Adversarial Training: addressing biases in non-topical classification and regression tasks

Anonymous ACL submission

Abstract

Many datasets for non-topical text classification contain topical shifts. Their presence in the data forces the classifiers to fit topics-related features instead of focusing on those relevant for the target class. The problem of topical shifts is also significant for the textual regression tasks. In our study, we estimate the effect of the topical shifts on performance of the classifiers and regressors in non-topical prediction tasks and try to reduce their impact by using adversarial methods. As two test tasks, we use sentiment analysis prediction on Amazon Reviews and identification of the education degree of the author on PASTEL. Each task is predicted as classification and regression. We show that Adversarial Domain Adaptation (ADA) helps to reduce the effect of topical shifts and to decrease the error in regression. Finally, we make a recommendation when to use ADA and how to select the hyperparameters for it.

1 Introduction

Non-topical text classification includes a wide range of tasks aimed at predicting a text property that is not connected directly to a text topic. For instance, predicting a text style, politeness, difficulty level, the age or the first language of its author, etc. Solutions for these tasks are applied in many areas such as information retrieval, language teaching, or linguistic research (Luu and Malamud, 2020).

One of the most challenging issues for the tasks of non-topical text classification is presence of topical shifts (Sharoff et al., 2010). This implicitly pushes the non-topical classifiers to rely on data features related to the topics instead of the ones relevant for the target non-topical variable. Besides, there are other kinds of shifts which could drastically affect performance of the classifiers. For instance, domain change or change of the distribution of a property related to demography. For example, shift of the gender distribution can cause gender-based biases (Dixon et al., 2018).

However, the problem of topical shifts is relevant not only for non-topical classification of texts but also for textual regression (Dayanik et al., 2022). Nevertheless, the problem of topical and distribution shifts is not widely researched for text regression.

One of the techniques that can potentially mitigate the topical biases of the non-topical text classification is causal models (Feder et al., 2020), (Maiya, 2021) because they have a functionality to make the classifiers less sensitive to the features that influence both the target variable and the text distribution, causing a spurious association. However, these methods require significant computational resources.

Another important algorithm is Adversarial Domain Adaptation (ADA) (Tzeng et al., 2017). It uses an adversarial loss to make the classification features less dependent on the domain of the training data. It supposes training a feature extractor, a domain discriminator, and a target classifier. The feature extractor and target classifier are trained to achieve high accuracy for the classification of the target class and at the same time deceive the domain discriminator to make it impossible to differentiate two domains. In contrast, the domain discriminator intends to classify the text domain correctly.

In our work, we compare adversarial models based on Adversarial Domain Adaptation (ADA) (Tzeng et al., 2017) with the baseline of BERT-based models. We show that usage of adversarial methods helps to increase the accuracy on the dataset under-represented in train and thereby reduce model reliance on the distribution shifts.

One group of the tasks for which the topical shifts affect significantly the quality of the predictions is text author profiling. It includes identification of the education degree of the author, gender classification, age regression. PASTEL (Kang et al., 2019) contains the information about the author genders, education degree, age, and even the polical views. It makes the dataset useful for evaluating the effects of the topical shifts.

In this study we:

1. show that the text classifiers and regressors based on the BERT architecture are sensitive to the topical shifts in the training data for non-topical classification tasks (subsection 6.1);

2. test ADA for sentiment analysis and education degree identification in order to decrease the deterioration of classification accuracy on the texts from the source under-represented in the train (section 5).

2 Related Studies

The problem of distribution shifts has a long history of research.

Some approaches (Basile, 2020) propose direct manipulations on the word embeddings. In contrast to our study, (Basile, 2020) does not apply adversarial methods and instead modifies the embeddings of the *weird* words - the words specific to the target domain, while we focus on contextual embeddings in pre-trained language models.

One of the methods for mitigating the distribution shifts is Adversarial Domain Adaptation (ADA) (Tzeng et al., 2017). In the original paper, the authors focus on transferring knowledge from a label-rich domain (source domain) to a label-scarce domain (target domain) for pervasive cross-domain for text classification, whilst our main objective is to minimise effect of the domain-related features. For Amazon reviews, it is the topical features. For PASTEL, it is the gender related ones.

There is a lot of research done for sentiment analysis on the Amazon Reviews dataset (Xie et al., 2020). However, all the studies released so far for Amazon Review train a classification model. Nevertheless, this task can be set as a regression problem, since the dataset has 5 possible labels and there is a clear order set on them. Moreover, there are no studies so far to reduce the effect of the topical shifts for both regression and classification on Amazon Reviews, although there are studies detection of the topical shifts for topical classification (Zirn et al., 2017).

3 Adversarial modification of the BERT-based architectures

3.1 Adversarial Domain Adaptation

ADA belongs to Unsupervised Domain Adaptation (Ramponi and Plank, 2020). It shows promising performance in numerous NLP tasks in recent years (Tzeng et al., 2017).

It usually consists of a shared feature extractor $f = G_f(x)$, a label predictor $y = G_y(x)$ and a domain discriminator $d = G_d(x)$. The domain discriminator d aims to distinguish the domain label between source and target, meanwhile the feature extractor f is trained to deceive the feature discriminator d. This adversarial training process can be formulated as

$$\min_{G_f, G_y} L_y(X_s, Y_s) - \lambda L_f(X_s, X_t),$$

$$\min_{G_d} L_d(X_s, X_t),$$

where L_y is the cross-entropy classification loss for the target label, L_f is the loss of the feature extractor.

4 Data

We take the Amazon Reviews dataset for training and testing our models for sentiment analysis. We select 24 topic categories with enough data from the original dataset and sample 8 thousand texts containing at least 50 words for each of them. Such a minimal length is selected to make the model training more stable. The threshold of 50 words approximately corresponds to the 10-percentile of the array of all the length in the dataset. We split randomly these datasets to train and test in a 3:1 ratio, so that the training subset contains 6000 texts, and the test one has 2000 texts.

For Amazon Reviews, we take the categories with >= 10K textual examples. We remove the texts containing less than 50 words. The remaining texts are randomly split to train and test. For each category, we take 6000 texts to train and 2000 texts to test. In the original dataset, the number of texts is higher than 6K for most categories. However, to exclude dependence of the metrics on the number of texts for different categories from Amazon Reviews, we randomly sample 6K examples for train and 2K examples for test for each category. It is shown in (Mosbach et al., 2021) that even with the training dataset of size 1000, BERT fine-tuning is stable enough. Hence, the sampling of a subset of

category	train					test				
	1	2	3	4	5	1	2	3	4	5
Arts	349	295	543	1021	3792	121	92	207	351	1229
Auto	429	337	569	1125	3540	148	115	175	349	1213
Books	213	275	656	1458	3398	76	96	234	480	1114
CDs	212	264	558	1375	3591	74	90	180	445	1211
Phones	476	426	661	1277	3160	139	148	196	440	1077
Clothing	326	436	775	1353	3110	123	148	251	441	1037
Music	140	225	448	1130	4057	50	66	156	415	1313
Electro	576	420	633	1321	3050	184	140	165	454	1057
Grocery	388	311	575	990	3736	137	115	193	325	1230
Home	491	365	563	1131	3450	156	115	192	377	1160
Industry	379	300	513	1159	3649	113	102	165	409	1211
Kindle	118	182	577	1603	3520	34	52	211	530	1173
Luxury Beauty	148	259	851	1821	2921	49	104	330	586	931
Movies	448	452	797	1484	2819	151	153	268	454	974
Musical Instruments	273	277	578	1287	3585	70	93	207	435	1195
Office	422	337	557	1258	3426	126	104	222	429	1119
Patio	518	325	571	1145	3441	175	113	197	379	1136
Pet	429	365	644	1048	3514	157	121	192	337	1193
Pantry	241	269	533	1104	3853	98	112	159	391	1240
SW	798	386	913	1641	2262	274	147	295	552	732
Sports	312	319	614	1291	3464	107	120	218	418	1137
Tools	384	358	567	1165	3526	140	113	182	378	1187
Toys	304	301	679	1300	3416	101	104	233	450	1112
Video Games	472	415	778	1502	2833	197	136	262	513	892

Table 1: The distribution of the review marks in the sampled Amazon Reviews Dataset. For each category, the train subset contains 6000 texts, the test subset contains 2000 texts.

Gender	train		test	
	NoDegree	Master	NoDegree	Master
Male	388	454	92	110
Female	571	310	138	75

Table 2: The distribution of the education degrees in the PASTEL Dataset

6K examples for each category should not affect the model performance.

Another dataset we use in our study is PASTEL (Kang et al., 2019). It contains a detailed information about the authors of the texts including the gender, age, education degree, country of origin, and even the political leaning. In the PASTEL dataset Table 2, the share of the texts written by the people with Master degree is higher among the male writers, although it is close to 50% for both genders.

5 Experiments

On the Amazon Reviews dataset, we solve a task of sentiment analysis. On PASTEL, we train a model for identification of the education degree of the author. We intend to train a classifier or regressor robust to the distribution shifts.

The main metric we use to compare the models for classification on PASTEL is f1 score. To evaluate the regression models, we use Mean Absolute Error (MAE). For sentiment analysis classification on the Amazon Reviews dataset, the macro f1 metric has a issue that it does not take into account the distance between the classes. To address it, we use Quadratic weighted kappa (QWK). It was actively used for ordinal classification (Amigó et al., 2020) and ordinal quantification (Sakai, 2021).

We test the robustness of the regressors and classifiers to shifts of the category/topic distribution in Amazon Reviews, and shift of the gender distribution in PASTEL. For each available category in Amazon Reviews and gender in PASTEL, we make a train dataset of 75% texts where this category/gender prevails and add 25% of texts from another category/gender, then compute the metrics of the trained model on the texts from category/gender under-represented in the train. In addition, we make the same experiment when the prevailing category/gender is present in 90% of the texts in train.

For all the experiments, we use multilingual BERT with the base configuration (12-layer, 768-hidden, 12-heads, 125M parameters) as a baseline for all the experiments. In all our experiments for classification and regression, learning rate=10^{-5}

is used, since this value is proposed in (Sun et al., 2019) and (Zou et al., 2021). For regression, we use SVR on top of the BERT embeddings. This technique is proved to be efficient for the regression tasks as essay scoring (Yang et al., 2020). It is helpful in situations when the embeddings spaces of the target classes are not linearly separable. We take the optimal value of the hyperparameter C from (Cherkassky and Ma, 2004): $C = mean(y) + 3\sigma(y)$, where y is the vector of the ground true labels, $\sigma(y)$ is the standard deviation.

Table 1 shows that the dataset for Amazon Reviews is highly unbalanced and the reviews with marks 4 and 5 prevail. For this reason, we perform upsampling to avoid degeneration of the classifier and regressor which cause the model to predict values in between of 4 and 5 otherwise.

6 Results

6.1 Model Sensitivity to the Confounder

We first evaluate the effect of the potential confounders before trying to reduce it. We train BERT-based classifiers to evaluate the f1 score for prediction of the category for Amazon Reviews and the gender for PASTEL.

The category classifier attains 59% f1 score when tries to identify the text category among the 24 possible variants Table 1. The PASTEL gender classifier attains around 80% accuracy. It reveals that BERT-based model is sensitive to the counfounder-related features. It could potentially cause the issue with the distribution shifts. For example, for Amazon Reviews it can negatively affect the performance of the category which was not present in train.

6.2 Amazon Reviews

For each category $category_i$, we train a baseline BERT model on it. We select two other categories $test\_category1_i$ and $test\_category2_i$, on which MAE of the trained model increases the most. For the category $category_i$, we create a train dataset consisting of 75% of the texts of $category_i$ and 25% of texts from $test\_category1_i$. The test subsets of categories $test\_category1_i$ and $test\_category2_i$ are used to test the model in subsubsection 6.2.2. We make an analogical experiment subsubsection 6.2.1 for classification but for QWK instead of MAE.

6.2.1 Classification

Table 3 shows the results for classification on Amazon Reviews. The values of QWK for ADA with $\lambda = 0.05$ are in general higher than those for the base BERT and ADA with $\lambda = 0.2$.

6.2.2 Regression

Table 4 shows the result for regression on Amazon Reviews when the prevailing category makes 75% or 90% texts of the train dataset correspondingly. When the share of the prevailing category is 75%, the best result on the test1 and test2 for most categories is attained with $\lambda = 0.2$ and $\lambda = 0.05$. In the case if the share of the prevailing category is 90%, the optimal lowest MAE for most categories is achieved with $\lambda = 0.5$.

We can see that for the vast majority of categories, ADA helps to reduce the MAE. In addition, the more shifted the trained dataset is, the higher value of λ is needed to get the optimal result. Moreover, regardless of the degree of bias of the training dataset, the values $\lambda = 0.05$ and $\lambda = 0.2$ for most categories still decrease the value of MAE.

6.3 PASTEL. Education

Table 5 shows that ADA improves the f1 score when tested on the female texts regardless of the train dataset. However, there is no improvement of the texts written by male authors.

The results for regression Table 6 on PASTEL show that usage of ADA decreases MAE significantly on the test dataset. In most cases, the value of λ when the MAE is the lowest is $\lambda = 0.2$. Moreover, the MAE decrease on the test dataset is more remarkable that the on the dev dataset. It matches the intuition that ADA improves the quality of regression on the data where the text distribution is different from that in train.

We also try different values of λ - the ones lower than 0.05 and those higher than 0.5. The close the λ to 0, the closer are the predictions to those of the BERT classifier. In contrary taking $\lambda > 0.5$ makes the predictions less similar to those of BERT but they do not improve the value attained by $\lambda = 0.2$.

7 Comparison with LLM

Nowadays, Large Language Models (LLMs) are getting more popular and are being used for solving a wider set of tasks. For example, the models like ChatGPT (Touvron et al., 2023b) and LLaMA (Touvron et al., 2023a) are used for zero-shot classification (Wang et al., 2023).

train	test1	test2	dev				test1				test2			
75%	25%	0%	0.0	0.05	0.2	0.5	0.0	0.05	0.2	0.5	0	0.05	0.2	0.5
Arts	Music	Kindle	0.667	-0.051	0.018	**0.029**	0.555	-0.035	0.042	**0.053**	0.525	-0.035	0.041	**0.069**
Auto	Music	Kindle	0.6	**0.011**	-0.027	0.005	0.583	**0.021**	-0.014	-0.013	0.537	-0.045	-0.189	**0.025**
Books	Music	Home	0.623	0.002	-0.023	**0.032**	0.596	-0.008	-0.019	**0.037**	0.543	0.004	-0.128	**0.015**
Cell	Music	Kindle	0.572	0.06	**0.096**	0.08	0.452	0.104	**0.141**	0.102	0.403	0.115	**0.126**	0.125
Cloth	Music	Kindle	0.646	**0.046**	0.032	-0.001	0.553	**0.06**	0.024	-0.021	**0.53**	-0.068	-0.0	-0.092
Music	Pet	Arts	0.613	**0.04**	0.03	0.034	0.586	**0.058**	0.04	0.046	0.607	**0.083**	0.054	0.062
Electro	Kindle	Music	0.648	0.007	-0.016	**0.025**	0.591	0.015	**0.018**	0.007	**0.541**	-0.028	-0.074	-0.034
Grocery	Music	Kindle	0.644	-0.028	**0.006**	-0.008	**0.592**	-0.006	-0.01	-0.021	0.53	-0.016	0.001	**0.08**
Home	Music	Kindle	0.647	**0.024**	-0.01	0.009	0.544	0.016	-0.027	**0.024**	0.522	0.008	-0.047	**0.06**
Industry	Kindle	Music	0.552	**0.089**	0.041	0.061	0.494	**0.118**	0.057	0.081	0.395	**0.105**	-0.028	0.102
Kindle	Music	Home	0.605	-0.017	-0.014	**0.023**	0.549	0.01	0.005	**0.041**	**0.585**	-0.074	-0.012	-0.041
Luxury	Music	Kindle	**0.633**	-0.036	-0.017	-0.017	0.576	**0.006**	-0.044	-0.029	0.519	-0.051	0.019	**0.032**
Movies	Music	Pet	**0.664**	-0.021	-0.006	-0.051	**0.643**	-0.039	-0.024	-0.026	**0.58**	-0.034	-0.058	-0.075
M. Inst	Music	Kindle	0.56	0.059	**0.065**	0.03	0.554	**0.062**	0.06	-0.003	0.441	0.05	**0.109**	0.07
Office	Music	Kindle	0.624	**0.015**	-0.02	-0.003	0.573	**0.017**	-0.045	0.006	0.514	0.036	**0.051**	-0.043
Patio	Music	Kindle	0.597	-0.005	-0.024	**0.006**	0.531	**0.033**	0.01	0.022	0.477	0.039	0.02	**0.046**
Pet	Music	Kindle	**0.635**	-0.081	-0.016	-0.013	**0.569**	-0.106	-0.024	-0.03	**0.556**	-0.259	-0.026	-0.015
Pantry	Music	Kindle	0.617	-0.018	-0.004	**0.001**	0.564	-0.009	0.002	**0.004**	0.493	-0.038	0.008	**0.032**
SW	Music	Kindle	0.687	**0.0**	-0.001	-0.043	**0.59**	-0.025	-0.028	-0.089	0.449	0.046	0.09	**0.109**
Sports	Music	Kindle	0.596	-0.005	-0.007	**0.0**	**0.558**	-0.057	-0.013	-0.02	**0.503**	-0.055	-0.052	-0.065
Tools	Music	Kindle	0.62	**0.027**	-0.021	0.021	**0.594**	-0.031	-0.077	-0.016	**0.578**	-0.035	-0.09	-0.041
Toys	Music	Kindle	**0.674**	-0.01	-0.026	-0.013	**0.602**	-0.017	-0.047	-0.012	**0.611**	-0.042	-0.014	-0.007
Video	Music	Kindle	0.642	**0.029**	0.015	-0.043	0.609	**0.019**	-0.009	-0.115	**0.568**	-0.015	-0.05	-0.107
CDs	Music	Pet	0.612	-0.022	**0.008**	-0.035	0.613	**0.008**	-0.008	-0.063	**0.521**	-0.115	-0.036	-0.163
		avg	0.624	0.005	0.003	0.005	0.57	0.009	0.0	-0.002	0.522	-0.018	-0.012	0.006
		best	4	9	4	7	7	10	2	5	10	2	3	9
90%	10%	0%	0.0	0.05	0.2	0.5	0.0	0.05	0.2	0.5	0	0.05	0.2	0.5
Auto	Music	Kindle	0.511	-0.043	**0.102**	0.049	0.456	-0.052	**0.13**	0.059	0.371	-0.094	**0.185**	0.054
Books	Music	Home	0.603	0.006	**0.025**	0.007	0.55	-0.018	**0.026**	0.009	0.524	0.063	**0.064**	0.024
Cell	Music	Kindle	0.629	**0.054**	-0.014	-0.022	0.512	**0.062**	-0.006	-0.029	0.474	**0.078**	-0.035	-0.053
Cloth	Music	Kindle	**0.664**	-0.002	-0.003	-0.01	**0.533**	-0.033	-0.002	-0.001	**0.563**	-0.046	-0.021	-0.083
Music	Pet	Arts	0.568	-0.046	0.015	**0.029**	0.538	-0.073	0.01	**0.024**	0.619	-0.089	-0.025	**0.014**
Electro	Kindle	Music	**0.662**	-0.006	-0.07	-0.037	0.551	**0.005**	-0.02	0.005	**0.519**	-0.069	-0.075	-0.09
Grocery	Music	Kindle	**0.631**	-0.007	-0.004	-0.0	**0.552**	-0.002	-0.016	-0.015	**0.586**	-0.066	-0.108	-0.091
Home	Music	Kindle	0.658	-0.013	-0.12	**0.017**	0.517	-0.001	-0.101	**0.019**	**0.504**	-0.115	-0.161	-0.028
Industry	Kindle	Music	0.604	**0.043**	0.022	-0.004	0.592	**0.021**	0.014	-0.033	**0.523**	-0.006	-0.061	-0.095
Kindle	Music	Home	0.529	0.023	**0.081**	0.035	0.465	0.042	**0.081**	0.053	0.469	0.041	**0.065**	0.052
Luxury	Music	Kindle	0.61	-0.062	-0.028	**0.003**	0.496	-0.083	-0.006	**0.042**	**0.523**	-0.335	-0.19	-0.017
Movies	Music	Pet	**0.656**	-0.023	-0.042	-0.049	**0.623**	-0.04	-0.057	-0.067	**0.588**	-0.054	-0.1	-0.146
M.Instr	Music	Kindle	0.597	-0.029	-0.003	**0.002**	0.562	-0.039	-0.004	**0.008**	**0.528**	-0.01	-0.095	-0.135
Office	Music	Kindle	**0.626**	-0.025	-0.072	-0.094	**0.54**	-0.04	-0.07	-0.114	**0.526**	-0.038	-0.053	-0.1
Patio	Music	Kindle	0.598	**0.003**	-0.003	0.002	0.495	**0.027**	0.018	0.026	0.497	-0.021	-0.012	**0.039**
Pet	Music	Kindle	0.401	0.212	**0.217**	0.141	0.301	**0.207**	0.202	0.13	0.318	0.109	**0.2**	0.082
Pantry	Music	Kindle	0.57	0.004	0.01	**0.025**	0.498	-0.012	0.001	**0.027**	0.456	**0.078**	0.049	0.035
SW	Music	Kindle	**0.7**	-0.007	-0.026	-0.024	**0.567**	-0.038	-0.001	-0.04	0.421	0.047	**0.15**	0.04
Sports	Music	Kindle	**0.636**	-0.071	-0.052	-0.119	**0.581**	-0.152	-0.097	-0.224	**0.525**	-0.12	-0.03	-0.078
Tools	Music	Kindle	0.608	**0.006**	-0.025	-0.025	0.528	**0.012**	-0.0	-0.028	0.456	0.034	**0.08**	-0.126
Toys	Music	Kindle	0.626	-0.001	**0.043**	0.041	0.528	-0.031	0.046	**0.046**	0.54	0.034	0.054	**0.059**
Video	Music	Kindle	0.58	0.01	-0.086	**0.029**	0.511	-0.023	-0.173	**0.017**	0.444	-0.063	0.032	**0.049**
CDs	Music	Pet	0.577	0.044	0.061	**0.067**	0.567	0.015	0.04	**0.064**	**0.514**	-0.068	-0.056	-0.022
		avg	0.602	0.003	0.001	0.003	0.525	-0.011	0.001	-0.001	0.499	-0.031	-0.006	-0.027
		best	7	4	5	7	6	6	3	8	11	2	6	4

Table 3: Classification on Amazon Reviews, Quadratic Weighted Kappa, for 75% and 90% of the train topic labels. For $\lambda = 0.0$ (vanilla BERT) we report the absolute value of Quadratic Weighted Kappa, for $\lambda = 0.05, 0.2, 0.5$ the difference of QWK between the corresponding absolute value of QWK and $\lambda = 0.0$ is reported. The best highest values of QWK and the highest increases in QWK correspondingly are highlighted in bold. For each value of λ, we also report the number of categories for which it attains the highest value of QWK

train	test1	test2	dev				test1				test2			
75%	25%	0%	0.0	0.05	0.2	0.5	0.0	0.05	0.2	0.5	0	0.05	0.2	0.5
Arts	Kindle	CDs	0.749	**0.044**	-0.005	-0.032	0.716	-0.017	**0.238**	-0.304	0.904	**0.275**	0.226	0.198
Auto	CDs	Kindle	**0.585**	-1.042	-0.105	-0.487	0.563	-0.474	**0.019**	-0.194	**0.574**	-0.921	-0.104	-0.135
Books	Arts	M.Instr	0.818	**0.147**	-0.225	0.034	0.789	0.122	-0.569	**0.186**	0.859	**0.205**	-0.499	0.203
CDs	Arts	Kindle	0.737	**0.107**	-0.087	-0.48	0.756	**0.194**	-0.169	-0.675	0.812	**0.207**	-0.046	0.121
Cell	CDs	Kindle	1.433	**0.73**	0.317	0.309	1.479	0.679	**0.687**	0.286	1.423	**0.868**	0.551	0.74
Clothing	Kindle	Music	0.713	-0.072	**0.102**	-0.142	0.694	0.023	-0.006	**0.138**	0.899	0.092	**0.332**	0.165
Music	Arts	Luxury	**0.497**	-0.065	-0.169	-0.324	**0.549**	-0.038	-0.027	-0.426	**0.573**	-0.095	-0.031	-0.313
Electro	CDs	Kindle	0.991	**0.169**	0.006	0.048	0.976	0.113	**0.41**	0.374	0.807	0.077	0.093	**0.201**
Grocery	Kindle	CDs	0.735	**0.083**	0.066	-0.241	0.685	**0.192**	0.125	-0.442	0.819	**0.227**	-0.018	0.008
Industry	CDs	Kindle	**0.711**	-0.035	-0.026	-0.312	0.735	**0.151**	-0.228	-0.629	0.762	**0.139**	0.085	0.099
Kindle	M.Instr	Arts	**0.583**	-0.076	-0.192	-0.49	0.646	0.072	**0.087**	-0.965	0.604	0.03	**0.061**	-0.71
Luxury	CDs	Kindle	0.834	-0.022	**0.203**	0.055	0.922	0.075	**0.287**	-0.03	0.876	0.063	0.149	**0.312**
Movies	Arts	M.Instr	0.828	0.094	**0.193**	0.187	0.733	0.138	0.185	**0.192**	0.843	**0.243**	0.059	0.23
M.Instr	Kindle	CDs	**0.651**	-0.064	-0.034	-0.456	0.654	**0.17**	-0.107	-0.199	0.657	**0.069**	0.049	-0.119
Office	Kindle	CDs	0.682	-0.103	**0.018**	-0.191	0.617	0.04	**0.105**	-0.038	0.809	**0.21**	0.207	0.025
Patio	CDs	Kindle	0.974	-0.315	-0.189	**0.019**	0.954	0.271	**0.312**	-0.286	0.849	-0.256	0.086	**0.239**
Pet	CDs	Kindle	1.259	-0.075	**0.354**	0.325	1.303	0.253	0.447	**0.606**	1.233	0.367	0.305	**0.411**
Pantry	CDs	Kindle	**0.777**	-0.105	-0.228	-0.3	0.807	**0.254**	0.04	-0.299	0.698	**0.143**	0.047	-0.284
SW	CDs	Music	0.971	**0.058**	-0.196	0.003	0.989	-0.156	**0.223**	0.108	0.965	-0.13	0.127	**0.137**
Sports	Kindle	CDs	0.889	0.044	0.097	**0.218**	0.851	**0.389**	0.317	0.311	0.927	**0.313**	0.308	0.139
Tools	CDs	Kindle	0.846	**0.036**	-0.085	-0.311	0.839	**0.272**	-0.41	0.038	0.842	**0.217**	0.208	0.147
Toys	Music	Kindle	0.746	-0.311	**0.084**	-0.063	0.708	-0.045	-0.018	**0.23**	0.72	0.013	**0.139**	-0.163
Video	Music	Kindle	0.888	0.129	0.226	**0.239**	0.893	0.4	**0.429**	0.309	0.743	**0.152**	0.015	-0.411
Home	Kindle	Music	1.05	0.158	**0.389**	0.021	1.019	0.191	0.289	**0.344**	0.982	0.231	**0.436**	0.17
		avg	1.247	-0.03	0.032	-0.148	1.242	0.204	0.167	-0.085	1.261	0.171	0.174	0.088
		best	6	8	7	3	1	7	10	6	2	13	4	5
90%	10%	0%												
Arts	Kindle	CDs	**0.544**	-0.124	-0.142	-0.135	**0.534**	-0.016	-0.163	-0.012	0.637	**0.005**	-0.059	-0.042
Auto	CDs	Kindle	**0.73**	-0.198	-0.124	-0.119	0.694	0.011	**0.121**	-0.324	0.691	0.031	0.015	**0.058**
Books	Arts	M. Instr	**0.601**	-0.405	-0.308	-0.639	**0.619**	-0.131	-0.171	-1.102	**0.721**	-0.067	-0.097	-1.002
CDs	Arts	Kindle	**0.805**	-0.856	-0.178	-0.144	0.806	-1.001	-0.378	**0.007**	0.841	-0.077	-0.058	**0.141**
Cell	CDs	Kindle	1.322	**0.585**	-0.45	0.285	1.33	**0.729**	0.505	0.602	1.305	**0.625**	-0.696	0.491
Clothing	Kindle	Music	**0.628**	-0.101	-0.017	-0.04	0.589	0.01	0.005	**0.101**	0.642	-0.046	-0.401	**0.049**
Music	Arts	Luxury	**0.542**	-0.093	-0.029	-0.192	**0.599**	-0.021	-0.121	-0.366	**0.59**	-0.066	-0.086	-0.263
Electro	CDs	Kindle	1.114	**0.188**	0.124	-0.174	1.083	0.334	**0.478**	0.364	1.056	**0.45**	0.095	0.295
Grocery	Kindle	CDs	**0.606**	-0.067	-0.023	-0.422	0.586	**0.099**	0.089	0.073	0.736	-0.048	**0.109**	0.025
Industry	CDs	Kindle	1.489	-0.071	**0.84**	0.536	1.486	-0.583	**0.826**	0.189	1.428	0.197	**0.912**	0.871
Kindle	M. Instr	Arts	0.668	-0.172	**0.054**	-0.116	0.764	-0.15	0.14	**0.153**	0.73	-0.236	0.051	**0.106**
Luxury	CDs	Kindle	**0.651**	-0.013	-0.066	-0.098	0.706	**0.075**	-0.054	-0.316	0.763	0.163	-0.309	**0.182**
Movies	Arts	M. Instr	1.117	-0.327	**0.341**	0.269	1.152	-0.797	0.521	**0.538**	1.109	-0.654	0.427	**0.513**
M. Instr	Kindle	CDs	0.677	-0.204	-0.438	**0.112**	0.678	-0.164	-0.488	**0.138**	**0.662**	-0.25	-0.03	-0.357
Office	Kindle	CDs	**0.579**	-0.376	-0.799	-0.185	**0.546**	-0.091	-1.461	-0.0	**0.595**	-0.183	-0.825	-0.082
Patio	CDs	Kindle	0.999	0.083	**0.16**	0.111	0.998	-0.007	0.219	**0.399**	0.966	**0.271**	0.264	0.24
Pet	CDs	Kindle	1.48	**0.733**	0.367	0.605	1.477	**0.858**	0.306	0.445	1.436	**0.815**	0.681	0.22
Pantry	CDs	Kindle	0.672	-0.007	-0.067	**0.075**	0.703	0.092	**0.14**	0.058	0.644	-0.041	**0.059**	-0.098
SW	CDs	Music	1.078	**0.407**	-0.203	0.251	1.122	0.463	0.169	**0.535**	1.096	0.448	0.18	**0.556**
Sports	Kindle	CDs	0.881	**0.121**	-0.156	-0.44	0.896	**0.433**	-0.588	-1.049	1.003	**0.383**	0.083	-0.119
Tools	CDs	Kindle	0.737	**0.002**	-0.034	-0.288	**0.736**	-0.011	-0.231	-0.367	0.705	-0.035	**0.137**	-0.369
Toys	Music	Kindle	1.2	0.466	0.431	**0.628**	1.149	0.62	**0.635**	0.605	1.185	0.487	0.243	**0.613**
Video	Music	Kindle	0.762	**0.132**	-0.141	-0.01	0.663	**0.149**	-0.349	0.074	0.755	**0.202**	0.002	-0.0
Home	Kindle	Music	**0.602**	-0.167	-0.3	-0.231	**0.584**	-0.419	-0.436	-0.193	**0.612**	-0.001	-0.634	-0.891
		avg	1.28	-0.029	-0.072	-0.023	1.281	0.03	-0.018	0.034	1.307	0.148	0.004	0.071
		best	10	7	4	3	6	6	5	7	5	7	4	8

Table 4: Regression on Amazon Reviews, MAE, for 75% and 90% of the train topic labels. For $\lambda = 0.0$ (vanilla BERT) we report the absolute value of MAE, for $\lambda = 0.05, 0.2, 0.5$ the difference of MAE between $\lambda = 0.0$ and the corresponding absolute value of MAE is reported. The lowest MAE and the biggest declines in MAE correspondingly are highlighted in bold. For each value of λ, we also report the number of categories for which it attains the lowest value of MAE

train on	test on	dev				test			
		0	0.05	0.2	0.5	0	0.05	0.2	0.5
Female 75%, Male 25%	Male	0.45	0.07	**0.12**	0.09	**0.886**	-0.001	-0.009	-0.146
Female 90%, Male 10%	Male	0.45	0.07	0.12	**0.19**	**0.894**	-0.029	-0.032	-0.021
Male 75%, Female 25%	Female	**0.9**	**0**	**0**	**0**	**0.123**	-0.001	-0.004	-0.005
Male 90%, Female 10%	Female	**0.9**	**0**	**0**	**0**	0.120	0.021	0.003	**0.030**
avg		0.675	0.04	0.06	0.07	0.506	-0.003	-0.01	-0.036
best		2	2	3	3	3	0	0	1

Table 5: Classification on PASTEL. For $\lambda = 0.0$ (vanilla BERT) we report the absolute value of f1 score, for $\lambda = 0.05, 0.2, 0.5$ the difference of f1 between the corresponding absolute value of f1 and $\lambda = 0.0$ is reported. The best highest values of f1 and the highest increases in f1 correspondingly are highlighted in bold. For each value of λ, we also report the number of categories for which it attains the highest value of f1 score.

train_on	test_on	dev				test			
		0	0.05	0.2	0.5	0	0.05	0.2	0.5
Female 75%, Male 25%	Male	0.385	**0.019**	**0.019**	0.017	0.18	0.013	**0.031**	0
Female 90%, Male 10%	Male	0.427	0.025	**0.039**	0.014	0.236	0.032	**0.039**	0.011
Male 75%, Female 25%	Female	0.165	0.007	**0.012**	-0.009	0.385	**0.019**	0.017	0.017
Male 90%, Female 10%	Female	0.160	**0.002**	-0.013	-0.017	0.386	0.011	**0.016**	**0.016**
avg		0.284	0.013	**0.014**	0.001	0.297	0.019	**0.026**	0.011
best		0	2	**3**	0	0	1	**3**	1

Table 6: Regression on PASTEL, relative improve of MAE. The lowest values of MAE are in bold. For $\lambda = 0.0$ (vanilla BERT) we report the absolute value of MAE, for $\lambda = 0.05, 0.2, 0.5$ the difference of MAE between $\lambda = 0.0$ and the corresponding absolute value of MAE is reported. The lowest MAE and the biggest declines in MAE correspondingly are highlighted in bold. For each value of λ, we also report the number of experiments for which it attains the lowest value of MAE.

model/prompt	M		F	
	f1	mae	f1	mae
LLaMA zero-shot	0.676	**0.327**	0.390	**0.366**
LLaMA 9M + 1F	**0.705**	0.490	**0.521**	0.648
LLaMA 1M + 9F	0.664	0.460	0.484	0.592

Table 7: PASTEL. LLaMA 3.2 3B results. The highest f1 score and the lowest MAE are in bold.

In order to estimate applicability of the LLMs to the tasks of textual regression and non-topical text classification, we run LLaMA3.2-3B to compare their MAE and f1 score to those of the BERT-based models as well as the inference time. We try both zero-shot and few-shot prompts to understand whether adding some instructions to the prompt is helpful for increasing the quality of the predictions. We try two variants of few-shot prompts: 9 male examples + 1 female example; 9 female examples + 1 male example.

To make the time evaluation representantive, we run all the experiments for PASTEL on the same GPU provided by Google Colab. Inference for LLaMA takes more time and consumes more computational resources. The time needed to get the responses on the test for the BERT-based model is 32 seconds. In contrast, LLaMA 3.2 needs 254 seconds (or 4 minutes) in the zero-shot mode and 1534 seconds (or 26 minutes) in the few-shot mode.

Table 7 shows the performance of LLaMA on the PASTEL dataset. It reveals that a model based on the base BERT architecture attains the quality comparable to the LLMs like LLaMA but with a lower consumption of the computational resources and with a much lower inference time.

It shows that for some tasks it is still efficient to fine-tune relatively small BERT-based models instead of using LLMs out-of-box. It also confirms (Benayas et al., 2024) claiming that for tasks involving natural language understanding, encoder-only models generally outperform decoder-only models, all while demanding a fraction of the computational resources.

8 Conclusion

By conducting a range of experiments, we make the following conclusions:

1. The adversarial methods are efficient in combatting the topical shifts for the task of text regression and non-topical classification.

2. However, when the topic of the text is completely absent from the training dataset, the quality of the non-topical classification with ADA is the same or slightly worse than that with the vanilla BERT.

3. The more the training dataset is shifted, the higher value of λ is needed to attain the lowest MAE. However, this value should still be ≤ 0.5.

4. The optimal value of the λ hyperparameter is around 0.2 for PASTEL. For Amazon Reviews, the optimal value of λ is eather 0.05 for test1 or 0.2 for test2.

5. The values of λ between 0.05 and 0.2 could be recommended to be used by default when the degree of the topical shifts in the training data is unclear.

Limitations

Our experiments were provided for the English language. However, in theory, the optimal hyperparameters may depend crucially on the language-based properties of the dataset. Moreover, our study we take two datasets for regression and classification: Amazon Reviews and PASTEL. They are taken from different sources and represent different genres of texts that makes our experiments more foundamental than if we took datasets of the same genre. However, the texts available on the internet may belong to a much wider variety of genres. Thereby, our study does not fully represent real-world language diversity.

Ethical Considerations

In our research, we do not label the data ourselves. Instead, we use public datasets Amazon Reviews and PASTEL, which are already labeled by their authors. Those datasets are publicly available and only include the information voluntarily shared by the authors of the texts. In addition, these datasets respect anonymity of the authors of the included texts and do not disclose information about the names of the authors and their contacts such as email, phone numbers or links to the social media. Besides, one of the frequent ethical problems in the modern NLP applications is potential biases about specific groups of people. In our research, we try to reduce the reliance of BERT-based models on the gender related features for prediction of the education degree. It helps to reduce the potential gender-based biases.

References

Enrique Amigó, Julio Gonzalo, Stefano Mizzaro, and Jorge Carrillo de Albornoz. 2020. An effectiveness metric for ordinal classification: Formal properties and experimental results. *ACL 2020*.

Valerio Basile. 2020. Domain adaptation for text classification with weird embeddings. *CEUR-WS*.

Alberto Benayas, Miguel Angel Sicilia, and Marçal Mora-Cantallops. 2024. A comparative analysis of encoder only and decoder only models in intent classification and sentiment analysis: navigating the trade-offs in model size and performance. *Language Resources and Evaluation*.

Vladimir Cherkassky and Yunqian Ma. 2004. Practical selection of svm parameters and noise estimation for svm regression. *Neural networks, 17(1):113–126, 2004*.

Erenay Dayanik, Ngoc Thang Vu, and Sebastian Padó. 2022. Bias identification and attribution in nlp models with regression and effect sizes. *Northern European Journal of Language Technology, Volume 8*.

Lucas Dixon, Jeffrey Sorensen, and Nithum Thain. 2018. Measuring and mitigating unintended bias in text classification. *2018 AAAI/ACM Conference*.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2020. Causalm: Causal model explanation through counterfactual language models. *arXiv preprint arXiv:2005.13407*.

Dongyeop Kang, Varun Gangal, and Eduard Hovy. 2019. (male, bachelor) and (female, ph.d) have different connotations: Parallelly annotated stylistic language dataset with multiple personas. *arXiv preprint arXiv:1909.00098*.

Alex Luu and Sophia A. Malamud. 2020. Non-topical coherence in social talk: A call for dialogue model enrichment. *ACL*.

Arun S. Maiya. 2021. Causalnlp: A practical toolkit for causal inference with text. *arXiv preprint arXiv:2106.08043*.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *ICLR 2021*.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in nlp—a survey. *Coling*.

Tetsuya Sakai. 2021. Evaluating evaluation measures for ordinal classification and ordinal quantification. *ACL 2021*.

Serge Sharoff, Zhili Wu, and Katja Markert. 2010. The Web library of Babel: evaluating genre collections. In *Proc Seventh Language Resources and Evaluation Conference, LREC*, Malta.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? *arXiv preprint arXiv:1905.05583*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *ArXiv*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023b. Summary of chatgpt-related research and perspective towards the future of large language models. *ArXiv*.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhiqiang Wang, Yiran Pang, and Yanbin Lin. 2023. Large language models are zero-shot text classifiers. *ArXiv*.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2020. Unsupervised data augmentation for consistency training. *NeurIPS 2020*.

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. *EMNLP 2020*.

Cacilia Zirn, Goran Glavas, Federico Nanni, Jason Eichorst, and Heiner Stuckenschmidt. 2017. Classifying topics and detecting topic shifts in political manifestos. *Proceedings of the International Conference on the Advances in Computational Analysis of Political Text*.

Han Zou, Jianfei Yang, and Xiaojian Wu. 2021. Unsupervised energy-based adversarial domain adaptation for cross-domain text classification. *ACL 2021*.

A Appendix

A.1 Amazon Categories

Table 8 provides the original names of the categories from Amazon Reviews.

A.2 The baseline BERT models

Figure 1 shows the difference between the MAE of the model trained on this category on Amazon Reviews dataset. The rows of the table denote the category on which the model was trained. The columns denote the category on which the model was tested. The number in the cell $(row\_id, column\_id)$ is the differences between the MAE on the test subset for the category number $column\_id$ of the regression model trained on the category row_id and the one trained on the category $column\_id$.

We can see that almost all the number are positive. It means that changing the category on which the model is trained deteriorates the performance on the test if the testing texts belong to a different category. Moreover, most numbers are lower than 1.0. It means that that in most cases the model makes a mistakes within one sentiment label and the prediction remains more or less adequate.

Besides, there are two categories called *Digital Music* and *Kindle Store* for which the MAE delta is the highest for most categories. It could mean that the texts of these categories are much different from those of other categories.

short	full
Arts	Arts Crafts and Sewing
Auto	Automotive
Books	Books
CDs	CDs and Vinyl
Cell	Cell Phones
Cloth	Clothing
Music	Digital Music
Electro	Electronics
Grocery	Grocery and Gourmet Food
Home	Home and Kitchen
Industry	Industrial and Scientific
Kindle	Kidle Store
Luxury	Luxury Beauty
Movies	Movies And TV
M. Instr	Musical Instruments
Office	Office Products
Patio	Patio Lawn and Garden
Pet	Pet Supplies
Pantry	Prime Pantry
SW	Software
Sports	Sports and Outdoors
Tools	Tools and Home Improvement
Toys	Toys and Games
Games	Video Games

Table 8: The full names of the categories for Amazon Reviews

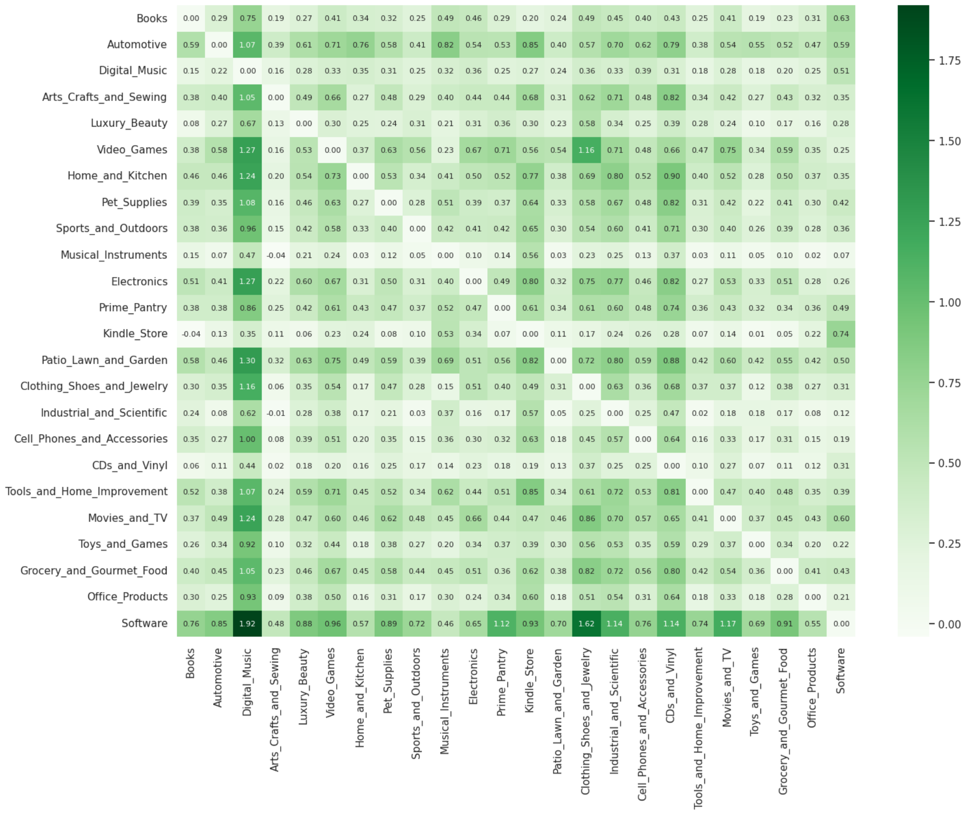

Figure 1: MAE delta on the Amazon Reviews test dataset. On the X-axis is the category on which the model was trained. On the Y-axis is the category where the model was tested.