Predicting independent living outcomes from written reports of social workers

Angelika Maier
Bielefeld University
Inspiration 1
33619 Bielefeld
amaier@uni-bielefeld.de

Philipp Cimiano
Bielefeld University
Inspiration 1
33619 Bielefeld
cimiano@cit-ec.uni-bielefeld.de

Abstract

In social care environments, the main goal of social workers is to foster independent living by their clients. An important task is thus to monitor progress towards reaching independence in different areas of their patients' life. To support this task, we present an approach that extracts indications of independence on different life aspects from the day-to-day documentation that social workers create. We describe the process of collecting and annotating a corresponding corpus created from data records of two social work institutions with a focus on disability care. We show that the agreement on the task of annotating the observations of social workers with respect to discrete independent levels yields a high agreement of .74 as measured by Fleiss' Kappa. We present a classification approach towards automatically classifying an observation into the discrete independence levels and present results for different types of classifiers. Against our original expectation, we show that we reach F-Measures (macro) of 95% averaged across topics, showing that this task can be automatically solved.

1 Introduction

Social workers are concerned with improving abilities and increasing confidence of their clients with the goal of supporting them in living their lives independently. A traditional taxonomy, called Metzler's taxonomy[1] (Metzler, 2001; Kommission, 2016), considers the following life categories in which independence is to be reached: *'everyday life', 'individual basic care', 'relationships', 'participation in cultural and social life', 'communication', 'emotional and psychological development'* and *'health promotion and maintenance'*. Each category has specific subcategories, with overall 34 categories. Social workers document their observations with respect to the behaviour of clients in day-to-day records that capture the whole trajectory of their clients. Interviews carried out with social workers have revealed that they would profit substantially from automatic summarization of the trajectories of patients, in particular their progress over time on reaching independence in the different categories defined by Metzler.

Towards this goal, in this paper we investigate whether it is possible to automatize this task. We frame the task as a classification problem in which each observation in the records about the patient is classified into a discrete independence level. In particular, we distinguish the following five independence levels: *1 ('able to accomplish a certain task alone')*, *2 ('able to accomplish a given task with help')*, *3 ('partly able to accomplish a task with help')*, *4 ('unable to accomplish a certain task')*. The neutral label 5 applies for documentations that do not allow a conclusion regarding the level of independence.

In Table 1, we provide one example for each of those independence levels for different Metzler categories. These examples are derived from the actual data, slightly rephrased to ensure anonymity.

As one contribution of this paper, we describe the process of collecting a corpus on the basis of data from two social care institutions. The corpus has been annotated by social work students that for our purposes can be regarded as domain experts. After several annotation rounds, the annotator agreement reached 0.74 as measured by Fleiss' Kappa, corresponding to a substantial agreement.

Building on this corpus, we train different classifiers on the annotated data and show that the

[1] https://www.soziales.niedersachsen.de/startseite/behinderte_menschen/eingliederungshilfe_behinderte_menschen/bedarfsfeststellung-des-bedarfs-gruppen-fuer-leistungsberechtigte-mit-vergleichbarem-bedarf-94870.html (as consulted online 11.10.2020), there, the taxonomy can be found in appendix 2.4

Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science, pages 139–148

https://doi.org/10.18653/v1/P17

classifiers can reach a macro F-Measure of 95% averaged across the 34 Metzler categories on the task of predicting independence levels. The best results overall are obtained with a linear SVM (one-vs-one) classifier. The best result is achieved on Metzler's categories *M-6 ('Managing money')*, *M-11 ('Get up / Go to bed')*, *M-17 ('Organization of free time / private activities')*, *M-19 ('Encounters in social groups')*, *M-22 ('Compensation of sensory impairments and communication disorders')*, *M-24* and *M-25* on orientation in familiar and unfamiliar surroundings, *M-27* and *M-28* on coping with psychological disorders (100% F-measure), while the worst result is achieved on Metzler's category *M-21 ('Development of future perspectives, life planning')* (92% F-Measure). We rely on a feature ablation experiment to determine the importance of different features on the task.

1 is able to accomplish a certain task

Category	M-7
Documentation	Client has filled and dispatched the notification for health insurance.

2 is able to accomplish a certain task with help

Category	M-15
Documentation	Client was reminded of the visit of his brother. He went on a trip with his caretaker and his brother.

3 is partly able to accomplish a certain task with help

Category	M-8
Documentation	Client had dinner. The food was brought to her. Sometimes she led the spoon to her mouth by herself.

4 is not able to accomplish a certain task

Category	M-27
Documentation	Client stayed in her room this morning. Didn't want to come down for lunch and didn't want to go to the meeting center.

5 neutral / irrelevant for category

Category	M-23
Documentation	Client says he has a cold and is not feeling well.

Table 1: Examples for each of the independence levels for different Metzler categories.
M-7: Deal with financial matters and social law matters, M-15: Relationship with relatives, M-8: Nutrition, M-27: Coping with drive disorders, M-23: Time orientation

The paper is structured as follows: In Section 2, we discuss related work for the problem of predicting extra-linguistic personal attributes. In Section 3, we describe how the data was obtained from two different disability care institutions. In Section 4, the method is described, including the development of guidelines for annotation and a description of the annotation process. We also describe which classifiers we use to automatize the classification. Before concluding, we discuss our results.

2 Related Work

There has been considerable work on the task of predicting personal attributes of users on the basis of their written contributions, e.g. social media posts. Personal attributes are typically extra-linguistic attributes that are not explicitly mentioned in texts, but can be inferred on the basis of analysis of style, grammar or vocabulary. One can distinguish between physical personal attributes that manifest themselves physically (such as age, ethnicity, gender, etc) on the one hand, and psychological personal attributes including mood, emotion, stress level, sentiment, resilience etc. The independence level that we consider in our work does not fully belong to one of this categories. While independence level is not strictly speaking a physically manifested attribute such as age or gender, it is related to the (observable) behaviour of a person and has thus an obvservable manifestation. Yet, it has similar characteristics to a psychological variable in the sense that it is not objectively measurable and is subject to interpretation.

An older study was carried out by Baumrind (1967) in the field of child care to manually analyze written documentations in combination with the environment of children to assess child care behaviour as a basis to predict their preschool behaviour patterns. Among the different attributes considered, level of independence is considered.

Another group where the level of independence is of relevance are old people (Araújo and Ceolim, 2007; Karakaya et al., 2009). Karakaya et al. collected data for 33 elderly people living in a nursing home and 25 elderly living at home. They measured the functional mobility, depressive symptoms, level of independence, and quality of life for both groups for comparison. The level of independence was evaluated by the Kahoku Aging Longitudinal Study Scale (KALS), which evaluates activities of elderly people in 12 areas. Each activity was rated on a 4-point scale (0: dependent, 1-2: some help, 3: independent). In this scale, higher scores indicate higher level of inde-

pendence. Our assessment of independence level follows in essence the proposal of the 4-point scale of Karakaya et al.

While there has been some work on (manual) independence level classification (e.g. the above mentioned works), we are the first to consider the automatic prediction using learned classifiers on textual documentations in the domain of disability care in social work. Furthermore, we investigate how the classification results vary depending on the Metzler category considered. An important difference to related work is that level of independence is not predicted from the self-reported experience of subjects, but from the written observations of a third party, in our case the social worker.

Due to the lack of work on automatic prediction of independence level, we discuss work on the related task of classifying personal attributes. There has been a lot of work on detecting personal attributes, in particular from social media posts from Facebook or Twitter (Kosinski et al., 2013; Yo and Sasahara, 2017).

Kosinski et al. (2013) use logistic and linear regression classifiers to detect psychological attributes related to *personality*, *intelligence*, *openness* and *happiness*, as well as physical attributes related to *sexual orientation*, *ethnicity*, *religious and political views*, *use of addictive substances*, *parental separation*, *age*, and *gender*, reaching accuracy levels between 85% and 88% on the data of 58,000 volunteers on Facebook. Yo and Sasahara (2017) tackle the prediction of the personal attributes *gender*, *occupation*, and *age groups* in tweets, reaching accuracy levels of 60-70%. A related task is the detection of stress levels, on which Lin et al. (2014) reach 83%-93% F-Measure on four different datasets derived from 350 million tweets data. Galatzer-Levy et al. (2018) extensively discuss the relevance of different machine learning algorithms and machine learning principles to the study of stress pathology, recovery, and resilience.

A related task is the classification of the emotion of a text or social media post. Bostan and Klinger (2018) compare models for prediction on annotated emotion corpora systematically, performing cross-corpus experiments by training classifiers on each dataset and evaluating them on others, reaching F-Measures (micro averaged) of 56% when training on all but one corpus and testing on a held-out corpus and 98% when training and testing on the same corpus. They predict the emotions: *joy, anger, sadness, disgust, fear, trust, surprise, love, confusion, anticipation* and *no emotion* (noemo). Cevher et al. (2019) constructed the AMMER corpus, triggering emotions such as *fear, anger, annoyance, insecurity, joy* and *no emotion* in the experimental setting of a car driving situation. They make use of off-the-shelf algorithms and a bidirectional LSTM, reaching a F-Measure of 76% with Transfer Learning. Schuff et al. (2017) have re-annotated the SemEval 2016 Stance Data set (Mohammad et al., 2016) with the emotion labels *anger, anticipation, disgust, fear, joy, sadness, surprise* and *trust*. They apply Maximum Entropy, Support Vector Machines, a LSTM, a Bidirectional LSTM, and a Convolutional Neural Network (CNN) to provide baseline results for their corpus. They reach best results of 77% F-Measure with a bidirectional LSTM.

3 Data

The data was collected from 2 different stationary disability care facilities of a major European welfare provider located in the German State of North Rhine-Westphalia in the context of a common project with Bielefeld University. Following contractual obligations, the institution will not be named in this paper. The dataset was collected for clients that gave their explicit consent, yielding the complete documentation for 22 clients. The data comprises of 731,601 records with 295,812 observations documented by the social workers in natural language.

Client	Category	Date
1235	M-15	12. März 2016

Text

ANONYM hatte heute Besuch von seiner Mutter.
Sie haben einen Spaziergang gemacht.
engl: ANONYM‘s mother visited him today.
They went for a walk.

Table 2: Example for a documentation of Metzler category *M-15 (‘Shaping social relationships with relatives’)*.

For each documentation, a client ID, a date, and a category is recorded. An example for a documentation is given in Table 2. Documentations can be very short containing only one or two words

up to several sentences to provide a detailed description for the social workers of the following shift. The Metzler category of a documentation is assigned by the social worker.

The 34 categories describe measures of care applied and originate from the Metzler taxonomy that is standard in social care contexts for the purpose of documentation. The seven top level categories of the taxonomy and the respective number of subcategories are given in Table 3.

Area	**Number of subcategories**
Everyday life	7
Individual basic care	6
Relationships	3
Participation in cultural and social life	5
Communication	4
Emotional and psychic development	4
Health promotion and maintenance	5

Table 3: Top-level Metzler categories and number of subcategories.

4 Method

In this section we discuss the development of guidelines for annotation and provide a description of the annotation process. We also describe which classifiers we use to automatize the classification.

4.1 Guideline development

We used two small datasets, comprising of 677 and 500 documentations, respectively, for the development of guidelines. The guidelines include general annotation principles and examples for each level of independence for each Metzler category and a list of abbreviations that are repeatedly used in documentations (e.g. MA: Mitarbeiter (engl. employee or co-worker)). Due to contractual obligations, we are only allowed to share the general principles for annotation but not the examples (even if they were generalized and made unfamiliar for the annotation guidelines). In the example section of the guidelines, for each level of independence a short definition is given, as well as some typical phrasing for documentations of this level. Annotators were asked to rely only on the information in the social workers' written comment when classifying the level of independence. Further, they were asked not to judge or evaluate the effect of independence on the quality of life of the individual. For example, they were asked not to judge whether in their view the effect of a client being able to spend money independently was positive or not, that is, not to take into account whether the money was spent wisely or not.

4.2 Annotation Process

For the annotation, we employed 3 advanced bachelor students (semester 3 and higher) in the field of social work, with practical experience in disability care institutions. The annotation was done in two steps. The first 5,000 data points were annotated by four subjects: the first author and the 3 students. A second batch of data comprising of 8,313 data points was annotated by the students only. The annotation has been conducted with 'OMEN - a collaborative, annotation platform' [2]. For annotation, a single documentation and the respective Metzler category are presented and annotators had to choose the appropriate level of independence. The annotators were instructed to use a short version of the annotation guidelines as well as the Metzler category for reference during the annotation process. The training of the annotators and optimization of the guidelines has been conducted in two iterations. In order to estimate the inter-annotator agreement, Fleiss's kappa was calculated (Wirtz and Caspar, 2002). In the first annotation round of 677 documentations (about 20 documentations per Metzler category), the agreement between the annotators reached a κ-value of 0.59 (on document level). After discussion, a further independent annotation round of new 500 documentations (about 15 documentations per Metzler category) resulted in κ= 0.66, showing that the annotators converged in their understanding of the task. The results from the second round were also discussed and used for further modification of the guidelines. The 15,719 documentations selected for annotation were annotated in two steps. In the first subsequent independent annotation step involving 5,000 documentations, an agreement of κ= 0.74 was reached, which can be regarded as a substantial agreement

[2] https://github.com/FrankGrimm/omen

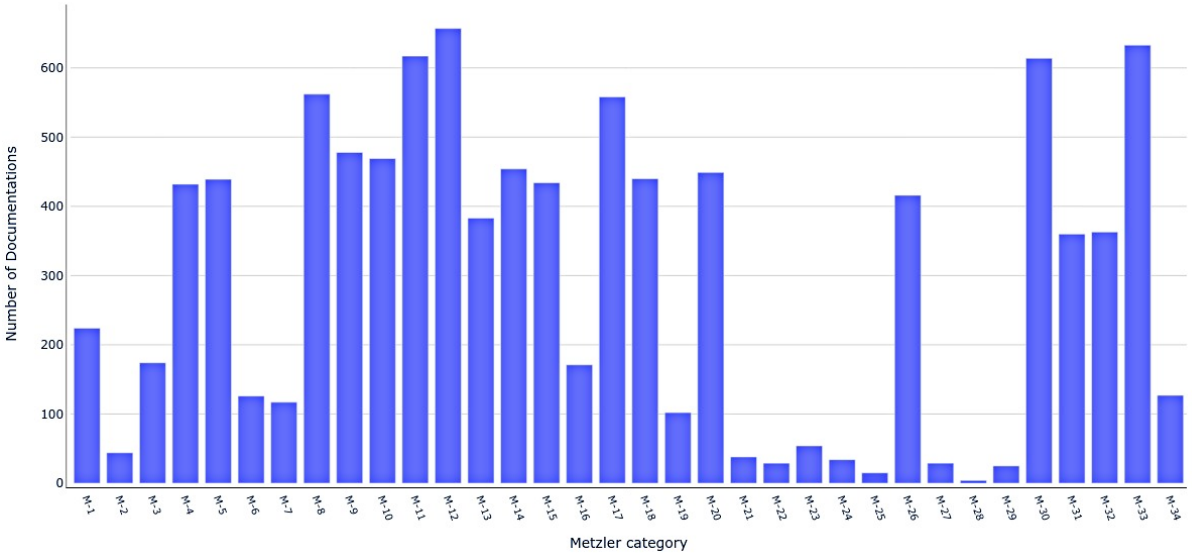

Figure 1: Number of documentations per Metzler category in annotated dataset selected with majority vote. N=10,071

in comparison to agreement by chance[3]. The remaining 8,313 documentations were annotated in a second subsequent annotation step by 3 annotators.

4.3 Classification models

On the basis of the annotated data described above, we train models to predict the level of independence in a supervised setting. For this purpose, we aggregate the annotated data with majority vote, i.e. only use documentations from the data annotated by 4 annotators where 3 or 4 of the annotators agree and only use documentations from the data annotated by 3 annotators where 2 or 3 of the annotators agree. The distribution of independence levels over the selected documentations can be found in Figure 2. Also, we report the number of documentations for each Metzler category in Figure 1.

We compare two settings: a setting where all the data is used to train one model that predicts level of independence irrespective of the Metzler category (category-agnostic classification) and one setting where there is one model per category (category-specific classification). We perform experiments with the following classification algorithms: linear Support Vector Machines (SVMs), AdaBoost, Random Forest, Logistic Regression, Perceptron, k-NN, Multinomial Naive Bayes and

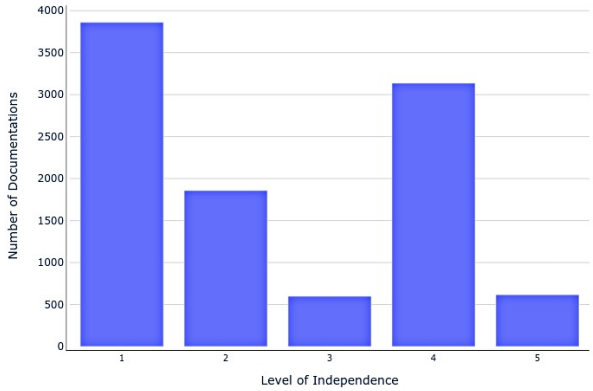

Figure 2: Number of documentations for each level of independence in annotated dataset selected with majority vote. N=10,071

Multi-layer Perceptron (MLP) as well as a convolutional neural network (CNN), with a convolutional window of 3.

We use the implemented version of these algorithms in the Python module scikit-learn (Pedregosa et al., 2011). For the CNN, we use the implementation in Keras (Chollet et al., 2015) with a convolutional window of 3, consisting of an embedding layer, two 1D convolutional layers, a dropout layer, a pooling layer and a dense layer. For the CNN, we use embeddings induced from our dataset rather than using pre-trained embeddings.

All names were substituted with the place-

[3] According to the scale of kappa value interpretation of Landis and Koch (1977)

holder 'ANONYM' before the annotation process. We pre-process the text data using standard pre-processing operations such as lemmatization, stemming and stopword removal. We use spaCy (Honnibal and Montani, 2017) to obtain the lemmata and POS-tags. All features are represented as tf.idf values in the models. POS-tags are glued to the words in a pre-processing step for this purpose. We compute embeddings specific for our dataset using all 295,812 documentations and all 636 unique texts for measures of care applied[4] (the wording over all data is very similar) to train our own word embeddings for our domain with the python module fasttext (Bojanowski et al., 2017). Also, we up-sample the datasets for training to have an equal distribution of the classes.

As a baseline, we consider models trained using unigram features weighted with tf.idf. We experimentally test different feature combinations including the following features for our models: trigrams, lemmata, stemmed words, word embeddings (only in CNN), POS-tags. This leads to 15 feature combinations (8 without POS tags and word embeddings for CNN).

5 Results

We adopt a 5-fold cross-validation setting in which in each fold we train with 80% of the data and evaluate on 20% of held-out data. We report the median F-measure over the five folds for each classifier. As a baseline, we rely on a model trained with unigrams weighted with tf.idf. We refer to this baseline as Bag-Of-Words (BOW) model. The only model with a different baseline is the CNN, where the convolutional window is of size 3, corresponding to the trigram features that we use on top of the BOW model. Therefore, we use the model with word embeddings that are trained on the raw text – rather than stemmed words or lemmata – as a baseline. The baseline score for the CNN is 6% F-Measure (macro) (median over five folds). The scores for the different classifiers using the BOW model are given in Table 4.

We report the F-Measure (macro) for the five best models in Table 5. The left part of the table shows the results for the category-specific setting in which we train one model per Metzler-category. The right part of the table shows the results for all

[4]Measures of care applied are another part of the data we obtained. It comprises of over 400,000 samples, but is more standardized than the documentations.

Baseline (BOW Model)	F1 (macro)
Linear SVM (one-vs-one)	0.86
Linear SVM (one-vs-rest)	0.85
Perceptron	0.85
LogisticRegression	0.85
KNN	0.81
Multinomial NB	0.80
RandomForestClassifier	0.69
AdaBoostClassifier	0.48
MLPClassifier	0.47

Table 4: Baseline scores with tf.idf of Bag-of-words : F-Measure (macro) (Median of 5-fold cross-validation) for models trained on all documentations; BOW=Bag-of-words

categories for the category-agnostic classification, that is the case where there is one single model that predicts independence levels independently of the Metzler category.

First of all, we observe that the Macro-average F_1 values in both settings are above the BOW baseline by 7% (category-specific setting) and 9% (category-agnostic setting) for the case of linear SVM and 7% and 10% for Perceptron, respectively. This trend is observed also for Logistic regression but not for KNN, where the value in the category-specific-setting increases 8% and drops 4% in the category-agnostic setting. For the CNN the value increases 82% in the category-specific setting and 52% in the category-agnostic setting. In general, we observe that the category-agnostic setting outperforms the category-specific setting for most classifiers , i.e. 2% for linear SVM, 3% for Perceptron, and 1% for Logistic regression. The exception is again the KNN classifier, where the value is 12% lower. Also, the CNN performs 30% better in the category-specific setting.

Considering the best-performing classifier, i.e. linear SVM (one-vs-one) for the specific categories, we observe that for 11 out of 34 categories, the category-agnostic setting performs better than the category-specific setting, for 9 categories both settings perform equally well and for 14 categories the category-specific setting performs better. The category-agnostic setting sometimes outperforms the category-specific setting in cases where there are a low number of (positive) examples for the corresponding category. This is the case for the categories *M-21*, *M-27* and *M-28*. In *M-28*, it is not possible to conduct 5-fold cross-validation

Metzler category	category-specific classification					category-agnostic classification					
	SVM OvO	**Per-cep-tron**	**LR**	**KNN**	**CNN**	**SVM OvO**	**Per-cep-tron**	**LR**	**KNN**	**CNN**	**Sup-port**
All	**0.97***	0.96	0.94	0.90	0.75	n/a	n/a	n/a	n/a	n/a	10071
M-1	**0.95***	0.94	0.90	0.87	0.85	**0.95***	0.93	0.89	0.76	0.75	224
M-2	0.91	0.91	0.91	0.86	**0.94***	**0.94***	**0.94***	0.70	0.63	0.50	44
M-3	0.97	**0.98***	0.97	0.95	0.91	**0.93**	0.90	0.86	0.74	0.56	174
M-4	**0.96***	0.94	0.95	0.93	0.95	**0.96***	**0.96***	0.94	0.85	0.77	432
M-5	**0.95**	0.94	0.92	0.92	**0.95**	**0.98***	0.97	0.94	0.82	0.70	439
M-6	**1.00***	0.99	0.97	0.96	0.99	0.94	**0.96**	0.91	0.79	0.43	126
M-7	**0.87**	**0.87**	0.78	0.77	0.79	**0.97***	0.96	0.94	0.80	0.47	117
M-8	**0.98***	0.97	0.96	0.95	0.96	**0.96**	0.95	0.95	0.79	0.78	562
M-9	**0.96***	**0.96***	0.95	0.89	**0.96***	**0.95**	**0.95**	0.93	0.78	0.80	478
M-10	**0.97**	**0.97**	0.96	0.95	0.96	0.97	**0.98***	0.95	0.87	0.75	469
M-11	**1.00***	0.99	0.99	0.97	0.99	**0.92**	0.87	0.88	0.77	0.62	617
M-12	**0.97***	0.96	0.95	0.95	0.92	**0.96**	0.93	0.94	0.81	0.80	657
M-13	**0.96***	**0.96***	0.92	0.93	0.92	**0.96***	0.95	0.90	0.85	0.72	383
M-14	**0.99***	0.97	0.97	0.96	0.96	**0.94**	0.90	0.89	0.71	0.63	454
M-15	**0.99***	**0.99***	0.97	0.96	0.98	**0.99***	0.97	0.93	0.85	0.70	434
M-16	**0.93**	0.92	0.88	0.85	0.92	0.92	**0.94***	0.90	0.77	0.58	171
M-17	**1.00***	0.99	0.99	0.98	0.98	**0.95**	0.92	0.89	0.74	0.55	558
M-18	**0.98***	**0.98***	0.97	0.95	0.96	**0.97**	0.96	0.94	0.86	0.76	440
M-19	**1.00***	**1.00***	0.99	0.99	0.98	**1.00***	0.95	**1.00***	0.87	0.84	102
M-20	**0.96***	0.95	0.94	0.91	0.94	**0.96***	**0.96***	0.90	0.71	0.65	449
M-21	0.84	0.82	**0.83**	0.82	0.69	**0.92***	**0.92***	**0.92***	0.74	0.44	38
M-22	0.95	**1.00***	0.95	0.94	0.93	**1.00***	**1.00***	0.82	0.82	0.00	29
M-23	**0.94***	**0.94***	0.84	0.73	0.79	0.63	**0.88**	0.63	0.54	0.73	54
M-24	**1.00***	**1.00***	**1.00***	**1.00***	**1.00***	**1.00***	**1.00***	0.75	0.82	0.33	34
M-25	**1.00***	**1.00***	**1.00***	**1.00***	**0.87**	**1.00***	**1.00***	**1.00***	0.80	0.00	15
M-26	**0.98***	**0.98***	0.97	0.95	0.97	**0.95**	**0.95**	0.93	0.74	0.68	416
M-27	**0.78**	0.75	0.67	0.67	0.58	**1.00***	**1.00***	0.97	0.69	0.50	29
M-28	n/a	n/a	n/a	n/a	n/a	**1.00***	**1.00***	**1.00***	0.50	0.00	4
M-29	**0.92**	0.90	0.83	0.90	0.80	**0.93***	0.89	0.89	0.60	0.00	25
M-30	**0.98***	**0.98***	0.97	0.96	0.96	**0.97**	**0.97**	0.94	0.86	0.58	614
M-31	**0.93**	0.91	0.87	0.82	0.87	0.96	**0.97***	0.93	0.75	0.66	360
M-32	**0.98***	0.97	0.96	0.95	**0.98***	**0.96**	0.94	0.95	0.79	0.67	363
M-33	**0.99***	**0.99***	0.98	0.97	0.96	**0.94**	**0.94**	0.91	0.81	0.75	633
M-34	**0.93**	**0.93**	0.91	0.87	0.92	**0.98***	**0.98***	0.96	0.82	0.88	127
macro F_1	**0.93**	0.92	0.90	0.89	0.88	**0.95***	**0.95***	0.91	0.77	0.58	

Table 5: Best F-Measure (macro) (Median of 5-fold cross-validation) and category-wise F-Measure (macro) (Median of 5-fold cross-validation) with model trained on all Metzler categories for best 5 models and macro average over categories, * marking values for best models per category; OvO=one-vs-one, SVM= Linear Support Vector Machine, LR= Logistic Regression, , KNN= k-NN classifier, CNN=Convolutional Neural Network, Support= Absolut number of annotated documentations (before up-sampling)
Since the Metzler taxonomy is only available in german, we provide an english translation of the mapping for the categories that are not mentioned in the text or in Table 1 in Appendix A.1.

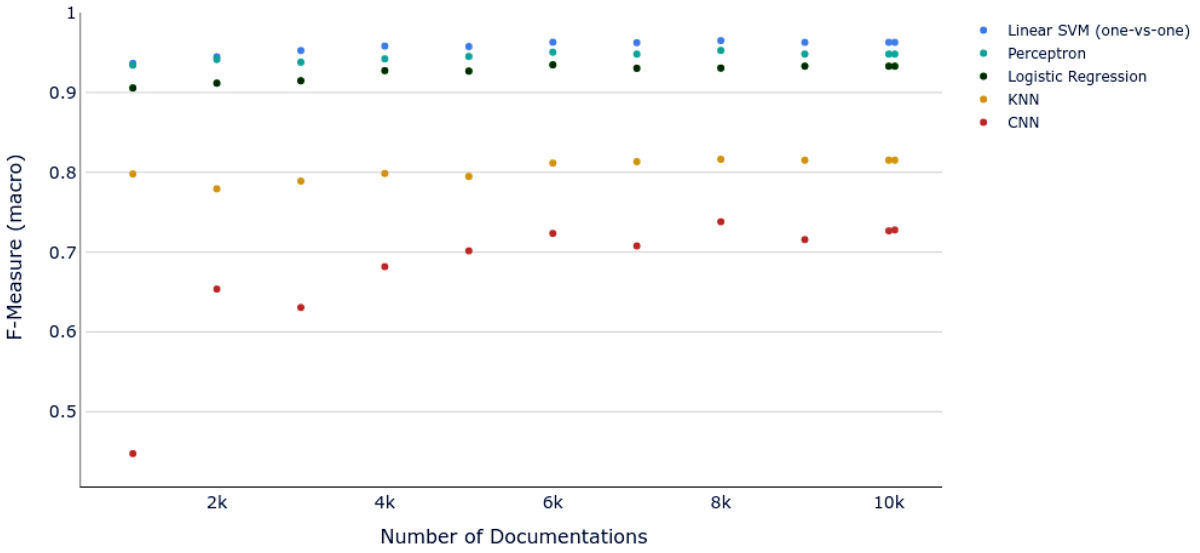

Figure 3: Impact of number of training samples in steps of 1,000 examples on the classification.

Feature combination	**SVM OvO**	**Perceptron**	**LR**	**KNN**	**CNN**
BOW	**0.89**	**0.89**	0.86	0.87*	-
ALL	0.89	**0.93***	0.91*	0.84	
-lemma	0.89	**0.92**	0.91*	0.85	-
-stem	**0.91**	**0.91**	0.89	0.84	-
-3grams	**0.93***	**0.93***	0.89	0.85	-
-POS	**0.89**	**0.89**	0.87	0.87*	0.75*

Table 6: Impact of leaving out different features on the classification performance, * marking values for best features per algorithm (3-grams = ngrams with n=3, POS= Part-Of-Speech tags, stem = stemmed words, lemma = lemmata)

with only 4 samples in the category-specific setting and the other categories occur so rarely that it is not possible to build an accurate category-specific classifier for them.

For category *M-23 ('time orientation')*, the category-agnostic setting has a rather low F-Measure in comparison to the other categories. One reason for this might be the fact that there are no examples for independence level *3 ('partly able to accomplish a task with help')* in *M-23*. Other categories where at least one label is missing are: *M-25 ('orientation in unfamiliar surroundings')*, *M-27* and *M-28*. Another reason for the better performance of the category-agnostic setting for some categories could be that these categories are more related to other sub-categories under the same top-level category, e.g. *M-2 ('Preparation of Snacks between main meals')* and *M-3 ('Preparation of main meals')* or categories *M-14* to *M-16* in the area *Relationships*. Categories with under 100 examples such as *M-27* and *M-28* on psychological disorders in addition to *M-2* seem to profit from their related categories in the category-agnostic setting and compensate for missing labels as well.

We experimentally investigate the impact of leaving out different features. In Table 6, we report the impact of leaving out different feature types from the full feature combination in comparison to the baseline models build with the unigram features weighted with tf.idf. For the SVM (one vs. one) a model with left out trigrams worked best. Only models with left out trigrams and left out stemmed words outperform the baseline. For Perceptron, the POS features have a great impact, because all models using POS outperform the baseline. The combinations of all features and leaving out lemmata worked best. The models build with Logistic regression outperform the baseline with every feature combination used on top of the unigrams. Here, the combination of all features except the trigrams worked best. For KNN no feature combination on top of the unigrams outperforms the baseline. Surprisingly, the worst F-Measures results from models using all features except the stemmed words. We also report the F-Measure for the CNN using stemmed words in combination

with lemmata and inherent trigrams. The baseline for the CNN is 6%; using stemmed words and lemmata increases F-Measure by 69% compared with using the word embeddings trained on raw text.

Finally, we investigate the impact of increasing the training data size on the performance of the classifiers. Figure 3 shows how F-Measures are affected by adding training data in steps of 1,000 examples. The diagram shows that the performance convergences from about 6,000 examples onwards, yielding no significant improvement from there. We conclude thus that our result can not be improved much more by simply increasing the number of annotated samples.

6 Conclusion

We have presented an approach to automatically classify observations in written reports by social workers into discrete independence levels reflecting the level of independence of their clients. We have described the construction of a corpus with substantial agreement on the task, comprising of over 15,000 documents. We have presented results on the task with different classifiers, showing that we can obtain F-measures of 95% macro-averaged over the different Metzler categories that we considered. We have also shown that a category-agnostic model outperforms a category-specific approach with one model per category on average over all categories. Furthermore, for some categories featuring smaller amounts of examples, the category-agnostic classifier performs better. The CNN model did achieve good results on the task but did not outperform SVM with one vs. one. While we did not investigate how to integrate more complex syntactic features other than POS information, this is an obvious avenue for future work. A further interesting question is which other psychological attributes can be predicted on the basis of the documentations considered here. An obvious category to explore is the mood of clients. In addition to classifying independence levels, we intend to perform further studies on the annotated data: One task is to present the trajectories and derived information for independence level per Metzler category to social workers to understand if it provides useful insights for them. Furthermore, it would be interesting to investigate in how far the observations in written reports by social workers reflect the actual condition or opinion of a client.

Acknowledgments

This research is part of the project MAEWIN and was supported by the Digital Society research program funded by the Ministry of Culture and Science of the German State of North Rhine-Westphalia. More information on the project can be found at https://www.uni-bielefeld.de/fakultaeten/technische-fakultaet/arbeitsgruppen/semantic-computing/projekte/maewin/
Many thanks to Frank Grimm for the development and technical support of the annotation tool.

References

Maria Odete Pereira Hidaldo de Araújo and Maria Filomena Ceolim. 2007. Assessment of the level of independence of elderly residents in long-term care institutions. *Revista da Escola de Enfermagem da USP*, 41(3):378–385.

Diana Baumrind. 1967. Child care practices anteceding three patterns of preschool behavior. *Genetic psychology monographs*.

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.

Laura Ana Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 2104–2119. Association for Computational Linguistics.

Deniz Cevher, Sebastian Zepf, and Roman Klinger. 2019. Towards multimodal emotion recognition in german speech events in cars using transfer learning. In *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019, Erlangen, Germany, October 9-11, 2019*.

François Chollet et al. 2015. Keras. https://keras.io.

Isaac R Galatzer-Levy, Kelly V Ruggles, and Zhe Chen. 2018. Data science in the research domain criteria era: relevance of machine learning to the study of stress pathology, recovery, and resilience. *Chronic Stress*, 2:2470547017747553.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Mehmet Gürhan Karakaya, Sevil Çuvalci Bilgin, Gamze Ekici, Nezire Köse, and Ayşe Saadet Otman. 2009. Functional mobility, depressive symptoms, level of independence, and quality of life of the elderly living at home and in the nursing home. *Journal of the American Medical Directors Association*, 10(9):662–666.

Gemeinsame Kommission. 2016. Niedersächsische anwendungshinweise zum hmb-w verfahren. verfahren der zuordnung von leistungsberechtigten zu gruppen für leistungsberechtigte mit vergleichbarem hilfebedarf (anlage 4 ffv lrv gem. §79 abs. 1 sgb 12). beschluss gk ffv lrv 8. sitzung.

Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15):5802–5805.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Huijie Lin, Jia Jia, Quan Guo, Yuanyuan Xue, Jie Huang, Lianhong Cai, and Ling Feng. 2014. Psychological stress detection from cross-media microblog data using deep sparse neural network. In *2014 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Heidrun Metzler. 2001. Hinweise zum verständnis des fragebogens zum „hilfebedarf". *HMB-W./Version*, 5.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*, pages 13–23. Association for Computational Linguistics.

Markus A Wirtz and Franz Caspar. 2002. *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Hogrefe.

Take Yo and Kazutoshi Sasahara. 2017. Inference of personal attributes from tweets using machine learning. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 3168–3174. IEEE.

A Appendices

A.1 English translation for the mapping of Metzler categories in Table 5 that are not mentioned in text or in Table 1

M-1: shopping
M-4: laundry care
M-5: keeping own area tidy
M-9: body care
M-10: toilet use / personal hygiene
M-12: bathing / showering
M-13: put on / take off clothes
M-14: relationships in the immediate vicinity
M-16: relationships with friends / partners
M-18: participation in offers / events
M-20: exploring areas of life outside the home
M-26: Coping with fear / anxiety / tension
M-29: dealing with / reducing significantly self-endangering and extraneous behaviour
M-30: carrying out medical or therapeutic prescriptions
M-31: Arrangement and implementation of medical appointments
M-32: special nursing requirements
M-33: observation and monitoring of the state of health, M-34: health promoting lifestyle