Exploring the Challenges of Behaviour Change Language Classification: A Study on Semi-Supervised Learning and the Impact of Pseudo-Labelled Data

Selina Meyer*, Marcos Fernández-Pichel†, David Elsweiler*, David E. Losada†
*Regensburg University
Universitätsstraße 31, 93053 Regensburg
†Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS),
Universidade de Santiago de Compostela
Rúa Jenaro de la Fuente s/n, 15782 Santiago de Compostela
{selina.meyer, david.elsweiler}@ur.de
{marcosfernandez.pichel, david.losada}@usc.es

Abstract

Automatic classification of behaviour change language can enhance conversational agents' capabilities to adjust their behaviour based on users' current situations and to encourage individuals to make positive changes. However, the lack of annotated language data of change-seekers hampers the performance of existing classifiers. In this study, we investigate the use of semi-supervised learning (SSL) to classify highly imbalanced texts around behaviour change. We assess the impact of including pseudo-labelled data from various sources and examine the balance between the amount of added pseudo-labelled data and the strictness of the inclusion criteria. Our findings indicate that while adding pseudo-labelled samples to the training data has limited classification impact, it does not significantly reduce performance regardless of the source of these new samples. This reinforces previous findings on the feasibility of applying classifiers trained on behaviour change language to diverse contexts.

Keywords: Behaviour Change, Semi-Supervised Learning, Low Resource Application Areas

1. Introduction

The way people talk about change can be an indicator for future success of their attempts to alter their behaviour (Magill et al., 2018; Moyers et al., 2007). Different types of language indicate varying levels of intent to change (Resnicow et al., 2012) and being able to automatically differentiate these language types could improve conversational agents (CAs) with the purpose of assisting behaviour change. For example, a CA could adapt its behaviour to a user's current situation and motivational level and elicit more favourable utterances in order to increase the user's resolve to change. Additionally, obtaining such information from patient texts, such as journals, could serve as a meaningful resource for practitioners, helping them gain deeper understandings of the patient's current situation (Kim et al., 2023)

Current CAs fail to use this information (Xu and Zhuang, 2022), not least because of a lack of annotated text around behaviour change. The availability of new datasets with labelled utterances would facilitate the construction of supervised learning solutions (Meyer and Elsweiler, 2022). However, such data is difficult to obtain for two main reasons. First, it is commonly sourced from transcripts of therapy sessions or counsellor training materials, leading to privacy and data security concerns that complicate the publication of datasets. This naturally limits the size of the data. Second, the costs of annotation and the necessary training of the assessors hinder the creation of large datasets with fine-grained annotations (Pérez-Rosas et al., 2016; Wu et al., 2022).

Additionally, behaviour change language exhibits certain peculiarities. It can be applied to different kinds of unrelated *target behaviours*, e.g. increasing physical activity or smoking cessation, which demands the ability of a classification algorithm to be able to transfer between contexts. Talking about behaviour change also naturally leads to highly imbalanced data. Certain types of utterances, such as reasons for or against change, tend to appear often while others, such as statements about specific commitments for the future, are less frequent (Lord et al., 2015).

Semi-supervised learning (SSL) has commonly been used to alleviate the limitations imposed on classification performance by a scarcity of training data. SSL has shown to be particularly successful in popular benchmarks (Van Engelen and Hoos, 2020; Duarte and Berton, 2023). In this paper, we explore the feasibility of using SSL in the context of classifying highly imbalanced text about behaviour change. We explore the inclusion of pseudo-labelled data, both from the original source and from new sources covering different behaviour change contexts. We evaluate to what extent including new pseudo-labelled data from different sources impacts the classifier's ability to correctly predict the utterance class of short texts. Furthermore, we work with several out-of-context test datasets and explore the trade-off between the amount of pseudo-labelled data added and the strictness of the inclusion criteria for the pseudo-labelled samples[1].

2. Background & Related Work

2.1. SSL for Text Classification

In their survey paper, Van Engelen and Hoos (2020) give an extensive overview of different SSL models and common application areas. A more recent review, by Duarte and Berton (2023), focuses specifically on the application of SSL methods to text classification. According to them, one of the most commonly explored types of SSL for text classification is self-learning, where a baseline classifier is used to assign pseudo-labels to new, unlabelled data. These pseudo-labelled examples are then included in the training data of the classifier. This classic approach is simple and has the advantage of being suited to be used in combination with any base learner (Van Engelen and Hoos, 2020).

Past work has shown the effectiveness of SSL in various domains, including health and well-being. For instance, Varma and Ré (2018) presented a tool for automatically generating weak supervision rules for data labelling. The authors demonstrated the effectiveness of this method in spam classification and medical diagnosis. In the same vein, Ratner et al. (2020) presented a tool to streamline the process of creating training data with weak supervision techniques. The usefulness of this tool, which allowed users to rapidly define labelling functions, was demonstrated in real-life applications such as medical information extraction and knowledge base construction.

Other studies have focused on mitigating weaknesses frequently associated with SSL techniques. For instance, there is often an inherent proneness to class imbalance, which is observable even when the baseline classifier is trained on balanced data (Wang et al., 2022). Real-world data is rarely balanced. Guo and Li (2022) addressed this problem by introducing a framework that supports adaptive thresholding for different classes. Their approach is effective without prior knowledge of a dataset's class distribution.

SSL has been frequently applied to publicly available and widely researched benchmarks. These experiments often yielded solid results (Van Engelen and Hoos, 2020; Duarte and Berton, 2023). However, recent studies have argued that performance on these datasets does not always equal reliability and robustness in real-world applications (Kiela et al., 2021; Schlegel et al., 2022; Church and Kordoni, 2022). It is hard to predict to which extent SSL is beneficial for a given situation (Van Engelen and Hoos, 2020), with many studies even reporting decreases in classification performance (Oliver et al., 2018; Li and Zhou, 2014). Because of this potential for deterioration, we chose to first evaluate the effect of SSL for behaviour change language using self-training, and leave the exploration of other, more sophisticated SSL methods to future work.

2.2. Behaviour Change Language

One way to formalise talk about behaviour change is the Motivational Interviewing Skill Code (*MISC*) (Miller et al., 2003). It helps to categorise utterances into different valences and topics around behaviour change across multiple target behaviours. While Motivational Interviewing (MI) was initially developed for addiction counselling, it has since been used for various topics, ranging from smoking cessation, over nutrition and fitness, to work-related behaviour (Miller and Rollnick, 2002; Clifford and Curtis, 2016; Page and Tchernitskaia, 2014; Güntner et al., 2019).

The *MISC* defines different categories for utterances, which we outline in Table 1. Based on the *MISC*, each user utterance that is not Follow/Neutral is assigned a valence and a topic. If the topic is *Reason*, the utterance is also assigned a reason type. This annotation framework can help to infer a person's intensity of commitment to behaviour change (Resnicow et al., 2012). For example, the *MISC* helps to understand how confident people feel about change, what type of rationale leads them to pursue change and whether they have already become active or are planning to do so in the near future.

Past research on classifying these behaviour codes has largely focused on the distinction between Change Talk, Follow/Neutral and Sustain Talk, and the few existing public MI-datasets do not contain topic and reason type annotations (Wu et al., 2022; Pérez-Rosas et al., 2016). This lack of fine-grained annotations hinders the development of more sophisticated classifiers that take into account the topic of user utterances and the types of reasons they voice for making a change.

An exception to this is the GLoHBCD, a German dataset that contains written forum data annotated with valences, topics, and reason types based on the *MISC* (Meyer and Elsweiler, 2022). The creators of the GLoHBCD demonstrated the feasibility of training transformer-based classifiers on the data, reaching macro F1 scores between 70% and 77% depending on the label-level. How-

[1] We make our code available on GitHub.

Level	Label	Description
Valence	Change Talk (+)	Utterances in favour of behaviour change
	Sustain Talk (-)	Utterances in favour of status quo
Topic	Reason	Reasons for/against change
	Taking Steps	Specific steps taken in the recent past
	Commitment	Agreement, intention, or obligation for the near future
Reason Type	Ability	Ability and degree of difficulty of the change
	Need	Necessity of change, or maintaining the status quo
	Desire	Desire for change, or current behaviour
	General	General justifications, incentives, or justifications
Follow/Neutral (FN)		Utterances not related to behaviour change

Table 1: Description of utterance classifications, based on (Miller et al., 2003; Meyer and Elsweiler, 2022)

ever, these experiments also showed that some label-levels are harder to classify and that the imbalanced nature of the data can be problematic. In further experiments, the same team showed that the classification of these utterances transfers to a certain extent between different target behaviours and conversational contexts (Meyer and Elsweiler, 2023).

With macro F1 scores ranging mostly between 60% and 90%, the classification results reached on out-of-context datasets suggest a certain degree of stability, but still leave much room for improvement. The GLoHBCD consists of only 4724 data points relevant to behaviour change, and the less represented classes include less than 200 samples, which makes it likely that introducing more data would lead to improved classification.

3. Datasets

In this paper, we intend to build on the results presented by Meyer and Elsweiler (2022, 2023) and determine the feasibility of applying SSL approaches to the GLoHBCD. We aim at increasing classification performance on the original dataset and, additionally, employing the classifiers on external chat-like conversational data about different target behaviours. To explore this, we have collected new data from the same source as the GLoHBCD, as well as from other sources. In this section, we first outline the main properties of the GLoHBCD (§3.1) and then give an overview of the data sources used for pseudo-labelling (§3.2) and the test sets used to evaluate transfer learning capabilities (§3.3).

3.1. GLoHBCD

The GLoHBCD is a dataset of forum posts, written by people trying to lose weight, which was annotated with labels based on the *MISC* (Meyer and Elsweiler, 2022). The data was collected in August 2020 and written between May 2006 and July 2020. It stems from two subforums of a large-scale German weight loss forum, which were initially pre-screened for utterances around motivation for weight loss, after which relevant posts were annotated on a sentence-level basis. Each data point consists of a single sentence from the forum, together with a valence, a topic, and, if the topic is reason, a reason type annotation, as defined in Table 1.

3.2. Data Used for Pseudo-labelling

We used three different datasets as sources for pseudo-labelled text, one of them stemming from the same source as the GLoHBCD, another coming from a different source with the same conversational context, and a third being sourced from spoken interactions about a variety of target behaviours. This allowed us to explore to what extent adding new data from different contexts, which likely introduces more linguistic variety, can be used to improve classification of new data.

Weight Loss Forum Data For the Weight Loss Forum Data, we collected new posts from the same source as the GLoHBCD. We collected all posts published after the extraction date of the initial GLoHBCD data (August 2020). There was no manual pre-filtering of this new data, which consists of 992 sentences and serves as in-domain data for pseudo-labelling.

Smoking Cessation Forum Data The Smoking Cessation Forum Data consists of data that is similar to the GLoHBCD, in the sense that it also consists of forum data. However, this dataset consists of reports of people attempting to quit smoking. As such, it represents data from the same type of source, but from a different context as the original dataset. The dataset was created by Meyer and Elsweiler (2023) and includes ground truth *MISC* annotations for each of the 662 sentences in the dataset. We can use these annotations to evaluate the effect of adding pseudo-labelled samples

Dataset	Domain/Target behaviour	Context	# sentences
Health Coaching Dialogue Corpus[2]	step count increase	Text conversations with health coach	508
Optifast Mock-Chatbot	weight loss	Text conversation with simulated motivational chatbot	90
DARN-CT-based Wizard of Oz Dialogues	New Year's resolutions	Text conversations with simulated motivational chatbot	80
Synthetic GPT-3 Data[3]	weight loss	User simulation through eliciting questions	74
GLoHBCD (test split)	weight loss	Forum - Interaction between peers	924
Smoking Cessation Forum (test split)	smoking cessation	Forum - Interaction between peers	199

Table 2: Overview of test datasets introduced in Meyer and Elsweiler (2023) to evaluate domain transfer capabilities of classifiers (table adapted from Meyer and Elsweiler (2023))

to the training data on the classification of data from different sources as the original training data. To explore this, we use 10% of this data collection as a test set to evaluate the performance of the final model. The remaining sentences are used as a source for pseudo-labelled data.

AnnoMI The AnnoMI dataset is a collection of transcribed MI sessions across a variety of behaviour change contexts (Wu et al., 2022). While still being language data related to behaviour change uttered by humans, this dataset differs both in context (topics range from weight loss and smoking cessation, across alcohol abuse to other issues) and source type, as the data is transcribed from spoken counselling sessions, whereas the GLoHBCD consists of peer-to-peer conversations in a written forum. As such, this dataset is the furthest away from the original dataset and could thus offer the largest increase in linguistic variance. Since the dataset only includes valence annotations of client utterances, we use all client utterances which are not annotated as Follow/Neutral as data to pseudo-label for our experiments. Since for the remaining datasets used in this study each sentence constitutes a single data point, we separate the utterances in the AnnoMI into sentences following the same approach as for the other datasets, resulting in 2481 sentences.

3.3. Data used for Testing the SSL Classifier

Finally, we use multiple test sets to evaluate the ability of the SSL classifier to predict the type of behaviour change utterance. This includes a broad range of collections, ranging from written chat-like conversations to forum and spoken interactions, assembled by Meyer and Elsweiler (2023) to evaluate transfer learning capabilities of classifiers trained on GLoHBCD data. In this way, we can evaluate the transfer learning capabilities of the SSL classifier. This is intended to give insights about the effects of adding pseudo-labelled data from the original source (and from other sources) on the ability of the classifier to recognise utterances under varying conditions. Introducing test data from such a broad variety of contexts tells us to which extent adding pseudo-labelled data from multiple sources benefits or hinders classification of new data with varying degrees of closeness to the GLoHBCD.

In Table 2 we give an overview of the datasets used for testing, their conversational context, and behaviour change domain. Following Meyer and Elsweiler (2023), we included synthetically generated chat data, which can be seen as stereotypical utterances about change. This acts as a sanity check, since a decrease in performance on this dataset after adding pseudo-labelled data would indicate a significant increase in noise. We also create an 80%-20% split of the GLoHBCD, using the 20% as a final test set, whereas the remaining 20% are used for training the baseline classifiers.

4. Experimental Setup

We ran experiments across four stages, which we will outline in this section. The first three stages are made up of fine-tuning experiments, whereas the fourth stage applies the findings to the test sets. For fine-tuning, we followed the following methodology: In 10-fold cross-validation, i) a BERT-based classifier is fine-tuned on the GLoHBCD training data (baseline classifier), ii) new data is pseudo-labelled, iii) GLoHBCD training data and pseudo-

[1] data based on Gupta et al. (2020) with annotations by Meyer and Elsweiler (2023)

[2] based on Meyer et al. (2022) with annotations by Meyer and Elsweiler (2023)

Valence		Topic		Reason Type	
System	Macro F1	System	Macro F1	System	Macro F1
baseline	72.65 (1.96)	baseline	74.05 (3.17)	baseline	75.63 (2.93)
NP, CT(0.5)	73.98 (2.29)	NP, CT(0.95)	75.89 (3.82)	P(0.7), CT(0.5), min	76.7 (2.96)
NP, CT(0.5), equal	73.73 (1.85)	P(0.5), CT(0.95)	75.46 (3.47)	P(0.7), CT(0.5)	76.28 (4.33)
NP, CT(0.1)	73.55 (2.36)	P(0.7), CT(0.95)	75.35 (3.15)	P(0.7), CT(0.4)	76.25 (3.35)

Table 3: Comparison of classification setups on gLoHBCD cross-validation splits with baseline (no SSL). Variants include Pre-filtering (P(.)) and No Prefiltering (NP). Confidence thresholds for sample incorporation and classification indicated as P(t) and CT(t) respectively. If threshold < 0.5, points labeled minority class if predicted confidence > threshold. Equal: equal samples, Min: only new minority class samples included.

labelled data are combined to fine-tune an SSL classifier, and iv) the SSL classifier and baseline classifier are evaluated against the validation split of the cross-validation. This process is repeated for each label-level (see Table 1). Figure 1 provides a visual overview of the experimental setup.

4.1. Stage 1: Pre-filtering and Confidence Thresholds

The careful selection of new data, for example, by excluding data points with low-confidence classifications with the help of a baseline classifier, has been shown to be essential for successfully applying SSL methods (Van Engelen and Hoos, 2020). To achieve this, i) we use a relevance filter supplied by the GLoHBCD authors[4] to weed out change-unrelated (Follow/Neutral) sentences, and ii) we compare different confidence thresholds for pre-filtering and pseudo-labelling. We test all combinations of three confidence thresholds (0.5, 0.7, 0.95) for both the relevance filter and the baseline classifier that is used to pseudo-label new data. To avoid noise, we use only Weight Loss Forum Data (§3.2) as a source for pseudo-labelled data at this stage, as it stems from the same source as the GLoHBCD.

4.2. Stage 2: Class Imbalance

Pseudo-labelled data is prone to class-imbalance even with a balanced baseline classifier (Wang et al., 2022). Such imbalance can severely impact performance (Guo and Li, 2022). In our first experimental stage, the majority class dominated pseudo-labelling, possibly suppressing SSL improvements. To address this, we tested additional strategies to boost minority-class representation.

From stage 1, we selected optimal pre-filter threshold combinations for each label-level. We then test the following variants: i) adding only minority-class pseudo-labelled samples, ii) adding equal amount of pseudo-labelled samples for all classes, based on the number of minority-class samples, and iii) pseudo-labelling as minority class even with low confidence (thresholds: 0.4, 0.3, 0.2, 0.1).

4.3. Stage 3: Amount and Domain of Pseudo-labelled Data

After initial proofs considering only Weight Loss Forum data, we wanted to assess to what extent the amount of new data added and the domain of pseudo-labelled data impact classification performance and transfer learning. To this end, we included the two other datasets described in §3.2, and tested the following combinations of datasets as providers of pseudo-labelled samples: i) Weight Loss Forum only, ii) Smoking Cessation Forum only, iii) AnnoMI only, iv) Weight Loss Forum + Smoking Cessation Forum, and v) Weight Loss Forum + Smoking Cessation Forum + AnnoMI.

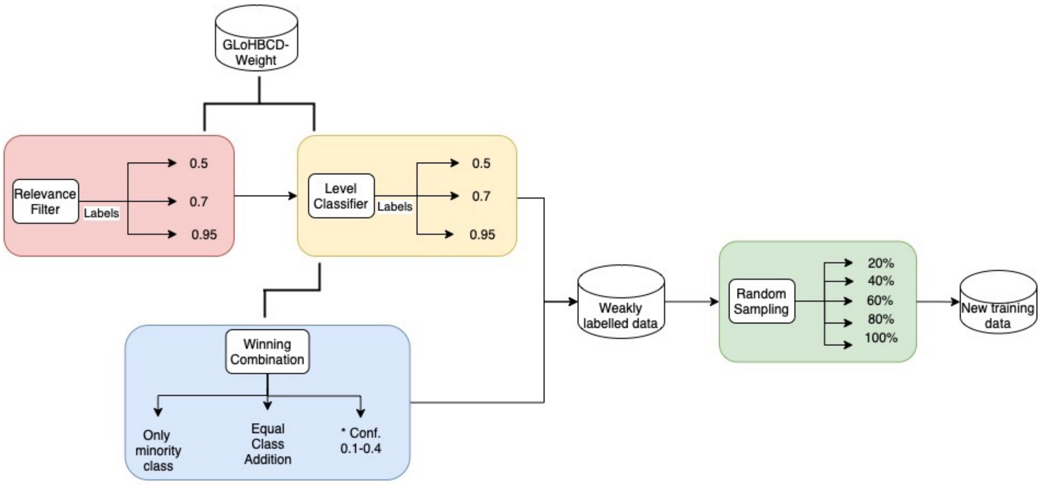
At this stage, the confidence thresholds were set to those that yielded the best results in Stages 1 and 2. We added varying shares of pseudo-labelled data to the original training data (between 20%-100% in 20% increments).

Since the data from the Smoking Cessation Forum contains ground truth labels, we incorporated these examples to the 10-fold cross-validation experiments (at each round 90% of them were pseudo-labelled and fed to the classifier and the remaining 10% of them were included into the validation fold along with the GLoHBCD validation data).

4.4. Stage 4: Application to Test Sets

In this final stage, we combine insights from stages 1-3 and applied the best performing system for each label-level to the independent test sets. Examining the SSL approach on data derived from chat-like conversational contexts and spanning various behaviour change domains aids in gauging its effectiveness and transfer learning capabilities. The main goal was to determine what kind of out-of-context data might benefit the most from

[4] https://huggingface.co/selmey/behaviour_change_prefilter_german



* Only includes winning relevance filter, overrides winning level classifier confidence threshold

Figure 1: Framework plot of experimental setup.

SSL. We also expected to build on previous results (Meyer and Elsweiler, 2023) and gain additional insights about the interaction between a dataset's properties and the difficulty of mining change behaviour cues from it.

5. Results

In Table 3 we summarise the classification results of the top three conditions from stages 1 and 2, in which only Weight Loss Forum data was pseudo-labelled, on the cross-validation splits of the GLoHBCD for each label-level compared to the baseline classifier. These results suggest that adding pseudo-labelled samples from the same source as the GLoHBCD has a minor yet discernible positive effect on classification. Most of the tested variants led to some improvements compared to the baseline, although the improvements were modest, and we did not observe any statistically significant differences between conditions. This outcome could potentially be attributed to the low amount of available new data.

Although there was no significant improvement in performance, none of the classification tasks experienced a decrease in performance when new pseudo-labelled data was added. When analysing the class-specific F1 scores and the amount of new data points added per class, we noticed that the F1 scores of the minority classes vary more than those of the majority classes. The amount of data labelled as the minority class is generally small, even in conditions where the confidence threshold for a sample to be labelled as the minority class was set lower than 0.5. In Figure 2, we show that this effect can be observed across all classification experiments (valence, topic, and reason type). Based on the results of stage 1 and 2, the systems chosen for the next stage of experiments were the following:

valence level: no prefilter, confidence threshold 0.5 (NP, CT(0.5)),

topic level: no prefilter, confidence threshold 0.5 (NP, CT(0.95)),

reason type level: prefilter with confidence 0.7, confidence threshold 0.5, and adding only minority samples (P(0.7), CT(0.5), min).

Applying those systems in stage 3 of experiments led to more stable classification results for the GLoHBCD validation sets compared to the Smoking Cessation Forum validation sets. This was expected since the smoking cessation data is from a different source and domain than the original training data, and has fewer samples.

However, regardless of the validation set and the type of pseudo-labelled data added, the results do not show a clear increase of performance when more data is added. Only in a few instances did adding out-of-domain data lead to improvements of in-domain classification. The effects of SSL seems to be slightly more apparent on the reason type level. For example, adding weight loss forum data led to slight improvements in reason type classifications of the GLoHBCD and Smoking Cessation Forum validation sets. The reason type classifiers work with few labelled data points from the original training data, thus presumably allowing pseudo-labelled samples to have more influence.

Based on the results obtained in stage 3, we included different shares of pseudo-labelled data from Weight Loss Forum, Smoking Cessation Forum and AnnoMI to predict on the test sets in Stage 4 (see Table 3). For valence and reason type classification, we included 20% of the pseudo-labelled

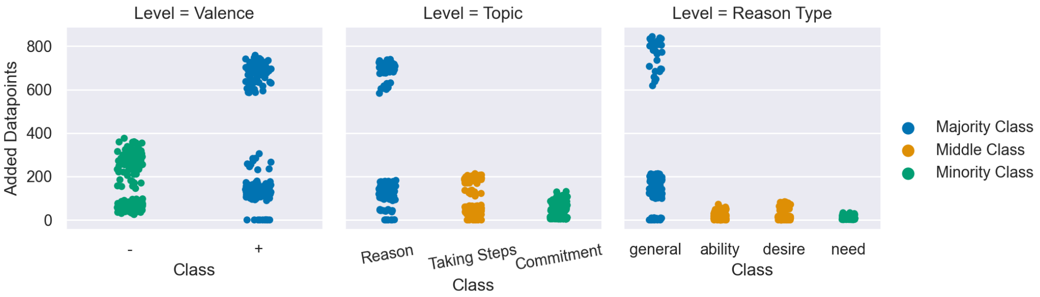


Figure 2: Amount of pseudo-labelled data points added to the training data across conditions by class.

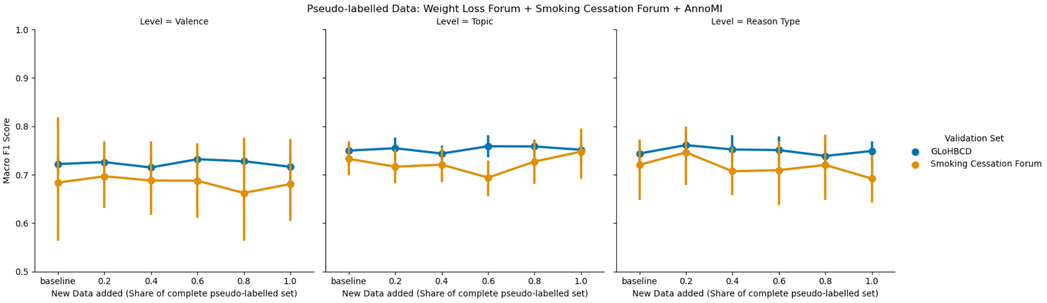


Figure 3: Change in classification performance when more pseudo-labelled data is added to training.

data, and for topic classification we included all pseudo-labelled samples. In Figure 4, we compare the classification performance of the baseline classifier without pseudo-labelled data and the best performing system from stages 1-3 for each label level.

For the topic level, all datasets but the Health Coaching Dialogue Corpus experienced some improvement in performance. Effects on valence and reason type were larger and more varied. The valence predictions by the SSL classifier on the Smoking Cessation Forum, the synthetic GPT-3 data, and the Health Coaching Dialogue Corpus were better than those of the baseline. Still, the SSL classifier produced poorer valence results for the Wizard of Oz dialogues, GLoHBCD and Optifast Data. For reason type, decreases in performance were observed for the synthetic GPT-3 data and the GLoHCBD test set, while performance on Optifast Data remained the same and all other datasets benefited from the inclusion of the pseudo-labelled data.

6. Discussion

Weak supervision has shown promising results in multiple previous studies working with curated benchmark datasets (Van Engelen and Hoos, 2020; Duarte and Berton, 2023). However, its effects appear to be more elusive when applied to imbalanced data. Although we found some slight improvements when applying the SSL-classifiers to test datasets, transfer learning did not improve for all out-of-context data. With the baseline classifier reaching F1 scores between 70% and 80% on in-context data, one potential reason for the lack of stable classification improvements could be unsteady behaviour of the baseline classifier when labelling new data.

In their survey study, Longpre et al. (2020) highlighted that simply augmenting the training data of large pre-trained transformer models is insufficient to enhance classification performance. The reason behind this limitation lies in the fact that augmentation alone does not introduce the necessary linguistic variety to impart new knowledge to these powerful models. Drawing from this argument, one possible explanation for the minimal impact observed when incorporating pseudo-labelled samples, regardless of their source or label level, could be attributed to the uniformity of language surrounding behaviour change across different conversational contexts and behaviour change topics.

This observation aligns with the findings presented by Meyer and Elsweiler (2023), who explored the transfer learning capabilities of classifiers for behaviour change language. In such a context, the addition of new training data, even from divergent sources, may not produce the required “newness” to improve classification performance. This is further exemplified by the fact that the AnnoMI, the largest dataset added during pseudo-labelling, not only stems from vastly differ-

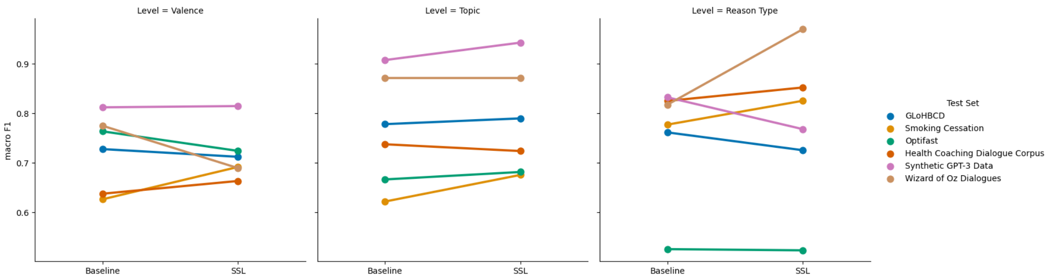


Figure 4: Comparison of classification results on multiple test sets. The plots present the performance of the baseline classifier (no semi-supervised learning) and a semi-supervised learning (SSL) classifier whose configuration was set based on the experiments of stages 1-3.

ent change scenarios, but even from a different modality (spoken conversation) compared to the GLoHBCD and this did not have a stable negative impact on classification results.

Generally, we discovered that the classification of behaviour change language remains stable and is not harmed by the inclusion of new data from alternative sources. These results speak in favour of the application of classifiers trained on behaviour change language to novel contexts. That being said, while our experiments do confirm that there are large parts of the data that seem to be very similar across contexts and target behaviours, there also seem to be some utterances that are more context-specific and might not be picked up correctly by the baseline classifiers used for pseudo-labelling.

Sentence: Aber dennoch heißt es heute, ganz besonders acht geben auf mich. (But still, the motto today is to take extra special care of myself.)
Potential Codes: C+, Rn+
Sentence: So komme ich wieder auf ein Fahrrad und mache mich etwas fitter. (So I get back on a bike and get a bit fitter.)
Potential Codes: C+, R+

Table 4: Example of an ambiguous sentence from the training data

Behaviour change language itself could also be a limiting factor for the success of the approach. Although this type of language has been shown to be rather stable across domains and target behaviours (Meyer and Elsweiler, 2023), the inter-rater reliability when labelling such data is often low compared to other annotation tasks even among trained professionals (Meyer and Elsweiler, 2022; Wu et al., 2022; Hershberger et al., 2021; Tanana et al., 2016; Pérez-Rosas et al., 2016). A task in which even human annotators with extensive training do not reach high consensus is likely to produce many samples that are highly contestable, or could even be correctly attributed to multiple classes (see Table 4). As such, relying on only one prediction per data point might never lead to excellent F1 scores, as they can be found in easier classification tasks.

Lastly, all test sets are annotated on a sentence to sentence basis, and no context is passed to the classifier. Especially in the case of chat-data, where some utterances might be replies to questions from the conversational partner, this way of labelling could lead to important information being missed by the classifier. This could additionally hinder robust classification and the potential of SSL-learning.

These results leave us pondering over the oft-debated issue of whether the current emphasis on SOTA-chasing (Church and Kordoni, 2022) is indispensable or advantageous for the effective deployment of algorithms in practical settings. In some domains, especially those with a high number of debatable labels, it might be preferable to accept mid-range classification performance. In our future work, we plan to explore to what extent the current effectiveness of the models is sufficient for practical applications.

7. Limitations

We did not add extensive amounts of data, and the size of each dataset used as a source for weakly labelled data was smaller than the size of the original dataset. We consider this as one of the main limitations of this work and intend to approach this problem in future work by adding large quantities of weakly labelled data from various sources. Our experiments so far have suggested that the source of pseudo-labelled data does not have a significant impact on classification performance. Consequently, we intend to explore the possibility of using web sources, such as relevant Reddit forums. These new sources could provide large amounts of textual data, although the noisy nature of these sources may necessitate a re-evaluation of our selection criteria, including the recalibration of confidence thresholds.

Another limitation was that some test sets, used in the final stage of experiments, were very small and in some cases included only few to no data points for the smaller classes. This could potentially have distorted our results and might have made the metrics more prone to outliers. Nonetheless, it is important to recognise that such imbalanced conditions may naturally occur when deploying these classifiers in real-life scenarios. In any case, we want to further explore the transfer capabilities of the solutions introduced here. For instance, by collecting and evaluating a larger dataset based on chat-like conversations around different target behaviours.

8. Conclusion

In this paper, we have attempted to shed light on the effectiveness of semi-supervised learning to increase both in-domain and transfer classification of written utterances concerning behaviour change. This is a low-resource classification task, where the learned classifiers can potentially be applied to data from various topics and across conversational contexts.

We found that adding pseudo-labelled data to the training sets had a stronger effect on the classification of smaller classes, whereas classification performance of the majority class remained fairly stable, regardless of the pre-filtering method or confidence thresholds. Observed effects were not stable across conditions, and adding larger amounts of data did not necessarily meant increased performance.

The transfer capabilities of the classifiers exhibited promising results in certain test scenarios. However, no consistent patterns or trends emerged when considering different label levels and target domains. Despite the lack of substantial performance enhancement through semi-supervised learning, there were also no noticeable deteriorations. This held true even when incorporating pseudo-labelled data from significantly distinct contexts, as evidenced by the AnnoMI collection. These findings highlight the robustness of the baseline classifier and its ability to effectively apply pseudo-labels to new data. Such outcomes could be attributed to the linguistic stability observed in the language pertaining to behaviour change across various contexts. These experiments underline the issue of unreliability of annotations in this domain hindering highly effective classification, leading us to question the need for high F1 scores in application areas like these.

9. Acknowledgements

The second and the fourth author thank: i) the financial support supplied by the Consellería de Cultura, Educación, Formación Profesional e Universidades (accreditation 2019-2022 ED431G-2019/04, ED431C 2022/19) and the European Regional Development Fund, which acknowledges the CiTIUS-Research Center in Intelligent Technologies of the University of Santiago de Compostela as a Research Center of the Galician University System, ii) the financial support supplied by project PID2022-137061OB-C22 (Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Proyectos de Generación de Conocimiento; supported by the European Regional Development Fund), iii) the financial support supplied by project PLEC2021-007662 (MCIN/AEI/10.13039/501100011033, Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Plan de Recuperación, Transformación y Resiliencia, Unión Europea-Next Generation EU).

The fourth author also thanks the financial support obtained from project SUBV23/00002 (Ministerio de Consumo, Subdirección General de Regulación del Juego).

10. Bibliographical References

Kenneth Ward Church and Valia Kordoni. 2022. Emerging trends: Sota-chasing. *Natural Language Engineering*, 28(2):249–269.

Dawn Clifford and Laura Curtis. 2016. *Motivational Interviewing in Nutrition and Fitness*. Guilford Publications.

José Marcio Duarte and Lilian Berton. 2023. A review of semi-supervised learning for text classification. *Artificial Intelligence Review*, pages 1–69.

Amelie V Güntner, Paul C Endrejat, and Simone Kauffeld. 2019. Guiding change: using motivational interviewing within organizations. *Gruppe. Interaktion. Organisation. Zeitschrift für Angewandte Organisationspsychologie (GIO)*, 50:129–139.

Lan-Zhe Guo and Yu-Feng Li. 2022. Class-imbalanced semi-supervised learning with adaptive thresholding. In *International Conference on Machine Learning*, pages 8082–8094. PMLR.

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*, pages 246–256, 1st virtual meeting. Association for Computational Linguistics.

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*, 12:613.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking Benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Taewan Kim, Seolyeong Bae, Hyun Ah Kim, Su-woo Lee, Hwajung Hong, Chanmo Yang, and Young-Ho Kim. 2023. Mindfuldiary: Harnessing large language model to support psychiatric patients' journaling. *arXiv preprint arXiv:2310.05231*.

Yu-Feng Li and Zhi-Hua Zhou. 2014. Towards making unlabeled data never hurt. *IEEE transactions on pattern analysis and machine intelligence*, 37(1):175–188.

Shayne Longpre, Yu Wang, and Chris DuBois. 2020. How Effective is Task-Agnostic Data Augmentation for Pretrained Transformers? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4401–4411.

Sarah Peregrine Lord, Doğan Can, Michael Yi, Rebeca Marin, Christopher W Dunn, Zac E Imel, Panayiotis Georgiou, Shrikanth Narayanan, Mark Steyvers, and David C Atkins. 2015. Advancing methods for reliably assessing motivational interviewing fidelity using the motivational interviewing skills code. *Journal of substance abuse treatment*, 49:50–57.

Molly Magill, Timothy R Apodaca, Brian Borsari, Jacques Gaume, Ariel Hoadley, Rebecca EF Gordon, J Scott Tonigan, and Theresa Moyers. 2018. A meta-analysis of motivational interviewing process: Technical, relational, and conditional process models of change. *Journal of consulting and clinical psychology*, 86(2):140.

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*, pages 2226–2235.

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 *Proceedings of the 5th International Conference on Conversational User Interfaces*, pages 1–13.

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*, pages 1–6.

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*.

William R Miller and Stephen Rollnick. 2002. *Motivational Interviewing, Second Edition: Preparing People for Change*. Applications of Motivational Interviewing Series. Guilford Publications.

Theresa B Moyers, Tim Martin, Paulette J Christopher, Jon M Houck, J Scott Tonigan, and Paul C Amrhein. 2007. Client language as a mediator of motivational interviewing efficacy: where is the evidence? *Alcoholism: clinical and experimental research*, 31:40s–47s.

Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. 2018. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31.

Kathryn M Page and Irina Tchernitskaia. 2014. Use of motivational interviewing to improve return-to-work and work-related outcomes: a review. *The Australian Journal of Rehabilitation Counselling*, 20(1):38–49.

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*, pages 42–51.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2020. Snorkel: Rapid training data creation with weak supervision. *The VLDB Journal*, 29(2-3):709–730.

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):474–480.

Viktor Schlegel, Erick Mendez-Guzman, and Riza Batista-Navarro. 2022. Towards Human-Centred Explainability Benchmarks For Text Classification. *arXiv preprint arXiv:2211.05452*.

Michael Tanana, Kevin A Hallgren, Zac E Imel, David C Atkins, and Vivek Srikumar. 2016. A Comparison of Natural Language Processing Methods for Automated Coding of Motivational Interviewing. *Journal of Substance Abuse Treatment*, 65:43–50.

Jesper E Van Engelen and Holger H Hoos. 2020. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440.

Paroma Varma and Christopher Ré. 2018. Snuba: Automating weak supervision to label training data. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 12, page 223. NIH Public Access.

Xudong Wang, Zhirong Wu, Long Lian, and Stella X Yu. 2022. Debiased learning from naturally imbalanced pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14647–14657.

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)*, pages 6177–6181. IEEE.

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