Zero-Shot Cross-Lingual Sentiment Classification under Distribution Shift: an Exploratory Study

Maarten De Raedt$^{\diamond\clubsuit}$ **Semere Kiros Bitew**$^{\clubsuit}$ **Fréderic Godin**$^{\diamond}$ **Thomas Demeester**$^{\clubsuit}$ **Chris Develder**$^{\clubsuit}$

$^{\diamond}$ Sinch Chatlayer $^{\clubsuit}$ Ghent University - imec

{maarten.deraedt, semerekiros.bitew, thomas.demeester, chris.develder}@ugent.be
frederic.godin@sinch.com

Abstract

The brittleness of finetuned language model performance on out-of-distribution (OOD) test samples in unseen domains has been well-studied for English, yet is unexplored for multilingual models. Therefore, we study generalization to OOD test data specifically in zero-shot cross-lingual transfer settings, analyzing performance impacts of both *language* and *domain* shifts between train and test data. We further assess the effectiveness of counterfactually augmented data (CAD) in improving OOD generalization for the cross-lingual setting, since CAD has been shown to benefit in a monolingual English setting. Finally, we propose two new approaches for OOD generalization that avoid the costly annotation process associated with CAD, by exploiting the power of recent large language models (LLMs). We experiment with 3 multilingual models, LaBSE, mBERT, and XLM-R trained on English IMDb movie reviews, and evaluate on OOD test sets in 13 languages: Amazon product reviews, Tweets, and Restaurant reviews. Results echo the OOD performance decline observed in the monolingual English setting. Further, (i) counterfactuals from the original high-resource language do improve OOD generalization in the low-resource language, and (ii) our newly proposed cost-effective approaches reach similar or up to +3.1% better accuracy than CAD for Amazon and Restaurant reviews.

1 Introduction

To solve Natural Language Processing (NLP) tasks in low-resource languages, using multilingual models is a much adopted strategy (Devlin et al., 2019; Artetxe and Schwenk, 2019; Conneau and Lample, 2019; Feng et al., 2022). A particularly popular paradigm is zero-shot cross-lingual transfer (Ruder et al., 2019; Artetxe et al., 2020b; Hu et al., 2020; Lauscher et al., 2020): pre-trained multilingual models are finetuned on downstream tasks with training data solely from a high-resource language
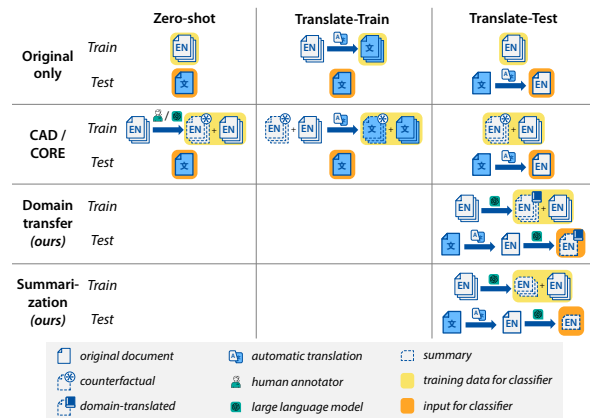

Fig. 1: **Zero-shot cross-lingual transfer setup.** Multiple transfer strategies, including our newly proposed *summarization* and *domain transfer* methods for boosting OOD generalization.

(e.g., English). The resulting finetuned model can then be applied on a low-resource language samples, i.e., without requiring costly training data in the low-resource language.

In such zero-shot cross-lingual transfer, linguistic discrepancy between training and test languages causes a challenge: typically, performance is subpar compared to monolingual models.[1] Several works have looked into narrowing the performance gap stemming from such language-based distribution shift (Liu et al., 2021; Yu and Joty, 2021; Zheng et al., 2021; Artetxe et al., 2023).

Yet, besides the language-based shift, in real-world settings there may also be a domain-shift between training and test samples, i.e., test samples may comprise out-of-distribution (OOD) data (Quiñonero-Candela et al., 2008). For example, a sentiment classifier to predict positive/negative appreciation by a consumer may be trained on movie reviews but applied on product reviews or tweets, where underlying sentiment features are assumed to be invariant (Arora et al., 2021).

---

[1]Admittedly, such monolingual models do need low-resource training data.

In a monolingual (English) setting, several studies unsurprisingly found a performance degradation when evaluating on OOD test data rather than on in-distribution (ID) data (Kaushik et al., 2019, 2020; Gardner et al., 2020; Katakkar et al., 2022). One of the underlying causes for that performance drop was found to be the classifier's reliance on spurious features, i.e., patterns that from a human perspective should not be indicative for the classifier's label (Poliak et al., 2018; Gururangan et al., 2018; Mc-Coy et al., 2019; Wang and Culotta, 2020; Joshi et al., 2022): e.g., Wang and Culotta (2020) found the occurrence of "*Spielberg*" to be important for a positive sentiment classification.

A strategy that has been shown to improve OOD generalization in the monolingual English setting is the use of counterfactually augmented data (CAD), where annotators minimally revise training data to flip their labels (Kaushik et al., 2019). Yet, constructing such annotations is costly: Kaushik et al. (2019) report 5 min/sample.

In this paper, we present an exploratory study of OOD generalization specifically in a *cross-lingual* context, since we found this not to be covered in related work (§2). Specifically, we (i) identify the impact of OOD data on zero-shot *cross-lingual* transfer performance, aiming to disentangle performance drops stemming from *language* vs. *domain* shifts between training and test data, and (ii) propose and analyze two new data augmentation strategies to improve OOD generalization that *avoid the costly annotations* associated with using counterfactuals. For both, we present an empirical study (§3) within the domain of binary sentiment analysis. We consider English IMDb reviews (Maas et al., 2011) as in-distribution training data, with out-of-distribution test data spanning 13 languages across the Amazon (Keung et al., 2020), Tweets (Barbieri et al., 2022), and Restaurants (Pontiki et al., 2016) datasets. We further experiment with pre-trained multilingual models mBERT (Devlin et al., 2019), XLM-R (Conneau and Lample, 2019), and LaBSE (Feng et al., 2022).

For (i), we answer a first research question, **(RQ1)** *How well do zero-shot cross-lingual methods trained with English sentiment data generalize to out-of-distribution samples in non-English languages?* To this end, we finetune the multilingual models on the English IMDb sentiment data, and evaluate their performance on OOD test samples in non-English languages.

For (ii), we answer **(RQ2)** *How can zero-shot cross-lingual transfer methods better generalize to out-of-distribution samples, including for non-English languages?* We will consider a CAD baseline as proposed by Kaushik et al. (2019), where annotators minimally revise training data to flip their labels, since training on both original and counterfactual data improves OOD generalization to unseen domains in the monolingual English setting. Specifically, we finetune the multilingual models on both the original English and counterfactually revised English IMDb reviews, and evaluate whether the OOD generalization gains observed in the monolingual setting translate also to OOD test samples in non-English languages.

We then propose (§3.3) two cost-effective alternatives for CAD, using Large Language Models (LLMs): (1) *domain transfer*, and (2) *summarization*, as illustrated in the 2 bottom rows of Fig. 1. For (1), we prompt an LLM to minimally edit both ID training and OOD test samples to map them onto the same, *hypothetical* domain, e.g., books. For (2), we prompt an LLM to abstractly summarize both ID training and OOD test data, since we hypothesize that summaries can capture the core essence of samples while removing non-essential, potentially spurious, information.

Our results (§4) show that in the OOD test setting for non-English languages, model performance of zero-shot cross-lingual transfer substantially declines, aligned with OOD generalization studies in a monolingual English setting. We further find that CAD improves OOD generalization for non-English samples, with gains up to +14.8%, +4.7%, and +7.9% for respectively LaBSE, mBERT, and XLM-R. Finally, our cost-effective *domain transfer* and *summarization* data augmentation methods similarly improve OOD generalization, on par with or even surpassing CAD for *Amazon* and *Restaurants* by up to +3.1% in accuracy.

2 Related Work

Zero-shot cross-lingual transfer: A large part of multilingual NLP research focuses on improving the transfer of multilingual models trained on high-resource language data to low-resource languages. This can be achieved either by (i) cross-lingual pre-training schemes that yield stronger multilingual models (Artetxe and Schwenk, 2019; Conneau and Lample, 2019; Conneau et al., 2020; Xue et al., 2021; Feng et al., 2022; Chi et al., 2022), or (ii) fine-

tuning strategies that facilitate better cross-lingual transfer (Liu et al., 2021; Yu and Joty, 2021; Zheng et al., 2021). Recently, Artetxe et al. (2023) revisited the *translate-test* and *translate-train* baselines (Shi et al., 2010; Duh et al., 2011; Artetxe et al., 2020a), where *test* samples are translated into English prior to evaluating them, or, respectively, the *training* samples are translated into the test languages for fine-tuning a multilingual model. Artetxe et al. found that using more recent machine translation systems, e.g., NLLB (Costa-jussà et al., 2022), further boosts performance and often surpasses strong zero-shot cross-lingual methods. Hence, we also experiment with *translate-test* and *translate-train* approaches.

Cross-lingual transfer under distribution shift: The limited research on the robustness of multilingual models has primarily focused on being robust against specific types of *noise*, e.g., adversarial perturbations for Japanese Natural Language Inference (Yanaka and Mineshima, 2021), a combination of general and task-specific text transformations based on manipulating synonyms, antonyms, syntax, etc. (Wang et al., 2021), and introducing errors and noise through Wikipedia edits (Cooper Stickland et al., 2023). Unlike these works, we will evaluate how well zero-shot cross-lingual transfer from English to non-English test samples can generalize in scenarios where there is a shift in *domain* from train to test data: the domain-specific features of test samples may change, whereas the semantic sentiment features remain invariant.

Counterfactually-augmented data (CAD): For English sentiment analysis, CAD is widely adopted to mitigate the effect of spurious patterns. For example, Kaushik et al. (2019, 2020) recruited Mechanical Turk workers to construct counterfactually revised samples by flipping labels with minimal editing, helping classifiers to learn real associations between samples and labels, thereby improving OOD generalization to unseen test domains. Building upon the success of CAD, several works have also studied how to automatically generate counterfactuals for English sentiment analysis (Wang and Culotta, 2021; Yang et al., 2021; Dixit et al., 2022; Howard et al., 2022; De Raedt et al., 2022). We adopt this CAD idea for OOD generalization in a zero-shot cross-lingual setting, which to the best of our knowledge has not been studied yet.

We start by exploring whether augmenting the English training data with the manually constructed counterfactuals from Kaushik et al. (2019) also benefits OOD generalization for non-English test samples. Additionally, we propose two new LLM-based methods as alternatives to constructing counterfactuals, aiming to specifically improve zero-shot transfer to non-English OOD test samples. We benchmark our new LLM-based methods against a CAD setup following Kaushik et al. (2019), thus assessing whether we can achieve similar OOD performance but avoid CAD's costly human annotations. We further contrast classifiers trained on data augmented by our two new LLM-based methods to those trained on counterfactuals generated by CORE (Dixit et al., 2022), the state-of-the-art method in automatic counterfactual generation. CORE first retrieves naturally occurring counterfactual edits from an unlabeled text corpus and then, based on these retrieved edits, instructs an LLM (GPT-3) to counterfactually revise training samples.

3 Experimental Setup

We describe the English ID training data and non-English OOD test data in §3.1. Next, we outline the pre-trained multilingual models and the transfer strategies we experiment with in §3.2. In §3.3, we present our LLM-based *domain transfer* and *summarization* data augmentation methods. We cover finetuning and evaluation in §3.4.

3.1 Datasets

In-distribution (ID) training data: We use the subset of 1,707 English reviews selected by Kaushik et al. (2019) from the IMDb sentiment dataset (Maas et al., 2011) as training data, as well as 245 English validation samples. To better assess the OOD generalization of cross-lingual transfer, we also report in-distribution results of all 13 considered languages on the IMDb test set with 488 samples. However, the test set of Kaushik et al. (2019) is English-only. Hence, we translate the 488 English test samples into each of the 12 other non-English languages, using OpenAI's `ChatGPT-turbo (v0301)` (Ouyang et al., 2022), as it achieves translation quality that is competitive to commercial machine translation tools (e.g., Google Translate or Microsoft Translation Suite) (Jiao et al., 2023; Hendy et al., 2023; Peng et al., 2023), while being more cost-effective. Since we aim to explore the benefits of English CAD for

		Original samples
	IMDB	If you haven't seen this, it's terrible. It is pure trash. I saw this about 17 years ago, and I'm still screwed up from it.
	TWEETS	She just didn't get them in areas were she needed them. Lots of voter supression going on. Hacking & tampering
	AMAZON	The straps are super small, for a very small wrist, and the closure is bad, easy to lose the watch.
	RESTAURANTS	The food is standard, but the person waiting at the door in the style of a manager is cold and unfriendly.
		Domain transferred samples
	IMDB	If you haven't read this book, it's terrible. It is pure trash. I read this about 17 years ago, and I'm still screwed up from it.
	TWEETS	She just didn't get the books in areas where she needed them. Lots of book censorship going on. Piracy & Plagiarism
	AMAZON	The binding of the book is super tight, suited for a compact size, and the cover is not secure, easy to lose the pages.
	RESTAURANTS	The books are average, but the person at the front desk in a manager-like role is distant and unapproachable.
		Summarized samples
	IMDB	Terrible and traumatizing movie, avoid it.
	TWEETS	Allegations of voter suppression and tampering.
	AMAZON	Small straps, bad closure, easy to lose.
	RESTAURANTS	Standard food, unfriendly manager.

Table 1: **LLM-based data-augmentation.** *Top:* original ID training and OOD test samples (including English translations). *Middle:* mapping of the diverse domain samples onto the *hypothetical* books domain. *Bottom:* demonstrates how *summarization* retains essential information while removing potentially spurious elements.

OOD generalization also to non-English test samples, we augment the respectively 1,707 and 488 original training and validation samples with their English counterfactually revised counterparts provided by Kaushik et al. (2019). All training, validation, and test sets are equally balanced between positive and negative samples.

Out-of-distribution (OOD) test data: Our OOD test data comprises three non-movie domains: *product reviews*, *tweets* and *restaurant feedback*. We use the MARC dataset (Keung et al., 2020) for Amazon *product reviews* in six languages: English, German, French, Spanish, Japanese, and Chinese. For *tweets*, we use the recent multilingual test sets provided by Barbieri et al. (2022), in eight languages: English, German, French, Spanish, Arabic, Hindi, Portuguese, and Italian. For *restaurant* reviews, we use the multilingual aspect-based sentiment classification dataset for the 2016 SemEval Task 5 (Pontiki et al., 2016), i.e., its restaurant domain data covering six languages: English, Dutch, French, Spanish, Russian, and Turkish. Since SemEval Task 5 concerns aspect-based sentiment, we apply a rule-based mapping to cast it as a binary classification task: included reviews are labeled either as *positive* (if all aspects are positive or a mix of neutral and positive) or *negative* (if all aspects are negative or a mix of neutral and negative). We undersample test examples from the majority sentiment to ensure that all test sets are balanced. Further dataset statistics are provided in Appendix A.

3.2 Zero-shot cross-lingual transfer

Pre-trained multilingual models: We consider the base-cased versions of two multilingual language models pre-trained on masked language model (MLM) objectives: mBERT, i.e., a multilingual variant of BERT (Devlin et al., 2019), and XLM-R, a RoBERTa-based multilingual model (Conneau and Lample, 2019). Additionally, we use the pre-trained multilingual sentence encoder LaBSE (Feng et al., 2022) that maps sentences to 768-dimensional single vector representations.

Transfer strategies: To transfer from the English ID training data to non-English test samples, we use 3 widely adopted strategies (Fig. 1, top row):

(1) *zero-shot*: finetunes the multilingual model on the English ID training and validation set, followed by directly evaluating the OOD test samples in the non-English languages.
(2) *translate-test*: finetunes the multilingual model on the English ID training and validation datasets. However, prior to making predictions for OOD test samples, the samples are translated into English.
(3) *translate-train*: first translates the English ID training and validation datasets to the target OOD test language. Subsequently, the multilingual model is trained on this translated data to then make predictions for the original, untranslated, OOD test samples in that non-English language.

Note that in case where both *translate-train* and CAD are used, the English CAD training and validation data are translated to the target OOD test language. For both *translate-test* and *translate-train*, we use OpenAI's ChatGPT-turbo (v0301) (Ouyang et al., 2022) as the LLM to translate from English to non-English languages and vice versa. We adopt OpenAI's default parameter values. See Appendix A for translation prompts.

3.3 LLM-based data-augmentation

We explore whether data augmentation using an LLM, as a cost-effective alternative to CAD, can also boost OOD generalization. We propose two such alternatives: (1) *domain transfer*, and (2) *summarization*. Our focus is on augmenting data for *translate-test*, as recent work has shown it to be more effective than *zero-shot* and *translate-train* (Artetxe et al., 2023). The multilingual models are finetuned on the original English ID, as well as the augmented ID training samples[2], with predictions made solely on augmented test samples. Table 1 provides illustrations for both strategies.

Domain transfer: We align the domains of both the original ID training and OOD test samples *translated* into English to a common *hypothetical* domain. To achieve this, we instruct ChatGPT-turbo (v0301) (Ouyang et al., 2022) to minimally change the samples so that they relate to the new *hypothetical* domain, for which we experiment with the domain of *books*. Note that rather than solely mapping OOD test samples to the ID training domain of movies, we use a *hypothetical* domain to transform both training and test samples with an LLM to avoid introducing a new distribution shift caused by the mismatch between the original human-based training and the LLM-generated test samples. See Appendix A for our domain transfer prompt.

Summarization: For our second augmentation strategy, we abstractly summarize both the original English training and the *translated* English OOD test samples. We hypothesize that such concise summaries can retain essential information while omitting non-essential and potentially spurious features, such as, e.g., specific syntax structures and lexical choices, thereby a priori preventing classifiers from relying on such features for prediction. Furthermore, transforming text with language models, i.e., through summarization, may have the added benefit of normalizing the background, non-sentiment related, features. Hence, summarizing the data can lead to more uniform syntax and word choice among test and training samples, potentially further narrowing the distribution mismatch between ID training and OOD test samples. Appendix A lists the exact prompt that we

---

[2]To ensure all strategies have the same number of training samples, we train the *original-only* models (without manual counterfactuals or LLM-augmented samples) on twice the number (3.4k) of original IMDb reviews (§3.4).

supply to ChatGPT-turbo (v0301) (Ouyang et al., 2022), using OpenAI's default parameter values.

3.4 Finetuning and evaluation

We finetune the MLM-based models, i.e., mBERT and XLM-R, by adding a classification head to the [CLS]-token. We use the Hugging Face Transformers library (Wolf et al., 2020) and train on a single Tesla V100 GPU for 20 epochs, with a batch size of 38, and a learning rate of $5e–6$. To select an optimal model, we employ early validation stopping with a loss threshold of 0.01 and a patience of 10. Since we are also interested in measuring the performance of a more compute-efficient model, we freeze LaBSE's parameters and train on CPU a logistic regression model on LaBSE's sentence vectors through five-fold cross-validation. We use the scikit-learn library (Pedregosa et al., 2011), with lbfgs (Liu and Nocedal, 1989) as the solver, and set the maximum number of iterations to 5,000.

To assess the performance of each transfer strategy, we report the mean accuracy over 5 randomly initialized training runs, i.e., with randomly selected weights and cross-validation folds for respectively mBERT/XLM-R and LaBSE.

Note that classifiers trained on CAD, as well as on data augmented by our two strategies, use respectively 1.7k extra manually constructed counterfactuals and 1.7k extra LLM-generated samples, in addition to the 1.7k original IMDb training samples. To ensure that the OOD generalization gains from CAD and our two augmentation strategies are not solely attributed to the increased number of training samples, we randomly sample an extra 1.7k original English IMDb reviews from the IMDb dataset of Maas et al. (2011) for the *original-only* strategy (i.e., models trained without counterfactuals or augmented data). As such, all considered strategies are trained on 3.4k samples

4 Experimental Results and Discussion

4.1 Zero-shot cross-lingual out-of-distribution generalization

We first address **(RQ1)**, on assessing OOD generalization to non-English samples. In Table 2, we present both ID and OOD accuracies of the *original only* method, which trains solely on (translated) English IMDb movie reviews without data augmentation.

We see that both for English and non-English, all models and transfer strategies decline in perfor-

| | IMDB | | AMAZON | | RESTAURANTS | | TWEETS | |
Method	EN	NON-EN	EN	NON-EN	EN	NON-EN	EN	NON-EN
LaBSE								
- ZSHOT	85.0	84.9	66.3	71.9	72.7	74.1	76.3	67.8
- TTRAIN	85.0	85.2	66.3	74.0	72.7	76.4	76.3	66.0
- TTEST	85.0	-	66.3	67.6	72.7	73.1	76.3	68.8
mBERT								
- ZSHOT	89.5	80.8	79.3	72.2	80.2	69.6	75.9	62.8
- TTRAIN	89.5	87.5	79.3	73.5	80.2	74.5	75.9	62.9
- TTEST	89.5	-	79.3	77.8	80.2	78.9	75.9	71.1
XLM-R								
- ZSHOT	92.4	88.4	86.3	85.0	86.0	79.2	84.3	69.2
- TTRAIN	92.4	90.7	86.3	86.0	86.0	83.0	84.3	72.5
- TTEST	92.4	-	86.3	85.6	86.0	81.5	84.3	71.7

Table 2: **In-distribution** vs. **out-of-distribution** test accuracies for the *original only* strategy trained solely on IMDb reviews (without CAD or data augmentation). Results are presented for English (EN) and non-English (NON-EN) test data, with the latter's accuracies averaged across all non-English languages per test set. Detailed results per language are provided in Appendix A. Note, for English, TTRAIN and TTEST do not involve any translation, hence their EN scores are equivalent to ZSHOT. Further, ID scores for TTEST are omitted as these would involve backtranslating the non-English ID samples (originally translated from English ID test data per §3.1) to English, which would largely assess back-translation quality.

mance when evaluated on OOD rather than ID test samples. For example, the *zero-shot* strategy's drop from English ID to English OOD ($ID_{EN} \rightarrow OOD_{EN}$) ranges from 8.7%–18.7% for LaBSE, 9.3%–13.6% for mBERT, and 6.1%–8.1% for XLM-R. Similarly, for non-English ($ID_{NON-EN} \rightarrow OOD_{NON-EN}$), the performance drops for LaBSE, mBERT, and XLM-R vary within the ranges of 10.8%–17.1%, 8.6%–18%, and 3.4%–19.2%, respectively. These findings suggest that model performance decline to OOD test samples in non-English is substantial, as was already known (and here confirmed again) for English. We do not, however, see a consistently stronger decline for non-English than for English, as may be intuitively expected. This is discussed in more detail in the next paragraph.

English vs. non-English OOD generalization: We assess whether multilingual models generalize better to English than non-English OOD test data. Overall, the EN versus NON-EN scores in Table 2 reveal that the MLM-based models mBERT and XLM-R generalize less well to non-English compared to English OOD test samples: the accuracies for non-English languages are lower in most cases. Surprisingly, the converse holds for LaBSE: it has consistently better non-English OOD accuracies compared to English on *Amazon* and *Restaurants*. Note, however, the absolute performance of the three models: LaBSE appears to be

the least accurate model in English in most cases. This is consistent with the fact that its encoder remains frozen during training in English, unlike the other encoders, whereas LaBSE's non-English performance is more on par with the other models. While our results suggest that performance decline to OOD test samples in non-English and English is substantial, the disparity among OOD model performance between non-English and English depends on the (i) pre-trained multilingual model or finetuning strategy, and (ii) the type of OOD data.

Impact of the pre-trained multilingual models: We compare the OOD generalization of LaBSE, mBERT, and XLM-R. The results in Table 2 show XLM-R as the top performer, consistently surpassing both LaBSE and mBERT. Despite having only 768 trainable parameters (frozen encoder with trainable logistic regression layer) against mBERT's 110M (fully tuned), it is surprising that LaBSE is at least on par with mBERT on non-English OOD data, except for *translate-test*. This suggests a stronger bias towards English in mBERT compared to LaBSE, also evidenced by an 8.7% drop in mBERT's ID *zero-shot* performance between English and non-English, whereas this difference is just 0.1% for LaBSE.

Impact of the transfer strategies: We assess the *translate-train* and *translate-test* strategies for

Method	LaBSE						mBERT						XLM-R					
	AMAZON		RESTAURANTS		TWEETS		AMAZON		RESTAURANTS		TWEETS		AMAZON		RESTAURANTS		TWEETS	
	EN	NON-EN	EN	NON-EN	EN	NON-EN	EN	NON-EN	EN	NON-EN	EN	NON-EN	EN	NON-EN	EN	NON-EN	EN	NON-EN
Original only																		
- ZSHOT	66.3	71.9	72.7	74.1	76.3	67.8	79.3	72.2	80.2	69.6	75.9	62.8	86.3	85.0	86.0	79.2	84.3	69.2
- TTRAIN	66.3	74.0	72.7	76.4	76.3	66.0	79.3	73.5	80.2	74.5	75.9	62.9	86.3	86.0	86.0	83.0	84.3	72.5
- TTEST	66.3	67.6	72.7	73.1	76.3	68.8	79.3	77.8	80.2	78.9	<u>75.9</u>	71.1	86.3	85.6	86.0	81.5	<u>84.3</u>	71.7
Original + CAD (Kaushik et al., 2019)																		
- ZSHOT	81.2	<u>82.9</u>	84.7	85.7	81.7	<u>74.5</u>	81.7	74.9	81.8	70.9	79.0	67.2	87.0	85.7	87.5	81.9	86.7	75.9
- TTRAIN	81.2	82.3	84.7	83.4	81.7	73.7	81.7	78.2	81.8	75.7	79.0	66.9	87.0	86.4	87.5	84.6	86.7	77.3
- TTEST	81.2	82.4	84.7	85.9	**81.7**	76.2	**81.7**	81.2	81.8	<u>81.2</u>	79.0	75.0	87.0	86.8	87.5	87.1	**86.7**	<u>79.6</u>
Original + CORE (Dixit et al., 2022)																		
- ZSHOT	81.0	82.0	85.0	84.9	77.4	71.1	80.2	74.1	80.4	69.6	73.6	64.8	86.8	87.0	89.7	87.5	83.9	77.9
- TTEST	81.0	81.7	<u>85.0</u>	86.3	<u>77.4</u>	74.3	80.2	79.9	80.4	79.9	73.6	72.8	86.8	87.0	<u>89.7</u>	<u>89.1</u>	83.9	**80.5**
Original + Domain transfer (ours)																		
TTEST+TRAN.	<u>81.7</u>	81.9	84.1	84.1	72.3	69.6	<u>81.3</u>	<u>80.3</u>	<u>83.3</u>	81.0	72.4	69.7	<u>87.1</u>	**87.1**	87.2	84.5	72.7	69.7
Original + Summarization (ours)																		
TTEST+SUM.	**86.2**	**84.7**	**91.6**	**88.8**	76.6	74.0	81.1	**81.2**	**87.3**	**84.3**	74.3	<u>73.8</u>	**87.8**	86.8	**92.8**	**90.2**	83.0	75.9

Table 3: **Out-of-distribution generalization with data augmentation.** *Original only*: baseline model trained solely on IMDb reviews, without CAD or data augmentation. *+CAD*: augments IMDb training samples with manually constructed counterfactuals. *+CORE*: augments training samples with automatically generated counterfactuals. *+Domain transfer* and *+Summarization* augment the training data with our newly proposed strategies. **Best** model in bold with the <u>runner-up</u> underlined.

OOD generalization against the *zero-shot* approach. The results in Table 2 reveal large OOD generalization gains for non-English languages using *translate-test* and mBERT, with accuracy gains between +5.6% and +9.3%. This supports our previous discussion of mBERT being more biased towards English. For LaBSE, *translate-train* is most effective on *Amazon* and *Restaurants*, with average accuracy boosts of +2.1% and +2.3% respectively, but not for *Tweets* (−1.8%). For XLM-R, *Restaurants* and *Tweets* benefit most from translation: *translate-train* (*translate-test*) surpass *zero-shot* with respective gains of +3.8% (+2.3%) and +3.3% (+2.5%). In conclusion, while translation-based strategies can further boost the OOD generalization zero-shot cross-lingual transfer, the benefits are dependent on the multilingual model and OOD test data.

4.2 Out-of-distribution generalization with data augmentation

To address (**RQ2**) on achieving better OOD generalization, we first analyze the effect of augmenting training data with the manually constructed counterfactuals of Kaushik et al. (2019). These counterfactuals will serve as an upper baseline against which we will subsequently compare the performance of models trained on (i) counterfactuals generated by the state-of-the-art in automatic counterfactual construction, i.e., CORE (Dixit et al., 2022), and (ii) our LLM *domain transferred* and *summa-*

rized augmented data.

Manually constructed counterfactuals: Comparing the *original + CAD* results in Table 3 to the corresponding *original only* results, reveals that augmenting training data with CAD consistently boosts OOD generalization, across all datasets and both for English and non-English test samples. Accuracy gains averaged over the non-English languages for OOD vary between 7%–14.8%, 1.2%–4.7%, and 0.4%–7.9% for respectively LaBSE, mBERT, and XLM-R. This confirms that the English OOD generalization gains of CAD based training (Kaushik et al., 2019) translate well to non-English OOD test data in a cross-lingual setting.

Impact of LLM-based data augmentation on cross-lingual OOD generalization: As an alternative to costly manually constructed counterfactuals, we investigate the viability of *automatic* data augmentation: CORE from Dixit et al. (2022) (replacing humans with the LLM for counterfactual creation), as well as our newly proposed *domain transfer* and *summarization* strategies described in §3.3. First, we compare the non-English OOD generalization of models trained with augmented data to models trained solely on original data. Table 3 shows clear non-English OOD improvements for all of LaBSE, mBERT, and XLM-R, with respective gains over *original only* ranging from: (i) 3.3%–14.1%, 0%–2.1%,

and 1.4%–8.8% for CORE, (ii) 0.8%–14.3%, −1.4%–2.5%, and −2.0%–3% for *domain transfer*, and (iii) 5.2%–17.1%, 2.7%–5.4%, and 1.3%–8.7% for *summarization*. The drops −1.4% and −2.0% for mBERT and XLM-R on *Tweets* suggest that *domain transfer* is less effective when the discrepancy between test and training domains is excessively large: the IMDb training data, similar to the *Amazon* and *Restaurant* domains, comprises reviews, whereas *Tweets* do not.

The bold and underlined scores in Table 3 denote the top two results. Our *summarization* strategy achieves the best non-English OOD generalization on *Amazon* and *Restaurants*, on par with (or surpassing) models trained on CAD. On *Tweets*, while *summarization* still improves models trained solely on the original data, training on CAD or CORE (XLM-R) yields the best results.

These findings support the efficacy of cost-effective data augmentation as a viable alternative to manually constructed counterfactuals for non-English test data. It is worth noting that our *summarization* and *domain transfer* methods scale linearly, only requiring a single transformation of training samples for each class. However, it is doubtful that CAD and CORE can be similarly expanded beyond binary sentiment classification due to their quadratic data complexity: counterfactuals have to be constructed among every pair of classes.

Impact of LLM-based data augmentation on mono-lingual OOD generalization: Thus far, our analysis has primarily focused on the generalization from English ID training data to non-English OOD test data. Here, we investigate whether our *summarization* and *domain transfer* strategies can also help classifiers generalize in the well-studied monolingual setup, i.e., from English training data to English OOD test data. In this setup, the *translate-test* step is omitted: both the English ID training reviews from IMDb and the English OOD test samples are summarized or domain transferred, without any prior translation.

Comparing the EN scores across the different transfer strategies in Table 3 for each of LaBSE, mBERT, and XLM-R, reveals findings similar to the OOD generalization to non-English languages. (i) For *Amazon* and *Restaurants*, all data augmentation approaches deliver classifiers that better generalize OOD compared to the *original only* classifiers trained without augmented data. Our *summarization* strategy achieves the best overall results,

Method	AMAZON		RESTAURANTS		TWEETS	
	EN	NON-EN	EN	NON-EN	EN	NON-EN
LaBSE						
ZSHOT	77.1	79.7	83.6	83.7	81.9	71.8
+TTEST	77.1	78.9	83.6	84.0	**81.9**	73.1
+SUM.	**86.2**	**84.7**	**91.6**	**88.8**	76.6	**74.0**
mBERT						
ZSHOT	80.7	73.6	82.4	72.5	77.8	63.5
+TTEST	80.7	79.6	82.4	80.0	**77.8**	72.0
+SUM.	**81.0**	**81.2**	**87.3**	**84.3**	74.3	**73.8**
XLM-R						
ZSHOT	87.8	87.7	89.4	84.9	86.3	75.1
+TTEST	87.8	**88.0**	89.4	87.1	**86.3**	**77.8**
+SUM.	87.8	86.8	**92.8**	**90.2**	83.0	75.9

Table 4: **Ablations** of our best data augmentation strategy: *summarization*. ZSHOT: trains on the original English and summarized English IMDb reviews. +TTEST: additionally translates test samples to English. +SUM.: further summarizes the English translated test samples prior inference.

surpassing both classifiers trained on CORE and manually constructed counterfactuals (CAD), except for mBERT and *Amazon*, where CAD results in a minor accuracy gain of 0.6% over *summarization*. (ii) Surprisingly, for *Tweets*, only classifiers trained on manually constructed CAD show consistent OOD generalization improvements over *original-only* classifiers. This is in contrast to the results observed for non-English, where CORE and our *summarization* augmentation approach were able to improve upon the *original-only* classifiers.

Overall, these results highlight that our *summarization* strategy can also benefit monolingual OOD generalization, surpassing classifiers augmented either with CAD or CORE generated counterfactuals for *Amazon* and *Restaurants*.

Ablations: We provide ablations in Table 4 for our most effective strategy, i.e., *summarization*, and find that:
(i) The benefits of translating test samples into English (*translate-test*) versus solely augmenting the training data with summaries (*zero-shot*) vary based on the multilingual and/or OOD test data: there are clear OOD improvements to non-English samples for mBERT and XLM-R, but results for LaBSE are mixed and comparable to the *zero-shot* strategy;
(ii) More importantly, further summarizing the English translated test samples improves OOD generalization more than solely translating them to

English, consistently boosting accuracies by up to +5% for LaBSE and +4.3% mBERT, across all datasets. For XLM-R, summarization slightly reduces accuracy, e.g., −1.2% for non-English languages on *Amazon* and −1.9% for *Tweets* compared to translation alone, yet still boosts OOD generalization to *Restaurants* by 3.1% over *translate-test*.

Cost-effectiveness of LLM-based augmentation: To assess the cost-effectiveness of our LLM-based augmentation, we discuss the costs of our best approach, i.e., *summarization*, and compare it to that cost of employing human workers to manually construct counterfactuals. Kaushik et al. (2019) report that human workers spent an average of 5 minutes revising a single IMDb review, with each worker earning $0.65 per revised review. Therefore, manually revising 1.7K training reviews incurs a total cost of ≈$1,105 and ≈141 hours of labor.

In contrast, our summarization strategy costs $0.0003 on average for summarizing a single training IMDb review, totaling $0.51 for all 1.7K training reviews. However, our best OOD generalization is achieved not only by summarizing training reviews, but also by using an LLM during inference to: (1) translate non-English test samples to English (*translate-test*), and (2) further summarize the English translated test samples. For (1), the cost is $0.00015 per OOD sample. For (2), an additional cost of $0.00007 is required per OOD sample.[3] The reported costs per test sample are taken as the average among all OOD test sets and non-English languages.

In conclusion, our *summarization* strategy costs $0.51 to summarize all 1.7K training samples, and $0.00022 (=(1)+(2)) per inference. Thus, for the same cost of employing human workers for CAD creation (≈ $1,105), our *summarization* strategy enables inference for 5M test samples. Note, however, that the best overall performance of classifiers augmented with CAD are achieved for *translate-test*. Therefore, if we also account for translation costs of the CAD-augmented classifiers, our *summarization* method can perform inference for 15M test samples for the same cost as employing human workers for CAD creation. This demonstrates the cost-effectiveness of our *summarization* approach when scaled up to 5M test samples as compared to

---

[3]Summarizing OOD test samples is less costly than summarizing IMDb training samples due to the test samples comprising fewer tokens.

zero-shot +CAD, and up to 15M when compared to *translate-test* +CAD. For future work, exploring open-source LLMs -or translation and summarization models could prove valuable for reducing inference costs.

5 Conclusions

We explored the generalization of zero-shot cross-lingual transfer to out-of-distribution (OOD) test data, considering both *language* and *domain* shifts. Our experiments on binary sentiment classification with pre-trained multilingual models LaBSE, mBERT, and XLM-R finetuned on English IMDb movie reviews and evaluated on non-English test samples comprising *Amazon* product reviews, *Restaurant* feedback, and *Tweets*, demonstrate that model performance substantially degrades, aligning with previous OOD generalization studies in a monolingual English setting. We also found that mBERT and XLM-R suffer more from performance reduction on OOD in non-English languages compared to English OOD degradation, while LaBSE's generalization strongly depends on the OOD dataset. Our experiments with models finetuned on original data augmented with manually constructed English counterfactual (CAD) IMDb reviews show that CAD's OOD generalization gains observed in a monolingual English setting also translate well to a zero-shot cross-lingual setup. Finally, to avoid costly manually constructed counterfactuals, we propose two new data augmentation approaches for OOD generalization based on large language models: (i) *domain transfer*, and (ii) *summarization*. Models trained with data augmented by our *summarization* strategy, show substantial gains across all datasets and models, and on *Amazon* and *Restaurants* surpassing models either augmented with (i) manually constructed CAD (Kaushik et al., 2019), or (ii) state-of-the-art generated CORE counterfactuals (Dixit et al., 2022).

Limitations

Task domain: In this exploratory study, we only presented results for zero-shot cross-lingual binary sentiment classification. To investigate whether our findings generalize beyond binary classification, and to other non-classification tasks, further analysis is required. Nevertheless, as mentioned in §4.2, our data augmentation approaches scale better for classification tasks with more than two classes, since it only requires summarizing/transferring the

training samples of each class once, whereas it is unclear how to scale counterfactuals to a larger number of classes.

Automatically translated in-distribution test data: Since we followed a similar setup as Kaushik et al. (2019), our experiments used the IMDb movie reviews as in-distribution sentiment data. While the main focus in our study is on out-of-distribution generalization, the in-distribution test set was only provided in English. Hence, we used translation tools to automatically translate the English IMDb test set to the considered non-English languages. This may have caused annotation artifacts in the translated in-distribution tests, making it unclear how well the reported in-distribution results for non-English languages match real-world test data for non-English languages.

Translate-test based on a multilingual model: As our aim was to analyze the out-of-distribution generalization of multilingual models and compare their performance, we did not include results for the *translate-test* based on a monolingual English model. We believe that using such a monolingual model could further boost the accuracy of *translate-test*, as well as for our *summarization* and *domain transfer* strategies. However, we leave exploration thereof for future work.

Applicability to low-resource languages: The effectiveness of the *translate-test* and *translate-train* approaches are highly dependent on the accuracy of the adopted machine translation system. In this study, we used ChatGPT-turbo (v0301) as our translation tool, and found it to produce high-quality translations for all languages considered in our experiments, i.e., boosting OOD generalization compared to the *zero-shot* strategy. However, such machine translations systems may not work well for low-resource languages that lack high-quality translation data.

Ethics Statement

Since our data augmentation methods use LLMs to generate summaries or create domain-transferred training (and test) samples, any biases present in the data used to train these LLMs could be transferred to the augmented data. We should therefore be careful to ensure that these biases do not carry over when training models on the augmented data, to avoid models that could discriminate against and/or potentially be harmful to certain demographics.

References

Udit Arora, William Huang, and He He. 2021. Types of out-of-distribution texts and how to detect them. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10687–10701, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting machine translation for cross-lingual classification. *arXiv preprint arXiv:2305.14240*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020a. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020b. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Bo Zheng, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2022. XLM-E: Cross-lingual language model pre-training via ELECTRA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6170–6182, Dublin, Ireland. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Asa Cooper Stickland, Sailik Sengupta, Jason Krone, Saab Mansour, and He He. 2023. Robustification of multilingual language models to real-world noise in crosslingual zero-shot settings with robust contrastive pretraining. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1375–1391, Dubrovnik, Croatia. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Maarten De Raedt, Fréderic Godin, Chris Develder, and Thomas Demeester. 2022. Robustifying sentiment classification by maximally exploiting few counterfactuals. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11386–11400, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tanay Dixit, Bhargavi Paranjape, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. CORE: A retrieve-then-edit framework for counterfactual data generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2964–2984, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kevin Duh, Akinori Fujino, and Masaaki Nagata. 2011. Is machine translation ripe for cross-lingual sentiment classification? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 429–433, Portland, Oregon, USA. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Phillip Howard, Gadi Singer, Vasudev Lal, Yejin Choi, and Swabha Swayamdipta. 2022. NeuroCounterfactuals: Beyond minimal-edit counterfactuals for richer data augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5056–5072, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is ChatGPT a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Nitish Joshi, Xiang Pan, and He He. 2022. Are all spurious features in natural language alike? an analysis through a causal lens. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9804–9817, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anurag Katakkar, Clay H. Yoo, Weiqin Wang, Zachary Lipton, and Divyansh Kaushik. 2022. Practical benefits of feature feedback under distribution shift. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 346–355, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Divyansh Kaushik, Amrith Setlur, Eduard H Hovy, and Zachary Chase Lipton. 2020. Explaining the efficacy of counterfactually augmented data. In *International Conference on Learning Representations*.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Dong C Liu and Jorge Nocedal. 1989. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528.

Zihan Liu, Genta Indra Winata, Andrea Madotto, and Pascale Fung. 2021. Preserving cross-linguality of pre-trained models via continual learning. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 64–71, Online. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *The Journal Of Machine Learning Research*, 12:2825–2830.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of ChatGPT for machine translation. *arxiv preprint*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. 2008. *Dataset shift in machine learning*. Mit Press.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Lei Shi, Rada Mihalcea, and Mingjun Tian. 2010. Cross language text classification by model translation and semi-supervised learning. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1057–1067, Cambridge, MA. Association for Computational Linguistics.

Xiao Wang, Qin Liu, Tao Gui, Qi Zhang, Yicheng Zou, Xin Zhou, Jiacheng Ye, Yongxin Zhang, Rui Zheng, Zexiong Pang, Qinzhuo Wu, Zhengyan Li, Chong Zhang, Ruotian Ma, Zichu Fei, Ruijian Cai, Jun Zhao, Xingwu Hu, Zhiheng Yan, Yiding Tan, Yuan Hu, Qiyuan Bian, Zhihua Liu, Shan Qin, Bolin Zhu, Xiaoyu Xing, Jinlan Fu, Yue Zhang, Minlong Peng, Xiaoqing Zheng, Yaqian Zhou, Zhongyu Wei, Xipeng Qiu, and Xuanjing Huang. 2021. TextFlint: Unified multilingual robustness evaluation toolkit for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 347–355, Online. Association for Computational Linguistics.

Zhao Wang and Aron Culotta. 2020. Identifying spurious correlations for robust text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3431–3440, Online. Association for Computational Linguistics.

Zhao Wang and Aron Culotta. 2021. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14024–14031.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Hitomi Yanaka and Koji Mineshima. 2021. Assessing the generalization capacity of pre-trained language models through Japanese adversarial natural language inference. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 337–349, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Linyi Yang, Jiazheng Li, Padraig Cunningham, Yue Zhang, Barry Smyth, and Ruihai Dong. 2021. Exploring the efficacy of automatically generated counterfactuals for sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 306–316, Online. Association for Computational Linguistics.

Tao Yu and Shafiq Joty. 2021. Effective fine-tuning methods for cross-lingual adaptation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8492–8501, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bo Zheng, Li Dong, Shaohan Huang, Wenhui Wang, Zewen Chi, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. 2021. Consistency regularization for cross-lingual fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3403–3417, Online. Association for Computational Linguistics.

Dataset	#Test												
	EN	DE	NL	FR	ES	IT	PT	TU	RU	JA	ZH	AR	HI
AMAZON	4,000	4,000	-	4,000	4,000	-	-	-	-	4,000	4,000	-	-
TWEETS	580	580	-	580	580	580	580	-	-	-	-	580	580
RESTAURANTS	980	-	960	1,268	760	-	-	780	1,012	-	-	-	-

Table 5: **Out-of-distribution** dataset statistics.

IMDB (EN)	#Train	#Val	#Test
Original	1,707	245	488
CAD	1,707	245	-

Table 6: **In-distribution** dataset statistics.

A Appendix

Datasets: Tables 5 and 6 summarize respectively the number of *out-of-distribution* test samples and the number of train, validation and test *in-distribution* test samples. Note that the number of samples for *translate-train* and *translate-test* exactly match those shown in the tables.

Prompts: Figs. 2 and 3 show our adopted prompts for instructing ChatGPT-turbo to translate (i) non-English out-of-distribution test samples into English for *translate-test*, and (ii) English in-distribution English training and validation samples into non-English for *translate-train*.

Detailed ID and OOD results per language: The in-distribution and out-of-distribution results per language are presented in Tables 7 and 8. As mentioned in §4.1, the *translate-test* in-distribution scores are not included for non-English languages. This is because these test sets are automatically translated versions of the original English test set. Including *translate-test* scores would require translating the already translated test samples back to English, which would evaluate the quality of back-translation rather than the *translate-test* performance itself. In our pilot experiments, we observed that the backtranslation quality was quite high. As such, small differences in accuracy between the performance of *translate-test* and the model performance on the original English test set appeared overly optimistic. Hence, we opted to exclude them.

Translate-test	**Translate-train**
Translate from {Language} to English. {**Language**}: {test sample} **English:**	Translate from English to {Language}. **English:** {train sample} {**Language**}:
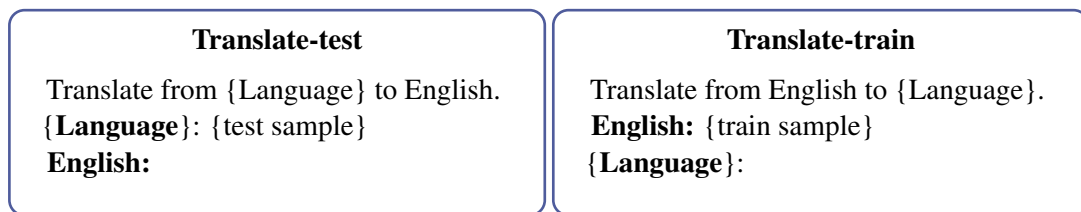
Fig. 2: **Translation prompts** for `ChatGPT-turbo (v0301)`.

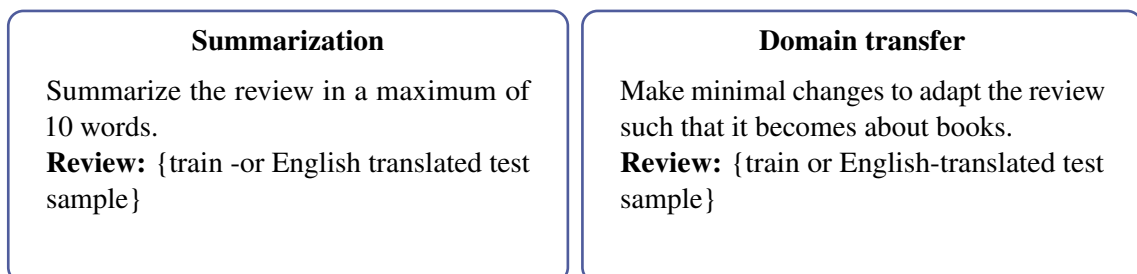| **Summarization**	**Domain transfer**
Summarize the review in a maximum of 10 words. **Review:** {train -or English translated test sample}	Make minimal changes to adapt the review such that it becomes about books. **Review:** {train or English-translated test sample}

Fig. 3: **Data augmentation prompts** for `ChatGPT-turbo (v0301)`. **Left:** *Summarization* prompt. **Right:** *Domain transfer* prompt.

IMDB

LaBSE

Method	EN	DE	NL	FR	ES	IT	PT	TU	RU	AR	HI	JA	ZH
Original only													
- ZSHOT	85.0	85.3	86.0	85.9	86.1	85.4	85.5	83.5	85.1	85.2	81.2	83.5	86.0
- TTRAIN	85.0	86.0	87.0	84.5	87.1	85.4	86.9	83.0	86.5	85.0	81.8	83.9	85.6
Original & CAD (Kaushik et al., 2019)													
- ZSHOT	81.4	82.0	80.1	82.0	82.6	81.6	81.6	80.5	80.1	79.3	80.1	80.3	79.5
- TTRAIN	81.4	83.0	80.7	82.4	83.0	81.8	83.8	82.0	80.7	78.7	78.7	80.7	79.1
Original & CORE (Dixit et al., 2022)													
- ZSHOT	80.1	77.9	80.3	79.3	81.4	79.3	78.7	79.1	78.3	79.9	75.4	79.5	79.1
Domain transfer (ours)													
- ZSHOT♠	83.3	84.5	84.5	84.4	86.0	85.4	85.5	82.3	83.8	84.6	79.0	82.7	83.3
+TRANS.	85.5	-	-	-	-	-	-	-	-	-	-	-	-
Summarization (ours)													
- ZSHOT♠	83.6	84.0	85.9	84.8	85.0	84.0	86.1	82.4	84.2	84.8	80.9	85.7	83.6
+SUM.	86.7	-	-	-	-	-	-	-	-	-	-	-	-

mBERT

Method	EN	DE	NL	FR	ES	IT	PT	TU	RU	AR	HI	JA	ZH
Original only													
- ZSHOT	89.5	84.0	77.8	84.2	86.9	83.4	83.2	76.1	80.0	75.2	72.2	81.9	84.8
- TTRAIN	-	87.2	89.1	89.1	90.2	88.7	88.8	87.4	87.8	84.1	81.9	87.1	88.5
Original & CAD (Kaushik et al., 2019)													
- ZSHOT	86.3	82.8	75.8	82.2	83.6	79.4	79.7	72.3	78.5	70.1	69.1	78.9	84.5
- TTRAIN	-	86.0	86.6	86.8	87.6	87.0	86.7	84.5	86.1	83.2	78.8	86.9	87.0
Original & CORE (Dixit et al., 2022)													
- ZSHOT	84.5	79.7	73.0	80.6	78.2	77.4	77.7	70.1	74.7	66.5	65.0	75.6	80.3
Domain transfer (ours)													
- ZSHOT♠	86.7	82.9	76.7	84.1	84.3	82.0	82.0	75.8	77.7	74.3	71.1	79.3	84.4
+TRANS.	87.8	-	-	-	-	-	-	-	-	-	-	-	-
Summarization (ours)													
- ZSHOT♠	87.2	83.1	74.4	82.3	84.4	81.1	82.3	74.4	77.3	73.6	71.0	80.9	82.9
+SUM.	88.2	-	-	-	-	-	-	-	-	-	-	-	-

XLM-R

Method	EN	DE	NL	FR	ES	IT	PT	TU	RU	AR	HI	JA	ZH
Original only													
- ZSHOT	92.4	90.4	90.9	89.9	89.8	89.5	90.7	88.5	89.4	84.7	82.3	85.4	89.6
- TTRAIN	-	91.4	92.2	91.7	91.6	91.3	91.8	91.0	90.9	89.2	86.4	89.2	91.1
Original & CAD (Kaushik et al., 2019)													
- ZSHOT	90.4	88.1	88.0	88.1	87.8	87.0	87.4	86.8	86.8	81.6	82.2	85.9	88.3
- TTRAIN	-	88.9	88.5	89.8	89.7	89.3	89.8	89.2	88.9	88.2	85.7	87.9	88.8
Original & CORE (Dixit et al., 2022)													
- ZSHOT	88.1	86.9	87.5	87.2	87.5	87.2	86.7	86.1	87.0	83.6	82.4	85.4	85.9
Domain transfer (ours)													
- ZSHOT♠	90.5	89.6	89.9	89.2	89.3	88.4	89.8	87.5	88.7	83.6	82.8	86.7	89.2
+TRANS.	91.1	-	-	-	-	-	-	-	-	-	-	-	-
Summarization (ours)													
- ZSHOT♠	91.4	89.5	90.3	89.9	89.5	89.1	88.8	88.3	88.8	83.6	81.7	85.1	89.7
+SUM.	89.9	-	-	-	-	-	-	-	-	-	-	-	-

Table 7: **In-distribution** accuracies for LaBSE, mBERT, and XLM-R. ♠: ablations. Scores for *translate-test* are omitted due to the English ID test sets being translated into the respective non-English languages. Note, for English, TTRAIN does not involve any translation, hence its EN scores are equivalent to ZSHOT.

LaBSE

Method	EN	DE	FR	ES	JA	ZH	AVG.	EN	NL	FR	ES	RU	TU	AVG.	EN	DE	FR	ES	AR	HI	IT	PT	AVG.
			IMDB → AMAZON							IMDB → RESTAURANTS								IMDB → TWEETS					
Original only																							
- ZSHOT	66.3	75.3	70.6	70.0	69.5	73.9	71.9	72.7	75.0	73.6	74.9	74.6	72.6	74.1	76.3	70.5	67.6	72.3	60.3	61.6	72.1	70.2	67.8
- TTRAIN	66.3	71.6	74.2	72.5	77.0	74.8	74.0	72.7	76.2	77.5	76.7	76.1	75.4	76.4	76.3	66.3	67.1	70.1	56.1	62.3	69.3	71.1	66.0
- TTEST	66.3	70.0	67.6	66.4	66.4	67.5	67.6	72.7	75.6	72.5	73.8	70.4	73.3	73.1	76.3	70.6	64.8	72.4	60.6	67.7	73.3	72.4	68.8
Original & CAD (Kaushik et al., 2019)																							
- ZSHOT	81.2	85.4	85.3	85.0	80.4	78.5	82.9	84.7	86.8	86.4	88.6	83.5	83.3	85.7	81.7	76.6	72.2	80.3	71.6	67.8	75.2	77.8	74.5
- TTRAIN	81.2	85.0	83.5	84.5	80.0	78.7	82.3	84.7	84.4	81.6	88.6	80.8	81.5	83.4	81.7	77.6	72.6	81.0	67.4	64.7	74.8	77.8	73.7
- TTEST	81.2	84.4	84.9	83.7	79.8	79.3	82.4	84.7	88.0	86.4	87.9	82.2	85.0	85.9	**81.7**	79.8	71.7	81.6	71.0	74.8	75.0	79.3	**76.2**
Original & CORE (Dixit et al., 2022)																							
- ZSHOT	81.0	84.8	84.2	84.6	80.2	76.3	82.0	85.0	84.6	85.4	88.7	84.7	81.2	84.9	77.4	71.2	67.6	76.0	66.9	64.0	75.7	76.0	71.1
- TTEST	81.0	84.4	83.9	83.2	79.8	77.1	81.7	85.0	86.5	85.3	89.5	84.1	86.2	86.3	77.4	77.9	69.8	80.5	65.3	72.8	76.0	77.8	74.3
Original & Domain transfer (ours)																							
- ZSHOT♠	76.0	82.5	79.5	79.1	77.7	75.8	78.9	81.4	83.1	81.2	82.6	82.0	78.4	81.5	80.9	72.3	68.1	76.2	65.2	64.7	74.8	74.3	70.8
+TTEST♠	76.0	80.6	79.8	79.2	76.6	75.6	78.4	81.4	84.4	82.8	81.6	80.7	81.6	82.2	80.9	72.7	69.1	75.4	66.3	74.1	74.3	74.3	72.3
+TRAN.	81.7	83.6	83.7	83.0	81.1	78.0	81.9	84.1	85.9	84.2	85.2	83.1	82.1	84.1	72.3	69.1	62.0	74.9	62.6	71.0	71.1	76.5	69.6
Original & Summarization (ours)																							
- ZSHOT♠	77.1	82.5	80.7	81.2	77.8	76.2	79.7	83.6	85.2	83.7	84.7	84.2	80.5	83.7	81.9	73.4	70.9	77.9	65.0	66.2	75.5	74.0	71.8
+TTEST♠	77.1	81.1	80.4	80.2	76.6	76.1	78.9	83.6	86.7	83.5	83.0	84.1	82.6	84.0	81.9	74.7	69.3	77.6	68.1	75.3	73.1	73.4	73.1
+SUM.	**86.2**↑	**86.3**↑	**87.6**↑	**87.5**↑	**82.6**↑	**79.7**↑	**84.7**↑	**91.6**↑	**89.5**↑	**89.1**↑	**89.5**↑	**89.2**↑	**86.5**↑	**88.8**↑	76.6↓	74.7↓	73.3↑	81.0↑	70.2↑	74.3↑	71.7↓	73.1↓	74.0

mBERT

Method	EN	DE	FR	ES	JA	ZH	AVG.	EN	NL	FR	ES	RU	TU	AVG.	EN	DE	FR	ES	AR	HI	IT	PT	AVG.
			IMDB → AMAZON							IMDB → RESTAURANTS								IMDB → TWEETS					
Original only																							
- ZSHOT	79.3	72.2	73.1	74.5	71.6	69.8	72.2	80.2	69.8	68.8	72.2	73.3	64.1	69.6	75.9	60.5	66.2	64.0	61.4	58.3	65.8	63.4	62.8
- TTRAIN	79.3	72.6	77.6	76.8	71.0	69.4	73.5	80.2	75.4	75.3	78.4	76.8	66.7	74.5	75.9	57.7	69.5	66.7	64.3	52.6	66.9	62.4	62.9
- TTEST	79.3	78.9	79.8	80.3	75.2	74.6	77.8	80.2	79.4	78.2	82.2	79.2	75.4	78.9	75.9	67.4	67.1	73.8	68.3	72.0	73.5	75.7	71.1
Original & CAD (Kaushik et al., 2019)																							
- ZSHOT	81.7	76.0	76.0	77.7	73.1	71.9	74.9	81.8	68.6	71.2	77.1	72.7	64.9	70.9	79.0	64.3	74.9	68.9	69.0	61.0	68.3	64.2	67.2
- TTRAIN	81.7	79.0	80.5	80.4	76.5	74.5	78.2	81.8	75.9	76.6	81.5	74.5	69.9	75.7	79.0	64.9	75.6	71.2	65.0	54.8	70.4	66.7	66.9
- TTEST	81.7	82.7	83.3	83.2	79.4	77.4	**81.2**	81.8	81.4	81.5	83.9	79.1	79.9	81.2	**79.0**	73.9	74.1	78.3	75.5	73.3	72.6	77.5	**75.0**
Original & CORE (Dixit et al., 2022)																							
- ZSHOT	80.2	74.3	75.3	77.2	73.6	70.2	74.1	80.4	65.3	72.1	75.3	71.2	63.9	69.6	73.6	59.4	72.0	70.3	62.7	59.3	68.3	61.5	64.8
- TTEST	80.7	81.3	80.4	82.5	79.2	76.3	79.9	80.4	79.2	79.7	82.9	78.5	79.4	79.9	73.6	70.6	70.0	77.9	73.0	70.1	73.0	75.1	72.8
Original & Domain transfer (ours)																							
- ZSHOT♠	79.6	73.2	74.8	76.4	72.3	71.0	73.5	80.2	70.8	70.4	73.6	73.1	63.9	70.4	78.1	60.5	69.0	63.8	62.6	58.8	66.0	64.9	63.7
+TTEST♠	79.6	80.3	81.0	80.8	76.8	75.8	78.9	80.2	78.2	77.8	80.9	78.3	76.4	78.3	78.1	68.8	68.2	73.9	72.3	72.7	72.9	75.6	72.1
+TRAN.	81.3	81.4	81.6	81.9	79.5	77.0	80.3	83.3	81.0	80.4	83.6	80.4	79.6	81.0	72.4	67.5	66.2	72.2	65.1	70.6	70.3	75.9	69.7
Original & Summarization (ours)																							
- ZSHOT♠	80.7	74.1	75.4	77.1	72.3	69.2	73.6	82.4	71.1	72.4	76.8	75.5	66.6	72.5	77.8	60.6	67.1	66.8	61.5	59.5	65.3	63.8	63.5
+TTEST♠	80.7	81.5	82.3	82.4	76.7	75.3	79.6	82.4	80.0	80.2	83.0	79.7	77.3	80.0	77.8	70.1	67.5	75.6	70.8	72.4	71.4	76.2	72.0
+SUM.	81.0↑	82.3↑	83.6↑	84.0↑	78.1↓	77.8↑	**81.2**↑	**87.3**↑	**84.6**↑	**85.5**↑	**87.3**↑	**83.6**↑	**80.4**↑	**84.3**↑	74.3↑	73.0↑	72.1↑	76.9↓	76.1↑	71.6↑	69.9↓	77.0↑	73.8

XLM-R

Method	EN	DE	FR	ES	JA	ZH	AVG.	EN	NL	FR	ES	RU	TU	AVG.	EN	DE	FR	ES	AR	HI	IT	PT	AVG.
			IMDB → AMAZON							IMDB → RESTAURANTS								IMDB → TWEETS					
Original only																							
- ZSHOT	86.3	86.7	85.0	83.9	86.9	82.4	85.0	86.0	81.2	78.6	80.7	81.9	73.4	79.2	84.3	75.5	66.0	72.9	68.4	63.6	70.0	68.0	69.2
- TTRAIN	86.3	86.9	86.5	88.2	87.1	81.4	86.0	86.0	85.9	79.2	86.7	85.5	77.9	83.0	84.3	75.4	66.9	82.1	71.3	66.6	71.6	73.6	72.5
- TTEST	86.3	86.7	87.6	88.5	85.5	81.4	85.6	86.0	81.6	82.2	86.0	79.8	79.8	81.5	84.3	76.6	67.5	77.3	70.2	70.0	69.4	71.2	71.7
Original & CAD (Kaushik et al., 2019)																							
- ZSHOT	87.0	86.9	86.3	86.3	86.2	82.7	85.7	87.5	82.5	81.8	83.3	82.1	79.6	81.9	86.7	77.6	76.1	82.7	78.2	67.9	74.2	74.6	75.9
- TTRAIN	87.0	87.6	87.8	88.4	87.0	81.0	86.4	87.5	85.3	83.5	87.6	85.0	81.7	84.6	86.7	80.4	75.1	85.0	79.6	68.4	75.6	77.0	77.3
- TTEST	87.0	87.8	88.8	88.4	86.9	82.1	86.8	87.5	87.3	86.5	89.2	85.8	86.9	87.1	**86.7**	81.4	77.6	84.3	79.6	76.0	77.8	80.6	79.6
Original & CORE (Dixit et al., 2022)																							
- ZSHOT	86.8	88.1	87.7	88.9	86.9	81.6	87.0	89.7	88.8	87.2	90.4	89.1	81.9	87.5	83.9	75.7	79.4	82.9	80.9	67.8	79.9	78.8	77.9
- TTEST	86.8	88.4	89.0	89.0	87.6	81.1	87.0	89.7	89.2	89.0	91.2	88.0	88.1	89.1	83.9	81.1	77.6	86.2	82.2	75.4	79.6	81.2	**80.5**
Original & Domain transfer (ours)																							
- ZSHOT♠	86.4	86.9	85.5	84.6	87.1	82.0	85.2	85.4	80.1	79.2	81.7	82.3	74.4	79.5	85.2	75.7	69.2	75.6	70.6	65.6	71.1	69.7	71.1
+TTEST♠	86.4	88.1	89.0	88.0	87.5	81.7	86.9	85.4	84.0	83.4	85.7	83.0	83.7	84.0	85.2	78.4	71.8	80.4	74.9	73.8	73.5	74.7	75.4
+TRAN.	87.1	88.3	89.2	88.4	87.1	82.5	87.1	87.2	84.3	85.0	87.0	82.8	83.4	84.5	72.7	72.4	66.0	73.7	65.8	70.0	66.4	73.9	69.7
Original & Summarization (ours)																							
- ZSHOT♠	87.8	89.1	89.3	88.7	88.1	83.3	87.7	89.4	86.1	83.8	86.5	86.5	81.7	84.9	86.3	76.6	71.7	81.6	75.8	69.0	75.7	75.2	75.1
+TTEST♠	87.8	89.5	90.5	89.5	88.0	82.4	**88.0**	89.4	87.5	87.7	88.6	85.8	85.7	87.1	86.3	79.8	73.7	83.0	77.1	75.7	75.1	80.4	77.8
+SUM.	**87.8**↑	87.6↓	89.7↑	89.2↑	86.1↓	81.2↓	86.8↓	**92.8**↑	**91.0**↑	**90.1**↑	**91.8**↑	**89.5**↑	**88.8**↑	**90.2**↑	83.0↓	78.0↓	74.6↓	80.0↓	76.0↓	74.1↓	71.4↓	77.0↑	75.9

Table 8: **Out-of-distribution** accuracies for LaBSE, mBERT, and XLM-R. **Best** model in bold with the runner-up underlined. ♠: ablations. For English, TTRAIN and TTEST do not involve any translation, hence their EN scores are equivalent to ZSHOT. Highlighted rows show a 1-on-1 comparison between classifiers augmented with (i) our (*summarization*) strategy, and (ii) the state-of-the-art generated CORE counterfactuals.