Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2226–2235
Marseille, 20-25 June 2022

GLoHBCD: A Naturalistic German Dataset for Language of Health Behaviour Change on Online Support Forums

Selina Meyer, David Elsweiler
Chair for Information Science, University of Regensburg
Universitätsstr. 31, 93051 Regensburg
{selina.meyer, david.elsweiler}@ur.de

Abstract

Health behaviour change is a difficult and prolonged process that requires sustained motivation and determination. Conversational agents have shown promise in supporting the change process in the past. One therapy approach that facilitates change and has been used as a framework for conversational agents is motivational interviewing. However, existing implementations of this therapy approach lack the deep understanding of user utterances that is essential to the spirit of motivational interviewing. To address this lack of understanding, we introduce the GLoHBCD, a German dataset of naturalistic language around health behaviour change. Data was sourced from a popular German weight loss forum and annotated using theoretically grounded motivational interviewing categories. We describe the process of dataset construction and present evaluation results. Initial experiments suggest a potential for broad applicability of the data and the resulting classifiers across different behaviour change domains. We make code to replicate the dataset and experiments available on Github.

Keywords: Conversational Agents, Motivational Interviewing, Behaviour Change, Language Resources

1. Introduction

Over the second half of the past century, illnesses resulting from poor health decisions, such as smoking, alcoholism and behaviours leading to obesity have emerged as a leading cause of death (Keeney, 2008; Johnson, N. B., Hayes, L. D., Brown, K., Hoo, E. C., Ethier, K. A., & Centers for Disease Control and Prevention (CDC), 2014). Despite their benefits and oftentimes necessity, health behaviour changes are difficult to put into practice and sustain (Kelly and Barker, 2016). A therapy approach developed to facilitate behaviour change is motivational interviewing (MI) (Miller and Rollnick, 2002). The automated delivery of MI by conversational agents (CA) has shown promise in the past (da Silva et al., 2018; Friederichs et al., 2015; Olafsson et al., 2019) and potentially offers multiple benefits such as constant availability and higher cost effectiveness when compared to a qualified counsellor. This would result in a lower entry barrier for first contact (Lisetti et al., 2015). Past studies have, however, largely disregarded the motivational state and utterances of the user, as the CAs were usually based on a rigid action framework and often limited user interaction to multiple choice entries with only few free text inputs.

This is problematic since an important part of MI is tailoring the conversation to the client's needs (Miller and Rollnick, 2002; Hall et al., 2012). MI is a very client-centered therapy approach and revolves around making a person aware of their own personal reasons for behaviour change by using open questions, reflections and affirmations. One of the central goals of MI is to elicit change talk, meaning language in favour of a behaviour change and limit sustain talk, language opposing behaviour change. This is intended to increase self-efficacy and readiness to change (Miller and Rollnick, 2002). It has been shown that there is a strong connection between a client's language in MI sessions and their ability to change their behaviour (Moyers et al., 2007). Depending on the client's voiced attitude towards change, different therapist reactions are most beneficial to support behaviour change (Beckwith and Beckwith, 2020; Clifford and Curtis, 2016).

To our knowledge, this complex interplay between user utterances and therapist behaviour has not yet been successfully implemented in CAs. One reason for this lack of personalisation might be the scarcity of publicly available datasets of MI counselling sessions or more general conversations around behaviour change, that specifically include fine-grained annotations of utterances by the person seeking change, brought about by the low availability of natural language psychotherapy corpora (Pérez-Rosas et al., 2018). This scarcity is especially apparent in the context of written language employed by chatbots, since MI is traditionally administered in a face-to-face setting (Miller and Rollnick, 2002). Resources are particularly scarce for non-English languages, and in the German language annotated MI data is, to our knowledge, completely unavailable.

The *Motivational Interviewing Skill Code (MISC)* (Miller et al., 2008) defines categories for client speech in MI conversations. Following this manual, generally speaking, a person's utterance around change can be assigned one of three valences: + for change talk, - for sustain talk and FN (Follow/Neutral) for utterances not related to the target behaviour. In addition to these valences, the *MISC* defines a number of content cat-

egories: *Reason, Need, Ability, Desire, Commitment, Taking Steps, and Other*, where *Need, Ability and Desire* are subcategories of *Reason* (see Table 1 for examples). These categories allow for the interpretation of the user's attitude towards behaviour change.

Here, we provide a novel language resource by manually annotating a sample of posts sourced from a popular German weight loss forum using the codes defined in the *MISC*. We chose the context of weight loss, since weight loss usually requires multiple health behaviour changes, such as increasing physical activity, changing nutrition habits, and other behavioural changes related to impulse control and coping mechanisms and success is often reliant on motivational factors and self-efficacy, both important factors of MI (Hauner et al., 2014; Elfhag and Rössner, 2005). We present the **German Language of Health Behaviour Change Dataset (GLoHBCD)**, a corpus of naturalistic written language around health behaviour change in an online context.

To our knowledge, the presented dataset is the first public resource of its kind in two aspects:

1. the application of MI client codes to non-counsellor mediated written online conversations around health behaviour change

2. the annotation of user utterances taking into account fine-grained client codes defined in the *MISC* in addition to valences

We find that fine-grained MI client codes defined in the *MISC* tend to naturally appear in unmediated conversations around weight loss and postulate that this dataset can offer useful information on language and self-disclosure online around health behaviour change. As initial evaluations and classification experiments yielded satisfactory results, we also expect the data to be useful as a language resource for motivational conversational agents, as well as the classification of change utterances in other health-related domains such as smoking cessation.

2. Related Work

Information technology systems and CAs have been used in various health and behaviour change settings (Milne-Ives et al., 2020; Pereira and Díaz, 2019; Oh et al., 2021; Brixey et al., 2017; Lee et al., 2017; Bharti et al., 2020), often with a focus on MI (Luo et al., 2021). In traditional MI, the therapist holds a number of instruments to nudge the client towards change, namely open questions, affirmations, reflections and summaries (Miller and Rollnick, 2002). For each of the instruments it is important to be as specific to the client and their situation as possible. Employing these instruments is essential to convey the spirit of MI and achieve treatment outcomes (Hall et al., 2012). Implementing this in a chatbot scenario will likely be a major challenge, since the CA would have to be constructed in a way that allows users the freedom to express themselves freely, while giving them the feeling of being truly understood, such that they are encouraged to reflect further on their behaviour. While empathetic response generation has shown progress in more general settings (Welivita et al., 2021; Shen et al., 2021; Majumder et al., 2020), the MI context requires all this to be achieved while maintaining control of the CA's replies such that MI concepts are applied correctly. Most existing implementations, therefore, mainly rely on open questions (Kocielnik et al., 2018) or heavily restrict user input (Bickmore et al., 2011; Nurmi et al., 2020; Gardiner et al., 2017). However, self-reflection is a big part of MI and being able to interpret a user's state of mind and reacting to nuanced utterances about motivation to change is indispensable when trying to enact the spirit of MI (Miller and Rollnick, 2002; Clifford and Curtis, 2016). In order to build a holistic MI chatbot, it is therefore necessary to be able to interpret a user's utterances with regards to their ability, commitment and reasons for behaviour change, as well as their general attitude towards change. Below, we report existing research and language corpora constructed with the goal to learn about language around behaviour change in the context of MI.

Almusharraf et al. (2020) designed a chatbot with the goal of learning through iterative interactions in the context of smoking. Participants were asked by the chatbot what they liked and disliked about smoking. The chatbot then attempted to allocate the participant's utterance to a category and asked the participant to correct if needed. They identified 21 unique reasons for or against smoking in 121 participant conversations. While knowing possible reasons for a health behaviour change is important, this approach only partially covers relevant user utterances in a MI setting and does not convey any information about more complex motivational factors that might appear when exploring possible health behaviour changes in-depth, such as a person's commitment, desires or confidence regarding change.

Pérez-Rosas et al. (2018) created a dataset of MI counselling sessions based on youtube and vimeo videos with the goal of identifying markers of high and low quality counselling. In their annotations they focused solely on the counsellor's behaviour, mainly the employment of open questions and reflections. Guntakandla and Nielsen (2018) take a similar approach, using Wizard of Oz conversations as their database and annotating different kinds of reflections. In a comparison of different natural language processing methods for the automated coding of MI, Tanana et al. (2016) create annotations for 341 psychotherapy sessions. They distinguish between 11 therapist codes and three client codes (Change Talk, Sustain Talk and Follow/Neutral, where Follow/Neutral are utterances not related to the change). Furthermore, Tavabi et al. (2020) present a method for the automatic coding of client behaviour in MI, using two clinical datasets

around alcohol consumption that also distinguish between the same three client codes as Tanana et al. (2016).

Hasan et al. (2019) annotated 37 transcripts of motivational interviews on the topic of weight loss with the purpose of using automated pattern analysis to detect effective communication sequences in MI. They differentiated between multiple counsellor codes and five user codes: change talk, sustain talk, high uptake weight, high uptake other, low uptake, where the uptake codes are focused on the development or progression of the conversation by the client rather than the content of the utterance.

We were unable to find datasets or studies that account for the fine-grained categories defined in the *MISC*. However, these categories could offer valuable information on a user's reasons, readiness and commitment to change. They could also be useful to understand which steps behaviour changers tend to take first. Understanding such information could be a vital component of a context and user aware motivational CA for behaviour change.

Thus, we attempt to fill this gap by annotating natural language data from a German weight loss forum. To the best of our knowledge, neither annotated nor un-annotated publicly available MI corpora exist for the German language. In contrast to actual MI session transcripts, the chosen medium for sourcing our data is readily available and reflects an example of naturalistic written online conversation. As a result, it might come closer to the input we might expect from chatbot users than transcripts from spoken MI-counselling session. Though unmediated by a professional counsellor, the resulting data can offer us valuable information about self-disclosure online in the context of behaviour change and could potentially be leveraged to model user utterances of a behaviour change CA.

3. Methods

We screened two thematically suited subforums of Germany's largest weight loss forum adipositas24.de[1] for the presence of change talk and sustain talk. At the time of data selection (August 2020) the two forums consisted of 7210 posts, written between May 2006 and July 2020. Most of the posts did not contain any instances of change and sustain talk but were focused more on obtaining factual information, offering emotional support to others, or relaying past experiences, for instance with specific weight loss programmes or clinics and psychological therapy. After screening, 1203 posts were identified and obtained for more in-depth analysis. We split these posts into sentences based on punctuation marks, after manually marking inappropriate punctuation (e.g. after abbreviations). Data cleaning was largely conducted by identifying suitable regex-statements since different users often made use of different abbreviations for the same word. Overall, the cleaned and split dataset consisted of 15,533 sentences.

One annotator annotated each of the resulting sentences with content categories and valences based on the *MISC* (Miller et al., 2008), with minor adaptations to account for the use case of unmediated online conversation as opposed to therapist-mediated oral conversations. For a detailed description of our annotation scheme and an overview of the code distribution see Table 1. Allusions to past behaviour changes were annotated as FN, since we wanted the resulting dataset to reflect language around *ongoing or planned* behaviour change only. As annotations were done sentence by sentence rather than on a semantic basis, we encountered some sentences that contained multiple categories or valences. These instances, however, only made up 2.1% of the complete dataset. More than two thirds of the data were coded as (*FN*) and 64.8% of the remaining sentences were annotated as *R* or one of its sublabels. The large share of *FN* and R in the dataset is in line with results found in the literature (Lord et al., 2015), where more than 80% of client utterances in annotated MI sessions were annotated as *FN*, with *R* and *O* appearing most frequently among the remaining utterances.

For copyright and privacy reasons, we do not publish the dataset itself. Instead, we provide code to obtain and process data to replicate the **GLoHBCD** and the results reported in section 4.4[2]. We provide the following information for each sentence in the dataset as a csv-file: the thread-id and post-id as defined in the forum's html-code, a sentence-id for each sentence in a post resulting from the provided script to split the sentences, and our annotations given in the three columns label (*FN, O, R, TS, C*, or a combination of them), sublabel (optional; *a, d, n*, or a combination of them) and valence (+ = *change talk,* - = *sustain talk*, or a combination of them).

Due to the low share of sentences with multiple labels and valences in the overall dataset and the large variability of code combinations we encountered in these few sentences, we did not include these 321 sentences in our evaluation. We also excluded the label *O* in our experiments, since the category was perceived as highly subjective and Miller et al. (2008) suggest that utterances labeled as *O* should always be discussed in a group setting. Lastly, we do not include *FN*-sentences in our tests for inter-rater reliability and analysis of the data. The *FN*-statements we encountered in the forum are not expected to be comparable to *FN* utterances in a counselling session, as forum users often provided fac-

[1] https://www.adipositas24.de/community/index.php?board/267-allgemeines
https://www.adipositas24.de/community/index.php?board/197-psychologische-therapie

[2] https://github.com/SelinaMeyer/GLoHBCD

Category		General Description	Keywords	Valence - Description	Example	N	Share (%)
Reason (R)		rationale, basis, incentive, justification or motive		+ - Reasons for behaviour change - Concern about current situation/behaviour - Possible advantages of changing behaviour - Disadvantages of the current situation	It's making me insecure to "present" myself in public (358829/25)	1339	13.9
				- - Reasons against behaviour change - Advantages of the current situation - Possible disadvantages of the behaviour change	My family loves me just the way I am (484073/8)	765	
	Desire (d)	Desire or will	"want", "desire", "like", "wish"	+ - Desire for change - Will to achieve a certain goal	I long for exercise, a life with less pain, more power, without scorn and ridicule from society and, above all, a feeling of being satiated, to banish food from the centre of my life (358829/49)	239	1.9
				- - Preferences for unhealthy things/behaviour - Desire to maintain behaviour	I'm craving chocolate and biscuits, even though I'm generally not a sweet eater (741454/26)	43	
	Ability (a)	Ability, degree of difficulty of the change	"can", "possible", "willpower" "ability", "hard"	+ - Confidence in own abilities to achieve the goal. - Statements that the behaviour change is easy	I've already made it this far on my own, then I'll be able to do the rest on my own too....(611970/16)	134	3.3
				- - Statements about the difficulty of changing behaviour. - Pessimism about own ability to change	Well, if it was that easy for me to change my diet, I would have done it of course (698138/1)	349	
	Need (n)	Need or necessity	"need", "must"	+ - Necessity of behaviour change - Necessity of a goal	But now I need to develop strategies with which I can overcome even these times (893200/13)	181	1.2
				- - Necessity of the negative behaviour	Knowing that I have food like that in my kitchen makes me crazy and I need to eat it (1895006/9)	10	
Commitment (C)		agreement, intention or obligation regarding future behaviour	"will", "promise", " intend"	+ - Planned steps towards behaviour change	Tonight I'm going to do 30 minutes on my cross trainer (674331/16)	433	2.9
				- - Planned steps away from change - - Intentionally unplanned steps towards the behaviour change	Tomorrow is my daughter's birthday, so I'll definitely be feasting on cake - I'll just treat myself (484969/7)	21	
Taking Steps (TS)		Specific steps that have been taken in the recent past	"have", "yesterday", "I did..."	+ - Specific steps taken towards behaviour change in the recent past	I also started doing a lot of sports (498588/5)	950	8.0
				- - Actions that have been taken and go against the behaviour change	since yesterday i've been eating like crazy again...(716757/5)	260	
Additional Categories used in original annotation (not included in inter-rater reliability and experiments)							
Other (O)		Utterances that reflect movement towards (+) or away (-) from behaviour change but do not fit into other categories (i.e. hypothetical statements, general statements of problem recognition)			Nevertheless, in my opinion it is just as you say, until now I was like an alcoholic who just didn't acknowledge his problem (+; 496082/10) therapy ok, but it probably won't do any good in our weight classes (-; 666268/3)	828	5.3
Follow/Neutral (FN)		Statements unrelated to behaviour change. Questions, reporting, non-commital statements, small talk			Did you even read until the end? (186373/25) When do you eat bread during the day? (484466/7)	9643	62.1
Multiple		Sentences containing multiple categories or valences were annotated with all codes that applied			I also want to switch to more tea right away, but I can't drink it without sugar and I don't like sweetener (C+/Ra-/Rd-) (1793127/6)	321	2.1

Table 1: Annotation scheme adapted from *MISC* with examples from the dataset (translated from German, referenced with Post-id/Sentence-id) sentence count and share of each category in the **GLoHBCD**

tual information and emotional support to other users, a behaviour that is not expected from clients in MI counselling. Future work could look at annotating such statements with MI-therapist codes, to learn about the different roles users embody in forum conversations. We also encountered a lot of chit-chat based on personal connections between forum users. As a result, the experiments we describe in the following section are all conducted on a subset of the **GLoHBCD** containing 4724 sentences, each annotated with exactly one label of *R, C, TS*, a maximum of one sublabel of *a, d, n* and one valence of +, -.

4. Evaluation

The sentences in the subset used for evaluation were written by 299 unique users. On average, the dataset contains 15.8 sentences per user with a standard deviation of 52.14. 75% of users account for fewer than 10 sentences in the dataset, while the three most active users accounted for 431 to 480 sentences each. 56.46% of the dataset are written by the 15 most active users.
We analyse our subset in different ways to ensure annotation consistency and establish the relevance of the data in the context of MI and behaviour change. In section 4.1 we report inter-rater reliability of the annotation categories. Next, in section 4.2 we examine how the assigned valence code relates to the sentiment of a sentence. In section 4.3 we report on a keyword analysis to identify words especially likely to occur in one category as compared to the others. Lastly, we apply machine learning techniques to evaluate whether it is possible to automatically predict the category and valence of an utterance in section 4.4.

4.1. Inter-Rater Reliability

To ensure consistency of the data, a second rater annotated a subset of 146 sentences from the **GLoHBCD** subset using the annotation scheme in Table 1. Sample sizes for the categories were stratified by occurrence in the data, with the restriction that, for each category, at least 10 sentences should be annotated by the second annotator. We chose unweighted *Cohen's* κ as a measure for inter-rater reliability as suggested by Artstein and Poesio (2008) for categorical data prone to annotator bias rated by two annotators. Agreement between the two annotators was calculated separately for labels, sublabels and valences. Sentences that were annotated with the label *R* and no sublabel are denoted as *R_* on the sublabel level. According to Landis and Koch (1977) and their interpretation of κ values, agreement can be evaluated as substantial for most categories, with the exception of *TS* and *R_*, for which only moderate agreement was achieved (see Table 2). Following the stricter interpretation of the metric introduced by McHugh (2012) would lead to evaluation of agreement as weak for *TS* and *R_* and moderate for the remaining categories.
These results are in line with comparable research in this area, which illustrates the issue of subjective interpretation in MI (Pérez-Rosas et al., 2016; Tanana et al., 2016; Hershberger et al., 2021). While in the mentioned studies, it is mainly the therapist's behaviour that was subject to annotation, the typical range of *Cohen's* κ values reported in the literature on MI seems to be between 0.4 and 0.6 and in some categories drops below the cutoff point for moderate or weak agreement. A potential explanation for this might be that utterances at times represent different labels to varying degrees, causing differing interpretations by raters.

Level	κ	**N**
Valence	**0.755**	
+		107
-		39
Label	**0.58**	
R	0.621	91
TS	0.491	31
C	0.625	24
Sublabel	**0.654**	
R_	0.579	45
Ra	0.681	16
Rd	0.662	16
Rn	0.768	14

Table 2: *Cohen's* κ for two Raters and 146 samples.

4.2. Sentiment Analysis

We postulate that sentiment analysis and valence are not interchangeable, since utterances with negative sentiment (i.e. "I hate the way I look") could represent change talk, while utterances with positive sentiment (i.e. "I love eating chocolate") could be instances of sustain talk. To evaluate the relationship between valence and sentiment, we used a pretrained german bert model for sentiment analysis [3]. The model had reached F1 scores higher than 90% for most datasets used during training (Guhr et al., 2020). We used a sample of 1000 sentences and their assigned valences and classified these sentences without further model fine-tuning. A Chi2 test revealed significant differences between expected and observed distribution of data (χ (2, N = 1000) = 51.21, $p < 0.0001$), indicating a correlation between annotated valences and predicted sentiments.

	negative		**neutral**		**positive**	
	exp	obs	exp	obs	exp	obs
-	165.3	217	111.1	81	44.6	23
+	349.7	298	234.9	265	94.4	116

Table 3: Expected and observed sentence distributions of sentiment and valence

[3] https://huggingface.co/oliverguhr/german-sentiment-bert

Valence	Top 10 Keywords per Category	N
+	1: do (mache) 2: hope (hoffe) 3: now (jetzt) 4: will (werde) 5: like (möchte) 6: **kilos (kilos)** 7: **kg (kg)** 8: goal (ziel) 9: finally (endlich)	9
-	1: not (nicht) 2: hard (schwer) 3: problem (problem) 4: unfortunately (leider) 5: find (fällt) 6: is (ist) 7: nothing (nichts) 8: believe (glaube)	8
Label		
TS	1: have (habe) 2: **eaten (gegessen)** 3: **eat (esse)** 4: was (war) 5: yesterday (gestern) 6: make (mache) 7: started (angefangen) 8: changed (umgestellt) 9: have (hab) 10: day (tag)	30
C	1: will (werde) 2: try (versuchen) 3: tomorrow (morgen) 4: sometime (mal) 5: first (erstmal) 6: today (heute) 7: continue (weiter) 8: committed (vorgenommen) 9: go (gehe) 10: next (nächsten)	19
R	1: is (ist) 2: am (bin) 3: **kg (kg)** 4: are (sind) 5: fear (angst) 6: feeling (gefühl) 7: yourself (sich) 8: satisfied (zufrieden)	8
Sublabel		
R_	1: have (habe) 2: was (war) 3: am (bin)	3
Ra	1: can (kann) 2: hard (schwer) 3: manage (schaffe) 4: not (nicht) 5: manage (schaffen) 6: difficult (schwierig) 7: find (fällt) 8: it (es) 9: know (weiß) 10: doable (machbar)	10
Rd	1: want to (will) 2 would like (möchte) 3: hope (hoffe) 4: I (ich) 5: gladly (gern) 6: like (mag) 7: wish (wünsche) 8: **cake (kuchen)**	8
Rn	1: must (muss) 2: have to (müssen) 3: important (wichtig) 4: need (brauche) 5: take care (aufpassen) 6: change (ändern) 7: work (arbeiten) 8: do (tun), 9: find (finden)	9

Table 4: Top 10 significant log-odds keywords for each category ($p < 0.05$, Bonferroni-corrected). German original in brackets. N displays the overall number of significant keywords for the category.

However, the contingency tables show that the sentiment of an utterance does not allow for conclusions to be drawn about its valence, since a large portion of both change and sustain talk was classified as neutral sentiment (see Table 3).

To further test our assumption, we treated the valence annotations as ground truth to calculate performance metrics for the sentiment predictions. The resulting Macro F1 of 27% confirmed that valence has to be considered separately of sentiment analysis. We draw the conclusion that while a general trend towards positive sentiment in change talk and negative sentiment in sustain talk exists, sentiment analysis is not enough to interpret a person's attitude towards attempting or sustaining a change in behaviour.

4.3. Keyword Analysis

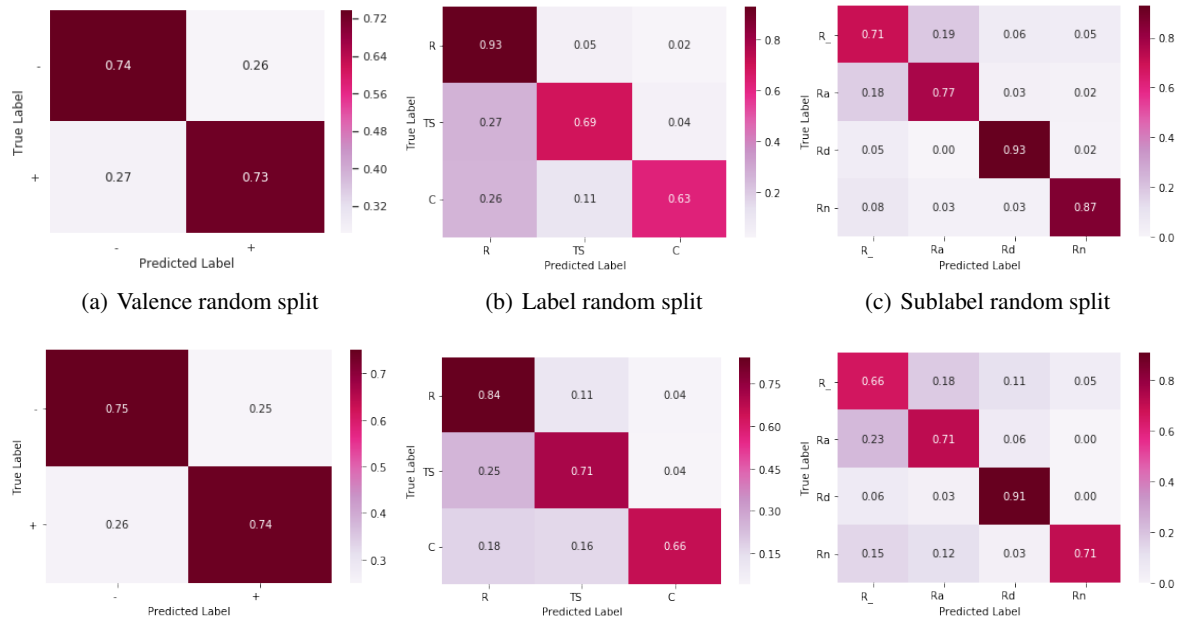
We were interested in whether the dataset could potentially be used to classify language for behaviour change topics beyond weight loss (i.e. smoking cessation). To this end we explored to what extent differences between annotation categories are specific to the topic of weight loss. We ran a log-likelihood analysis using the odds-ratio as effect size measure to identify keywords specific to the different labels, sublabels and valences. In Table 4 we display the top 10 keywords for each category compared to other categories on the same level. All keywords in the table are significant at bonferroni-adjusted $p < 0.05$.

We note that most of the keywords are function words or other words not specific to weight loss, with each category including a maximum of two weight or nutrition related keywords in the top 1 (indicated in bold in Table 4). This indicates that the classification of categories could potentially be applied to different topics, as long as the goal is behaviour change. We also find that *R_*, the sublabel with the lowest *Cohen's* κ score has only three significant keywords, all of them purely functional and conveying little discriminating information. This is not the case for *TS*, the other category with low inter-rater agreement, which actually yields the most keywords, including more nutrition specific words at the lower ranks (15: fish (fisch) 17: cheese (käse) 28: salad (salat) 30: vegetables (gemüse)). As such, both annotation categories with low κ scores bear some anomalies concerning the significant keywords when compared to the other categories. As a result we can expect, that a classifier trained on this data might generalise less well to other topics for the annotation categories *TS* and *R_* than for the other categories.

4.4. Machine Learning

To test how well the categories can be classified automatically, we used GermanBERT [4]. We created three separate classifiers, one for labels, sublabels and polarization, respectively. We randomly split the data into train and test sets using an 80/20 split and stratifying the data by class labels. As valence, labels and sublabels were highly unbalanced, we undersampled the largest class (*+/R/R_*) to the same size as the second largest class (*-/TS/Ra*) in the training set. The training

[4] https://huggingface.co/bert-base-german-cased



(a) Valence random split (b) Label random split (c) Sublabel random split

(d) Valence split by user activity level (e) Label split by user activity level (f) Sublabel split by user activity level

Figure 1: Confusion matrix of BERT classifications of an independent test set after fine-tuning.

set was then used for fine-tuning the BERT model using 10-fold cross validation across three epochs with a learning rate of 5e-05. The fine-tuned model was then used to predict labels of the test set.

In a second pass, to account for the fact that a large amount of data is provided by the most active users (the 65 most active users are responsible for 80% of the data), we split the data such that all sentences written by these 65 users represent the training set and all sentences written by the remaining 234 users make up the test set. This ensures that we can determine whether user specific utterances influence classification. Again, we undersampled the classes *R, R_* and + for the valence and label training set. We present the metrics for the different splits in Table 5.

	CV		**Test Set**		
	F1	**Std**	**Pre**	**Rec**	**F1**
Random Split					
Valence	73.97	2.63	70.42	73.31	70.87
Labels	74.16	3.22	79.64	74.87	**76.96**
Sublabels	79.49	2.69	66.20	81.89	**71.53**
Split by user activity level					
Valence	75.11	2.24	72.39	74.76	**72.86**
Labels	76.31	3.78	71.38	73.71	72.46
Sublabels	79.43	2.6	62.84	74.76	66.69

Table 5: 10-fold cross-validation and test set performances (%). We use Macro-F1 to evaluate performance.

The results indicate that classifications are reliable independently of user specific conversational styles. There were no significant differences between splits in cross-fold validation. Although test set results of the three classifiers varied across the different splits, this did not lead to significant differences for the mean performance of the three classifiers in cross-validation. The confusion matrices of the test set predictions across conditions show that true positive rate increased for both conditions for the valence classifier when splitting by user activity level. This was also the case for most classes in the label test set, with exception for the class *R*, which decreased by 9 percentage points compared to the random train-test split. Differences between conditions were most apparent in predictions for the sublabel test set, where each class performed worse in the split by user activity (see figure 1). Since the sublabels were only applicable to sentences with label R and we undersampled the largest class *R_*, this resulted in a very small training set of only 1151 samples in the random split and 1096 samples in the split by activity level, which might lead to less reliable results.

5. Limitations

There are a number of limitations to our study, which we reflect on here. By choosing a weight loss forum as our data source, our data might reflect some data bias towards people who are highly motivated to make a behaviour change or have already begun an attempt to change. As such, the dataset might not sufficiently represent utterances by people who struggle with their decision to change or are not aware of unhealthy behaviour.

Furthermore, as mentioned, a substantial share of the data we reviewed consisted of users offering emotional

support or giving information rather than talking about their own current behaviour change. As such, the forum users could be said to take on the role of client and therapist simultaneously. Thus annotating therapist's codes in addition to client codes might have yielded more information on how different utterances are elicited by others, and what information is disclosed without prompt. We also do not claim reliability for utterances containing multiple labels or valences, or utterances containing the label *O*. However, we are aware that in real conversation, user utterances might encompass more than one label or valence. While our full dataset does include such sentences, the tests conducted for inter-rater reliability as well as model training used only sentences that could distinctly be allocated to a label and valence. We thus primarily recommend the usage of the subset used in evaluation for further analysis rather than the full dataset.

6. Conclusion

We presented a novel dataset that applies fine-grained concepts from Motivational Interviewing to written conversational data around health behaviour changes in the specific context of weight loss. Exploration of the data reveals that automatic classification of the data is reliable and that discriminating features between the categories are specific to the topic of weight loss and the conversational style of users only to a small degree, which speaks in favour of a broader applicability of the dataset to different contexts.

In future work we will explore to what extent the models based on this data are applicable to other behaviour change topics such as smoking cessation. We also plan to collect data from simulated chatbot motivational interviews and test how our models can be used to predict user utterance categories. The dataset contains a lot of information that could be leveraged in further analyses. For instance, one could look at how language around behaviour change develops over time by exploring utterances by the most active users. This is facilitated by the fact that a large portion of the dataset is written by the most active users. Another angle would be to look at the grammar of the different categories and identify category-specific phrases.

To our knowledge, the presented dataset is the first to apply fine-grained client codes from the context of MI to written, unmediated conversation around health behaviour change, thus presenting novel insights in self-disclosure of behaviour changers on the internet, making it a valuable resource for motivational conversational agents to build upon MI-concepts.

7. Acknowledgements

We would like to thank the German Academic Scholarship Foundation for funding parts of this research project. We also thank Professor Udo Kruschwitz for his support in the writing of this paper.

8. Bibliographical References

Almusharraf, F., Rose, J., and Selby, P. (2020). Engaging Unmotivated Smokers to Move Toward Quitting: Design of Motivational Interviewing–Based Chatbot Through Iterative Interactions. *Journal of Medical Internet Research*, 22(11):e20251.

Artstein, R. and Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Beckwith, V. Z. and Beckwith, J. (2020). Motivational Interviewing: A Communication Tool to Promote Positive Behavior Change and Optimal Health Outcomes. *NASN School Nurse*, 35(6):344–351.

Bharti, U., Bajaj, D., Batra, H., Lalit, S., Lalit, S., and Gangwani, A. (2020). Medbot: Conversational Artificial Intelligence Powered Chatbot for Delivering Tele-Health after Covid-19. In *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, pages 870–875. IEEE.

Bickmore, T. W., Schulman, D., and Sidner, C. L. (2011). A Reusable Framework for Health Counseling Dialogue Systems Based on a Behavioral Medicine Ontology. *Journal of Biomedical Informatics*, 44(2):183–197.

Brixey, J., Hoegen, R., Lan, W., Rusow, J., Singla, K., Yin, X., Artstein, R., and Leuski, A. (2017). Shihbot: A Facebook Chatbot for Sexual Health Information on HIV/AIDS. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 370–373.

Clifford, D. and Curtis, L. (2016). *Motivational Interviewing in Nutrition and Fitness*. Guilford Publications.

da Silva, J. G. G., Kavanagh, D. J., Belpaeme, T., Taylor, L., Beeson, K., and Andrade, J. (2018). Experiences of a Motivational Interview Delivered by a Robot: Qualitative Study. *Journal of Medical Internet Research*, 20(5):e116.

Elfhag, K. and Rössner, S. (2005). Who Succeeds in Maintaining Weight Loss? A Conceptual Review of Factors Associated with Weight Loss Maintenance and Weight Regain. *Obesity Reviews*, 6(1):67–85.

Friederichs, S. A., Oenema, A., Bolman, C., Guyaux, J., Van Keulen, H. M., and Lechner, L. (2015). Motivational Interviewing in a Web-Based Physical Activity Intervention: Questions and Reflections. *Health Promotion International*, 30(3):803–815.

Gardiner, P. M., McCue, K. D., Negash, L. M., Cheng, T., White, L. F., Yinusa-Nyahkoon, L., Jack, B. W., and Bickmore, T. W. (2017). Engaging Women with an Embodied Conversational Agent to Deliver Mindfulness and Lifestyle Recommendations: A Feasibility Randomized Control Trial. *Patient Education and Counseling*, 100(9):1720–1729.

Guhr, O., Schumann, A.-K., Bahrmann, F., and Böhme, H. J. (2020). Training a Broad-Coverage German Sentiment Classification Model for Dialog Systems. In *Proceedings of the 12th Language Resources and*

Evaluation Conference (LREC 2020), pages 1627–1632.

Guntakandla, N. and Nielsen, R. (2018). Annotating Reflections for Health Behavior Change Therapy. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 4002–4007.

Hall, K., Gibbie, T., and Lubman, D. I. (2012). Motivational Interviewing Techniques: Facilitating Behaviour Change in the General Practice Setting. *Australian Family Physician*, 41(9):660–667.

Hasan, M., Carcone, A. I., Naar, S., Eggly, S., Alexander, G. L., Hartlieb, K. E. B., and Kotov, A. (2019). Identifying Effective Motivational Interviewing Communication Sequences Using Automated Pattern Analysis. *Journal of Healthcare Informatics Research*, 3(1):86–106.

Hauner, H., Moss, A., Berg, A., Bischoff, S., Colombo-Benkmann, M., Ellrott, T., Heintze, C., Kanthak, U., Kunze, D., Stefan, N., et al. (2014). Interdisziplinäre Leitlinie der Qualität S3 zur „Prävention und Therapie der Adipositas". *Adipositas-Ursachen, Folgeerkrankungen, Therapie*, 8(04):179–221.

Hershberger, P. J., Pei, Y., Bricker, D. A., Crawford, T. N., Shivakumar, A., Vasoya, M., Medaramitta, R., Rechtin, M., Bositty, A., and Wilson, J. F. (2021). Advancing Motivational Interviewing Training with Artificial Intelligence: ReadMI. *Advances in Medical Education and Practice*, 12:613.

Johnson, N. B., Hayes, L. D., Brown, K., Hoo, E. C., Ethier, K. A., & Centers for Disease Control and Prevention (CDC). (2014). CDC National Health Report: Leading Causes of Morbidity and Mortality and Associated Behavioral Risk and Protective Factors—United States, 2005–2013. *MMWR supplements*, 63(4):3–27.

Keeney, R. L. (2008). Personal Decisions are the Leading Cause of Death. *Operations Research*, 56(6):1335–1347.

Kelly, M. P. and Barker, M. (2016). Why is Changing Health-Related Behaviour so Difficult? *Public Health*, 136:109–116.

Kocielnik, R., Xiao, L., Avrahami, D., and Hsieh, G. (2018). Reflection Companion: a Conversational System for Engaging Users in Reflection on Physical Activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(2):1–26.

Landis, J. R. and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, pages 159–174.

Lee, D., Oh, K.-J., and Choi, H.-J. (2017). The Chatbot Feels You-a Counseling Service Using Emotional Response Generation. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 437–440. IEEE.

Lisetti, C., Amini, R., and Yasavur, U. (2015). Now all Together: Overview of Virtual Health Assistants Emulating Face-to-Face Health Interview Experience. *KI-Künstliche Intelligenz*, 29(2):161–172.

Lord, S. P., Can, D., Yi, M., Marin, R., Dunn, C. W., Imel, Z. E., Georgiou, P., Narayanan, S., Steyvers, M., and Atkins, D. C. (2015). Advancing Methods for Reliably Assessing Motivational Interviewing Fidelity Using the Motivational Interviewing Skills Code. *Journal of Substance Abuse Treatment*, 49:50–57.

Luo, T. C., Aguilera, A., Lyles, C. R., and Figueroa, C. A. (2021). Promoting Physical Activity through Conversational Agents: Mixed Methods Systematic Review. *Journal of Medical Internet Research*, 23(9):e25486.

Majumder, N., Hong, P., Peng, S., Lu, J., Ghosal, D., Gelbukh, A., Mihalcea, R., and Poria, S. (2020). MIME: MIMicking Emotions for Empathetic Response Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 8968–8979.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Miller, W. R. and Rollnick, S. (2002). *Motivational Interviewing, Second Edition: Preparing People for Change*. Applications of Motivational Interviewing Series. Guilford Publications.

Miller, W. R., Moyers, T. B., Ernst, D., and Amrhein, P. (2008). *Manual for the Motivational Interviewing Skill Code (MISC), Version 2.1*. Center on Alcoholism, Substance Abuse, and Addictions, The University of New Mexico.

Milne-Ives, M., Lam, C., De Cock, C., Van Velthoven, M. H., and Meinert, E. (2020). Mobile Apps for Health Behavior Change in Physical Activity, Diet, Drug and Alcohol Use, and Mental Health: Systematic Review. *JMIR mHealth and uHealth*, 8(3):e17046.

Moyers, T. B., Martin, T., Christopher, P. J., Houck, J. M., Tonigan, J. S., and Amrhein, P. C. (2007). Client Language as a Mediator of Motivational Interviewing Efficacy: Where is the Evidence? *Alcoholism: Clinical and Experimental Research*, 31:40s–47s.

Nurmi, J., Knittle, K., Ginchev, T., Khattak, F., Helf, C., Zwickl, P., Castellano-Tejedor, C., Lusilla-Palacios, P., Costa-Requena, J., Ravaja, N., et al. (2020). Engaging Users in the Behavior Change Process With Digitalized Motivational Interviewing and Gamification: Development and Feasibility Testing of the Precious App. *JMIR mHealth and uHealth*, 8(1):e12884.

Oh, Y. J., Zhang, J., Fang, M.-L., and Fukuoka, Y. (2021). A Systematic Review of Artificial Intelligence Chatbots for Promoting Physical Activity, Healthy Diet, and Weight Loss. *International Journal of Behavioral Nutrition and Physical Activity*, 18(1):1–25.

Olafsson, S., O'Leary, T., and Bickmore, T. (2019).

Coerced Change-Talk with Conversational Agents Promotes Confidence in Behavior Change. In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pages 31–40.

Pereira, J. and Díaz, Ó. (2019). Using Health Chatbots for Behavior Change: a Mapping Study. *Journal of Medical Systems*, 43(5):1–13.

Pérez-Rosas, V., Mihalcea, R., Resnicow, K., Singh, S., and An, L. (2016). Building a Motivational Interviewing Dataset. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 42–51.

Pérez-Rosas, V., Sun, X., Li, C., Wang, Y., Resnicow, K., and Mihalcea, R. (2018). Analyzing the Quality of Counseling Conversations: the Tell-Tale Signs of High-Quality Counseling. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3742–3748.

Shen, L., Zhang, J., Ou, J., Zhao, X., and Zhou, J. (2021). Constructing Emotional Consensus and Utilizing Unpaired Data for Empathetic Dialogue Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3124–3134.

Tanana, M., Hallgren, K. A., Imel, Z. E., Atkins, D. C., and Srikumar, V. (2016). A Comparison of Natural Language Processing Methods for Automated Coding of Motivational Interviewing. *Journal of Substance Abuse Treatment*, 65:43–50.

Tavabi, L., Stefanov, K., Zhang, L., Borsari, B., Woolley, J. D., Scherer, S., and Soleymani, M. (2020). Multimodal Automatic Coding of Client Behavior in Motivational Interviewing. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20)*, pages 406–413.

Welivita, A., Xie, Y., and Pu, P. (2021). A Large-Scale Dataset for Empathetic Response Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1264.