Whodunit? Learning to Contrast for Authorship Attribution

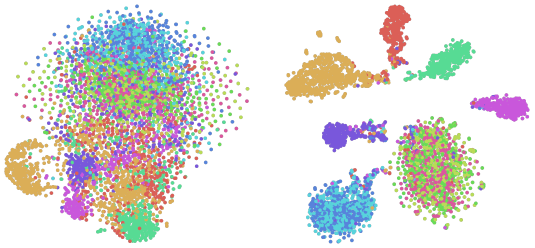
Anonymous ACL submission

Abstract

Authorship attribution is the task of identifying the author of a given text. Most existing approaches use manually designed features that capture a dataset's content and style. However, this dataset-dependent approach yields inconsistent performance. Thus, we propose to fine-tune pretrained language representations using a combination of contrastive learning and supervised learning (Contra-X). We show that Contra-X advances the state-of-the-art on multiple human and machine authorship attribution benchmarks, enabling improvements of up to 6.8%. We also show Contra-X to be consistently superior to cross-entropy fine-tuning across different data regimes. Crucially, we present qualitative and quantitative analyses of these improvements. Our learned representations form highly separable clusters for different authors. However, we find that contrastive learning improves overall accuracy at the cost of sacrificing performance for some authors. Resolving this tension will be an important direction for future work. To the best of our knowledge, we are the first to analyze the effect of combining contrastive learning with cross-entropy fine-tuning for authorship attribution.[1]

1 Introduction

Authorship attribution (AA) is the task of identifying the author of a given text. AA systems are commonly used to identify the authors of anonymous email threats (Iqbal et al., 2010) and historical texts (Mendenhall, 1887), and to detect plagiarism (Gollub et al., 2013). With the rise of neural text generators that are able to create highly believable fake news (Zellers et al., 2019), AA systems are also increasingly employed in machine-generated-text detection (Jawahar et al., 2020). When performed on texts generated by human and machine writers, AA can also act as a type of *Turing Test* for Natural Language Generation (Uchendu et al., 2021, 2020).



(a) BERT (b) Contra-BERT

Figure 1: t-SNE visualization of the fine-tuned representations (a: baseline; b: Contra-X). Each color denotes one author in the Blog10 dataset. Our contrastive method effectively creates a tighter representation spread for each author and increased separation between authors. Best viewed in color.

Traditional AA methods design features that characterize texts based on their content or writing style (Jafariakinabad and Hua, 2019; Zhang et al., 2018; Sapkota et al., 2015b; Sari et al., 2018). However, the features useful for distinguishing authors are often dataset-specific, yielding inconsistent performance under varying conditions (Sari et al., 2018). In contrast, learning features from large corpora of data aims to produce general pretrained models (Devlin et al., 2018) that improve performance on many core natural language processing (NLP) tasks, including AA (Fabien et al., 2020). However, it is unclear if basic fine-tuning makes full use of the information in the training data. We seek to augment the learning process.

Contrastive learning is a technique that pulls similar samples close and pushes dissimilar samples apart in the representation space (Gao et al., 2021). It has proven useful in tasks that require distinguishing subtle differences (Tian et al., 2020; Kawakami et al., 2020). This makes it highly suited to encouraging the learning of distinct author subspaces. However, no prior work has investigated its relevance to the AA task. To this end, we seek to under-

[1] Code will be made available at `<redacted>`

stand its impact on the learning of author-specific features under the supervised learning paradigm.

To achieve this, we combine **CONTRA**stive learning with **CROSS**-entropy finetuning (**Contra-X**) and demonstrate its efficacy via evaluation on multiple AA datasets. Unlike previous AA work, we evaluate not only on human writing corpus, but also on machine-generated texts. There are three major reasons. First, this can show the generality of our approach. Secondly, performing AA on human and machine authors reflects the increased importance of identifying machine-generated text sources. Thirdly, this potentially reveals information about how differently machines write compared to humans. In addition, we study the performance of our method under different data budgets. We find Contra-X to consistently improve model performance and yield distinct author subspaces. Finally, we analyze the performance gains vis-à-vis a number of AA-specific stylometric features. To the best of our knowledge, we are the first to investigate the use of contrastive learning for authorship attribution.

2 Related Work

Authorship attribution. AA techniques fall under two broad categories: feature-based or learning-based approaches. Traditional methods manually engineer stylometric features relevant for identifying authors (Sari et al., 2018). These include term frequency–inverse document frequency (TF-IDF) (Rahgouy et al., 2019), letter and digit frequency (Sari et al., 2018), and character n-grams (Sapkota et al., 2015a). However, as various datasets have different latent properties, e.g., topical or content biases, dataset-specific design is often required for optimal performance (Sari et al., 2018).

In contrast, we use a more general solution of *learning* task-specific feature representations. BertAA (Fabien et al., 2020) showed that simply fine-tuning pre-trained language models with a cross-entropy loss can produce excellent AA performance. This suggests that pre-trained general representations are a promising starting point. However, there remains the challenge of learning representations that effectively model author-specific characteristics. Our work makes use of a contrastive learning objective to achieve this goal.

Contrastive Learning. Contrastive learning aims to learn discriminative features by pulling semantically similar samples close and pushing dissimilar samples apart. This encourages the learning of highly separable features that can be easily operated on by a downstream classifier. Unsupervised contrastive learning has been used to improve the robustness and transferability of speech recognition (Kawakami et al., 2020) and to learn semantically meaningful sentence embeddings (Gao et al., 2021). It has also been combined with supervised learning for intent detection (Zhang et al., 2021), punctuation restoration (Huang et al., 2021), machine translation (Gunel et al., 2021), and dialogue summarization (Tang et al., 2021). However, to the best of our knowledge, we are the first to study its efficacy and limitations on authorship attribution.

Detection of Machine Generated Text. Natural Language Generation (NLG) models can generate texts indistinguishable from human writings (Radford et al., 2019; Brown et al., 2020; Zellers et al., 2019). Given the potential for malicious use (Solaiman et al., 2019), identifying machine-generated text sources has become increasingly important. Detecting artificial texts can be viewed as a special case of authorship attribution where there is a mix of human and non-human authors. In addition, authorship attribution models can serve as *Turing tests* for the NLG models (Uchendu et al., 2021), and advances in AA can also improve the evaluation of NLG models. We hence also evaluate our approach on TuringBench, which contains texts from both human and machine authors.

3 Methodology

3.1 Problem formulation

Authorship attribution is a classification task where the input is some text, t, and the target is the author, a. Formally, given a corpus $\mathcal{D}$, where each sample is a text-author pair $\langle t, a\rangle$, we aim to learn a predictor, p, that maximizes the prediction accuracy:

$$Acc = \mathop{\mathbb{E}}_{\langle t,a\rangle \in D} \mathbb{1}_{argmax(p(t))=a} \tag{1}$$

Conventionally, this is achieved by optimizing a surrogate cross-entropy loss function via mini-batch gradient descent. Assuming we have a mini-batch containing N texts $\{t_i\}_{i=1:N}$ and corresponding authors $\{a_i\}_{i=1:N}$, the loss function is:

$$\mathcal{L}_{CE} = -\sum_i a_i \log(p(t)_{a_i}) \tag{2}$$

However, we hypothesize that $\mathcal{L}_{CE}$ does not adequately reflect the key challenge of the task, which is to learn highly discriminative representations for the input texts such that authorship can be clearly identified. Thus, we propose to augment the loss with a contrastive learning objective.

3.2 Contra-X for Authorship Attribution

We conjecture that the key to the authorship attribution task is to learn highly author-specific representations that capture each author's characteristics. Specifically, this requires representations to be similar for samples from the same authors, but distinct for samples from different authors. We adopt two specific strategies to achieve this goal:

- Unlike most previous work that hand-crafts features and then learns a predictor p from scratch, we fine-tune the general representations acquired from the large-scale unsupervised pre-training. Specifically, we decompose p as $p = \phi \circ h$ where ϕ is the pre-trained language model and h is a classifier layer. As shown by BertAA (Fabien et al., 2020), the learned representation is a decent starting point for the task.
- However, different to BertAA that fine-tunes the model $p = \phi \circ h$ with cross entropy, we use an additional contrastive objective to encourage ϕ to capture the idiosyncrasies of each author. We conjecture that this can better utilize the information in the training data.

We use an additional contrastive objective to more fully utilize the information in the training data. Intuitively, the contrastive loss encourages the model to ***maximize*** the representational similarity of texts written by the same author, i.e., positive pairs, and ***minimize*** the representational similarity of texts written by different authors, i.e., negative pairs. Formally, given a mini-batch containing N texts $\{t_i\}_{i=1:N}$ and their authors $\{a_i\}_{i=1:N}$, we feed them into a pre-trained language model ϕ to obtain a batch of embeddings $\{e_i\}_{i=1:N}$, where $e_i = \phi(t_i)$. Embeddings of two samples by the same author $\langle e_i, e_j \rangle_{a_i=a_j}$ are a positive pair, and embeddings of two samples by different authors $\langle e_i, e_j \rangle_{a_i \neq a_j}$ are a negative pair. We construct a similarity matrix $\mathcal{S}$ in which the entry (i, j) denotes the pairwise similarity between e_i and e_j. Formally,

$$\mathcal{S}_{i,j} = \cos(e_i, e_j) = \frac{e_i \cdot e_j}{\|e_i\| \|e_j\|} \tag{3}$$

To encourage the abovementioned pairwise constraints, we define the contrastive objective as:

$$\begin{aligned}\mathcal{L}_{CL} &= -\sum_i \log(\frac{\sum_{a_i=a_j} \exp(\cos(e_i, e_j)/\tau)}{\sum_k \exp(\cos(e_i, e_k)/\tau)}) \\ &= -\sum_i \log(\frac{\sum_{a_i=a_j} \exp(\mathcal{S}_{i,j}/\tau)}{\sum_k \exp(\mathcal{S}_{i,k}/\tau))}),\end{aligned} \tag{4}$$

where τ is the temperature. The loss could be viewed as applied on a softmax distribution to maximize the probability that e_i and e_j come from a positive pair, given $a_i = a_j$. However, it is different from $\mathcal{L}_{CE}$ in that it explicitly enforces pairwise constraints in the representation space $\phi(\cdot)$. During training, we jointly optimize $\mathcal{L}_{CE}$ and $\mathcal{L}_{CL}$:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \cdot \mathcal{L}_{CL}, \tag{5}$$

where λ is a balancing coefficient. This joint optimization, Contra-X, improves upon $\mathcal{L}_{CE}$ by mining richer knowledge in the training data via encouraging meaningful pairwise relations in the representation space $\phi(\cdot)$. We conjecture that the model learn discriminative features in alignment with the classification objective. The effectiveness will be empirically examined (Section 4 and Section 5) and qualitatively analyzed (Section 6.2).

3.3 Implementation Details

We implement ϕ with two pre-trained transformer encoders, BERT (Devlin et al., 2018) and DeBERTa (He et al., 2021). BERT is a commonly used text classification baseline and DeBERTa, its more recent counterpart. We use the `bert-base-cased` and `deberta-base` checkpoints from the `transformers` library (Wolf et al., 2019). For all datasets, the input length is set to 256 and the embedding length per token is 768. The transformer generates embeddings which are then passed to the classifier h.

We implement the classifier h as a 2-layer Multi-Layer Perceptron (MLP) with a dropout of 0.35. As described in Section 3.2, the final model p is a composition of the pre-trained language model and the MLP classifier, i.e., $p = \phi \circ h$.

In all experiments, we use the AdamW optimizer (Loshchilov and Hutter, 2019) with an initial learning rate of $2e-5$ and a cosine learning rate schedule (Loshchilov and Hutter, 2017). We train for 8 epochs with a batch size of 24. We set λ to 1.0 and τ to 0.1. Training takes 2-12 hours depending on the dataset size with $4 \times$ RTX2080Ti. No model- or dataset-specific tuning was done for fair comparison and to show the robustness of the approach.

Model	Blog10	Blog50	IMDb62
Token SVM (Seroussi et al., 2014)	-	-	92.5
Char-CNN (Ruder et al., 2016)	61.2	49.4	91.7
Continuous N-gram (Sari et al., 2017)	61.3	52.8	95.1
N-gram CNN (Shrestha et al., 2017)	63.7	53.1	95.2
Syntax CNN (Zhang et al., 2018)	64.1	56.7	96.2
BertAA (Fabien et al., 2020)	65.4	59.7	93.0
BERT *(our baseline)*	60.4	55.2	97.2
Contra-BERT	66.3 (5.9↑)	62.0 (6.8↑)	97.9 (0.7↑)
DeBERTa *(our baseline)*	69.1	64.7	98.1
Contra-DeBERTa	**69.7 (0.6↑)**	**68.4 (3.7↑)**	**98.2 (0.1↑)**

Table 1: Results on human AA datasets, measured in accuracy.[2] Results in top section are from their respective papers. Improvements over the baselines are indicated in parentheses. The best model for each dataset is **bolded**.

4 Human Authorship Attribution

We first investigate the impact of contrastive learning on models for human authorship attribution.

4.1 Experiment setup

Models. We experiment with four different models: two baselines BERT and DeBERTa, fine-tuned with cross-entropy, and their Contra-X versions, where X is the model name. These baselines allow us to isolate the effect of the proposed contrastive learning objective $\mathcal{L}_{CL}$.

Datasets. Following prior work (Ruder et al., 2016; Zhang et al., 2018; Fabien et al., 2020), we use the Blog (Schler et al., 2006) and IMDb (Seroussi et al., 2014) corpora for evaluation. For Blog, we take the top 10 and 50 authors with the most entries to form the Blog10 and Blog50 datasets respectively. For IMDb, we take a standard subset of 62 authors (Seroussi et al., 2014) (IMDb62). More details are in Appendix A.

Evaluation. Following standard evaluation protocol, we divide each dataset into train/validation/test splits with an 8:1:1 ratio, and report the test split results here. Hyperparameter tuning, if any, is performed on the validation set. For easy comparison, we also present results on the 8:2 train/test splits used by Fabien et al. (2020) in Appendix B. We do not observe any significant differences.

4.2 Results

From Table 1, we observe that the inclusion of contrastive learning improves the baseline performance across the board, allowing us to beat the previous state-of-the-art on all human AA datasets. We observe that the largest performance improvements come from Blog10 and Blog50 datasets where there is substantial room for progress, i.e., up to 6.8% for BERT and 3.7% for DeBERTa. In contrast, the performance gains on IMDb62 are marginal due to diminishing returns, with the baseline models already achieving close to 100% accuracy. These results suggest that contrastive learning is empirically useful for fine-tuning pre-trained language models on the authorship attribution task, when the baseline performance is not approaching an asymptotic maximum.

5 Synthetic Text Authorship Attribution

We investigate our proposed models on authorship attribution datasets with neural-generated text.

5.1 Experimental Setup

Models. We test the same four models from Section 4: BERT, Contra-BERT, DeBERTa, and Contra-DeBERTa.

Dataset. We use the TuringBench (Uchendu et al., 2021) dataset. This corpus contains human-written news articles, collectively categorized as a single human author, and machine-generated texts from 19 different neural language generators. The models generate the texts based on the titles of the human-written articles. This controls for topic differences between samples by different authors. There are a total of 200,000 texts from 20 authors. Additional statistics are available in Appendix A.

Evaluation. We use the 7:1:2 train/validation/test splits provided by Uchendu et al. (2021) and report

Model	TuringBench
Random Forest	61.47
SVM (3-grams)	72.99
WriteprintsRFC	49.43
OpenAI Detector	78.73
Syntax CNN	66.13
N-gram CNN	69.14
N-gram LSTM-LSTM	68.98
BertAA	78.12
BERT	80.78
RoBERTa	81.73
BERT *(our baseline)*	79.46
Contra-BERT	80.59 (1.13↑)
DeBERTa *(our baseline)*	82.00
Contra-DeBERTa	**82.53 (0.53↑)**

Table 2: Results on human and machine authorship attribution (accuracy). Results in top section are from the TuringBench paper. Improvements over the baselines are indicated in parentheses. Best model is **bolded**.

the results on the test set.

5.2 Results

Table 2 shows the results of the synthetic authorship attribution benchmark.[3] Contrastive learning provides a small improvement in accuracy over the baseline models, in particular allowing Contra-DeBERTa to set a new state-of-art. These results suggest that the use of general language representations and contrastive learning is generalizable to synthetic authorship attribution.

6 Discussion

In this section, we study the following questions:

- How does data availability affect the performance with and without contrastive learning?
- How does contrastive learning qualitatively affect the representations learned?
- When does Contra-X succeed and fail?

6.1 Performance vs. Dataset Size

Due to the often-adversarial nature of real-world AA problems, the availability of appropriate data is a concern. Therefore, it is important to examine the impact of data availability on potential AA systems.

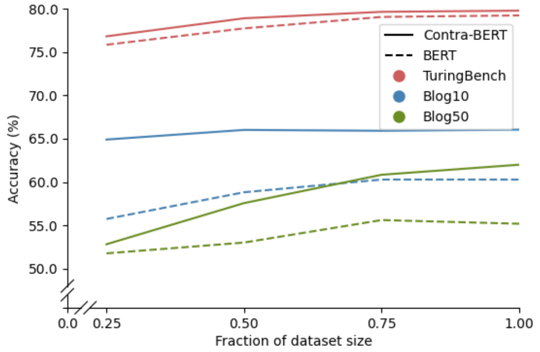


Figure 2: Comparison of performance between BERT and Contra-BERT under different data regimes.

To do this, we construct 4 subsets of the Blog10, Blog50, and TuringBench datasets with stratified sampling by author. Each subset is 25%, 50%, 75%, and 100% the size of the original dataset. We use the same setup as in Section 4.1 to train BERT and Contra-BERT on each subset.

Figure 2 plots accuracy vs. dataset size to illustrate the performance under different dataset sizes. On Blog10, Contra-BERT maintains a surprisingly consistent level of accuracy while BERT suffers significant degradation in performance as data decreases. On Blog50, Contra-BERT shows more substantial performance gains compared to BERT as the dataset size increases. We hypothesize that the task is intrinsically harder due to the larger number of authors, requiring a larger amount of data to learn well. Even so, Contra-X improves the performance of both BERT and DeBERTa by 6.8% and 3.7%, respectively, on the full dataset. On TuringBench, the difference in accuracy is less obvious, although Contra-BERT maintains the advantage. A possible explanation is that even the smaller subsets are sufficiently large.

From the above statistics, we notice consistent improvements across different data regimes. A possible explanation is that the contrastive objective explicitly encourages the model to focus on inter-author differences as opposed to irrelevant features.

6.2 Qualitative Representational Differences

Next, we visualize the learned representations to understand the qualitative effect of the contrastive learning objective. We embed the test samples from the Blog50 dataset and visualize the result using t-SNE (van der Maaten and Hinton, 2008).

Qualitatively, it is clear that Contra-BERT produces more distinct and tighter clusters compared

[3] Results of previous methods are from TuringBench (Uchendu et al., 2021). For consistency, we report results to 2 decimal places. For full results for other metrics, i.e., precision, recall, and F1-score, see Appendix F.

	Feature Type				Performance Improvement (Acc.)	
Dataset	Content	Style	Hybrid	Topic	BERT	DeBERTa
Blog10	0.82472	**0.33766**	**0.59218**	0.85465	5.9	0.6
Blog50	1.0000	1.0000	1.0000	**0.81145**	6.8	3.7
TuringBench	**0.60842**	0.56926	0.91988	1.0000	1.13	0.53

Table 3: Inter-author difference on different feature metrics (improvements from each contrastive model listed for reference). The smaller the value, the higher the similarity measured by that feature. For consistency, each column is linearly scaled such that the maximum is 1. The smallest value for each feature is **bolded**.

to BERT (Figure 1). Since $\mathcal{L}_{CL}$ is the only independent variable in the experiment, differences in representation can be attributed to the contrastive objective. The improvement is expected, because the objective $\mathcal{L}_{CL}$ explicitly encourages the representation to be similar for intra-author samples (i.e., tight clusters) and different for inter-author samples (i.e., larger distance between clusters). This supports our conjecture in Section 3.2.

However, we observe that some clusters still overlap and are inseparable by t-SNE. This suggests that the model still faces some difficulty in distinguishing between specific authors.

6.3 When Does Contra-X Succeed and Fail?

To understand the conditions in which Contra-X succeeds and fails, we follow Sari et al. (2018) and extract 4 stylometric features from the dataset: topic, style, content, and hybrid features. Detailed descriptions for each feature are in Appendix C. For this set of features, $\mathcal{F}$, the corresponding feature extractors are ϕ_f, $f \in \mathcal{F}$. We can then represent each author, A_i, with a feature. Given an author A_i with N documents $\{t_i\}_{i=1:N}$, we define the representation of A_i to be the mean of the vector representations of the N documents:

$$v^f_{A_i} = \frac{1}{N}\sum_{i=1}^{N} \phi_f(t_i). \tag{6}$$

We analyze the relation between model performance and dataset characteristics below. We exclude IMDb62 from this analysis since the maximum margin for improvement on the dataset is too small ($< 3\%$). Performing analysis on these datasets may introduce confounding factors.

Dataset-Level Analysis. Here, we wish to quantify the difficulty of distinguishing any two authors in each dataset and compare them against performance improvements. We define the inter-author dissimilarity of a dataset $\mathcal{D}$ in a feature space $f \in \mathcal{F}$ to be the mean pairwise difference across all author pairs $\langle A_i, A_j\rangle$ measured by the feature f:

$$v^f_{\mathcal{D}} = \frac{1}{|A|^2}\sum_{A_i, A_j \in \mathcal{D}} d(v^f_{A_i}, v^f_{A_j}), \tag{7}$$

where d is a distance metric for a pair of vectors:

$$d(v^f_{A_i}, v^f_{A_j}) = \begin{cases} JSD(v^f_{A_i}, v^f_{A_j}) & \text{if} f = topic \\ 1 - \cos(v^f_{A_i}, v^f_{A_j}) & \text{otherwise.} \end{cases} \tag{8}$$

where JSD is the Jenson-Shannon Divergence (Nathanson, 2013) and cos is the cosine similarity. The lower the value, the harder it is to distinguish the authors in a dataset in the corresponding feature space, on average.

From Table 3, we observe that Blog50 has both the highest degree of topical similarity and largest improvement from contrastive learning, while TuringBench has the least topical similarity and also the least improvement. This suggests that contrastive learning may be more useful when authors in a dataset write about highly similar topics. On the other hand, the opposite is true for content similarity: TuringBench has the highest content similarity and yet the least improvement.

Inadequacy of NLG Models? We also note the high topical dissimilarity of TuringBench. This is unexpected since this corpus is generated by querying each NLG model with the same set of titles as prompts (Section 5.1). Following Sari et al. (2018), we model topical similarity using Latent Dirichlet Allocation (LDA; Blei et al., 2003). LDA represents a text as a distribution over latent topics, where each topic is represented as a distribution over words. This observation suggests that some NLG models may struggle to write on topic.[4]

[4] See Appendix D for a brief analysis.

Author-Level Analysis. Next, we analyze how author characteristics affect model performance on these authors. Specifically, we examine the correlation between the similarity of specific authors and how well the models distinguish between them. We define the distance between two authors to be the mean distance across all representation spaces:

$$PD(A_i, A_j) = \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} \frac{1}{C_f} d(v^f_{A_i}, v^f_{A_j}), \quad (9)$$

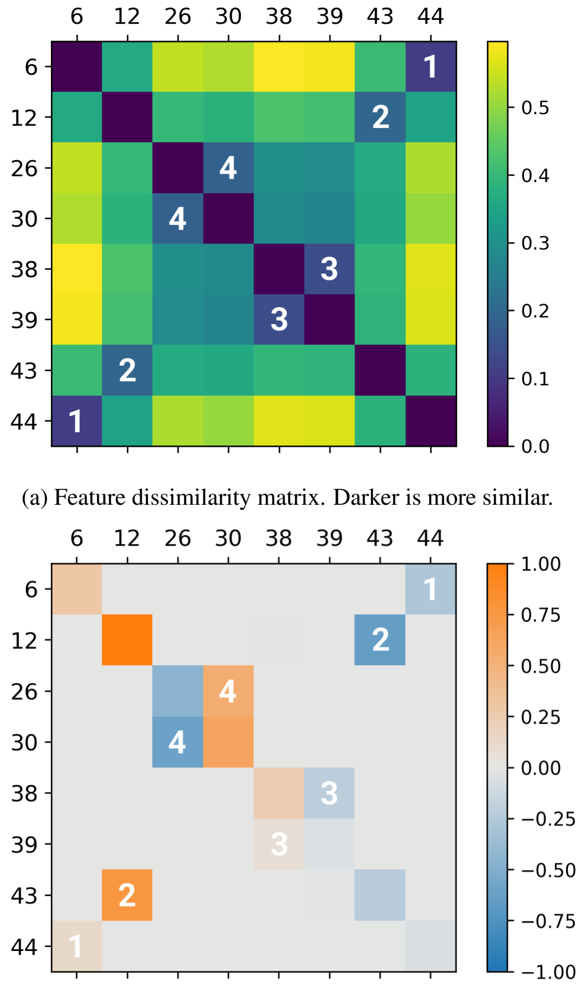
where C_f is a normalization term, defined as

$$C_f = \max_{A_i, A_j \in \mathcal{D}} d(v^f_{A_i}, v^f_{A_j}). \quad (10)$$

We plot the similarity matrix for selected Blog50 authors in Figure 3a. The authors are selected such that they form pairs that are highly indistinguishable by the above metrics. The cells numbered 1-4 represent the most similar author pairs (i.e., darker-coloured cells). Performance-wise, on each of these pairs, Contra-BERT shows significant improvements in overall class-level accuracy over BERT.[5] This is consistent with the intuition that contrastive learning is more useful for distinguishing author pairs that are more similar.

Increased bias. The pairwise improvement mentioned above shows a curious property of being biased towards one of the authors in the pair. To visualize this, we subtract the confusion matrix of BERT from that of Contra-BERT and name the result the *relative confusion matrix* (Figure 3b). Each cell in the matrix indicates the increase in the probability that an author A_i is classified as A_j from BERT to Contra-BERT. For example, the blue cell at $(12, 43)$ shows that Contra-BERT confused A_{12} as A_{43} less than BERT, while the orange cell at $(43, 12)$ shows that Contra-BERT confused A_{43} as A_{12} more frequently.

Note first the intuitive link between the similarity and confusion matrices: similar authors are more likely to be confused by one of the models for each other. Observe also that the pairs in the confusion matrix are always present in light-dark pairs. In other words, if BERT misclassifies more samples from A_i as A_j (e.g., A_{12} as A_{43}), then Contra-BERT mislabels more samples from A_j as A_i (i.e., A_{43} as A_{12}). This suggests that as Contra-BERT learns to classify samples from A_i better, it sacrifices the ability to identify A_j samples. Note that although this sometimes stems from training on an imbalanced dataset, in our case, A_i and A_j contain similar numbers of samples.[6] Thus, the observation is unlikely to be due to class imbalance.



(a) Feature dissimilarity matrix. Darker is more similar.

(b) Relative confusion matrix. This is obtained by subtracting the confusion matrix of BERT from that of Contra-BERT.

Figure 3: Feature similarity matrix and relative confusion matrix between BERT and Contra-BERT on selected authors. In both figures, (i, j) denotes the cell at the i-indexed row and j-indexed column. In (a), (i, j) denotes $d(A_i, A_j)$, the feature dissimilarity between the two authors. In (b), a lower value (blue) of (i, j) indicates Contra-BERT confused A_i for A_j less than BERT.

Nevertheless, the cumulative accuracy across A_i and A_j is always higher for Contra-BERT compared to the baseline, e.g., 33.6% vs 23.1% for A_{12} and A_{43} combined, leading to an overall performance improvement on the whole dataset. This shows that the model implicitly learns to make trade-offs to optimize the contrastive objective, i.e., it chooses to learn specialized representations that

[5] See Appendix E.1 for exact values. This trend also holds for Contra-DeBERTa and DeBERTa; see Appendix E.2.

[6] See Appendix E.1 for exact sample counts.

are particularly biased against some authors but improve the average performance over all authors. This shows that Contra-X captures certain features that enable the model to distinguish a subset of the authors. However, to obtain consistent improvement, we need a deeper understanding of the difference between easily-confused authors and incorporate that insight into the contrastive learning algorithm (Wolpert and Macready, 1997). This can be potentially achieved by constructing more meaningful negative samples. However, this is beyond the scope of our paper and left to future work.

6.4 Potential Ethical Concerns

In this subsection, we discuss potential ethical concerns related to the previous discussion on the increased bias in author-level performance.

Decreased Fairness? With classification models, fairness in predictions across classes is an important consideration. We want to, for instance, avoid demographic bias (Hardt et al., 2016), which may manifest as systematic misclassifications of authors with specific sociolinguistic backgrounds.

Having observed increased bias against certain authors, we seek to find out if this trend holds across the entire dataset. We quantitatively evaluate this by computing the variance in class-level accuracy across all authors. The results show that the improvements from our contrastive learning objective appear to incur a penalty in between-author fairness. Contra-BERT on Blog10 and Blog50, and Contra-DeBERTa on Blog50 achieve substantial gains in accuracy, and also produce notably higher variance than their baseline counterparts.[7] In contrast, for models where the improvements are marginal, the differences in variance are insignificant. A potential direction for future work is investigating whether the use of contrastive learning consistently exacerbates variances in class-level accuracy. Studying the characteristics of the classes that the model is biased against may boost not just overall performance, but also predictive fairness.

7 Conclusion

Successful authorship attribution necessitates the modeling of author-specific characteristics and idiosyncrasies. In this work, we make the first attempt to study the effect of combining contrastive learning with supervised learning on the authorship attribution task. We jointly optimized the contrastive and cross-entropy losses (Contra-X), demonstrating empirical improvements in authorship attribution on both human-written and machine-generated text. We also showed the generality of the method across data regimes via consistent improvement over conventional fine-tuning across various dataset sizes. Critically, we contributed analyses of how and when Contra-X works, in the context of the AA task. At the dataset level, we qualitatively showed that Contra-X creates a tighter representation spread of each author and increased separation between authors. Within each dataset, at the author level, we found that Contra-X is able to distinguish between highly similar author pairs, at the cost of hurting the performance on other authors. This points to a potential direction for future work, as resolving it would lead to better overall improvement and increased fairness of the final representation.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Maël Fabien, Esaú Villatoro-Tello, Petr Motlícek, and Shantipriya Parida. 2020. BertAA : BERT fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing, ICON 2020, Indian Institute of Technology Patna, Patna, India, December 18-21, 2020*, pages 127–137. NLP Association of India (NLPAI).

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on*

[7] See Appendix G for exact values.

Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, pages 6894–6910. Association for Computational Linguistics.

Tim Gollub, Martin Potthast, Anna Beyer, Matthias Busse, Francisco Rangel Pardo, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. 2013. Recent trends in digital text forensics and its evaluation. In *CLEF*, volume 8138, pages 282–302.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3315–3323.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Qiushi Huang, Tom Ko, H. Lilian Tang, Xubo Liu, and Bo Wu. 2021. Token-level supervised contrastive learning for punctuation restoration. *CoRR*, abs/2107.09099.

Farkhund Iqbal, Hamad Binsalleeh, Benjamin C. M. Fung, and Mourad Debbabi. 2010. Mining writeprints from anonymous e-mails for forensic investigation. *Digit. Investig.*, 7(1-2):56–64.

Fereshteh Jafariakinabad and Kien A. Hua. 2019. Style-aware neural model with application in authorship attribution. In *18th IEEE International Conference On Machine Learning And Applications, ICMLA 2019, Boca Raton, FL, USA, December 16-19, 2019*, pages 325–328. IEEE.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2020. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2296–2309, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kazuya Kawakami, Luyu Wang, Chris Dyer, Phil Blunsom, and Aäron van den Oord. 2020. Learning robust and multilingual speech representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1182–1192. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Thomas Corwin Mendenhall. 1887. The characteristic curves of composition. *Science*, (214s):237–246.

Michael Nathanson. 2013. Review: Elements of information theory. john wiley and sons, inc., hoboken, nj, 2006, xxiv + 748 pp., ISBN 0-471-24195-4, $111.00. by thomas m. cover and joy a. thomas. *Am. Math. Mon.*, 120(2):182–187.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Mostafa Rahgouy, Hamed Babaei Giglou, Taher Rahgooy, Mohammad Karami Sheykhlan, and Erfan Mohammadzadeh. 2019. Cross-domain authorship attribution: Author identification using a multi-aspect ensemble approach. In *CLEF (Working Notes)*.

Sebastian Ruder, Parsa Ghaffari, and John G. Breslin. 2016. Character-level and multi-channel convolutional neural networks for large-scale authorship attribution. *CoRR*, abs/1609.06686.

Upendra Sapkota, Steven Bethard, Manuel Montes, and Thamar Solorio. 2015a. Not all character n-grams are created equal: A study in authorship attribution. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–102, Denver, Colorado. Association for Computational Linguistics.

Upendra Sapkota, Steven Bethard, Manuel Montes-y-Gómez, and Thamar Solorio. 2015b. Not all character n-grams are created equal: A study in authorship attribution. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 93–102. The Association for Computational Linguistics.

Yunita Sari, Mark Stevenson, and Andreas Vlachos. 2018. Topic or style? exploring the most useful features for authorship attribution. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 343–353. Association for Computational Linguistics.

Yunita Sari, Andreas Vlachos, and Mark Stevenson. 2017. Continuous n-gram representations for authorship attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association* *for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 267–273. Association for Computational Linguistics.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. 2006. Effects of age and gender on blogging. In *Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium, Technical Report SS-06-03, Stanford, California, USA, March 27-29, 2006*, pages 199–205. AAAI.

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

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*, pages 669–674, Valencia, Spain. Association for Computational Linguistics.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. 2019. Release strategies and the social impacts of language models. *CoRR*, abs/1908.09203.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2021. CONFIT: toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. *CoRR*, abs/2112.08713.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI*, volume 12356 of *Lecture Notes in Computer Science*, pages 776–794. Springer.

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8384–8395. Association for Computational Linguistics.

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2001–2016, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

David H. Wolpert and William G. Macready. 1997. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.*, 1(1):67–82.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9051–9062.

Jianguo Zhang, Trung Bui, Seunghyun Yoon, Xiang Chen, Zhiwei Liu, Congying Xia, Quan Hung Tran, Walter Chang, and Philip S. Yu. 2021. Few-shot intent detection via contrastive pre-training and fine-tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1906–1912. Association for Computational Linguistics.

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, Brussels, Belgium, October 31 - November 4, 2018*, pages 2742–2753. Association for Computational Linguistics.

A Dataset Statistics

Table 4 presents statistics of the Blog10, Blog50, IMDb62, and Enron100 datasets.

B Human Authorship Attribution Results with 8:2 Split

Following Fabien et al. (2020), we divide the datasets into train-test splits at an 8:2 ratio for Blog10, Blog50, and IMDb62 and follow the default split for TuringBench. We show the results on the test set in Table 5.

C Similarity Metrics

Following Sari et al. (2018), we use four key metrics to analyze the characteristics of individual datasets (i.e., samples written by a particular author, or all samples in a corpus). We describe these metrics in detail below:

Content. We measure the frequencies of the most common word unigrams, bigrams, and trigrams to produce a feature vector that represents an author's content preferences over each document.

Style. We combine multiple stylometric features, i.e., average word length, number of short words, percentage of digits, percentage of upper-case letters, letter frequency, digit frequency, vocabulary richness, and frequencies of function words and punctuation, into a feature vector representing an author's writing style in a given document.

Hybrid. We measure the frequencies of the most common character bigrams and trigrams, to capture both content and style preferences of the author (Sapkota et al., 2015a) in a given document.

Topic. We use Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to generate a probability distribution over an author's possible topics. We run LDA with 20 topics, as in Sari et al. (2018), and fit the data over 500 iterations.

D TuringBench Dataset Analysis

Closer examination of the TuringBench dataset reveals that some models appear to produce fairly incoherent text. Table 6 contains snippets from various models. Qualitatively, it is difficult to identify what the topic of each text is supposed to be; there appear to be multiple topics referenced in each text. This suggests that some of these models do not write on-topic, and consequently may explain why LDA reflects a high degree of topical dissimilarity between models.

On the other hand, at the phrase level, these models largely put out sensible phrases, e.g., “strong economic growth”, “stunning game”, “suspicious clicks”. We hypothesize that this is why the content similarity on TuringBench is comparatively higher, since the content metric measures word *n*-gram frequencies.

E Analysis of Similar Author Pairs

E.1 BERT and Contra-BERT

Figure 4 shows the individual similarity matrices for the four feature types. The general pattern of the highlighted pairs being darker (i.e., more similar) than their surrounding cells can be seen across all the matrices. Table 8 shows the exact prediction accuracies for the four highlighted pairs. As noted previously, Contra-BERT always achieves a higher total accuracy (defined as total correct predictions over total samples) over both authors in a pair compared to BERT.

E.2 DeBERTa and Contra-DeBERTa

Figure 5 shows the feature similarity matrices and the relative confusion matrix for selected authors for DeBERTa and Contra-DeBERTa. Note that some of the author pairs are the same as those shown for BERT (i.e., 6 & 44, 38 & 39) while other pairs are different. Similar to Figure 3(b), the colour of a given cell (i, j), $i \neq j$, indicates whether Contra-DeBERTa confused A_i for A_j more or less often than DeBERTa. For instance, the blue-coloured $(1, 15)$ shows that Contra-DeBERTa confused A_1 as A_{15} less than DeBERTa, while the orange $(15, 1)$ shows that Contra-DeBERTa confused A_{15} as A_1 more times.

Table 9 shows the exact prediction accuracies for the highlighted pairs. As with Contra-BERT, Contra-DeBERTa achieves a higher total accuracy on each pair than DeBERTa.

F Full TuringBench results

Table 7 shows the precision, recall, F1, and accuracy scores on TuringBench.

G Class-Level Accuracy Variance

Table 10 shows the exact class-level accuracy variances for our four models on Blog10, Blog50, and TuringBench.

	Blog10	Blog50	IMDb62	TuringBench
# authors	10	50	62	20
# total documents	23498	73275	61973	149561
avg char / doc (no whitespace)	407	439	1401	1063
avg words / doc	118	124	341	188

Table 4: Statistics of the four datasets used in our experiments.

Model	Blog10	Blog50	IMDb62
Token SVM (Seroussi et al., 2014)	-	-	92.5
Char-CNN (Ruder et al., 2016)	61.2	49.4	91.7
Continuous N-gram (Sari et al., 2017)	61.3	52.8	95.1
N-gram CNN (Shrestha et al., 2017)	63.7	53.1	95.2
Syntax CNN (Zhang et al., 2018)	64.1	56.7	**96.2**
BertAA (Fabien et al., 2020)	**65.4**	**59.7**	93.0
BERT	60.3	55.6	97.2
Contra-BERT	66.0 (5.7↑)	62.2(6.6↑)	97.7(0.5↑)
DeBERTa	68.0	65.0	98.1
Contra-DeBERTa	**69.9(1.9↑)**	**69.7(4.7↑)**	**98.2(0.1↑)**

Table 5: Results of human authorship attribution - 8:2 train/test split

Model	**Text**
CTRL	“apple gives tim cook $384 million stock grant... steve jobs is set to receive an additional $1.4 billion in cash... recovery needs but it also requires p le with skills not just on paper or through education training but, crucially, real work experience. those are two things which can only come if we have strong economic growth...”
FAIR_WMT19	“antoine helps real sociedad draw with valladolid... sociedad’s goal in a 1-1 was highlight of stunning game played on night terrorist bombing attack manchester. tuesday, two bombs exploded central manchester arena during popular outdoor concert, killing 22 p le and injuring hundreds more..."
GROVER_MEGA	“...the messages, which along message some will choose avoid draft, ready for qualification training are fake, according public affairs. do not respond spoof, requires suspicious clicks, pictures, or notes function, an official memo from issued thursday reads...”
TRANSFORMER_XL	“carlos ghosn, mum on tokyo escape, unleashes a rambling defense of the state student-teacher training program in japan... as 2015, three universities (hiroshima, izumo, kawachi) accept all two degrees; they have also accepted each other. nevertheless, buddhist monks maintain that their colleges provide admission hindu traditions rather than admitting any religious instruction.”

Table 6: Sample text snippets from various NLG models in the TuringBench dataset.

Model	**Precision**	**Recall**	**F1**	**Accuracy**
Random Forest	58.93	60.53	58.47	61.47
SVM (3-grams)	71.24	72.23	71.49	72.99
WriteprintsRFC	45.78	48.51	46.51	49.43
OpenAI detector[8]	78.10	78.12	77.14	78.73
Syntax CNN	65.20	65.44	64.80	66.13
N-gram CNN	69.09	68.32	66.65	69.14
N-gram LSTM-LSTM	6.694	68.24	66.46	68.98
BertAA	77.96	77.50	77.58	78.12
BERT	80.31	80.21	79.96	80.78
RoBERTa	82.14	81.26	81.07	81.73
BERT *(our baseline)*	78.56	78.81	78.53	79.46
Contra-BERT	80.10 (1.66↑)	79.99 (1.88↑)	79.84 (1.31↑)	80.59 (1.13↑)
DeBERTa *(our baseline)*	82.16	81.84	81.82	82.00
Contra-DeBERTa	**82.84 (0.68↑)**	**82.04 (0.20↑)**	**81.98 (0.17↑)**	**82.53 (0.53↑)**

Table 7: Full results across four metrics on human and machine authorship attribution. Results in top section are from the TuringBench paper. Improvements over the baselines are indicated in parentheses. Best model is **bolded**.

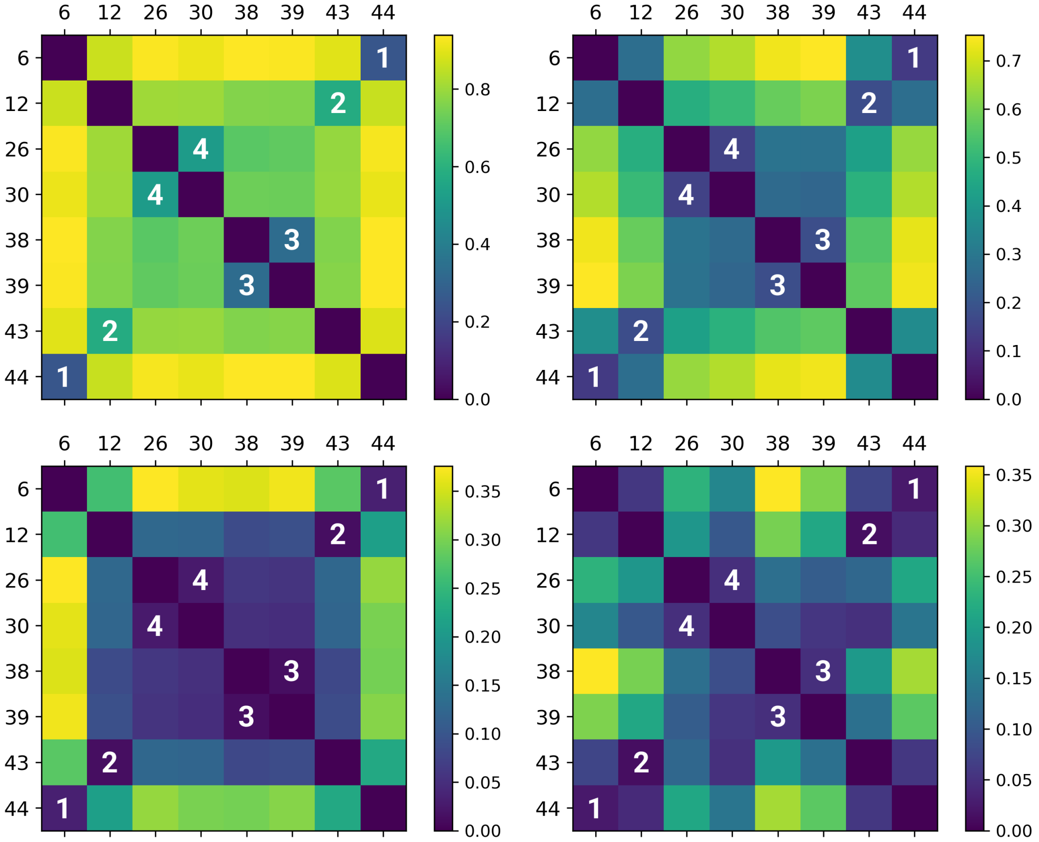


Figure 4: (Clockwise from top left) Similarity metrics between authors A_i (i-indexed row) and A_j (j-indexed column) for content, topic, hybrid, and style features respectively for selected authors on Blog50.

	Author 1			**Author 2**			Total
Model	#	Samples	Correct	#	Samples	Correct	Accuracy (%)
BERT	12	229	2	43	225	47	10.8
Contra-BERT			209			0	**46.0**
BERT	30	153	8	26	154	92	32.6
Contra-BERT			135			0	**44.0**
BERT	6	116	35	44	113	18	23.1
Contra-BERT			73			4	**33.6**
BERT	38	112	48	39	112	8	25.0
Contra-BERT			96			0	**42.9**

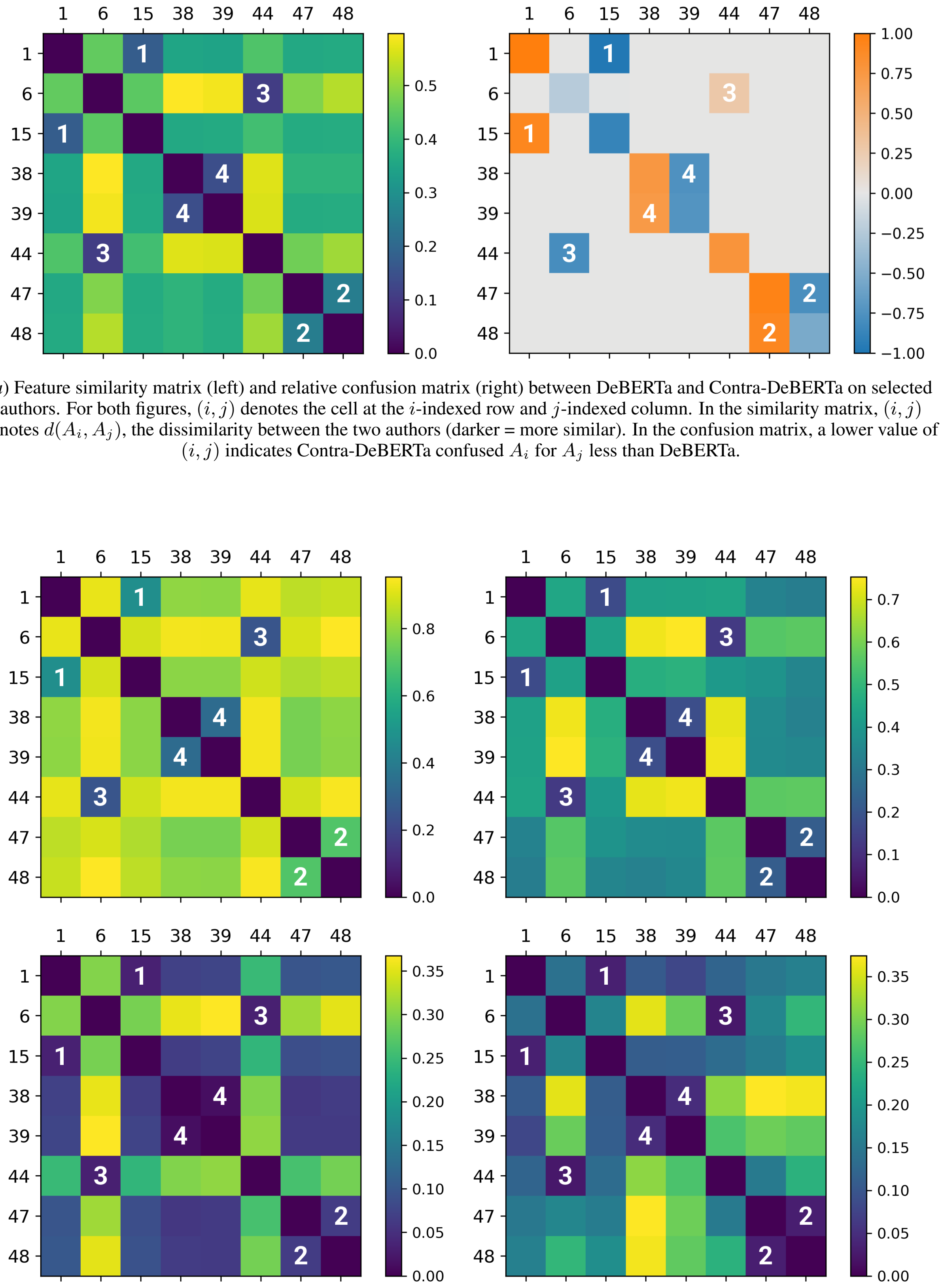
Table 8: Performance of BERT and Contra-BERT on selected author pairs of Blog50. Higher accuracy for each pair is **bolded**.

	Author 1			**Author 2**			Total
Model	#	Samples	Correct	#	Samples	Correct	Accuracy (%)
DeBERTa	1	109	0	15	103	94	44.3
Contra-DeBERTa			107			0	**50.5**
DeBERTa	47	105	0	48	104	61	29.2
Contra-DeBERTa			102			4	**50.7**
DeBERTa	44	113	24	6	116	28	22.7
Contra-DeBERTa			108			3	**48.5**
DeBERTa	38	112	0	39	112	90	40.2
Contra-DeBERTa			81			12	**41.5**

Table 9: Performance of DeBERTa and Contra-DeBERTa on selected author pairs of Blog50. Higher accuracy for each pair is **bolded**.

	Blog10	**Blog50**	**TuringBench**
BERT	0.15494	0.10430	0.06747
Contra-BERT	**0.17698** (Acc. +5.9)	**0.12087** (Acc. +6.8)	**0.06772** (Acc. +1.13)
DeBERTa	0.19735	0.13267	**0.05191**
Contra-DeBERTa	**0.20029** (Acc. +0.6)	**0.14343** (Acc. +3.7)	0.05126 (Acc. +0.53)

Table 10: Variance in class-level accuracy (accuracy increase by each contrastive model is listed for reference). The higher the variance, the more the model performance varies between different classes. For each dataset, higher variance for each baseline/contrastive pair is **bolded**.



(*a*) Feature similarity matrix (left) and relative confusion matrix (right) between DeBERTa and Contra-DeBERTa on selected authors. For both figures, (i, j) denotes the cell at the i-indexed row and j-indexed column. In the similarity matrix, (i, j) denotes $d(A_i, A_j)$, the dissimilarity between the two authors (darker = more similar). In the confusion matrix, a lower value of (i, j) indicates Contra-DeBERTa confused A_i for A_j less than DeBERTa.

(*b*) (Clockwise from top left) Similarity metrics between authors A_i (i-indexed row) and A_j (j-indexed column) for content, topic, hybrid, and style features respectively for selected authors on Blog50.

Figure 5: Visualizations for selected author pairs for DeBERTa and Contra-DeBERTa on Blog50.