Towards Cross-Content Conversational Agents for Behaviour Change: Investigating Domain Independence and the Role of Lexical Features in Written Language Around Change

Selina Meyer
selina.meyer@ur.de
University of Regensburg
Regensburg, Bavaria, Germany

David Elsweiler
University of Regensburg
Regensburg, Bavaria, Germany
david.elsweiler@ur.de

ABSTRACT

Valuable insights into an individual's current thoughts and stance regarding behaviour change can be obtained by analysing the language they use, which can be conceptualized using Motivational Interviewing concepts. Training conversational agents (CAs) to detect and employ these concepts could help them provide more personalized and effective assistance. This study investigates the similarity of written language around behaviour change spanning diverse conversational and social contexts and change objectives. Drawing on previous research that applied MI concepts to texts about health behaviour change, we evaluate the performance of existing classifiers on six newly constructed datasets from diverse contexts. To gain insights in determining factors when identifying change language, we explore the impact of lexical features on classification. The results suggest that patterns of change language remain stable across contexts and domains, leading us to conclude that peer-to-peer online data may be sufficient to train CAs to understand user utterances related to behaviour change.

CCS CONCEPTS

• **Computing methodologies** → *Natural language processing*; *Discourse, dialogue and pragmatics*; • **Applied computing** → *Psychology*.

KEYWORDS

NLU, behaviour change language, transfer learning, conversational contexts

ACM Reference Format:
Selina Meyer and David Elsweiler. 2023. Towards Cross-Content Conversational Agents for Behaviour Change: Investigating Domain Independence and the Role of Lexical Features in Written Language Around Change. In *ACM conference on Conversational User Interfaces (CUI '23), July 19–21, 2023, Eindhoven, Netherlands.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3571884.3597136

1 INTRODUCTION

Motivational interviewing (MI) is a client-centred counselling approach that helps individuals resolve ambivalence about behaviour change [31]. It involves active listening, open-ended questions, and reflective responses to develop an individual's intrinsic motivation for change and help them develop a plan for taking action towards their goals [9, 31]. The Motivational Interviewing Skill Code (*MISC*) [30] allows language around behaviour change to be conceptualized by assigning valence and content labels to determine a person's current stance and thoughts about change.

Despite these concepts being inherently valuable to contextualize behaviour change and their potential to inform the choice of suitable actions for conversational agents (CAs), they remain largely absent in research around the application of conversational AI to behaviour change contexts. Applications tested in the context of MI have ranged from content-specific rule-based systems [18] to more open-domain generative agents [3, 45]. While some CAs model general emotions, or employ sentiment analysis, they have so far failed to account for the user's current situation and state of mind regarding change specifically [47]. Recently, the GLoHBCD, an annotated dataset containing natural language texts about behaviour change for weight loss, was released along with classifiers trained on the data and an analysis of its properties [27]. This dataset applies client utterance codes defined in the *MISC* to written and non-therapist-guided conversations, making it a potentially valuable resource in training CAs to be more attentive to the user and allow them to personalize content based on a user's current state of mind regarding their behaviour change.

While the *MISC* concepts used for annotation in the GLoHBCD can provide valuable information for CA's, the data itself is sourced from a peer-to-peer online forum, which might limit its usefulness for the purpose of training a CA. Although the GLoHBCD was annotated on a sentence-to-sentence rather than a post basis [27], which potentially increases its similarity to shorter messages as they would be likely to appear in conversation with a text-based CA, the language in the GLoHBCD might still significantly differ from such user utterances. Furthermore, the GLoHBCD focuses entirely on language around weight loss, whereas other behaviour change topics are disregarded. Nevertheless, analyses of the dataset hint that the language patterns learned by classifiers might not depend on the "target behaviour" (the behaviour the user wishes to change), since class-specific keywords were largely unrelated to weight loss specifically [27]. This indicates that classification may prove to be stable across different change contexts. Consequently, our research explores to what extent different conversational contexts and topics have an influence on the nature of written language around behaviour change by leveraging the GLoHBCD and the

CUI '23, July 19–21, 2023, Eindhoven, Netherlands

ACM ISBN 979-8-4007-0014-9/23/07.
https://doi.org/10.1145/3571884.3597136

Level	Code	Description	Example
Valence	**+**	Change Talk – Utterances in favour of behaviour change	I *need to stop* smoking
	–	Sustain Talk – Utterances in favour of status quo	I *don't want to* quit
Content Label	**R**	Reason – Reasons for/against change	I need to stop smoking
	TS	Taking Steps – Specific steps taken in the recent past	I *threw away* all my cigarettes
	C	Commitment – agreement, intention, or obligation for the near future	I'm *going to* throw away all of my cigarettes
Reason Sublabel	**a**	ability – ability and degree of difficulty of the change	I *can* quit
	n	need – necessity of change, or maintaining the status quo	I *need* to stop smoking
	d	desire – desire for change, or current behaviour	I don't *want to* quit
	general	general justifications, incentives, or justifications	Protecting my health is the most important thing to me

Table 1: Description of sentence codes based on [27, 30]. Each sentence is assigned a valence and content label. Sentences of category Reason are also assigned a reason sublabel. Markers that speak for the assignment of a specific code are *italicized*.

classifiers trained on it and by expanding on it with new, out-of-domain data.

This work is driven by the following research questions:

RQ1: Is the distinction between utterances belonging to the different behaviour change concepts defined in the *MISC* sufficiently stable across conversational contexts and change topics to allow context-independent classification?
RQ2: Which utterance types/*MISC* codes are hardest to predict across contexts?
RQ3: To what extent do different conversational contexts and change topics restrict classification transferability?
RQ4: Which lexical features are the most important for the classification of written utterances around behaviour change?

To answer these questions, we employ the same annotation strategy as Meyer and Elsweiler [27] to assemble six new datasets from different conversational, social and behaviour change contexts. Next, we evaluate how classifiers trained on the GLoHBCD perform on new datasets, taking into account the properties of the datasets that emphasize significant conceptual distinctions between the GLoHBCD and real-time chat conversations, as well as variations in target behaviour. Finally, since keyword analysis of the different GLoHBCD classes indicated that function words may play a bigger role for classification than content words [27], we explore the role of different lexical properties on classification performance across domains in an attempt to gain insight on how written change language is typically constructed.

Our results suggest that transfer learning between conversational domains to understand user utterances in the context of behaviour change is a feasible approach for the training of CAs. This means we can leverage domains with abundant training data to enable models to function in domains where there is a scarcity of labelled data available, of which there are many. Moreover, our findings evidence the plausibility of using *MISC* codes in such behavioural change settings which, if implemented, would enable CAs to react in a more context- and situation-appropriate way, as would be expected in a counselling session.

2 BACKGROUND AND RELATED WORK

Our research fits within the general tradition of systems for behaviour change [12, 13, 29, 37], with a focus on applying Motivational Interviewing concepts and techniques [9, 30, 31] to inform the design of a persuasive conversational agent. The *MISC* is traditionally used to evaluate MI sessions and includes behaviour codes both for the therapist and the client [30]. The client codes can be conceptualized along the DARN-CAT continuum (**D**esire, **A**bility, **R**eason, **N**eed, **C**ommitment, **A**ctivation, **T**aking Steps) [36, 38], a framework that describes client talk in ascending intensity of how strongly they speak for (or against) change.[1] For instance, Taking Steps-statements are a stronger marker of change than utterances about desire or ability. In addition to these codes, each client utterance is also assigned a valence, which defines whether the utterance speaks in favour of change (change talk) or against change (sustain talk). Utilizing these codes and their order of intensity in a persuasive CA can provide valuable insights into a user's readiness to change and help guide the CA in selecting appropriate counsellor turns. In this work, we focus on the identification of the codes in different conversational scenarios. For an overview of codes and what utterances they apply to, see Table 1.

2.1 Motivational Interviewing and Computational Methods

Existing attempts at automatically replicating processes of MI with the help of conversational agents are typically focused on specific target behaviours (e.g. smoking cessation) and do not account for user's thoughts about change [18, 32, 47]. One reason for this lack of focus on user-specific states of mind might be the fact that attempts at automating MI-annotation have so far mainly focused on therapist codes [32] or, if client codes were part of the project, merely distinguished between the valence of client utterances, rather than the content labels defined in the *MISC* [7, 42–44]. Moreover, (automatic) *MISC* annotation is mostly applied to spoken language, often as a means to evaluate MI sessions or to facilitate counsellor training [19, 23].

As such, existing resources with annotations of *MISC* client behaviour codes focus on applying valences to spoken data, such as transcripts from MI-Sessions [33, 46]. Such client utterances are likely not representative of the way users would interact with a CA and are missing annotations of *MISC* content-labels. Thus, they may be able to help identify whether a user is generally thinking positively or negatively about change, but cannot help in identifying either the user's reasons, perceived ability, willingness, or urgency

[1] Since the *MISC* does not include a code for activation utterances, we subsequently refer to the continuum as DARN-CT for the purpose of this paper.

to change, or what the user has done to achieve change in the recent past or is planning to do in the near future. Such concepts, however, are important for the successful application of persuasive technology [12, 13].

2.2 Differences Between Language Across (Online) Conversational Contexts

The GLoHBCD applies MI-annotation methods, which are typically applied to spoken data such as transcripts of MI sessions, to written, unmediated interactions between multiple parties on the internet. Since the GLoHBCD's publication, another group of researchers has applied *MISC* codes to written chat interactions, focusing on therapist behaviour codes [40]. They identified new codes, such as "inappropriate" or "chit-chat" that are unique to this type of online discourse when compared to in-person MI-sessions. They do not account for client codes in their analysis and focus on peer-counsellor chats between two people, which are likely to be more structured and mediated by the counsellor than peer-to-peer interactions.

Written online interactions are significantly different from spoken interactions between people in various ways, since written text is missing much of the information that is transmitted in spoken, or even face-to-face settings [14]. Giles et al. [14] explored an incomplete list of distinctions between spoken and online data, such as the potential for editing text before sending it in online conversations, the fact that online data is typically archived and readily available in retrospect, making it easier to refer back to, software-specific features, conventions, or restraints (e.g. character limits), the visibility of the interaction to others and the distinction between the number of interactants and audience members.

A lot of these distinctions can be applied not only to differentiate between spoken and online conversational data, but also to online conversational and other social media data from different contexts, which typically differs across two main dimensions: the synchronicity of the interactions, and the number of people involved in the conversation [1]. Baldwin et al. [2] compared different kinds of social media texts, namely YouTube comments, Twitter posts, forum posts, blogs, and Wikipedia articles based on their relative similarity, as well as lexical and grammatical features. They also compared these sources with the British National Corpus as an example of typical written texts. They found that conversation online, even when highly asynchronous, tends to be shorter and, by extension, less complex than other texts. Their results show that the average word length of forum data is comparable with that of YouTube comments and tweets and that forum posts are the most similar to all other explored corpora. This indicates that classifiers trained on the forum posts annotated in the GLoHBCD should be able to transfer to other online sources fairly well, especially when the new sources also constitute public conversations between multiple people.

We are not aware of any research exploring the differences between forum data and more synchronous chat interactions, or comparisons between forum data and conversations between humans and chatbots. There have, however, been explorations of differences between human-human and human-chatbot text conversations. Hill et al. [20] found that messages to conversational agents tended to be shorter and utilize a more restricted vocabulary. This could speak for the feasibility of applying the classifiers trained on the GLoHBCD to chatbot conversations, as it would imply that these conversations are of lower complexity than the training data. However, Rapp et al. [35] found in their meta-review that conversations with open-domain chatbots tend to include more grammatical and spelling errors, more profanity and more random keystrokes than other conversational data, a fact that could hinder classification transferability.

While the literature shows that there are significant differences between different modes of conversation even in written online contexts, it is not clear to what extent these differences are relevant to the classification of the behaviour change codes defined in the *MISC*. To explore this, we account for differences between conversational data identified in the literature in the construction of our datasets and add to existing research by including direct comparisons between human-chatbot conversations and other social media interactions.

2.3 Transfer Learning in Conversational AI

Transfer learning is a popular approach to dealing with low resource data in deep learning contexts and has found frequent application to text data, which is often prone to property inconsistencies between training and test data [4]. Past research has applied the technique to various health, wellbeing, and behavioural contexts, for instance to detect opinions and behavioural intentions to COVID-19 vaccines in tweets [26], and to learn from Twitter data in order to classify clinical patient messages [41].

In the context of conversational AI, Das et al. [11] applied transfer learning to improve text generation for psychotherapy. To achieve this, they fine-tuned existing LLMs with counselling based data from Reddit and psychotherapy training sources. Rajan et al. [34] used graph-based transfer learning to filter the most useful pieces of information in CA-based conversations to enable less time-intensive, more personalized and context-specific conversations. In the scope of the ConvAI2 competition at NeurIPS 2018, Golovanov et al. [15] trained a transformer-based generative model on the benchmark chat-datasets PersonaChat [48] and DailyDialog [24] before fine-tuning on the competition data. They also experimented with the inclusion of Reddit forum data, but found it led to a deterioration in automated performance metrics, which led to the exclusion of this data in the final model. However, they note the large discrepancy between automated metrics and human evaluation of generated dialogues, which could mean that Reddit data may have been more suited for training their model than the metrics lead to believe.

In our case, we use transfer learning to determine the feasibility of applying classifiers trained on forum conversations to CA-human conversational data in order to classify and thus extract information from user utterances that can later be used to steer CA behaviour in conversations. In the following section, we outline the main dimensions to compare conversational text data on as derived from the related work, and introduce the datasets we constructed based on these dimensions to evaluate the transferability of the GLoHBCD data to other conversational contexts.

Dataset	Domain/Target Behaviour	Context	synchronous	multi-party	human-human	# sentences
GLoHBCD	weight loss	Forum - Interaction between peers		x	x	4724
Smoking Cessation Forum	smoking cessation	Forum - Interaction between peers		x	x	683
Health Coaching Dialogue Corpus	step count increase	Text conversations with health coach			x	508
Optifast Mock-Chatbot	weight loss	Text-based conversation with simulated motivational chatbot	x			90
DARN-CT-based Wizard of Oz Dialogues	New Year's resolutions	Text-based conversations with simulated motivational chatbot	x			80
Synthetic GPT-3 Data	weight loss	user simulation through eliciting questions				74
Instagram data	weight loss	Instagram posts - Community interaction		x	x	918

Table 2: Overview of datasets constructed to evaluate the GLoHBCD

3 CONSTRUCTING DATASETS WITH VARYING BACKGROUNDS

To explore the stability of the way *MISC* codes are worded in different written conversational contexts, we first identify three main differences between the GLoHBCD data and user interactions with CAs, based on Androutsopoulos [1] and Hill et al. [20]:

- **Conversation Mode** - Being sourced from a peer-to-peer forum, the GLoHBCD constitutes multi-party conversations, whereas a chatbot conversation would only involve two parties. For the purpose of this project, we also take this as an indicator for the visibility of the interactions to third parties, where multi-party conversations are generally public to outside audiences and conversations between two parties are seen as private interactions.
- **Synchronicity** - While still being somewhat asynchronous compared to spoken interactions, conversations between bot and human happen in a faster, live-chat-like fashion, as opposed to forum posts which often have multiple days between messages.
- **Interaction Type** - This refers to the difference between human-human and human-bot conversations.

In addition to these differences, we also want to explore the stability of classification across different behaviour change topics, leading to a fourth dimension to investigate:

- **Domain** - the target behaviour in the conversation, meaning the specific change the user is intending to make or trying to achieve.

Based on these four dimensions, we collected and annotated data from six sources representing varying degrees of each dimension. In this section, we describe the collection process and give an overview and comparison of these datasets. Each dataset was split into sentences and annotated by the authors[2] using a script and annotation scheme supplied by Meyer and Elsweiler [27][3]. The annotation scheme is adapted from the *MISC* and applies labels across three levels to each sentence, which are described in Table 1.

3.1 Smoking Cessation Forum

This collection from a smoking cessation forum[4] is the most similar to the original dataset in terms of conversational style. We chose the subforums "nichtraucher-bald-bin-ich-soweit" (nonsmoker, I'll be there soon) and "mein-rauchfrei-tagebuch" (my smoke-free diary), both containing accounts of individuals at different stages in their behaviour change process. While the first subforum is mainly used by individuals who are preparing to quit smoking or are in the early stages of a quit attempt, the second subforum is frequented more by nicotine-dependent individuals who have quit smoking within the last few days or weeks. We randomly selected threads from both subforums and, following the approach outlined in Meyer and Elsweiler [27], screened the posts for presence of change and sustain talk. Posts that fit the criteria were then annotated on a sentence-to-sentence basis. We annotated posts, until the smallest code class (**n**) in the GLoHBCD had occurred over 20 times, representing >10% of the amount in the GLoHBCD. This amounted to 167 annotated posts (1675 sentences), distributed over 36 threads, written by 46 users. 102 posts stemmed from the subforum "nichtraucher-bald-bin-ich-soweit" (1024 sentences) and 65 posts from the subforum "mein-rauchfrei-tagebuch" (652 sentences). In line with Meyer and Elsweiler, we excluded sentence that were not related to behaviour change or sentences that represented a combination of multiple content labels [27]. The resulting dataset contains 683 sentences in our evaluation which are annotated with exactly one content label, a maximum of one reason sublabel and one valence label.

3.2 Health Coaching Dialogue Corpus

Gupta et al. [16] released a corpus of text messages between certified health coaches and patients in the context of goal setting. The corpus includes conversations with 26 patients over the course of several weeks, with conversations focusing around setting and evaluating daily step count goals. For target setting, the S.M.A.R.T.

[2]the first author has completed a fully certified Motivational Interviewing training programme

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

[4]https://www.endlich-nichtraucher-forum.de/

strategy was used, which focuses around setting achievable, actionable and timely goals. The original dataset is in English. We used the free DeepL API[5] to translate them into German before annotating participants' utterances. This dataset is different from the original forum data in multiple aspects: it reflects direct conversations between two humans, as opposed to the more indirect conversational style between multiple parties in a forum. It also focuses on a different, much more specific target, as increasing a step count does not necessarily have to relate to weight loss.

3.3 Optifast Mock-Chatbot

To collect this data, we developed a chatbot-style survey where we prompted users to respond to MI-style questions in natural language text. The aim of this survey was to gauge the users' responses to an unresponsive chatbot, in order to assess their potential interactions with chatbots in this context. This survey was geared towards participants of the Optifast-52 programme in Regensburg, a weight loss programme for people with clinical obesity. The survey was constructed using botpress [6]. The conversation consisted of three phases. First, the mock-chatbot introduced itself, and gave an overview of the scope and goal of the conversation, before asking for permission to continue. When the participant agreed to go on, this triggered the elicitation phase, in which the bot first asked the participant to explain what brought them here. Following this, participants were asked alternating questions aimed at eliciting either change talk (e.g. *What didn't you like about your lifestyle before you started Optifast? How has Optifast improved your life?*), or sustain talk (e.g. *You must have had good reasons why you didn't decide to do an Optifast course earlier. What were those reasons?*). In this phase, the bot had two versions, where the first version began with a change talk question and ended with a sustain talk question and the second version displayed a negative/positive counterpart for each of the first version's questions (e.g. *Suppose you (don't) maintain your current behavioural changes. What would your life look like in 5 years?*). Both versions contained questions explicitly asking for reasons to change, commitments for the near future and recently taken steps in favour or against change. They also included questions that called for participants to envision their future if they were to succeed or fail in their endeavour, and look back to what their life was like before starting the Optifast programme. In the concluding phase, participants were asked to estimate how likely it is they will reach their goals on a scale from one to ten and whether they feel that reflecting about their change in this way has helped them, before the chatbot thanked them for the conversation, wished them success for their further weight loss and said goodbye.

10 Participants who were enrolled in the Optifast-52 programme in Regensburg used the chatbot in the course of a related study. Due to timing reasons, all participants were in the final phase of the Optifast programme and had already achieved most of their weight-loss goals and were working to maintain their new weight. The resulting annotated dataset contains 90 relevant user utterances, the majority of which represent change talk. Although the Mock-Chatbot was not very reactive regarding specific user inputs, this dataset represents a more synchronous, question-answer style conversation between a system and a human. Participants were told that they were conversing with a simple chatbot before starting the conversation.

3.4 Synthetic GPT-3 Data

As a sanity check, we use data generated by Meyer et al. [28] prompting GPT-3 questions based on the DARN-CT framework described in section 2, meaning questions regarding desire, ability, reasons, need, commitment and recently taken steps to lose weight [28, 38]. We annotated 120 of the generated sentences. We expect that the GPT-3 outputs represent highly typical and clichéd answers to the prompted questions, and should thus be very easy to classify. This allows us to gain insights into the potential of using classifiers trained on the forum data on very simple live-chat-like utterances. Since the prompts used for data generation were more QA-like [28], the data more closely resembles conversations between a client and a coach than forum posts. Low classification results would mean that a classification transfer between these two conversational styles is not feasible, whereas high results would indicate stable patterns between language of change across forum posts and more conversational settings.

3.5 DARN-CT-based Wizard of Oz Dialogues

This data was collected as a Wizard of Oz study, which simulated a chatbot and asked 14 study participants about changes they want to make in their lives, also following the DARN-CT structure. The wizard asked predefined questions sampled from the questions used to generate GPT-3 data referred to in 3.4 and geared towards eliciting participant utterances for each code defined in the *MISC*. The study focused on inducing a live-chat conversational style. It was conducted shortly after new year's and participants were asked to talk about their new year's resolutions, which led to the inclusion of various different behaviour change intentions (e.g. "procrastinate less", "enjoy live more", "use the car less", "increase physical activity"). Compared to the Optifast Mock-Chatbot, it included more and different change domains and was not restricted to clinically obese users in treatment. It was also more interactive, as the wizard in some cases summarized what the participant had said in previous utterances, as a counsellor in a MI-session would [31].

3.6 Instagram Data

This dataset consists of data crawled from Instagram using different hashtags related to weight loss and constitutes a new conversational style of a single person interacting with a community in the form of short and precise messages. This puts it between forum posts and chat conversations. Parts of the dataset (220 sentences) were separately annotated by one of the authors and a second researcher (not an author) to ensure inter-annotator agreement, the remaining data were annotated by the second researcher alone, resulting in a total of 918 annotated sentences. The dataset does not include valence annotations. In line with Meyer and Elsweiler [27], inter-annotator agreement was calculated separately for content labels and reason sublabels. The resulting Kappa scores were more stable than for the GLoHBCD, ranging between $\kappa = 0.6$ and $\kappa = 0.79$.

[5] https://www.deepl.com/en/docs-api/
[6] https://botpress.com/docs/

			Smoking Cessation Forum	Health Coaching Dialogue Corpus	Optifast Mock-Chatbot	DARN-CT-based Wizard of Oz Dialogues	Synthetic GPT-3 Data	Instagram Data	GLoHBCD
R	**general**	+	30.3%	8.7%	50%	21.3%	13.5%	26.9%	28.3%
		-	15.7%	5.1%	12.2%	1.3%	21.6%		16.2%
	d	+	8.9%	6.7%	2.2%	12.5%	2.7%	7%	5.1%
		-	0.9%	0.6%	0%	0%	1.4%		0.9%
	a	+	2.8%	8.1%	1.1%	0%	8.1%	5.8%	2.8%
		-	3.7%	2%	4.4%	1.3%	9.5%		7.4%
	n	+	2.6%	2%	0%	8.8%	14.9%	3.3%	3.8%
		-	0.3%	0.4%	0%	0%	1.4%		0.2%
C		+	11%	34.1%	18.9%	41.3%	8.1%	16.1%	9.2%
		-	0.4%	1.8%	0%	0%	1.4%		0.4%
TS		+	18%	25.2%	7.8%	12.5%	10.8%	41%	20.1%
		-	5.4%	5.5%	3.3%	1.3%	6.8%		5.5%
		+	73.6%	84.6%	80%	96.3%	58.1%		69.3%
		-	26.4%	15.4%	20%	3.8%	41.9%		30.7%

Table 3: Label Distribution of out-of-domain and out-of-context corpora compared to GLoHBCD

3.7 Overall Comparison of Datasets

The different datasets introduced in this section complement the GLoHBCD in different ways by incorporating different domains, contexts, and modes of conversations. The differences and similarities in the datasets are summarized in Table 2. We define datasets as synchronous, if the conversation is started and completed in a single session, and closely resembles a live-chat scenario [25]. Per this definition, only the Optifast Mock-Chatbot and the Wizard of Oz Dialogues qualify as synchronous interactions, since all other data does not require or provide instantaneous replies. We base the distinction between human-human and human-bot off the user's perception [10]. Since the participants of the study collecting the DARN-CT-based Wizard of Oz Dialogues were not aware that the conversation was controlled by a person, these qualify as human-bot interactions.

Different contexts led to different distribution of labels (see Table 3). The questions asked in the Wizard of Oz Study were primarily focused on eliciting change talk, which led to only three sustain talk statements. Furthermore, for easy to achieve targets, as they were the goal of the study which led to the creation of the Health Coaching Dialogue Corpus, the share of ability and commitment change talk was much higher than in the GLoHBCD, and the Instagram data included a particularly high share of commitment and taking steps statements, which might reflect the focus of the social media platform, which is primarily photography-based, on sharing experiences and actions rather than thoughts. This indicates that the way people address behaviour change in different modes of written language depends not only on the context, but also the perceived difficulty of the attempted behaviour change. Still, the labels in the annotation scheme seem to adequately reflect and account for these differences.

4 METHODS

Using the newly constructed datasets and the GLoHBCD data, we employ different strategies to address our research questions. We begin by classifying the new datasets using classifiers fine-tuned on the GLoHBCD, before focusing on specific dataset properties and exploring the most important lexical features across contexts. In this section, we outline our approach to answering each of our research questions and report and interpret the main results.

4.1 RQ1: Feasibility of Context-Independent Classification

Meyer and Elsweiler [27] supply scripts to fine-tune three bert-base-german-cased [8] instances to classify the three label-levels (valence, content labels, and reason sublabels). In order to explore to what extent stable classification across different contexts is feasible, we adapt these scripts to predict on each of the datasets outlined in section 3, as well as a test-split of the GLoHBCD using 10-fold cross-validation and record the Macro F1 of predictions for each test set.

Since the resulting data does not meet assumptions of normality, we run separate Friedman tests for each classification level. We find significant differences between the Macro F1 results for the different datasets on each label-level (Valence: $\chi^2(5) = 37.83, p < 0.001$; Content Labels: $\chi^2(5) = 55.11, p < 0.001$; Reason Sublabels: $\chi^2(5) = 49.89, p < 0.001$) and use bonferroni-corrected pairwise Wilcoxon signed-rank tests as post-hoc test to identify for which datasets the classification performance was significantly higher or lower than for the GLoHBCD. The results indicate that the valence classifier leads to the most stable results across datasets, with significant differences at $p < 0.05$ only for the Smoking Cessation Forum and the Health Coaching Dialogue Corpus, which were classified significantly worse and the Synthetic GPT-3 data, which was classified significantly more reliably than the GLoHBCD. In comparison, the content label and reason sublabel classification led to more variation in classification results across datasets, with significant differences in Macro F1 across all datasets compared to the GLoHBCD for the content label classifier and all datasets except the Synthetic GPT-3 data and the DARN-CT Wizard of Oz data for the reason sublabel classifier.

It is worth noting that classification results for the out-of-context datasets were often significantly better than for the GLoHBCD, a trend that was most pronounced for the sublabel classifier. We also notice that the content labels and reason sublabels of the Optifast Mock-Chatbot data and Instagram data are the most difficult to classify, whereas the valence classification transfers least well to the Health Coaching Dialogue Corpus (for a visual comparison of Macro F1 scores across datasets, see Figure 1). This leads us to conclude that the valence of utterances seems to exhibit the most stable patterns across datasets, whereas the reason sublabels seem to be the most easily identifiable in out-of-context data.

4.2 RQ2: Differences in Prediction by *MISC* Codes

While the results outlined in section 4.1 tell us about the feasibility of classification across contexts and code-levels, they do not allow us to draw conclusions on whether specific *MISC* codes are easier to predict across contexts. To explore this, we use the resulting classifiers to predict on each test set and record the F1 score for each class separately. This allows us to make inferences about potential biases

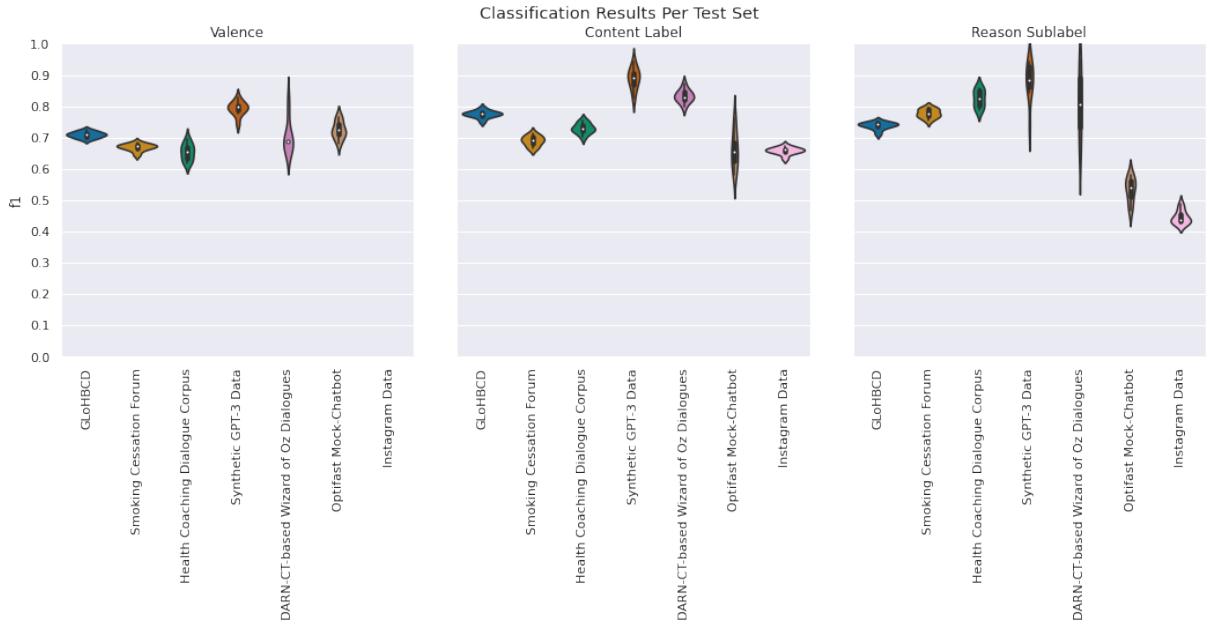

Figure 1: Classification performance of classifiers trained on GLoHBCD on test sets in 10-fold cross validation. At each fold, the classifiers were used to predict on the complete test sets

	Valence		Content Label			Reason Sublabel			
Test Set	**+**	**-**	**R**	**C**	**TS**	**general**	**d**	**a**	**n**
GLoHBCD	83.59	60.66	90.13	69.36	74.83	86.65	76.67	57.31	67.53
Smoking Cessation Forum	81.50	51.77	85.14	56.06	63.81	92.68	82.25	69.88	75.68
Health Coaching Dialogue Corpus	91.65	41.94	75.62	72.91	77.19	78.38	86.84	68.18	1
Optifast Mock-Chatbot	88.52	56.25	87.27	52.17	66.67	92.47	66.67	66.67	0
DARN-CT-based Wizard of Oz Dialogues	97.06	33.34	81.58	66.67	85.71	90.90	95.24	1	0.8
Synthetic GPT-3 Data	81.81	73.34	96.15	93.34	82.76	87.5	1	83.87	95.65
Instagram data	––	––	71.87	55.56	68.15	80.36	34.41	37.84	42.55

Table 4: Class-wise F1 scores (%) for datasets and code-levels. Classifiers were trained on GLoHBCD data and used to predict on the test sets

towards certain classes learned by the classifiers, or whether particular codes exhibit greater variation across different datasets, which could explain differences in classification performance among the datasets.

We show the resulting F1 score for each class-label and dataset in Table 4. We see that the majority classes of each code-level (+, R, general) are predicted more reliably than the other classes across datasets and that the distribution of F1 scores for the different classes is similarly distributed for the different datasets. The results also indicate that commitment utterances seem to be harder to identify in some datasets (Smoking Cessation Forum, Instagram Data, and Optifast Mock-Chatbot).

This could mean that the way commitment to take a certain action is voiced is more context-dependent than the wording of other labels. For instance, commitment statements on Instagram, where users tend to present themselves in a positive light [17, 22], might look different to commitment statements on self-help forums. Furthermore, the fact that the Optifast participants who used the Mock-Chatbot were towards the end of their programme might have led to different kinds of commitment statements than the ones voiced in the other datasets, where people were at the beginning or in the middle of making a change. We also find that the F1 scores across all reason sublabels were very low for the Instagram data, indicating that the positive and performative nature of the social media platform might have a bigger impact on the way people write about change than other conversational contexts.

4.3 RQ3: Investigating the Role of Different Dataset Properties

We wanted to investigate the effect of the distinguishing dimensions of Conversation Mode, Synchronicity, Interaction Type, and Domain on classification more closely. To explore this, we group the newly constructed datasets based on their properties as outlined in Table 2 and compare the classification performance between in-context and out-of-context groups. For each dimension, a group

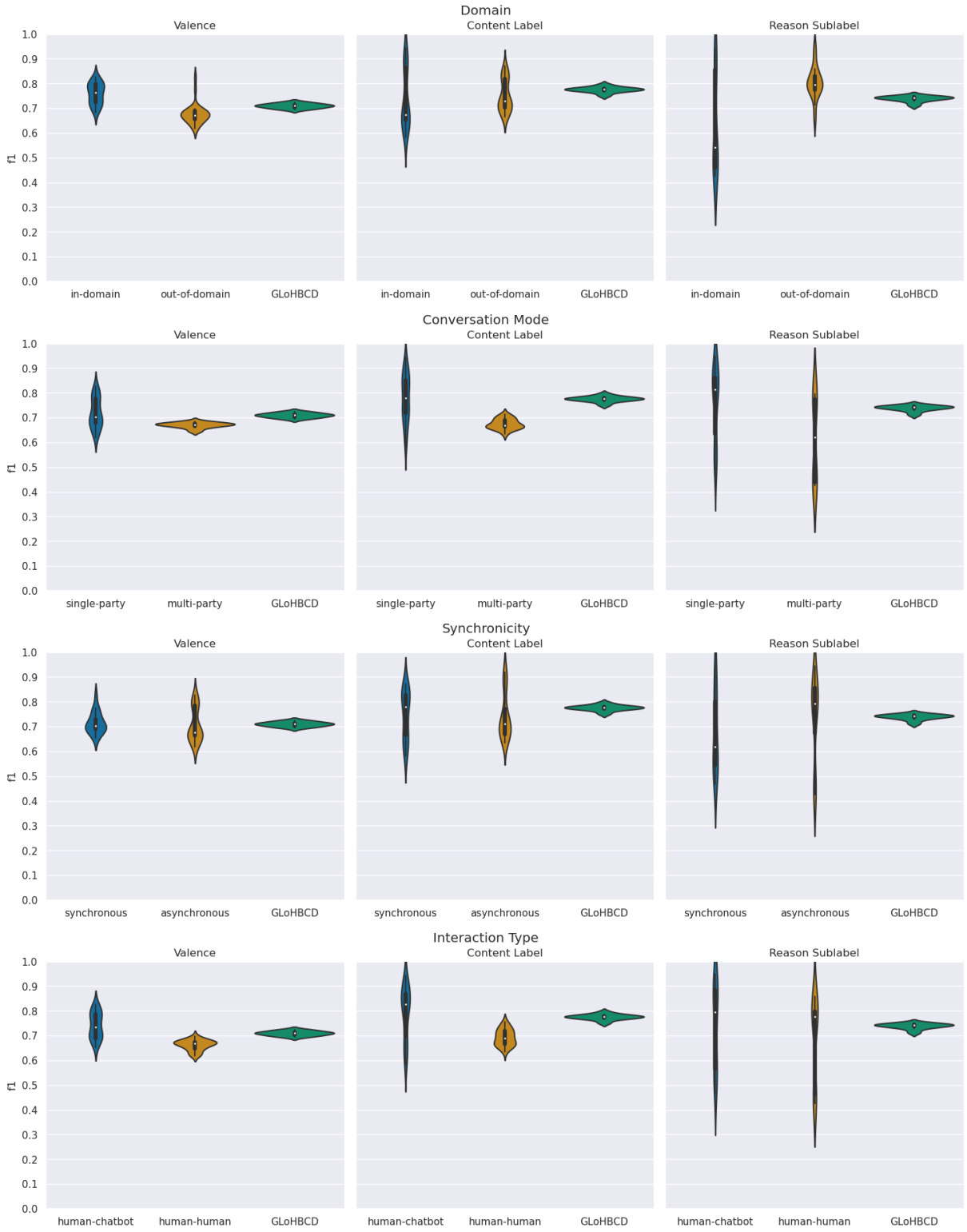

Figure 2: Classification results based on dataset properties (in-context versus out-of-context) compared to classification performance on GLoHBCD data

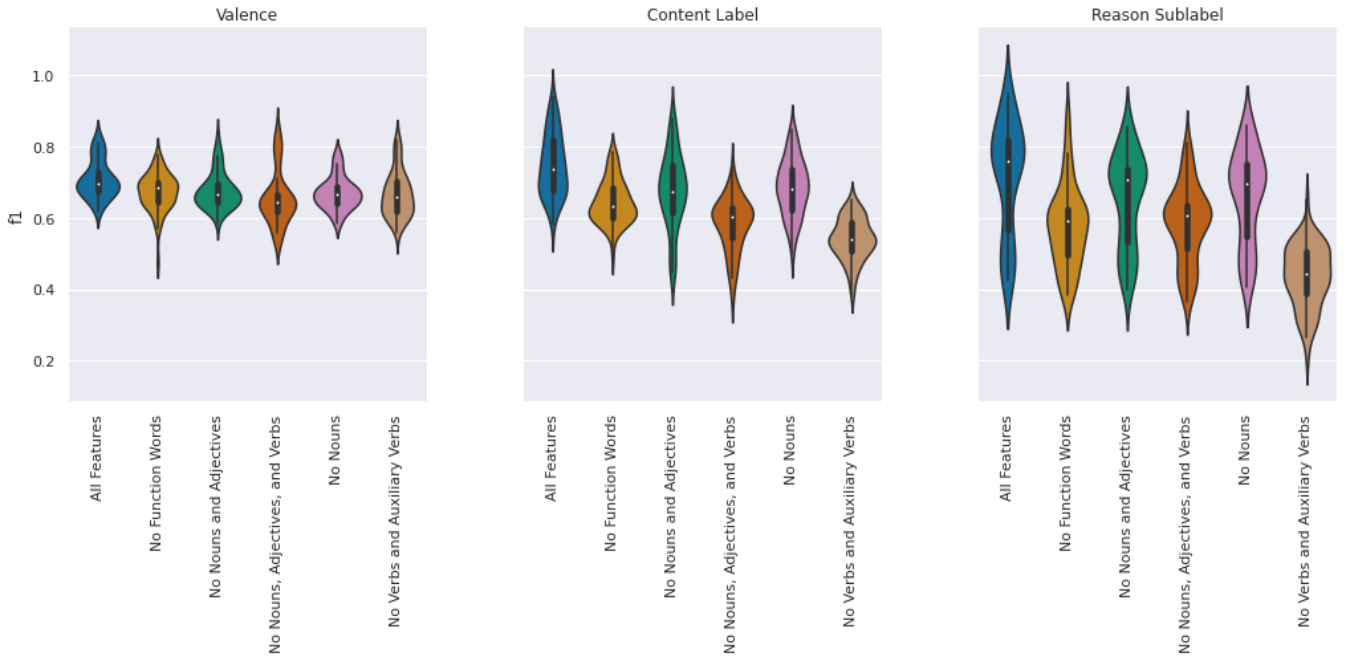

Figure 3: Classification results across all datasets when removing different lexical features in text preprocessing

is in-context if it has the same property as the GLoHBCD. For instance, for the Domain-dimension, the in-context group is made up of the Optifast Mock-Chatbot, the Synthetic GPT-3 Data, and the Instagram Data, since, like the GLoHBCD, they all focus on weight loss. The Smoking Cessation Forum, DARN-CT-based Wizard of Oz Dialogues and Health Coaching Dialogue Corpus are grouped as out-of-context data for this dimension.

For each dimension, we compare Macro F1 scores between the in-context and out-of-context condition (e.g., synchronous vs. asynchronous) using Mann-Whitney-U tests and plot the results, keeping the GLoHBCD results separate for reference (see Figure 2). Except for the Synchronicity-dimension and the content label classifications on the Domain-dimension, we find significant differences at $p < 0.05$ between in-context and out-of-context data for all dimensions and code-levels. Classification results on out-of-context data are significantly higher for all code-levels in the Interaction Type and Conversation Mode-dimensions, as well as the reason sublabels in the Domain-dimension. The only condition, in which classification results for out-of-context data are significantly lower than for in-context data, are valence codes in the Domain-dimension. This leads us to believe that the properties which distinguish language data from different sources do not have a big impact on the classification of *MISC* codes.

Including the GPT-3 data into the domain comparisons might have skewed the results slightly, since the GPT-3 outputs simulate potentially highly stereotypical utterances. For this reason, we compared each condition again after removing the GPT-3 classification results from the data. There was no condition in which removing the GPT-3 data led to significantly worse results for the out-of-context condition compared to the in-context condition, though the difference between groups was rendered insignificant for the reason sublabels on the Interaction Type-dimension. In addition, the out-of-context valence classifications on the Synchronicity-dimension had significantly higher scores than the in-context classifications when GPT-3 classification scores were removed.

Based on these results, it can be inferred that peer-to-peer forum data, when annotated using *MISC* labels on a sentence-to-sentence basis, exhibits similar characteristics to chatbot conversations, at least in terms of the properties which typically distinguish text data from different sources. We interpret this as strong evidence for the feasibility of transfer learning, i.e. classification between different conversational contexts and change topics.

4.4 RQ4: Exploring the Importance of Lexical Features for Cross-Context Classification

Finally, to investigate the role of lexical features for the identification of behaviour change classes on the different code-levels, we preprocess our data in different ways before re-running the classification experiments outlined in section 4.1. The experiments were, to some extent, guided by the analyses presented in Meyer and Elsweiler [27], which demonstrated that function words such as auxiliary verbs, rather than nouns, full verbs, or adjectives, were the primary keywords in the GLoHBCD that distinguished between the various classes. This led to the assumption that the semantic content of an utterance might play a smaller role in classification than the grammatical structure. We are thus focusing on removing different lexical features from the data in order to gauge their effect on classification. The aim is to better understand the language used when talking about behaviour change and determine if specific word types are hindering out-of-context classification. We preprocessed all datasets in five different ways using spaCy [21]:

- No Nouns - Remove only nouns

- No Nouns and Adjectives – Remove nouns, proper nouns and adjectives
- No Nouns, Adjectives, and Verbs – Remove nouns, proper nouns, adjectives, and verbs
- No Verbs and Auxiliary Verbs - Remove only verbs and auxiliary verbs
- No Function Words – Remove auxiliary verbs, adpositions, determiners, conjunctions, and pronouns

For each condition, we fine-tuned bert-base-german-cased on the preprocessed GLoHBCD training set using 10-Fold Cross-Validation. At each fold, the fine-tuned classifier was used to predict on all preprocessed test sets. The results across conditions are shown in Figure 3.

Again, the data violated the assumption of normal distribution, so we ran a Friedman's test for each code-level to identify differences between classification performances of classifiers when incorporating different lexical features. For all cases, Friedman's test showed significant differences (Valence: $\chi^2(5) = 28.51, p < 0.001$, Labels: $\chi^2(5) = 48.8, p < 0.001$; Sublabels: $\chi^2(5) = 48.29, p < 0.001$) and post-hoc analysis using pairwise Wilcoxon signed-rank test revealed that removing any lexical information led to a significant decrease in classification performance compared to the complete sentence. Removing verbs and auxiliary verbs led to the worst results on the content label and reason sublabel-level, whereas the worst results on the valence-level were obtained after removing nouns, adjectives, and verbs. The valence-level was the only condition, where removing function words led to better results than removing nouns. For content labels and reason sublabels, removing nouns and removing nouns and adjectives led to the best results of all conditions in which features were removed. This confirms the assumption voiced in Meyer and Elsweiler [27] that nouns, which are arguably likely to be the most topic-specific words in utterances about change, are less relevant for classification than function words. We also find that removing nouns, adjectives, and verbs leads to better results than removing verbs and auxiliary verbs for content labels and reason sublabels, even though more information is removed in the "No Nouns, Adjectives, and Verbs" condition. This indicates that auxiliary verbs are of specifically high importance when classifying content labels and reason sublabels. This result can be explained by the nature of the annotation scheme, where, for instance, commitment statements are usually oriented towards the future, whereas taking steps is typically directed at the past (see Table 1). Similarly, for reason sublabels, auxiliary verbs such as "must", "need", "want" are common markers for a class. Overall, these results show that semantically less and structurally more meaningful words which are likely to be stable across contexts are the most important for classifying *MISC* concepts for the content labels and reason sublabels, whereas valence classification might be slightly more reliant on topic-specific semantically meaningful words.

Since the content labels showed more classification variance depending on feature selection than the valence-level and the data was more balanced between content label classes than for the reason sublabels, we used the content label-level as an example to look at the classification results divided by dataset for each feature classifier (see Figure 4). We see variation in the results across datasets, which suggests that the most important lexical features for classification vary depending on dataset type. For all datasets except the Optifast Mock-Chatbot, removing nouns, or nouns and adjectives had the least effect on classification performance, with no significant differences between classification performance between the two conditions. Removing verbs and auxiliary verbs had the biggest impact on classification performance across datasets except for the Instagram Data, the Optifast Mock-Chatbot data, and the DARN-CT-based Wizard of Oz conversations. For these three datasets, the worst results were obtained for the "No Noun, Adjectives, and Verbs" condition, which is also the condition that removed the largest amount of information. This shows that full verbs alone, whose exclusion led to comparably high classification results for most datasets, do play a less important role in classification than auxiliary verbs. Still, all kinds of verbs are much more needed for classification than nouns and adjectives, and in live-chat contexts, auxiliary verbs appear to be of less importance than in less synchronous interactions, an effect that might also be caused by the more Question-Answer like structure of the Optifast Mock-Chatbot and the DARN-CT-based Wizard of Oz Dialogues. The Optifast Mock-Chatbot seemed to show the most irregularities compared to the other datasets, since nouns, adjectives, and verbs seemed to be a significant factor for correct classification of this dataset.

5 DISCUSSION

We collected datasets that represent different distinguishing properties between forum texts and interactions between humans and chatbots. These datasets were then used to evaluate the feasibility of classification transfer. Our findings indicate a strong probability of the practicality of cross-domain classification and, by extension, the presence of consistent patterns of written language regarding behaviour change across diverse conversational domains. This effect was more pronounced for the content labels and reason sublabels, which seemed to transfer more easily to different topics and contexts.

Our experiments revealed that valence utterances exhibit greater variability across diverse contexts, which can be partially elucidated by examining the classification outcomes obtained when certain lexical features were eliminated. These outcomes suggest that the reason for this disparity may be due to the fact that valence classification relies more on context-specific words, such as nouns and verbs, while the primary discriminative feature of content labels and reason sublabels are auxiliary verbs, which are expected to remain consistent across conversational contexts. This hypothesis is further corroborated by the observation that, for the domain dimension, the classification outcomes for out-of-context data were significantly poorer than those for in-context data, but only for the valence-level. In all other conditions, there was either no significant difference, or the out-of-context datasets had higher classification scores. These observations can also be explained by the annotation scheme, in which content labels and reason sublabels are clearly defined and often marked by specific words (such as *must*, *want to*, *need*, *will*), whereas the assignment of the valence label is more elusively guided by determining if the utterance is positive or negative regarding the change in question [27, 30].

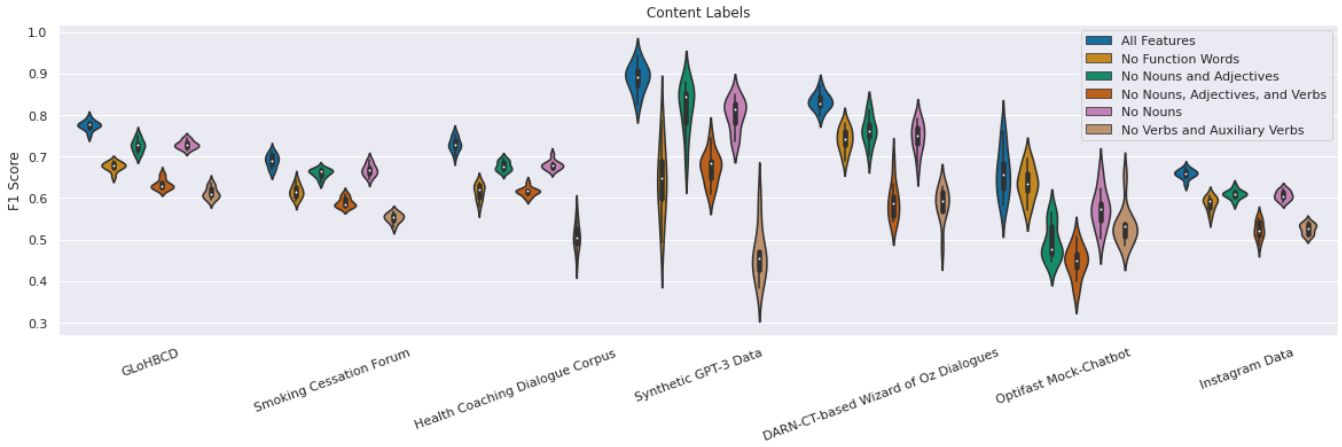

Figure 4: Classification results on the label-level for each dataset when removing different lexical features in data preprocessing

Judging by classification performance, the Optifast Mock-Chatbot and Instagram data seem to be the least similar to the GLoHBCD data. In the case of the Mock-Chatbot data, this can likely be explained by the fact that participants' were towards the end of the programme and had already lost a lot of weight. At the point of data collection, all participants were aiming to maintain their current (reduced) weight, putting them in a different stage of behaviour change than the forum users, who were mostly in the process of losing weight, or planning to do so in the near future [27]. The finding that individuals at different stages of change articulate their progress in diverse ways highlights the importance of considering users' present circumstances and mental states when encouraging and assisting their behaviour change efforts via conversational AI.

For the Instagram Data, the increased dissimilarity to the training data could be explained by the nature of the platform. Since Instagram is primarily used as a tool for self-promotion by many users, leading to a prevalence of positive content that puts posters into a good light and represents their "ideal self" [17, 22], it is likely that it leads to less honest interactions compared to a self-help forum or private conversations. Furthermore, since Instagram is frequently used to post selfies, the interactions on this platform are less anonymous and more publicly visible to broader audiences than those in the other datasets, which has been identified as a key differentiating factor in online discourse, along with platform-specific conventions and audience demographics [2, 14].

Overall, we find that the way people talk about change on peer-to-peer forums when separated into sentences closely resembles the way they talk about change in live-chat or short message scenarios as symbolized by the DARN-CT-based Wizard of Oz chats and the Health Coaching Dialogue Corpus, provided they are at a similar stage of behaviour change, and that the GLoHBCD appears to be sufficiently similar to chatbot-like conversational data across all dimensions, which we initially identified as potential distinguishing factors between conversational contexts, including the behaviour change objective. Although the GLoHBCD change topic is health focused, the conversational data collected in the DARN-CT-based Wizard of Oz conversations focused on New Year's resolutions, which also included non-health related topics such as increased productivity or gaining a happier mindset. The high classification scores reached when predicting on this dataset lead us to believe that the classifiers can be used for the reliable interpretation of change utterances even beyond health-related behaviour change contexts. Judging from this, we conclude that classifiers trained on the GLoHBCD can be used to identify users' thoughts around change in human-chatbot conversations across various change topics.

Applying these findings to conversational AI has the potential for more user-aware and personalized conversations around behaviour change. Incorporating the classification of *MISC* codes in the conversational design of CAs for behaviour change can expand existing efforts of automating MI [3, 18, 39] to cater more to the user's current situation and could strengthen motivation to change and self-efficacy to a larger degree than more general CAs. They can also serve as expansions for conversational mood-logging [6], as the *MISC* concepts include information implicitly related to users' moods and likelihood to change their behaviour, such as their perceived ability to change their behaviour or their desires [12], which are likely not recognized through mood-logging or traditional forms of sentiment analysis. As such, this study lays the groundwork for more flexible and realistic counselling-style CAs for behaviour change and facilitates the gathering of information from free text user inputs, advancing more naturalistic conversations between humans and CAs.

6 LIMITATIONS

Due to the difficulty of collecting data of behaviour change conversations in the different contexts and the high cost of annotation, some of our constructed datasets are small and only include very few samples for the less represented classes, which leads to a higher degree of F1 score variation, especially for the reason sublabels. We counteract this potential bias by repeating classification 10 times, each time training on 90% of the GLoHBCD training data. Still, cases where only few samples were present for a class tended to have more extreme F1 scores (either very high or very low). This is

to be expected, since in such cases a single misclassification has a much higher impact on the overall score than for larger classes. The data imbalances found in our datasets are likely to also appear in the wild, since users are likely to converse with a persuasive CA for behaviour change in different ways and for a different number of turns depending on their current situation (such as where they are in their behaviour change, what type of support they are looking for, how much they are lacking motivation at a particular moment).

Generally, this study has shown the feasibility of applying GLoHBCD classifiers to data that is more similar to chatbot conversations and spans different change topics. However, to elucidate the impact of the data imbalances in the datasets we used, we plan to further validate our findings in future work, by focusing on using the classifiers to guide conversational actions of a chatbot for behaviour change for different change topics. This will allow the collection of a larger dataset that can be used to draw further conclusions on how users converse about behaviour change with a CA and to what extent such interactions differ from the datasets introduced in this work. It could also be used to further increase domain and context independence of the existing classifiers, for instance by applying semi-supervised learning techniques [5].

7 CONCLUSION AND FUTURE WORK

This paper showcases the effectiveness of applying the insights gained from earlier studies on the language used for behaviour change in peer-to-peer forums for weight loss to different conversational contexts and topics. Consequently, it lays the groundwork for the creation of more user-centric conversational agents in the broader field of behaviour change. We see this work as the foundation for the usage of the tested classifiers to automatically interpret user utterances in conversations with conversational user interfaces. These interpretations can, in turn, be used to navigate automated MI-style conversations with users attempting behavioural changes across various contexts. Our future work will focus on applying this paper's findings to construct a conversational framework for agents for behaviour change that are grounded in the user's current state of mind and thoughts about change, and, following the implementation of a CA, test the effectiveness of employing these concepts in user studies.

ACKNOWLEDGMENTS

We thank the German Academic Scholarship Foundation for funding parts of this research project.

REFERENCES

[1] Jannis Androutsopoulos. 2013. Online data collection. *Data Collection in Sociolinguistics: Methods and Applications* (05 2013), 236–250. https://doi.org/10.4324/9780203136065

[2] Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources?. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. 356–364.

[3] Erkan Başar, Iris Hendrickx, Emiel Krahmer, Gert-Jan de Bruijn, and Tibor Bosse. 2022. Hints of independence in a pre-scripted world: on controlled usage of open-domain language models for chatbots in highly sensitive domains. In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence*. 401–407.

[4] Samar Bashath, Nadeesha Perera, Shailesh Tripathi, Kalifa Manjang, Matthias Dehmer, and Frank Emmert Streib. 2022. A data-centric review of deep transfer learning with applications to text data. *Information Sciences* 585 (2022), 498–528.

[5] Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022. A survey on data augmentation for text classification. *Comput. Surveys* 55, 7 (2022), 1–39.

[6] Robert Bowman, Benjamin R Cowan, Anja Thieme, and Gavin Doherty. 2022. Beyond Subservience: Using Joint Commitment to Enable Proactive CUIs for Mood Logging. In *Proceedings of the 4th Conference on Conversational User Interfaces*. 1–6.

[7] Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Srikumar. 2019. Observing Dialogue in Therapy: Categorizing and Forecasting Behavioral Codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5599–5611.

[8] Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's Next Language Model. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 6788–6796. https://doi.org/10.18653/v1/2020.coling-main.598

[9] Dawn Clifford and Laura Curtis. 2016. *Motivational interviewing in nutrition and fitness*. Guilford Publications.

[10] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies: why and how. In *Proceedings of the 1st international conference on Intelligent user interfaces*. 193–200.

[11] Avisha Das, Salih Selek, Alia R Warner, Xu Zuo, Yan Hu, Vipina Kuttichi Keloth, Jianfu Li, W Jim Zheng, and Hua Xu. 2022. Conversational bots for psychotherapy: A study of generative transformer models using domain-specific dialogues. In *Proceedings of the 21st Workshop on Biomedical Language Processing*. 285–297.

[12] BJ Fogg. 2009. A Behavior Model for Persuasive Design. In *Proceedings of the 4th International Conference on Persuasive Technology* (Claremont, California, USA) *(Persuasive '09)*. Association for Computing Machinery, New York, NY, USA, Article 40, 7 pages. https://doi.org/10.1145/1541948.1541999

[13] Brian J Fogg. 2009. The behavior grid: 35 ways behavior can change. In *Proceedings of the 4th international Conference on Persuasive Technology*. 1–5.

[14] David Giles, Wyke Stommel, Trena Paulus, Jessica Lester, and Darren Reed. 2015. Microanalysis Of Online Data: The methodological development of "digital CA". *Discourse, Context & Media* 7 (2015), 45–51. https://doi.org/10.1016/j.dcm.2014.12.002

[15] Sergey Golovanov, Alexander Tselousov, Rauf Kurbanov, and Sergey I Nikolenko. 2020. Lost in conversation: A conversational agent based on the transformer and transfer learning. In *The NeurIPS'18 Competition: From Machine Learning to Intelligent Conversations*. Springer, 295–315.

[16] Itika Gupta, Barbara Di Eugenio, Brian Ziebart, Aiswarya Baiju, Bing Liu, Ben Gerber, Lisa Sharp, Nadia Nabulsi, and Mary Smart. 2020. Human-Human Health Coaching via Text Messages: Corpus, Annotation, and Analysis. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 1st virtual meeting, 246–256. https://aclanthology.org/2020.sigdial-1.30

[17] Elspbeth Harris and Aurore C Bardey. 2019. Do Instagram profiles accurately portray personality? An investigation into idealized online self-presentation. *Frontiers in Psychology* 10 (2019), 871.

[18] Linwei He, Erkan Basar, Reinout W Wiers, Marjolijn L Antheunis, and Emiel Krahmer. 2022. Can chatbots help to motivate smoking cessation? A study on the effectiveness of motivational interviewing on engagement and therapeutic alliance. *BMC Public Health* 22, 1 (2022), 726.

[19] Paul J Hershberger, Yong Pei, Dean A Bricker, Timothy N Crawford, Ashutosh Shivakumar, Miteshkumar Vasoya, Raveendra Medaramitta, Maria Rechtin, Aishwarya Bositty, and Josephine F Wilson. 2021. Advancing motivational interviewing training with artificial intelligence: ReadMI. *Advances in Medical Education and Practice* (2021), 613–618.

[20] Jennifer Hill, W. Randolph Ford, and Ingrid G. Farreras. 2015. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior* 49 (2015), 245–250. https://doi.org/10.1016/j.chb.2015.02.026

[21] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. (2020). https://doi.org/10.5281/zenodo.1212303

[22] Yuheng Hu, Lydia Manikonda, and Subbarao Kambhampati. 2014. What we instagram: A first analysis of instagram photo content and user types. In *Eighth International AAAI conference on weblogs and social media*.

[23] Zac E Imel, Brian T Pace, Christina S Soma, Michael Tanana, Tad Hirsch, James Gibson, Panayiotis Georgiou, Shrikanth Narayanan, and David C Atkins. 2019. Design feasibility of an automated, machine-learning based feedback system for motivational interviewing. *Psychotherapy* 56, 2 (2019), 318.

[24] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 986–995.

[25] M Beatrice Ligorio. 2001. Integrating communication formats: Synchronous versus asynchronous and text-based versus visual. *Computers & Education* 37, 2 (2001), 103–125.

[26] Siru Liu, Jili Li, and Jialin Liu. 2021. Leveraging transfer learning to analyze opinions, attitudes, and behavioral intentions toward COVID-19 vaccines: social media content and temporal analysis. *Journal of Medical Internet Research* 23, 8 (2021), e30251.

[27] Selina Meyer and David Elsweiler. 2022. GLoHBCD: A Naturalistic German Dataset for Language of Health Behaviour Change on Online Support Forums. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 2226–2235.

[28] Selina Meyer, David Elsweiler, Bernd Ludwig, Marcos Fernandez-Pichel, and David E Losada. 2022. Do We Still Need Human Assessors? Prompt-Based GPT-3 User Simulation in Conversational AI. In *Proceedings of the 4th Conference on Conversational User Interfaces*. 1–6.

[29] Susan Michie, Maartje M Van Stralen, and Robert West. 2011. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. *Implementation science* 6, 1 (2011), 1–12.

[30] William R Miller, Theresa B Moyers, Denise Ernst, and Paul Amrhein. 2003. Manual for the motivational interviewing skill code (MISC). *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico* (2003).

[31] William R Miller and Stephen Rollnick. 2002. *Motivational Interviewing, Second Edition: Preparing People for Change*. Guilford Publications.

[32] Juanan Pereira and Óscar Díaz. 2019. Using health chatbots for behavior change: a mapping study. *Journal of medical systems* 43, 5 (2019), 1–13.

[33] Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2016. Building a motivational interviewing dataset. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*. 42–51.

[34] Mohammad Hasnain Rajan, Keith Rebello, Yajur Sood, and Sunil B Wankhade. 2021. Graph-Based Transfer Learning for Conversational Agents. In *2021 6th International Conference on Communication and Electronics Systems (ICCES)*. IEEE, 1335–1341.

[35] Amon Rapp, Lorenzo Curti, and Arianna Boldi. 2021. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies* 151 (2021), 102630. https://doi.org/10.1016/j.ijhcs.2021.102630

[36] Ken Resnicow, Fiona McMaster, and Stephen Rollnick. 2012. Action reflections: a client-centered technique to bridge the WHY–HOW transition in motivational interviewing. *Behavioural and cognitive psychotherapy* 40, 4 (2012), 474–480.

[37] Lee M Ritterband, Frances P Thorndike, Daniel J Cox, Boris P Kovatchev, and Linda A Gonder-Frederick. 2009. A behavior change model for internet interventions. *Annals of Behavioral Medicine* 38, 1 (2009), 18–27.

[38] Stephen Rollnick, William R Miller, Christopher C Butler, and Mark S Aloia. 2008. *Motivational interviewing in health care: helping patients change behavior*. Taylor & Francis.

[39] Ahson Saiyed, John Layton, Brian Borsari, Jing Cheng, Tatyana Kanzaveli, Maksim Tsvetovat, and Jason Satterfield. 2022. Technology-Assisted Motivational Interviewing: Developing a Scalable Framework for Promoting Engagement with Tobacco Cessation Using NLP and Machine Learning. *Procedia Comput. Sci.* 206, C (jan 2022), 121–131. https://doi.org/10.1016/j.procs.2022.09.091

[40] Raj Sanjay Shah, Faye Holt, Shirley Anugrah Hayati, Aastha Agarwal, Yi-Chia Wang, Robert E Kraut, and Diyi Yang. 2022. Modeling Motivational Interviewing Strategies On An Online Peer-to-Peer Counseling Platform. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–24.

[41] Mary Sun. 2022. *Natural Language Processing for Health System Messages: Deep Transfer Learning Approach to Aspect-Based Sentiment Analysis of COVID-19 Content*. Ph. D. Dissertation. Harvard University.

[42] Michael Tanana, Kevin Hallgren, Zac Imel, David Atkins, Padhraic Smyth, and Vivek Srikumar. 2015. Recursive neural networks for coding therapist and patient behavior in motivational interviewing. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*. 71–79.

[43] Leili Tavabi, Kalin Stefanov, Larry Zhang, Brian Borsari, Joshua D Woolley, Stefan Scherer, and Mohammad Soleymani. 2020. Multimodal automatic coding of client behavior in motivational interviewing. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 406–413.

[44] Leili Tavabi, Trang Tran, Kalin Stefanov, Brian Borsari, Joshua D Woolley, Stefan Scherer, and Mohammad Soleymani. 2021. Analysis of behavior classification in motivational interviewing. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, Vol. 2021. NIH Public Access, 110.

[45] Lu Wang, Munif Ishad Mujib, Jake Williams, George Demiris, and Jina Huh-Yoo. 2021. An Evaluation of Generative Pre-Training Model-based Therapy Chatbot for Caregivers. *arXiv preprint arXiv:2107.13115* (2021).

[46] Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. 2022. Anno-MI: A Dataset of Expert-Annotated Counselling Dialogues. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6177–6181.

[47] Bei Xu and Ziyuan Zhuang. 2022. Survey on psychotherapy chatbots. *Concurrency and Computation: Practice and Experience* 34, 7 (2022), e6170.

[48] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too?. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2204–2213.