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

Anonymous ACL submission

Abstract

Modern classifiers, especially neural networks, excel at leveraging faint and subtle signals competing with many other signals in the data. When such potentially noisy setups lead to high accuracy rates (e.g., 90%+), it produces concerns about the authenticity of the results, raising questions about potential spurious correlations – a phenomenon often referred to as "Clever Hans". We explore this phenomenon in the context of translationese classification, where previous work has found indirect and episodic evidence that a high-performance BERT classifier learns to use spurious topic information rather than just translationese signals. In this paper, we first use probing to provide direct evidence that high-performance translationese classifiers pick up unknown potentially spurious topic correlations. We then introduce adversarial training as a strategy to mitigate any such potentially spurious topic correlations, where previous work was only able to mitigate specific known (episodic) Clever Hans. We demonstrate the effectiveness of our approach on translationese classification tasks on two translation pairs.

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 be manifested 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, 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 BERT-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. Amponsah-Kaakyire et al. (2022) show that the differences between these methods are due to learned features rather than classifier differences.

Using Integrated Gradients (Sundararajan et al., 2017), (Amponsah-Kaakyire et al., 2022) also found that BERT 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 BERT uses for O/T classification are geographical place names: German-based place names for O and Spanish-based place 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 potentially spurious but 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), present the

topic floor, average weighted alignment of documents in 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) also mitigated known topic signals in the form of location-named entities (NEs) (Amponsah-Kaakyire et al., 2021) by masking NEs in the training and test data.

However, (Borah et al., 2023) provided only indirect evidence that BERT uses topic signals in O/T classification by showing that in principle a BERT classifier can learn LDA/Bertopic clusters as target labels and that masking known spurious topics such as location and other NEs in the data reduces O/T classification accuracy. Showing that if told to do so, BERT can learn topics is not the same as showing that a BERT 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 BERT 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 place NEs. In addition, (Borah et al., 2023) only address known spurious topic mitigation (geographic place 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 BERTtopic (over and above geographic place NEs) may carry spurious information with respect to the O/T target label classification.

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. 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 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? Resolving this will help mitigate "Clever Hans" in translationese classification.

In this paper, we address the two questions using probing for the first and adversarial training for the second. We probe BERT's encoder layers to test whether a high-performance BERT-based O/T classifier can identify any potentially spurious topic correlations with target classifications captured by LDA. We compare three BERTs - 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 BERT model not fine-tuned on any further data. The logic is that if BERT O/T classifiers learn and use spurious LDA topic correlations with O/T target labels, then probing BERT O/T classifiers for LDA topics should yield higher accuracy/F1 than an MLM BERT and an off-the-shelf BERT. If this is observed, this constitutes direct evidence that a BERT O/T classifier learns and uses spurious unknown topic information and that the "topic floor" proposed by (Borah et al., 2023) is real. If not, then it is unclear whether the classifier learns and uses spurious unknown topic information and this raises doubts about the "topic floor". For our second research question of extending Clever Hans mitigation beyond known spurious correlations (such as location NEs), we utilize adversarial training to suppress LDA-based potentially spurious unknown topic signals 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. Additionally, adversarially-trained classifiers should generalize better to test data that differs in various ways from what the classifier has seen in training.

Our contributions include:

1. We use probing to directly show that a BERT 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 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 mitigating spurious topic-target

O/T label correlations using adversarial training leads to O/T classifiers with substantially improved generalization (robustness) to test data that differs markedly from training data of the classifiers.

4. We present empirical results for Clever Hans mitigation in translationese classification for two different language pair settings from (Amponsah-Kaakyire et al., 2021): *de-es* (half German originals and half Spanish-German translations), and *de-en* (half German originals and English-German translations), extending previous work on *de-es* only (Borah et al., 2023).

5. We use Integrated Gradients (IG) based Explainable AI to compute the top tokens adversarial BERT uses for translationese classification, and show that topic reliance is mitigated.

Translationese classification is a prototypical instance of classification using weak signals competing with many other signals in the data. We expect our contributions to be useful in many similar 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 explainable AI (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 an unsupervised clustering-based approach.

More recent research uses feature and representation learning approaches (sometimes augmented with hand-crafted features) based on neural networks (Sominsky and Wintner, 2019; Pylypenko et al., 2021). (Pylypenko et al., 2021) show that representation learning-based approaches like BERT perform much better than handcrafted and feature engineering approaches and this is due to feature learning rather than the classifiers (Amponsah-Kaakyire et al., 2022). Using Explainable AI (XAI) approaches like IG (Sundararajan et al., 2017), (Amponsah-Kaakyire et al., 2022) found that BERT exploits spurious topic signals in the form of location names correlated with the O/T classification labels in the data.

Translationese signals are subtle and spurious correlations between O and T classification targets and topic signals in the data may impact (and inflate) the classification results of neural networks. (Borah et al., 2023) use translationese classification as a setting to measure and mitigate Clever Hans in a classification task where signals are weak and competing with many other signals. They consider unknown and known spurious correlations in the form of topic signals. The basic idea is simple: when topic signals are unknown, they use unsupervised topic clustering, LDA and BERTtopic, 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 the 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. The topic floor is 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.

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

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) propose Clever Hans mitigation, albeit 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., 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 BERT can be trained (i.e. told) to learn LDA (and BERTopic) topics as target classes. This, however, is not the same as showing that a BERT 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 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.

Translationese is not just a topic in basic linguistic research: many 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., 2020), 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). 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). In our paper, we use probing to find direct evidence that BERT learns and uses spurious topic signals as provided by unsupervised topic modeling approaches 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. If x is the input to the GRL and y is the output, then during backpropagation, if $\frac{\partial L}{\partial y}$ is the gradient of the loss function with respect to y, then:

$$\frac{\partial L}{\partial x} = -\lambda \frac{\partial L}{\partial y}$$

where $\frac{\partial L}{\partial x}$ is the gradient of the loss with the respect to x, and λ controls the amount of gradient reversal. (Stacey et al., 2020) e.g. used an ensemble adversarial technique to reduce hypothesis-only bias in Natural Language Inference (NLI) appearing 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 potentially spurious topic signals by BERT in O/T target label classification. To the best of our knowledge, this is the first time adversarial training has been explored in Clever Hans mitigation in translationese classification.

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 where 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. The MDPE subsets of the Europarl data we use here contain only data from before 2004, since relay translations were included in 2004, where it may not be known whether or not the source language is already the result of a translation (Bogaert, 2011). For our experiments on how our adversarial and our regular models generalize, we use the literature translationese corpus (Rabinovich et al., 2018), which consists of literature classics (originals and translations) originating in the 18th-20th centuries authored by English or German writers. For this, our O/T classification models fine-tuned on the MPDE DE-EN translationese dataset are tested on the test set of the literature translationese corpus, which consists of 4.3k paragraphs with half German originals and half translations from English to German.

4 Unsupervised Clustering

We use Latent Dirichlet Allocation (LDA) (Blei et al., 2001) as our unsupervised 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, which determines the per-document topic distribution, and β which determines the per-topic word distribution. LDA assigns a latent topic to every word through iteration by computing a topic word distribution (θ) in the data. We need to specify the number of topics n that we want LDA to output. In our experiments we explore $n = 2, 3, 5, 10$, and 20, as these all show high topic floor scores in the range $[0.55, 0.60]$ (Borah et al., 2023). After performing LDA, we assign each data point (i.e. paragraph) 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. We use the Gensim (Rehurek and Sojka, 2011) implementation of LDA for our experiments.

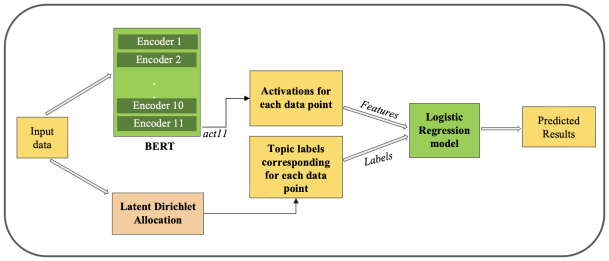

Figure 1: Probing Pipeline

5 Probing for Topics in O/T Classification

5.1 Probing Experiment Design

In this section, we present our probing-based approach to show whether a high-performance BERT-based translationese classifier learns to use spurious correlations in the form of LDA-based topics.

We probe three different BERTs for topic classification:

1. **[BERT+OTD+CL]**: a BERTforSequenceClassification model fine-tuned on MPDE translationese data containing original/translated labels for an O/T classification task.

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

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

Each of the three BERT models are pre-trained on the same data. The logic behind our experiment is the following: BERT finetuned on O/T data and trained for O/T classification [BERT+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 BERT should exhibit better performance on topic probes compared to a BERT fine-tuned on the same O/T data with the regular BERT MLM objective but not trained for O/T classification [BERT+OTD] and better than a simple BERT out of the box [BERT] not fine-tuned at all on the O/T data.

We perform topic classification probing using BERT encoder activations as features and LDA topics as the target labels of a simple logistic regression probe. We take the [CLS] activations of

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

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

the last encoder layer output (768 dimensional). 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 identified by LDA. Training and hyperparameter details are provided in Appendix A.1. The probing pipeline is displayed in Fig 1.

5.2 Probing Results

To account for the stochastic nature of LDA, we perform probing experiments on three different runs of LDA on the data. The logistic regression model in the probe is deterministic, hence for each single LDA run probe results are deterministic.

Table 1 shows the probing results for all numbers of LDA topics n averaged over 3 runs of LDA in the data. Compared to [BERT+OTD] and [BERT], probing [BERT+OTD+CL] yields the highest topic scores in terms of accuracy and, even more pronounced, F1 scores. This shows that O/T classification makes BERT learn spurious topic information and that this does not happen (to the same extent) for BERT finetuned on the same O/T data with just the MLM objective and without O/T classification and similarly for BERT out of the box. Table 7 in Appendix D shows the same trend for probing *de-en*.

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 topic clustering methods to mitigate "Clever Hans" in translationese. Adversarial training (Ganin et al., 2016) uses an additional objective function to provide model generalization for both adversarial data and clean data.

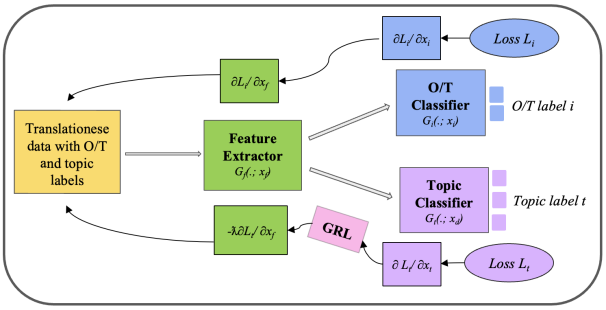

Figure 2: Adversarial Training Pipeline

In our experiments, 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 BERT blind to spurious topics and reduce "Clever Hans" identified by unsupervised topic modeling techniques for translationese classification. Training and hyperparameter details are provided in Appendix A.2. The adversarial training pipeline is displayed in Fig 2

6.2 Adversarial Training Results

This section shows the results of our adversarial training experiments. Results are averaged over 20 bootstrapped samples of the test data, and 95% confidence scores of the F1 scores are displayed in Table 2.

Table 2 shows a comparison of O/T accuracies and F1, and topic accuracies and F1 for the adversarial and non-adversarial BERT. 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 reduced for the adversarial model. This is expected as we maximize the loss of topic classification while minimizing the loss of the O/T label classification. Results for $n = 3$ are the only exception to the general pattern: while, as expected, topic accuracy for non-adversarial (0.458) is greater than for the adversarially trained model (0.399), the corresponding F1 scores are reversed (0.288 against 0.379). Table 5 in Appendix E displays the results for the *de-en* pair and fully shows the expected pattern for both

n	Adversarial		Non-Adversarial	
	O/T acc, F1 (95% confidence F1)	Topic acc, F1 (95% confidence F1)	O/T acc, F1 (95% confidence F1)	Topic acc, F1 (95% confidence 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])	0.399, 0.379 ([0.37, 0.39])	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 2: Adversarial and Non-Adversarial O/T classification and topic label classification results for *de-es*

accuracy and F1 scores. The topic accuracies and F1 scores for the adversarial model are lower than the non-adversarial model.

As expected, the topic label accuracies and F1 scores reduce overall as the number of topic labels increases. In comparison, for the non-adversarial BERT model finetuned on O/T data topic label accuracies and F1 scores are higher than for the adversarial model, while the O/T accuracies remain almost consistent across both models. We note the scores for O/T acc and F1 are constant across all n (except n=3, discussed above) for the non-adversarial BERT model since it is only fine-tuned for translationese classification and not adversarially "finetuned" against topic classification.

Adversarial training results for the *de-en* pair are presented in the Table 5 in Appendix E.

7 Integrated Gradients and Topic Traces

7.1 IG Experiment Design

IG (Sundararajan et al., 2017) is an explainable AI (XAI) technique that computes the gradient of a model's output to its input features and can be applied to any differentiable model processing images, text, etc. and requires no modification to the model to be explained. We use 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 BERT uses some spurious location name topic signals as short-cuts in the data for translationese classification. For example, German T data translated from Spanish contain Spanish (or South American) location NEs in the top tokens identified by IG, e.g., 'Nicaragua', 'Bilbao', 'Colombia' etc. (Borah et al., 2023) used IG on the BERT 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 paper, we use IG to compute the top tokens used by the adversarial BERT model to investigate the mitigation of topic signals in translationese classification. The expectation is to have a reduced number of topic-related tokens, like named entities (as found by (Amponsah-Kaakyire et al., 2022)) in the top tokens computed by IG.

7.2 IG Results

Table 3 shows the top 20 tokens with the highest IG attribution scores used by the adversarial and non-adversarial models for the O and T-test sets for the *de-es* dataset. There is only one South American Spanish language location token in the top 20 tokens in the test set for the adversarial case - *arequipa* in the translated class. By contrast, in the non-adversarial case, 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 cannot be considered as a direct spurious correlation with the O set language. Overall, there is mitigation of topic-related NE cues in the adversarial model as expected.

Table 8 in Appendix F shows the same trend for the *de-en* pair.

8 Model Generalization

8.1 Experiment Design Generalization

Mitigating "Clever Hans" using adversarial training should make our models generalize better to unseen data that may not come with the same spurious topic information as the training data our O/T classifier was originally trained on. We test the performance of our adversarially trained O/T classification model finetuned on the MPDE corpus translationese data (Amponsah-Kaakyire et al., 2021) vs. a non-adversarially trained O/T classification model finetuned on the same data. To do this, we deliberately use test data from a different domain, the literature translationese corpus from (Rabinovich et al., 2018). The corpus consists of

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

Model	**O/T acc, 95% conf.score**	**O/T F1, 95% conf.score**
Adversarial	0.538, [0.53, 0.54]	0.526, [0.52, 0.53]
Non-Adversarial	0.483, [0.47, 0.49]	0.461 [0.45, 0.47]

Table 4: Test set results of the adversarial model vs non-adversarial model on the *de-en* literature translationese dataset

literature classics between the 18th and 20th centuries authored by English and German writers. We utilize the German originals and translations from English to German texts, so that they align with the translation direction of *de-en* corpus from the MPDE corpus. The literature test set consists of 4.3k paragraphs (discarding paragraphs with less than 20 words) with half of them being German originals and the other half translations from English to German.

8.2 Results Generalization

Table 4 presents the test set performance of the adversarial model and the non-adversarial model (BERT O/T classifier finetuned on MPDE translationese data) on the literature translationese dataset. Table 4 shows that accuracy and F1 scores for the adversarial model are higher than for the non-adversarial model. This shows that the adversarial model finetuned to suppress spurious topic signals generalizes better to unseen data on O/T translationese classification. (Note that the adversarial models considered here for comparison are trained for maximizing n=2 topic labels). However, note that our adversarial classifier, although clearly more robust under domain shift and better than our non-adversarial classifier, still does not perform well on the new dataset. The results still show high dependence on the domain the classifier was trained on with substantial further scope for improvement in the generalization experiments.

9 Conclusion

In this paper, we focus 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 generalize previous work in (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) that we can mitigate any (unknown) spurious topic correlation using adversarial training with LDA topic labels as adversarial targets in classification. We show 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 conjecture that our contributions are generic in the sense that they should be useful in many other classification settings that are subject to similar general constraints. Our research shows that the topic floor proposed by (Borah et al., 2023) is real, and that adversarial training maintains high target classification accuracy while diminishing Clever Hans. We present translationese classification experiments on two language pairs, we use integrated gradients to spot-check the effect of adversarial training on known spurious correlations (location NEs) and we show that adversarially trained translationese classifiers are more robust in the sense that they generalize better to data in a domain different from what the classifier saw in training. Future research includes zooming in on specific LDA topics that exhibit high alignment with target labels, exploring other topic modeling approaches, and building models with better generalization abilities by further mitigating stronger spurious signals used in classification tasks that require classifying subtle signals like translationese.

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 high topic floor scores for these settings. That said we should explore topic models other than LDA, e.g. BERTtopic (Grootendorst, 2022) etc. While our adversarially trained models clearly generalize better to unseen data in a different domain, overall O/T classification results on the literature data set leave much to be desired and present considerable scope for improvement.

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.

Caroline Bogaert. 2011. *Is absolute multilingualism maintainable? The language policy of the European Parliament and the threat of English as a lingua franca*. Ph.D. thesis, MA thesis, Universiteit Gent.

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.

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

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.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35.

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*.

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.

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.

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. 2018. 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.

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 Implementation Details

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

A.1 Probing

For [BERT+OTD+CL], we use a BertForSequenceClassification model fine-tuned on the O/T data for O/T label classification. For [BERT+OTD], we use a BertForMaskedLM model fine-tuned on the O/T data for MLM task. For [BERT], we use BERT out-of-the-box with pretrained weights from huggingface. We use BERT-base-multilingual-uncased for our experiments which is pretrained on 104 languages with the largest Wikipedia on an MLM objective. 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 BERT models for 4 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.

A.2 Adversarial Training

We use the uncased version multilingual BERT (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 BERT models for 4 epochs. For our LDA topic labels, we experiment with $n = 2$, 3, 5, 10, and 20.

A.3 Computational resources

All experiments are run on NVIDIA RTX2080 GPUs. Each BERT (adversarial and non-adversarial) training experiment takes 1.5 GPU hours. We do not use GPU for our other experiments, like, LDA, probing using logistic regression, and BERT embedding extraction experiments.

B 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.

n	Adversarial		Non-Adversarial	
	O/T acc, F1 (95% confidence F1)	Topic acc, F1 (95% confidence F1)	O/T acc, F1 (95% confidence F1)	Topic acc, F1 (95% confidence 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 5: Adversarial training results for the *de-en* dataset

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 6: BERT fine-tuned on translationese data for O/T classification for two language-pairs

C Translationese fine-tuned BERT results for different language-pairs

Here, we present the accuracy and F1 scores of BERT fine-tuned on the MPDE translationese dataset for the two datasets (*de-es* and *de-en*). Note that the results are for non-adversarial BERT which is not trained to suppress any topic signals. Table 6

n	Model	Accuracy	F1-score
2	[BERT+OTD+CL]	0.564	0.667
	[BERT+OTD]	0.556	0.606
	[BERT]	0.561	0.659
3	[BERT+OTD+CL]	0.409	0.538
	[BERT+OTD]	0.397	0.483
	[BERT]	0.397	0.479
5	[BERT+OTD+CL]	0.306	0.434
	[BERT+OTD]	0.290	0.379
	[BERT]	0.295	0.381
10	[BERT+OTD+CL]	0.254	0.405
	[BERT+OTD]	0.252	0.393
	[BERT]	0.256	0.392
20	[BERT+OTD+CL]	0.142	0.236
	[BERT+OTD]	0.129	0.199
	[BERT]	0.134	0.200

Table 7: Probing results (last encoder layer as features) on the *de-en* datasets

D Probing on other language pairs

In this section, we present the results of probing experiments on the *de-en* set. Table 7 displays the probing experiments for different n values. As observed in the de-es dataset in Section 5.2, we find

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

Table 8: Top 5 tokens for adversarial model trained on *de-es* and *de-en* datasets for different n

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.

E Adversarial Training on other language pairs

Here, we present the results of adversarial training on the *de-en* language-pair. Table 5 shows the results of adversarial training for different n values. Similar to the de-es language pair, 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.

F Integrated Gradients on other language pairs

Table 8 presents the results of integrated gradients 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. This may be an indicator that spurious topic correlation signals using location NEs are reduced in our adversarial model.