Focused Crawling and Model Evaluation in the field of Conversational Agents and Motivational Interviewing

Gergana Rosenova Tsakova
Tutors: David E. Losada Carril y Marcos Fernández Pichel

Master's Thesis, University of Santiago de Compostela
Master in Massive Data Analisys Tecnologies: Big Data

Abstract—**The exploitation of Motivational Interviewing concepts when analysing individuals' speech contributes to gaining valuable insights into their perspectives and attitudes towards behaviour change. The scarcity of labelled user data poses a persistent challenge and impedes technical advancements in research in non-English language scenarios. To address the limitations of manual data labelling, we propose a semi-supervised learning method as a means to augment an existing training corpus. Our approach leverages machine-translated user-generated data sourced from social media communities and employs self-training techniques for annotation. We conduct an evaluation of multiple classifiers trained on various augmented datasets. To that end, we consider diverse source contexts and employ different effectiveness metrics. The results indicate that this weak labelling approach does not yield significant improvements in the overall classification capabilities of the models. However, notable enhancements were observed for the minority classes. As part of future work, we propose to enlarge the datasets only with new examples from the minority classes. We conclude that several factors, including the quality of machine translation, can potentially bias the pseudo-labelling models. The imbalanced nature of the data and the impact of a strict pre-filtering threshold are other important aspects that need to be taken into account.**

1. Introduction

Health behaviour change is a difficult process that requires people to alter their habits and daily routine. Sustained motivation and determination are fundamental to achieving actual change. Motivational interviewing (MI) is a therapy approach that facilitates behaviour change by exploring the language changes that occur when an individual undergoes transformative experiences in their life [1]. The primary goal of this approach is to enhance individuals' self-awareness regarding their motivations for change and to strengthen their personal commitment towards achieving a goal.

Recent advances in Natural Language Processing (NLP) and Artificial Intelligence (AI) have led to the widespread availability and utilisation of Conversational Agents (CAs), systems with the capability to emulate human conversations using text or spoken language. Extensive evidence has demonstrated the potential advantages of utilising CAs for health-related purposes, supporting the process of change [2].

This work stems from a collaborative project between Universidad de Santiago de Compostela (USC) and Universität Regensburg (UR), which explores the utilisation of a CA that implements MI to enhance motivation and foster behaviour change. The CA employed in this project encounters significant language limitations as textual resources are particularly scarce for non-English languages. In the case of German language, annotated MI data are completely unavailable [2].

Scarcity of user-generated data poses a significant challenge with far-reaching implications in domain-specific NLP applications. The limited availability of labelled data hinders or slows down the development of robust NLP models, leading to potential limitations in their performance. In such scenarios, researchers often face limited relevant data, compelling them to resort to costly and time-consuming approaches to advance their studies. These approaches involve laborious collection and annotation of existing texts or opting for out-of-domain data that may not align with the task's objectives [3].

To mitigate the resource-intensive nature of traditional data collection methods, faster and more cost-effective alternatives are often explored to augment the training corpora. Two prominent alternatives are data augmentation, which involves applying a variety of transformations to the existing labelled data to create synthetic samples, and semi-supervised learning, an efficient solution that leverages unlabelled data, which is typically more abundant than labelled data [4]. This is achieved through techniques such as self-training, where the model, trained on a limited amount of labelled data, generates pseudo-labels for unlabelled data. These weakly labelled samples are used to further refine the model's predictions. This approach enables models to learn from a broader range of examples, enhancing generalisation and performance.

Semi-supervised learning tends to be effective when we have limited labelled data but have access to a large amount of unlabelled data [4]. We start from a small collection of user-generated text data in German. This original dataset was annotated with labels related to behaviour change. We aim to explore a viable way to augment this labelled dataset. Given the nature of the problem at hand, labelling subject's utterances, real user-generated data is more valuable than automatically generated samples or synthetic data (e.g., derived from modifying the original available dataset, using augmentation techniques [5]). The exponential growth of user-generated content published on Internet, coupled with the fact that the original training dataset is sourced from a peer-to-peer online forum [6], are factors in favour of exploring the potential of English-language online communities and forums as valuable sources of high-quality user data.

This work is therefore guided by the following research question: Are machine-translated user data, sourced from social media communities, and annotated via semi-supervised learning, a viable solution to augment an existing training corpus and to increase the base classifier's performance for cataloging behaviour change utterances?

To address this question, an existing human-labelled German-language dataset is used as baseline in this work. This dataset applies utterance codes defined in the Motivational Interviewing Skill Code (MISC) [7], which allows conceptualization of change-related speech through the assignment of valence, content labels and sublabels.

Data collected from online communities in English language are initially segmented into sentences. Next, a transformer model [8] is utilised to translate the segmented data into German. To ensure the relevance of data, a pre-filter classifier, which has been fine-tuned with on-topic and off-topic sentences, is used to identify sentences relevant to behaviour change[1]. Next, semi-supervised learning is employed to produce weak labels, assembling multiple new datasets from diverse behaviour change contexts. To explore the viability of the semi-supervised learning approach, a number of classification experiments are conducted. Classifiers are trained on the newly constructed datasets and their performance is evaluated on a held-out test set. In doing so, we obtain valuable insights about the potential benefits of this method to address data scarcity in this specific domain.

2. Related Work

This work falls within the scope of an ongoing collaborative project between the University of Santiago de Compostela (USC), Spain and the University of Regensburg (UR), Germany. In [6], the creators of the original collection, named GLoHBCD, provided a thorough explanation on the construction and evaluation of this behaviour change dataset. We aim at augmenting this dataset and its creators provided a detailed methodology for replicating the original experiments. This involves the creation of three classifiers across different behaviour change domains.

2.1. Motivational Interviewing and Behaviour Change

Motivational Interviewing (MI) is a client-centered approach used in Psychology and Healthcare to facilitate behaviour change. It aims to elicit motivation in individuals through goal-oriented communication to make changes in their lives. Traditionally, therapists employ various techniques such as open questions, affirmations, reflections, and summaries to guide clients towards change [1]. MI has found wide application in domains like substance abuse treatment, weight loss, and mental health interventions.

The increased adoption of voice assistants presents an opportunity for conversational agents to support health management. Our research focuses on assisting in the design of persuasive CAs by applying MI concepts and techniques [6]. Conversational agents employing MI techniques leverage natural language processing and machine learning algorithms to understand and respond to user's input, simulating human-like conversation. Recent developments in technology-assisted behaviour change incorporate MI annotation techniques. The exploitation of the Motivational Interviewing Skill Code helps to evaluate individuals' utterances and to measure the quality and fidelity of MI interventions [7].

Previous research have explored the creation of automated counseling systems in which clients interact with an embodied conversational agent that acts as a virtual counselor [9], the development of agent-based interventions to increase motivation and confidence to promote physical activity [10] and the design of specialized CAs to support parents' strategies tailored to healthy eating goals [11]. The evaluation results of these systems show promising results. For example, increased motivation was observed in surveyed individuals who had interactions with the CAs. However, the construction of CAs tailored to non-English speakers remains largely unexplored [6].

2.2. Semi-supervised Learning and Data Augmentation

Many studies have attempted to automatically augment corpora in various fields, including computer vision [12], audio augmentation [13] and speech recognition [14]. In these studies, automated data augmentation resulted in enhanced performance and more robust models, particularly in scenarios where limited data were available. Recent research suggests that this approach applied to language data could lead to substantial improvements in multiple classification tasks. Various surveys [15], [16] presented textual data augmentation methods such as synonym and embedding replacement, structure-based transformations, sentence replacement by round-trip translation, and so forth. Data augmentation in the context of behaviour change utterances was previously explored [5] by replacing and enhancing user data with synthetic data generated by ChatGPT. The performance of the resulting classifiers was tested on different combinations of synthetic and real user data.

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

Data augmentation focuses on enriching the original existing dataset by introducing synthetic variations. An alternative path to expand training sets consists of employing semi-supervised learning (SSL), which leverages both labelled and unlabelled data. SSL is concerned with situations where there is a scarcity of labelled data but an abundance of unlabelled data. SSL has emerged as a popular method for addressing data scarcity in deep learning contexts, with text data being a common domain of application. Among the various types of SSL, we focus on self-training, which is one of the pioneering SSL approaches and has demonstrated state-of-the-art performance in multiple tasks including neural machine translation [17].

The classic self-training methodology involves employing a pre-trained classifier to generate pseudo-labels for unlabelled data. These pseudo-labelled examples are then combined with the original corpus to create an augmented dataset, which is subsequently utilized to retrain a new model [18], [19]. In recent years, different alternatives have been explored to improve self-training with weak supervision [20], regularization [21], contrastive learning [22] and consistency learning [23].

Recent work on leveraging data augmentation and pseudo-labelling in the context of sleep-related issues showed promising results [24]. Similarly, [25] employed SSL to label new datasets with classifiers fine-tuned on the GLoHBCD and investigated the characteristics of written language about behaviour change.

3. Methods

The proposed methodology in this work involves the use of machine translation on data in English sourced from Reddit, the application of self-training to generate pseudo-labels and the re-training of the original models with the goal of improving the base classifiers.

3.1. Data extraction

With the increasing popularity of Internet, individuals often share online their experiences and challenges related to mental health and behaviour change. This public exposure is supported by various online platforms such as forums and blogs [26]. This represents an opportunity for researchers to explore and analyse abundant amounts of user-generated data, which can be exploited to feed machine and deep learning algorithms.

Reddit has gained an important role in scientific research due to its popularity among a large and diverse user base. Its communities, named subreddits, focus on specific interests and have a significant volume of publications. This allow researchers to target relevant scientific topics [27]. Previous studies have suggested that Reddit is a feasible source of data for creating domain-specific training datasets, particularly in the area of health. For example, Reddit data has been employed to detect signs of anxiety or depression from individuals' interactions [26], [27].

Reddit's publicly available API facilitates the retrieval and extraction of user-generated content [28]. In addition, Reddit is an appropriate source of data to expand the GLoHBCD corpus. In fact, there are some similarities between the platform used to construct GLoHBCD and Reddit, both being peer-to-peer communication forums where the conversational style is indirect among multiple parties.

Six different subreddits, topic-specific online communities within Reddit, were utilised for collecting data for our research. Each one of them is somehow related to behaviour change, but covers a specific topic. This allowed for the original corpus, that is centred on weight loss, to be expanded with samples of behaviour change focusing on other topics. The subreddits and topics are the following:

- *r/loseit* – healthy methods to lose weight and maintain progress.
- *r/smokingcessation* – encouragement to quit smoking and motivation for those who have already stopped.
- *r/leaves* – support for users trying to stop drug abuse.
- *r/stopdrinking* – motivation for controlling or stopping alcohol abuse.
- *r/selfimprovement* - inciting change in all aspects of individuals' life.
- *r/DecidingToBeBetter* – dedicated to self-improvement.

Using Reddit's API, we extracted a sample of thousand posts of the "top" category of each subreddit (top rated publications). According to the platform's voting system, posts in the "top" section of a subreddit are highly regarded by the community. We assume such posts are of higher quality and particularly useful for performing our experiments. The average length of a post was 292 words.

3.2. Data pre-processing

The availability of abundant user-generated data on the web has led to substantial progress in diverse NLP tasks. However, leveraging unstructured data is challenging and requires the usage of NLP tools to pre-process the datasets before they reach the training stage [26]. First, we need to segment the posts into sentences, as we work at sentence-level in this project. The average length of sentences in the corpus is 87 tokens per sentence. Next, suitable regular expression operations are applied to remove URLs, HTML tags, special symbols, and emojis from the textual data.

3.3. Machine Translation with Transformers

After the pre-processing stage, data need to go through a machine translation stage. Data scarcity is an omnipresent issue in many NLP applications, especially in non-English language projects. As Motivational Interviewing is relevant to individuals across the globe, it is crucial to design technological solutions for languages such as German, which is the target language for our project.

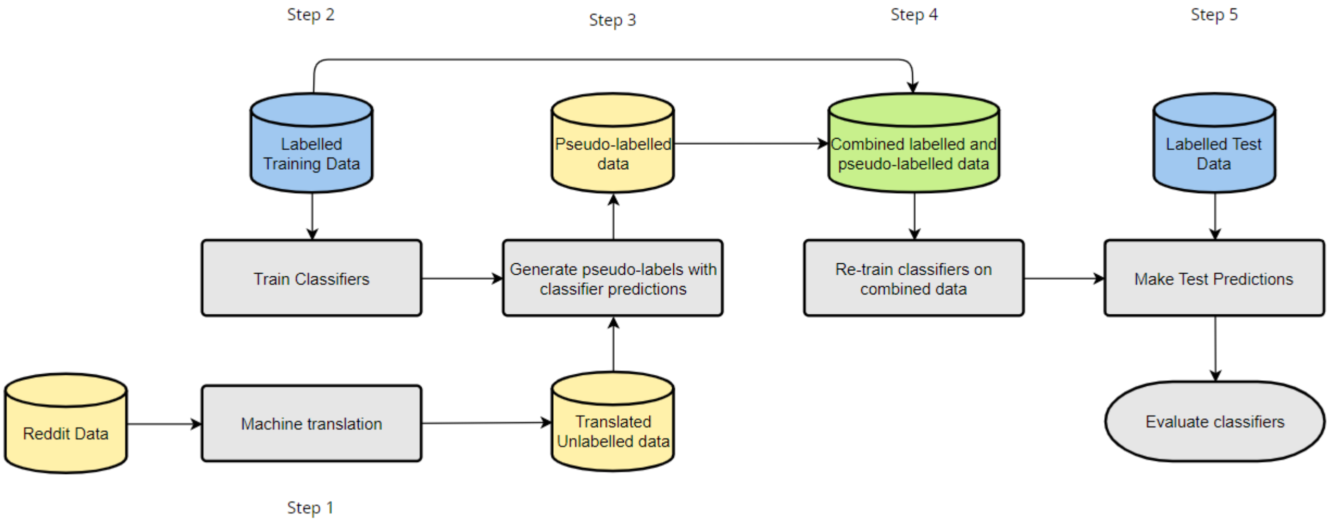

Figure 1. Proposed semi-supervised learning pipeline.

Previous studies suggest that data augmentation through machine translation is a promising technique. [29] generated new data by translating an annotated corpus from English to Urdu for fake news detection and [30] proceeded the same way between English and French for a reading comprehension task. Other recent approaches employed zero-shot multilingual MT techniques to improve end-to-end speech translation models [31] or machine translation in the context of speech recognition for telephone conversations [32].

The English-German translations represent a necessary step in the proposed SSL pipeline and produce data in a target language that faces data scarcity problems. However, the successful exploitation of data in widespread languages like English relies on the quality of the translation module. Thus, a primary issue we seek to answer in our experiments is whether or not the present quality of translations for the English-German language pair is good enough.

To carry out the MT an OPUS-MT pre-trained model developed by the Language Technology Research Group at the University of Helsinki was employed. The model is based on state-of-the art transformer-based neural machine translation (NMT) and trained on freely available parallel corpora collected in the OPUS repository [33]. The Reddit corpus was translated sentence by sentence and we made no manual post-edition (as it would be as costly as manual translation, which is a laborious effort that we are trying to avoid).

3.4. Pre-filtering

The semi-supervised approach we aim to apply requires in-domain unlabelled data. To promote the incorporation of relevant new data, we choose as sources a few subreddits related to behaviour change, translate the publications into German, apply a German BERT language model and, additionally, run a pre-filtering stage. This step classifies each in-domain German text as either related or not related to behaviour change. According to previous studies, the macro F1 effectiveness of this topic classifier is 72.67% [34]. This step enables the identification of on-topic sentences that could be used to infer information about users' change behaviour. A strict 0.99 confidence threshold on these relevance predictions is established. Examples that do not surpass the confidence threshold are discarded, reducing the data to instances classified as "Change Related" with very high confidence. The pre-filtering model was fine-tuned on a dataset related to weight loss, and it exhibits a tendency to identify a reduced number of topic-related sentences when applied to data from other domains. Table 1 reports the size of each subreddit collection before and after this relevance filtering.

TABLE 1. Size of subreddit datasets before and after pre-filtering.

subreddit	Size	change related	% change related
loseit	25226	6135	24
smokingcessation	6747	1332	20
leaves	12475	1884	15
stopdrinking	13628	1955	14
selfimprovement	16105	1848	11
DecidingToBeBetter	14669	2279	16

3.5. Base classifiers

To build the original base classifier, the first step consists of a fine-tuning step on the original dataset, GloHBCD. GloHBCD is an existing corpus composed of German-language texts abound behaviour change. The extracts are annotated with content categories and valences based on the MISC codes [7]. Each sentence represents a person's utterance around change, and it is annotated with one valence label ("+" for change talk and "-" for sustain talk). In addition to the valences, the sentences are assigned one of three possi-

TABLE 2. Test set performances (%) of baseline classifiers.

	Test set			
	Accuracy	Macro F1	Precision	Recall
Valence	75.97	69.79	71.53	68.84
Label	84.50	77.31	79.37	75.72
Sublabel	80.54	73.68	71.77	76.15

ble content labels: Reason (R) that encompasses the basis, incentives, justification or motives for change, Taking Steps (TS) representing specific steps that have been taken towards change and Commitment (C) that includes agreement, intention, or obligation regarding future behaviour. The instances in the Reason category additionally receive one of four additional sublabels, indicating the nature of the reason for change: general (R_) represents the examples with no sublabel, ability (Ra) encompasses ability and degree of difficulty of the change, desire (Rd) being desire or will and need (Rn) represents need or necessity. All these labels were assigned manually, following an annotation scheme driven by keywords [6].

GLoHBCD was separated into three training sets, each one corresponding to one level of classification: valence, content label or reason sublabel. Three separate base classifiers were developed by fine-tuning a pre-trained German BERT base model for each classification level. BERT, a deep bi-directional transformer, enables the construction of effective classification models in multiple text classification tasks, especially in the medical field [35].

The test sets were created by using a 80/20 random stratified split. The fine-tuning was performed using 10-fold cross validation across three epochs. The resulting models are then used to make predictions on the test sets.

The test set contains 929 examples for valence and label. Macro-averaged F1 is an appropriate measure because the datasets are imbalanced. This metric is more suitable for evaluating the models than Micro-F1, because it reflects the true model performance even when the classes are skewed. We also report classification accuracy and averaged recall and precision. Table 2 reports the performance of the three base classifiers.

In addition, we will sometimes report F1, precision, and recall for each class separately. These metrics are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{3}$$

where TP, FP, and FN represent true positives, false positives, and false negatives of each class, respectively.

TABLE 3. Distribution of labels after assigning pseudo-labels.

	Valence		Content Label			Reason Label			
subreddit	%-	%+	%R	%TS	%C	%general	%a	%d	%n
loseit	17	83	60	32	7	74	13	9	4
smokingcessation	27	73	70	23	7	68	15	13	4
leaves	25	75	75	18	7	70	16	11	4
stopdrinking	24	76	66	25	9	74	14	9	3
selfimprovement	25	75	69	22	9	67	19	9	5
DecidingToBeBetter	19	81	64	25	12	68	16	12	5

3.6. Pseudo-labelling

We aim to effectively augment the training sets for the three base classifiers, following the methodology used in [20] and [36]. The next step in the self-training approach is to use the base models to generate pseudo-labels for the unlabelled data. The fine-tuned base models are employed to annotate the translated in-domain sentences, obtaining weakly labelled instances. Pseudo-labelling requires multiple training sessions, however recent work suggests that the most efficient scenario is to conduct pseudo-labelling only 1-2 times [24], thus we chose to do only one iteration. Table 3 shows label distribution after applying pseudo-labelling.

3.7. Augmenting the training sets

To validate the proposed method, various experiments with different datasets were conducted. The main goal of these experiments was to evaluate to what extent SSL can contribute to enhance the performance of the original classifiers. To that end, the original performance of the base models was compared to the performance of each model after being re-trained with new data.

We combine the instances from the original training dataset, GLoHBCD, with the newly labelled Reddit data, working with each subreddit separately. Following [24], we establish three different confidence thresholds for the pseudo-label prediction: 0.5, 0.75 and 0.99. We only include new examples, with a positive or negative label, that are classified with a confidence score higher than the threshold. The examples classified with low confidence are therefore ignored. Note that with the 0.5 threshold all examples are incorporated (with either positive or negative pseudo-label).

The different subreddits complement the base training set in different ways by integrating diverse topics. We believe that such diversity will increase the generalisation abilities of the models and increase their performance. Besides testing the incorporation of instances from each individual subreddit, we also test a mixed configuration where new instances come from all subreddits. This leads to 21 different variants –(6 subreddits + all subreddits) * 3 confidence thresholds– applied on each base classifier. As the experiments are conducted for three classification tasks, this results in a total number of 63 new training sets.

Table 4 shows the distribution of pseudo-labels assigned by the base classifier to the data extracted from each subreddit.

TABLE 4. Overview of augmented datasets and label distributions.

		Valence			Label				Sublabel				
Training set	threshold	size	% -	% +	size	%R	%TS	%C	size	%R_	%Ra	%Rd	%Rn
GLoHBCD	-	3703	31	69	3696	65	25	10	2411	69	16	9	6
loseit	50	9838	22	78	9831	62	30	8	6120	72	14	9	5
	75	9683	22	78	9534	63	29	8	5924	73	14	9	5
	99	8788	20	80	8182	65	27	8	4903	78	11	8	4
smokingcessation	50	5035	30	70	5028	66	25	9	3345	68	16	10	6
	75	4979	29	71	4943	67	24	9	3286	69	15	10	6
	99	4738	28	72	4677	67	24	9	3018	71	14	10	6
leaves	50	5587	29	71	5580	68	23	9	3828	69	16	10	5
	75	5526	29	71	5499	69	23	8	3728	69	15	10	5
	99	5208	28	72	5178	70	22	8	3321	73	13	9	5
stopdrinking	50	5658	28	72	5651	65	25	9	3697	70	15	9	5
	75	5599	28	72	5557	66	25	9	3625	71	15	9	5
	99	5266	27	73	5186	66	24	9	3303	73	13	9	5
selfimprovement	50	5551	29	71	5544	66	24	9	3684	68	17	9	6
	75	5503	28	72	5428	67	24	9	3587	69	16	9	6
	99	5211	27	73	5104	68	23	9	3214	71	14	9	6
DecidingToBeBetter	50	5982	26	74	5975	65	25	10	3862	68	16	10	6
	75	5931	26	74	5858	65	25	10	3773	69	15	10	6
	99	5626	25	75	5429	66	24	10	3351	72	13	9	6
mixed	50	19136	23	77	19129	65	26	9	12481	70	15	10	5
	75	18706	22	78	18339	66	26	8	11868	72	14	10	4
	99	16322	19	81	15276	69	23	8	9055	79	9	8	3

The dataset sizes differ based on the subreddit used to augment the original corpus, with an average of 6010 examples for valence and label, nearly double the size of the GLoHBCD dataset. This results in approximately half the sets being composed of labelled data and the other half of unlabelled data. As per the creation of the sublabel datasets, all examples annotated as Reason (R), 69,5% of the data on average, were taken into consideration. The sublabel datasets are smaller, averaging 3865 examples, in contrast to the original sublabel dataset, which consists of 2411 examples. All these statistics do not take into account the mixed dataset, which is assembled by incorporating examples from all subreddits.

All datasets are imbalanced, mirroring the pattern observed in the original datasets. The distribution of valence, labels and sublabels across datasets is similar, which suggests that the way users address behaviour change in written language remains consistent regardless of the context. A higher threshold leads to a higher percentage of examples for the majority class.

3.8. Re-training

We perform classic self-training by incorporating the pseudo-labelled data without implementing any additional transformation. Under this approach, a BERT German-cased language model is fine-tuned from each of the newly constructed training sets. This training process is exactly the same as the original training done to build the base models. It follows a 10-fold cross validation over three epochs, we do not change the hyper-parameter settings either.

4. Experimental results

We conducted separate experiments on all label-levels, employing different base classifiers, and comparing the newly obtained SSL-based classifiers with the original classifiers. Additionally, we investigate how the confidence thresholds influence model's performance. Our findings reveal that the new models produce varying effects depending on the classification task considered. Specifically, we observed slight improvements in effectiveness for the valence and sublabel classifiers, while the label classifiers exhibit a decline in performance when pseudo-labelled data are introduced.

4.1. Valence

For the valence level of classification, we face a binary classification scenario. The model assigns a positive or negative label to the individuals' speech. The minority class is "sustain" (or negative) and the majority class is "change", which is represented by the positive label. The results are shown in Table 5. A modest overall improvement of approximately 1-2% is observed across datasets. Generally, classifiers trained on datasets with higher (stricter) thresholds yield better performance. Notably, there is a significant improvement in performance in the minority class. The subreddits about alcohol and drug abuse and general life changes (*stopdrinking*, *leaves*, *DecidingToBeBetter*) are the most promising, as we obtain here the highest performance results (datasets *DecidingToBeBetter 99* and *leaves 75*).

4.2. Label

In the case of the label classifier, we address a multi-class classification scenario with three classes: Reason (R)

TABLE 5. Results of valence classifier.

Valence

		Accuracy	F1-score			Precision			Recall		
Training set	thr		avg	-	+	avg	-	+	avg	-	+
GLoHBCD	-	75.97	69.79	56.13	83.46	71.53	62.83	80.23	68.84	50.71	86.96
loseit	50	75.54	69.32	55.51	83.13	70.94	61.84	80.03	68.42	50.36	86.49
	75	75.43	**70.21**	**57.73**	82.68	70.79	60.31	**81.26**	**69.76**	**55.36**	84.16
	99	**76.84**	**71.56**	**59.32**	**83.81**	**72.56**	**63.41**	**81.71**	**70.87**	**55.71**	86.02
smokingcessation	50	**77.38**	**71.67**	**58.94**	***84.39***	**73.40**	***65.50***	**81.29**	**70.65**	**53.57**	***87.73***
	75	75.54	**70.31**	**57.84**	82.77	70.92	60.55	**81.29**	**69.84**	**55.36**	84.32
	99	**77.16**	**71.80**	**59.50**	**84.10**	**73.01**	**64.32**	**81.70**	**71.00**	**55.36**	86.65
leaves	50	75.87	**70.06**	**56.87**	83.25	71.33	62.03	**80.64**	**69.26**	**52.50**	86.02
	75	***77.49***	***72.17***	***60.00***	**84.34**	***73.44***	**65.00**	***81.87***	***71.34***	**55.71**	86.96
	99	75.65	69.72	**56.31**	83.12	71.05	61.70	**80.41**	**68.91**	**51.79**	86.02
stopdrinking	50	**77.06**	**71.43**	**58.75**	**84.11**	**72.92**	**64.53**	**81.30**	**70.52**	**53.93**	**87.11**
	75	**76.19**	**70.83**	**58.33**	83.33	**71.73**	62.10	**81.36**	**70.20**	**55.00**	85.40
	99	**77.06**	**71.57**	**59.07**	**84.06**	**72.89**	**64.29**	**81.49**	**70.72**	**54.64**	86.80
selfimprovement	50	75.76	69.74	**56.25**	83.23	71.21	62.07	**80.35**	**68.88**	**51.43**	86.34
	75	75.76	**70.82**	**58.82**	82.82	71.21	60.61	**81.82**	**70.50**	***57.14***	83.85
	99	**76.52**	**70.58**	**57.37**	**83.79**	**72.24**	**63.76**	**80.72**	**69.63**	**52.14**	**87.11**
DecidingToBeBetter	50	75.76	**69.81**	**56.42**	83.21	71.20	61.97	**80.43**	**68.98**	**51.79**	86.18
	75	75.65	69.72	**56.31**	83.12	71.05	61.70	**80.41**	**68.91**	**51.79**	86.02
	99	***77.49***	***72.17***	***60.00***	**84.34**	***73.44***	**65.00**	***81.87***	***71.34***	**55.71**	86.96
mixed	50	**76.19**	**70.42**	**57.36**	**83.48**	**71.76**	62.71	**80.81**	**69.60**	**52.86**	86.34
	75	75.87	**70.54**	**58.00**	83.07	71.32	61.35	**81.28**	**69.97**	**55.00**	84.94
	99	75.87	**70.27**	**57.36**	83.17	71.32	61.73	**80.91**	**69.57**	**53.57**	85.56

TABLE 6. Results of label classifier.

Label

		Accuracy	F1-score				Precision				Recall			
Training set	thr		avg	R	TS	C	avg	R	TS	C	avg	R	TS	C
GLoHBCD	-	84.50	77.31	90.21	75.45	66.28	79.37	87.27	82.18	68.67	75.72	93.36	69.75	64.04
loseit	50	83.42	76.68	89.16	73.80	**67.07**	79.02	86.91	76.82	**73.33**	74.78	91.53	**71.01**	61.80
	75	83.96	77.20	89.16	**75.77**	**66.67**	***80.55***	86.31	79.63	***75.71***	74.67	92.19	**72.27**	59.55
	99	83.96	**76.28**	89.53	76.27	63.03	78.68	86.88	80.75	68.42	74.35	92.36	**72.27**	58.43
smokingcessation	50	83.42	76.33	89.07	74.44	65.48	78.57	86.29	79.81	**69.62**	74.52	92.03	69.75	61.80
	75	82.02	73.87	88.67	71.49	61.45	75.81	86.44	74.77	66.23	72.27	91.03	68.49	57.30
	99	83.42	76.86	89.00	73.76	**67.84**	78.93	86.16	79.90	**70.73**	75.23	92.03	68.49	***65.17***
leaves	50	84.39	75.96	***90.37***	**76.17**	61.35	78.68	**87.42**	81.04	67.57	73.85	***93.52***	**71.85**	56.18
	75	82.88	74.45	89.34	73.41	60.61	76.57	86.95	76.96	65.79	72.74	91.86	**70.17**	56.18
	99	83.75	76.51	89.46	74.21	65.85	**79.52**	86.15	80.39	**72.00**	74.20	93.02	68.91	60.67
stopdrinking	50	83.32	76.05	89.14	75.05	63.95	77.16	**87.64**	77.58	66.27	75.06	90.70	**72.69**	61.80
	75	84.39	***77.91***	89.85	**75.59**	**68.29**	**80.12**	***87.92***	77.78	**74.67**	**76.10**	91.86	**73.53**	62.92
	99	83.21	75.37	89.23	73.47	63.41	78.33	85.87	79.80	**69.33**	73.12	92.86	68.07	58.43
selfimprovement	50	83.75	75.82	89.55	74.94	62.96	79.00	86.18	80.98	**69.86**	73.41	93.19	69.75	57.30
	75	83.10	75.65	88.96	74.25	63.75	78.23	86.98	75.88	**71.83**	73.67	91.03	**72.69**	57.30
	99	83.42	76.57	88.94	74.11	**66.67**	**79.64**	85.91	79.05	**73.97**	74.20	92.19	69.75	60.67
DecidingToBeBetter	50	83.10	75.64	89.21	73.48	64.24	77.68	87.16	76.13	**69.74**	73.97	91.36	**71.01**	59.55
	75	83.96	77.00	89.85	73.26	**67.90**	**80.23**	86.59	78.74	**75.34**	74.55	93.36	68.49	61.80
	99	83.53	75.91	89.55	73.94	64.24	78.39	86.76	78.67	**69.74**	73.94	92.52	69.75	59.55
mixed	50	84.39	77.16	89.59	***77.29***	64.60	**79.93**	87.13	80.45	**72.22**	75.00	92.19	***74.37***	58.43
	75	82.78	74.50	89.18	73.50	60.82	76.14	86.79	78.20	63.41	73.15	91.69	69.33	58.43
	99	82.99	75.24	88.60	75.38	61.73	77.69	86.30	78.28	68.49	73.30	91.03	72.69	56.18

as the majority class, and Taking steps (TS) and Commitment (C) as the minority classes. The results are shown in Table 6. Contrary to the observations made with the valence classifier, we do not observe here an overall improvement in performance. Only with the *stopdrinking* dataset and a threshold of 0.75, we found a marginal improvement of less than 1%, which we consider insignificant. On average, these classifiers perform one percentage point lower than the baseline classifier. There appears to be an increase in the effectiveness metrics of the minority classes, but this improvement comes at the cost of a decline in the majority class. The most substantial improvement is observed in the Commitment class, which exhibits increases in precision and recall in the range of 5%-7%. Different thresholds do not significantly impact on the results. The highest figures are achieved when using data from subreddits *leaves* and *stopdrinking* with a threshold 0.75.

4.3. Sublabel

The task of sublabel classification consists of multilabel classification with four labels, see Table 7. The sublabel classifiers exhibit behaviour that is similar to the one achieved by the label classifiers. The resulting Macro F1 varies by approximately ±1%, although we do not claim statistical significance for this difference. Notably, it is observed an improvement in F1 and recall of the majority class

TABLE 7. Results of sublabel classifier.

		Label														
		Acc.	F1-score					Precision					Recall			
Training set	thr		avg	R_	Ra	Rd	Rn	avg	R_	Ra	Rd	Rn	avg	R_	Ra	Rd	Rn
GLoHBCD	-	80.54	73.68	86.60	61.29	75.59	71.23	71.77	88.58	60.64	67.61	70.27	76.15	84.71	61.96	85.71	72.22
loseit	50	78.19	71.31	84.75	56.38	73.85	70.27	68.97	87.37	55.21	64.86	68.42	74.46	82.28	57.61	85.71	72.22
	75	79.03	71.70	85.68	55.06	75.00	71.05	69.58	87.19	56.98	66.67	67.50	74.55	84.22	53.26	85.71	**75.00**
	99	79.36	71.91	85.89	54.14	**80.00**	67.61	70.74	86.85	55.06	**72.46**	68.57	73.54	**84.95**	53.26	**89.29**	66.67
smokingcessation	50	80.54	**73.88**	**86.63**	58.38	**78.05**	**72.46**	**73.26**	87.59	58.06	**71.64**	**75.76**	74.88	**85.68**	58.70	85.71	69.44
	75	80.37	72.60	**86.76**	59.09	75.20	69.33	71.08	87.62	***61.90***	**68.12**	66.67	74.65	**85.92**	56.52	83.93	72.22
	99	80.20	73.40	86.21	58.24	**78.74**	70.42	**72.06**	87.50	58.89	**70.42**	**71.43**	75.32	**84.95**	57.61	**89.29**	69.44
leaves	50	80.03	**74.71**	85.68	58.51	**78.69**	**75.95**	**71.94**	87.98	57.29	**72.73**	69.77	***78.08***	83.50	59.78	85.71	***83.33***
	75	***81.21***	**74.73**	**87.06**	57.47	**78.33**	***76.06***	***74.58***	86.75	**60.98**	**73.44**	***77.14***	75.16	**87.38**	54.35	83.93	**75.00**
	99	80.54	**73.73**	**86.73**	57.14	**80.00**	71.05	**72.02**	87.81	57.78	**75.00**	67.50	75.73	**85.68**	56.52	85.71	**75.00**
stopdrinking	50	**80.87**	**74.24**	**86.90**	57.78	**79.03**	**73.24**	**73.27**	87.65	59.09	**72.06**	**74.29**	75.60	**86.17**	56.52	**87.50**	72.22
	75	***81.21***	***74.92***	**86.94**	58.43	**80.33**	**73.97**	**73.79**	87.47	60.47	**74.24**	**72.97**	**76.36**	**86.41**	56.52	**87.50**	**75.00**
	99	80.20	73.03	86.48	55.37	**80.00**	70.27	**71.97**	86.80	57.65	**75.00**	68.42	74.34	**86.17**	53.26	85.71	72.22
selfimprovement	50	80.54	73.58	**86.86**	54.55	**79.67**	**73.24**	**72.91**	87.07	57.14	**73.13**	**74.29**	74.64	**86.65**	52.17	**87.50**	72.22
	75	79.03	72.01	85.19	56.67	**77.78**	68.42	69.91	86.68	57.95	**70.00**	65.00	74.72	83.74	55.43	**87.50**	72.22
	99	79.70	72.77	86.21	55.03	**81.36**	68.49	71.52	87.50	53.61	**77.42**	67.57	74.16	**84.95**	56.52	85.71	69.44
DecidingToBeBetter	50	80.03	73.13	85.96	58.24	***81.67***	66.67	71.15	87.25	58.89	**76.56**	61.90	75.51	84.71	57.61	**87.50**	72.22
	75	80.03	73.27	86.00	58.76	**77.27**	71.05	70.91	87.85	**61.18**	67.11	67.50	**76.70**	84.22	56.52	***91.07***	**75.00**
	99	**81.04**	73.29	***87.29***	54.76	**80.67**	70.42	**73.60**	86.26	60.53	**76.19**	**71.43**	73.38	***88.35***	50.00	85.71	69.44
mixed	50	79.36	72.09	85.78	52.94	**78.74**	70.89	69.97	86.63	57.69	**70.42**	65.12	75.23	**84.95**	48.91	**89.29**	**77.78**
	75	79.03	71.93	85.57	53.93	77.17	71.05	69.82	86.97	55.81	69.01	67.50	74.72	84.22	52.17	87.50	75.00
	99	78.69	72.46	84.97	53.97	80.65	70.27	70.39	87.02	52.58	73.53	68.42	74.99	83.01	55.43	89.29	72.22

(R_). One of the minority classes, Rd, gets improvements in F1, precision, and recall across multiple datasets. However, these improvements comes at the expense of a decline in the remaining classes. The best performing classifiers are those fine-tuned on datasets from the *leaves*, *stopsmoking* and *DecidingToBeBetter* subreddits. Once again, the results indicate that the confidence threshold of the pseudo-labels is not crucial when it comes to the performance on the test set.

5. Discussion and future work

After conducting a series of experiments over machine-translated and pseudo-labelled datasets, the obtained results suggest that the proposed approach of combining MT and SSL does yield moderate improvements in some specific instances. It was observed a general trend of improvement of the minority class across three different classification tasks. Although certain classes showed improvements in individual metrics, the general predictive power of the SSL classifiers remained within the range of the performance achieved by the baseline models. Various confidence thresholds for incorporating pseudo-labels were explored, under the hypothesis that higher thresholds would result in better performance. However, we observed a trend where the average F1 score either remained the same or slightly decreased with a higher threshold. Furthermore, the best results did not often correspond with the highest, more stringent, thresholds. These findings suggest that the confidence level of the pseudo-label does not have a strong influence on the predictive capability of the SSL model.

Another important aspect is the training set size. In general, the new datasets are comparable to the original ones, except from the mixed datasets that incorporate pseudo-examples from all subreddits. However, the mixed classifiers did not yield better results, despite having been fine-tuned with larger datasets. This result implies that the size of the dataset does not play a significant role in the classification performance.

The topic of the subreddit used to obtain the augmented training set appears to play a crucial role in the classification improvement. The highest scores were obtained from datasets of subreddits focused on alcohol abuse, drug usage and general life improvements. These subreddits deal with aspects that are not related to weight-loss but it seems that they supply complementary utterances about life change that are effective as additional signs for the classifiers. Additionally, it is worth noting that the original dataset is derived from a forum where participants were aiming to lose weight or were in the process of doing so. The subreddits mentioned before include individuals who are in the phase of attempting to maintain the changes they have achieved, thus placing them in a different stage of behaviour change.

Fine-tuning language models using SSL is challenging due to the presence of noisy labels. Traditional self-training mechanisms overlook the base model's weakness during the pseudo-labelling process [37]. We argue that the effectiveness of the approach is affected by the robustness of the initial models. In our case, the base classifiers were trained with limited datasets, with 3700 samples only. Related to this, we observed the scarcity of pseudo-labelled data imputed to the minority classes. This produces additional data imbalance in the re-training phase and results in poor performance.

Another limitation that must be taken into consideration is the quality of machine translation. Previous research has highlighted the potential inaccuracies of machine translation [29], which may contribute to lowered performance of the BERT models on the test data. Some inconsistencies can be found in the translated data, for example, the translation of "fast" in English to "schnell" (instead of the appropriate translation "fasten" in the given context). Moreover, MT could alter the natural word order, resulting in uncommon combinations of words. This could be an issue for accurately labelling user utterances towards change. In the future, we will further analyse the MT module, compare multiple solutions and, possibly, incorporate pseudo-labelled data from multiple language sources.

In general, we observed promising improvements in metrics for the minority classes. A potential SSL approach that could be considered in future work consists of including only training pseudo-labelled samples for the minority labels. This could help to mitigate data imbalance and obtain higher overall results. In addition, as performance depends on the subreddit's topics, data from other (non-health) change related topics could be sourced.

Another future line of research could be oriented to lowering the threshold for the pre-filtering classifier. Such an approach would feed more examples to the subsequent modules and it could help to acquire a larger number of examples.

6. Conclusion

Limited labelled data is a common issue in many machine learning applications. This work has addressed this problem in the context of an ongoing project centred around a Conversational Agent employing Motivation Interviewing techniques. Our proposed approach aims to mitigate the scarcity of MI annotated data in German language. To that end, we created a semi-supervised learning pipeline, consisting of data scraping, machine translation and self-training to reduce the costly and labor-intensive process of labelling data manually. The objective was to develop robust classification models for annotating users' change-related speech according to the MISC code. Oriented to these goals, our study consisted of a series of experiments with pre-trained transformer-based models and three different classification tasks.

We hypothesised that English language data from the platform Reddit from diverse topic could be translated into German and then used to augment a pre-existing German language corpus and, as a consequence, enhancing the predictive accuracy of behaviour change utterance detection. However, the experimental results demonstrated that this approach did not yield significant improvements in the classification capabilities, compared to the baseline models. Several factors might contribute to these findings, including inaccuracies in the automated translation, potential biases of the pseudo-labelling models due to imbalanced training datasets, the presence of noisy labels, and the established strict pre-filtering threshold. Our study illustrates the type of challenges encountered in text classification with non-English languages and underscores the need for further research in addressing these limitations.

References

[1] W. Miller and S. Rollnick, "Motivational interviewing: Preparing people for change, 2nd ed." *Journal For Healthcare Quality*, vol. 25, p. 46, 05 2003.

[2] L. Laranjo, A. G. Dunn, H. L. Tong, A. B. Kocaballi, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A. Y. S. Lau, and E. Coiera, "Conversational agents in healthcare: a systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 9, pp. 1248–1258, 07 2018.

[3] N. Varshney, S. Mishra, and C. Baral, "Interviewer-candidate role play: Towards developing real-world nlp systems," 2021.

[4] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2021.

[5] S. Meyer, D. Elsweiler, B. Ludwig, M. Fernández-Pichel, and D. Losada, "Do we still need human assessors? prompt-based gpt-3 user simulation in conversational ai," 07 2022.

[6] S. Meyer and D. Elsweiler, "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*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 2226–2235.

[7] J. Hettema, J. Steele, and W. R. Miller, "Motivational interviewing," *Annual Review of Clinical Psychology*, vol. 1, no. 1, pp. 91–111, 2005, pMID: 17716083.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[9] D. Schulman, T. Bickmore, and C. Sidner, "An intelligent conversational agent for promoting long-term health behavior change using motivational interviewing." 01 2011.

[10] S. Olafsson, T. O'Leary, and T. Bickmore, "Coerced change-talk with conversational agents promotes confidence in behavior change," in *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, ser. PervasiveHealth'19. New York, NY, USA: Association for Computing Machinery, 2019, p. 31–40.

[11] D. Smriti, J. Y. Shin, M. Mujib, M. Colosimo, T.-S. Kao, J. Williams, and J. Huh-Yoo, "Tamica: Tailorable autonomous motivational interviewing conversational agent," in *Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare*, ser. PervasiveHealth '20. New York, NY, USA: Association for Computing Machinery, 2021, p. 411–414.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, p. 84–90, may 2017.

[13] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Interspeech 2015*, 2015, pp. 3586–3589.

[14] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 5582–5586.

[15] M. Bayer, M.-A. Kaufhold, and C. Reuter, "A survey on data augmentation for text classification," *ACM Computing Surveys*, vol. 55, no. 7, pp. 1–39, dec 2022.

[16] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow, "A survey on recent approaches for natural language processing in low-resource scenarios," 2021.

[17] J. He, J. Gu, J. Shen, and M. Ranzato, "Revisiting self-training for neural sequence generation," 2020.

[18] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning*. The MIT Press, 2006.

[19] D.-H. Lee, “Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks,” 2013.

[20] G. Karamanolakis, S. Mukherjee, G. Zheng, and A. H. Awadallah, “Self-training with weak supervision,” 2021.

[21] C. Wei, K. Shen, Y. Chen, and T. Ma, “Theoretical analysis of self-training with deep networks on unlabeled data,” 2022.

[22] Y. Yu, S. Zuo, H. Jiang, W. Ren, T. Zhao, and C. Zhang, “Fine-tuning pre-trained language model with weak supervision: A contrastive-regularized self-training approach,” 2021.

[23] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le, “Unsupervised data augmentation for consistency training,” 2020.

[24] H. Shim, S. Luca, D. Lowet, and B. Vanrumste, “Data augmentation and semi-supervised learning for deep neural networks-based text classifier,” in *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, ser. SAC ’20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1119–1126.

[25] S. Meyer and D. Elsweiler, “Towards cross-content conversational agents for behaviour change: Investigating domain independence and the role of lexical features in written language around change,” 05 2023.

[26] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, “Detection of depression-related posts in reddit social media forum,” *IEEE Access*, vol. 7, pp. 44 883–44 893, 2019.

[27] J. H. Shen and F. Rudzicz, “Detecting anxiety through Reddit,” in *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*. Vancouver, BC: Association for Computational Linguistics, Aug. 2017, pp. 58–65.

[28] A. N. Medvedev, R. Lambiotte, and J.-C. Delvenne, “The anatomy of reddit: An overview of academic research,” in *Dynamics On and Of Complex Networks III*, F. Ghanbarnejad, R. Saha Roy, F. Karimi, J.-C. Delvenne, and B. Mitra, Eds. Cham: Springer International Publishing, 2019, pp. 183–204.

[29] M. Amjad, G. Sidorov, and A. Zhila, “Data augmentation using machine translation for fake news detection in the Urdu language,” in *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 2537–2542.

[30] A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le, “Qanet: Combining local convolution with global self-attention for reading comprehension,” 2018.

[31] T. A. Dinh, D. Liu, and J. Niehues, “Tackling data scarcity in speech translation using zero-shot multilingual machine translation techniques,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6222–6226.

[32] G. Huang, A. Gorin, J.-L. Gauvain, and L. Lamel, “Machine translation based data augmentation for cantonese keyword spotting,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6020–6024.

[33] J. Tiedemann and S. Thottingal, “OPUS-MT – building open translation services for the world,” in *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. Lisboa, Portugal: European Association for Machine Translation, Nov. 2020, pp. 479–480.

[34] “Gbert finetuned on on-topic vs. off-topic sentences of the GLoHBCD dataset,” $https://huggingface.co/selmey/behaviour_change_prefilter_german$.

[35] M. V. Koroteev, “Bert: A review of applications in natural language processing and understanding,” 2021.

[36] J. Du, E. Grave, B. Gunel, V. Chaudhary, O. Celebi, M. Auli, V. Stoyanov, and A. Conneau, “Self-training improves pre-training for natural language understanding,” 2020.

[37] S. Mukherjee and A. Awadallah, “Uncertainty-aware self-training for few-shot text classification,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 21 199–21 212.