Contrastive Disentanglement for Authorship Attribution

Anonymous ACL submission

Abstract

Authorship Attribution (AA) aims to identify the authorship of texts by analyzing distinctive writing styles. While current AA methods have yielded promising performance, these approaches commonly exhibit suboptimal performance in contexts where the subject matter varies significantly (i.e., topic-shift scenarios). This limitation stems from their inadequacy in differentiating between the topical content and the author's stylistic signature. Additionally, existing studies predominantly focus on AA at an individual level, thereby neglecting the exploration of regional-level AA, which could reveal common linguistic patterns influenced by cultural and geographical factors. Addressing these gaps, this paper introduces ContratDistAA, a novel framework employing contrastive learning coupled with mutual information maximization to segregate content from stylistic features in latent representations for AA tasks. Our comprehensive experimental evaluations reveal that ContratDistAA outperforms existing state-of-the-art models in both individual and regional-level AA scenarios. This advancement not only enhances the accuracy of authorship attribution but also expands its applicability to encompass regional linguistic analysis, thus contributing significantly to the broader field of computational linguistics.

1 Introduction

Motivation. Authorship Attribution (AA) is an extensively researched area (Zheng and Jin, 2023). The goal of AA is to identify the author of a piece of text based on distinctive linguistic characteristics inherent in their writing style. Applications of AA span a broad range of domains, including digital forensics (Iqbal et al., 2008) and plagiarism detection (Stamatatos and Koppel, 2011).

Existing methods in AA can be broadly categorized into two groups: traditional stylometric approaches (Seroussi et al., 2011; Bevendorff et al., 2019) and machine learning-based techniques (Zhang et al., 2018; Saedi and Dras, 2021). Traditional stylometric methods exploit features such as word lengths, sentence lengths, and function words to attribute authorship. Machine learning-based methods, particularly deep learning techniques, were leveraged to capture intricate patterns in writing styles, often surpassing the performance of stylometric methods (Rivera-Soto et al., 2021; Wang et al., 2023).

Despite these advancements, a significant challenge persists in scenarios involving a shift in topics, particularly when the testing phase encompasses topics not present in the training dataset (Sapkota et al., 2014; Hu et al., 2023). TThis issue primarily arises from the conflation of topic-related content and the author's unique writing style. Consequently, standard stylistic features employed in AA may inadvertently reflect topical variations rather than the author's stylistic nuances, leading to inaccuracies in authorship determination based solely on writing style.

Moreover, the majority of existing research in AA predominantly concentrates on the individual author level, thereby overlooking the potential of regional-level AA. Exploring AA at the regional level could reveal distinct linguistic styles shared by authors within the same geographical region, influenced by cultural nuances. For example, in Singapore, the widespread use of English is distinctively marked by local cultural influences and slang, offering a unique dimension essential for effective AA at a regional scale. This warrants further investigation to fully understand and utilize the nuances of regional linguistic variations for the AA task.

Research Objectives. In this paper, we propose ContrastDistAA, a novel AA approach that leverages contrastive learning and mutual information to disentangle topic and style information in the latent space. This allows us to handle topic shift settings and conduct AA at both individual and regional levels. To facilitate our investigations, we construct a new dataset to support the regional-level AA task. We conduct extensive experiments to evaluate ContrastDistAA against state-of-the-art baselines on both regional-level and individual-level AA tasks.

Contributions. Our work makes the following contributions: (i) We introduce a new regional-level AA task and a dataset to support the evaluation of AA methods on this new task. (ii) We propose ContrastDistAA, which can disentangle content and style information to improve AA performance. (iii) We conduct extensive experiments to benchmark ContrastDistAA against state-of-the-art AA methods. Our experiment results demonstrate ContrastDistAA's superior performance in both individual-level and regional-level AA tasks. This study not only fills a gap in the AA literature but also sheds light on the intricate interplay between linguistic styles and cultural elements within the realm of AA, offering new perspectives and understanding in the field.

2 Related Work

2.1 Authorship Attribution

AA has been extensively researched, with recent surveys providing comprehensive overviews of seminal works and advancements in the field (Zheng and Jin, 2023; Tyo et al., 2022). Researchers primarily relied on heuristic and statistical approaches in the nascent stages of AA. These involved the usage of basic stylometric features such as word lengths, sentence lengths, and function words (Neal et al., 2017; Ding et al., 2017). This phase evolved with the training of classical machine learning algorithms as classifiers to link these stylometric features with author identities (Boenninghoff et al., 2019b,a; Theóphilo et al., 2019). The emergence of deep learning marked a significant shift in AA, enabling the learning of more complex writing patterns (Shrestha et al., 2017; Hu et al., 2020; Jafariakinabad et al., 2019; Liu et al., 2021). The introduction of pre-trained language models like BERT (Devlin et al., 2019) further revolutionized AA, achieving state-of-the-art results through fine-tuning for specific AA tasks (Rivera-Soto et al., 2021; Manolache et al., 2021; Reimers and Gurevych, 2019; El Boukkouri et al., 2020). However, these techniques often performed poorly in topic-shift scenarios, where the topics under evaluation during the testing phase are not represented in the training data (Altakrori et al., 2021). Our ContrastDistAA approach aims to overcome this challenge by employing contrastive learning and mutual information to separate content (i.e., topic) and linguistic style in latent space for AA.

2.2 Contrastive Learning

Contrastive Learning has emerged as a pivotal approach in forming embedding spaces, where it clusters similar data points together while distancing dissimilar ones. Its efficacy is particularly evident in computer vision, as seen in the work of Chen et al. (2020) with their data augmentation framework, and He et al. (2020) through the Momentum Contrast (MoCo) for enhanced representation learning. In Natural Language Processing (NLP), contrastive learning has been instrumental in refining sentence representations, exemplified by the methodologies of Giorgi et al. (2021) and Gao et al. (2022), which utilize contrastive loss for learning textual embeddings. Additionally, significant progress has been made in formulating strategies for generating positive and negative samples, with Robinson et al. (2021) addressing the challenge of hard negatives through user-controlled sampling.

2.3 Disentangled Representation Learning

Disentangled Representation Learning, a method that isolates distinct attributes of data into separate variables, has significantly influenced various fields. In computer vision, it is exemplified by CycleGAN, which uses latent embeddings for image translation without paired examples (Zhu et al., 2020). In speech processing, this approach involves using mutual information minimization to separate voice style from content (Yuan et al., 2021). In NLP, models like ADNet, which combine motivational and adversarial losses, effectively disentangle style and meaning in text embeddings (Romanov et al., 2019). Notable developments include the multi-decoder model of Fu et al. (2017) for text transfer tasks with limited parallel corpora and Shen et al. (2020)'s use of denoising objectives for sentence reconstruction. Inspired by these advances, our work adopts a similar approach to meticulously disentangle content and style information in textual data for the AA task.

3 Methodology

This section outlines our proposed model, ContrastDistAA, designed to learn a disentangled representation of writing style for AA. As depicted in Figure 1, ContrastDistAA is structured in two

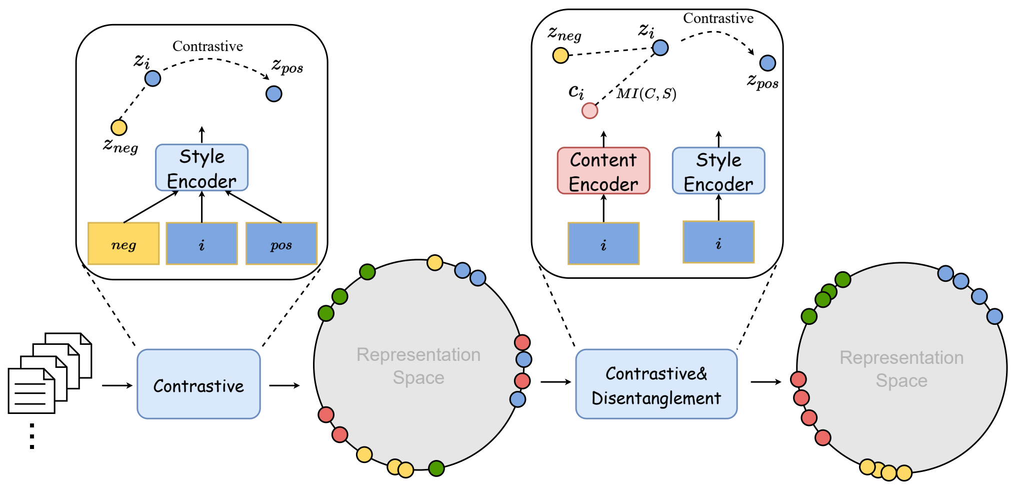

Figure 1: The architecture overview of ContrastDistAA model. The proposed models contains two-stages training process: (i) training using contrastive loss, and (ii) training using both contrastive loss with disentanglement loss.

distinct phases. The initial phase employs supervised contrastive loss to extract key stylistic features from labeled data. However, given the potential for content-related information to be intertwined with style, thus impacting the robustness of AA models, the subsequent phase of ContrastDistAA introduces a mutual information-based approach. This technique aims to separate style and content representations in the latent space, thereby enhancing the effectiveness of contrastive learning by clearly differentiating between style and content-specific attributes, including topical elements.

In subsequent sections, we will first review the contrastive learning component and the associated contrastive losses. This is followed by an introduction to mutual information, which is applied to disentangled representation learning for AA.

3.1 Contrastive Learning

Self-supervised representation learning has seen considerable progress in recent years, largely attributable to the application of contrastive learning (Wu et al., 2018; Hénaff et al., 2020; Oord et al., 2018; Chen et al., 2020). The fundamental mechanism of contrastive learning involves drawing an anchor and a "*positive*" sample closer in an embedding space, while simultaneously distancing the anchor from multiple "*negative*" samples, thus yielding meaningful representations. Specifically for AA tasks, we define "*positive pair*" consists of a text sample authored by the same individual as the anchor within a minibatch. In contrast, "*negative pairs*" are formed by aligning the anchor with randomly chosen samples from different authors within the same minibatch.

The initial phase of ContrastDistAA involves applying contrastive learning to train a style encoder, which extracts style features from texts authored by individuals or authors from specific regions. We utilize BERT (Devlin et al., 2019), acclaimed for its proficiency in capturing writing styles, as the style encoder. This encoder transforms discrete text into representations within latent space. Following this, supervised contrastive loss is applied to align representations of texts by the same author or from the same region more closely, while simultaneously distinguishing those from different authors or regions. This methodology enhances the style encoder's ability to discern and learn discriminative style representations.

3.1.1 Supervised Contrastive Loss for AA

In the ContrastDistAA model, we implement a supervised contrastive loss for AA. Consider a batch consisting of N textual samples from distinct authors. Let $i \in I \equiv \{1, 2, \cdots, N\}$ represent an individual sample in the minibatch, and let $A(i) \equiv I \backslash \{i\}$ denote the set of other texts excluding i. The negative samples for anchor i, denoted as $NEG(i) \equiv \{neg \in A(i) : y_{neg} \neq y_i\}$, are those not sharing the same author as i, while $POS(i) \equiv \{pos \in A(i) : y_{pos} = y_i\}$ represents the positive samples, sharing the same author as i. The supervised contrastive loss is particularly effective in scenarios where multiple samples belong to the same class, as it utilizes the available labels (Khosla et al., 2021). The formulation of the supervised contrastive loss for AA tasks is as follows:

$$L^{sup} = \sum_{i \in I} \frac{-1}{|POS(i)|} \sum_{pos \in POS(i)} log \frac{exp(z_i \cdot z_{pos}/\tau)}{\sum_{neg \in NEG(i)} exp(z_i \cdot z_{neg}/\tau)} \quad (1)$$

where $z_i = StyleEncoder(x_i)$, the $\cdot$ symbol denotes the inner product, $\tau \in \mathcal{R}^+$ is a scalar temperature parameter, $POS(i) \equiv \{pos \in A(i) : y_{pos} = y_i\}$ is the set of indices of all positive samples distinct from i, and $|POS(i)|$ is its cardinality.

3.2 Mutual Information for Style-Content Disentanglement

The style encoder, trained using supervised contrastive loss, becomes proficient at extracting representations that encapsulate both style and content attributes. Therefore, to refine the style encoder's focus on capturing writing style more distinctly, we integrate mutual information with contrastive learning. This synergy aims to separate style and content information within the latent space.

Mutual information, a fundamental concept in information theory, measures the dependence between two random variables. For our model, mutual information between style (z) and content (c) representations is crucial. Its mathematical definition involves the expectation of the logarithm of the ratio of the joint distribution of z and c to their respective marginal distributions, which can be expressed as follows:

$$I(z;c) = \mathbb{E}_{p(z,c)}[log \frac{p(z,c)}{p(z)p(c)}] \quad (2)$$

In practice, accurately calculating mutual information is challenging due to the intractability of the integral involved (Chen et al., 2016; Belghazi et al., 2018; Poole et al., 2019). To address this, we employ the Contrastive Log-ratio Upper Bound (CLUB) estimation method (Cheng et al., 2020). This approach is particularly suitable when conditional distributions such as $p(z|c)$ or $p(c|z)$ are not explicitly available. We approximate $p(z|c)$ using a variational distribution $q_\theta(z|c)$, parameterized by θ, leading to the definition of the variational CLUB term (vCLUB) as follows:

In disentangled representation learning, a common objective is to minimize the mutual information between varying types of embeddings, aligning with our training target (Poole et al., 2019). However, determining the exact value of mutual information presents challenges in practical settings, as the integral in Eq. 2 is often intractable. To overcome this, several mutual information estimation methods have been proposed (Chen et al., 2016; Belghazi et al., 2018; Poole et al., 2019). We employ the estimation method known as the Contrastive Log-ratio Upper Bound (CLUB) (Cheng et al., 2020), which is suitable for the scenario where the conditional distributions $p(z|c)$ or $p(c|z)$ is not provided. A variational distribution $q_\theta(z|c)$ with parameter θ is used to approximate $p(z|c)$.Consequently, a variational CLUB term (vCLUB) is defined as follows:

$$I_{vCLUB}(z;c) := \mathbb{E}_{p(z,c)}[logq_\theta(z|c)] - \mathbb{E}_{p(z)}\mathbb{E}_{p(c)}[logq_\theta(z|c)] \quad (3)$$

The unbiased estimator for vCLUB is derived from a set of samples, effectively quantifying the mutual information in a computationally feasible manner. unbiased estimator for vCLUB with sample $\{z_i, c_i\}$ is expressed as follows:

$$\hat{I}_{vCLUB} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} [logq_\theta(z_i|c_i) - logq_\theta(z_j|c_i)] = \frac{1}{N} \sum_{i=1}^{N} [logq_\theta(z_i|c_i) - \frac{1}{N} \sum_{j=1}^{N} logq_\theta(z_j|c_i)]. \quad (4)$$

In summary, to facilitate style-content disentanglement in ContrastDistAA, we first deploy a content encoder, also a BERT model, to extract content representations, denoted as c. Meanwhile, the pretrained style encoder from the first stage extracts style representations, denoted as z. Each post i thus has two distinct representations: the content representation c_i and the style representation s_i. Here, we apply the vCLUB estimator to minimize the mutual information between these content and style representations, refining the distinctiveness of each. Concurrently, the supervised contrastive loss continues to enhance the style encoder's ability to capture writing style nuances. During the evaluation phase, only the style encoder is used to extract style representations from posts authored by individuals or from specific regions. The regional or individual author style representations are then calculated by averaging the post-style representations, facilitating a comprehensive and nuanced assessment of writing styles.

Dataset	#Users	#Train	#Valid	#Test
Regional Tweets	87,836	382,598	42,513	42,513
CCAT50	50	1,766	442	465
Twitter1000	1,000	6,000	2,000	2,000
IMDB62	62	37,200	12,400	12,400

Table 1: Statistics of datasets

4 Experiments

4.1 Experimental Settings

Datasets. To evaluate ContrastDistAA effectively on both individual and regional AA tasks, we utilize four datasets in our experiments. The statistical distributions of the datasets are shown in Table 1.

Regional Tweets: This dataset, aimed at exploring regional writing styles, comprises English tweets from Southeast Asia, collected using the Twitter API from 2021 to 2022. It includes 425,111 tweets from 87,836 users across six regions: Singapore, Kuala Lumpur, Manila, Jakarta, Hanoi, and Bangkok. The selection criteria focused on English tweets with more than three words for better data quality. The dataset is divided into training, validation, and testing sets in an 8:1:1 ratio.

CCAT50: A subset of the Reuters Corpus and a prominent resource in AA research, the CCAT50 dataset (Liu et al., 2012) focuses on the top 50 contributors in the CCAT (corporate/industrial) subtopic. It consists of 5,000 texts (50 per author) divided into distinct training, validation, and testing sets following a 6:2:2 ratio, based on the processed version by (Tyo et al., 2022).

Twitter1000: Derived from a larger Twitter dataset used in AA research (Shrestha et al., 2017; Schwartz et al., 2013), Twitter1000 includes tweets from the top 1,000 authors by volume, with 100 tweets randomly selected from each. The dataset is organized into training, validation, and testing subsets, also following a 6:2:2 ratio.

IMDB62: Recognized for long-text AA studies (Seroussi et al., 2014), the IMDB62 dataset includes contributions from 62 authors, each providing 1,000 texts. Similar to the others, this dataset is partitioned into training, validation, and testing sets in a 6:2:2 ratio.

Evaluation Metrics. Following existing AA studies, we adopt Macro-F1 and Micro-F1 as the evaluation metrics in our experiments.

4.2 Baselines

We benchmark our model against commonly used and state-of-the-art AA models. These baselines are trained or fine-tuned to perform both the regional-level and individual-level AA tasks.

LR-Stylo: This logistic regression model, leveraging stylometric features as inputs, is grounded in prior research (Sari, 2018; Aborisade and Anwar, 2018). Based on (Fabien et al., 2020), it uses ten different stylometric features like text length and word count for classification.

LR-TF-IDF: Employing Term Frequency - Inverse Document Frequency (TF-IDF) at the word level, this logistic regression classifier follows the approach of (Fabien et al., 2020). Pre-processing includes stemming and stop-word removal before constructing the TF-IDF features.

LR-Char: This model uses character N-gram-based features, shown to be effective in AA (Bischoff et al., 2020; Shrestha et al., 2017; Altakrori et al., 2021). Following (Tyo et al., 2022), the logistic regression classifier is trained with a mix of character N-gram, part-of-speech N-gram, and summary statistics.

LSTM: An LSTM model, inspired by recent studies (Oliva et al., 2022), incorporates a dense layer followed by a max pooling layer. It focuses on the hidden states of the LSTM for AA tasks.

BertAA: Utilizing a pre-trained BERT language model, BertAA (Fabien et al., 2020) is fine-tuned specifically for AA, integrating a dense layer and softmax activation function for AA classification.

DistilBert: Known for its efficiency as a compact language model, DistilBERT (Sanh et al., 2019) is fine-tuned for AA tasks.

Roberta: Employing the Roberta model (Liu et al., 2019), we follow the original hyperparameters and fine-tune it on AA datasets over a specific number of epochs.

4.3 Implementation.

Our experiments were carried out on a system operating on Ubuntu 20.04.3 LTS, equipped with robust hardware specifications including 24 CPU cores, 128 GB of RAM, and a base clock speed of 2.9 GHz. To facilitate efficient training of the pre-trained models, Nvidia GTX 3090 graphics cards were utilized. BERT, with its pre-trained weights, served dual roles as both the style and content encoders in our experiment, which was divided into two distinct stages.

Method	Regional Tweet		CCAT50		Twitter1000		IMDB62	
	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1
LR-Stylo	0.176	0.251	0.013	0.037	0.019	0.035	0.013	0.037
LR-TF-IDF	0.402	0.446	0.554	0.554	0.566	0.566	0.554	0.554
LR-Char	0.252	0.308	0.180	0.209	0.077	0.128	0.503	0.503
LSTM	0.186	0.290	0.244	0.274	0.124	0.126	0.307	0.326
BertAA	0.433	0.472	0.518	0.512	0.226	0.249	0.627	0.654
DistilBERT	0.407	0.449	0.453	0.447	0.213	0.242	0.402	0.441
Roberta	0.476	0.522	0.466	0.497	0.622	0.626	0.735	0.749
ContrastDistAA	**0.510**	**0.550**	**0.578**	**0.584**	**0.960**	**0.961**	**0.813**	**0.816**
ContrastDistAA (w/o dist)	0.505	0.508	0.552	0.566	0.960	0.916	0.803	0.816

Table 2: Macro and Micro F1 scores for baselines and ContrastDistAA on four benchmark datasets.

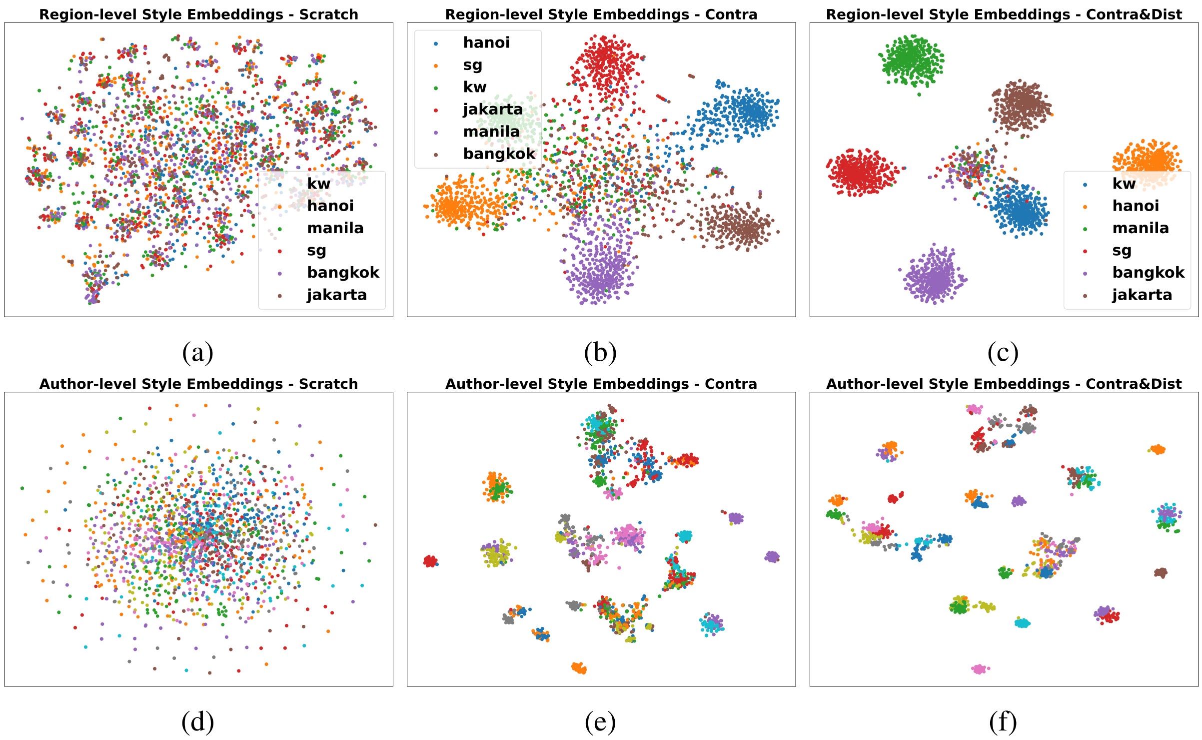

Figure 2: t-SNE visualization of posts from Regional Tweets and CCAT50. Specifically, we select 100 posts from each region in the Regional Tweets dataset and 50 posts from each author in CCAT50. The top three visualizations display the posts from Regional Tweets, while the bottom three pertain to CCAT50.

We train ContrastDistAA in two stages. In the first stage, the style encoder was the sole focus, trained using supervised contrastive loss over 30 epochs. The subsequent stage marked the joint training of both the content and style encoders. This phase, extending for an additional 20 epochs, employed supervised contrastive loss alongside a mutual information estimator. The implementation of the mutual information estimator was based on the source code[1] provided by (Cheng et al., 2020). Consistency in training parameters was maintained throughout, with a learning rate set at 1e-3 and a batch size of 32 for both stages. This setup ensured a balanced and rigorous training process for the ContrastDistAA model.

[1] https://github.com/Linear95/CLUB

4.4 Experiment Results

In our study, the efficacy of the ContrastDistAA model was thoroughly assessed on both regional and individual-level datasets, with its performance benchmarked against a range of established baseline models. The comparative results, evaluated using F1 scores, are detailed in Table 2.

The ContrastDistAA model consistently exhibited superior performance across these datasets. For instance, within the Regional Tweets dataset, it attained a Micro F1 score of 0.55, representing a notable 7% improvement compared to the next closest model, BertAA. In the context of the CCAT50 dataset, ContrastDistAA surpassed all baselines in every evaluated metric, achieving a significant 16% improvement in Micro F1 scores. The model also demonstrated exceptional performance on the Twitter1000 dataset, registering a substantial 29% increase in F1 scores. Furthermore, on the IMDB62 dataset, ContrastDistAA achieved a 6.7% improvement in performance, indicative of its robustness even in the presence of textual complexity. These results collectively affirm the ContrastDistAA model's capability in effectively discerning writing styles at both regional and individual levels, thereby establishing it as a state-of-the-art benchmark in the AA tasks.

Interestingly, we also noted that the models' F1 scores are generally lower for the Regional Tweets dataset, suggesting the difficulty of the region-level AA task. The individual authors typically have more distinct and consistent writing styles compared to a group of authors from a region. This uniqueness in individual writing styles makes it easier for models to attribute authorship accurately, leading to higher F1 scores. In contrast, regional-level AA deals with broader, less distinct writing styles shared by a group, which can be more challenging to differentiate.

4.5 Ablation Study

We also conduct an ablation study, which aimed to assess the impact of the dual-stage training process on ContrastDistAA. This study involved comparing the model's performance after the initial training phase, which utilized solely contrastive loss, against its performance following the second training stage that integrated both contrastive and disentanglement losses. The results, detailed in last two rows of Table 2, emphasize the significant contribution of representation disentanglement learning to the model's efficacy.

Crucially, the findings reveal that ContrastDistAA demonstrates an improvement in F1 scores when the disentanglement loss is incorporated in the second training stage, compared to the model trained only with contrastive loss. This improvement underscores the value of the second training stage in enhancing the model's capability. By effectively separating content-related elements from style-related information in the training process, the model becomes more adept at isolating and recognizing distinctive stylistic features inherent to different regional writings. This separation is key to the improved performance, illustrating the effectiveness of the comprehensive two-stage training approach in ContrastDistAA.

4.6 Qualitative Analysis

To demonstrate the efficacy of ContrastDistAA, we employed the t-SNE algorithm (Van der Maaten and Hinton, 2008) to visually represent post style embeddings in two-dimensional space. This visualization aimed to show how different training methodologies influence the distribution of post representations. We selected 100 posts from each region in the Regional Tweets dataset and 50 posts per author from the CCAT50 dataset, extracting their latent representations using three approaches: (i) BERT in its basic form, (ii) a style encoder trained with contrastive loss, and (iii) a style encoder trained using both contrastive loss and mutual information.

Figure 2 presents these representations. The first three visualizations pertain to posts from the Regional Tweets dataset, while the latter three focus on the CCAT50 dataset. Notably, with the application of contrastive loss, distinct clusters emerge, indicating the style encoder's ability to capture style information effectively. However, challenges are evident, such as the central clustering in Figure 2 (b), reflecting the limitations of contrastive learning with complex samples. The incorporation of mutual information for disentangling content and style in latent space results in more distinct clustering patterns, as seen in Figure 2 (c). This suggests that the integration of both contrastive and disentanglement learning notably enhances the style encoder's capability to discern style information, thereby improving its application in AA tasks.

4.7 Case Studies for Regional AA

To highlight the unique writing styles prevalent in different regions, we conducted a linguistic analysis of posts from these areas. This involved selecting three posts from each region and calculating the cosine similarity between their representations and the corresponding regional style representations, providing insights into how closely these posts align with predominant regional writing styles.

Our analysis revealed distinct linguistic features characteristic of each region, often embodied in specific words or expressions that encapsulate regional nuances and evoke emotional responses. For instance, authors from Bangkok frequently conclude sentences with unique words such as "*kub*", "*naka*", "*krub*", or "*na*" adding an expressive and emotive quality to their writing. In Jakarta, au-

Regions	Sentences	Similarity
Bangkok	1. @USER thank u naa	0.825
	2. @USER You're very welcome I feel honored and very happy . ka pleading_face two_hearts	0.977
	3. @USER You make all of us lazy people feel ashamed on a Sunday morning na krub .	0.990
Hanoi	1. isit Indonesian #Booth in Ly Thao To Park , DATE	0.995
	2. Those light is fierce ! #welldone @USER Trang Tien Plaza HTTPURL	0.996
	3. try some coconut coffee hot_beverage USER Cong Caphe HTTPURL	0.996
Jakarta	1. @USER Serem amat :loudly_crying_face:	0.991
	2. @USER batman who laughs lumayan lah atleast	0.951
	3. Mantul the babbies nyusul the daddies	0.992
Manila	1. Salamat sa live selling at unboxing ! Lol char . Love you bestie ! Congratulations ! HTTPURL	0.996
	2. Wow , salamat po sa Dios To God be the Glory sparkles Are Your Prayers Heard #PureDoctrinesOfChrist HTTPURL	0.996
	3. DATE nabudol ako sa film life . Excited for youuuuuu . @USER Stay Broke , Shoot Film . HTTPUR . HTTPURL	0.959
Singapore	1. STOp . the tarot card readings gotta STOOOOOOOop pls lah	0.857
	2. So much things on my mind rn ! Inshallah all goes well	0.651
	3. @USER i no have scandal lehhh u my one and only	0.984
Kuala Lumpur	1. Pusing lah kot mana pun , no one else is calling it democratic . Except PN of course	0.908
	2. Say goodbye to grainy spycam footage . Tak main lah video quality Nokia	0.501
	3. adut saya order 138 utk pastikan bontot staff saya 8p m 5pm tak ke Pavilion	0.964

Table 3: Examples showcasing the unique writing expressions (highlighted in yellow) from each region. The similarity score is the cosine similarity between the post representation and the region style embedding.

thors use expressions like "lumayan" to indicate a moderate experience, "seem amat" for excitement, and "mantul" to denote something extraordinary, showcasing the rich and diverse writing style of this region. Hanoi's writing style, influenced by the modern Latin script and its use of diacritical marks, often features Vietnamese words without these marks. This use reflects a blend of traditional and contemporary linguistic practices, allowing for effective communication while honoring the linguistic heritage and subtleties of the region. These findings underscore the distinct linguistic identities of each region, as mirrored in their writing styles.

5 Conclusion

In this study, we introduced ContrastDistAA, a model designed to effectively separate content and style information, thereby enhancing AA performance. A significant contribution of our research is the introduction of the regional-level AA task, along with a dedicated dataset to evaluate AA methods in this new context. Through comprehensive experiments, ContrastDistAA was benchmarked against state-of-the-art AA techniques, demonstrating its superior performance in both individual-level and regional-level AA tasks.

The results from our case studies indicate that ContrastDistAA is adept at identifying unique linguistic features indicative of regional writing styles. Specifically, the contrastive learning and representation disentanglement approach have helped to effectively segregate content from stylistic features for AA tasks. This capability is crucial for understanding how linguistic styles and cultural influences interplay in AA. Our research addresses a previously unexplored aspect of AA and offers fresh perspectives on the relationship between linguistic styles and cultural elements.

For future work, we will focus on further exploring regional and cultural writing styles. We aim to include a broader range of cultural characteristics and regional diversity, thereby enhancing the understanding of AA in diverse linguistic and cultural contexts. This ongoing research will continue to expand the horizons of AA, contributing to a deeper understanding of the intricate relationship between authorship, language, and culture.

6 Limitations

This study makes noteworthy contributions to the field of Authorship Attribution (AA), but it also acknowledges two key limitations. The first limitation pertains to the methodology of obtaining style representations for regions and authors, which is based on averaging post representations. This approach, while practical, is susceptible to the cluster center shift problem, especially when outliers are included in the calculations. Outliers can significantly skew the average, leading to potential misrepresentations of the typical writing style of a region or an author.

The second limitation is the geographical scope of the dataset used. The Regional Tweets dataset is confined to six regions within Southeast Asian countries, which, while providing valuable regional insights, limits the broader applicability and generalizability of the study's findings. To enhance the scope and robustness of future research in AA, it would be beneficial to include more diverse regions from various countries and cultural areas. This expansion would offer a more comprehensive understanding of the diverse linguistic and stylistic nuances that characterize writing styles globally, and contribute to the development of AA methods that are universally relevant and sensitive to regional and cultural variations.

References

Opeyemi Aborisade and Mohd Anwar. 2018. Classification for authorship of tweets by comparing logistic regression and naive bayes classifiers. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 269–276. IEEE.

Malik Altakrori, Jackie Chi Kit Cheung, and Benjamin C. M. Fung. 2021. The topic confusion task: A novel evaluation scenario for authorship attribution. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4242–4256, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR.

Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2019. Generalizing unmasking for short texts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 654–659.

Sebastian Bischoff, Niklas Deckers, Marcel Schliebs, Ben Thies, Matthias Hagen, Efstathios Stamatatos, Benno Stein, and Martin Potthast. 2020. The importance of suppressing domain style in authorship analysis. *arXiv preprint arXiv:2005.14714*.

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

Benedikt Boenninghoff, Robert M Nickel, Steffen Zeiler, and Dorothea Kolossa. 2019b. Similarity learning for authorship verification in social media. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2457–2461. IEEE.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. 2020. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pages 1779–1788. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Steven HH Ding, Benjamin CM Fung, Farkhund Iqbal, and William K Cheung. 2017. Learning stylometric representations for authorship analysis. *IEEE transactions on cybernetics*, 49(1):107–121.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Jun'ichi Tsujii. 2020. CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6903–6915, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Maël Fabien, Esau Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. BertAA : BERT fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2017. Style transfer in text: Exploration and evaluation.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2022. Simcse: Simple contrastive learning of sentence embeddings.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. Declutr: Deep contrastive learning for unsupervised textual representations.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735.

Xinyu Hu, Weihan Ou, Sudipta Acharya, Steven HH Ding, Ryan D'Gama, and Hanbo Yu. 2023. Tdrlm: Stylometric learning for authorship verification by topic-debiasing. *Expert Systems with Applications*, 233:120745.

Zhiqiang Hu, Roy Ka-Wei Lee, Lei Wang, Ee-peng Lim, and Bo Dai. 2020. Deepstyle: User style embedding for authorship attribution of short texts. In *Web and Big Data: 4th International Joint Conference, APWeb-WAIM 2020, Tianjin, China, September 18-20, 2020, Proceedings, Part II 4*, pages 221–229. Springer.

Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. 2020. Data-efficient image recognition with contrastive predictive coding.

Farkhund Iqbal, Rachid Hadjidj, Benjamin CM Fung, and Mourad Debbabi. 2008. A novel approach of mining write-prints for authorship attribution in e-mail forensics. *digital investigation*, 5:S42–S51.

Fereshteh Jafariakinabad, Sansiri Tarnpradab, and Kien A Hua. 2019. Syntactic recurrent neural network for authorship attribution. *arXiv preprint arXiv:1902.09723*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2021. Supervised contrastive learning.

Jianbo Liu, Zhiqiang Hu, Jiasheng Zhang, Roy Ka-Wei Lee, and Jie Shao. 2021. A syntax-aware encoder for authorship attribution. In *Web Information Systems Engineering–WISE 2021: 22nd International Conference on Web Information Systems Engineering, WISE 2021, Melbourne, VIC, Australia, October 26–29, 2021, Proceedings, Part I 22*, pages 403–411. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhi Liu, Zongkai Yang, Sanya Liu, and Wenting Meng. 2012. A novel random subspace method for online writeprint identification. *J. Comput.*, 7(12):2997–3004.

Andrei Manolache, Florin Brad, Elena Burceanu, Antonio Barbalau, Radu Ionescu, and Marius Popescu. 2021. Transferring bert-like transformers' knowledge for authorship verification. *arXiv preprint arXiv:2112.05125*.

Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying stylometry techniques and applications. *ACM Computing Surveys (CSuR)*, 50(6):1–36.

Christian Oliva, Santiago Palmero Muñoz, Luis F Lago-Fernández, and David Arroyo. 2022. Improving lstms' under-performance in authorship attribution for short texts. In *Proceedings of the 2022 European Interdisciplinary Cybersecurity Conference*, pages 99–101.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. 2019. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

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*, pages 913–919, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*.

Alexey Romanov, Anna Rumshisky, Anna Rogers, and David Donahue. 2019. Adversarial decomposition of text representation.

Chakaveh Saedi and Mark Dras. 2021. Siamese networks for large-scale author identification. *Computer Speech & Language*, 70:101241.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Upendra Sapkota, Thamar Solorio, Manuel Montes, Steven Bethard, and Paolo Rosso. 2014. Cross-topic authorship attribution: Will out-of-topic data help? In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1228–1237.

Yunita Sari. 2018. *Neural and Non-neural Approaches to Authorship Attribution*. Ph.D. thesis, University of Sheffield.

Roy Schwartz, Oren Tsur, Ari Rappoport, and Moshe Koppel. 2013. Authorship attribution of micro-messages. In *Proceedings of the 2013 Conference on empirical methods in natural language processing*, pages 1880–1891.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2011. Authorship attribution with latent dirichlet allocation. In *Proceedings of the fifteenth conference on computational natural language learning*, pages 181–189.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2014. Authorship attribution with topic models. *Computational Linguistics*, 40(2):269–310.

Tianxiao Shen, Jonas Mueller, Regina Barzilay, and Tommi Jaakkola. 2020. Educating text autoencoders: Latent representation guidance via denoising.

Prasha Shrestha, Sebastian Sierra, Fabio A 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*, pages 669–674.

Efstathios Stamatatos and Moshe Koppel. 2011. Plagiarism and authorship analysis: introduction to the special issue. *Language Resources and Evaluation*, 45:1–4.

Antônio Theóphilo, Luís AM Pereira, and Anderson Rocha. 2019. A needle in a haystack? harnessing onomatopoeia and user-specific stylometrics for authorship attribution of micro-messages. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2692–2696. IEEE.

Jacob Tyo, Bhuwan Dhingra, and Zachary C Lipton. 2022. On the state of the art in authorship attribution and authorship verification. *arXiv preprint arXiv:2209.06869*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Andrew Wang, Cristina Aggazzotti, Rebecca Kotula, Rafael Rivera Soto, Marcus Bishop, and Nicholas Andrews. 2023. Can authorship representation learning capture stylistic features? *Transactions of the Association for Computational Linguistics*, 11:1416–1431.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742.

Siyang Yuan, Pengyu Cheng, Ruiyi Zhang, Weituo Hao, Zhe Gan, and Lawrence Carin. 2021. Improving zero-shot voice style transfer via disentangled representation learning.

Richong Zhang, Zhiyuan Hu, Hongyu Guo, and Yongyi Mao. 2018. Syntax encoding with application in authorship attribution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2742–2753, Brussels, Belgium. Association for Computational Linguistics.

Wanwan Zheng and Mingzhe Jin. 2023. A review on authorship attribution in text mining. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(2):e1584.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2020. Unpaired image-to-image translation using cycle-consistent adversarial networks.