Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3637–3649
Marseille, 20-25 June 2022

Cross-lingual Approaches for the Detection of Adverse Drug Reactions in German from a Patient's Perspective

Lisa Raithel[1,2,3], **Philippe Thomas**[1], **Roland Roller**[1], **Oliver Sapina**[1],
Sebastian Möller[1,2], **Pierre Zweigenbaum**[3]
[1]Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI) Berlin, 10559 Berlin, Germany
[2]Technische Universität Berlin, 10623 Berlin, Germany
[3]Université Paris-Saclay, CNRS, Laboratoire interdisciplinaire des sciences du numérique (LISN),
91405 Orsay, France
lisa.raithel@dfki.de

Abstract

In this work, we present the first corpus for German Adverse Drug Reaction (ADR) detection in patient-generated content. The data consists of 4,169 binary annotated documents from a German patient forum, where users talk about health issues and get advice from medical doctors. As is common in social media data in this domain, the class labels of the corpus are very imbalanced. This and a high topic imbalance make it a very challenging dataset, since often, the same symptom can have several causes and is not always related to a medication intake. We aim to encourage further multi-lingual efforts in the domain of ADR detection and provide preliminary experiments for binary classification using different methods of zero- and few-shot learning based on a multi-lingual model. When fine-tuning XLM-RoBERTa first on English patient forum data and then on the new German data, we achieve an F1-score of 37.52 for the positive class. We make the dataset and models publicly available for the community.

Keywords: pharmacovigilance, text classification, adverse drug reactions

1. Introduction

Adverse drug reactions (ADRs) are a major and increasing public health problem and exist all over the world, mostly under-reported (Alatawi and Hansen, 2017). They describe an unanticipated, negative reaction to a medication. Of course, new drugs are tested extensively when being developed, but certain vulnerable groups, like pregnant or elderly people, are rarely part of clinical trials and still might be in need of medication at some point. Furthermore, clinical trials and also physicians prescribing medications cannot cover every potential use case. Therefore, an efficient monitoring of ADRs from different angles is important and necessary for improving public health and safety in medication intake.

One of those angles is the classification of documents containing mentions of ADRs. While there exist various resources for training NLP models on the task of ADR classification, for example scientific publications or drug leaflets, the world wide web provides a more up-to-date and diverse source of information, especially social media and patient fora (Sarker and Gonzalez, 2015). Here, patients freely write about their experiences during medication therapy. By using their own non-expert language, e.g. using laymen terms for describing their situation, they create a rich data source for pharmacovigilance from a patient's perspective in a wide variety of languages. Therefore, there is theoretically a huge amount of text to leverage. On the other side, however, this text is noisy (e.g. spelling mistakes), incomplete (e.g. deleted messages) and the use of laymen terms and abbreviations (e.g. "AD" as a generic name for any anti-depressant) complicates the processing of these data (Seiffe et al., 2020; Basaldella et al., 2020). It is furthermore difficult to collect these resources (e.g. finding informative keywords to search Twitter) and researchers are soon confronted with privacy issues when handling social media data. Thus, only a small amount of (annotated) data is publicly available for research and unfortunately, most of these data are in English.

Another issue arises when looking at the label distributions of the described resources. Depending on where the dataset originates from and how it was collected, the distribution of labels for text classification is skewed in either direction: for instance, data from the CADEC corpus (Karimi et al., 2015) tend to include a lot more positive documents (that is, documents containing adverse drug reactions) than negative ones, while for example HLP-ADE (Magge et al., 2021), a corpus of tweets, contains 13 times more negative than positive examples.

Thus, before working on the more specific task of extracting adverse effects and their corresponding drugs, and also the relations between them, there needs to be a reliable way to distinguish texts mentioning ADRs from those without ADRs. This first step is still necessary even in the era of deep learning (Magge et al., 2021). Therefore, we manually annotated a dataset of German health-related user posts and employed it for zero- and few-shot experiments using a Language Model (LM) equipped with a binary classification head. With this, we want to investigate whether it is possible to use an LM fine-tuned on one language (the source language) to classify documents from another language (the target language) by using only a small number of available documents in the ADR domain. We do this in a “true” few-shot scenario (Kann et al., 2019; Perez et al., 2021), assuming that we do not only have a small number of shots for the fine-tuning training set, but also only a small number of shots for the development and test set. We focus our efforts on finding documents containing ADRs, i.e. the positive class (label 1), since these are the most interesting to us. Our contributions are as follows:

Dataset We provide a binary annotated corpus of German documents containing adverse drug reactions. The dataset is challenging, since its class and topic distributions are imbalanced, and the texts are written in everyday language.

Few-Shot Classification We experiment with different model settings and data combinations to transfer knowledge between English and German. In several few-shot settings, we come close to the performance of fine-tuning on the full German (training) data while achieving a higher recall, making the few-shot models a potentially better filter for unlabeled data.

Error Analysis We conduct an error analysis on the predicted results to provide future directions of research for handling imbalanced, small and user-generated data.

Models, code and dataset can be found on github[1].

2. Related Work

Before social media became popular, the detection and extraction of ADRs was mostly conducted on electronic health records (EHRs) and clinical reports (Uzuner et al., 2011; Lo et al., 2013; Sarker and Gonzalez, 2015). Nowadays, however, not all drug effects are reported to healthcare professionals, but are also widely discussed online. This, and the rise of deep learning, spurred the collection of datasets (mostly in English) and the introduction of shared tasks and challenges, such as the SMM4H series (Weissenbacher et al., 2018; Weissenbacher et al., 2019; Klein et al., 2020). The methods of choice for tackling these tasks often included rule-based approaches and ensembles of statistical classifiers, e.g. Support Vector Machines based on static word embeddings (Sarker and Gonzalez, 2015; Nikfarjam et al., 2015). Now, the majority of approaches uses deep neural nets (Minard et al., 2018; Wunnava et al., 2019) and with the introduction of transformer models (Vaswani et al., 2017), especially BERT (Devlin et al., 2019) and all its variants are taking over.
For example, Chen et al. (2019) use BERT combined with a knowledge-based approach to achieve first place in the SMM4H 2019 task with an F1-score of 62.89 for ADR classification. In SMM4H 2020 Task 2, the best system employed a RoBERTa model and achieved an F1-score of 64.0 for the positive class (Wang et al., 2020).
Also, joint learning approaches are getting more attention. `DeepADEMiner` (Magge et al., 2021), for instance, is a pipeline for classifying, extracting and normalizing ADRs from Twitter data in one go. They also publish a naturally distributed dataset of tweets (7% positive documents) on which they achieve an F1-score of 63.0 for the task of document classification using a RoBERTa model (Liu et al., 2019).
Raval et al. (2021), on the other hand, employ a T5 model (Raffel et al., 2020) to jointly learn ADR classification and extraction of ADRs, drugs and drug dosages as a sequence-to-sequence problem. Their proposed method applies the original T5 task balancing strategies, but adds dataset balancing to account for different dataset sizes, domains and label distributions. On CADEC (Karimi et al., 2015) and the SMM4H 2018 tasks 1 and 2 (Weissenbacher et al., 2018), the authors achieve an F1-score for the positive class of 98.7, 69.4 and 89.4 respectively (for ADR detection). Finally, the authors also apply their model on the very imbalanced French SMM4H 2020 dataset (Klein et al., 2020) in a zero-shot fashion and achieve an F1-score of 20.0.
The latter is one of the few approaches to tackle ADR data that is not in English. Just recently, however, we can see an increased effort to publish non-English data for the detection of ADRs which we will describe in the next section.

3. Datasets

We first describe the new German dataset we provide and then shortly summarize other available, non-German datasets.

3.1. The *Lifeline* corpus

For populating our corpus, we choose the German *Lifeline* forum[2], a forum where people write about health issues, but also about other topics. Users can only participate in the discussion if they are registered, but all questions and answers are freely accessible without registration. The forum is anonymous as well. With permission of the forum administrators, we downloaded the existing user threads[3] and gave the accordingly prepared documents to one of our annotators, a final year student in pharmacy with some practical experience in the handling of medications.
Note that one document corresponds to a complete forum post and thus (usually) contains more than one sentence. This, of course, comes with a caveat: some-

[1] `https://github.com/DFKI-NLP/cross-ling-adr`. The dataset will only be accessible via a data protection agreement.

[2] `https://fragen.lifeline.de/forum/`

[3] Data downloaded in July 2021.

German	English Translation	Label
Ach ja, vielleicht für einige noch interessant. Hatte in meiner Verzweiflung ja auch ein AD ausprobiert. Bei mir hat es den TSH voll hochgetrieben! Davon hatte der Psychiater noch nie gehört...echt verrückt. Geholfen hat es übrigens überhaupt gar nicht. Konnte es zum Glück problemlos ausschleichen.	By the way, this might be interesting for some. In my despair, I had also tried an AD. It drove my TSH all the way up! The psychiatrist had never heard of this...really crazy. By the way, it didn't help at all. Fortunately, I was able to phase it out without any problems.	1
Hallo. Man muss sich nur mal eine Zigarettenschachtel ansehen. Die Warnhinweise sind nicht umsonst drauf. Ich bin seit Ewigkeiten Raucherin mit 5 jähriger Unterbrechung, aber ich Depp habe wieder angefangen. Es ging mir 100 Mal besser ohne Glimmstengel. Seit ich in den WJ bin, wird mir nach der ersten Kippe am Tag schummelig. Das sagt wohl alles. Ich verteufel meine Sucht. Lg	Hi. You only have to look at a packet of cigarettes. The warnings are not there for nothing. I've been a smoker for ages with a 5 year break, but I started again. I felt a 100 times better without the cigarettes. Since I've been in MP, I get woozy after the first smoke of the day. I guess that says it all. I'm demonizing my addiction. Cheers	0

Table 1: A positive (top) and a negative example (bottom) from the *Lifeline* corpus. A positive document is one containing an adverse effect of a drug. Medication and adverse effect for the positive sample are color-coded. Note the use of (very general) abbreviations (AD = anti-depressant, TSH = Thyroid-stimulating hormone, MP = meno pause) and the descriptions in colloquial speech.

topic	train/dev	test
women's health	2541	634
cosmetic OPs	166	47
skin	129	36
bones	125	29
gen. med.	117	20
heart	92	26
nerves	44	11
men's health	22	3
sports	21	6
infections	21	5
nutrition	19	9
int. organs	15	3
allergies	8	2
life	7	2
gastroint. system	7	2
avg #tokens	110.6	109.9
avg #sentences	8.3	8.2

Table 2: The distribution of topics for train/dev and test set in the *Lifeline* corpus. Bottom: the average number of tokens and sentences per document.

times, documents contain mentions of adverse reactions to one drug, but also positive reactions to another drug, leaving the final class label unclear. However, to facilitate annotation, we kept the guidelines very simple: each document containing at least one adverse reaction was to be labeled as positive (Class 1), all others were to be labeled as negative (Class 0). The only exceptions were documents that contained speculations in which people only knew about side effects from hearsay or rumors and did not experience them themselves. Those samples were flagged by the annotator and removed for future investigation. After one round of training and discussing the annotations, our annotator labeled 4,169 documents (flagged ones already excluded) using the annotation tool Prodigy[4]. The average progress was 100 documents in about one hour.

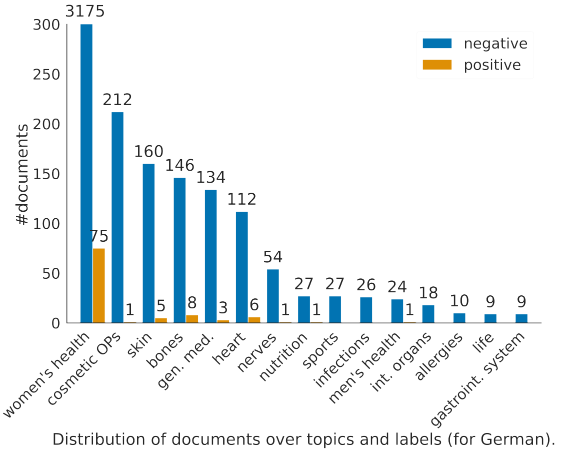

Figure 1: The distribution of topics in the new German corpus. The blue bar represents the negative documents, the orange bar the positive ones. Note the huge difference in the number of documents per topic especially in the *women's health* thread.

The resulting *Lifeline* dataset for binary classification of ADRs contains 101 positive and 4,068 negative examples (positive to negative ratio $\sim 1:40$). In Table 1 we show one positive and one negative document as an example. Figure 1 shows the distribution of topics over the entire dataset. Note the huge amount of documents in *women's health* (3,175 documents) compared to the other topics. In Table 2, we show the distribution of topics divided into train/dev and test set. The average number of tokens per document is approximately 110.6 for the train/dev set and 109.9 for the test set.

[4] https://prodi.gy/docs

lang.	overall	neg	pos	ratio	type	annotation	authors
es	400	235	165	1.4 : 2	forum	entities	(Segura-Bedmar et al., 2014)
fr	3033	2984	49	61 : 1	Twitter	binary	(Klein et al., 2020)
ru	370	-	-	-	drug reviews	multi-label	(Alimova et al., 2017)
ru	500	-	-	-	drug reviews	multi-label + entities	(Tutubalina et al., 2021)
ru	9515	8683	842	10 : 1	Twitter	binary	(Klein et al., 2020)
ja+en	169	-	-	-	forum	entities	(Arase et al., 2020)
de	4169	4068	101	40 : 1	forum	binary	ours

Table 3: Other non-English social media corpora for the detection (and partially extraction) of ADRs. en=English, fr=French, ru=Russian, ja=Japanese, de=German. Note the label imbalance in all datasets.

Further, the documents contain about 8.3 sentences in the train/dev set and 8.2 sentences in the test set.

3.2. Other available data

Apart from the already mentioned corpora, other (English) datasets might contain for instance tweets (Magge et al., 2021), case reports (Yada et al., 2020) or the content of PubMed abstracts (Gurulingappa et al., 2012). For our experiments, we combine the datasets CADEC (Karimi et al., 2015) and PsyTAR (Zolnoori et al., 2019), since both corpora are based on a patient forum[5]. We used only those two, because we assumed data from another domain, e.g. Twitter or more structured sources, might complicate the transfer between the English and the German data. Together, they comprise 2,137 documents, with 1,683 positive and 454 negative examples.

Finally, we show in Table 3 the non-English social media datasets that are (at least partially[6]) publicly available, including our new German corpus. Three of those comprise data from patient fora (Spanish, Japanese, German), two are created from Twitter messages (French, Russian) and two are collected drug reviews (Russian). Note the strong label imbalance for all corpora, and the reverse labels distribution for the Spanish data (and also for the English data we use). Moreover, the corpora vary strongly in their overall sizes and length of documents: for example, Twitter messages are usually rather short, while forum posts might contain several hundred tokens. As we can see in Figure 2 comparing the German and the English dataset we use for our experiments, the length of a user post also varies with its content: documents containing ADRs tend to be longer than those without ADRs.

Since English and German are typologically closer than English and the other languages listed in Table 3, we focus the experiments on the transfer of the English knowledge (*source language data*) to the German data (*target language data*). The target language data contains the German forum posts as described above. Note the reverse class label distribution for both datasets.

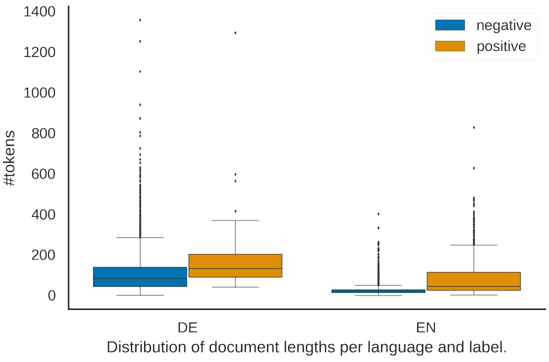

Figure 2: The number of tokens per label and per corpus (not divided by train/dev and test). "DE" corresponds to the *Lifeline* corpus, while "EN" corresponds to the combination of CADEC and PsyTAR. For both corpora, the documents containing no ADRs tend to be shorter.

4. Experiments

In the following, we describe the experimental setting and models we used. Details for the specific fine-tuning (hyper-) parameters can be found in the appendix. Our goal is to classify target language documents into those containing ADRs (the positive Class 1) and those that do not contain any ADR (the negative Class 0).

As baseline, we use a Support Vector Machine (SVM, (Boser et al., 1992)) and for the neural network approach, we chose a transformer model (Vaswani et al., 2017), since these are capable of incorporating contextual knowledge to a certain degree. Since we only have a small number of positive samples in the target data, we first fine-tune a model on the source data and then try to transfer the learned knowledge via (i) zero-shot "learning", (ii) a second fine-tuning on *all* the target language training data (henceforth called *full model*), and (iii) few-shot learning. For (iii), we experiment with different "modes": per-class few-shot learning (per_class), adding negative examples to the shots (add_neg), and adding negative as well as source language examples to the shots (add_source). The modes are explained in more detail in Section 4.3.

[5] www.askapatient.com

[6] For the French and Russian Twitter datasets, the test sets are unfortunately not available, they were part of the SMM4H 2020 shared task.

4.1. Pre-processing

Pre-processing is handled in a very simple way: Before feeding the documents to the model, we replace URLs, user names, dates, and similar occurrences with generic names, e.g. <URL>, using ekphrasis (Baziotis et al., 2017). For the baseline models, the documents are then tokenized simply by white space and for the transformer models, tokenization is done by the wordpiece tokenizer (Wu et al., 2016) of the respective model. Documents with less than four tokens are filtered out, those that are longer than 300 tokens are truncated. This setup achieved the best results during preliminary experiments. Both datasets are divided into a training/development (train/dev, 80%) and a test set (20%) via a stratified split corresponding to the distribution of labels.

4.2. Baselines

We train a Support Vector Machine on the target language data using fasttext embeddings (Bojanowski et al., 2017) and sklearn (Pedregosa et al., 2011). The embeddings for one document are calculated as the average over the word vectors. The models are trained with class weights ("balanced") and otherwise default parameters, and tested on the same (German) test set on which the transformer models are tested. For comparison with the neural approach, we use the available shot data, add negatives to the shots and also add negatives plus source language data to the shots, using aligned embeddings (Joulin et al., 2018). Further, we train an SVM on the entire target training data (full SVM).

4.3. Two-Stage Fine-Tuning

A perfect transformer model to start with would be a multi-lingual model trained on *health-related*, *user-generated* texts. However, since this rare combination does not exist (yet), we explore XLM-RoBERTa (Conneau et al., 2020), a multi-lingual model trained on general domain texts (henceforth XLM-R), and BioRedditBERT (Basaldella and Collier, 2019), a biomedical (English) model fine-tuned on user posts from Reddit[7] (henceforth BRB).

First, for both model types, we conduct a hyper-parameter search to find the best parameters for our task on the source language data with respect to macro F1-score on the development set. For this, we use the Weights & Biases sweeps framework (Biewald, 2020). Using the determined hyper-parameters and ten different seeds for model initialisation, we fine-tune (fine-tuning 1) ten source language models on the *source data* to account for the instability of language models (Devlin et al., 2019): XLM-R$_1$ – XLM-R$_{10}$ (the same for the BRB model).

We hypothesise that using all target language data will bias the model towards the negative class (the negative to positive ratio is 40 : 1) and want to counter-act this by testing several few-shot scenarios. Here, however, we want to apply a "true" few-shot setting (Perez et al., 2021) and thus do not use an extensive development set to optimize the models on. Thus, the dev set has always the exact same number of examples and classes as the train set. We evaluate the following few-shot scenarios:

per_class We use the exact same number of documents for both labels in train and dev set. For example, if we use *ten shots*, we have five positive and five negative examples in both the train and dev set.

add_neg We refer with the term "few-shot" only to the positive examples and add a certain amount of negative samples to the train and dev set. Using the example of *ten shots* again, we construct a train set of ten positive and {100, 200, 300, 400} negative examples. The same goes for the dev set. This approach serves to approximate the "natural" label distribution of the target data.

add_source We again refer only to the positive examples when using the term "few-shot", add {100, 200, 300, 400} negative examples and *additionally* {100, 200, 300, 400} random samples from the source language for both train and dev set. We assume that this approach might help to reduce the *catastrophic forgetting* (McCloskey and Cohen, 1989, i.a.) of language models.

Note that we only use shots of 10 and 40 to reduce the amount of experiments and since there are only 101 positive samples in the train/dev set of the target language in total, we can only experiment with up to 40 shots. The corresponding test set then contains 21 positive examples. We then proceed as follows:

1. We choose five seeds for sampling from the *target train/dev set*, creating five different train/dev sets to sample the shots from.
2. We freeze all *source language* models except for their classifier and fine-tune (fine-tuning 2) on the five sampled sets with the few-shot approaches described above. This results in {XLM-R$_{fine1}$, ..., XLM-R$_{fine10}$} for every seed and every scenario (again, the same goes for the BRB models).
3. Each model in each scenario is applied to the fixed target language test set; the final prediction is decided by majority voting of all ten models.
4. Finally, the performance per scenario is averaged over the five seeds.

The experimental setup is visualized in Figure 3 in the Appendix. To compare the few-shot scenarios, we also fine-tune the ten XLM-R models (XLM-R$_1$, ..., XLM-R$_{10}$) on *all* available target language training data, called the *full model* (XLM-R$_{full}$), and finally also apply the *not* fine-tuned XLM-R and BRB models in a zero-shot fashion to the target language.

[7] https://www.reddit.com/

Finally, we also try boosting the performance via rule-based post-processing, since we observed that many false positives discuss health issues that are similar to adverse reactions. After calculating the voting winners, we use an extensive medication list[8] and a self-created shorter list related to women's health topics (the biggest topic in the dataset) and abbreviations to check each document's predicted label. If it is positive but *does not* include a drug name from the medication list, we switch the label to negative. Conversely, if the label is positive and *does* include a word from the women's health list, we switch the label to negative as well. Both checks are performed independently and we calculate the final scores for each approach.

5. Results

We now present the results, ordered by language and fine-tuning approach.

5.1. Source Language (English)

The results for the first round of fine-tuning, i.e. fine-tuning XLM-R and BRB on the source language data, can be found in Table 5 and 6 in the Appendix. For both models, we can see a clear tendency to perform better for the majority class—in this case it is the positive one. XLM-R achieves an average F1-score of 91.03 (±0.67), while for the negative class, the average F1 is 65.20 (±2.81).

5.2. Target Language (German)

The results on the target language can be found in Table 4: The second block (rows 6 and 7) shows the zero-shot approach, the third block (rows 8 - 13) shows the results of XLM-R and the last block (rows 14 - 17) shows the performance of BRB. Note that we only show the best performance(s) with respect to F1-score of the positive class for every setting and leave out classifiers that did not predict any positives.

Baseline The SVM baseline achieves a performance of 17.39 F1 for the positive class when we train the model on *all available* German training data. This performance is comparable to the *few-shot* performance of the transformer models. When using shot data, the F1-score for the positive class decreases by at least 4 points, depending on the applied few-shot method. However, the SVM model combined with the per_class strategy and 40 shots achieves the highest recall for the positive class overall. The results of the baselines can be found in the first block of Table 4.

Zero-Shot The zero-shot approach mainly stands out because of the second highest recall (95.23) and third-best AUC score for the positive class achieved by XLM-R (see the second block in Table 4). In contrast to that, the zero-shot approach using BRB is much more biased towards the negative class.

Full fine-tuning Against our assumptions, a second fine-tuning of the already fine-tuned XLM-R$_{full}$ with the *full target training dataset* achieves the highest F1-score for the positive class overall (37.52 ±6.65) (Table 4, third block, first row). It also achieves the highest precision for the positive class, as well as the highest recall for the negative class. Note, however, the rather low AUC score (64.0 ±3.68) and the high standard deviations for both precision and recall of Class 1. Also, the recall for Class 1 (28.57 ±7.53) is amongst the lowest for all experiments.

per_class In the per-class scenario, we can see that the 40-shot approach achieves a slightly higher F1-score for the positive class for both models. This is somewhat surprising, we would have expected a higher difference in this scenario because of the additional positive samples during fine-tuning. Note the high AUC for the 40-shot setting of XLM-R. Otherwise, this approach ranges in the lower area with respect to F1-score for Class 1.

add_neg Adding negatives during fine-tuning does not help in the 10-shot setting: here, for both models, we do not find any positives (and therefore, the results are not presented in the table). For 40 shots, we find that adding 100 random negative samples works best. Here, XLM-R clearly wins over BRB.

add_source Adding negative target *and* random source language data achieves (with 40 shots) the second best results after the full training. For XLM-R this works best for 10 shots when adding 100 negative and 200 source language examples, while the 40-shot scenario is improved by 300 negative and 300 source language examples. Moreover, this setting achieves the best AUC score overall (78.95 ±2.67 for 10 shots), even compared to the full-data model, and has a lower standard deviation than when only adding negatives. BRB does not find any positive examples with only 10 shots, but when adding 100 negative and 200 source language examples, it comes close to XLM-R's performance. Also, the described post-processing can improve the F1-scores for the positive class: XLM-R increases its F1-score from 15.03 (±1.26) to 21.22 (±1.03) for 10 shots and from 22.55 (±3.42) to 28.56 (±2.29) for 40 shots; BRB increases its performance from 22.23 (±4.06) to 25.33 (±4.29).

5.3. Error Analysis

We now provide an error analysis of the best performing model XLM-R$_{full}$ with respect to the falsely predicted documents. Out of the 824 documents (21 positives, 803 negatives), the best model predicted 8/21 positives and 796/803 negatives correctly, leaving 20 documents predicted incorrectly.

For the 13 false negatives, we can find no clear indication why those were missed except for one ex-

[8] 22,827 medication names copy-pasted from a German information website about health-related topics (https://www.apotheken-umschau.de/medikamente/arzneimittellisten/medikamente_a.html)

	method	data	P_0	R_0	F1_0	P_1	R_1	F1_1	P_m	R_m	F1_m	AUC
SVM	full	all	98.73	86.92	92.5	10.26	57.14	17.39	54.49	72.03	54.92	72.03
SVM	per_class	10	99.34 ± 1.01	35.52 ± 20.03	49.48 ± 23.39	3.39 ± 0.65	85.71 ± 24.28	6.51 ± 1.25	51.36 ± 0.7	60.62 ± 7.64	27.99 ± 11.88	60.62 ± 7.64
SVM	per_class	40	**99.90** ± **0.22**	22.24 ± 4.73	36.17 ± 6.67	3.23 ± 0.17	**99.05** ± **2.13**	6.26 ± 0.31	51.57 ± 0.14	60.64 ± 2.23	21.22 ± 3.48	60.64 ± 2.23
SVM	add_neg	10 + 200 neg	98.27 ± 0.17	87.25 ± 3.33	92.4 ± 1.77	7.94 ± 1.16	40.95 ± 7.97	13.16 ± 1.3	53.11 ± 0.56	64.1 ± 2.64	52.78 ± 1.38	64.1 ± 2.64
SVM	add_neg	40 + 400 neg	98.98 ± 0.33	71.63 ± 2.82	83.08 ± 1.8	6.16 ± 0.28	71.43 ± 10.1	11.33 ± 0.59	52.57 ± 0.3	71.53 ± 3.68	47.21 ± 0.68	71.53 ± 3.68
BRB	Zero-shot	0	97.55	99.00	98.27	11.11	4.76	6.67	54.33	51.88	52.47	51.88
XLM-R	Zero-shot	0	99.77	54.42	70.42	5.18	95.23	9.82	52.48	74.83	40.13	74.83
XLM-R	full	all	98.16 ± 0.19	**99.43** ± **0.23**	**98.79** ± **0.07**	**57.64** ± **7.14**	28.57 ± 7.53	**37.52** ± **6.65**	**77.9** ± **3.55**	64.00 ± 3.68	**68.15** ± **3.33**	64.00 ± 3.68
XLM-R	per_class	10	99.15 ± 0.91	66.45 ± 9.91	79.21 ± 6.27	5.24 ± 1.25	75.24 ± 31.66	9.75 ± 2.55	52.2 ± 1.08	70.84 ± 10.88	44.48 ± 1.9	70.84 ± 10.88
XLM-R	per_class	40	99.71 ± 0.13	61.34 ± 5.95	75.82 ± 4.69	6.04 ± 0.87	93.33 ± 2.61	11.33 ± 1.55	52.87 ± 0.48	77.34 ± 3.65	43.58 ± 3.11	77.34 ± 3.65
XLM-R	add_neg	40 + 100 neg	98.22 ± 0.73	94.5 ± 8.62	96.12 ± 4.46	26.37 ± 15.67	32.38 ± 30.38	19.81 ± 12.47	62.29 ± 7.67	63.44 ± 11.33	57.97 ± 6.94	63.44 ± 11.33
XLM-R	add_source	10 + 100 neg + 200 source	99.39 ± 0.27	75.99 ± 4.5	86.07 ± 2.84	8.29 ± 0.8	81.9 ± 8.52	15.03 ± 1.26	53.84 ± 0.36	**78.95** ± **2.67**	50.55 ± 1.99	**78.95** ± **2.67**
XLM-R	add_source	40 + 300 neg + 300 source	98.72 ± 0.43	90.91 ± 4.82	94.59 ± 2.4	15.84 ± 6.53	54.29 ± 18.01	22.55 ± 3.42	57.28 ± 3.1	72.6 ± 6.7	58.57 ± 2.8	72.6 ± 6.7
BRB	per_class	10	98.03 ± 0.12	75.54 ± 5.29	85.25 ± 3.35	4.38 ± 0.7	41.9 ± 5.22	7.91 ± 1.12	51.21 ± 0.39	58.72 ± 1.98	46.58 ± 2.14	58.72 ± 1.98
BRB	per_class	40	99.14 ± 0.23	56.59 ± 8.68	71.72 ± 7.39	4.74 ± 0.62	80.95 ± 6.73	8.94 ± 1.11	51.94 ± 0.35	68.77 ± 3.35	40.33 ± 4.2	68.77 ± 3.35
BRB	add_neg	40 + 100 neg	97.67 ± 0.2	99.00 ± 1.01	98.33 ± 0.4	21.91 ± 20.74	9.52 ± 8.69	11.00 ± 8.6	59.79 ± 10.38	54.26 ± 3.86	54.66 ± 4.12	54.26 ± 3.86
BRB	add_source	40 + 100 neg + 200 source	97.94 ± 0.11	98.23 ± 0.48	98.08 ± 0.23	24.21 ± 5.6	20.95 ± 4.26	22.23 ± 4.06	61.07 ± 2.83	59.59 ± 2.06	60.16 ± 2.08	59.59 ± 2.06

Table 4: Target language (German): results of the best runs for every scenario. We excluded those that scored an F1 of 0.0 for the positive class. BRB = BioRedditBERT, XLM-R = XLM-RoBERTa. **_0** and **_1** represent the negative and positive class, respectively. **P** is precision, **R** is recall and **F1** is F1-score, and **_m** indicates the macro scores.

ample where the document was cut off before the user was talking about side effects. We notice some spelling mistakes and unclear formulations but nothing "human-unreadable". Some adverse reactions are mentioned implicitly, though, or only very briefly. On the other side, some of the false negatives are very clearly describing the problems, even mentioning the word "side effects", thus it is not obvious to us why the document was classified as negative. One document describes the reactions partially in a positive light (weight gain), this sentiment might also mislead a model that is biased towards more negative sentiments with respect to adverse drug effects.

Regarding the 7 false positives, we find several examples of persons talking about side effects they experienced *before* taking the new drug the post is about. Also, we find posts describing health issues that can be easily confused with side effects, and also one occurrence where the reactions came from *not* taking the drug.

Against our expectations, we cannot see problems in the predictions with respect to the topic distribution, i.e. there is no clear bias of the models towards performing better on documents from the women's health topic. For most documents it is not clear why they were not correctly classified. Here, a larger test set would probably help in the analysis.

6. Discussion

We find that, unsurprisingly, label imbalance (more positives than negatives in the first round of fine-tuning, vice-versa in the second round of fine-tuning) has the strongest influence on performance. This is evident in both the baselines and the neural approaches. It is interesting, however, that the models seem to perform better when having *more, but imbalanced* data than when having carefully balanced data, as is the case for the per-class setting. This might also be interrelated with the small number of samples overall.

Reminding the models of the original data (in our case the source data) gave the best results apart from the full data model, confirming again the phenomenon of catastrophic forgetting. Adding source examples to the full data model training might therefore also improve the performance in the full data scenario.

Compared to almost all other settings, the full data model has a very low recall for the positive class (28.57 ±7.53) as well as one of the lowest AUC scores. This is both interesting and unfortunate, since it probably cannot be used as a filter for subsequent tasks, such as ADR span identification or ADR-drug relation extraction.

Further, we conclude that the multi-lingual aspect of XLM-R seems to be of more importance than the user-data content of BRB, since in most cases XLM-R outperforms BRB. However, it is still interesting to see that even BRB, pre-trained purely on English data, can produce some results on the German data, coming close to the performance of XLM-R in the add_source setting.

We can also see a very high instability in the models, coming from the Language Models themselves but also from the sampled data (influenced by the seeds). Thus, it might help to carefully select the samples we use for few-shot training in case we have enough to choose from. However, we do not know if a model prefers the same examples as we humans do to learn better.

Further, one of the bigger issues with the presented models and dataset is that the model is often not able to distinguish side effects induced by drugs from accompanying effects of menopause (or other health-related issues). We tried to counteract this by applying the described post-processing, and it helped, at least for the add_source scenario, but not enough. Admittedly, this might be improved by a better medication list: our list contains mostly the original drug names, and those are often not used by patients.

Of course, since German is very close to English, it remains to be seen whether the presented scenarios are also applicable to more distant languages. There might also be some cultural aspects in the handling of adverse effects in online user forums for the different languages.

7. Conclusion

We have presented the first German corpus for the detection of adverse drug reactions from a patient's perspective, that is, a corpus created from user posts in a health-related online forum. Further, we described a series of experiments to find the documents containing adverse drug reactions using a multi-lingual approach and comparing it with few-shot scenarios. We experimented with a user-centred model and a more general multi-lingual model and found that classification performance benefits from the multi-lingual aspect.

The classifiers with a high recall can still be used as a filter for improving the downstream performance for ADR recognition and ADR-drug relation extraction, even though their F1-scores are not the best ones. Furthermore, it might be interesting to try character-based models, like e.g. CharacterBERT (El Boukkouri et al., 2020). Although this is not a multi-lingual model it might perform better on rare words like medication names, and it might handle spelling mistakes better than wordpiece-based models. Including negation of ADRs (Scaboro et al., 2021) and other corpora, e.g. the TLC corpus (Seiffe et al., 2020), to disambiguate user terms using a mapping from technical to laymen terms and vice versa might also be beneficial for the performance.

8. Acknowledgements

This research was funded by the Agence Nationale de la Recherche (ANR, French National Research Agency) – 20-IADJ-0005-01 and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 442445488 under the trilateral ANR-DFG-JST call for project KEEPHA. We would also like to thank the *Lifeline* administrators for allowing us to use their data, explosion.ai for providing us with a researchers licence for Prodigy, and the anonymous reviewers for their detailed and helpful feedback.

9. Bibliographical References

Alatawi, Y. M. and Hansen, R. A. (2017). Empirical estimation of under-reporting in the U.S. Food and Drug Administration Adverse Event Reporting System (FAERS). *Expert Opinion on Drug Safety*, 16(7):761–767, July.

Basaldella, M. and Collier, N. (2019). BioReddit: Word embeddings for user-generated biomedical NLP. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 34–38, Hong Kong. Association for Computational Linguistics.

Baziotis, C., Pelekis, N., and Doulkeridis, C. (2017). Datastories at SemEval-2017 Task 4: Deep

LSTM with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada. Association for Computational Linguistics.

Biewald, L. (2020). Experiment tracking with weights and biases.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory.*, pages 144–152.

Chen, S., Huang, Y., Huang, X., Qin, H., Yan, J., and Tang, B. (2019). HITSZ-ICRC: A report for SMM4H Shared Task 2019-automatic classification and extraction of adverse effect mentions in tweets. In *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 47–51, Florence, Italy. Association for Computational Linguistics.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

El Boukkouri, H., Ferret, O., Lavergne, T., Noji, H., Zweigenbaum, P., and Tsujii, J. (2020). CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6903–6915, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H., and Grave, E. (2018). Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

Kann, K., Cho, K., and Bowman, S. R. (2019). Towards realistic practices in low-resource natural language processing: The development set. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3340–3347, Hong Kong, China. Association for Computational Linguistics.

Klein, A., Alimova, I., Flores, I., Magge, A., Miftahutdinov, Z., Minard, A.-L., O'Connor, K., Sarker, A., Tutubalina, E., Weissenbacher, D., and Gonzalez-Hernandez, G. (2020). Overview of the fifth social media mining for health applications (#SMM4H) shared tasks at COLING 2020. *Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 27–36.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692 [cs]*, July.

Lo, H. Z., Ding, W., and Nazeri, Z. (2013). Mining adverse drug reactions from electronic health records. In *2013 IEEE 13th International Conference on Data Mining Workshops*, pages 1137–1140, December.

Magge, A., Tutubalina, E., Miftahutdinov, Z., Alimova, I., Dirkson, A., Verberne, S., Weissenbacher, D., and Gonzalez-Hernandez, G. (2021). DeepADEMiner: A deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter. *Journal of the American Medical Informatics Association*, 28(10):2184–2192, September.

McCloskey, M. and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press.

Minard, A.-L., Raymond, C., and Claveau, V. (2018). IRISA at SMM4H 2018: Neural network and bagging for tweet classification. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 50–51, Brussels, Belgium. Association for Computational Linguistics.

Nikfarjam, A., Sarker, A., O'Connor, K., Ginn, R., and Gonzalez, G. (2015). Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681, March.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Perez, E., Kiela, D., and Cho, K. (2021). True few-shot learning with language models. *35th Conference on Neural Information Processing Systems*, page 17.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Raval, S., Sedghamiz, H., Santus, E., Alhanai, T., Ghassemi, M., and Chersoni, E. (2021). Exploring a unified sequence-to-sequence transformer for medical product safety monitoring in social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3534–3546, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Sarker, A. and Gonzalez, G. (2015). Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics*, 53:196–207, February.

Scaboro, S., Portelli, B., Chersoni, E., Santus, E., and Serra, G. (2021). NADE: A benchmark for robust adverse drug events extraction in face of negations. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 230–237, Online, November. Association for Computational Linguistics.

Uzuner, Ö., South, B. R., Shen, S., and DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 18(5):552–556.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *31st Conference on Neural Information Processing Systems (NIPS 2017*, page 11.

Wang, C.-K., Dai, H.-J., Zhang, Y.-C., Xu, B.-C., Wang, B.-H., Xu, Y.-N., Chen, P.-H., and Lee, C.-H. (2020). ISLab system for SMM4H shared task 2020. *Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 42–45.

Weissenbacher, D., Sarker, A., Paul, M. J., and Gonzalez-Hernandez, G. (2018). Overview of the third social media mining for health (SMM4H) shared tasks at EMNLP 2018. In *Proceedings of the 2018 EMNLP workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 13–16, Brussels, Belgium. Association for Computational Linguistics.

Weissenbacher, D., Sarker, A., Magge, A., Daughton, A., O'Connor, K., Paul, M. J., and Gonzalez-Hernandez, G. (2019). Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019. In *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 21–30, Florence, Italy. Association for Computational Linguistics.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv:1609.08144 [cs]*, October.

Wunnava, S., Qin, X., Kakar, T., Sen, C., Rundensteiner, E. A., and Kong, X. (2019). Adverse drug event detection from electronic health records using hierarchical recurrent neural networks with dual-level embedding. *Drug Safety*, 42(1):113–122, January.

10. Language Resource References

Alimova, I., Tutubalina, E., Alferova, J., and Gafiyatullina, G. (2017). A machine learning approach to classification of drug reviews in Russian. In *2017 Ivannikov ISPRAS Open Conference (ISPRAS)*, pages 64–69, November.

Arase, Y., Kajiwara, T., and Chu, C. (2020). Annotation of adverse drug reactions in patients' Weblogs. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 6769–6776.

Basaldella, M., Liu, F., Shareghi, E., and Collier, N. (2020). COMETA: A corpus for medical entity linking in the social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3122–3137, Online, November. Association for Computational Linguistics.

Gurulingappa, H., Rajput, A. M., Roberts, A., Fluck, J., Hofmann-Apitius, M., and Toldo, L. (2012). Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45(5):885–892, October.

Karimi, S., Metke-Jimenez, A., Kemp, M., and Wang, C. (2015). Cadec: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, 55:73–81, June.

Segura-Bedmar, I., Revert, R., and Martínez, P. (2014). Detecting drugs and adverse events from Spanish social media streams. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, pages 106–115, Gothenburg, Sweden. Association for Computational Linguistics.

Seiffe, L., Marten, O., Mikhailov, M., Schmeier, S., Möller, S., and Roller, R. (2020). From Witch's Shot to Music Making Bones - resources for medical laymen to technical language and vice versa. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6185–6192, May.

Tutubalina, E., Alimova, I., Miftahutdinov, Z., Sakhovskiy, A., Malykh, V., and Nikolenko, S. (2021). The Russian Drug Reaction Corpus and neural models for drug reactions and effectiveness detection in user reviews. *Bioinformatics*, 37(2):243–249, April.

Yada, S., Joh, A., Tanaka, R., Cheng, F., Aramaki, E., and Kurohashi, S. (2020). Towards a versatile medical-annotation guideline feasible without heavy medical knowledge: Starting from critical lung diseases. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4565–4572, May.

Zolnoori, M., Fung, K. W., Patrick, T. B., Fontelo, P., Kharrazi, H., Faiola, A., Shah, N. D., Shirley Wu, Y. S., Eldredge, C. E., Luo, J., Conway, M., Zhu, J., Park, S. K., Xu, K., and Moayyed, H. (2019). The PsyTAR dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications. *Data in Brief*, 24, June.

Appendix

		Class 0			Class 1			Macro Average			
model	**seed**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**AUC**
XLM_R_1	78	66.67	71.26	68.89	92.42	90.77	91.59	79.55	81.02	80.24	81.02
XLM_R_2	99	68.57	55.17	61.15	88.95	93.45	91.15	78.76	74.31	76.15	74.31
XLM_R_3	227	62.77	67.82	65.19	91.49	89.58	90.53	77.13	78.70	77.86	78.70
XLM_R_4	409	66.25	60.92	63.47	90.09	91.96	91.02	78.17	76.44	77.24	76.44
XLM_R_5	422	70.77	52.87	60.53	88.55	94.35	91.35	79.66	73.61	75.94	73.61
XLM_R_6	482	64.89	70.11	67.40	92.10	90.18	91.13	78.50	80.15	79.27	80.15
XLM_R_7	485	59.48	79.31	67.98	94.14	86.01	89.89	76.81	82.66	78.94	82.66
XLM_R_8	841	61.22	68.97	64.86	91.69	88.69	90.17	76.46	78.83	77.52	78.83
XLM_R_9	857	67.90	63.22	65.48	90.64	92.26	91.45	79.27	77.74	78.46	77.74
XLM_R_10	910	71.43	63.22	67.07	90.75	93.45	92.08	81.09	78.34	79.58	78.34
	mean	66.00	65.29	65.20	91.08	91.07	91.03	78.54	78.18	78.12	78.18
	std	3.94	7.89	2.81	1.67	2.55	0.67	1.45	2.82	1.44	2.82

Table 5: Source language data (English): results for XLM-RoBERTa in precision, recall and F1 score per class and macro-averaged. The models have the same configuration and are trained and tested on the exact same data, but have a different seed for initialization. Support for class 0: 87, support for class 1: 336

		Class 0			Class 1			Macro Average			
model	**seed**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**AUC**
BRB_1	78	75.41	52.87	62.16	88.67	95.54	91.98	82.04	74.20	77.07	74.20
BRB_2	99	68.82	73.56	71.11	93.03	91.37	92.19	80.92	82.47	81.65	82.47
BRB_3	227	66.67	73.56	69.95	92.97	90.48	91.70	79.82	82.02	80.82	82.02
BRB_4	409	60.16	85.06	70.48	95.67	85.42	90.25	77.91	85.24	80.36	85.24
BRB_5	422	64.36	74.71	69.15	93.17	89.29	91.19	78.76	82.00	80.17	82.00
BRB_6	482	73.26	72.41	72.83	92.88	93.15	93.02	83.07	82.78	82.92	82.78
BRB_7	485	62.50	63.22	62.86	90.45	90.18	90.31	76.47	76.70	76.59	76.70
BRB_8	841	61.22	68.97	64.86	91.69	88.69	90.17	76.46	78.83	77.52	78.83
BRB_9	857	63.54	70.11	66.67	92.05	89.58	90.80	77.80	79.85	78.73	79.85
BRB_10	910	78.33	54.02	63.95	88.98	96.13	92.42	83.66	75.08	78.18	75.08
	mean	67.43	68.85	67.40	91.96	90.98	91.40	79.69	79.92	79.40	79.92
	std dev	6.32	9.79	3.79	2.11	3.23	1.01	2.64	3.64	2.10	3.64

Table 6: Source language data (English): results for BioRedditBERT in precision, recall and F1 score per class and macro-averaged. The models have the same configuration and are trained and tested on the exact same data, but have a different seed for initialization. Support for class 0: 87, support for class 1: 336

model	**data**	**learning rate**	**batch size**	**freeze**	**train sampler**
XLM-R	English	0.00001056	7	1	random
BRB	English	0.00001584	8	1	random
XLM-R	German (full)	0.00001056	7	0	weighted

Table 7: Specifications of the best models. The first and second lines correspond to the basis for the few-shot experiments where we trained 10 versions, the bottom one is XLM-RoBERTa again fine-tuned on the German full dataset. For the first two, a random sampler and freezing all layers except the classifier worked best, while not freezing any layers and using a weighted training sampler achieved the best performance for the third model.

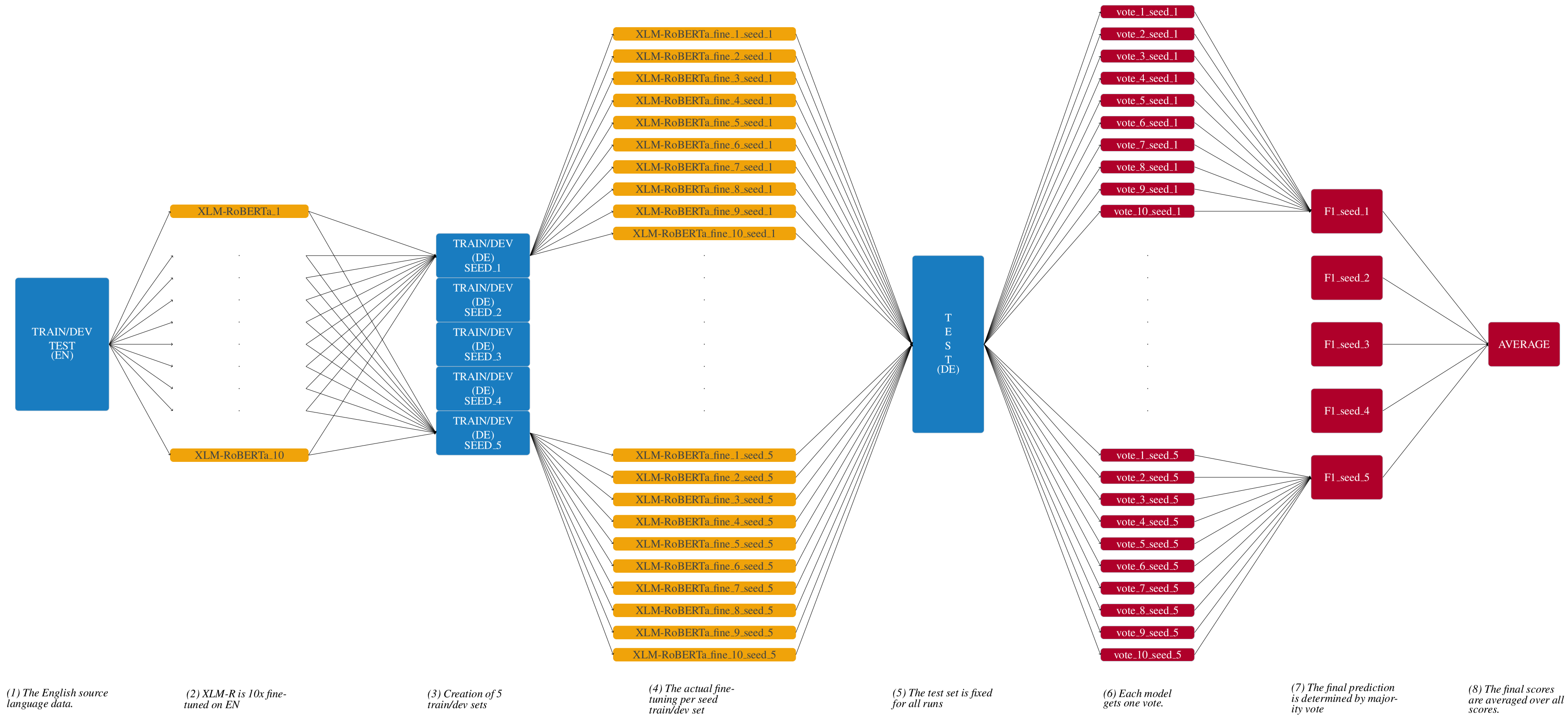
TRAIN/DEV TEST (EN)
XLM-RoBERTa_1
XLM-RoBERTa_10
TRAIN/DEV (DE) SEED_1
TRAIN/DEV (DE) SEED_2
TRAIN/DEV (DE) SEED_3
TRAIN/DEV (DE) SEED_4
TRAIN/DEV (DE) SEED_5
XLM-RoBERTa_fine_1_seed_1
XLM-RoBERTa_fine_2_seed_1
XLM-RoBERTa_fine_3_seed_1
XLM-RoBERTa_fine_4_seed_1
XLM-RoBERTa_fine_5_seed_1
XLM-RoBERTa_fine_6_seed_1
XLM-RoBERTa_fine_7_seed_1
XLM-RoBERTa_fine_8_seed_1
XLM-RoBERTa_fine_9_seed_1
XLM-RoBERTa_fine_10_seed_1
XLM-RoBERTa_fine_1_seed_5
XLM-RoBERTa_fine_2_seed_5
XLM-RoBERTa_fine_3_seed_5
XLM-RoBERTa_fine_4_seed_5
XLM-RoBERTa_fine_5_seed_5
XLM-RoBERTa_fine_6_seed_5
XLM-RoBERTa_fine_7_seed_5
XLM-RoBERTa_fine_8_seed_5
XLM-RoBERTa_fine_9_seed_5
XLM-RoBERTa_fine_10_seed_5
TEST (DE)
vote_1_seed_1
vote_2_seed_1
vote_3_seed_1
vote_4_seed_1
vote_5_seed_1
vote_6_seed_1
vote_7_seed_1
vote_8_seed_1
vote_9_seed_1
vote_10_seed_1
vote_1_seed_5
vote_2_seed_5
vote_3_seed_5
vote_4_seed_5
vote_5_seed_5
vote_6_seed_5
vote_7_seed_5
vote_8_seed_5
vote_9_seed_5
vote_10_seed_5
F1_seed_1
F1_seed_2
F1_seed_3
F1_seed_4
F1_seed_5
AVERAGE
(1) The English source language data.
(2) XLM-R is 10x fine-tuned on EN
(3) Creation of 5 train/dev sets
(4) The actual fine-tuning per seed train/dev set
(5) The test set is fixed for all runs
(6) Each model gets one vote.
(7) The final prediction is determined by majority vote
(8) The final scores are averaged over all scores.

Figure 3: The setup for the few-shot experiments: (1 + 2) We fine-tune 10 XLM-R models on the English source language data (fine-tuning 1). (3) Then, we choose 5 seeds and create 5 train/dev sets, from which we sample the shots. (4) We fine-tune (fine-tuning 2) each XLM-R model on each seed data, obtaining 10 XLM-R_fine models for every seed. (5) For every seed, each model is applied to the test set, and (6) we vote on the final results. (7) We obtain 5 results, one for every seed (F1-scores etc). (8)Those 5 results per setting are averaged.