Counterfactual Augmentation for Robust Authorship Representation Learning

Hieu Man
University of Oregon
Computer Science
Eugene, OR, USA
hieum@uoregon.edu

Thien Huu Nguyen
University of Oregon
Computer Science
Eugene, OR, USA
thien@cs.uoregon.edu

ABSTRACT

Authorship attribution is a task that aims to identify the author of given pieces of writing. Authorship representation learning using neural networks has been shown to work in open-set environment settings with hundreds of thousands of authors. However, the performance of authorship attribution models often degrades significantly when texts are from different domains than the training data. In this work, we propose addressing this issue by adopting a novel causal framework for authorship representation learning. Our key insight is to use causal interventions during training to make models robust to differences in domains. Specifically, we introduce generating style-counterfactual examples by retrieving the most similar content texts by different authors on the same topics/domains. This exposes the model to challenging examples with similar content but distinct styles. Furthermore, we introduce causal masking of topic-indicative words to generate content-counterfactual examples. Content-counterfactuals hide topic content to encourage focusing on writing style. Experiments on three disparate domains - Amazon reviews, fanfiction stories, and Reddit comments - demonstrate that our approach significantly outperforms previous state-of-the-art methods for authorship attribution.

CCS CONCEPTS

• **Computing methodologies** → Natural language processing.

KEYWORDS

Authorship Attribution, Domain Generalization, Counterfactual Learning

ACM Reference Format:
Hieu Man and Thien Huu Nguyen. 2024. Counterfactual Augmentation for Robust Authorship Representation Learning. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3626772.3657956

1 INTRODUCTION

Untangling distinctive stylistic elements of an author's writing from their content has proven to be an exceptionally difficult task in computational linguistics over the years [20, 29]. While recent years have witnessed significant progress in authorship attribution through machine learning, robustly separating stylistic features from content remains an open research question. Authorship attribution has important real-world applications, e.g., plagiarism detection, moderating online content, analyzing historical documents [12, 27], and assisting forensic investigations [25].

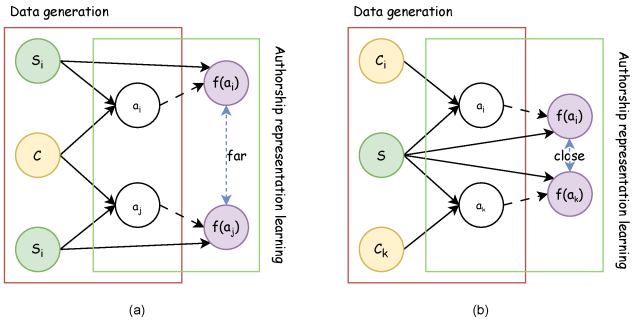

Figure 1: Causal graph models assumptions of relations about content, style, and authorship in: collections with similar content but different authors *(a)*, and collections by the same author with different content *(b)*.

Prior work has relied heavily on stylometric feature engineering to model linguistic properties like punctuation, character patterns, and function word usage [24–26]. While effective on small, homogeneous datasets, these feature-based methods often fail to generalize to large corpora covering diverse topics and domains. Recently, neural network approaches seek to learn stylistic representations directly from texts via contrastive learning [7, 9, 20, 23]. However, these methods might latch onto superficial content cues rather than capture robust markers of individual writing styles. Consequently, they display limited generalization across different topics/domains. More recently, [22] has proposed to reduce the topic bias in authorship attribution by distilling topic-regularized probabilities from a base model into a target model. However, it requires additional topic labels and training of multiple models, adding complexity for large-scale real-world applications.

To tackle these limitations, we introduce a novel causal framework for learning authorship representations that can be invariant to content. At its core, we analyze the problem of disentangling writing style from content through a causal lens [16, 17]. This allows us to mitigate the topic bias by exploiting inherent causal mechanisms. Specifically, our approach leverages two causal interventions during training: (i) retrieving style-counterfactual examples

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07.
https://doi.org/10.1145/3626772.3657956

that hold content constant but vary authorship, and (ii) generating content-counterfactual via causal masking of topic-indicative words, which conceal content cues and forces the models to rely on subtler style markers. By treating counterfactuals as hard examples in contrastive learning, we can promote the disentanglement of stylistic fingerprints from transient content signals. As such, superficial topic/domain correlations can be effectively eliminated, leading to more robust style representations that can better generalize across topics and domains. We validate our approach through experiments on Amazon reviews, fanfiction, and Reddit comments. Our results show significant improvements in authorship attribution performance compared to previous methods. Using R@8 and MRR metrics, our approach achieves average gains of 10.9% and 15.7% for in-domain testing, and 15.3% and 18.2% for cross-domain testing. This highlights the robustness of our causal learning framework for authorship attribution.

2 METHODOLOGY

Problem Formulation. We consider a collection of set of documents $A = \{a_1, a_2, ..., a_n\}$, where each $a_i = \{t_1, t_2, ..., t_{a_i}\}$ represents the set of documents written by the same author. For the convenience of notation, we still denote a_i for the author. Following prior work [20, 22], our goal is to learn a function f that maps a collection a_i to an authorship representation $r_{a_i} \in \mathbb{R}^h$ such that the representations of document collections by the same author have higher cosine similarity compared to representations of collections by different authors.

2.1 Causal Interpretation

To leverage intuitive assumptions about how text data is generated, we propose formalizing the problem of authorship representation learning by using a causal graph. We start from three key assumptions: (i) Texts are generated from latent content and style variables. (ii) Only style is relevant for authorship; content can vary freely across authors. (iii) Style and content are independent causal factors. Concretely, different collections by the same author should reflect a consistent underlying style, despite having different content. Meanwhile, collections with similar content or topic but written by different authors should exhibit distinct styles. Next, we concisely represent these assumptions using a causal graph [16, 18].

We model content C and style S as separate causes of the observed text collection a. Importantly, only style directly affects the authorship representation $f(a)$, while content can vary freely across authors. Specifically, for collections a_i and a_j with similar content C but different author styles S_i and S_j, the representations $f(a_i)$ and $f(a_j)$ should be distinct due to the differing styles and independent of the content C (Figure 1a). Conversely, for collections a_i and a_k by the same author but different content C_i and C_k, their corresponding representations $f(a_i)$ and $f(a_k)$ should be similar due to the same style S (Figure 1b). The graph provides a principled way to guide the learning process toward style representations by discounting misleading content correlations. Specifically, this causal structure enables us to inject knowledge through two key interventions: retrieving style-counterfactual examples that control content while varying style, and masking words highly indicative of topic to generate content-counterfactual examples that vary the content while having same style.

2.2 Causal Authorship Representation Learning

Causal Interventions. The high-level idea behind our causal interventions is to create counterfactual examples by varying content or style while keeping the other fixed. We obtain such counterfactual data by augmenting the original data.

To produce style-counterfactual examples that control content while intervening on style, we propose retrieving texts from the same domain/topic that have similar content but different authors. Specifically, for each collection of training, $a_i \in A$, we use an external retrieval system to find the top-k most semantically similar collections from other authors a_j i.e., $a_j \neq a_i$. This way, we obtain texts that have high content overlap with the original collection, but different writing styles. Formally, we define the style-counterfactual examples set A_i^S of a collection a_i as $A_i^S = Top_k(sim(a_i, A))$. Where $sim(.)$ is the similarity score computed by the external retrieval system and $Top_k(.)$ selects the top-k most similar collections.

To generate content-counterfactual examples that preserve style while modifying content, we propose causal masking of topic-related words. In particular, we first train an unsupervised topic model such as LDA [5] on the dataset. Significant topical terms, W^T, are then extracted to serve as the topic word list for the dataset. Next, for each collection a_i, we generate the content-counterfactual examples by randomly masking the topic words of the original collection. This results in the content-counterfactual examples A_i^C that change content while retaining the author's style, $A_i^C = Mask(a_i, W^T)$. Where $Mask(.)$ is randomly masking.

Contrastive Learning. We learn authorship style representations using a contrastive learning approach [8, 11]. Given a collection of texts $a_i \in A$, we first generate two types of counterfactual examples: style-counterfactual examples A_i^S, and content-counterfactual examples A_i^C. These act as negative and positive examples, respectively. Let z_i be the L2 normalization of the representation $f(a_i)$. We minimize the following contrastive loss over the full set of data $a_i \in A$:

$$L_{contrastive} = \sum_{i=1}^{n} \sum_{a_j \in A_i^C} \log \frac{\exp\left(z_i \cdot z_j / \tau\right)}{\sum\limits_{a_k \in A_i^S \cup A_i^C} \exp\left(z_i \cdot z_k / \tau\right)}$$

where τ is a temperature parameter set to 0.01. This encourages the model to learn to extract authorship style features that are invariant to changes in content.

Causal Invraiant Regularization. To further enforce invariant representations, we utilize causal invariant learning [1] to learn authorship representations such that the distribution over the hidden features is invariant under content interventions. Following [14], for a collection a_i with authorship representation $r_{a_i} = f(a_i)$, the invariant criteria is: $p(r_{a_j} \mid a_j, f) = p(r_{a_k} \mid a_k, f) \quad \forall a_i, a_k \in A_i^C$. To achieve this, we use invariance regularizer:

$$\sum_{a_j, a_k} KL\left(p(r_{a_j} \mid a_j, f), p(r_{a_k} \mid a_k, f)\right)$$

Our overall training objective combines contrastive learning and invariant regularization:

$$L = L_{contrastive} + \lambda \sum_{a_j, a_k} KL\left(p(r_{a_j} \mid a_j, f), p(r_{a_k} \mid a_k, f)\right)$$

where λ weights the invariance penalty. This objective will encourage authorship invariant representations under content changes, as enforced by causal invariant regularization.

3 EXPERIMENTS

Data. Following [20, 22], we conduct our experiments on three domains: Amazon reviews [15], fanfiction stories [3, 4], and Reddit comments [2]. We use the same data splits as [20] for Amazon reviews and Reddit comments. For fanfiction, we use the PAN-20-small dataset [4] with about 52K authors for training and the PAN-21 test set with 27K authors [3] for testing. We use the PAN-20 test data [4] as the validation data for fanfiction.

Hyperparameters. Our model is based on the architecture proposed by [20], which employs a pre-trained sBERT [19] as an encoder. We use the following hyperparameters for training our model: a minibatch size of 128, a learning rate of 2*e*-5 with the AdamW [13] optimizer. For retrieval, we use BM25 [21] and for unsupervised topic modeling, we use LDA [6].

Baselines. We evaluate our approach against four baselines that use transformer-based models: Multiclass log loss (MLL) [9] and Contrastive loss (CL) [11, 20]. These methods learn document representations by optimizing different objectives. We also compare with ARR [22], a technique that reduces topic bias by distillation. Furthermore, we include the TF-IDF vector representation of the concatenated text content of a document collection as a simple baseline. We use single words as tokens for this model.

3.1 Overall Perfromance

Train dataset	Models	Test dataset: Amazon R@8	Amazon MRR	PAN21 R@8	PAN21 MRR	Reddit MUD R@8	Reddit MUD MRR
Amazon	TF-IDF	31.6	24.8	27.8	20.4	7.7	5
	CL	82.5	69	41.7	30.9	23.7	15.4
	CL + ARR	84.2	70.8	40.1	29.7	26.5	17.9
	Our	**96.8**	**93**	**54.9**	**47.3**	**40.2**	**29.5**
PAN	TF-IDF	30.7	21.6	18.9	10.1	7.1	4.3
	CL	55.7	40.2	30.4	20.5	10.8	5.6
	CL + ARR	54.9	40.1	28.9	19.2	9.6	5
	Our	**84.2**	**76.5**	**42.6**	**35.4**	**16.1**	**10.7**
Reddit	TF-IDF	7.7	5.1	6.5	4.1	10.3	6.8
	CL	68.9	55.6	47.5	39.7	65.6	50.4
	CL + ARR	70.1	57.3	50.3	40.9	63.2	49.9
	Our	9 **3.6**	**89.9**	**54**	**46.3**	**71.9**	**58.6**

Table 1: Recall at 8 (R@8) and Mean Reciprocal Rank (MRR) results for zero-shot transfer experiments

Our framework demonstrates strong performance on both within-domain and cross-domain evaluations. We trained separate models for each domain and then evaluated on the test sets for all the domains. As shown in Table 1, our approach outperforms baselines on all datasets across both R@8 and MRR metrics. For in-domain testing, our model achieves average gains of 10.9% on R@8 and 15.7% on MRR compared to state-of-the-art, i.e., CL+ARR. More impressively, we observe even greater improvements under cross-domain conditions, with average gains of 15.3% on R@8 and 18.2% on MRR over CL+ARR. These consistent and sizable improvements on both in-domain and cross-domain benchmarks highlight the effectiveness of our framework for learning across diverse domains.

3.2 Ablation Study

	Amazon R@8	Amazon MRR	PAN21 R@8	PAN21 MRR	Reddit MUD R@8	Reddit MUD MRR
Our	**96.8**	**93**	**54.9**	**47.3**	**40.2**	**29.5**
Content-counterfactual	91.7	90.6	50.9	41.1	29.3	20.1
Random Masking	84.5	70.1	42.8	28.2	24.5	17.3
Style-counterfactual	93.1	91.3	51.9	44.8	36.3	25.4
w/o Invariant Regularization	95.1	92.2	53.3	46.1	38.2	28.4
Our$_{Contriever}$	95.3	91.7	52.1	45.9	39.8	29

Table 2: Recall at 8 (R@8) and Mean Reciprocal Rank (MRR) results for ablation study training on Amazon reviews

We perform an extensive ablation study to analyze the contribution of each component of our proposed model, which is trained on Amazon reviews. The results in Table 2 show the performance of our full model compared to several ablated variants.

Specifically, we consider the following ablated models: (1) **Content-counterfactual**: using only content-counterfactual examples as positive examples and random sampling negative examples; (2) **Random Masking**: using only examples with randomly masked words as positive examples random sampling negative examples; (3) **Style-counterfactual**: using only style-counterfactual examples as negative examples and random sampling positive examples; (4) **w/o Causal Invariant Regularization**: using the objective function without Causal Invariant Regularization; (5) **Our**$_{Contriever}$: using Contriver [10] as the retrieval model.

Ablating each component causes noticeable drops in performance, demonstrating their contribution. With the in-domain case, removing style-counterfactual examples results in the largest decrease, R@8 and MRR reduced by 5.1% and 2.4% respectively. The differences are even more pronounced on the challenging out-of-domain datasets. On Reddit, our full model substantially outperforms all ablated versions, with 10.9% higher R@8 and 9.4% higher MRR compared to the baseline without Style-counterfactual. Similar trends are observed on PAN21, with our full model achieving 4% and 6.2% better in R@8 and MRR versus ablated models. The substantial gaps in both in- and cross-domain performance verify that all our proposed techniques are crucial for generalization.

Comparing the "Random Masking" and "Content-counterfactual" baselines reveals some insightful differences. While both aim to make the representation more robust to content variations, the content-counterfactual examples provide more meaningful and challenging augmentations. As a result, the content-counterfactual baseline substantially outperforms random masking across all datasets, with 7.2% higher R@8 and over 20.5% higher MRR on Amazon. The gap is even larger for cross-domain generalization, with content-counterfactual achieving 8-12% better R@8 and 13-21% better MRR than random masking. This highlights the importance of generating

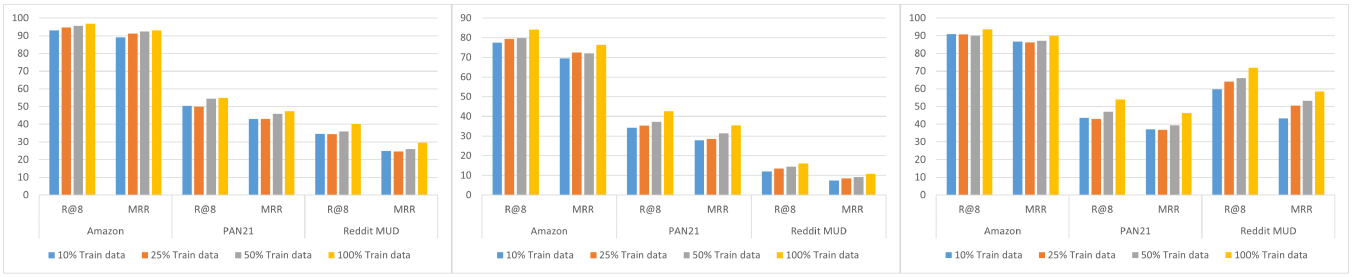

Figure 2: Performance with different training data ratios on Amazon (left), Fanfiction (middle), and Reddit comments (right).

high-quality content variations, rather than just random noise, to improve the author's representation. The controlled counterfactual transformations used in our approach are more effective at retaining author style while modifying content. Furthermore, the causal invariant regularization provides noticeable gains in domain generalization capability. Removing this regularization term hurts performance across datasets, confirming its importance for learning representations. Finally, using Contriever as the retriever instead of BM25 decreases performance on both in-domain and cross-domain datasets. This indicates the benefits of a sparse retriever like BM25 for authorship representation learning, as compared to dense retrieval methods like Contriever.

3.3 Partial Training Study and Author-representation Visualizations

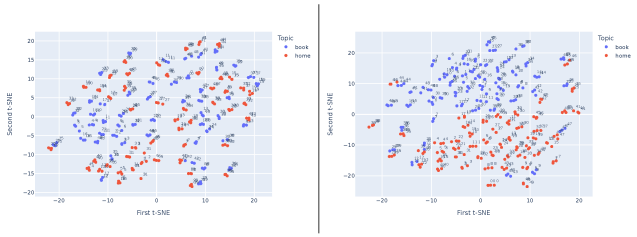

Figure 3: Two-dimensional projections of the author representations of Our (left) and CL model (right).

To evaluate the data efficiency of our proposed model, we train variants with 10%, 25%, 50% of full training sets. Models are tested on full test sets. We then assess the performance of these models on the full test sets across all domains. Figures 2 present the results using truncated training sets. Remarkably, our model achieves comparable performance to the full data setting even when trained on just 10% of data. This demonstrates the exceptional data efficiency and generalization capacity of our approach. As more training data is added, performance steadily improves across metrics and domains. However, the margins are noticeably small between 50% training data and the full dataset. The robustness of our model under low resource conditions can be attributed to the proposed training techniques. By extensively augmenting the training data and regularizing for representation invariance, our model learns effectively from limited examples. Exposing the model to diverse stylistic and content variations improves generalization.

To better demonstrate the quality of the author representations, we visualized the embeddings using t-SNE [28]. Figures 3 show two-dimensional projections of embeddings for 50 authors with comments across the book and home categories from the Amazon reviews dataset. Author IDs are displayed adjacent to the corresponding data points. As can be seen, collections written by the same author cluster more tightly with our model compared to the contrastive loss baseline. Importantly, data points of authors with writings across categories appear in closer proximity in our model but not the CL. This suggests our model learns more generalized author representations that capture stylistic similarities across contexts rather than simply relying on the category.

4 CONCLUSION

In this work, we proposed a novel causal framework to learn authorship representations invariant to content. The key insight is using causal interventions during training to disentangle style from content. We introduced two techniques: retrieving style-counterfactuals to control content while varying style, and generating content-counterfactuals via causal masking to hide content cues. Treating counterfactuals as hard examples for contrastive learning induces robust author embeddings capturing subtler style markers. Through extensive experiments on three distinct domains - Amazon reviews, fanfiction stories, and Reddit comments - we showed effectiveness of our approach. Our model significantly outperformed previous state-of-the-art, with average gains of 10.9% and 15.7% on in-domain authorship attribution. More impressively, even greater improvements of 15.3% and 18.2% under challenging cross-domain conditions. Consistent sizable gains highlight versatility for few-shot transfer learning across diverse topics and domains. An exciting avenue for future work is exploring the transferability of our author embeddings to downstream tasks through fine-tuning. We hope our insights will pave the way for authorship attribution models that continue to work robustly even as the world constantly evolves.

ACKNOWLEDGEMENTS

This research has been supported by the Army Research Office (ARO) grant W911NF-21-1-0112, the NSF grant CNS-1747798 to the IUCRC Center for Big Learning, and the NSF grant # 2239570. This research is also supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract 2022-22072200003.

REFERENCES

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2020. Invariant Risk Minimization. arXiv:1907.02893 [stat.ML]

[2] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit Dataset. arXiv:2001.08435 [cs.SI]

[3] Janek Bevendorff, Berta Chulvi, Gretel Liz De La Peña Sarracén, Mike Kestemont, Enrique Manjavacas, Ilia Markov, Maximilian Mayerl, Martin Potthast, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, Benno Stein, Matti Wiegmann, Magdalena Wolska, and Eva Zangerle. 2021. Overview of PAN 2021: Authorship Verification, Profiling Hate Speech Spreaders on Twitter, and Style Change Detection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings*. Springer-Verlag, Berlin, Heidelberg, 419–431. https://doi.org/10.1007/978-3-030-85251-1_26

[4] Janek Bevendorff, Bilal Ghanem, Anastasia Giachanou, Mike Kestemont, Enrique Manjavacas, Ilia Markov, Maximilian Mayerl, Martin Potthast, Francisco Rangel, Paolo Rosso, Günther Specht, Efstathios Stamatatos, Benno Stein, Matti Wiegmann, and Eva Zangerle. 2020. Overview of PAN 2020: Authorship Verification, Celebrity Profiling, Profiling Fake News Spreaders on Twitter, and Style Change Detection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings* (Thessaloniki, Greece). Springer-Verlag, Berlin, Heidelberg, 372–383. https://doi.org/10.1007/978-3-030-58219-7_25

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3, null (mar 2003), 993–1022.

[7] Benedikt Boenninghoff, Steffen Hessler, Dorothea Kolossa, and Robert M Nickel. 2019. Explainable authorship verification in social media via attention-based similarity learning. In *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 36–45.

[8] Jacob Goldberger, Geoffrey E Hinton, Sam Roweis, and Russ R Salakhutdinov. 2004. Neighbourhood Components Analysis. In *Advances in Neural Information Processing Systems*, L. Saul, Y. Weiss, and L. Bottou (Eds.), Vol. 17. MIT Press.

[9] Julien Hay, Bich-Lien Doan, Fabrice Popineau, and Ouassim Ait Elhara. 2020. Representation learning of writing style. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*. Association for Computational Linguistics, Online, 232–243. https://doi.org/10.18653/v1/2020.wnut-1.30

[10] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised Dense Information Retrieval with Contrastive Learning. https://doi.org/10.48550/ARXIV.2112.09118

[11] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 18661–18673.

[12] Moshe Koppel and Shachar Seidman. 2013. Automatically Identifying Pseudepigraphic Texts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, 1449–1454. https://aclanthology.org/D13-1151

[13] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101 [cs.LG]

[14] Jovana Mitrovic, Brian McWilliams, Jacob Walker, Lars Buesing, and Charles Blundell. 2020. Representation Learning via Invariant Causal Mechanisms. arXiv:2010.07922 [cs.LG]

[15] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 188–197. https://doi.org/10.18653/v1/D19-1018

[16] Judea Pearl et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress* 19, 2 (2000), 3.

[17] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

[18] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

[19] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[20] Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. Learning Universal Authorship Representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 913–919. https://doi.org/10.18653/v1/2021.emnlp-main.70

[21] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (apr 2009), 333–389. https://doi.org/10.1561/1500000019

[22] Jitkapat Sawatphol, Nonthakit Chaiwong, Can Udomcharoenchaikit, and Sarana Nutanong. 2022. Topic-Regularized Authorship Representation Learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 1076–1082. https://doi.org/10.18653/v1/2022.emnlp-main.70

[23] Prasha Shrestha, Sebastian Sierra, Fabio González, Manuel Montes, Paolo Rosso, and Thamar Solorio. 2017. Convolutional Neural Networks for Authorship Attribution of Short Texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, 669–674. https://aclanthology.org/E17-2106

[24] Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *J. Am. Soc. Inf. Sci. Technol.* 60, 3 (mar 2009), 538–556.

[25] Efstathios Stamatatos. 2017. Authorship Attribution Using Text Distortion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, 1138–1149. https://aclanthology.org/E17-1107

[26] Ariel Stolerman, Rebekah Overdorf, Sadia Afroz, and Rachel Greenstadt. 2014. Breaking the Closed-World Assumption in Stylometric Authorship Attribution. In *Advances in Digital Forensics X*, Gilbert Peterson and Sujeet Shenoi (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 185–205.

[27] Justin Anthony Stover, Yaron Winter, Moshe Koppel, and Mike Kestemont. 2016. Computational authorship verification method attributes a new work to a major 2nd century African author. *Journal of the Association for Information Science and Technology* 67, 1 (2016), 239–242. https://doi.org/10.1002/asi.23460 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23460

[28] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html

[29] Andrew Wang, Cristina Aggazzotti, Rebecca Kotula, Rafael Rivera Soto, Marcus Bishop, and Nicholas Andrews. 2023. Can Authorship Representation Learning Capture Stylistic Features? arXiv:2308.11490 [cs.CL]