Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2025"

April 23–25, 2025

Causal Models and Adversarial Training: Selecting the right properties for robust non-topical text classification

Mikhail Lepekhin
Moscow Institute of Physics and Technology
Russia
lepehin.mn@phystech.edu

Serge Sharoff
University of Leeds
UK
s.sharoff@leeds.ac.uk

Abstract

The vast majority of datasets for non-topical classification of texts contain distribution shifts. In most cases, those are 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. It causes a dramatic decrease in the accuracy of the trained models when the test data are taken from a different data source. To address this problem, we experiment with two techniques: causal models and adversarial domain adaptation. In our work, we apply CausalLM, Adversarial Domain Adaptation (ADA), and Energy-based ADA (EADA) for gender classification and compare the results. The results are novel for the non-topical classification task. We show that both causal and adversarial methods manage to make the model more resilient to the distribution shifts although it causes a decrease of accuracy when tested on the domain prevailing in the training dataset. Moreover, we describe the first attempt to reduce the impact of topical shifts in the task of non-topical classification with usage of causal methods. Besides, we provide a link to the GitHub repository with the code of our experiments to ensure their reproducibility: https://github.com/MikeLepekhin/CausalAndAdversarialMethods.

Keywords: adversarial, causal models, gender classification, non-topical classification, bert
DOI: 10.28995/2075-7182-2025-23-224-233

Каузальные модели и состязательные методы: выбор правильных свойств для надежной нетематической классификации текстов

Аннотация

Подавляющее большинство наборов данных для нетематической классификации текстов содержат смещения. В большинстве случаев это тематические сдвиги. Их наличие в данных стимулирует классификаторы выучивать признаки, релевантные для предсказания тем, вместо фокусирования на признаках, относящихся к целевому классу. Это приводит к резкому снижению точности обученных моделей, когда тестовые данные берутся из другого источника данных. Для решения этой проблемы мы экспериментировали с двумя методами: каузальными (причинно-следственными моделями) и состязательными методами доменной адаптации для классификации гендера автора текста. Мы рассматриваем CausaLM, Adversarial Domain Adaptation (ADA) и Energy-Based ADA (EADA) и сравниваем результаты. Мы показываем, что как причинно-следственные, так и состязательные методы позволяют сделать модель более устойчивой к изменениям распределения, хотя это приводит к снижению точности при тестировании на текстах из источника, преобладающего в обучающем датасете. Кроме того, мы предоставляем ссылку на репозиторий GitHub с кодом наших экспериментов, чтобы обеспечить их воспроизводимость.

Ключевые слова: adversarial, каузальные модели, классификация гендера, нетематическая классификация, bert

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).

Unlike topical classification of texts, non-topical classification has a handful of additional difficulties. The most significant of them is that the target variable in these tasks is more complex than a topic. For example, every topic has a set of keywords, and, therefore, one can define whether a text belongs to a topic or not based on the occurrence of the keywords in the text. In contrast, the gender of the author cannot be defined just by a set of keywords. The classification of texts by genres and difficulty level has the same issue.

Complexity of the target variable in tasks of non-topical classification makes it vulnerable to distribution shifts. It implicitly pushes the classifiers to train on the irrelevant data features instead of the relevant ones. One of the most popular issues of non-topical classification of texts is the presence of topical shifts (biases) in the data (Sharoff et al., 2010).

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 more sensitive to the relevant features and to attend less to those that influence both the target variable and the text distribution, causing a spurious association. In causal inference, such features are called *confounders*. For example, (Feder et al., 2020) proposes a causal loss, which contains a negative summand corresponding to the head of the classifier related to the confounder. The efficiency of causal approaches for model de-biasing for the task of sentiment analysis is shown in (Feder et al., 2020). We assume that the approach can also be successful in application to more complex data distribution shifts in non-topical classification tasks.

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.

Energy-based ADA or EADA (Zou et al., 2021) is a modification of ADA. The authors show that that energy-based models exhibit free energy biases when training and test data come from different distributions and present a novel loss combined with an active learning strategy to assist knowledge transfer in the target domain, dubbed active domain adaptation. They show that domain gap can be implicitly diminished by aligning the free energy of target data compact around the source domain via a regularization term.

In our work, we compare causal models based on the CausaLM framework, adversarial models based on Adversarial Domain Adaptation (ADA) (Tzeng et al., 2017) and Energy-based ADA (EADA) (Zou et al., 2021) with the baseline of BERT-based models. We show that usage of both causal and adversarial methods helps to increase the accuracy on the dataset under-represented in train and thereby reduce model reliance on the distribution shifts.

In this study we:

1. Show that the BERT classifiers are sensitive to the source of the training data (subsection 6.1);
2. apply causal and adversarial mechanisms in order to decrease the deterioration of classification accuracy on the texts from the source under-represented in the train; (subsection 6.2);
3. make an ablation study (subsection 6.3)

2 Related Studies

The problem of domain adaptation has a long history of research.

Some approaches (Basile, 2020) propose direct manipulations on the textual embeddings. In contrast with out study, (Basile, 2020) does not apply adversarial methods in any form and instead modifies the embeddings of the *weird* words - the words specific to the target domain.

We use the architecture and mechanism from CausaLM (Feder et al., 2020). However, the authors do not have an objective to maximize the classification accuracy with the causal loss. They estimate

significance of multiple textual features instead. In contrast, we apply the causation mechanism in order to incent our classifiers to pay less attention to the confounders. We use the causal loss which consists of a positive summand for the cross-entropy of gender classification and a negative summand for cross-entropy of the confounder.

(Zhou and He, 2023) is a more advanced causal method which adjusts a model for latent covariates and takes into account the non-confounding covariates, which are relevant only to either the treatment or the outcome. Similarly to CausaLM (Feder et al., 2020), the authors only intend to estimate the causal effect of the covariate and do not increase accuracy of the model. However, similarly to CausaLM (Feder et al., 2020), this method could be potentially adjusted for raising the classification accuracy in tasks of non-topical classification.

Another widely used framework for causal models is CausalNLP (Maiya, 2021). Similarly to CausaLM (Feder et al., 2020), its general objective is estimation of causal dependences of the target variable on the confounders.

We use both Adversarial Domain Adaptation (ADA) and Energy-based Adversarial Domain Adaptation (EADA) from (Zou et al., 2021). However, the authors solve a different task in their paper. They 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.

(Han et al., 2021) propose a novel meta-learning framework integrated with an adversarial domain adaptation network, aiming to improve the adaptive ability of the text classifiers. The authors add a new component to the BERT-based model - meta-knowledge generator on the base of BiLSTM. The goal of this module is not only to make the final classification results better, but also to confuse the domain discriminator as much as possible. Unlike our study, (Han et al., 2021) mostly focused on the tasks of topical classification. The authors solve the task of sentiment analysis on the basis of Amazon Reviews, and classify the news from the datasets Reuters-21578 and 20 Newsgroups.

3 Causal and adversarial modification of the BERT-based architectures

3.1 CausaLM

Figure 1 shows the architecture implemented in (Feder et al., 2020). It is based on the BERT architecture (Devlin et al., 2018), but includes additional layers on top of it - for classification of the confounder and masked language model.

$$L(\theta_{bert}, \theta_{mlm}, \theta_{nsp}, \theta_{cc}, \theta_{tc}) =$$
$$= \frac{1}{n}\Bigg(\sum_{i=1}^{n} L^i_{mlm}(\theta_{bert}, \theta_{mlm}) + \sum_{i=1}^{n} L^i_{nsp}(\theta_{bert}, \theta_{nsp})$$
$$+ \sum_{i=1}^{n} L^i_{cc}(\theta_{bert}, \theta_{cc}) - \lambda \sum_{i=1}^{n} L^i_{tc}(\theta_{bert}, \theta_{tc}) \Bigg)$$

Where θ_{bert} denotes all the BERT parameters, except those devoted to θ_{mlm}, θ_{nsp}, θ_{tc} and θ_{cc}. L_{cc} - cross-entropy loss for classification of the target variable. In our case, it is the gender of the text author. L_{tc} is the loss of the confounder classification. λ is a hyper-parameter which controls the weight of the adversarial task.

The model in CausaLM is trained in 3 stages:

1. Training the model on MLM;
2. Training the confounder classifier;
3. Training to classify the target class (in our case, it is the gender of the author) with loss including Cross-Entropy components for MLM, confounder, and the target class label.

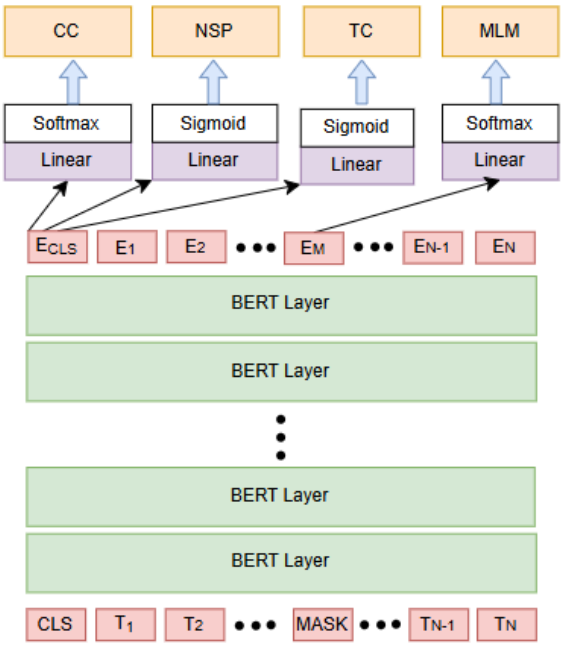

Figure 1: Architecture of a BERT-based causal model

3.2 Adversarial Domain Adaptation

ADA method 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 $d = G_d(x)$. In addition to the standard full supervision learning process in the source domain, a minimax game is designed between the feature extractor f and the domain discriminator d. 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 (in our study, it is the gender of the text author). L_f is the loss of the feature extractor. It denotes the cross-entropy of the classification of the text source. Both L_y and L_f are calculated and optimised with freezing of weights of the domain discriminator. L_d is similar to L_f. However, when it is calculated and optimised, the weights of the feature extractor and the label predictor are frozen.

3.3 Energy-based Adversarial Domain Adaptation

The adversarial training objective of three modules forms a minimax game, that is defined by:

$$\min_{G_f, G_y} L_{CE}(X_s, Y_s) + \gamma L_{AE}(X_t),$$

$$\min_{G_a} \big(L_{AE}(X_s) + max(0, m - L_{AE}(X_t))\big),$$

$$\text{where } L_{AE}(x_i) = ||G_a(G_f(x; \theta_f); \theta_a) x_i||$$

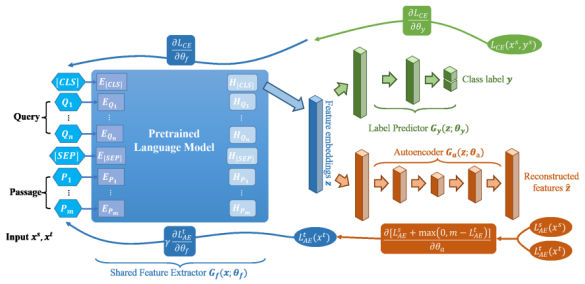

Figure 2: Architecture of a model for Energy-based Adversarial Domain Adaptation. The original image: (Zou et al., 2021)

Dataset	M freq	W freq	mean len	len perc 10	len perc 25	len median	len perc 75	len perc 90
MAIL	3236	6764	217	71	83	115	188	370
AWD	5984	4016	84	13	21	39	76	144

Table 1: Datasets for training and testing the classifiers

where γ is a hyperparameter to control the effectiveness of G_a. L_{CE} is cross-entropy loss for the target label classifier. L_{AE} is the loss of auto-encoder. m is the margin between the representations from the source domain and the target domain. The autoencoder G_a can be considered as an energy function that associates lower energies to the observed samples in a binary classification problem (Zou et al., 2021). The shared feature extractor G_f maps both labeled source data X_s and unlabeled target data X_t to a latent feature space. Both G_f and the label predictor G_y are trained with full supervision using the labeled data in the source domain.

The architecture is shown in Figure 2.

4 Experiments

The main metric we use to compare the models in the experiments is accuracy.

Our goal is to train a reliable classifier for which a domain shift deteriorates the model accuracy as little as possible. We compute the accuracy on the texts from , then compute the accuracy on the texts from the under-represented data source. We degine the difference of accuracies on the texts from the data source prevailing in the training dataset, and on the texts from the data source under-represented in train as *the difference of accuracies*. We denote it as δ (or delta).

We train a set of BERT-based models for CausaLM, ADA, and EADA. We use multilingual BERT with base configuration (12-layer, 768-hidden, 12-heads, 125M parameters, google-bert/bert-base-multilingual cased in HuggingFace) as a baseline for all the experiments. We compare it with the large version (24-layer, 1024-hidden, 16-heads, 355M parameters, google-bert/bert-large-multilingual-cased in HuggingFace). In all our experiments, learning rate=10^{-5} is used, since this value is proposed in (Sun et al., 2019) and (Zou et al., 2021).

5 Data

In our study, we work on the task of gender classification of the text author. It is a kind of non-topical classification, since the target classification variable is not a topic or a topical feature but a more complicated concept which cannot be described by certain keywords..

For the experiments, we use two datasets. The first one contains texts from Mail.Ru Blogs, the second one does from AWD. Each dataset contains nearly 10000 texts. In Table 1, the distribution of the genders and the text lengths is shown. In the Mail.Ru Blogs dataset, around 32% texts are written by men, and

mail share	acc
25	0.901
50	0.931
75	0.920
90	0.905

Table 2: BERT accuracies when trained to predict the source of the texts

the rest are written by women. In AWD, the gender distribution is different, since around 60% texts are written by men. Moreover, the texts from these datasets have different length distributions.

In all the experiments, we take 8000 examples to the train and 2000 texts to the test. Both Mail.Ru Blogs and AWD are Russian social media platforms. However, their content and target audience are different. Mail.Ru Blogs is a general-purpose platform which includes a wide variety of topics including sport, politics, technology, health, science, tourism and so on. At the same time, AWD is a platform about tourism. It causes presence of significant topical shifts in the AWD dataset.

These datasets have the etalon labels for the genders of the text authors, since the gender label is indicated by users of the platforms themselves.

6 Results and discussion

6.1 Prediction of the Text Domain

We use the number of dataset (mail or awd) as the confounder for both adversarial methods and CausaLM. It is a binary label. 0 means awd, 1 means mail. To make sure that the distribution of texts in these two data sources differs significantly, we conduct a range of experiments with classifiers of the source of the text. We indicate how the accuracy of the source classifiers is affected by shares of the data sources in the training dataset. We train multilingual bert classifiers on datasets containing 10%, 25%, 50%, and 75% of the Mail.Ru texts.

Table 2 shows that for each train/test split proportion the accuracy of the source classifier is higher than 90%. It means that the text distributions in Mail.Ru and AWD differ a lot and even BERT is able to notice the difference. We assume that it is caused by topical shifts in AWD given the specialization of this website. Moreover, the high accuracies shows vulnerability of the base BERT model to such a kind of topical shifts.

6.2 Causal and Adversarial Methods

The difference of the accuracy on the overrepresented data source in the train data and the accuracy on the underrepresented one, called *delta* is used as the second key metric to evaluate the vulnerability of the model to the topic shifts in the test data.

We show in Table 3 that Adversarial Domain Adaptation (ADA) helps to decrease the accuracy delta between the test on the text source prevailing on train and testing on the texts from the under-represented source. The increase in γ corresponds to the decrease in delta. At the same time, the accuracy on the test dataset, prevailing on train, decreases whilst the accuracy on the shifted dataset increases. It makes models more resilient to the domain shifts.

As we show in Table 3, dependence of the CausaLM performance on the λ is similar to that for the adversarial methods. The higher the λ, the lower the λ and the higher the accuracy on the texts from the source under-represented on the train.

In Table 3, it is clear that CausaLM achieves a lower delta than the adversarial methods. However, the decrease in accuracy on the texts from the over-represented source is more significant than that for the ADA method.

metod	m	mail share	λ	mail acc	awd_acc	delta
BERT	-	25	0	0.801	0.772	-0.029
ADA	-	25	0.05	*0.810*	0.761	-0.049
ADA	-	25	0.2	**0.818**	0.759	*-0.059*
EADA	4	25	0.05	0.806	0.753	-0.053
EADA	4	25	0.2	0.809	0.746	**-0.063**
BERT	-	75	0	0.838	0.716	0.122
CausaLM	-	75	0.05	0.785	0.725	*0.060*
CausaLM	-	75	0.2	0.781	*0.728*	**0.054**
ADA	-	75	0.05	0.830	0.725	0.105
ADA	-	75	0.2	0.825	**0.731**	0.094
ADA	-	75	0.5	0.815	0.719	0.096
EADA	2	75	0.05	0.819	0.688	0.131
EADA	2	75	0.2	0.819	0.673	0.146
EADA	4	75	0.05	0.819	0.692	0.127
EADA	4	75	0.2	0.815	0.685	0.130
EADA	8	75	0.05	0.832	0.694	0.138
EADA	8	75	0.2	0.823	0.685	0.138
BERT	-	90	0.0	0.840	0.702	0.138
ADA	-	90	0.05	0.836	*0.711*	0.125
ADA	-	90	0.2	0.825	**0.712**	*0.113*
ADA	-	90	0.5	0.694	0.600	**0.094**
EADA	4	90	0.05	0.832	0.699	0.133
EADA	4	90	0.2	0.806	0.600	0.206

Table 3: The results of the Adversarial Domain Adaptation. We compare the models on accuracy. The models with the highest accuracy on the source under-represented in train are highlighted in bold. The models with the second highest accuracy are highlighted in italic. For delta, we select the same way, but the smallest values.

We also find Table 3 that the accuracy on the texts from Mail.Ru is higher than that on the texts from AWD even if AWD prevails on the train. It can be explained by higher length of the texts from Mail.Ru. It is known (Baillargeon and Lamontagne, 2022) that the text length affects the accuracy of transformer-based models. It results in negative values of δ. In this case, the value of δ is less informative than the accuracy on Mail.Ru Blogs - the data source under-represented in the train dataset. The highest and the second highest accuracy on Mail.Ru Blogs is attained with ADA. EADA shows a result worse that ADA but better than the base BERT.

For the base BERT, ADA and EADA we also apply a one-sided t-test to evaluate the statistical significance of the improvements in the accuracy achieved by usage of the adversarial methods. For each setting (25% mail + 75% awd, 75% mail + 25% awd, and 90% mail + 10% awd), we make 5 random splits of mail and awd to train and test in the corresponding proportions and calculate the accuracy on each test sample. We calculate the p-values and find out that it is lower than 0.05 for ADA on the splits 75% mail + 25% awd, and 90% mail + 10% awd. There is not statistical significance for the results in the setting of 25% mail + 75% awd, since the p-value is 0.18. We suppose that it happens because of the length distribution of the AWD dataset prevailing in this setting.

In most cases, the increase of the accuracy on the test dataset is statistically significant, that reiterate applicability of the methods of adversarial domain adaptation for combatting the effects of the topical shifts.

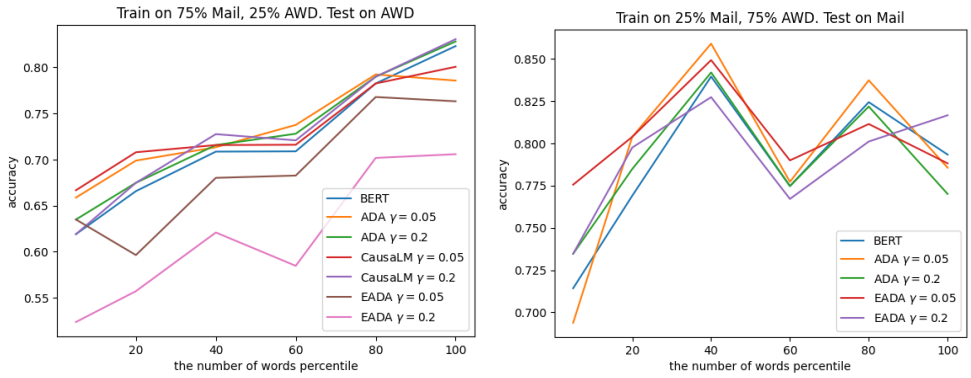

Figure 3: Effect of the text length on the performance of the causal and adversarial methods. Splits: 75% mail + 25% awd and 25% mail + 75% awd

6.3 Ablation Study

We show in Figure 3 and Figure 3 how the model performance is affected by the length of the texts. The texts are split to the groups of the same size by the number of words of them. It can be seen that for the splits with Mail.ru share equal to 75, the ADA model with $\gamma = 0.05$ beats the BERT-based baseline on every text length. However, when we take 25% of the Mail.ru texts and 75% of the AWD texts to the train, it gets less stable on the shortest and the longest texts.

In contrast to ADA, EADA has an additional hyperparameter m, the margin between the representations from the source domain and the target domain. The default value recommended in (Zou et al., 2021) is 4. We also try $m = 2$ and $m = 8$ for train 75% Mail.Ru + 25% AWD. Since, they all show a worse result than $m = 4$.

EADA attains higher accuracy than the baseline only when trained on with 25% mail and 75% awd.

The plot for training on 75% Mail.ru and 25% AWD shows a clear pattern that the model accuracy has a strongly positive correlation with the length of the text in the test. Moreover, it can be seen for all the models we trained on this mail/awd split. It confirms the conclusions from (Baillargeon and Lamontagne, 2022).

But this pattern is not that clear when the models are trained on 25% Mail.ru + 75% AWD and tested on Mail.ru. We assume that it can be caused by the difference in the length distributions of the Mail.Ru and AWD datasets, since the classifiers see too few long texts during the training.

7 Training Time

model	train data	epoch time
BERT	mail	**117**
ADA	mail	*128*
EADA	mail	132
CausaLM	mail	234

Table 4: Training time of all the models on the Mail.Ru Blogs dataset. The model that is trained the fastest is highlighted in bold. The second fastest model is shown in italics.

We calculate the time required for training model of each architecture we mention in our experiments. All the models were trained on single Nvidia TITAN RTX-based GPU. Available GPU capacity: 24 GB.

Table 4 shows that addition of the adversarial loss (both ADA and EADA) to the BERT models does not increase the training time significantly. However, the situation is completely different for CausaLM, where training is done in 3 stages.

The inference time of the classifiers depends solely on the model architecture and is not affected by adding of causal losses in any framework. Hence, there is no statistical significant in difference between evaluation times for the causal models and the non-causal ones.

EADA (Zou et al., 2021) has an additional hyperparameter, m - the margin between the representations from the source domain and the target domain. The default value proposed in (Zou et al., 2021) is 4. We carry out experiments with $m = 2$ and $m = 8$ for $75\%mail + 25\%awd$ to potentially improve the accuracy on the test. However, the default value $m = 4$ turns out to be the best, so we fix it for the splits $25\%mail + 75\%awd$ and $90\%mail + 10\%awd$.

We show Table 3 that the addition of causal loss improves the accuracy of the gender classification task in awd in case when the texts from mail.ru prevail in the training. It partially matches our expectations that causal models are capable of improving accuracy when tested on a dataset from a different domain.

8 Conclusion

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

1. The topical shift make a considerable impact on non-topical classification tasks.
2. The addition of either adversarial or causal loss slightly increases the model accuracy when tested on texts from a different domain.
3. Although the delta is decreased more by usage of CausaLM, the ADA method is able to do it without that significant decrease of accuracy on the source dataset.
4. The time spent on training causal models using ADA is significantly less than when using CausaLM. It makes usage of ADA more rational in terms of the required time and computational resources.
5. Although it is a more complex algorithm, EADA shows an improvement only on the split with 75% mail and 25% awd data. Hence, the method turns out to be less efficient than ADA.

Our results indicate that both causal and adversarial methods are useful for improving the quality of non-topical text classification in the presence of topical shifts and domain changes. Overall, the use of ADA can be recommended for cases when there is a significant domain shift on test data. This is an important practical result, given the prevalence and relevance of non-topical classification tasks in the modern world.

References

Jean-Thomas Baillargeon and Luc Lamontagne. 2022. Assessing the impact of sequence length learning on classification tasks for transformer encoder models. *The Florida AI Research Society*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

ChengCheng Han, Zeqiu Fan, Dongxiang Zhang, Minghui Qiu, Ming Gao, and Aoying Zhou. 2021. Meta-learning adversarial domain adaptation network for few-shot text classification. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

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*.

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

Serge Sharoff, Zhili Wu, and Katja Markert. 2010. The Web library of Babel: evaluating genre collections. *// 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*.

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

Yuxiang Zhou and Yulan He. 2023. Causal inference from text: Unveiling interactions between variables. *Findings of the Association for Computational Linguistics: EMNLP 2023*.

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