is defined as the ratio between the vocabulary size and the number of tokens from a text ($LR = |V|/|T|$). Thus, a value close to 1 indicates a higher LR, which means vocabulary terms are used only once, while values near to 0 represent a higher number of tokens used more frequently (i.e., more repetitive).

Two main observations can be done at this point. On the one hand, notice that for the three partitions (i.e., *train, dev, and test*), textual descriptions are very short, on average 20 tokens with a vocabulary of 18 words, resulting in a very high LR (0.92). The high LR value means that very few words are repeated within each textual description, i.e., very few redundancies. On the other hand, globally speaking, the complete dataset has a low LR (0.08 for *train* and 0.13 for *dev* and *test*). Although these values are not directly comparable due to the size of each partition, they indicate, to some extent, that information across texts is very repetitive, i.e., simi-

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

[2] During our experimentation a total of 13 instances were removed from the training partition due to its lack of label, leaving 167,187 instances.

lar types of words are being used by tested subjects for describing different images, even though they belong to different classes (motives and levels). We are aware of the necessity from a deeper analysis of the data in order to reach concrete conclusions about the nature of the texts; however, this initial analysis helped us to envision the complexity and nature of the data.

3 Methodology

We aim to automate the annotation of participant responses for the OMT task by training a machine learning model. Machine learning (ML) models as such cannot use raw text as input. Therefore it is necessary to transform the input to a feature representation understandable by the model. Accordingly, we evaluate two ML approaches for solving the OMT task: fine-tuning of transformers based architectures (Section 3.1), and a traditional fully-connected neural network (Section 3.2).

It is important to mention that instead of facing the OMT task as a 30 class classification problem, we split the problem into two separate classification tasks: motives (5 classes), and levels detection (6 classes). For each of classification problem, we applied the exact same methodology as described in the following sections. Finally, in order to produce the required output by the organizers, we merge the predicted motive and the predicted level for every instance.

3.1 Simple Transformer

The transformer model (Vaswani et al., 2017) introduces an architecture that is solely based on attention mechanism and does not use any recurrent networks but yet produces results superior in quality to Seq2Seq (Sutskever et al., 2014) models, incorporating the advantage of addressing the long term dependency problem found in Seq2Seq model.

For our experiments using Simple Transformers (ST) architectures, we setup three different configurations:

1. Bert (Devlin et al., 2019): we use a pre-trained model referred as `bert-base-german-cased`, with 12-layer, 768-hidden, 12-heads, 110M parameters.[3] The model is pre-trained on German Wikipedia dump (6GB of raw text files), the OpenLegalData dump (2.4 GB), and news articles (3.6 GB). We refer to this configuration as ST-Bert in our experiments.
2. XLM (Conneau and Lample, 2019): for this configuration we use a model with 6-layer, 1024-hidden, 8-heads, which is an English-German model trained on the concatenation of English and German Wikipedia documents (`bert-base-german-cased`). We refer to this configuration as ST-XLM in our experiments.
3. DistilBert (Sanh et al., 2019): fir this model we used a model with 6-layer, 768-hidden, 12-heads, 66M parameters (`distilbert-base-german-cased`). We refer to this configuration as ST-DistilBert in our experiments.

For all the previous configurations, in order to perform the fine-tuning of the ST architecture, we added an untrained layer of neurons on the end, and re-train the model for the OMT classification task. To perform these experiments, we used the Simple Transformers library which allows us to easily implement the proposed idea.[4] For all the experiments done using simple transformers architecture we set the `max_length` parameter to 90, and we re-trained the models up to 2 epochs. Further details of employed models can be found at huggingface web page.[5]

Hyper Parameter	Range
number of layers	3
number of hidden layers	1
nodes in hidden layer	16
activation function	ReLU

Table 2: Fully connected neural network configuration parameters.

3.2 Fully Connected Neural Network

As an additional classification method, we configured a fully connected neural network (FC). This type of artificial neural network is configured such that all the nodes, or neurons, in one layer, are connected to all neurons in the next layer. The network and configuration parameters are mentioned in Table 2.

For our performed experiments using FCs, we passed as input features to the FC the sentence rep-

[3] https://deepset.ai

[4] https://pypi.org/project/simpletransformers

[5] https://huggingface.co/transformers/pretrained_models.html

Method	*Configuration type*	*Configuration sub-type*	**F1-macro** *(dev)*	**F1-macro** *(test)*
ST	Bert	bert-base-german-cased	**0.694**	**0.698**
ST	XLM	xlm-mlm-ende-1024	0.688	0.686
ST	DistilBert	distilbert-base-german-cased	0.692	0.688
FC	Bert (pre-trained)	LHL	0.589	0.589
FC	Bert (pre-trained)	Concat4LHL	0.616	0.579
FC	Bert (fine-tuned)	LHL	0.673	0.671
FC	Bert (fine-tuned)	Concat4LHL	0.675	0.230
Baseline	SVM	*tf-idf*	0.639	0.644
1st place	–	–	–	0.704

Table 3: Obtained results on the *dev* and *test* partitions of the OMT classification task. Results are reported in terms of the F1 macro measure. Baseline and 1st place results were extracted from the companion paper (Johannßen et al., 2020).

resentation generated using Bert encoding. Thus, to generate the representation of the sentence, we evaluate several configurations, namely: last hidden layer (LHL), concatenation of the 4 last hidden layers (Concat4LHL), min, max and mean pool of the last hidden layers. However, we only report the best performances obtained during the validation stage, i.e., LHL and Concat4LHL configurations. On the one hand, for generating the Concat4LHL representation we concatenate the last four layers values from the token CLS. As known, the CLS token at the beginning of the sentence is treated as the sentence representation. On the other hand, for the LHL configuration, we preserve as the sentence representation the values of the last hidden layer from the token CLS.

For the reported experiments under the FC method, two configurations of Bert were tested for generating the LHL and Concat4LHL representation: i) pre-trained German encodings of Bert (`distilbert-base-german-cased`), referred as Bert(pre-trained); and ii) resultant fine-tuned Bert encodings from the re-training we explained in Section 3.1, referred as Bert(fine-tuned).

4 Experiments and Results

The results of each considered method are shown in Table 3. The proposed baseline by the GermEval 2020 organizers, is a linear Support Vector Classifier (SVC) using as a form of representation of the documents a traditional *tf-idf* strategy, specifically a 30 (combined motive/level labels) binary SVCs (one-vs-all) classifiers. Results are reported in terms of F1-macro, for both *dev* and *test* partitions. As can be observed in table 3, the proposed baseline obtains an F1=64.4%, representing a strong base method. During the competition, the best reported performance was an F1 macro of 70.4% (last row of Table 3).

During the validation stage, the best result using the FC method was obtained under the Concat4LHL configuration, i.e., when texts are represented using as features the concatenation of the four last hidden layers from 'Bert (fine-tuned)' model. However, notice that the same configuration obtained the worst performance during the test stage (23%). We think that some errors occurred during the setup of the output file, or at worst, maybe some error occurred during final training, provoking some overfitting situation. In spite of this result, the 'Bert fine-tuned' consistently improves the performance of the experiments using a fully connected neural network. Particularly, during the development stage, both experiments using the fine-tuned version of Bert outperformed the same configuration that uses the pre-trained version of Bert. Except for the FC(Bert pre-trained-Concat4LHL), a similar situation occurred during the test phase, i.e., adjusting the attention of Bert to the OMT task, helped the FC method for obtaining a more relevant results.

Finally, the best performance was obtained by the simple transformers architectures. As expected, the best performance is obtained when the Bert model is employed, followed by DistilBert and XML models. Generally speaking, ST-BERT configuration obtains a relative improvement of 7.4% over the competition baseline. Overall, the obtained performance by the three considered configurations exhibits marginal differences, thus, the performance obtained by the DistilBert could be considered a very good alternative given that represents a significantly smaller, faster, cheaper and lighter transformer model.

5 Conclusion

This paper describes Idiap's participation at the GermEval 2020 shared task on the Classification and Regression of Cognitive and Motivational Style from the text. Our participation aimed at analyzing the performance of recent NLP technologies for solving the OMT classification task. To this end, we performed a comparative analysis among Simple Transformers based architectures, e.g., Bert, XLM, and DistilBert, and traditional machine learning techniques. Notably, transformers based methods exhibit the best empirical results, obtaining a relative improvement of 7.7% over the baseline suggested as part of the GermEval 2020 challenge. Overall, our system obtained the second-best place in terms of the F1 macro among participant teams during the GermEval 2020 edition.

As future work, we plan to evaluate the impact of hyperparameter tuning through optimization methods, such as Bayes optimizer (Snoek et al., 2012), and to perform further analysis on how the attention-mechanism from the transformers architecture is working in the OMT task.

Acknowledgments

The work was supported by an innovation project (under an InnoSuisse grant) oriented to improve the automatic speech recognition and natural language understanding technologies for German. Title: "SM2: Extracting Semantic Meaning from Spoken Material" funding application no. 29814.1 IP-ICT and EU H2020 project "Real-time network, text, and speaker analytics for combating organized crime" (ROXANNE), grant agreement: 833635. The first author, Esaú Villatoro-Tello is supported partially by Idiap, SNI-CONACyT, CONACyT project grant CB-2015-01-258588, and UAM-C Mexico during the elaboration of this work.

References

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067.

Ferdinand Denzinger and Veronika Brandstätter. 2018. Stability of and changes in implicit motives. a narrative review of empirical studies. *Frontiers in psychology*, 9:777.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

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.

Dirk Johannßen and Chris Biemann. 2018. Between the Lines: Machine Learning for Prediction of Psychological Traits - A Survey. LNCS-11015:192–211. Part 2: MAKE-Text.

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, Steffen Remus, Timo Baumann, and David Sheffer. 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*, pages 1–9.

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.

Julius Kuhl and David Scheffer. 1999. Der operante multi-motiv-test (omt): Manual [the operant multi-motive-test (omt): Manual]. *Germany: University of Osnabrück*.

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

H.A. Murray. 1943. *Thematic Apperception Test*. Harvard University Press.

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.

James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959.

Kurt Sokolowski, Heinz-Dieter Schmalt, Thomas A Langens, and Rosa M Puca. 2000. Assessing achievement, affiliation, and power motives all at once: The multi-motive grid (mmg). *Journal of personality assessment*, 74(1):126–145.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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