IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 31 January 2026, accepted 16 February 2026, date of publication 6 March 2026, date of current version 18 March 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3671578

RESEARCH ARTICLE

Adversarial Training and Contrastive Loss: Addressing Biases in Non-Topical Classification and Text Regression Tasks

MIKHAIL LEPEKHIN[1], SERGE SHAROFF[2], AND ILYA MAKAROV[3,4]
[1]Moscow Independent Research Institute of Artificial Intelligence, 141700 Moscow, Russia
[2]University of Leeds, LS2 9JT Leeds, U.K.
[3]AXXX, 105064 Moscow, Russia
[4]Trusted AI Research Center, RAS, 109004 Moscow, Russia

Corresponding author: Mikhail Lepekhin (mikhail.lepekhin@corsearch.com)

The work of Ilya Makarov on Section 2 was supported by a grant, provided by the Ministry of Economic Development of the Russian Federation (agreement dated June 20, 2025 No. 139-15-2025-011, identifier 000000C313925P4G0002).

ABSTRACT Topical shifts—differences between the topic distribution in training and test data—can substantially reduce the performance of non-topical text classification and regression models. In this work, we propose a modified sampling procedure for the contrastive loss to make it more robust to the topical shifts. We also investigate how to increase robustness to such shifts by combining Adversarial Domain Adaptation (ADA) with a modified contrastive learning objective. The method jointly minimizes label prediction loss, adversarial domain loss, and a contrastive loss designed to reduce the dependence on topic-related confounders. We evaluate the approach on two tasks: sentiment prediction in Amazon Reviews and user education degree identification in the PASTEL dataset. Across both tasks, the proposed method improves robustness under topical shifts, reducing the mean absolute error in text regression and increasing the Quadratic Weighted Kappa (QWK) in text classification. These results demonstrate that jointly applying ADA and confounder-aware contrastive learning can mitigate topic sensitivity and yield more reliable text prediction models.

INDEX TERMS Adversarial learning, Bert, contrastive loss, non-topical classification, text regression, text classification.

I. INTRODUCTION

Non-topical text classification refers to NLP tasks that predict a latent property of a text that is not directly tied to its topical content. Typical examples include style or politeness prediction, readability estimation, and author profiling tasks such as age, gender, or first-language identification [1]. These tasks are widely used in information retrieval, computer-assisted language learning, and linguistic analysis.

A key challenge in non-topical classification is the presence of *topical shifts*, i.e., variations in topic distributions between training and test data [2], [3], even when using the latest large language models (LLMs) [4]. Such shifts implicitly encourage models to rely on spurious topical cues rather than on features genuinely associated with the target non-topical variable. Beyond topical variation, other forms of distributional shift (e.g., domain changes or demographic imbalance) can substantially degrade classifier performance. For instance, a shift in the gender distribution may amplify gender-related biases [5]. Similar issues arise in *text regression* tasks, where distribution shifts can lead to biased or unstable predictions [6]. However, in contrast to classification, the effects of topical and distributional shifts on text regression remain underexplored.

Let X denote a text input and Y denote a non-topical target variable (categorical or continuous), and D denote a domain or topic label. We consider the supervised learning setting in which a model is trained on data drawn from one or more source domains D_s and evaluated on data from an underrepresented or unseen target domain D_t. The objective

The associate editor coordinating the review of this manuscript and approving it for publication was Sajid Ali.

is to learn a predictor $f_\theta(X)$ that is *invariant* to shifts in D, i.e., robust to changes in topical or domain distributions, while accurately predicting Y.

Domain adaptation techniques offer a promising direction for reducing model reliance on topic-specific features. In particular, *Adversarial Domain Adaptation* (ADA) [7] aims to learn domain-invariant representations by jointly training (i) a feature extractor, (ii) a domain discriminator, and (iii) a target classifier or regressor. The feature extractor and classifier are optimized to achieve high prediction quality while simultaneously preventing the discriminator from distinguishing source and target domains.

The problem of topical shifts affects not only text classification tasks but also tasks such as Named Entity Recognition (NER) [8], [9] and Question Answering (QA) [10]. For these tasks, the problem of topical shifts can be aggravated by label sparsity, which has led to benchmarks such as ColdQA [11] and CLIFT [12] for testing the robustness of models for Question Answering. However, the embeddings used for text classification and regression differ from those trained for NER and QA. Because question answering and NER require fine-grained, context-dependent representations, methods optimized for these tasks do not necessarily promote the invariant, holistic features needed for robust text classification or regression.

Datasets, such as *Amazon Reviews* [13] provide a natural testbed for evaluating robustness to topical shifts due to their coverage of more than twenty product categories. The strong class imbalance and ordinal nature of sentiment labels motivate the use of metrics that account for ordered misclassification; hence we adopt Cohen's Quadratic Weighted Kappa (QWK) [14], [15].

Topical shifts also substantially affect author profiling tasks. The PASTEL dataset [16], which contains demographic and socio-political annotations (e.g., gender, age, education level), enables systematic evaluation of such effects in both classification and regression settings.

This work investigates the sensitivity of Transformer-based models to topical and distribution shifts in non-topical text classification and regression. We further evaluate the extent to which adversarial domain adaptation mitigates these effects. Our main contributions are as follows:

1) We demonstrate that BERT-based classifiers and regressors exhibit significant performance degradation under topical shifts across multiple non-topical tasks (subsection VII-A).
2) We evaluate ADA-based adversarial models for sentiment regression and education-degree classification, showing that adversarial training improves robustness on underrepresented domains and reduces reliance on spurious topical features.
3) We modify the sampling algorithm for Contrastive Loss so that it is adjusted to reduce the effect of the topical shifts and show that it improves the metrics of the BERT-based models for rating review classification and regression and education degree classification and regression.
4) To the best of our knowledge, we are the first to train regressors and classifiers with a joint loss combining ADA and the modified Contrastive Loss to reduce the effect of topical shifts and demonstrate a 3-4 point improvement in QWK and up to a 5% decrease in MAE for the tasks of classification and regression, respectively.

Overall, our findings highlight persistent vulnerabilities of state-of-the-art NLP models to distribution shifts and show that adversarial approaches constitute a viable direction for developing more domain-invariant predictors.

II. RELATED WORK

The problem of distribution shifts has been extensively studied across natural language processing and machine learning. Prior work has explored a wide range of techniques for mitigating mismatches between training and deployment domains, including embedding manipulation, adversarial learning, causal modeling, and robust classification frameworks.

Early efforts to address distribution shifts examine direct manipulations of word embeddings. For example, [17] modifies the embeddings of domain-specific "*weird*" words that are characteristic of a target corpus. While effective for lexical-level adaptation, such methods do not incorporate adversarial objectives and operate on static embeddings. In contrast, our work focuses on contextual representations produced by pre-trained language models, and leverages adversarial training to suppress domain-related features.

There are various methods to evaluate robustness of models for NLP tasks. Certain approaches [18] focus on exposing the models to adversarial attacks and estimating how the quality of predictions changes with the intensity of the attacks. There are studies [19] that emulate the adversarial attacks using dropouts. However, our primary focus is to test the model robustness to natural rather than adversarial topical shifts, in which the test distribution is different in the absence of an attack. This approach is similar to [3] and [20] where the authors perform testing on controllably biased datasets. For example, in our case, we control the topical shift in Amazon Reviews by training a sentiment classifier on reviews from one category and testing it on a different category.

Adversarial Domain Adaptation (ADA) [7] is a widely used approach for reducing the discrepancy between source and target distributions. The original ADA formulation addresses knowledge transfer from a label-rich source domain to a label-scarce target domain for cross-domain text classification. In the last few years, numerous improvements to Adversarial Domain Adaptation have appeared [21], including Energy-based ADA [22], Meta ADA [23]. However, the general principle of those methods is the same. Unlike ADA, which aims to improve target-domain performance through representation alignment, our objective is

to minimize the influence of domain-specific attributes—topical information in Amazon Reviews and gender-related features in PASTEL—rather than to perform classical domain transfer.

A substantial body of work investigates sentiment analysis on the Amazon Reviews dataset [24]. Traditional approaches rely on classical machine learning models. For instance, [25] employs Bag-of-Words vectorization paired with Decision Tree and Logistic Regression classifiers. Other studies explore a range of statistical and neural architectures. Reference [26] compares Logistic Regression, Naïve Bayes, SVM, CNN, and RNN models for 5-star rating prediction and further introduces the Hybrid Sequential Binary Classification (HSBC) framework. However, their analysis does not incorporate modern Transformer-based architectures such as BERT.

More recent work leverages BERT-based models for sentiment prediction [27], [28], [29]. While these studies align with our use of contextualized language models, they do not employ adversarial mechanisms to suppress domain-specific biases or to improve robustness under distribution shifts.

Existing approaches on Amazon Reviews predominantly treat the problem as a classification task, despite the inherently ordinal structure of the five rating labels. Moreover, prior work evaluates models using accuracy or F1 score, metrics that do not account for error severity. In contrast, we adopt Cohen's Quadratic Weighted Kappa [14], which better reflects the ordered nature of the labels. It has been actively used for ordinal classification [30] and ordinal quantification [31].

Furthermore, none of the reviewed studies explicitly target the mitigation of topical shifts for either regression or classification on Amazon Reviews. Although topical shift detection has been examined in other domains, such as political topic classification [32], methods for *reducing* topical bias in sentiment prediction remain underexplored.

Causal models [33], [34], [35] and large causal models [36], [37] describe mechanisms for reducing spurious correlations by removing the influence of confounding variables that affect both the text distribution and the target label. Such methods offer a principled way to mitigate topical biases. However, state-of-the-art causal NLP techniques typically require significant computational resources and complex model architectures. Due to these practical constraints, we do not incorporate causal modeling in our experiments, though it remains a promising complementary direction.

There have already been studies combining adversarial and contrastive learning [38], [39], [40]. However, those studies were applied to the problem of domain adaptation where the goal was to increase accuracy on a specific domain instead of making the text embeddings domain- or topic-invariant. Moreover, the authors do not propose a modification of the contrastive loss that stimulates the text embeddings to be domain-invariant. Finally, those studies do not consider applying the combination of adversarial loss and contrastive learning to the task of regression.

Overall, prior work has examined adversarial adaptation, embedding manipulation, and sentiment analysis on Amazon Reviews, but no existing study addresses the reduction of topical biases in both classification and regression settings using adversarial and contrastive training on contextualized embeddings. Our work fills this gap by introducing a modified contrastive loss that explicitly suppresses domain-related features while maintaining predictive QWK score. In addition, we propose a joint mechanism combining the adversarial and contrastive loss functions. The main novelty of this paper lies in a confounder-aware positive sampling strategy for contrastive learning, its integration into a joint training objective, and its application to text regression tasks.

III. METHODOLOGY

To mitigate the effects of distributional and topical shifts, we propose an adversarial modification to BERT-based architectures. This section details the components of our approach, including Adversarial Domain Adaptation, Contrastive Loss, and our proposed joint training objective.

A. ADVERSARIAL DOMAIN ADAPTATION

Adversarial Domain Adaptation (ADA) is a technique rooted in Unsupervised Domain Adaptation (UDA) [41]. It has demonstrated promising performance across numerous NLP tasks in recent years [7].

The architecture typically consists of a shared feature extractor $f = G_f(x)$, a label predictor $y = G_y(f)$, and a domain discriminator $d = G_d(f)$. The domain discriminator d aims to distinguish between the source and target domains, while the feature extractor f is trained to deceive the discriminator, thereby learning domain-invariant representations. This adversarial training process can be formulated as a minimax game:

$$\min_{G_f, G_y} \mathcal{L}_y(X_s, Y_s) - \lambda_{\text{ADA}} \mathcal{L}_d(X_s, X_t), \tag{1}$$

$$\min_{G_d} \mathcal{L}_d(X_s, X_t), \tag{2}$$

where $\mathcal{L}_y$ denotes the cross-entropy classification loss for the target label, and $\mathcal{L}_d$ represents the domain classification loss. The hyperparameter λ_{ADA} controls the intensity of the adversarial adaptation.

B. CONTRASTIVE LOSS

Our contrastive approach is based on the loss function proposed in [42]. The primary motivation is to enforce compactness in the embedding space for texts sharing the same target label.

Consider a batch I. Let $A(i) = I \setminus \{i\}$ denote the set of indices for all elements in the batch excluding i. Furthermore, let $P(i)$ denote the set of indices for all texts where the target label matches that of text i. The contrastive loss is defined as:

$$\mathcal{L}_{CL} = \sum_{i \in I} \frac{-1}{|P(i)|} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)}, \tag{3}$$

TABLE 1. The distribution of the review labels in the sampled Amazon reviews dataset. For each category, the training subset contains 6000 texts, the test subset contains 2000 texts.

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

where z_i, z_p, and z_a are the embeddings of the anchor text x_i, the positive sample x_p, and the negative sample x_a, respectively. The parameter τ denotes the temperature; a higher τ reduces the sensitivity of the loss function to the dot product $z_i \cdot z_p$.

The total loss function for this component is formulated as:

$$\mathcal{L}_{total} = (1 - \lambda_{\text{CL}})\mathcal{L}_{ce} + \lambda_{\text{CL}}\mathcal{L}_{cl}, \tag{4}$$

where $\mathcal{L}_{ce}$ is the standard cross-entropy loss.

The hyperparameters for this method include:

- λ_{CL}: The weight assigned to the contrastive loss term.
- $K = |P(i)|$: The number of positive elements to sample.
- τ: The temperature scaling factor.
- Sampling Strategy: The method for selecting positive examples (e.g., random sampling or *similarity capping*). As shown in [42], model performance is comparable between random sampling and similarity capping.

Similarity capping involves selecting positive examples whose embeddings are closest to the anchor text *i* in the vector space.

In this work, we propose the following modifications to the standard contrastive framework:

- **Confounder-aware Positive Sampling:** We extend the definition of positive pairs. We select examples that share the same target label, but have the *opposite* confounder value (see 1). This encourages the model to learn label-relevant features that are invariant to the confounding variable.
- **Loss Normalization:** To ensure consistent scaling across different loss components and to remove dependence on batch size and the number of positive examples per text, we normalize the contrastive loss $\mathcal{L}_{CL}$. This is achieved by dividing $\mathcal{L}_{CL}$ by a normalization factor

Algorithm 1 Confounder-Aware Sampling for Contrastive Loss

Require: I - the batch to process
Require: *last_epoch_output*
Require: *last_epoch_emb*
Require: *last_epoch_confounder*
Require: *ground_true*
for each $i \in I$ **do**
  potential_positives $\leftarrow$ []
  for $j := 1$ to n **do**
    if $ground\_true_j == ground\_true_i$ and $last\_epoch\_output_j \neq ground\_true_j$ and $last\_epoch\_confounder_j \neq last\_epoch\_confounder_i$ **then**
      potential_positives.add(j)
    end if
  end for
end for
if *len*(*potential_positives*) == 0 **then**
  return []
end if
allow_replacement $\leftarrow$ (*len*(*potential_positives*) $< k$)
$P \leftarrow$ *random_sample*(*potential_positives*, *size* $= k$, *replacement* $=$ *allow_replacement*)
return P

N_{norm}, which bounds its magnitude by 1:

$$N_{norm} = \frac{B}{k} \cdot \max(|\log m|, |\log M|), \tag{5}$$

where B represents the batch size and

$$m := \frac{\exp(-1/\tau)}{(B-1)\cdot\exp(1/\tau)}, \tag{6}$$

$$M := \frac{\exp(1/\tau)}{(B-1)\cdot\exp(-1/\tau)}, \tag{7}$$

We normalize the contrastive loss in this way only to simplify the process of selecting the hyperparameters. Otherwise, a loss without normalization becomes wildly different for different tasks and text domains, making it impossible to compare the optimal value of λ_{CL}.

- **Label Grouping (Amazon Reviews):** To adapt the contrastive loss for the '5'-star Amazon Reviews dataset, we aggregate classes '1' and '2' into a *low score* subgroup, and classes '4' and '5' into a *high score* subgroup. An example is considered positive if its target class belongs to the same subgroup as the anchor. The label grouping follows Amazon's official feedback definition: ratings of '4' and '5' stars correspond to positive feedback, ratings of '1' and '2' stars correspond to negative feedback, and '3'-star ratings are treated as neutral. This grouping aligns the contrastive objective with the semantic interpretation of sentiment in the Amazon Reviews platform.

C. JOINT OBJECTIVE: CONTRASTIVE LOSS + ADA

We further propose a combined architecture integrating both Contrastive Loss and Adversarial Domain Adaptation.

Let $\mathcal{L}_{ADA}$ denote the loss function minimized with ADA Equation 1.

$$\mathcal{L}_{ADA} = \mathcal{L}_y(X_s, Y_s) - \lambda_{\text{ADA}}\mathcal{L}_d(X_s, X_t), \tag{8}$$

The final objective function is defined as:

$$\min_{G_f, G_y} \lambda_{CL}\mathcal{L}_{cl} + (1-\lambda_{CL})\mathcal{L}_{\text{ADA}} \tag{9}$$

$$\min_{G_d} \mathcal{L}_d(X_s, X_t), \tag{10}$$

This hybrid method incorporates the hyperparameters from both the ADA and Contrastive Loss frameworks to simultaneously align domains and refine class boundaries.

We evaluate all values of $\lambda_{ADA} \in \{0.01, 0.05, 0.2, 0.5, 0.8\}$ and $\lambda_{CL} \in \{0.02, 0.05, 0.1, 0.2, 0.5, 0.9\}$. The best combination we found is $\lambda_{ADA} = 0.05$, $\lambda_{CL} = 0.05$.

IV. DATA

We use the Amazon Reviews dataset to train and test our models for sentiment analysis. We select 24 topic categories with enough data from the original dataset and sample 8,000 texts containing at least 50 words for each of them. Such a minimum length is selected to stabilize the model training. The threshold of 50 words approximately corresponds to the 10th percentile of the array of all the lengths in the dataset.

TABLE 2. The distribution of the education degrees in the PASTEL dataset.

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

We randomly split these datasets into training and test sets in a 3:1 ratio, so that the training subset contains 6000 texts, and the test one has 2,000 texts.

For Amazon Reviews, we select the categories with at least 10K textual examples. We remove the texts containing less than 50 words. The remaining texts are randomly split into training and test sets. For each category, we select 6000 texts to train and 2000 texts to test. In the original dataset, the number of texts exceeds 6K for most categories. However, to eliminate 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 [43] that even with a training dataset of size 1,000, BERT fine-tuning is stable enough. Hence, sampling a subset of 6K examples for each category should not affect the model performance.

Another dataset we use in our study is PASTEL [16]. It contains 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 texts written by people with a Master's degree is higher among the male writers, although it is close to 50% for both genders.

V. METRICS

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 the issue that it does not take into account the distance between the classes. To address it, we use Quadratic Weighted Kappa (QWK) [14]. Its formula:

$$\kappa = 1 - \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\, O_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\, E_{ij}} \tag{11}$$

$$w_{ij} = \frac{(i-j)^2}{(N-1)^2}, \tag{12}$$

$$E_{ij} = \frac{(\sum_k O_{ik})(\sum_l O_{lj})}{\sum_{i,j} O_{ij}}, \tag{13}$$

where O is the observed rating histogram (confusion matrix), E is the expected histogram under the assumption of independence, w_{ij} is the quadratic weight between rating categories i and j, N is the total number of rating categories. For Amazon Reviews, $N = 5$.

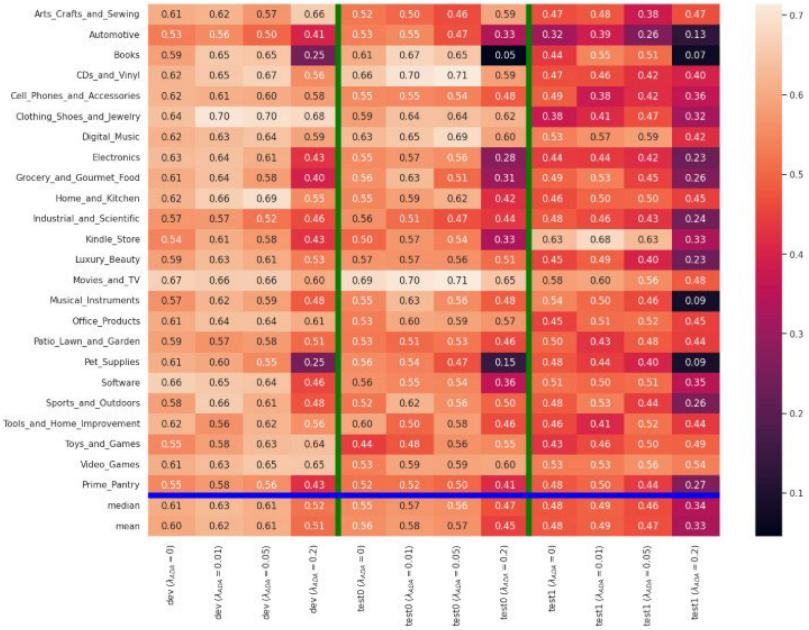


FIGURE 1. Rating classification with Adversarial Domain Adaptation (ADA) on Amazon Reviews, Quadratic Weighted Kappa.

VI. EXPERIMENT SETUP

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

We test the robustness of the regressors and classifiers to shifts in the category/topic distribution in Amazon Reviews and to shifts in the gender distribution. For each available category in Amazon Reviews and gender in PASTEL, we construct a training dataset of 75% of the texts where this category/gender prevails and add 25% of the texts from another category/gender, then compute the metrics of the trained model on the texts from category/gender underrepresented in the training data. In addition, we carry out 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. In our experiments, we use Adam with learning rate of 10^{-5} for both classification and regression, since this value is proposed in [22] and [44]. For all the experiments with Contrastive Loss, we use $\tau = 0.5, K = 5$ because these values show the best performance in [42]. In order to reduce the computational cost of our experiments, we avoid applying similarity capping.

For regression on PASTEL, we use Support Vector Regression (SVR) [45] on top of the BERT embeddings. The process consists of two parts. We first train both our vanilla BERT and adversarial regressors end-to-end as usual. Then we extract the embeddings of those models and train SVR on them using exactly the same training dataset as for training the neural regressors. This technique is proven to be efficient for the regression tasks such as essay scoring [46]. This approach is useful when the embedding spaces of the target classes are not linearly separable. We select the optimal value of the hyperparameter C from [47]: $C = mean(y) + 3\sigma(y)$, where y is the vector of the ground-truth labels, $\sigma(y)$ is the standard deviation.

We apply the end-to-end approach to classification for PASTEL and both classification and regression tasks on Amazon Reviews, as otherwise the end-to-end regression model on PASTEL does not converge. Whether we use ADA, Contrastive Loss, and their combination or not, the MAE on both male and female texts for the end-to-end model is around 0.4, which is only slightly better than a random classifier. We suppose that the reason for such an outcome is the noise in the PASTEL dataset because its authors relied heavily on crowd-sourced labels. There are studies [48], [49] showing that SVR can be helpful for text regression on noisy datasets.

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

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 training 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 VII-B2. We conduct an analogous

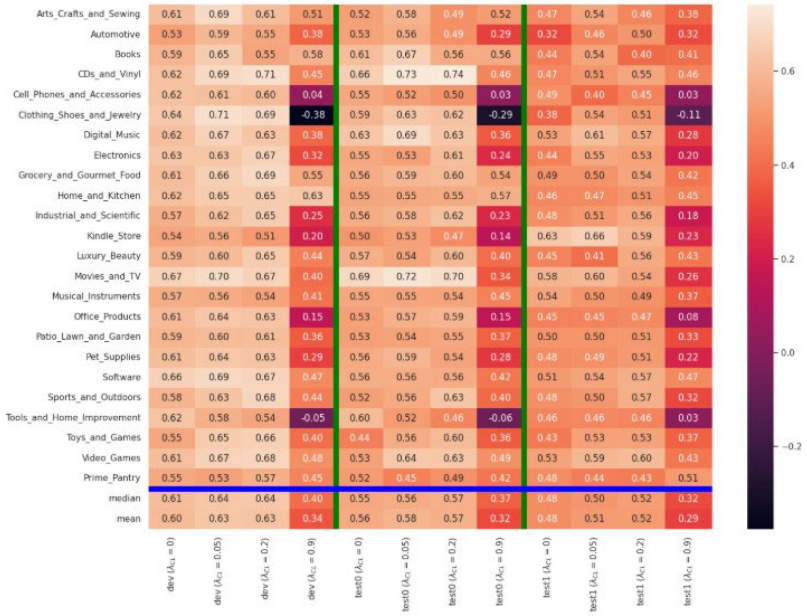


FIGURE 2. Rating classification with contrastive loss on Amazon Reviews, Quadratic Weighted Kappa.

experiment for classification (see subsubsection VII-B1), but using QWK instead of MAE.

VII. RESULTS

A. MODEL SENSITIVITY TO THE CONFOUNDER

We first evaluate the effect of the potential confounders before attempting to reduce it. We train BERT-based classifiers to evaluate macro F1 score for prediction of the category for Amazon Reviews and the gender for PASTEL.

The category classifier attains 59% macro F1 score when it tries to identify the text category among the 24 possible variants Table 1. The PASTEL gender classifier attains around 80% F1 score. This reveals that BERT-based models are sensitive to confounder-related features. It could potentially cause distributional shift issues. For example, for Amazon Reviews it can negatively affect the performance of the category which is not present in the training data.

B. AMAZON REVIEWS

1) RATING CLASSIFICATION

Figure 1 and Figure 2 show the results for classification on Amazon Reviews. The values of QWK for ADA with $\lambda_{ADA} = 0.05$ are generally higher than those for the base BERT and ADA with $\lambda_{ADA} = 0.2$. Moreover, ADA improved the QWK score for 15 categories, whilst Contrastive Loss improved the QWK score for 19 categories. It can be seen that beyond a certain value of λ_{ADA}, the classification quality begins to decline. For Contrastive Loss, a significant decline in QWK occurs with $\lambda_{CL} \geq 0.8$.

The optimal range of values of λ_{CL} is between 0.01 and 0.2. For ADA, it is between 0.01 and 0.05.

We also experiment with applying ADA and Contrastive Loss together. The best combination we found is $\lambda_{ADA} = 0.05$, $\lambda_{CL} = 0.05$. According to Figure 3, combining ADA and Contrastive Loss increases the value of QWK. It shows that the adversarial methods can improve performance of each other when applied together.

2) RATING REGRESSION

According to Figure 4 and Figure 5, Adversarial Domain Adaptation and contrastive loss reduces Mean Absolute Error.

Figure 6 and Figure 3 show that the model trained with joint adversarial and contrastive loss attains the results either comparable with the best of these methods separately or even slightly better. The optimal pair of hyperparameters (λ_{ADA} and λ_{CL}) is slightly different, but still close to the optimal values selected separately for each method.

Figure 4 shows the result for regression on Amazon Reviews when the prevailing category makes up 75% of the training dataset. When the share of the prevailing category is 75%, the best result on the test1 and test2 for most categories is attained with $\lambda_{ADA} = 0.2$ and $\lambda_{ADA} = 0.05$. In the case where the share of the prevailing category is 90%, the optimal lowest MAE for most categories is achieved with $\lambda_{ADA} = 0.5$.

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

Figure 10 and Figure 11 show that Contrastive Loss attains more accurate results for regression than the vanilla BERT, but ADA is still better. For Contrastive Loss the optimal λ_{CL} is around 0.2.

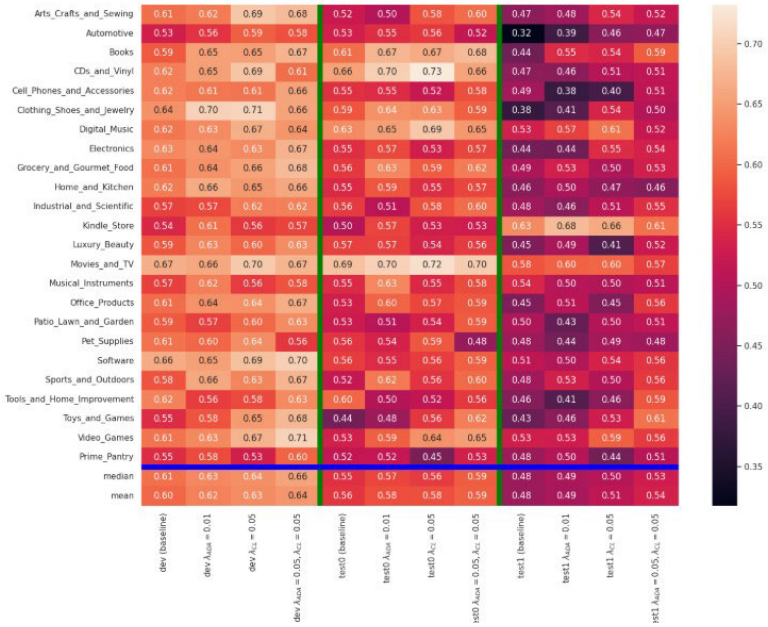


FIGURE 3. Rating classification with ADA, Contrastive Loss, and the joint loss function on Amazon Reviews, Quadratic Weighted Kappa.

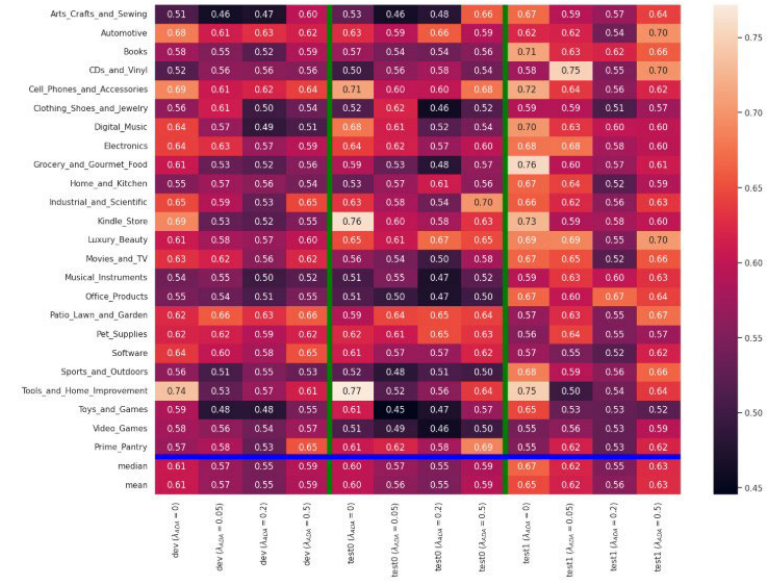


FIGURE 4. Rating regression with adversarial Domain Adaptation (ADA) on Amazon Reviews, Mean Absolute Error. The results on dev and test for $\lambda_{ADA} = 0.0, 0.05, 0.02, 0.5$.

C. PASTEL: CLASSIFICATION AND REGRESSION OF THE EDUCATION DEGREE

Figure 7 shows that ADA improves the F1 score when tested on the female texts regardless of the training dataset. However, there is no improvement for texts written by male authors. For contrastive learning, the outcome is similar: Figure 8.

The results for regression Figure 9 on PASTEL show that usage of ADA decreases MAE significantly on the test dataset. In most cases, the value of λ_{ADA} when the MAE is the lowest is $\lambda_{ADA} = 0.2$. Moreover, the MAE decrease on the test dataset is more remarkable than that 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 evaluate different values of λ_{ADA} - the ones lower than 0.05 and those higher than 0.5. The closer λ_{ADA} is to 0, the closer the predictions are to those of the BERT classifier. In contrast, selecting $\lambda_{ADA} > 0.5$ pushes the predictions closer to those of BERT, but does not improve the accuracy and and F1 scores attained by $\lambda_{ADA} = 0.2$.

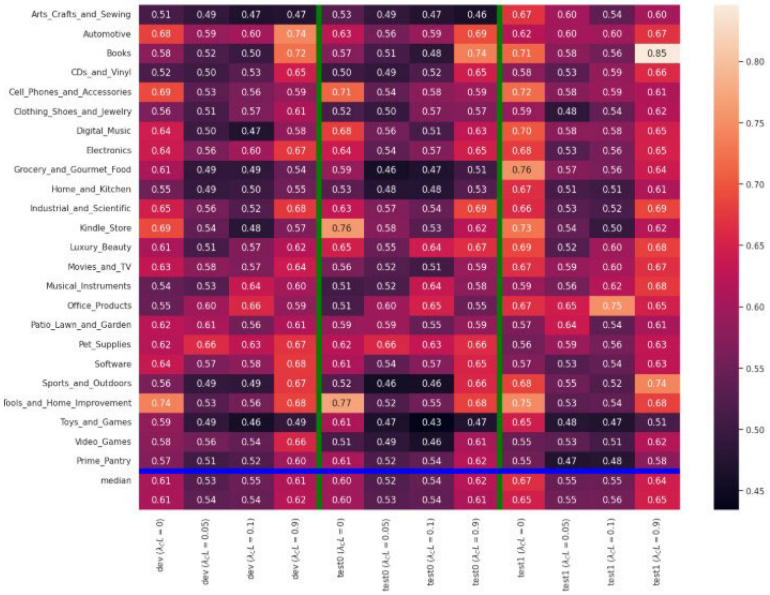


FIGURE 5. Rating regression with contrastive loss on Amazon Reviews, MAE. The results on dev and test for $\lambda_{CL} = 0.0, 0.05, 0.01, 0.5$.

FIGURE 6. Rating regression with ADA, Contrastive Loss, and the joint loss on Amazon Reviews, MAE. The results on dev and test for the best values of λ_{ADA} and λ_{CL}.

On Amazon Reviews, Contrastive Loss shows better performance than ADA. On PASTEL, the situation is the opposite. It might be caused by noise from the PASTEL dataset.

VIII. ABLATION STUDY

In this section, we conduct a series of ablation and diagnostic experiments to better understand the training data and the behavior of the proposed methods and to disentangle the contributions of their individual components. We analyze model errors using confusion matrices and qualitative failure examples, examine the structure of learned feature spaces through embedding visualizations, and compare the confounder-aware contrastive loss with its standard variant. In addition, we investigate robustness under varying degrees of distributional shift and analyze performance asymmetries across demographic subgroups. Together, these experiments provide insights into the mechanisms underlying the observed performance gains and the robustness properties of adversarial and contrastive training.

We present confusion matrices Figure 12 for the classification models we trained. The results show that most classification errors occur when the model predicts a class neighboring the ground-truth label rather than a distant

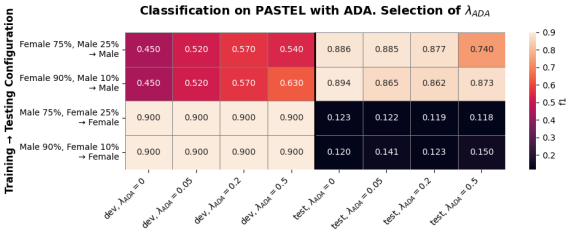


FIGURE 7. Text classification of the author's educational degree on PASTEL using Adversarial Domain Adaptation. The plot shows the dependence of MAE on the training configuration and the parameter λ_{ADA}. When $\lambda_{ADA} = 0.0$, the model corresponds to the standard BERT.

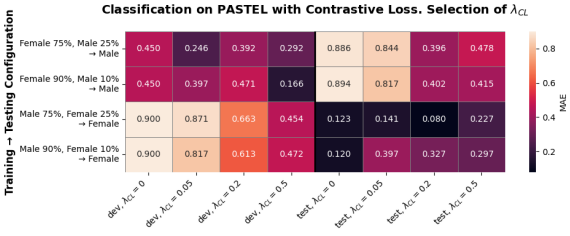


FIGURE 8. Text classification of the author's educational degree on PASTEL using contrastive loss. The plot shows the dependence of MAE on the training configuration and the parameter λ_{CL}. When $\lambda_{CL} = 0.0$, the model corresponds to the standard BERT.

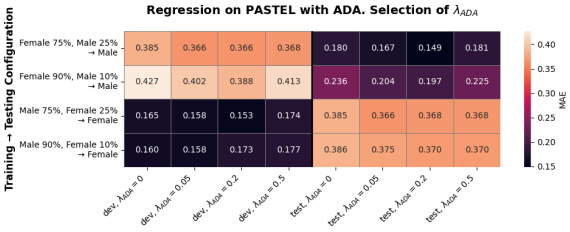


FIGURE 9. Text regression of the author's educational degree on PASTEL using Adversarial Domain Adaptation. The plot shows the dependence of MAE on the training configuration and the parameter λ_{ADA}. When $\lambda_{ADA} = 0.0$, the model corresponds to the standard BERT.

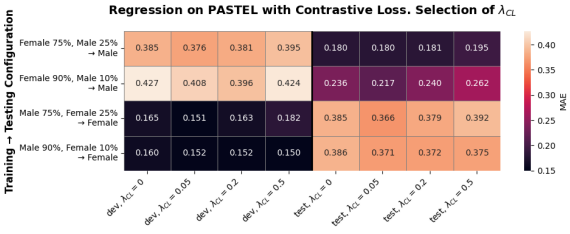


FIGURE 10. Text regression of the author's educational degree on PASTEL using contrastive loss. The plot shows the dependence of MAE on the training configuration and the parameter λ_{CL}. When $\lambda_{CL} = 0.0$, the model corresponds to the standard BERT.

one. There are relatively few cases in which texts with ground-truth labels 4 or 5 are misclassified as 1 or 2. According to the Amazon Reviews labeling scheme, ratings of 1 and 2 correspond to negative feedback, whereas ratings of 4 and 5 indicate positive feedback. Distinguishing between ratings 4 (rather positive) and 5 (strongly positive) can be challenging, as the language features used in the corresponding reviews are highly similar. Overall, the confusion matrices support the adequacy of the trained classifiers. In particular, ADA+CL improves accuracy for

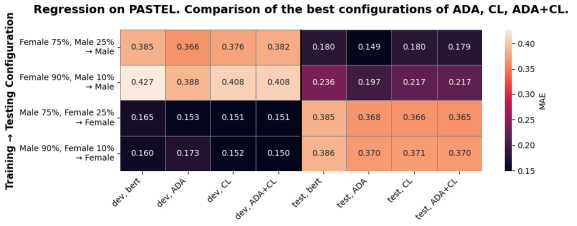


FIGURE 11. Text regression of the author's educational degree on PASTEL using Adversarial Domain Adaptation, contrastive loss, and joint loss. Comparison of the best configurations ($\lambda_{ADA} = 0.2$ for ADA, $\lambda_{CL} = 0.05$ for CL, $\lambda_{ADA} = 0.2$, $\lambda_{CL} = 0.1$ for ADA+CL).

TABLE 3. PASTEL. LLaMA 3.2 3B results. The highest F1 score and the lowest MAE are in bold.

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 4. Examples of texts from Amazon reviews that are correctly classified by the model trained with ADA and missclassified by the vanilla BERT. 75% of the training dataset are the texts about video games, 25% texts in train are about digital music.

Digital Music	Kindle Store
Just a good old boys having fun kinda song. I am not at all a Dierks Bentley fan. You will find not one of his songs in my vast music collection. So that should say something right there about the song! Had never heard of Cole Swindell. Saw it on CAM liked it and as they say the rest is history!	This was okay. Tegan was not my favorite. She told everyone everything. She was strong one minute and then (in my opinion) did ridiculous things in the next. There is pining for the boyfriend, Matthew. I was not able to connect with Tegan and Ethan, the hero. There were limited amount of scenes with them. I felt like this book did a lot of background and set-up for the rest of the books in the series. Not sure yet if I will read the other books in the series. I have loved each and every one of L.H. Cosway books and was excited to read this paranormal. I am not sure if the heroine improves in the following books.
Ground truth: 5 Predicted: 2	Ground truth: 3 Predicted: 2

texts with ground-truth labels 3 and 4, while slightly reducing performance along the main diagonal for texts with labels 1 and 5. This behavior is expected, as positive and negative reviews often contain clear lexical cues that facilitate their identification. The reduced confusion between neutral and positive reviews and between neutral and negative reviews observed for ADA+CL indicates that models trained with adversarial and contrastive objectives learn more nuanced features beyond simple keyword associations. As a result, models incorporating adversarial training exhibit increased robustness to topical shifts.

In addition, we analyze the mistakes of the classifiers. Table 4 shows examples of texts that were misclassified by the vanilla BERT, but were classified correctly by the model trained with ADA. The first text is from the Digital Music category and the second one is from the Kindle Store category. 25% of the training data for both models consists of Digital Music texts, and there are no texts in the training set from the Kindle Store category. The texts contain clear topical shifts and a reader can easily identify that the first text

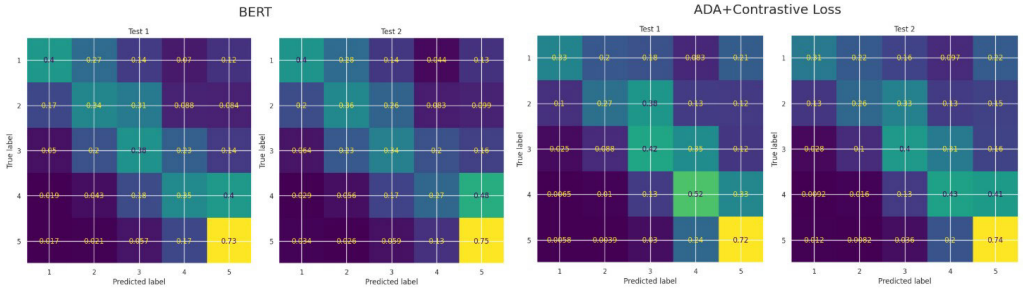


FIGURE 12. Confusion Matrices for vanilla BERT and ADA + Contrastive Loss for all the Amazon Reviews categories.

concerns music and the second one is about a book. The most challenging for classifiers is to distinguish the neutral reviews (marked as 3) for the positive and negative ones since the lexicon used to identify textual neutrality is relatively limited. It corresponds to the fact that the neutral reviews are often misclassified as negative or positive ones as shown in the confusion matrices Figure 12.

To better understand the structure of the feature spaces of our models, we visualize the embeddings Figure 13 of the classifiers on Amazon Reviews trained on the texts from the Movies category. We use the Scikit-learn [50] implementation of t-SNE [51]: $TSNE(random\_state = 42, n\_iter = 4000)$. The visualized embeddings of the textual reviews for all the models form two clusters: one consisting mostly of negative reviews (labels 1 and 2) and the one consisting mostly of positive ones (labels 4 and 5). It can be seen that the neutral reviews are not separated clearly from the positive and negative ones. This observation is consistent with the conclusions drawn from the confusion matrix analysis.

Since we introduce a change to the standard version of the contrastive loss by making it confounder-aware, it is important to understand how it performs compared to the standard loss. We perform an experiment for all the Amazon Reviews categories. We observe that our confounder-aware modification increases the QWK score by around 2.3% for classification and decreases the MAE by around 1.4% for regression. It shows that our modification of the contrastive loss is effective for combating the effect of distribution shifts. We also conduct an experiment using a positive-pair sampling strategy with similarity capping; however, it yields only a minor improvement of 0.2% while nearly doubling the training time.

For classification on the Amazon Reviews category named Books, we also provide experiments on different train/test split scenarios to investigate how the intensity of the distribution shifts affects robustness of our models. In addition to mixture ratio=75, we also evaluate mixture ratio=67, 70, 80, 90, 95 with 10 different random seeds. We compute the mean value and std for QWK and perform a t-test for dependent samples to understand whether the improvement of ADA and CL over the vanilla BERT is statistically significant. We find that for mixture share=67, 70, and 80 all the methods show an improvement over the vanilla BERT with p-values < 0.04. Moreover, the less bias in the dataset, the higher is the advantage of the adversarial methods over the vanilla BERT. However, for mixture ratio $\geq$ 90 there is no statistical improvement achieved by the adversarial and contrastive methods. We conduct the analogous experiment for regression and get the same conclusion with the only difference that the p-values for regression are slightly higher. The reason for lack of improvement when trained on the heavily biased datasets could be that the discriminator for the ADA and ADA+CL erodes to almost always predicting a constant value since $\geq$ 90% examples it sees are of the same class.

In addition, we analyze the performance of our classifiers and regressors trained on the PASTEL dataset. Figure 7 and Figure 9 show a significant asymmetry between classification and regression performance on texts written by male and female authors in predicting their education level. Figure 7 indicates that adversarial methods yield a larger increase in the F1 score for male-authored texts than for female-authored ones. We observe a significant correlation between text length and author gender in the PASTEL dataset. For example, the median text length is 38 words for the male authors, compared to 49 words for the female authors. Previous work [52] shows that the performance of BERT-based classifiers deteriorates when evaluated on texts longer than those seen during training.

Moreover, for male authors, we identify a strong correlation between text length and educational attainment. Specifically, the median number of words in texts written by male authors with a Master's degree is 23, whereas it is 43 for male authors without a degree. In contrast, no such correlation is observed for female authors: the median text length is 51 words for authors with a Master's degree and 46 words for those without a degree. This correlation for male authors induces a spurious dependency between text length and educational attainment and is highly likely to affect the predictions of BERT-based classifiers and regressors, which are known to be vulnerable to shortcut features [53]. This phenomenon partly explains why performance metrics are higher for male-authored texts.

IX. COMPARISON WITH LARGE LANGUAGE MODELS

LLMs are becoming more popular and are being used for a wider set of tasks. For example, the models like ChatGPT [54] and LLaMA [55] are used for zero-shot classification [56].

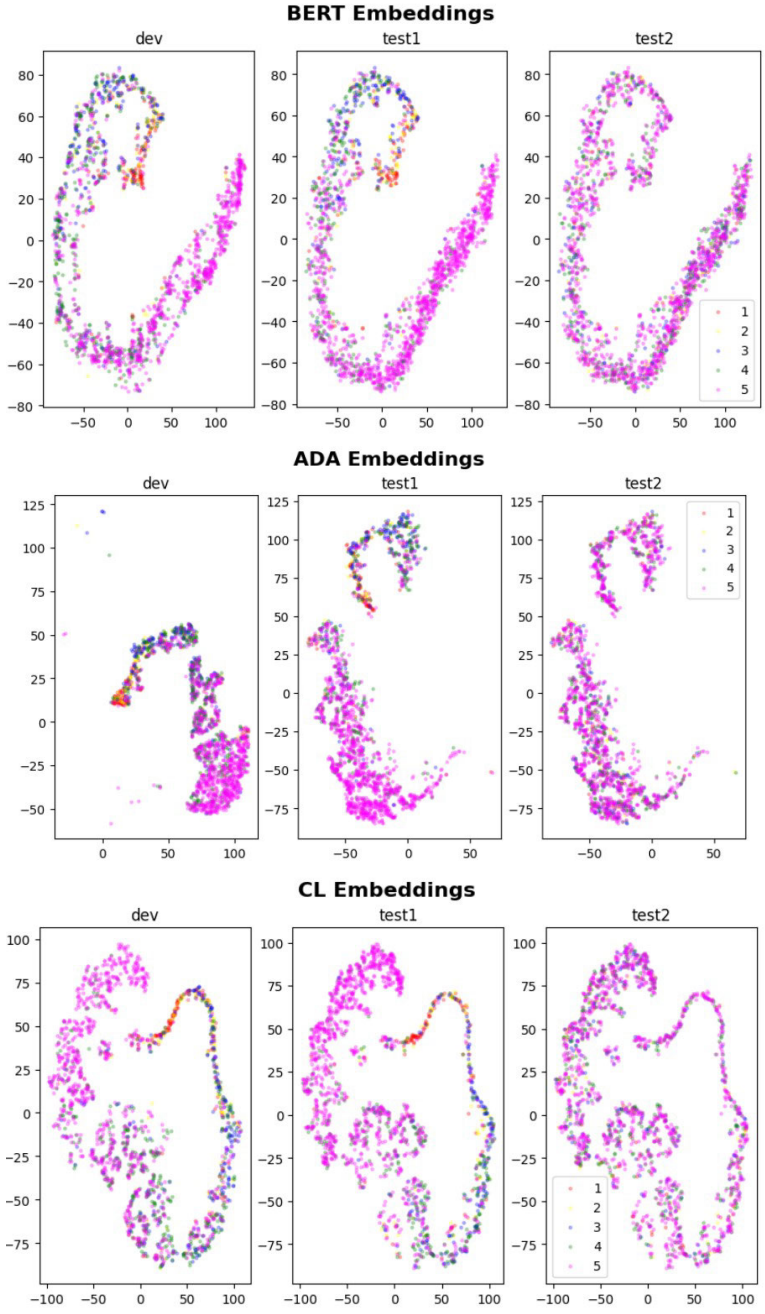


FIGURE 13. The embeddings for vanilla BERT, and ADA-only and CL-only models trained on 75% texts of Movies and 25% texts of Arts.

In order to estimate applicability of the LLMs to the tasks of text 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 apply 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 evaluate two variants of few-shot prompts: 9 male examples + 1 female example; 9 female examples + 1 male example.

For a more representative time evaluation, we run all the experiments for PASTEL on the same GPU provided by Google Colab. Inference for LLaMA requires more time and consumes more computational resources. The time needed to get 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 3 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 the box. It also confirms [57] which claims that for tasks involving natural language understanding, encoder-only models generally outperform decoder-only models, all while demanding a fraction of the computational resources.

X. CONCLUSION AND FUTURE WORK

In this work, we addressed the problem of distribution shifts—in particular, topical and gender-related shifts—in both text regression and non-topical text classification. We proposed an adversarial training framework based on Adversarial Domain Adaptation (ADA) and Contrastive Loss to reduce the influence of domain-specific features in pre-trained language models. Our experiments on the Amazon Reviews and PASTEL datasets demonstrate that adversarial approaches effectively mitigate topical biases and improve robustness to distribution shifts.

Our results show that ADA significantly enhances performance when the training data contains topical variation; however, when the topical signal is entirely absent from the training set, ADA performs similarly or slightly worse than vanilla BERT. We observe that the optimal ADA weighting parameter λ_{ADA} depends on the degree of dataset shift: more heavily shifted datasets require higher values of λ_{ADA}, yet values above 0.5 tend to degrade performance. For PASTEL, the optimal λ is approximately 0.2, while for Amazon Reviews it is around 0.05. As a practical guideline, values near 0.05 can be recommended when the degree of topical shift is unknown.

For Contrastive Loss, the optimal weighting lies between 0.05 and 0.2, and this method outperforms ADA when the training data is relatively clean. Moreover, for classification tasks, combining ADA with Contrastive Loss yields higher QWK than using either method individually.

Overall, our study shows that adversarial and contrastive techniques provide effective mechanisms for suppressing domain-related biases in both regression and classification settings.

Future work may investigate more computationally demanding causal-inference modeling techniques, adaptive and dynamically learned hyperparameter-scheduling strategies, and cross-domain generalization using larger and more diverse datasets. In particular, the choice of hyperparameter-scheduling mechanism is likely to exert a substantial influence on the effectiveness of adversarial training—especially when adversarial and contrastive objectives are jointly optimized. Although in this study we focus on using the adversarial and contrastive approaches for the English language, it is important to investigate how the methods described in this study perform when the language is treated as the confounder in the multilingual and cross-lingual setup. We hypothesize that their application can help improve the performance of the BERT-related regressors and classifiers given that Adversarial Domain Adaptation is known for its efficiency in cross-lingual settings [58], [59]. However, the exact values of the optimal hyperparameters are likely to be highly dependent on the language prevailing in train.

LIMITATIONS

Our experiments were conducted for the English language. However, in theory, the optimal hyperparameters may depend crucially on the language-based properties of the dataset. Moreover, in our study we use two datasets for regression and classification: Amazon Reviews and PASTEL. They are taken from different sources and represent different genres of texts, making our experiments more fundamental 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. Moreover, the PASTEL dataset exhibits a spurious correlation between text length and educational attainment for male authors. We hypothesize that this effect arises from the dataset's reliance on crowd-sourced annotation. To mitigate the risk of training biased models, it may therefore be beneficial to incorporate data from additional sources rather than relying exclusively on the PASTEL dataset.

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 addresses, phone numbers, or links to the social media. Besides, one frequent ethical problem in modern NLP applications is potential biases against specific groups. In our research, we aim 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.

APPENDIX

A. AMAZON CATEGORIES

Table 5 lists the original names of the categories from Amazon Reviews.

B. THE BASELINE BERT MODELS

Figure 14 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*.

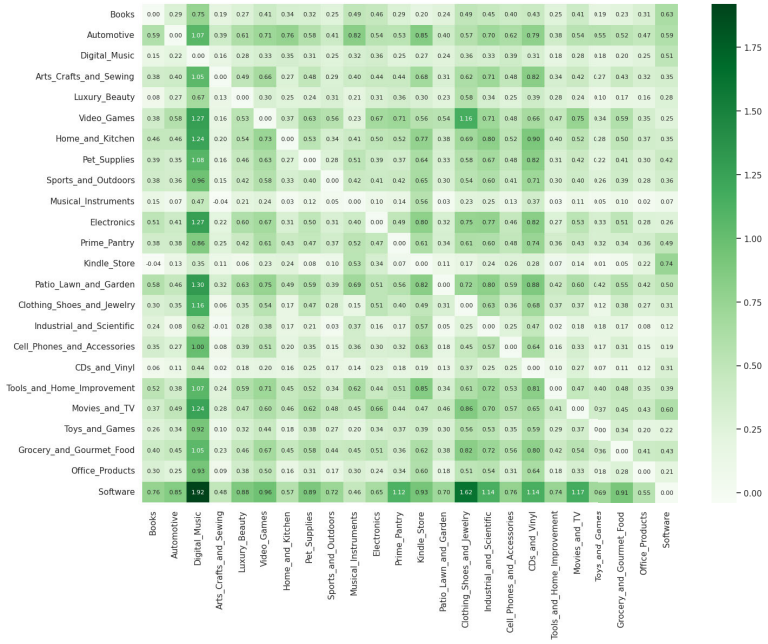


FIGURE 14. MAE delta on the Amazon Reviews test dataset. On the X-axis is the category on which the model was trained. On the Y-axis is the category where the model was tested.

TABLE 5. The full names of the categories for amazon reviews.

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	Kindle 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

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 in most cases the model mislaccifies 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.

C. DERIVATION OF THE NORMALIZATION FACTOR

Assume that all embeddings z_i are ℓ_2-normalized. Then, for any $x, y \in I$, we have

$$|z_x \cdot z_y| \leq \|z_x\| \, \|z_y\| = 1. \tag{14}$$

This implies

$$-1 \leq z_x \cdot z_y \leq 1. \tag{15}$$

Since the exponential function is monotonic, it follows that

$$\exp\left(-\frac{1}{\tau}\right) \leq \exp\left(\frac{z_x \cdot z_y}{\tau}\right) \leq \exp\left(\frac{1}{\tau}\right). \tag{16}$$

We define the constant

$$m := \frac{\exp(-1/\tau)}{(B-1)\exp(1/\tau)}. \tag{17}$$

Since $|A(i)| = B - 1$ by definition, we obtain

$$m \leq \frac{\exp(z_i \cdot z_p/\tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a/\tau)} \leq M, \tag{18}$$

where

$$M := \frac{\exp(1/\tau)}{(B-1)\exp(-1/\tau)}. \tag{19}$$

Because the logarithm is a monotonic function, we can bound the absolute value of the log-ratio as

$$\left|\log \frac{\exp(z_i \cdot z_p/\tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a/\tau)}\right| \leq \max\left(|\log m|, |\log M|\right). \tag{20}$$

Using this bound, we obtain the following inequality for the contrastive loss defined in Equation 3:

$$|\mathcal{L}_{CL}| = \left| \sum_{i \in I} \frac{-1}{|P(i)|} \log \frac{\exp(z_i \cdot z_p/\tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a/\tau)} \right| \tag{21}$$

$$\leq \sum_{i \in I} \frac{1}{|P(i)|} \max\big(|\log m|, |\log M|\big) \tag{22}$$

$$= \frac{B}{k} \max\big(|\log m|, |\log M|\big), \tag{23}$$

where $k = |P(i)|$ denotes the number of positive samples per anchor.

REFERENCES

[1] A. Luu and S. A. Malamud, "Non-topical coherence in social talk: A call for dialogue model enrichment," in *Proc. ACL*, 2020, pp. 118–133.

[2] S. Sharoff, Z. Wu, and K. Markert, "The web library of Babel: Evaluating genre collections," in *Proc. 7th Lang. Resour. Eval. Conf.*, 2010, pp. 3063–3070.

[3] P. Petrenz and B. Webber, "Stable classification of text genres," *Comput. Linguistics*, vol. 34, no. 4, pp. 285–293, 2010.

[4] D. Roussinov, S. Sharoff, and N. Puchnina, "Controlling out-of-domain gaps in LLMs for genre classification and generated text detection," in *Proc. 31st Int. Conf. Comput. Linguistics*, Jan. 2025, pp. 3329–3344.

[5] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman, "Measuring and mitigating unintended bias in text classification," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Dec. 2018, pp. 67–73.

[6] E. Dayanik, N. T. Vu, and S. Padó, "Bias identification and attribution in NLP models with regression and effect sizes," *Northern Eur. J. Lang. Technol.*, vol. 8, 2022, doi: 10.3384/nejlt.2000-1533.2022.3505.

[7] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2962–2971.

[8] X. Li and P. Groth, "How different is different? Systematically identifying distribution shifts and their impacts in NER datasets," *Lang. Resour. Eval.*, vol. 59, no. 2, pp. 1111–1150, Jun. 2025.

[9] A. Nayak and H. Prasad Timmapathini, "Wiki to automotive: Understanding the distribution shift and its impact on named entity recognition," 2021, *arXiv:2112.00283*.

[10] N. Calderon, N. Porat, E. Ben-David, A. Chapanin, Z. Gekhman, N. Oved, V. Shalumov, and R. Reichart, "Measuring the robustness of NLP models to domain shifts," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, Nov. 2024, pp. 126–154.

[11] H. Ye, Y. Ding, J. Li, and H. T. Ng, "Robust question answering against distribution shifts with test-time adaption: An empirical study," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, Dec. 2022, pp. 6179–6192.

[12] A. Pal, "CLIFT: Analysing natural distribution shift on question answering models in clinical domain," in *Proc. NeurIPS Workshop Robustness Sequence Model.*, 2022. [Online]. Available: https://openreview.net/pdf?id=9PQFROOfqm

[13] N. Shrestha and F. Nasoz, "Deep learning sentiment analysis of Amazon.Com reviews and ratings," *Int. J. Soft Comput., Artif. Intell. Appl.*, vol. 8, no. 1, pp. 01–15, Feb. 2019.

[14] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit," *Psychol. Bull.*, vol. 70, no. 4, pp. 213–220, 1968.

[15] A. Kwako and C. Ormerod, "Can language models guess your identity? Analyzing demographic biases in AI essay scoring," in *Proc. 19th Workshop Innov. Use NLP Building Educ. Appl. (BEA)*, Jun. 2024, pp. 78–86.

[16] D. Kang, V. Gangal, and E. Hovy, "(Male, bachelor) and (female, Ph.D) have different connotations: Parallelly annotated stylistic language dataset with multiple personas," 2019, *arXiv:1909.00098*.

[17] V. Basile, "Domain adaptation for text classification with weird embeddings," in *Proc. CEUR-WS*, 2020, pp. 18–24.

[18] R. Babbar and B. Schölkopf, "Data scarcity, robustness and extreme multi-label classification," *Mach. Learn.*, vol. 108, nos. 8–9, pp. 1329–1351, Sep. 2019.

[19] E. Tsymbalov, K. Fedyanin, and M. Panov, "Dropout strikes back: Improved uncertainty estimation via diversity sampling," 2020, *arXiv:2003.03274*.

[20] E. Schultheis, M. Wydmuch, R. Babbar, and K. Dembczynski, "On missing labels, long-tails and propensities in extreme multi-label classification," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2022, pp. 1547–1557.

[21] M. HassanPour Zonoozi and V. Seydi, "A survey on adversarial domain adaptation," *Neural Process. Lett.*, vol. 55, no. 3, pp. 2429–2469, Jun. 2023.

[22] H. Zou, J. Yang, and X. Wu, "Unsupervised energy-based adversarial domain adaptation for cross-domain text classification," in *Proc. Findings Assoc. Comput. Linguistics, ACL-IJCNLP*, 2021, pp. 1208–1218.

[23] C. Han, Z. Fan, D. Zhang, M. Qiu, M. Gao, and A. Zhou, "Meta-learning adversarial domain adaptation network for few-shot text classification," in *Proc. Findings Assoc. Comput. Linguistics, ACL-IJCNLP*, 2021, pp. 1664–1673.

[24] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le, "Unsupervised data augmentation for consistency training," in *Proc. NeurIPS*, 2020, pp. 6256–6268.

[25] M. A. Kausar, S. O. Fageeri, and A. Soosaimanickam, "Sentiment classification based on machine learning approaches in Amazon product reviews," *Eng., Technol. Appl. Sci. Res.*, vol. 13, no. 3, pp. 10849–10855, Jun. 2023.

[26] H. Zhang, "Model comparison in sentiment analysis: A case study of Amazon product reviews," *Highlights Sci., Eng. Technol.*, vol. 16, pp. 23–31, Nov. 2022.

[27] H. Ali, E. Hashmi, S. Y. Yildirim, and S. Shaikh, "Model comparison in sentiment analysis: A case study of Amazon product reviews," *Electronics*, vol. 16, pp. 23–31, Jul. 2024.

[28] G. Nkhata, U. Anjum, and J. Zhan, "Sentiment analysis of movie reviews using BERT," 2025, *arXiv:2502.18841*.

[29] M. S. Sayeed, V. Mohan, and K. S. Muthu, "BERT: A review of applications in sentiment analysis," *HighTech Innov. J.*, vol. 4, no. 2, pp. 453–462, Jun. 2023.

[30] E. Amigo, J. Gonzalo, S. Mizzaro, and J. Carrillo-De-Albornoz, "An effectiveness metric for ordinal classification: Formal properties and experimental results," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3938–3949.

[31] T. Sakai, "Evaluating evaluation measures for ordinal classification and ordinal quantification," in *Proc. ACL*, 2021, pp. 2759–2769.

[32] C. Zirn, G. Glavas, F. Nanni, J. Eichorst, and H. Stuckenschmidt, "Classifying topics and detecting topic shifts in political manifestos," in *Proc. Int. Conf. Adv. Comput. Anal. Political Text*, 2017, pp. 88–93.

[33] A. Feder, N. Oved, U. Shalit, and R. Reichart, "CausaLM: Causal model explanation through counterfactual language models," 2020, *arXiv:2005.13407*.

[34] A. S. Maiya, "CausalNLP: A practical toolkit for causal inference with text," 2021, *arXiv:2106.08043*.

[35] Y. Zhou and Y. He, "Causal inference from text: Unveiling interactions between variables," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, 2023, pp. 10559–10571.

[36] J. Ma, "Causal inference with large language model: A survey," in *Proc. Findings Assoc. Comput. Linguistics, NAACL*, 2025, pp. 5886–5898.

[37] S. Mahadevan, "Large causal models from large language models," 2025, *arXiv:2512.07796*.

[38] C. Azuma, T. Ito, and T. Shimobaba, "Adversarial domain adaptation using contrastive learning," *Eng. Appl. Artif. Intell.*, vol. 123, Aug. 2023, Art. no. 106394.

[39] N. Yadav, M. Alam, A. Farahat, D. Ghosh, C. Gupta, and A. R. Ganguly, "CDA: Contrastive-adversarial domain adaptation," 2023, *arXiv:2301.03826*.

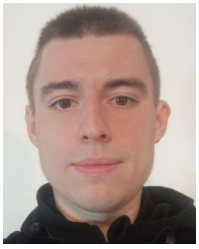
[40] J. Chen, Z. Zhang, L. Li, B. Shahrasbi, and A. Mishra, "Contrastive adversarial training for unsupervised domain adaptation," 2024, *arXiv:2407.12782*.

[41] A. Ramponi and B. Plank, "Neural unsupervised domain adaptation in NLP—A survey," in *Proc. Coling*, 2020, pp. 6838–6855.

[42] T. Jiang, "Learn from failure: Causality-guided contrastive learning for generalizable implicit hate speech detection," in *Proc. Coling*, 2025, pp. 8858–8867.

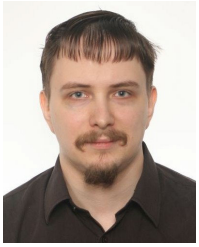
[43] M. Mosbach, M. Andriushchenko, and D. Klakow, "On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines," in *Proc. ICLR*, 2021, pp. 1–19.

[44] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?" 2019, *arXiv:1905.05583*.
[45] Y. Liu, L. Wang, and K. Gu, "A support vector regression (SVR)-based method for dynamic load identification using heterogeneous responses under interval uncertainties," *Appl. Soft Comput.*, vol. 110, Oct. 2021, Art. no. 107599.
[46] R. Yang, J. Cao, Z. Wen, Y. Wu, and X. He, "Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, 2020, pp. 1560–1569.
[47] V. Cherkassky and Y. Ma, "Practical selection of SVM parameters and noise estimation for SVM regression," *Neural Netw.*, vol. 17, no. 1, pp. 113–126, Jan. 2004.
[48] J. H. Park, P. Xu, and P. Fung, "PlusEmo2 Vec at SemEval-2018 task 1: Exploiting emotion knowledge from emoji and #hashtags," in *Proc. 12th Int. Workshop Semantic Eval.*, Jun. 2018, pp. 264–272.
[49] T. B. Trafalis and S. A. Alwazzi, "Support vector regression with noisy data: A second order cone programming approach," *Int. J. Gen. Syst.*, vol. 36, no. 2, pp. 237–250, Apr. 2007.
[50] 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, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.
[51] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008.
[52] Y. Zhou, S. Dai, Z. Cao, X. Zhang, and J. Xu, "Length-induced embedding collapse in PLM-based models," in *Proc. 63rd Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2025, pp. 28767–28791.
[53] Y. Zhou, R. Tang, Z. Yao, and Z. Zhu, "Navigating the shortcut maze: A comprehensive analysis of shortcut learning in text classification by language models," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, Nov. 2024, pp. 2586–2614.
[54] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Summary of chatgpt-related research and perspective towards the future of large language models," 2023, *arXiv:2304.01852*.
[55] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Roziére, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.
[56] Z. Wang, Y. Pang, and Y. Lin, "Large language models are zero-shot text classifiers," 2023, *arXiv:2312.01044*.
[57] A. Benayas, M. A. Sicilia, and M. Mora-Cantallops, "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," *Lang. Resour. Eval.*, vol. 59, no. 3, pp. 2007–2030, Sep. 2025.
[58] S. Latif, J. Qadir, and M. Bilal, "Unsupervised adversarial domain adaptation for cross-lingual speech emotion recognition," 2019, *arXiv:1907.06083*.
[59] A.-M. Avram, M. Lupaşcu, D.-C. Cercel, I. Mironică, and Ş. Trăuşan-Matu, "UniBERT: Adversarial training for language-universal representations," *Neural Comput. Appl.*, vol. 37, no. 26, pp. 21473–21492, Sep. 2025.



MIKHAIL LEPEKHIN received the B.S. and M.S. degrees in computer science from Moscow Institute of Physics and Technology, in 2020 and 2022, respectively. He is currently pursuing the Ph.D. degree with Moscow Independent Research Institute of Artificial Intelligence. Since 2024, he has been a Machine Learning Researcher and a Developer with Corsearch, working on counterfeit detection systems for online marketplaces. His publications from the last few years are mostly related to text genre classification and non-topical classification of texts. His research interests include large language models, machine translation, and text classification techniques.

SERGE SHAROFF received the Ph.D. degree from Lomonosov Moscow State University, in 1997. He was appointed as a Postdoctoral Researcher with Russian Research Institute for Artificial Intelligence, from 1997 to 2000, and the Humboldt Research Fellowship at the University of Bielefeld, from 2001 to 2002. He joined the University of Leeds, in 2003. His research interests are related to three domains: linguistics (primarily computational linguistics and corpus linguistics), cognitive science, and communication studies. His recent research is on digital curation of representative corpora automatically collected from the Web, i.e., their annotation in terms of genres, domains, or morphosyntactic categories. His current set of resources includes very large corpora for Arabic, Chinese, English, French, German, Italian, Polish, Portuguese, Russian, and Spanish.



ILYA MAKAROV received the M.S. degree in mathematics and the Ph.D. degree in mechanics and mathematics from Lomonosov Moscow State University, in 2011 and 2015, respectively, and the Ph.D. degree from the University of Ljubljana, in 2021. Since 2011, he has been a Researcher and a Senior Lecturer with the School of Data Analysis and Artificial Intelligence, HSE University. He has published more than 60 refereed articles in related areas. His awards and honors include the Best Teacher Award from HSE University, in 2022 and from 2017 to 2018. He is currently a Leading Research Fellow with AXXX and the Trusted AI Center, RAS. His research interests include a wide set of domains in artificial intelligence: network science, computer vision: image and video super-resolution (including multi-frame and HDR), semantic segmentation, pose estimation, action recognition, augmented reality, virtual reality, autonomous vehicles, and game artificial intelligence.