More "Clever" than "Hans": Probing and Adversarial Training in Translationese Classification

Anonymous ACL submission

Abstract

Modern classifiers, especially neural networks, excel at leveraging subtle signals competing with many other signals in the data. When such noisy setups lead to accuracy rates of 90%+, as is for instance the case with current high-performance neural translationese classifiers, it raises concerns about potential spurious correlations in the data with the target labels – a phenomenon often referred to as "Clever Hans". Recent research has indeed found evidence that high-performance multi-lingual BERT translationese classifiers use spurious topic information in the form of location names, rather than just translationese signals. In this paper, we address two difficult open problems associated with confounding signals in translationese classification. First, we use probing to provide *direct* evidence that these classifiers learn and use spurious topic correlations, some potentially unknown. Second, we introduce adversarial training as a strategy to mitigate *any* spurious topic correlation, including those unknown apriori. We show the effectiveness of our approach on translationese classification using three multi-lingual models, two language pairs, and four translationese data sets, as well as on a non-translationese classification task: occupation classification.

1 Introduction

"Translationese" describes the systematic linguistic differences between originally authored, non-translated texts in a given language, and texts translated into the same language, in the same genre and style (Gellerstam, 1986). Translationese effects can manifest at all levels of linguistic representation including vocabulary, syntax, semantics, and discourse. Five factors have been identified in the literature as the primary causes of translationese: source language interference, over-adherence to target language norms, explicitation, implicitation, and simplification (Toury, 1980; Baker et al., 1993; Teich, 2012; Volansky et al., 2013).

In this paper, we focus on translationese classification, which refers to classifying text in a given language as Original (O) or Translated (T). Translationese signals can be very subtle, and are often competing with many other signals in the data including genre, style, topic, author, bias, and so on.

Current methods for translationese classification are mostly based on representation learning neural networks and large language models (Sominsky and Wintner, 2019; Pylypenko et al., 2021). These models perform exceedingly well on the task: Pylypenko et al. (2021) show that mBERT-based approaches (Devlin et al., 2019) perform much better than traditional manual feature engineering-based classification models (e.g. SVMs) by as much as 15-20 accuracy points.

Using Integrated Gradients (Sundararajan et al., 2017), (Amponsah-Kaakyire et al., 2022) found that mBERT uses some spurious topic-based correlations as short-cuts for translationese classification instead of only proper translationese signals, showing evidence of "Clever Hans" (Hernández-Orallo, 2019; Lapuschkin et al., 2019). Using a subset of the MPDE dataset (Amponsah-Kaakyire et al., 2021), containing half German original sentences, and half translations from Spanish to German, (Amponsah-Kaakyire et al., 2022) show that some of the top tokens mBERT uses for O/T classification are geographical location names: German-based location names for O and Spanish-based location names for T. These are clearly topic and not translationese signals. Recently, (Borah et al., 2023) presented an approach to quantify and mitigate the impact of “Clever Hans” in translationese classification. They focus on quantifying any potentially spurious but possibly unknown topic information in the data aligned with O/T target labels and, using unsupervised topic modeling techniques like LDA (Blei et al., 2001) and BERTopic (Grootendorst, 2022), and present the *topic floor*, average weighted alignment of documents in any of the topics with target classification labels, as a worst-case upper bound to which a classifier may exploit spurious topic information aligned with O/T target labels. The topic floor provides a spurious topic information-based baseline for classification models. (Borah et al., 2023) were only able to mitigate *known* topic signals in the form of location-named entities (NEs) (Amponsah-Kaakyire et al., 2022) by masking NEs in the training and test data.

From a methodological point of view, (Borah et al., 2023) provided only *indirect* evidence that mBERT uses topic signals in O/T classification by showing that in principle a mBERT classifier can learn LDA/BERTopic clusters as target labels and that masking known spurious topics such as location and other NEs in the data established by manual analysis of the output of attribution methods reduces O/T classification accuracy. Showing that if told to do so, mBERT can learn topics is not the same as showing that a mBERT O/T classifier is learning and using spurious topics as information in O/T classification all by itself. Furthermore, masking NEs in data changes the data (compared to the data without masking) and this may be the reason for reduced classification accuracy. In sum, even though it is likely that it does, evidence that mBERT uses Clever Hans in the form of spurious topic information in O/T classification provided in (Borah et al., 2023) is only *indirect* and at best *episodic* for location NEs. In addition, (Borah et al., 2023) can only address *known* spurious topic mitigation (geographic location and other NEs), even though spurious topics may be manifest in lexical, morpho-syntactic, and semantic information, and, more importantly, many more of the (unknown) topics established by LDA or mBERTtopic (over and above geographic location NEs) may carry spurious information with respect to the O/T target label classification.

Thus, two important questions regarding "Clever Hans" in translationese classification remain unanswered. First, there is no direct evidence that spurious topic signals in translationese data are actually learned and used by the target label O/T classifiers all by themselves. It is not clear whether the Clever Hans spurious "topic floor" posited by (Borah et al., 2023) is real in the sense that it is learned and used by the O/T classifiers. How can we obtain *direct* evidence for this? Second, how can we leverage unsupervised topic information from any LDA/BERTopic clusters to mitigate the impact of all potentially spurious unknown topic correlations with the desired target label classification, beyond the potentially problematic and limited scope masking of specific NEs for known spurious topic information in the data as established by manual analysis?

In this paper, we address the two questions using probing for the first and adversarial training for the second. We probe mBERT's encoder layers to test whether a high-performance mBERT-based O/T classifier can identify any potentially spurious topic correlations with target classifications captured by LDA, crucially unlike (Borah et al., 2023) without training BERT to learn topics. We compare three mBERTs - one fine-tuned on the MPDE translationese data with O/T labels as a translationese classifier, another fine-tuned on the same data but without O/T labels as a simple masked language model (MLM, and not a classifier), and an off-the-shelf mBERT model not fine-tuned on any further data. The logic is that if mBERT O/T classifiers learn and use spurious LDA topic correlations with O/T target labels, then probing mBERT O/T classifiers for LDA topics should yield higher accuracy/F1 than an MLM mBERT and an off-the-shelf mBERT. If this is observed, this constitutes *direct* evidence that an mBERT O/T classifier learns and uses spurious unknown topic information all by itself and that the "topic floor" proposed by (Borah et al., 2023) is real. For our second research question of extending Clever Hans mitigation beyond manually established known spurious correlations (such as location NEs), we utilize adversarial training to suppress any LDA-based potentially spurious unknown topic signals (whatever they are) in translationese classification. If this is successful, we should see adversarially-trained O/T classifiers with high O/T prediction accuracy and low LDA topic probing results. Our contributions include:

1. We use probing to prove that an mBERT O/T classifier learns and uses spurious topic correlations in the data as represented by LDA topics with the classification targets.

2. To the best of our knowledge, we are the first to show that domain adversarial training mitigates unknown Clever Hans signals across the board in the form of LDA topics while ensuring strong O/T classification performance.

3. We show that our LDA and adversarial training based "Clever Hans" mitigation generalizes to different languages (*de-es*, *de-en* and

en-fr), translationese data sets (MPDE, Ted, Political Commentary and Literature), models (`mBERT`, `XLM-R` and `mBART`) and tasks (translationese and occupation classification).

4. We compare our automatic version of LDA and adversarial training based Clever Hans mitigation with manual known spurious correlation mitigation based on attribution approaches (Wang et al., 2022) and (Borah et al., 2023).

Our probing and adversarial training based methodology to detect and mitigate 'Clever Hans' is depicted in Fig 1. Translationese classification is a paradigmatic instance of classification using weak signals competing with many other signals in the data. Our occupation classification experiment indicates that our approach is useful in other classification scenarios where the possibility of Clever Hans spurious correlations is at stake. [1]

2 Related Work

2.1 Clever Hans and Translationese Classification

Previous work on identifying Clever Hans in machine learning models includes (Lapuschkin et al., 2019), who introduced Layer-wise Relevance (LRA) to unmask Clever Hans behavior and understand what machines can learn. (Hernández-Orallo, 2019) presented limitations of LRA and issues with evaluating the performance of explainability methods. Unmasking and mitigating Clever Hans is an active area of research in XAI (Mohseni et al., 2021) but to date rarely addressed in NLP (Heinzerling, 2020; Niven and Kao, 2019; McCoy et al., 2019).

Early efforts in translationese classification focused on exploring hand-crafted, linguistically inspired features, manual feature engineering and classical supervised machine learning classifiers like Support Vector Machines (SVMs) and Decision Trees etc. (Ilisei et al., 2010; Baroni and Bernardini, 2005; Volansky et al., 2013; Rubino et al., 2016; Avner et al., 2016). (Rabinovich and Wintner, 2015) present unsupervised clustering-based approaches.

More recent research uses feature and representation learning approaches based on neural networks (Sominsky and Wintner, 2019; Pylypenko et al., 2021). (Pylypenko et al., 2021) show that representation learning-based approaches like mBERT outperform handcrafted and feature engineering approaches and this is due to feature learning rather than the classifiers (Amponsah-Kaakyire et al., 2022). Manually inspecting output from Explainable AI (XAI) approaches like IG (Sundararajan et al., 2017), (Amponsah-Kaakyire et al., 2022) found that mBERT exploits topic signals in the form of location names spuriously correlated with the O/T classification labels.

(Borah et al., 2023) use translationese classification as a setting to measure and mitigate Clever Hans in classification where signals are weak and competing with many other signals. The basic idea is simple: when, as is generally the case, topic signals in the data are unknown, they use unsupervised topic clustering, LDA and mBERTtopic, and measure overlap between the documents in a given topic and the target O/T classes, i.e. they count how many of the documents in a topic are O and how many are T. A topic that is perfectly aligned with O and T is either 100% O or 100% T, and a topic that is maximally undecided between O and T is 50% O and 50% T. The "topic floor" of the topics in a data set for classification targets O and T is then simply the weighted average of the alignments of the topics with O and T, defined using an alignment measure. The alignment of a topic top_i with O and T is given by

$$align_{O,T}(top_i) = \frac{\max(|top_i \cap O|, |top_i \cap T|)}{|top_i|}$$

The weighted average over n topics top is:

$$avg\_align_{O,T}(top) = \sum_{i=1}^{n} w_i \times align_{O,T}(top_i)$$

where a weight $w_i = |top_i|/|Data|$ is just the proportion of paragraphs in topic top_i divided by the total number of paragraphs in the data.

The "topic floor" is proposed as an upper bound of what spurious topic correlations may contribute to target classification results and as a baseline for translationese classifiers. They also show that their alignment measure is the same as *cluster purity* (Zhao, 2005), although cluster purity was not intended to quantify Clever Hans. (Borah et al., 2023) present Clever Hans mitigation, but only for known topic spurious correlations: they mask location NEs in the data as a known spurious topic correlation signal from the work of (Amponsah-Kaakyire et al.,

[1] Code and data at http://www.anonymized/for/review

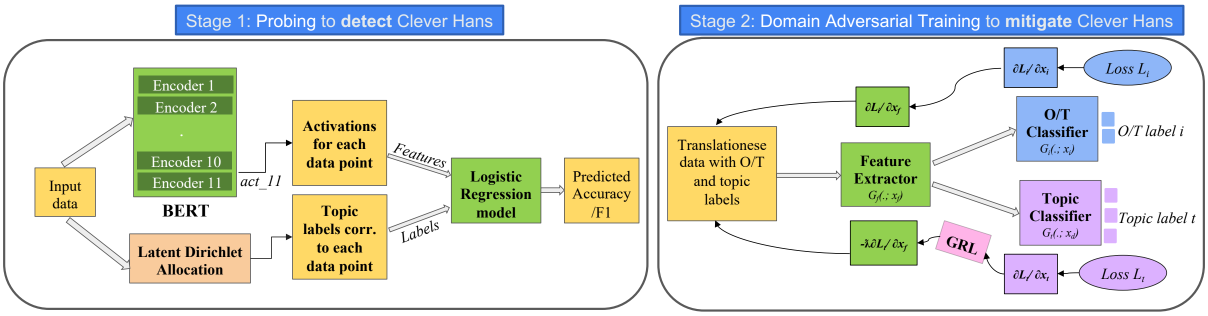

Figure 1: Probing and Adversarial Training Based Method

2022) and similar to (Dutta Chowdhury et al., 2022) also experiment with full PoS-based data masking. While the research presented in (Borah et al., 2023) is thought-provoking and makes an important contribution to an area that is understudied, namely quantifying Clever Hans in classification, it is lacking in two major respects: first, it only shows *indirectly* that topic-based spurious correlations are indeed learned and used by O/T classifiers by showing that mBERT can be trained (i.e. told) to learn LDA (and BERTopic) topics as target classes. This, however, is not the same as showing that an mBERT O/T classifier on its own accord (all by itself) picks up and uses any potentially spurious topic information as represented by LDA topics. Second, Clever Hans mitigation is only presented for manually established *known* spurious topic correlations and via data masking. This is both limiting and unfortunate as masking interferes with the data. In this paper, we address both shortcomings.

Finally, translationese is not just an important topic in basic linguistic research: many practical cross-lingual and multi-lingual applications are affected by translationese (Zhang and Toral, 2019; Singh et al., 2019; Artetxe et al., 2020; Clark et al., 2020b), and translationese is regarded as one of the final frontiers of high-resource machine translation (Freitag et al., 2019, 2020; Ni et al., 2022). The effects of translationese on machine translation (MT) training and evaluation were studied in many prior works (Kurokawa et al., 2009; Lembersky et al., 2012; Toral, 2019; Graham et al., 2019; Freitag et al., 2019, 2020). Further, building better translationese classifiers may lead to better MT training and evaluation and improved flagging of (human or machine) translated data while scraping the web (Thompson et al., 2024).

2.2 Probing

Early work on probing neural networks focused on extracting properties like gender, tense, and PoS using linear classifiers (Hupkes et al., 2018). Probing into inner layers of deep neural networks in NLP and Computer Vision was introduced by (Ettinger et al., 2016), (Shi et al., 2016) and (Alain and Bengio, 2018) respectively. In our paper, we use probing to find direct evidence that mBERT learns and uses spurious topic signals as provided by unsupervised topic modeling approaches (LDA) in translationese classification.

2.3 Domain-Adversarial Training

Domain Adversarial Training was introduced by (Ganin and Lempitsky, 2015) for domain adaptation where models learn features helpful for a target task but invariant to changes in the domain. Training is jointly performed with two objectives: one to predict target class labels and one to predict the domain and then regularising the former model to decrease the accuracy of the latter using a gradient reversal layer (GRL). The GRL multiplies the gradient by a certain negative constant during backpropagation, so that the loss of the domain classifier is maximized while training. (Stacey et al., 2020) used an ensemble adversarial technique to reduce *known* hypothesis-only bias in Natural Language Inference (NLI) due to spurious correlations between natural language utterances and their respective entailment classes. In our paper, we train our model adversarially to the topic classifier to reduce the use of any (and not just specific *known*) potentially spurious topic signals by mBERT in O/T target label classification. To the best of our knowledge, this is the first time adversarial training has been explored in *unknown* topic-based 'Clever Hans' mitigation in translationese classification.

We provide a more comprehensive analysis on previous and current work on detecting and miti-

N	Model	Accuracy	F1-score
2	[mBERT+OTD+CL]	0.531	0.635
	[mBERT+OTD]	0.515	0.544
	[mBERT]	0.521	0.556
3	[mBERT+OTD+CL]	0.412	0.563
	[mBERT+OTD]	0.392	0.457
	[mBERT]	0.389	0.468
5	[mBERT+OTD+CL]	0.327	0.483
	[mBERT+OTD]	0.313	0.414
	[mBERT]	0.318	0.424
10	[mBERT+OTD+CL]	0.242	0.387
	[mBERT+OTD]	0.224	0.320
	[mBERT]	0.229	0.331
20	[mBERT+OTD+CL]	0.164	0.275
	[mBERT+OTD]	0.149	0.227
	[mBERT]	0.153	0.243

Table 1: Probing results (last encoder layer as features) for LDA Topics = n topic prediction on the *de-es* dataset

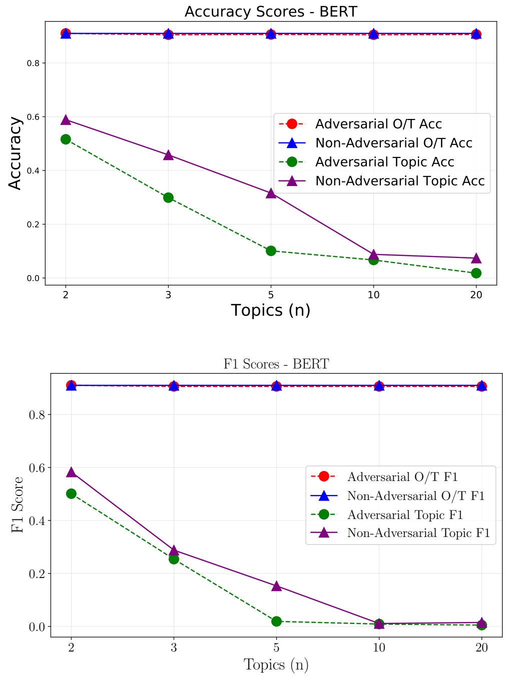

Figure 2: mBERT-Adv Acc and F1 on MPDE *de-es*

gating spurious correlations in Appendix E. We reproduce (Wang et al., 2022), a recent on mitigating spurious correlation (see Appendix F) in sentiment and occupation classification across datasets as it presents a competitive performance across datasets. (Wang et al., 2022) utilize attention scores to find top spurious tokens and mitigate them by masking the data. We found that, although mitigation using Cross Dataset Analysis proposed by (Wang et al., 2022) performs well in translationese classification, however, it does not effectively mitigate spurious topic signals as seen using our IG experiments (Table 22).

3 Data

We use the Multilingual Parallel Direct Europarl (MPDE) corpus (Amponsah-Kaakyire et al., 2021), which is a multilingual corpus with parallel data from the Europarl proceedings where the translation direction is known and all source data are originally authored (i.e. not already the result of translations from other languages themselves). We utilize two language pairs from the MPDE corpus: (1) *de-es*: a monolingual German dataset consisting of half German (DE) originals and half translations from Spanish (ES) to German and (2) *de-en*: a monolingual German dataset consisting of half German (DE) originals and half translations from English (EN) to German. Each of these datasets consists of 42k paragraphs, half of which are O and half are T. The average length (in terms of tokens) per training example (paragraph) is 80.

4 Unsupervised Clustering

We use Latent Dirichlet Allocation (LDA) (Blei et al., 2001) using (Rehurek and Sojka, 2011) as our unsupervised automatic topic modeling approach in our experiments. LDA performs topic modeling using two assumptions: (1) documents are a mixture of topics, and (2) topics are a mixture of words. Using these assumptions, LDA generates a document-term matrix that consists of documents as rows and terms or words corresponding to each document as columns. The parameters used in LDA are α and β, which determine the per-document topic distribution, and the per-topic word distribution respectively. We need to specify the number of topics n for LDA to generate. In our experiments we explore $n = 2, 3, 5, 10$, and 20, as these consistently show high topic floor scores in the range $[0.55, 0.60]$ (Borah et al., 2023). After performing LDA, we assign each data point in our dataset to the topic to which it belongs with the highest probability. We use the topics as labels for our probing and adversarial training experiments.

5 Probing for Topics in O/T Classification

5.1 Probing Experiment Design

Below, we present our probing-based approach to show whether a high-performance mBERT-based

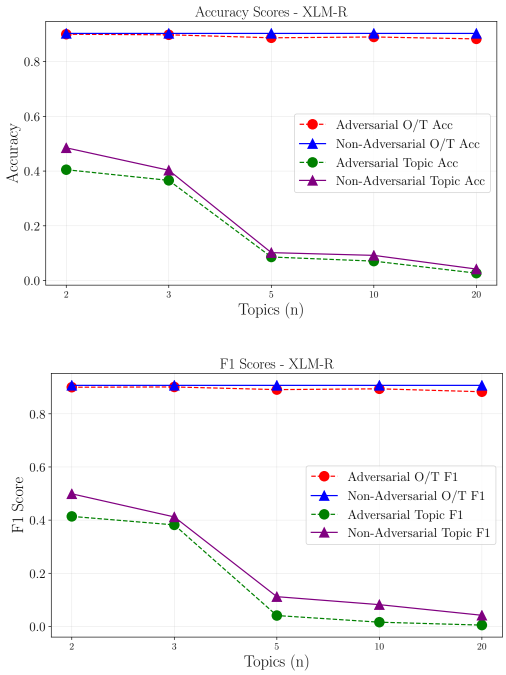

Figure 3: XLM-R-Adv Acc and F1 on MPDE *de-es*

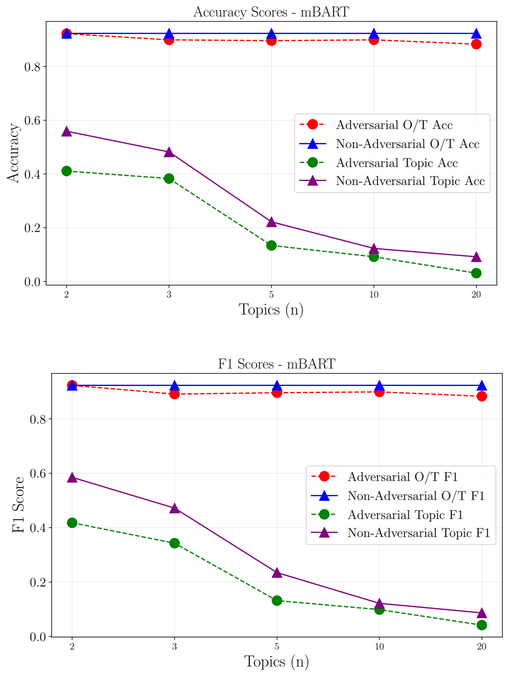

Figure 4: mBART-Adv Acc and F1 on MPDE *de-es*

translationese classifier learns to use spurious correlations in the form of LDA-based topics. We probe three mBERTs for topic classification:

1. **[mBERT+OTD+CL]**: a BERTforSequenceClassification fine-tuned on MPDE translationese data with original/translated labels for O/T classification.

2. **[mBERT+OTD]**: a BERTforMaskedLM fine-tuned on the same MPDE data for a MLM task but without O/T classification.

3. **[mBERT]**: a BERTforSequenceClassification off-the-shelf, without any fine-tuning on MPDE or O/T classification.

Each of the mBERT models is pre-trained on the same data. The logic behind our experiment is: mBERT finetuned on O/T data and trained for O/T classification [mBERT+OTD+CL] will learn and use spurious topic information only if this information is useful to O/T classification. If this is the case, then this mBERT should exhibit better performance on LDA topic probes compared to a mBERT fine-tuned on the same O/T data with the regular MLM objective but not trained for O/T classification [mBERT+OTD] and better than a simple mBERT out of the box [mBERT] not fine-tuned at all on the O/T data.

We perform topic classification probing using mBERT's last layer activations as features and LDA topics as the target labels of a simple logistic regression probe. For topics, we take the clusters found by LDA, and assign each data point the topic it belongs to with the highest probability. We perform experiments by setting $n = 2, 3, 5, 10$, and 20. Training and hyperparameter details are provided in Appendix G.1.

5.2 Probing Results

To account for the stochastic nature of LDA, we perform probing experiments on three different runs of LDA and average the results. We keep the same seed for logistic regression across runs. Table 1 shows the probing results for all numbers of LDA topics n. Compared to [mBERT+OTD] and [mBERT], probing [mBERT+OTD+CL] yields the highest topic scores in terms of accuracy and, even more pronounced, F1 scores. This shows that O/T classification makes mBERT learn spurious topic information and that this does not happen (to the same extent) for mBERT finetuned on the same O/T data with just the MLM objective and without O/T classification and similarly for mBERT out of the box. Table 6 in Appendix A.1 shows the same trend for probing *de-en*.

MODEL	NON-ADVERSARIAL		ADVERSARIAL	
	Original	**Translated**	**Original**	**Translated**
mBERT	situations	entstand	ppm	italo
	.	virus	uks	domino
	ria	inti	andersson	##unta
	##lk	sagte	prosa	##inne
	##iet	entdeckte	monterrey	arequipa
	golden	gras	prvni	moliere
	sak	buts	##ibe	brachten
	turn	nicaragua	hang	and
	##emeb	rekord	##tero	##saka
	orange	bilbao	plastik	giorgio
XLM-R	Serie	inn	Visa	PAD
	:	Bali	PG	definition
	happening	Nicaragua	download	uit
	DDR	mete	!	tru
	TH	Hyper	Fro	elementar
	_kat	stie	_Pea	lav
	Geschmack	Amen	istic	2008
	igo	Paradox	Statistika	ember
	bestellen	schrieb	straff	st##adte
	plural	Colombia	Digital	site
mBART	_ble	Colombia	app	Tob
	_boy	_studier	so	assis
	entes	_Sanchez	inge	dad
	!	cio	_SEA	Anna
	_Schreib	GO	esse	tan
	_regenera	trop	72	nsic
	_back	Ecuador	dien	Inv
	_traditionell	_Mu	_Vis	Earth
	donner	ringe	_ros	ibi
	stop	’	_frustr	hw

Table 2: Top 10 tokens for non-adversarial and adversarial (n=2) models trained on *de-es* MPDE dataset

6 Adversarial Training vs. Clever Hans

6.1 Adversarial Training Experiment Design

We employ Adversarial Training to utilize the spurious topic signals as identified by the unsupervised automatic topic clustering methods to mitigate "Clever Hans" in translationese.

We take topic labels as adversarial data, and O/T translationese labels as clean data. While training the model, we minimize the loss for O/T signals, while maximizing the loss for the topic signals. Our goal is to improve O/T accuracy while minimizing topic accuracy. As a consequence, this should make our model blind to spurious topics and reduce "Clever Hans" identified by unsupervised topic modeling techniques for translationese classification. To show that results generalise to different architectures, we experiment with three models: `multilingual-mBERT` (as we used previously for probing), `XLM-R`, and `mBART`. Training and hyperparameter details are provided in Appendix G.2.

6.2 Adversarial Training Results

Results are averaged over five different random seeds, and displayed in Figs 2, 3 and 4 for `mBERT`, `XLM-R`, and `mBART` respectively on the MODE *de-es* dataset.

The figures show a comparison of O/T and topic accuracies and F1 for the adversarially and non-adversarially trained models. Results show that the accuracies and F1 scores for translationese classification are maintained at a high level while the topic accuracies and F1 scores are consistently reduced for the adversarial model for all n. This shows that adversarial training is able to mitigate unknown automatically established spurious topic correlations. The accuracy and F1 scores with confidence scores for all models are displayed in Tables 8, 10, and 11 in Appendix A.2.

Table 9 in Appendix A.2 displays the results for the *de-en* pair using mBERT and fully shows the expected pattern for both accuracy and F1 scores. Table 15 of Appendix C contains the results of adversarial training for three other translationese corpora.

7 Integrated Gradients and Topic Traces

7.1 Integrated Gradients Experiment Design

We use Integrated Gradients (IG) to compute the tokens that have the highest attribution scores during translationese classification of the test set, in a similar fashion as (Amponsah-Kaakyire et al., 2022; Borah et al., 2023). (Amponsah-Kaakyire et al., 2022) used IG attribution scores to show that mBERT uses some spurious location name topic signals for translationese classification. (Borah et al., 2023) used IG on the mBERT O/T model fine-tuned on NE-masked data to show that the number of location tokens in the top tokens was reduced, thus resulting in some mitigation of ‘Clever Hans’. In our work, we use IG to compute the top tokens used by the three adversarial models[2] to capture known specific Clever hans as in location NEs in translationese classification.

7.2 IG Results

Table 2 shows the top 10 tokens with the highest IG attribution scores used by the adversarial and non-adversarial models for the O and T-test sets for the MPDE *de-es* dataset by the three models. For `mBERT`, there is only one South American Spanish language location token among the top tokens for the adversarial case - *arequipa* in the translated class. By contrast, in the non-adversarial case, as presented by (Amponsah-Kaakyire et al.,

[2]We use the encoder embeddings to compute IG results for mBART

Setting	O/T Acc, CI	O/T F1, CI	Topic Acc, CI	Topic F1, CI
Non-adversarial	0.975 [0.96,0.96]	0.961 [0.96,0.98]	0.518 [0.50, 0.51]	0.492 [0.49, 0.51]
Adversarial	0.970 [0.96,0.98]	0.954 [0.95,0.96]	0.459 [0.44,0.46]	0.430 [0.42,0.44]

Table 3: Adversarial (n=2) and Non-adversarial results on Occupation Classification Task. We highlight the lower topic accuracies and F1.

Non-adversarial		Adversarial	
Non-Surgeon	Surgeon	Non-Surgeon	Surgeon
herself	duren	concern	filed
wiki	bateau	underwent	museum
di	##lande	eligible	instant
databases	tn	band	soul
##virus	his	baseball	wikipedia

Table 4: Top 5 IG tokens: occupation classification task

2022)[3], there are several German location NEs in O (e.g. *##wald*, *stuttgart*) and Spanish in T (e.g., *Nicaragua*, *Bilbao*, *Colombia*). We find one location NE in the O for the adversarial model - *monterrey*, however, it is not a German-dominated area, hence this is not considered as a direct spurious correlation with the O set language. Table 2 shows similar trends for the other two models: `XLM-R` and `mBART`. Table 12 in Appendix A.3 provides the same trend for the *de-en* pair by `mBERT`. Table 13 provides similar trends by `XLM-R` and `mBART` on the MPDE *de-es* dataset. We also provide IG results for different translationese corpora in Table 17 of Appendix C.

8 Occupation Classification Task

To investigate whether our 'Clever Hans' mitigation approach generalizes to other classification tasks that involve subtle signals competing with many other signals but are not translationese, we run our experiments on another task: *occupation classification*. Using the dataset by (Pruthi et al., 2020), the task consists of English biographies of surgeons and non-surgeons (physicians) from (De-Arteaga et al., 2019). The training data consists of 17,629 biographies and the dev set consists of 2,519 samples. We utilize adversarially and non-adversarially trained `mBERT` on the occupation classification data for our experiments. Using IG, we then find the top tokens with the highest attribution score for occupation classification.

[3]We do not provide the full list here, please check Table 7 for the list of top 20 tokens with the highest IG attribution

Results. Table 3 shows that adversarial training on occupation classification reduces topic dependency while maintaining O/T classification performances. Table 4 shows the top IG tokens for the surgeon and non-surgeon classes. For the non-adversarial setting, we find pronouns like *herself* and *his* as top tokens for the non-surgeon and surgeon classes respectively. This shows a spurious correlation of gendered pronouns with occupations, indicating gender bias. With adversarial training, the top five tokens do not contain any gender-related information, mitigating the use of spurious correlations in occupation classification. We provide full performance (accuracy and F1 scores) and IG results for other *n* in adversarial and non-adversarial settings in Appendix D.

9 Conclusion

In this paper, we focused on an under-researched area: "Clever Hans", i.e., spurious correlations in the data with target classification labels, in the form of topic information in classification scenarios where target signals are weak and competing with many other signals in the data. We generalized previous work by (i) providing *direct* evidence using prompting that feature and representation learning-based neural classifiers learn and use spurious topic correlations in the data; and (ii) by showing that we can mitigate any *unknown* spurious topic correlation using adversarial training with LDA topic labels as adversarial targets in the classification. We showed this in translationese classification, a prototypical example of a classification setting where target signals are weak and competing with many other signals in the data. We showed that our approach generalises to three language pairs (de-es, de-en, en-es), three models (mBERT, XML-R and mBART) and a non-transationese task: occupation classification.

Future research includes zooming in on specific LDA topics that exhibit high alignment with target labels as well as exploring other topic modeling approaches.

10 Limitations

Our research on unknown spurious topics is based on LDA. If a topic is not in LDA, it cannot be probed nor mitigated by adversarial training. LDA requires us to set the number of topics n. We explore $n = 2, 3, 5, 10, 20$, based on findings by (Borah et al., 2023) that show consistently high topic floor scores for these settings. That said we should explore topic models other than LDA, e.g. BERT-topic (Grootendorst, 2022) etc.

11 Ethical Considerations

We experiment with three multi-lingual models: `mBERT`, `XLM-R`, and `mBART` trained on a variety of data, these models may contain harmful social biases and use them for translationese classification. As we see in the occupation classification task, explainability results using IG suggest that language models like BERT indeed use gender biases as spurious correlations.

Additionally, the translationese corpora may also contain biases related to culture and language, and historical and social biases.

References

Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes.

Kwabena Amponsah-Kaakyire, Daria Pylypenko, Cristina España-Bonet, and Josef van Genabith. 2021. Do not rely on relay translations: Multilingual parallel direct Europarl. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 1–7, online. Association for Computational Linguistics.

Kwabena Amponsah-Kaakyire, Daria Pylypenko, Josef Genabith, and Cristina España-Bonet. 2022. Explaining translationese: why are neural classifiers better and what do they learn? In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 281–296, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. 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.

Ehud Alexander Avner, Noam Ordan, and Shuly Wintner. 2016. Identifying translationese at the word and sub-word level. *Digit. Scholarsh. Humanit.*, 31:30–54.

Mona Baker, Gill Francis, and Elena Tognini-Bonelli. 1993. *Text and technology: in honour of John Sinclair*. John Benjamins Publishing.

Marco Baroni and Silvia Bernardini. 2005. A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. *Literary and Linguistic Computing*, 21(3):259–274.

David Blei, Andrew Ng, and Michael Jordan. 2001. Latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

Angana Borah, Daria Pylypenko, Cristina España-Bonet, and Josef van Genabith. 2023. Measuring spurious correlation in classification: "clever hans" in translationese. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 196–206, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2020a. Learning to model and ignore dataset bias with mixed capacity ensembles. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3031–3045, Online. Association for Computational Linguistics.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020b. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. 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.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

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.

Koel Dutta Chowdhury, Rricha Jalota, Cristina España-Bonet, and Josef Genabith. 2022. Towards debiasing translation artifacts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3983–3991, Seattle, United States. Association for Computational Linguistics.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139, Berlin, Germany. Association for Computational Linguistics.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France. PMLR.

Martin Gellerstam. 1986. Translationese in swedish novels translated from english. *Translation studies in Scandinavia*, 1:88–95.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. Translationese in machine translation evaluation.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142, Hong Kong, China. Association for Computational Linguistics.

Benjamin Heinzerling. 2020. Nlp's clever hans moment has arrived. *Journal of Cognitive Science*, 21(1).

José Hernández-Orallo. 2019. Gazing into clever hans machines. *Nature Machine Intelligence*, 1(4):172–173.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. 61(1).

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. 2010. Identification of translationese: A machine learning approach. In *International conference on intelligent text processing and computational linguistics*, pages 503–511. Springer.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. In *Proceedings of Machine Translation Summit XII: Papers*, Ottawa, Canada.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1–8.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Adapting translation models to translationese improves SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 255–265, Avignon, France. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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.

Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Trans. Interact. Intell. Syst.*, 11(3–4).

Jingwei Ni, Zhijing Jin, Markus Freitag, Mrinmaya Sachan, and Bernhard Schölkopf. 2022. Original or translated? a causal analysis of the impact of translationese on machine translation performance. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5303–5320, Seattle, United States. Association for Computational Linguistics.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of* the Association for Computational Linguistics, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

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

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793, Online. Association for Computational Linguistics.

Daria Pylypenko, Kwabena Amponsah-Kaakyire, Koel Dutta Chowdhury, Josef van Genabith, and Cristina España-Bonet. 2021. Comparing feature-engineering and feature-learning approaches for multilingual translationese classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8596–8611, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ella Rabinovich and Shuly Wintner. 2015. Unsupervised Identification of Translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432.

Ella Rabinovich, Shuly Wintner, and Ofek Luis Lewinsohn. 2018a. A parallel corpus of translationese. In *Computational Linguistics and Intelligent Text Processing*, pages 140–155, Cham. Springer International Publishing.

Ella Rabinovich, Shuly Wintner, and Ofek Luis Lewinsohn. 2018b. A parallel corpus of translationese. In *Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing 2016, Konya, Turkey, April 3–9, 2016, Revised Selected Papers, Part II 17*, pages 140–155. Springer.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Raphael Rubino, Ekaterina Lapshinova-Koltunski, and Josef van Genabith. 2016. Information density and quality estimation features as translationese indicators for human translation classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 960–970, San Diego, California. Association for Computational Linguistics.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. Xlda: Cross-lingual data augmentation for natural language inference and question answering.

Ilia Sominsky and Shuly Wintner. 2019. Automatic detection of translation direction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1131–1140, Varna, Bulgaria. INCOMA Ltd.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Sebastian Riedel, and Tim Rocktäschel. 2020. Avoiding the Hypothesis-Only Bias in Natural Language Inference via Ensemble Adversarial Training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8281–8291, Online. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Elke Teich. 2012. *Cross-Linguistic Variation in System and Text*. De Gruyter Mouton, Berlin, Boston.

Brian Thompson, Mehak Preet Dhaliwal, Peter Frisch, Tobias Domhan, and Marcello Federico. 2024. A shocking amount of the web is machine translated: Insights from multi-way parallelism.

Antonio Toral. 2019. Post-editese: an exacerbated translationese. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 273–281, Dublin, Ireland. European Association for Machine Translation.

Gideon Toury. 1980. *In search of a theory of translation*. Porter Institute for Poetics and Semiotics, Tel Aviv University.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2013. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. 2022. Identifying and mitigating spurious correlations for improving robustness in NLP models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1719–1729, Seattle, United States. Association for Computational Linguistics.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

Ying Zhao. 2005. *Criterion functions for document clustering*. Ph.D. thesis, USA. AAI3180039.

A mBERT Results for MPDE $de - es$ and $de - en$ language-pairs

Here, Table 5 presents the accuracy and F1 scores of mBERT fine-tuned on the MPDE translationese dataset for two language pairs (*de-es* (as discussed previously in the paper) and *de-en*). Note that the results are for non-adversarial mBERT which is not trained to suppress any topic signals.

A.1 Probing on two language pairs

In this section, we present the results of probing experiments on the *de-en* set. Table 6 displays the probing experiments for different n values. As observed in the de-es dataset in Section 5.2, we find that Model 1 finetuned on the O/T labels performs the best among all the models. The differences are more dominant in terms of F1 scores. The results are consistent for *de-en*, with topic label accuracies and F1 scores decreasing as we increase n.

A.2 Adversarial Training on two language pairs

We use the uncased version multilingual mBERT (Devlin et al., 2019) for our adversarial model by specifying two classification objectives: one for O/T classification and the other for topic label classification. We use a batch size of 16, a learning rate of $4 \cdot 10^{-6}$, and an Adam optimizer with epsilon $1 \cdot 10^{-5}$ to train our adversarial mBERT models for 4 epochs. For our LDA topic labels, we experiment with $n = 2$, 3, 5, 10, and 20.

Here, we present the results of adversarial training on different language pairs: *de-es* and *de-en* language-pairs. Tables 8 and 9 shows the results of adversarial training for different n values. We find the O/T accuracies and F1 scores are high whereas the topic accuracies and F1 scores are low and decrease with an increase in the value of n for both language pairs on MPDE.

A.3 Integrated Gradients on Two Language Pairs

We first present the top 20 tokens having the highest attribute scores utilized by mBERT for translationese classification on the *de-es* MPDE dataset in Table 7. The non-adversarial results are taken from (Amponsah-Kaakyire et al., 2022). We find a number of NEs in the non-adversarial results, namely *##wald*, *##stuttgart* in Original, and *nicaragua*, *bilbao* and *colombia* in Translated. With adversarial mitigation, we find that the number of NEs belonging to German or Spanish areas in O/T sets respectively are reduced.

Table 12 shows the results of IG given by the adversarial models for the two datasets for different values of n. The top 5 tokens with the highest average attribution for the test set data of each dataset are displayed. Although we see some location tokens, most of these are not related to the location where that language is spoken, i.e. we have *Venezuela*, *Pakistan*, and *Monterrey* in the original set, where German is not predominantly spoken.

B Adversarial Mitigation for 'Clever Hans' by different models

Apart from mBERT, we perform the same experiments using other multi-lingual language models like `XLM-R`(Conneau et al., 2020) and `mBART`(Liu et al., 2020). We first perform translationese classification on the MPDE dataset. Post that, we perform domain adversarial training to reduce topic dependency for translationese classification by the models. Here we extend the results from Section 6.2 in the paper.

Tables 10 and 11 present the results of accuracy, F1 scores and confidence scores for translationese classification on MPDE *de-es* dataset for `XLM-R` and `mBART` respectively. We find that adversarial training leads to almost similar translationese accuracies and F1 scores for `mBERT`, `XLM-R` and `mBART`, while reducing topic accuracies for all *n*. We further look into the top attribution tokens using IG to look for topic-related tokens for different *n*. Table 13 show that both XLM-R and mBART contain topic-related NEs that post adversarial training do not appear in the top 5 tokens used for translationese classification. This shows that adversarial training mitigates spurious topic signals utilized by different models for translationese classification. Our approach shows a robust performance for different multilingual models.

C Different Translationese Corpora

Here, we present our results on different translationese corpora, namely, TED talks, political commentary, and Literature corpora by (Rabinovich et al., 2018a). The corpora details are present in Table 15.

LANGUAGE PAIRS	O/T ACC, 95% CONFIDENCE SCORE	O/T F1, 95% CONFIDENCE SCORE
de-es	0.910, [0.90, 0.91]	0.910, [0.90, 0.92]
de-en	0.863, [0.85, 0.87]	0.872, [0.86, 0.88]

Table 5: mBERT fine-tuned on translationese data for O/T classification for two language-pairs from MPDE

N	MODEL	ACCURACY	F1-SCORE
2	[mBERT+OTD+CL]	0.564	0.667
	[mBERT+OTD]	0.556	0.606
	[mBERT]	0.561	0.659
3	[mBERT+OTD+CL]	0.409	0.538
	[mBERT+OTD]	0.397	0.483
	[mBERT]	0.397	0.479
5	[mBERT+OTD+CL]	0.306	0.434
	[mBERT+OTD]	0.290	0.379
	[mBERT]	0.295	0.381
10	[mBERT+OTD+CL]	0.254	0.405
	[mBERT+OTD]	0.252	0.393
	[mBERT]	0.253	0.392
20	[mBERT+OTD+CL]	0.142	0.236
	[mBERT+OTD]	0.129	0.199
	[mBERT]	0.134	0.200

Table 6: Probing results (last encoder layer as features) on the *de-en* datasets

ADVERSARIAL		NON-ADVERSARIAL	
Original	**Translated**	**Original**	**Translated**
ppm	italo	situations	entstand
uks	domino	.	virus
andersson	##unta	ria	inti
prosa	##inne	##lk	sagte
monterrey	arequipa	##iet	entdeckte
prvni	moliere	golden	gras
##ibe	brachten	sak	buts
hang	and	turn	nicaragua
##tero	##saka	##emeb	rekord
plastik	giorgio	orange	bilbao
domain	fut	hand	verfugte
##istes	olan	##wald	bol
diri	##rennen	1732	colombia
rasa	intra	dobe	nis
propose	uga	##pas	och
Stevenson	850	profits	vorkommen
versie	##izione	stuttgart	oecd
eingegliedert	boyko	soja	;
##ging	errichteten	r	erklarte
siehe	besuchte	ruth	clinton

Table 7: Top 20 tokens with highest attribution scores by IG for adversarial model ($n = 2$) and non-adversarial model fine-tuned on *de-es* dataset

We perform our adversarial training experiments on different translationese corpora, namely, TED talks, *political commentary*, and *Literature corpora* by (Rabinovich et al., 2018b). The Ted corpora is based on the subtitles of the TED talks delivered in English and translations to English of TEDx talks originally given in French. Therefore, it contains half English originals, and half translations from French. The political commentary corpus contains articles, commentary, and analysis on world affairs and international relations. These articles were collected from Project Syndicate[4]. It contains half German originals and half translations from English to German. The literature translationese corpus consists of literature classics (originals and translations) originating from the 18th to 20th centuries authored by English or German writers. It contains half German originals and half translations from English to German. We perform these experiments to understand the effectiveness of our spurious correlation mitigation approach using adversarial training. We utilize `mBERT` for all experiments in this section.

In Table 14, we find that mBERT performs well on the *Literature* corpora for translationese classification. However, it performs poorly on the *Ted* and *Politics* corpora. The smaller sizes of these corpora may be attributed to these lower performances. After adversarial training (for n=2), we find the O/T accuracies and F1 do not decrease drastically from the non-adversarial model, while topic accuracies and F1 are reduced (as expected).

Table 17 displays the results of IG for non-adversarial and adversarial mBERT on different corpora. For Ted dataset, NEs like *bowie*, *robbins*, *clayton* in O which are some common English names and *richelieu*, an industry based in Montreál, where French is predominantly spoken in T; occur in the top 5 tokens with the highest attribution scores. However, for the adversarially trained model, we find one NE token: *prada*. For politics, we find NEs like *calcutta*, *barbosa*, *bogota*, *tibet* in the top tokens, not necessarily belonging to the regions the languages are spoken in. However, the number of NEs in the adversarially trained model is reduced. Finally, for the Literature dataset, we find tokens like *watt*, *timothy*, *westminster*, and *lancaster* in T, common NEs in England; and also

[4]http://www.project-syndicate.org

n	Adversarial		Non-Adversarial	
	O/T acc, F1 (95% CI F1)	Topic acc, F1 (95% CI F1)	O/T acc, F1 (95% CI F1)	Topic acc, F1 (95% CI F1)
2	0.910, 0.910 ([0.90, 0.92])	0.516, 0.501 ([0.49, 0.51])	0.910, 0.910 ([0.90, 0.92])	0.589, 0.583 ([0.57, 0.59])
3	0.905, 0.906 ([0.90, 0.91])	0.299, 0.254 ([0.25, 0.26])	0.910, 0.910 ([0.90, 0.92])	0.458, 0.288 ([0.28, 0.29])
5	0.906, 0.906 ([0.90, 0.91])	0.101, 0.019 ([0.01, 0.02])	0.910, 0.910 ([0.90, 0.92])	0.316, 0.153 ([0.15, 0.15])
10	0.905, 0.906 ([0.90, 0.91])	0.088, 0.018 ([0.01, 0.02])	0.910, 0.910 ([0.90, 0.92])	0.067, 0.011 ([0.01, 0.01])
20	0.906, 0.906 ([0.90, 0.91])	0.050, 0.005, ([0.00, 0.00])	0.910, 0.910 ([0.90, 0.92])	0.074, 0.015 ([0.01, 0.02])

Table 8: Adversarial and Non-Adversarial. O/T classification and topic label classification results on MPDE *de-es*. Lower topic accuracies and F1 are highlighted. Note the scores for O/T acc and F1 are constant across all n for the non-adversarial models since it is only fine-tuned for translationese classification and not adversarially "finetuned" against topic classification.

n	Adversarial		Non-Adversarial	
	O/T acc, F1 (95% CI F1)	Topic acc, F1 (95% CI F1)	O/T acc, F1 (95% CI F1)	Topic acc, F1 (95% CI F1)
2	0.905, 0.903 ([0.90, 0.91])	0.489, 0.490 ([0.48. 0.50])	0.863, 0.872 ([0.86, 0.88])	0.572, 0.575 ([0.57, 0.59])
3	0.897, 0.897 ([0.89, 0.90])	0.365, 0.332 ([0.32, 0.34])	0.863, 0.872 ([0.86, 0.88])	0.379, 0.344 ([0.34, 0.35])
5	0.901, 0.899 ([0.89, 0.90])	0.138, 0.082 ([0.08, 0.09])	0.863, 0.872 ([0.86, 0.88])	0.159, 0.084 ([0.08, 0.09])
10	0.902, 0.901 ([0.89, 0.91])	0.054, 0.006 ([0.01, 0.01])	0.863, 0.872 ([0.86, 0.88])	0.077, 0.022 ([0.02, 0.02])
20	0.904, 0.903 ([0.90, 0.91])	0.048, 0.005 ([0.00, 0.00])	0.863, 0.872 ([0.86, 0.88])	0.063, 0.015 ([0.01, 0.02])

Table 9: Adversarial and Non-Adversarial O/T classification and topic label classification results on MPDE *de-en*. Lower topic accuracies and F1 are highlighted

other NEs: *pascal*, *welch*. The adversarially trained has only two NEs: *warner* and *marianne*. Therefore, topic-related tokens reduce after adversarial training showing the efficiency of our methodology on corpora belonging to different domains in translationese.

D Results on another task: Occupation Classification

Here, we present the results of adversarially and non-adversarially trained mBERT trained for the occupation classification task (extending section 8 in the paper). In Table 18, we find that topic accuracies are reduced for different n in the adversarial setting (as expected). Our adversarially trained model is able to mitigate the influence of potentially spurious topical information in occupation classification.

In Table 16, we find that the named entities in different topics, and also gendered pronouns are very low (not pertaining to previously described gender bias where males were associated with surgeon class and females with non-surgeon class: *sister* is present in the 'surgeon' class), showing the effectiveness of our 'Clever Hans' mitigation approach.

E Comparison to other works in NLP

Here, we present how our work compares to other work in detecting and mitigating spurious correlations.

Table 19 shows different studies that focus on spurious correlation detection in NLP. Earlier work focused on known shortcuts, however, recent work has been focusing more on unknown shortcuts.

Table 20 shows studies focused on mitigating spurious correlations. Different approaches have been proposed, with just one other approach that focuses on adversarial mitigation (Stacey et al., 2020). They experimented with NLI, which does not directly involve subtle signals like translationese. Our approach applied domain adversarial training for translationese classification and occupation classification (which utilizes spurious correlations like gender bias, as seen before).

N	Adversarial		Non-Adversarial	
	O/T acc, F1 (95% CI F1)	Topic acc, F1 (95% CI F1)	O/T acc, F1 (95% CI F1)	Topic acc, F1 (95% CI F1)
2	0.900, 0.900 ([0.90, 0.91])	0.405, 0.414 ([0.41, 0.42])	0.903, 0.907 ([0.90, 0.91])	0.485, 0.499 ([0.49, 0.50])
3	0.898, 0.901 ([0.90, 0.91])	0.366, 0.382 ([0.37, 0.38])	0.903, 0.907 ([0.90, 0.91])	0.403, 0.412 ([0.41, 0.42])
5	0.887, 0.891 ([0.89, 0.91])	0.086, 0.041 ([0.04, 0.04])	0.903, 0.907 ([0.90, 0.91])	0.102, 0.112 ([0.11, 0.12])
10	0.890, 0.894 ([0.89, 0.89])	0.071, 0.016 ([0.01, 0.02])	0.903, 0.907 ([0.90, 0.91])	0.092, 0.082 ([0.08, 0.08])
20	0.883, 0.884 ([0.88, 0.89])	0.027, 0.005, ([0.00, 0.00])	0.903, 0.907 ([0.90, 0.91])	0.054, 0.042 ([0.014, 0.04])

Table 10: Adversarial and Non-Adversarial O/T classification and topic label classification by XLM-R on MPDE *de-es*. Lower topic accuracies and F1 are highlighted

N	Adversarial		Non-Adversarial	
	O/T acc, F1 (95% CI F1)	Topic acc, F1 (95% CI F1)	O/T acc, F1 (95% CI F1)	Topic acc, F1 (95% CI F1)
2	0.922, 0.923 ([0.92, 0.93])	0.411, 0.418 ([0.41, 0.42])	0.923, 0.924 ([0.92, 0.93])	0.485, 0.499 ([0.49, 0.50])
3	0.899, 0.892 ([0.89, 0.90])	0.383, 0.343 ([0.34, 0.34])	0.923, 0.924 ([0.92, 0.93])	0.403, 0.412 ([0.41, 0.41])
5	0.896, 0.896 ([0.89, 0.90])	0.134, 0.132 ([0.13, 0.13])	0.923, 0.924 ([0.92, 0.93])	0.102, 0.112 ([0.11, 0.12])
10	0.899, 0.868 ([0.86, 0.88])	0.092, 0.099 ([0.09, 0.10])	0.923, 0.924 ([0.92, 0.93])	0.092, 0.082 ([0.08, 0.08])
20	0.883, 0.889 ([0.88, 0.89])	0.031, 0.042 ([0.03, 0.04])	0.923, 0.924 ([0.92, 0.93])	0.054, 0.042 ([0.04, 0.04])

Table 11: Adversarial and Non-Adversarial O/T classification and topic label classification by mBART on MPDE *de-es*. Lower topic accuracies and F1 are highlighted

N	DE-ES		DE-EN	
	Original	**Translated**	**Original**	**Translated**
2	ppm uks andersson prosa moonterrey	italo domino ##unta ##inne arequipa	acta unterstutzte ##oster ##ging asean	osterreichs parole workshops ungern !
3	• β stamme tras fet	fue often widmete kraftwerk kirche vendee	! nordlich ##sstraße ##ival ##ke	 thessaloniki ansonsten willy alfonso
5	started heading angegeben ernannt ##gemeinde	gerne mochte colombia ##indi bitter	nochmals legales revanche ##lasse ##hier	q schweizer mochte ##poru vieira
10	mochte ##ohe tunis altar pea	veroffentlichte widmete berichtet gelangte ##tierte	determiner bible skinner physik venezuela	quei cork ##shire mosaik barone
20	venezuela pakistan ##ids italia oost	##rennen ##list ##verk hast quebec	thuringen beaten philippine ##beni pohja	##mble roy angels romBERT earl

Table 12: Top 5 tokens for adversarial model trained on *de-es* and *de-en* datasets for different n

N	XLM-R		mBART	
	Original	**Translated**	**Original**	**Translated**
2	Visa PG download ! Fro	PAD definition uit tru elementar	app so inge _SEA esse	Tob assis dad Anna tan
3	_Pea 850 Physik _ros Qu	protest _idyll _gemeint Joker ski	Saison) latura _lai fil	_f##u _thema ables speciale begin
5	_utopi _glatt frustr bekomme ##ofter	_kolon _Installation triumph _idyll _esot	xon app _Essen _Mental rre	UR _Már _224 _studier _nostra
10	objekt MT _boy gner _m##ochte	quez _m##ochte _lett Expert _m##ochten	_ mán MUS une _IS	, _Bring dimension _2001 Amen
20	cool _restaur Adam _Slo modi	_Nee _01 _friss ordina 31	kur _alarm _push _schicken app	" iler _Trip amour bha

Table 13: Top 5 tokens for adversarial XLM-R and mBART trained on *de-es* dataset for different n

Corpora	Non-Adversarial			
	O/T Acc	O/T F1	Topic Acc	Topic F1
Ted	0.719 [0.71,0.72]	0.715 [0.71,0.72]	0.529 [0.52,0.53]	0.531 [0.53,0.53]
Politics	0.595 [0.58,0.60]	0.460 [0.46,0.47]	0.463 [0.46,0.47]	0.458 [0.45,0.46]
Literature	0.868 [0.86,0.88]	0.899 [0.88,0.90]	0.587 [0.58,0.59]	0.610 [0.61,0.61]
	Adversarial			
Ted	0.684 [0.67,0.69]	0.688 [0.68,0.69]	0.140 [0.13,0.14]	0.121 [0.12,0.14]
Politics	0.548 [0.54,0.55]	0.414 [0.41,0.43]	0.314 [0.30,0.32]	0.303 [0.30,0.32]
Literature	0.827 [0.82,0.84]	0.850 [0.84,0.85]	0.538 [0.52,0.54]	0.560 [0.55,0.56]

Table 14: O/T and topic accuracies and F1 for adversarially (n=2) and non-adversarially trained mBERT for different corpora

Dataset	Train Set	Dev Set	Test Set	MTL
MPDE (*de-es,de-en*)	29580	6336	6344	80.16
Ted (*en-fr*)	5752	1438	1998	17.88
Politics (*de-en*)	8900	1482	1484	20.67
Literature (*de-en*)	25211	5000	5888	49.90

Table 15: Corpora Stats (number of examples for each set) for different Translationese Corpora (MTL: Mean Token Length). We also include details of the MPDE corpora for comparison

N	Non-Surgeon	Surgeon
3	hayward dance ##bring lc kids	collar night comment hour ##wen
5	longtime motivation afghanistan nbc russo	excel philip mike border nii
10	facultad ##school streets pubmed conservative	olav sister ##usa fold ede
20	wi legislative hospital biography minister	kolkata apollo americans fold typically

Table 16: IG results for occupation classification for different n

F Reproducing work from a study on spurious correlation mitigation on translationese classification

Apart from (Borah et al., 2023)'s results on the same MPDE corpus, we reproduce results from another work(Wang et al., 2022) for comparison with our mitigation approach. that focuses on mitigating spurious correlation on two tasks: sentiment classification and occupation classification. We utilize this work as they are recent in the domain of spurious correlation mitigation.

We utilize their proposed approach for *Cross-Data Analysis (CDA)* for spurious correlation mitigation in O/T classification. The work proposes CDA, that is, identify spurious tokens from several corpora in a test and later masking them for mitigation. Spurious tokens are found by identifying the most important tokens having the highest attention scores contributing to [CLS] tokens across different heads. For translationese classification, we consider the three corpora: *MPDE*, *Politics* and *Literature*, having 'de-en' data. After finding the most important tokens across these datasets, we mask these tokens in the MPDE dataset and perform the experiments. This approach is a more manual approach where top tokens are found and then masked in a similar manner as (Borah et al., 2023).

Data	Non-adversarial		Adversarial	
	Original	Translated	Original	Translated
Ted (en-fr)	jimbo bowie robbins ##dini clayton	richelieu 1916 noticias 1755 bolivia	wow newspapers track knock pendant	cosmic deti alzheimer 2006 prada
Politics (de-en)	ncaa calcutta f1 marines hurricanes	rouen barbosa bogota tibet associations	jura astronaut philosophie ##ibil ##erz	metropole astronaut indes yoko ##fio
Literature (de-en)	r pascal russe welch ##sper	watt ##bari timothy Westminster lancaster	warner st tomba chim base	##tow marianne konsul ##familien sokol

Table 17: Top 5 tokens for non-adversarial and adversarial (n=2) mBERT trained for different translationese corpora

N	Setting	O/T Acc, CI	O/T F1, CI	Topic Acc, CI	Topic F1, CI
2	Non-adversarial Adversarial	0.975 [0.96,0.96] 0.970 [0.96,0.98]	0.961 [0.96,0.98] 0.954 [0.95,0.96]	0.518 [0.50, 0.51] 0.459 [0.44,0.46]	0.492 [0.49, 0.51] 0.430 [0.42,0.44]
3	Non-adversarial Adversarial	0.975 [0.96,0.96] 0.968 [0.96,0.97]	0.961 [0.96,0.98] 0.952 [0.95,0.95]	0.450 [0.45, 0.45] 0.303 [0.30, 0.30]	0.422 [0.42, 0.42] 0.330 [0.33, 0.34]
5	Non-adversarial Adversarial	0.975 [0.96,0.96] 0.967 [0.96,0.98]	0.961 [0.96,0.98] 0.950 [0.95,0.96]	0.209 [0.20, 0.21] 0.143 [0.14,0.15]	0.213 [0.21, 0.21] 0.150 [0.14,0.15]
10	Non-adversarial Adversarial	0.970 [0.96,0.98] 0.970 [0.96,0.97]	0.954 [0.95,0.96] 0.954 [0.95,0.97]	0.110 [0.11, 0.11] 0.046 [0.04,0.04]	0.102 [0.10, 0.11] 0.032 [0.03,0.03]
20	Non-adversarial Adversarial	0.975 [0.96,0.96] 0.970 [0.97,0.98]	0.961 [0.96,0.98] 0.954 [0.95,0.95]	0.005 [0.00, 0.01] 0.001 [0.00,0.00]	0.005 [0.01, 0.01] 0.001 [0.00,0.00]

Table 18: Adversarial and Non-adversarial results (Acc(uracy), F1 score, CI(Confidence Score)) by mBERT on Occupation Classification Task. Lower topic accuracies and F1 scores are highlighted.

Paper	Approach	Task(s)
(He et al., 2019)	Known shortcuts - biased model that only uses features known to relate to dataset bias	NLI
(Clark et al., 2019)	Known shortcuts - using features correlated with training labels and not correlated with test labels	NLI, VQA, and QA
(Clark et al., 2020a)	Unknown shortcuts - lower capacity model to capture shallow correlations	Textual entailment, VQA, Image recognition task
(Wang et al., 2022)	Unknown shortcuts - attention scores (interpretability technique), cross-dataset stability analysis, knowledge aware perturbation	Sentiment Classification, Occupation Classification
(Amponsah-Kaakyire et al., 2022)	Unknown shortcuts - Integrated Gradients	Translationese Classification
(Borah et al., 2023)	Unknown shortcuts - 'Topic floor' measure	Translationese Classification
Our work	Unknown shortcuts - Topic Modeling Approaches, Probing to uncover Gender Bias	Translationese Classification, Occupation Classification

Table 19: Comparison to work on Spurious Correlation Detection in NLP

Paper	Approach	Task(s)
(He et al., 2019)	Biased model using dataset bias features + Debiased model	NLI
(Clark et al., 2019)	Biased model to capture spurious correlations + Robust model	NLI, QA and VQA
(Clark et al., 2020a)	Lower capacity model - trained with higher capacity model to capture shallow correlations	Textual entailment, Visual question answering, Image recognition tasks
(Stacey et al., 2020)	Ensemble Adversarial Mitigation	NLI
(Wang et al., 2022)	Masking spurious tokens	Sentiment Classification, Occupation Classification
(Borah et al., 2023)	Masking spurious tokens	Translationese Classification
Our work	Domain Adversarial Training	Translationese Classification, Occupation Classification to uncover Gender Bias

Table 20: Comparison to work on Spurious Correlation Mitigation in NLP

METHOD	O/T ACC, 95% CONFIDENCE SCORE	O/T F1, 95% CONFIDENCE SCORE
(Wang et al., 2022)	0.910, [0.91, 0.91]	0.915, [0.91, 0.92]
(Borah et al., 2023)	0.890, [0.88, 0.89]	0.890, [0.88, 0.88]
Domain Adversarial Training (Ours)	0.910, [0.90, 0.91]	0.910, [0.90, 0.92]

Table 21: Results of spurious correlation mitigation in the MPDE *de-en* translationese dataset

N	ORIGINAL	TRANSLATED
CDA (Wang et al., 2022)	tunis ! belarus republika thuringen	bilbao bale miranda zarangoza valencia
(Borah et al., 2023)	besukhte entdeckte veroffentlichte gehorten fuhrte	. alpen apo profits ##nova
Domain Adversarial Training (Ours)	ppm uks andersson prosa monterrey	italo domino ##unta ##inne arequipa

Table 22: IG results for comparing different studies on spurious correlation mitigation

Table 21 shows that the CDA method proposed by (Wang et al., 2022) has a similar performance as ours for O/T translationese classification. We further perform IG using the reproduced model and present the results in Table 22. We find that using CDA, there are several location NEs in the top tokens that are associated with the regions where the languages are spoken, for example: *thuringen*, *bilbao*, *zarangoza*, and *valenia*, even though the O/T classification performance is high. This shows that spurious tokens are utilized by the model with the proposed mitigation approach. Whereas, our approach has just one NE in the top 5 tokens with a more automatic approach.

G Implementation Details

This section contains training and hyperparameter details for probing and adversarial training experiments.

G.1 Probing

For [mBERT+OTD+CL], we use a multilingual BERTForSequenceClassification (base)model fine-tuned on the O/T data for O/T label classification. For [mBERT+OTD], we use a BERTForMaskedLM model fine-tuned on the O/T data for MLM task. For [mBERT], we use mBERT out-of-the-box with pre-trained weights from huggingface. We use mBERT-base-multilingual-uncased for our experiments which is pre-trained on 104 languages with the largest Wikipedia on an MLM objective. For BERT Sequence Classifier [BERT+OTD+CL], we use a batch size of 16, a learning rate of $4 \cdot 10^{-5}$, and an Adam optimizer with epsilon $1 \cdot 10^{-8}$ to train our mBERT models for 4 epochs. For the BERTForMaskedLM model - we use - learning rate: $1 \cdot e^{-5}$ and epsilon $1 \cdot 10^{-8}$, and trained for 3 epochs. For our LDA topic labels, we experiment with $n = 2, 3, 5, 10$, and 20.

For the probing experiments, we use a simple logistic regression model using the scikit-learn(Pedregosa et al., 2011) library, with an ’l2’ penalty.

G.2 Adversarial Training

We use the uncased version of `mBERT-base`[5] (like our experiments for probing and all other subsequently) for our adversarial model by specifying two classification objectives: one for O/T classification and the other for topic label classification. For `XLM-R`, we use the `multilingual XLM-Roberta`[6] from huggingface. For `mBART`, we used the `mBART-large-50`[7] model from huggingface. We use a batch size of 16, a learning rate of $4 \cdot 10^{6}$, and an Adam optimizer with epsilon $1 \cdot 10^{5}$ to train our all our adversarial models for 4 epochs. For our LDA topic labels, we experiment with n = 2, 3, 5, 10, and 20.

G.3 Computational resources

Experiments were run on NVIDIA RTX2080 and NVIDIA-A40 GPUs. mBERT and `XLM-R`(adversarial and non-adversarial) were run on NVIDIA RTX2080 GPUs training experiment takes 1.5 GPU hours. `mBART` was run on

[5] https://huggingface.co/google-bert/bert-base-multilingual-uncased

[6] https://huggingface.co/docs/transformers/en/model_doc/xlm-roberta

[7] https://huggingface.co/docs/transformers/en/model_doc/mbart

NVIDIA-A40 and training took around 2 hours. We do not use GPU for our other experiments, like, LDA, probing using logistic regression, and mBERT embedding extraction experiments.

H Reproducibility

We open-source our codes and datasets, which are both uploaded to the submission system. We include commands with hyperparameters in our codes. This would help future work to reproduce our results.