Topics to Avoid: Demoting Latent Confounds in Text Classification

Sachin Kumar[◇] **Shuly Wintner**[♣] **Noah A. Smith**[♡♠] **Yulia Tsvetkov**[◇]
[◇]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[♣]Department of Computer Science, University of Haifa, Haifa, Israel
[♡]Paul G. Allen School of Computer Science & Engineering,
University of Washington, Seattle, WA, USA
[♠]Allen Institute for Artificial Intelligence, Seattle, WA, USA
sachink@cs.cmu.edu, shuly@cs.haifa.ac.il
nasmith@cs.washington.edu, ytsvetko@cs.cmu.edu

Abstract

Despite impressive performance on many text classification tasks, deep neural networks tend to learn frequent superficial patterns that are specific to the training data and do not always generalize well. In this work, we observe this limitation with respect to the task of *native language identification*. We find that standard text classifiers which perform well on the test set end up learning topical features which are confounds of the prediction task (e.g., if the input text mentions Sweden, the classifier predicts that the author's native language is Swedish). We propose a method that represents the latent topical confounds and a model which "unlearns" confounding features by predicting both the label of the input text and the confound; but we train the two predictors adversarially in an alternating fashion to learn a text representation that predicts the correct label but is less prone to using information about the confound. We show that this model generalizes better and learns features that are indicative of the writing style rather than the content.[1]

1 Introduction

Text classification systems based on neural networks are biased towards learning frequent spurious correlations in the training data that may be confounds in the actual classification task (Leino et al., 2019). A major challenge in building such systems is to discover features that are not just correlated with the signals in the training data, but are true indicators of these signals, and therefore generalize well.

For example, Kiritchenko and Mohammad (2018) found that sentiment analysis systems implicitly overfit to demographic confounds, systematically amplifying the intensity ratings of posts written by women. Zhao et al. (2017) showed that visual semantic role labeling models implicitly capture actions stereotypically associated with men or women (e.g., *women are cooking* and *men are fixing a faucet*), and in cases of higher model uncertainty assign stereotypical labels to actions and objects, thereby amplifying social biases found in the training data.

We focus on the task of *native language identification* (L1ID), which aims at automatically identifying the native language (L1) of an individual based on their language production in a second language (L2, English in this work). The aim of this task is to discover stylistic features present in the input that are indicative of the author's L1. However, a model trained to predict L1 is likely to predict that a person is, say, a native Greek speaker, if the texts authored by that person mention Greece, because the training data exhibits such topical correlations (§2).

This problem is the focus of our work, and we address it in two steps. First, we introduce a novel method for representing *latent* confounds. Recent relevant work in the area of domain adaptation (Ganin et al., 2016) and deconfounding for text classification (Pryzant et al., 2018; Elazar and Goldberg, 2018) assumes that the set of confounds is known a priori, and their values are given as part of the training data. This is an unrealistic setting that limits the applicability of such models in real world scenarios. In contrast, we introduce a new method, based on log-odds ratio with Dirichlet prior (Monroe et al., 2008), for identifying and representing latent confounds as probability distributions (§3). Second, we propose a novel alternating learning procedure with multiple adversarial discriminators, inspired by adversarial learning (Goodfellow et al., 2014), that demotes latent confounds and results in textual representations that are invariant to the confounds (§4).

[1]The code is available at: https://github.com/Sachin19/adversarial-classify

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 4153–4163,
Hong Kong, China, November 3–7, 2019.

Note that these two proposals are task-independent and can be extended to a vast array of text classification tasks where confounding factors are not known a priori. For concreteness, however, we evaluate our approach on the task of L1ID (§5). We experiment with two different datasets: a small corpus of student written essays (Malmasi et al., 2017) and a large and noisy dataset of Reddit posts (Rabinovich et al., 2018). We show that classifiers trained on these datasets without any intervention learn spurious topical correlations that are not indicative of style, and that our proposed deconfounded classifiers alleviate this problem (§6). We present an analysis of the features discovered after demoting these confounds in §7.

The main contributions of this work are:

1. We introduce a novel method for representing and identifying variables which are confounds in text classification tasks.

2. We propose a classification model and an algorithm aimed at learning textual representations that are invariant to the confounding variable.

3. We introduce a novel approach to adversarial training with multiple adversaries, to alleviate the problem of drifting parameters during alternating classifier–adversary optimization.

4. Finally, we analyze some linguistic features that are not only predictive of the author's L1 but are also devoid of topical bias.

2 Motivation

We study the general effect of *topical* confounds in text classification. To motivate the need to demote them, we introduce as a case study the L1ID task, in which the goal is to predict the native language of a writer given their texts in L2.

We begin with a subset of the L2-Reddit corpus (Rabinovich et al., 2018), consistsing of Reddit posts by authors with 23 different L1s, most of them European languages. Some of the posts come from Europe-related forums (e.g. r/Europe, r/AskEurope, r/EuropeanCulture), whereas others are from unrelated forums. We view the latter as out-of-domain data and use them to evaluate the generalization of our models. We use a subset of this corpus, with only the 10 most frequent L1s, to guarantee a large enough balanced training set. We remove all the posts with fewer than 50 words and sample the dataset to obtain a balanced distribution of labels: from this balanced dataset, we randomly sample 20% of examples from each class and divide them equally to create development and test sets. In total, there are around 260,000 examples in the training set and 32,000 examples each in the development, the in-domain test set, and the out-of-domain test set.

We trained a standard (non-adversarial) classifier, with a bidirectional LSTM encoder followed by two feedforward layers with a $\tanh$ activation function and a softmax in the final layer (full experimental details are given in §5.2). We refer to this model as NO-ADV. The results are shown in Table 1. Notice the huge drop in accuracy on the out-of-domain data, which indicates that the model is learning topical features.

To further verify this claim, we used *log-odds ratio with Dirichlet prior* (Monroe et al., 2008)—a common way to identify words that are statistically overrepresented in a particular population compared to others—to identify the top-K words that were most strongly associated with a specific L1 in the training set. (We refer the reader to (Monroe et al., 2008) for the details about the algorithm.) We experimented with $K \in \{20, 50, 100, 200\}$. Table 2 shows the top-10 words in each class; observe that almost all of these words are geographical (hence, topical) terms that have nothing to do with the L1.

Next, we masked such topical words (by replacing them with a special token) and evaluate the trained classifier on masked test sets. Accuracy (Table 1) degrades on both the in-domain and out-of-domain sets, even when only 20 words are removed. The drop in accuracy with the out-of-domain dataset is smaller since these data do not include many instances where the presence of topical words would help in identifying the label. These experiments confirm our hypothesis that the baseline classifier is primarily learning topical correlations, and motivate the need for a deconfounded classification approach which we describe next.

	In-Domain	**Out-of-Domain**
NO-ADV	52.5	25.7
+MASK TOP-20	32.8	21.0
+MASK TOP-50	31.6	20.4
+MASK TOP-100	30.1	19.7
+MASK TOP-200	28.5	18.7

Table 1: Motivation: accuracy (%) of L1ID on the L2-Reddit dataset.

3 Representing Confounds

Latent Dirichlet allocation (LDA; Blei et al., 2003) is a probabilistic generative model for discovering abstract topics that occur in a collection of documents. Under LDA, each document can be considered a mixture of a small (fixed) number of topics—each represented as a distribution over words—and each word's presence is assumed to be attributed to one of the document's topics. More precisely, LDA assigns each document a probability distribution over a fixed number of topics K.

LDA topics are known to be poor features for classification (McAuliffe and Blei, 2008), indicating that they do not encode all the topical information. Moreover, they can encode information which is not actually topical and can be a useful L1 marker. Motivated by our case study (§2), we propose a novel method to represent topic distributions, based on log-odds scores (Monroe et al., 2008), and compare it to LDA as a baseline.

For each class label y and each word type w, we calculate a log-odds score $lo(w, y) \in \mathbb{R}$. The higher this score, the stronger the association between the class and the word. As we saw in §2, the highest scored words are mostly topical and hence constitute superficial features which we want the classification model to "unlearn." We therefore define a distribution which assigns high probability to a document containing these high scoring words. For a label $y \in \mathcal{Y}$ and an input document $x = \langle w_1, \ldots, w_n \rangle$, we define $p(y \mid x)$:

$$p(y \mid x) \propto p(y) \cdot p(x \mid y) = p(y) \cdot \prod_{i=1}^{n} p(w_i \mid y)$$

The above expansion assumes a bag of words representation. When the dataset is balanced, $p(y)$ is equal for each label and can be omitted. Finally, we define $p(w_i \mid y) \propto \sigma(lo(w_i, y))$, where $\sigma(.)$ is the sigmoid function, which squashes the log-odds scores (whose values are in $\mathbb{R}$) to the range $[0, 1]$. We normalize the sigmoid values over the vocabulary to convert them to a probability distribution. In this distribution, the number of "topics" equals the number of labels, m.

4 Deconfounded Text Classification

We now formalize the task setup and the classification model. We are given N labeled documents in the training set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$, where x_i is a document with label $y_i \in \mathcal{Y}$, where $m = |\mathcal{Y}|$ is the number of labels. For each document x_i, we represent latent (topical) confounds—domain-specific and superficial document features—as a K-dimensional multinomial distribution $t_i \in \{(t_1, \ldots, t_K) \mid \sum_{j=1}^{K} t_j = 1\}$. In our task, the confounds are *topics*, so that each t_j represents the proportion of document i associated with topic j but these topics are not given a priori. In this work, the number of topics K, equals m, but the methods presented in this work are valid for any number of topics.

Our goal is to train a classifier f, parameterized by θ, which learns to accurately predict the target label, while ignoring superficial topical correlations present in the training set. That is, for a text x we wish to predict $\hat{y} = f_\theta(x)$ which doesn't encode any information about t. Following Ganin et al. (2016), Pryzant et al. (2018), and Elazar and Goldberg (2018), we input x to an encoder neural network $h(x; \theta_h)$ to obtain a hidden representation $\mathbf{h}_x$ (see Figure 1), followed by two feedforward networks: (1) $c(h(x); \theta_c)$ to predict the label y; and (2) an adversary network $\text{adv}(h(x); \theta_a)$ to predict the topics. Departing from prior work which used predefined binary confounds, our adversary predicts the topic distribution t. If $\mathbf{h}_x$ does not encode any information to predict t, then $c(h(x))$ will not depend on t. Concretely, we want to optimize the following quantity:

$$\min_{c,h} \frac{1}{N} \sum_{i=1}^{N} \text{CE}(c(h(x_i)), y_i) + \text{CE}(\text{adv}_h^*(h(x_i)), \mathbb{U}_K)$$

where CE denotes cross-entropy loss, and

$$\text{adv}_h^* = \arg\min_{\text{adv}} \frac{1}{N} \sum_{i=1}^{N} \text{CE}(\text{adv}(h(x_i)), t_i),$$

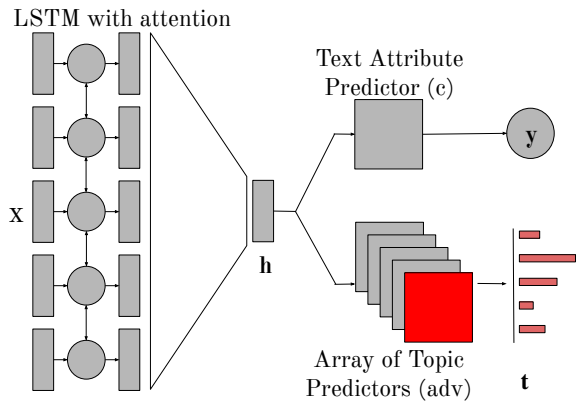
and $\mathbb{U}_K = (\frac{1}{K}, \ldots, \frac{1}{K})$. This objective seeks a representation $\mathbf{h}_x$ which is maximally predictive of the class label but not of the topical distribution (ideally, it should output a uniform topic distribution for every input).

4.1 Learning Schedule: Alternating Optimization of Classifier and Adversary

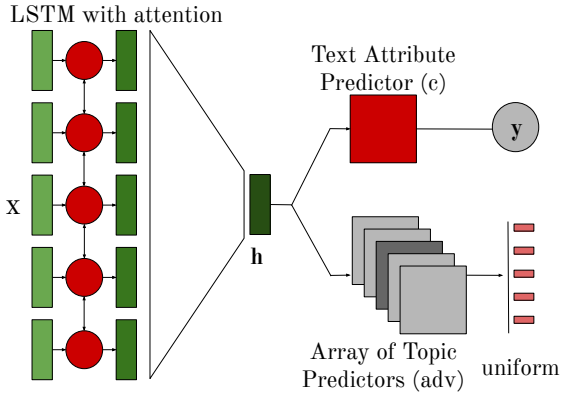
In practice, this optimization is done in an alternating fashion by minimizing the following two

English	ireland irish british britain russia scotland england states american london brexit
Finnish	finland finnish finns helsinki swedish finn nordic sweden sauna nokia estonian
French	french france paris sarkozy macron fillon hollande gaulle hamon marine valls breton
German	german germany austria merkel refugees asylum germans bavaria austrian berlin also
Greece	greek greece greeks syriza macedonia athens turkey macedonians fyrom turkish ancient
Dutch	dutch netherlands amsterdam wilders rotterdam holland rutte belgium bike hague
Polish	poland polish poles warsaw lithuanian lithuania judges jews ukranians imho tusk
Romanian	romania romanian romanians moldova bucharest hungarian hungarians transistria
Spanish	spain catalan spanish catalonia catalans madrid barcelona independence spaniards
Swedish	sweden swedish swedes stockholm swede malmo danish nordic denmark finland

Table 2: Top words based on log-odds scores for each label in the L2-Reddit dataset.



(a) Weights of the LSTM and of the discriminator are fixed. A new topic predictor is trained by minimizing the cross entropy of the output and the distribution of the input document over latent topics as described in §3.



(b) Weights of all the topic predictors are fixed, but the encoder is trained. The model is jointly minimizing the cross-entropy of the classifier and encouraging the topic predictor toward uniformity.

Figure 1: We alternate between training the topic predictor (left; (1)) and the deconfounded classifier/encoder (right; (2)). Pretraining is not shown in the figure.

quantities:

$$\min_{\text{adv}} \frac{1}{N} \sum_{i=1}^{N} \text{CE}(\text{adv}(h(x_i)), t_i) \quad (1)$$

$$\min_{c,h} \frac{1}{N} \sum_{i=1}^{N} \text{CE}(c(h(x_i)), y_i) + \text{CE}(\text{adv}(h(x_i)), \mathbb{U}_K) \quad (2)$$

The training schedule is critical in adversarial setups where the loss has two competing terms (Mescheder et al., 2018; Arjovsky and Bottou, 2017; Roth et al., 2017); here, these terms minimize classification loss while maximizing the topic prediction loss. Algorithm 1 details our proposed alternating learning procedure.

Inspired by generative adversarial networks (GANs; Goodfellow et al., 2014), the training procedure alternates between training the classifier and the adversary (see Figure 1). First (*pretraining*), we train the encoder along with the classifier using only classification loss, until convergence. After pretraining, $\mathbf{h}_x$ has encoded topical information which it uses for classification (as shown in our analysis in §2). Now, we train only $\text{adv}(h(x))$ to (accurately) predict t, keeping the parameters of $h(.)$ fixed. Once $\text{adv}(.)$ is trained, it should be able to successfully extract a topic distribution from $\mathbf{h}_x$ (topic training, see Figure 1a). The goal now is to modify $\mathbf{h}_x$ in such a way that $\text{adv}(\mathbf{h}_x)$ produces a uniform distribution (that is, fooling the adversary; similar to fooling the discriminator in GANs). We do that by keeping the weights of $\text{adv}(.)$ fixed, and training the network to produce the class label and a uniform topic distribution (*topic forgetting*, see Figure 1b). We then repeat this procedure for a fixed number of steps which was tuned using the validation set.

Algorithm 1: Alternating optimization of classifier and adversary.

Result: $\theta_h, \theta_c, \theta_{a_1}, \ldots, \theta_{a_T}$
Randomly initialize θ_h, θ_c;
while *not converged* **do**
  Sample a minibatch of b training samples;
  Update θ_h and θ_c using gradients with respect to $\frac{1}{b}\sum_{i=1}^{b} \mathrm{CE}(c(h(x_i)), y_i)$.;
end
$j = 1$;
for *number of training iterations* T **do**
  Randomly initialize θ_{a_j};
end
for t *steps* **do**
  Sample a minibatch of b training samples;
  Fix θ_h and θ_c, update θ_{a_j} using gradients with respect to $\frac{1}{b}\sum_{i=1}^{b} \mathrm{CE}(\mathrm{adv}_{\theta_{a_j}}(h(x_i)), t_i)$;
end
for c *steps* **do**
  Sample a minibatch of b training samples;
  Fix θ_{a_u} for $u \in_R \{1, \ldots, j\}$ and update θ_c and θ_h using gradients with respect to $\frac{1}{b}\sum_{i=1}^{b} \mathrm{CE}(c(h(x_i)), y_i) + \mathrm{CE}(\mathrm{adv}_{\theta_{a_u}}(h(x_i)), \mathbb{U}_K)$;
end
$j \longleftarrow j + 1$;

4.2 Multiple Adversaries

In our experiments, we observe that after every "topic forgetting" stage, $\mathrm{adv}(.)$ does end up producing a uniform distribution, but in the next "topic training" phase, $\mathrm{adv}(.)$ is able to reproduce the topical distribution accurately. This is because, during "topic forgetting," the classifier does not really forget the topics in $\mathbf{h}_x$; it just encodes them in a different way.[2] This is a general problem in setups with alternating classifier-adversary optimization. To solve this issue, we propose using *multiple adversaries*, inspired by the "experience replay" approach used in reinforcement learning (O'Neill et al., 2010; Mnih et al., 2015). During the ith "topic training" phase, we train a new adversary adv_i (with parameters θ_{a_i} instead of retraining only one adversary over and over again. In the next "topic forgetting" phase, at each training step we pick adv_j at random from the pool of previously learned adversaries, $j \in_R \{1, \ldots, i\}$. By using multiple adversaries, we make it difficult for the classifier to encode topical information anywhere.

[2] We observe this in our analysis where the most salient features encoded after pretraining and topic forgetting phrase are the same.

5 Experimental Setup

5.1 Datasets

We evaluate our topical confound demotion method on the L1ID task. We show experiments with two datasets where L2 is English: the L2-Reddit dataset described in §2, and TOEFL17, a collection of essays authored by non-native English speakers who apply for academic studies in the US (Malmasi et al., 2017). This corpus reflects eleven L1s: Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish. The training data include 11,000 authors (1,000 per L1) and the development set has 1,100 essays per L1. We evaluate on the development set. Each essay is also marked with a prompt ID which was given to the authors to write the essay. There are 8 prompts in total, based on which we construct 8 versions of train and test set. In each version, we remove essays marked with one of the prompts from both the train and the development sets, and consider the removed essays from the development set an "out-of-domain" test set. We refer to the version where prompt "PK" is out-of-domain as "–PK" in the results (Table 3), $K \in \{0, \ldots, 7\}$.

5.2 Implementation Details

We tokenized and lowercased all the text using spaCy. Limiting our vocabulary to the most frequent 30,000 words in the training data, we replaced all out-of-vocabulary words with "UNK." We encoded each word using a word embedding layer (initialized at random and learned) and passed these embeddings to a bidirectional LSTM encoder (one layer for each direction) with attention ($h(x)$; Pryzant et al., 2018). Each LSTM layer had a hidden dimension of 128. We used two layered feed forward networks with a $\tanh$ activation function in the middle layer (of size 256), followed by a softmax in the final layer, as $c(.)$ and $\mathrm{adv}(.)$.

5.3 Baselines

We consider several baselines that are intended to capture the stylistic features of the texts, explicitly avoiding content.

Linear classifier with content-independent features (LR) Replicating Goldin et al. (2018), we trained a logistic regression classifier with three types of features: function words, POS trigrams, and sentence length, all of which are reflective of the style of writing. We deliberately avoided using content features (e.g., word frequencies).

Classification with no adversary on masked texts (LO-TOP-K) We mask the top-K words (based on log-odds scores) in *both* the train and the test sets (as in §2); we train the classification model again without training $\mathrm{adv}(.)$. After masking the top words, we expect patterns of writing style (and, therefore, L1) to become more apparent.

Adversarial training with gradient reversal (GR-LO) A common method of learning a confound-invariant representations is to use a gradient reversal layer (Beutel et al., 2017; Ganin et al., 2016; Pryzant et al., 2018; Elazar and Goldberg, 2018). The output of the encoder, $\mathbf{h}_x$, is passed through this layer before applying $\mathrm{adv}(.)$. This training setup usually proves too difficult to optimize, and often results in poor performance. That is, even if the performance of $\mathrm{adv}(.)$ is weak, $\mathbf{h}_x$ still ends up leaking information about the confound (Lample et al., 2019; Elazar and Goldberg, 2018). In the forward pass, this layer acts as identity whereas in the backward pass it multiplies the gradient values by $-\lambda$, essentially reversing the gradients before they go into the encoder. λ controls the intensity of the reversal (we used $\lambda = 0.2$).

LDA topics as confounds (ALT-LDA) We trained LDA on the training set and for each example in the training set, generated a probability distribution (over 50 topics), and used it as topical confound with our proposed learning setup, alternating classifier-adversary training.

6 Results

6.1 TOEFL17 Dataset

We begin with experiments on the TOEFL17 dataset, where predicting L1 is an easier task due to the lower proficiency of the authors. Table 3 reports the accuracy of our proposed model, denoted **ALT-LO**, compared to the logistic regression baseline (**LR**), and two adversarial baselines: one demotes latent log-odds-based topics via gradient reversal (**GR-LO**), and another uses our proposed novel learning procedure but demotes baseline LDA topics (**ALT-LDA**). We report both in-domain accuracy and out-of-domain results; the latter is obtained by averaging the accuracy of each set "–PK" over $K \in \{0, \ldots, 7\}$.

	In-Domain	Out-of-Domain
LR	55.3	50.9
GR-LO	12.7	13.6
ALT-LDA	59.1	50.1
ALT-LO	**61.9**	**60.4**

Table 3: Classification accuracy with topic-demoting methods, TOEFL dataset.

Our model strongly outperforms all baselines that demote confounds, in both classification setups. We observe in our experiments that gradient reversal is especially unstable and hyperparameter sensitive: it has been shown to work well with categorical confounds like domain type or binary gender, but in demoting continuous outputs like a topic distribution, we observe it is not effective. The proposed alternating training with multiple discriminators obtains better results, and replacing LDA with log-odds-based topics also improves both in-domain and (much more substantially) out-of-domain predictions, confirming the effectiveness of our proposed innovations.

A vanilla classifier without demoting confounds (denoted in §2 as NO-ADV) yields in-domain and out-of-domain accuracies of 62.0 and 58.3, respectively. We would expect that the better generalization power of our proposed model would come at a price of lower accuracy in-domain. Our goal is to capture the true signals of L1, rather than superficial patterns that are more frequent in the data and artificially boost the performance in NO-ADV settings. This is indeed what we observe.

For example, the text "…i agree with you on the prolonged war if the plc heartland (poland proper) was not as rich as it was i dont really see how we would been …" in the dataset is labeled as "Polish" instead of the gold label "Swedish" by the NO-ADV classifier, likely because of the mention of the term "poland", but the ADV-LO model predicts it correctly since it likely picks on other features that indicate non-fluency, like "we would been". Such naive classification errors become especially costly in making predictions about peo-

ple's demographic attributes: ethnicity, which often correlates with L1, but also gender, race, religion, and others (Hardt et al., 2016; Beutel et al., 2017).

6.2 L2-Reddit Dataset

Next, we experiment with L2-Reddit, a larger and more challenging dataset (since many speakers in the dataset are highly fluent, and the signal of their native language is weaker). The performance of the simple baselines on this dataset is shown in Table 4. The accuracy of the linear classifier is poor (compared to Table 1), perhaps because it fails to capture some contextual features learned by the neural network models. With LO-TOP-20, the performance on both test sets improves. It slightly degrades when more words are removed, perhaps because some words indicative of L1 are also removed.

	In-Domain	Out-of-Domain
LR	21.2	18.5
LO-TOP-20	38.7	21.9
LO-TOP-50	36.4	21.4
LO-TOP-100	35.8	21.2
LO-TOP-200	34.7	20.8

Table 4: Baseline classification accuracy on L2-Reddit.

Finally, we evaluate the impact of our novel training procedure and the quality of our proposed topical confound identification method. We compare our proposed solution, denoted ALT-LO, with two alternatives, as before, one with a different learning setup (GR-LO) and one with a different confound representation (ALT-LDA). Table 5 summarizes the results: our proposed learning procedure ALT-LO performs better than both the alternatives. Unsurprisingly, the model trained with gradient reversal (GR-LO) performs particularly poorly; this was our primary motivation to explore better learning techniques.

	In-Domain	Out-of-Domain
GR-LO	22.5	15.7
ALT-LDA	46.2	21.9
ALT-LO	**48.8**	**22.9**

Table 5: Classification accuracy with topic-demoting methods, L2-Reddit dataset.

To further confirm that the ALT-LO model is not learning topical features, we repeat the experiment presented in Table 1—masking the top K topical words (based on log-odds scores) from the test sets, but not retraining the models—now, with our proposed model ALT-LO. Table 6 shows that in contrast to standard models that do not demote topical confounds (as in Table 1), there is less degradation in the performance of ALT-LO. We conjecture that our model is stable to demoting topics because it learns relevant stylistic features, rather than spurious correlations.

	In-Domain	Out-of-Domain
ALT-LO	48.8	22.9
+MASK TOP-20	38.7	21.6
+MASK TOP-50	36.2	21.5
+MASK TOP-100	33.5	21.2
+MASK TOP-200	31.9	20.4

Table 6: Accuracy on the L2-Reddit dataset; the proposed model (ALT-LO) with different settings of the test sets.

7 Analysis

We present an analysis of what the models are learning, based on words they attend to for classification. We focus on the L2-Reddit dataset.

Following Pryzant et al. (2018), we generated a lexicon of most attended words by (1) running the model on the test set and saving the attention score for each word; and (2) for each word, computing its average attentional score and selecting the top-k words based on this score.

What emerges from this lexicon (Table 7) is a dramatic difference between the top indicative words in the various models. Whereas in the baseline model *all* the most indicative words are proper nouns, the ALT-LO model highlights exclusively function words. The proper nouns in the baseline model are all geographical terms directly associated with the L1s reflected in the L2-Reddit dataset: they are easy giveaways of the authors' L1s, but they are meaningless linguistically. In contrast, the function words highlighted in the ALT-LO model are mostly prepositions and determiners; it is well known that nonnative speakers are challenged by the use of prepositions (in any L2, English included). The distribution of determiners is also a challenge for nonnatives, and

the correct usage of *the* in particular is quite hard for learners to master. These challenges are evident from the most indicative words of our model. Observe also that the LO-TOP-50 model is somewhere in the middle: it includes some proper nouns (including geographical terms such as *eu* or *us*) but also several function words. A more detailed analysis of these observations is left for future work.

Recently, there has been a debate on whether attention can be used to explain model decisions (Serrano and Smith, 2019; Jain and Wallace, 2019; Wiegreffe and Pinter, 2019), we thus present additional analysis of our proposed method based on saliency maps (Ding et al., 2019). Saliency maps have been shown to better capture word alignment than attention probabilities in neural machine translation. This method is based on computing the gradient of the probability of the predicted label with respect to each word in the input text and normalizing the gradient to obtain probabilities. We use saliency maps to generate lexicons similar to the ones generated using attention. As shown in table 8, the top indicative words for baseline and LO-TOP-50 follow a similar pattern as the ones obtained with attention scores. In line with results in Table 7, salient words for ALT-LO are determiners and prepositions. However, saliency maps also reveal that our proposed approach still attends to some geographical terms that were not demoted by our classifier.

8 Related Work

Controlling for confounds in text Controlling for confounds is an active field of research, especially in the medical domain, where the common solution is to do random trials or propensity score matching (Rosenbaum and Rubin, 1985). Paul (2017) tackled the problem of learning causal associations between word features and class labels using propensity matching for the task of sentiment analysis. This method is not scalable to large text datasets as it involves training a logistic regression model for every word type. Tan et al. (2014) built models to estimate the number of retweets of Twitter messages and addressed confounding factors by matching tweets of the same author and topic. Reis and Culotta (2018) proposed a statistical technique called Pearl's back-door adjustment for text classification (Pearl, 2009). All these works focused on a bag-of-words model with lexical features only.

Adversarial training in text Much recent work focuses on learning textual representations that are invariant to selective properties of the text. This work used domain adaptation and transfer learning (Ganin et al., 2016; Tzeng et al., 2014; Xie et al., 2017), either to remove sensitive attributes such as demographic information (Li et al., 2018; Elazar and Goldberg, 2018; Beutel et al., 2017), or to understand costumer behavior for social science applications (Pryzant et al., 2018). Most of the work in this area, however, focuses on cases where these confounds are known in advance and their values are given along with the training data.

Native language identification The L1ID task was introduced by Koppel et al. (2005), who worked on the International Corpus of Learner English (Granger, 2003). The same experimental setup was adopted by several other authors (Tsur and Rappoport, 2007; Wong and Dras, 2009, 2011). Since the release of nonnative *TOEFL* essays by the Educational Testing Service (Blanchard et al., 2013), the task gained popularity and this dataset has been used for two L1ID Shared Tasks (Tetreault et al., 2013; Malmasi et al., 2017).

Malmasi and Dras (2017) report that the state of the art is a linear classifier with character n-grams and lexical and morphosyntactic features.

The best accuracy under cross-validation on the TOEFL17 dataset, which includes 11 native languages (with a rather diverse distribution of language families), was 85.2%.

The above works all identify the L1 of *learners*. Identifying the native language of advanced, fluent speakers is a much harder task. Goldin et al. (2018) addressed this task, using the L2-Reddit dataset with as many as 23 different L1s, all of them European and many which are typologically close, which makes the task even harder. They experimented with a variety of features, using logistic regression as the classifier, and achieved results as high as 69% accuracy with cross-validation; however, when testing their classifier outside the domain it was trained on (Reddit forums focusing on European issues), accuracy dropped to 36%.

9 Conclusion

We introduced a method to represent unknown confounds in text classification using topic models and log-odds scores, and a new general method

NO-ADV	sweden france greece finland poland spain greek germany french eu romania polish dutch german spanish swedish netherlands finnish
LO-TOP-50	eu 's 're 'm ' & uk us because 've am its nt english these usa nt here 'll especially correct pis de within
ALT-LO	the in to of that a i is and 't as from with by ? on but & they are about at because like was would have you

Table 7: The highest scoring words in lexicons generated using attention scores.

NO-ADV	poland greek romania greece france spain french sweden finland polish dutch spanish netherlands finnish german
LO-TOP-50	on 're even 'd up less things 'll doesn living majority sense talk level 've rights took number north
ALT-LO	the of to i a in greece romania france finland that for is french & you 't finnish

Table 8: The highest scoring words in lexicons generated using saliency maps.

with alternating optimization to learn textual representations which are invariant of confounds. We evaluated the proposed solution on the task of native language identification, and showed that it learns to make predictions using stylistic features, rather than focus on topical information.

The learning procedure we presented is general and applicable to other tasks that require learning invariant representations with respect to some attribute of text (some of which are discussed in §8). We plan to evaluate our proposed solution on other tasks where topics can be latent confounds, like predicting gender bias (Voigt et al., 2018). We leave this exploration for future work.

Acknowledgments

The authors acknowledge helpful input from the anonymous reviewers. This work was supported in part by NSF grants IIS-1812327 and IIS-1813153, by grant no. 2017699 from the United States-Israel Binational Science Foundation (BSF), and by grant no. LU 856/13-1 from the Deutsche Forschungsgemeinschaft. Finally, the authors also thank Anjalie Field, Biswajit Paria, Ella Rabinovich, and Gili Goldin for helpful discussions.

References

Martin Arjovsky and Léon Bottou. 2017. Towards principled methods for training generative adversarial networks. In *Proc. ICLR*.

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. 2017. Data decisions and theoretical implications when adversarially learning fair representations. In *Proc. FATML*.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native English. *ETS Research Report Series*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *JMLR*.

Shuoyang Ding, Hainan Xu, and Philipp Koehn. 2019. Saliency-driven word alignment interpretation for neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation*.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proc. EMNLP*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *JMLR*.

Gili Goldin, Ella Rabinovich, and Shuly Wintner. 2018. Native language identification with user generated content. In *Proc. EMNLP*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NeurIPS*.

Sylviane Granger. 2003. The International Corpus of Learner English: a new resource for foreign language learning and teaching and second language acquisition research. *Tesol Quarterly*.

Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *NeurIPS*.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *Proc. NAACL*.

Svetlana Kiritchenko and Saif M. Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proc. NAACL*.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Automatically determining an anonymous author's native language. *Intelligence and Security Informatics*.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *Proc. ICLR*.

Klas Leino, Matt Fredrikson, Emily Black, Shayak Sen, and Anupam Datta. 2019. Feature-wise bias amplification. In *Proc. ICLR*.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proc. ACL*.

Shervin Malmasi and Mark Dras. 2017. Native language identification using stacked generalization. ArXiv:1703.06541.

Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel Tetreault, Robert Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. 2017. A report on the 2017 native language identification shared task. In *Proc. Workshop on Innovative Use of NLP for Building Educational Applications*.

Jon D. McAuliffe and David M. Blei. 2008. Supervised topic models. In *NeurIPS*.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. 2018. Which training methods for GANs do actually converge? In *Proc. ICML*.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature*.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*.

Joseph O'Neill, Barty Pleydell-Bouverie, David Dupret, and Jozsef Csicsvari. 2010. Play it again: reactivation of waking experience and memory. *Trends in Neurosciences*.

Michael J. Paul. 2017. Feature selection as causal inference: Experiments with text classification. In *Proc. CoNLL*.

Judea Pearl. 2009. *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Reid Pryzant, Kelly Wang, Dan Jurafsky, and Stefan Wager. 2018. Deconfounded lexicon induction for interpretable social science. In *Proc. NAACL*.

Ella Rabinovich, Yulia Tsvetkov, and Shuly Wintner. 2018. Native language cognate effects on second language lexical choice. *TACL*.

Virgile Landeiro Dos Reis and Aron Culotta. 2018. Robust text classification under confounding shift. *Journal of Artificial Intelligence Research*.

Paul R. Rosenbaum and Donald B. Rubin. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*.

Kevin Roth, Aurelien Lucchi, Sebastian Nowozin, and Thomas Hofmann. 2017. Stabilizing training of generative adversarial networks through regularization. In *NeurIPS*.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proc. ACL*.

Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. In *Proc. ACL*.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proc. Workshop on Building Educational Applications Using NLP*.

Oren Tsur and Ari Rappoport. 2007. Using classifier features for studying the effect of native language on the choice of written second language words. In *Proc. Workshop on Cognitive Aspects of Computational Language Acquisition*.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014. Deep domain confusion: Maximizing for domain invariance. ArXiv:1412.3474.

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. RtGender: A corpus for studying differential responses to gender. In *Proc. LREC*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proc. EMNLP*.

Sze-Meng Jojo Wong and Mark Dras. 2009. Contrastive analysis and native language identification. In *Proc. Australasian Language Technology Association Workshop*.

Sze-Meng Jojo Wong and Mark Dras. 2011. Exploiting parse structures for native language identification. In *Proc. EMNLP*.

Qizhe Xie, Zihang Dai, Yulun Du, Eduard H. Hovy, and Graham Neubig. 2017. Controllable invariance through adversarial feature learning. In *NeurIPS*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proc. of EMNLP*.