Can Authorship Representation Learning Capture Stylistic Features?

Andrew Wang[1*], **Cristina Aggazzotti**[1*], **Rebecca Kotula**[2], **Rafael Rivera Soto**[3], **Marcus Bishop**[2], **Nicholas Andrews**[1†]
[1] Johns Hopkins University, USA [2]Department of Defense, USA
[3]Lawrence Livermore National Laboratory, USA

Abstract

Automatically disentangling an author's style from the content of their writing is a longstanding and possibly insurmountable problem in computational linguistics. At the same time, the availability of large text corpora furnished with author labels has recently enabled learning authorship representations in a purely data-driven manner for authorship attribution, a task that ostensibly depends to a greater extent on encoding writing style than encoding content. However, success on this surrogate task does not ensure that such representations capture writing style since authorship could also be correlated with other latent variables, such as topic. In an effort to better understand the nature of the information these representations convey, and specifically to validate the hypothesis that they chiefly encode writing style, we systematically probe these representations through a series of targeted experiments. The results of these experiments suggest that representations learned for the surrogate authorship prediction task are indeed sensitive to writing style. As a consequence, authorship representations may be expected to be robust to certain kinds of data shift, such as topic drift over time. Additionally, our findings may open the door to downstream applications that require stylistic representations, such as style transfer.

1 Introduction

Knowing something about an author's writing style is helpful in many applications, such as predicting who the author is, determining which passages of a document the author composed, rephrasing text in the style of another author, and generating new text in the style of a particular author. The trouble is that fully characterizing something as complex as writing style has proven too unwieldy to admit fine-grained human annotations, which leaves the possibility of directly learning explicit and interpretable representations of writing style practically beyond reach. Instead, research in this area has largely focused on specific stylistic attributes, such as formality, toxicity, politeness, gender, simplicity, and humor, which are more straightforward to annotate (Rao and Tetreault, 2018; Rao and Tetreault, 2020; Madaan et al., 2020; Li et al., 2018; Jin et al., 2022). Unfortunately, the reliance on human labels and the narrow focus of such stylistic distinctions severely limit the utility of such representations in tasks related to authorship, such as those listed above.

In this paper, we focus instead on the *authorship prediction task*, which enjoys the benefit of not requiring manually elicited labels, since metadata in many corpora include either explicit author labels or usernames that may serve as proxies for latent authorship. As a result, the vast scale of data available for training authorship prediction models opens the door to learning *generalizable* authorship representations using deep learning. We specifically consider similarity learning approaches that aim to produce vector representations of documents, where the distance between two vectors is inversely related to the likelihood that the corresponding documents were composed by the same author (Boenninghoff et al., 2019; Andrews and Bishop, 2019).

However, achieving high accuracy in the authorship prediction task does not necessarily imply that stylistic features have been successfully learned. For example, in a given corpus, correctly predicting that two writing samples were composed by the same author may be possible on the basis of non-stylistic signal, such as the topic of conversation. Therefore, this work is concerned with obtaining a better understanding of the nature of representations learned for the authorship prediction task.

Unfortunately, because deep learning models behave like black boxes, we cannot directly

*These authors contributed equally to this work.
†Corresponding author: noa@cs.jhu.edu.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 1416–1431, 2023. https://doi.org/10.1162/tacl_a_00610
Action Editor: David Bamman. Submission batch: 3/2023; Revision batch: 6/2023; Published 11/2023.

interrogate a model's parameters to determine what information such representations contain. For example, one might hope to employ attention-based approaches that provide post hoc explanations through token saliency maps (Sundarajan et al., 2017). However, such methods provide no guarantee of the fidelity of their explanations to the underlying model. Furthermore, the subjective interpretation required to deduce the reasons that such methods highlight certain spans of text makes it nearly impossible to systematically draw conclusions about the model.

Instead, we propose targeted interventions to probe representations learned for the surrogate authorship prediction task. First, we explore masking content words at training time in §5, an operation intended to gauge the degree to which a representation relies on content. Then we explore automatic paraphrasing in §6, an operation intended to preserve meaning while modifying how statements are expressed. Finally, in §7 we explore the capacity of these representations to generalize to unseen tasks, specifically topic classification and coarse style prediction.

Taken together, and despite approaching the research question from various points of view, our experiments suggest that representations derived from the authorship prediction task are indeed substantially stylistic in nature. In other words, success at authorship prediction may in large part be explained by having successfully learned discriminative features of writing style. The broader implications of our findings are discussed in §8.

2 Related Work

Perhaps the work most closely related to our study is that of Wegmann and Nguyen (2021) and Wegmann et al. (2022), who propose measuring the stylistic content of authorship representations through four specific assessments, namely, formality, simplicity, contraction usage, and number substitution preference. Our work differs in two main respects. First, we regard style as an abstract constituent of black-box authorship representations rather than the aggregate of a number of specific stylistic assessments. Second, the works above deal with stylistic properties of *individual sentences*, whereas we use representations that encode longer spans of text. Indeed, we maintain that the writing style of an author manifests itself only after observing a sufficient amount of text composed by that author. For example, it would be difficult to infer an author's number substitution preferences after observing a single sentence, which is unlikely to contain multiple numbers. The same is true of other stylometric features, such as capitalization and punctuation choices, abbreviation usage, and characteristic misspellings.

In another related work, Sari et al. (2018) find that although content-based features may be suitable for datasets with high topical variance, datasets with lower topical variance benefit most from style-based features. Like the works mentioned above, Sari et al. explicitly identify a number of style-based features, so writing style is more of a premise than the object of study. In addition, experiments in this previous work are limited to datasets featuring a small number of authors, with the largest dataset considered containing contributions of only 62 authors.

A number of end-to-end methods have been proposed to learn representations of authorship (Andrews and Bishop, 2019; Boenninghoff et al., 2019; Saedi and Dras, 2021; Hay et al., 2020; Huertas-Tato et al., 2022). A common thread among these approaches is their use of *contrastive learning*, although they vary in the particular objectives used. They also differ in the the domains used in their experiments, the numbers of authors considered for training and evaluation, and their open- or closed-world perspectives. As discussed in §3, we use the construction introduced in Rivera-Soto et al. (2021) as a representative neural method because it has shown evidence of capturing stylistic features through both its success in the challenging open-world setting in multiple domains and its performance in zero-shot domain transfer.

3 Authorship Representations

In this paper, an *authorship representation* is a function mapping documents to a fixed Euclidean space. The fact that such representations are useful for a number of authorship-related tasks is generally attributed to their supposed ability to encode author-specific style, an assertion we aim to validate in this paper. In this section, we describe how these representations arise and how they are intended to be used.

Our analysis centers around representations f implemented as deep neural networks and trained

using a *supervised contrastive objective* (Khosla et al., 2020). At training time this entails sampling pairs of documents x, x' composed by the same author (resp., by *different* authors) and minimizing (resp., *maximizing*) the distance between $f(x)$ and $f(x')$. Therefore, we may assume at inference time that $f(x), f(x')$ are closer together if x, x' were composed by the same author than they would be if x, x' were composed by different authors. No meaning is ascribed to any attribute of $f(x)$, such as its coordinates, its length, or its direction. Rather, $f(x)$ is meaningful only in relation to other vectors.

In all the experiments of this paper, f is an instance of the *Universal Authorship Representation* (UAR) introduced in Rivera-Soto et al. (2021).[1] Notwithstanding the merits of a number of other recent approaches discussed in §2, we argue that because of its typical neural structure and typical contrastive training objective, UAR serves as a representative model. The same paper also illustrates that UAR may be used for zero-shot transfer between disparate domains, suggesting a capacity to learn generalizable features, perhaps of a stylistic nature, thereby making it a good candidate for our experiments.

Note that UAR defines a *recipe* consisting of an architecture and a training process that must be carried out in order to arrive at a representation, with the understanding that care must be taken in assembling appropriate training datasets. Specifically, we consider a diverse set of authors at training time in an effort to promote representations that capture *invariant* authorship features, chiefly writing style, rather than time-varying features, such as topic. Invariance is a desirable feature of authorship representations because it improves the likelihood of *generalization* to novel authors or even to novel domains. However, there is no guarantee that invariant features are exclusively stylistic in any given corpus, or that any training process we might propose will result in representations capturing exclusively invariant features. Therefore, this work is concerned with estimating the *degree* to which authorship representations are capable of capturing stylistic features, with the understanding that completely disentangling style from topic may be beyond reach.

[1] An open-source implementation is available at `https://github.com/LLNL/LUAR`.

4 Experimental Setup

Mirroring Rivera-Soto et al. (2021), we conduct experiments involving three datasets, each consisting of documents drawn from a different domain. For the reader's convenience, we present further details and some summary statistics of these datasets in §A.1.

To evaluate an authorship representation, we use the common experimental protocol described below. The objective is to use the representation to retrieve documents by a given author from among a set of candidate documents, which are known as the *targets*, on the basis of the distances between their representations and the representation of a document by the desired author, which is known as the *query*. To this end, each evaluation corpus has been organized into queries and targets, which are used to calculate the *mean reciprocal rank (MRR)*. Following is a friendly description of this metric, with a more elaborate formulation presented in §A.2.

An authorship representation may be used to sort the targets according to the distances between their representations and that of any fixed query. In fact, this ranking is often seen as the primary outcome of an authorship representation. Because one would need to manually inspect the targets in the order specified by the ranking, it would be desirable for any target composed by the same author as the query to appear towards the beginning of this list. The MRR is the expectation of the reciprocal of the position in the ranked list of the first target composed by the same author as a randomly chosen query. This metric ranges from 0 to 1, with higher values indicating a greater likelihood of finding documents composed by an author of interest within a large collection in the first few search results.

Following Rivera-Soto et al., the queries and targets in all our experiments are *episodes*, each consisting of 16 comments or product reviews contiguously published by the same author in the Reddit or Amazon domains, respectively, or 16 contiguous paragraphs of the same story in the fanfic domain.

In order to conduct a wide variety of experiments in a time-efficient manner, we train all representations on one GPU for 20 epochs, although we acknowledge that better results may be obtained by training with more data, on multiple GPUs, or for longer than 20 epochs.

5 Masking Content Words

Our first series of experiments aims to illustrate through a simple training modification that authorship representations are capable of capturing style. Specifically, the strategy of *masking* training data in a way that preserves syntactic structure, something which is known to relate to style, while removing thematic or topical information, has been effective, particularly in cross-domain authorship experiments (Stamatatos, 2018). To this end, we propose training authorship representations with restricted access to topic signal by masking varying proportions of content-related words in the training data. Evaluating each of these representations and comparing its ranking performance with that of a representation trained on the same *unmasked* data reveals the capacity of the representation to capture style.

Words may be roughly divided into two categories: *content words* and *function words*. Content words primarily carry topic signal. They tend to include nouns, main verbs, adjectives, and adverbs. Function words serve syntactic roles and convey style through their patterns of usage. They tend to include auxiliary verbs, prepositions, determiners, pronouns, and conjunctions (Mosteller and Wallace, 1964). These observations suggest masking words according to their parts of speech (POS), a process we call *Perturbing Linguistic Expressions* or *PertLE*.

5.1 The PertLE schema

In our PertLE masking schema, we replace all words belonging to certain POS categories with a distinguished masking token. This approach stems from the observation that content words may often be distinguished from function words on the basis of POS. However, this is simply a heuristic and there are many exceptions. For instance, although many adverbs may be categorized as content words, such as *happily*, others play a functional role, such as *instead*. Because masking on the basis of POS is an imperfect strategy to eliminate content, we introduce the following *levels* of the PertLE schema. In our *PertLE Grande* schema we mask all nouns, main verbs, adjectives, and adverbs. This is a greedy approach intended to mask words that could possibly convey content, at the expense of occasionally masking some function words. In contrast, in our *PertLE Lite* schema we mask only nouns, which are most likely to carry content information.[2] In a follow-up reported in §A.3 we repeat the main experiment below using a masking schema based on TF-IDF scores rather than POS.

Procedure To identify POS categories, we use the Stanford NLP Group's Stanza tokenizer and POS tagger (Qi et al., 2020) due to their efficiency, flexibility, versatility, and capacity for handling other languages. We use the Universal POS (UPOS) tagset because it distinguishes between main verbs (VERB) and auxiliary verbs (AUX), labels *not* and *-n't* as particles rather than adverbs, and tags many foreign language words with their correct POS category rather than labeling them as foreign words. For both masking levels, we replace each word to be masked with SBERT's masking token, `<mask>`, preserving any contracted particles (e.g., `gonna` ⇝ `<mask>na`). As an example, Figure 1 illustrates both levels of the PertLE schema applied to the same statements.

Using the procedure described in Rivera-Soto et al. (2021), for each domain we train multiple authorship representations on that domain's training corpus: one with the training corpus masked according to PertLE Grande, one masked according to PertLE Lite, and one unmasked to serve as a baseline. We evaluate each representation on each *unmasked* evaluation corpus to afford a fair comparison of the effects of the masking level for each combination of training and evaluation domain.

Note that for representations trained on *masked* data, this evaluation introduces a *mismatch* between training and evaluation datasets, although the baseline representations remain unaffected. In cases where masking results in a large *degradation* in performance, this setup makes it impossible to distinguish between our interventions and the train-test mismatch as the cause of the degradation. On the other hand, this distinction is immaterial in the case that masking does *not* degrade performance, and in fact, this case is the desired outcome of the experiment, as it would suggest that

[2] We also tried masking every word belonging to certain POS categories with a distinguished *pseudoword* specific to its POS. These pseudowords were selected to be morphologically similar to other words in their POS categories but not appear in our corpora. However, we adopt the simpler masking approach described above because it surprisingly produced very similar ranking results.

Unmasked:

- Hold me closer, tiny dancer. Count the headlights on the highway. Lay me down in sheets of linen. You had a busy day today.
- Just a small-town girl, livin' in a lonely world. She took the midnight train going anywhere.
- All I wanna do is have a little fun before I die, says the man next to me out of nowhere.

PertLE Grande:

- <mask> me <mask>, <mask> <mask>. <mask> the <mask> on the <mask>. <mask> me <mask> in <mask> of <mask>. You <mask> a <mask> <mask> <mask>.
- <mask> a <mask> <mask>, <mask> in a <mask> <mask>. She <mask> the <mask> <mask> <mask> <mask>.
- All I <mask>na <mask> <mask> <mask> a <mask> <mask> before I <mask>, <mask> the <mask> <mask> to me out of <mask>.

PertLE Lite:

- Hold me closer, tiny <mask>. Count the <mask> on the <mask>. Lay me down in <mask> of <mask>. You had a busy <mask> <mask>.
- Just a small-town <mask>, livin' in a lonely <mask>. She took the <mask> <mask> going anywhere.
- All I wanna do is have a little <mask> before I die, says the <mask> next to me out of nowhere.

Figure 1: Various levels of the PertLE masking schema applied to the same statements.

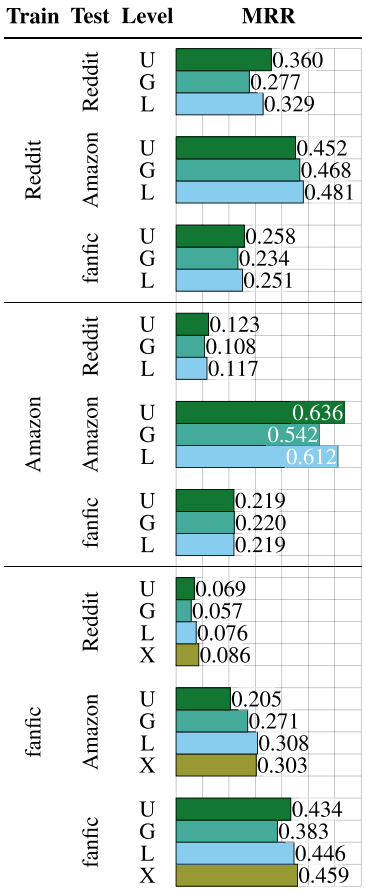

Figure 2: MRR results for models trained on unmasked data (U), or data masked according to the PertLE Grande (G), the PertLE Lite (L), or, additionally for fanfic, the PertLE Xtra-Lite (X) schema.

the corresponding representation does not benefit significantly from the information withheld by masking content words.

The results of the experiment are shown in Figure 2. For each corpus and each masking level, we independently trained *three* representations in an effort to reduce variance. Each number reported is the sample mean of the MRR according to each of the three independent representations, where 0.014 is the maximum sample standard deviation over all experiments reported in the figure.

Discussion A one-way ANOVA was performed for each combination of training and evaluation domain, showing that the masking schema had a statistically significant impact on the mean values of the MRR reported in Figure 2 with $p < 0.01$ in all cases except the case with training domain Amazon and evaluation domain fanfic. In that case, we conclude that masking words at training time had no significant effect on ranking performance, as desired. In the other cases the change in performance *was* significant but relatively minor.

In cases where performance *improved*, we believe the most likely explanation is that, deprived of content words at training time, the model was forced to discover other authorship features, which turned out to be more useful than content words in the corresponding evaluation domains. In cases where performance *dropped*, it appears that the model was unable to compensate for the loss of content words. However, we emphasize that in these cases the drop in performance is surprisingly small in light of the fact that PertLE Grande masks nearly *half* of all training tokens. See Table 4 in §A.3 for the exact proportions of tokens masked in each domain or Figure 1 for a qualitative example. Another possibility is that, as discussed above, PertLE Grande obscures writing style to some extent, which could also account for the small drop in ranking performance if the representations were primarily style-focused.

For all three training and evaluation domains, the MRR of the Lite model is quite close to that

of the unmasked model. This suggests that masking words most likely to convey content changes ranking performance very little, and even improves it in some settings. We also observe that although the MRR of each Grande model is generally less than that of the corresponding Lite model, it is not dramatically so. This suggests that increasing the proportion of tokens masked appears to eventually impair ranking performance, but not to the degree one might expect given the considerable proportion of words masked.

We know of no way to *completely* redact the content of a document while retaining its writing style. We doubt that this is even possible, least of all in an automated fashion. It follows that the representations trained on data masked according to the PertLE schema (as well as the TertLE schema discussed §A.3) probably *do* encode a small amount of content. Being trained to distinguish authors on the basis of such masked text, these models are therefore likely to learn to use that information to their advantage when appropriate, which would mean that the representations considered in this paper *do* convey a small amount of topical signal, an observation which is corroborated by the experiments in §6 and §7.

Nevertheless, the experiment shows that PertLE obscures *much* of the content of a training corpus, which in turn affects ranking performance only marginally. We argue that those representations are therefore likely to have learned to avail of features other than content, thereby illustrating their *capacity* to avail of writing style.

5.2 PertLE Xtra-Lite

As observed in Rivera-Soto et al. (2021), the representation trained on the fanfic corpus generalizes poorly to the other two domains, something which is probably due to the comparatively small size and lack of topical diversity of that dataset. This suggests that representations trained on fanfic stand to improve the most by a targeted inductive bias. Indeed, the Lite model trained on the fanfic dataset improves performance in all three evaluation domains. This may be explained by the observation that the fanfic domain may contain more jargon and specialized language appearing in the form of proper nouns representing names, places, and things. This is borne out by the observation that in the Reddit, Amazon, and fanfic domains, around 22%, 20%, and 35%, respectively, of all nouns are proper.

To further explore this observation we introduce the *Xtra-Lite* level of the PertLE schema, in which we mask only proper nouns, the POS category most likely to convey content information. Repeating the same procedure as before, we train a PertLE Xtra-Lite (X) representation on the fanfic domain and evaluate it on each unmasked evaluation dataset. The results in Figure 2 show that the Xtra-Lite model not only outperforms the unmasked and Grande models in all three domains, but also outperforms the Lite model in the Reddit and fanfic domains and performs nearly as well in the Amazon domain, confirming that representations trained on fanfic benefit from a targeted inductive bias.

6 Removing Style by Paraphrasing

In contrast with the experiments in §5 that aim to assess the ability of authorship representations to capture style by removing *content*, our next group of experiments aims to make the same assessment by instead removing *style*. For this purpose we consider automatic paraphrasing, which ideally introduces stylistic changes to a document while largely retaining its original content. If an authorship representation avails of stylistic features, then we expect paraphrasing a document to *impair* its ability to match the document with other documents by the same author.

6.1 Implementation Details

To generate paraphrases, we use the STRAP paraphraser developed by Krishna et al. (2020), which consists of a fine-tuned GPT-2 language model trained on ParaNMT-50M (Wieting and Gimpel, 2018), a large dataset consisting of paraphrases of English sentences.

Because automatic paraphrasing models provide no guarantee that the proposed paraphrases of a document retain its meaning, we need to check this explicitly. For this purpose we adopt the BERTScore (Zhang et al., 2019) as our primary similarity metric, rescaled to the unit interval. Unlike n-gram-matching metrics like Bleu, BERTScore leverages contextual embeddings to make predictions about semantic similarity.

Because STRAP acts on *sentences* rather than *episodes*, we apply it independently to each sentence comprising an episode, with the following

Domain	Model	Orig	Para	Δ
fanfic	**UAR**	0.325	0.139	**0.186**
	SBERT	0.203	0.167	0.036
Reddit	**UAR**	0.263	0.026	**0.237**
	SBERT	0.043	0.026	0.017
Amazon	**UAR**	0.266	0.025	**0.241**
	SBERT	0.165	0.069	0.096

Table 1: The impact on ranking performance of paraphrasing queries. Paraphrasing drastically impairs ranking performance of the UAR model relative to the baseline SBERT model, suggesting a reliance on stylistic features.

caveat. Preliminary experiments revealed that the degree to which automatic paraphrasing retained meaning varied widely, an issue that we mitigate as follows. For each of the sixteen constituent documents x of an episode, we paraphrase the sentences within x to obtain x' and calculate the mean BERTSCORE of each sentence of x with its paraphrase in x'. We discard the eight x' with lowest mean BERTSCORE to form the paraphrased episode, and also drop the eight corresponding x from the original episode for comparability.

6.2 Impact of Paraphrasing on Ranking

For each domain, we train an authorship representation in the usual way, perform the primary ranking experiment described in §4, and repeat the experiment after paraphrasing all the queries in the manner described in §6.1. In Table 1 we report the MRR for the original experiment (Orig), the MRR for the paraphrased variation (Para), and the change in performance (Δ) for each domain. To serve as a baseline, we repeat the entire experiment with the SBERT model in place of the trained authorship representation, which is denoted by UAR in Table 1.

For each domain and each model, the MRR substantially decreased for the paraphrased queries relative to the original queries, which confirms that both models rely to some extent on author style. However, the performance degradation was much more pronounced for UAR than for SBERT, suggesting that UAR captures style to a much greater extent than SBERT.

For each domain and each model, a paired t test shows that the decrease in MRR of the paraphrased queries relative to that of the original queries is significant with $p < 0.01$. Additionally, for each domain, a further t test shows that the difference between the two models of the differences in MRR between the original and paraphrased queries is significant with $p < 0.01$.

We also present the results of this experiment in a more qualitative way in Figure 3. Recall from §A.2 that for each query q_i and its corresponding target t_i, our primary ranking experiment entails ranking all the targets $t_1, t_2, \ldots, t_N$ according to their similarity to q_i and reporting the position r_i of t_i in the ranked list. In Figure 3 we plot r_i against r'_i for each $1 \leq i \leq M$, where r'_i is the position of t_i in the list ranked according to similarity to q'_i, the paraphrase of q_i.

Examples for which $r_i \approx r'_i$ correspond to points near the diagonal line shown, whereas examples for which $r_i > r'_i$ correspond to points above this line. Note that for the UAR model, most points lie above the diagonal line, while for SBERT, the points are more evenly distributed across both sides of the line. This again suggests that paraphrasing impairs the ranking performance of UAR much more dramatically than that of SBERT.

6.3 Quality of Paraphrases

If the premise that our paraphrases retain meaning but alter style were not satisfied, then we would not be able to infer that paraphrasing the queries in §6.2 is responsible for the drop in ranking performance. To confirm that the premise is largely satisfied, we propose the following metrics, both averaged over all the sentences comprising a query. To assess the degree of content preservation between a query q and its paraphrase q', we calculate their BERTSCORE. To confirm that q' significantly modifies the style of q, rather than making only minor changes, we calculate their *normalized edit distance*. While neither metric is perfect, together they provide a reasonable estimate of paraphrase quality.

As a baseline, we calculate the same metrics for the Microsoft Research Paraphrase Corpus (MRPC) (Dolan and Brockett, 2005). Figures 4 and 5 show that the distributions of both scores overlap substantially with those for the MRPC corpus restricted to sentence pairs deemed to constitute paraphrases by human annotators, which is labeled by MRPC+. As a further baseline, Figure 4 also shows the distribution of MRPC scores restricted to pairs deemed *not* to constitute

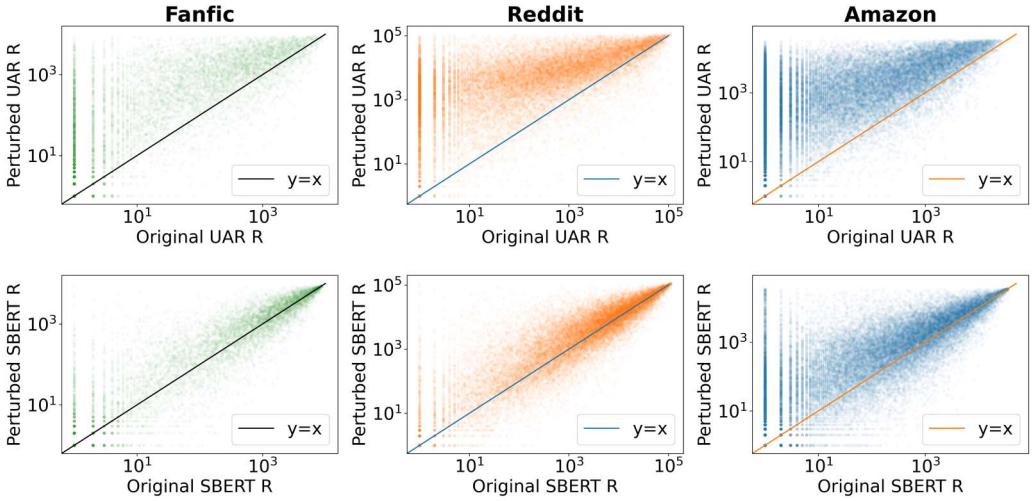

Figure 3: Rankings r_i against r'_i. UAR has more points above the diagonal line $r = r'$ than SBERT, which correspond with queries for which paraphrasing hurts ranking performance.

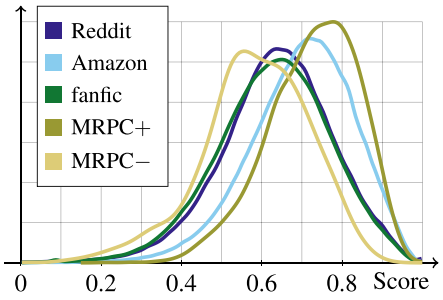

Figure 4: Distribution of BERTScores comparing documents to their paraphrases.

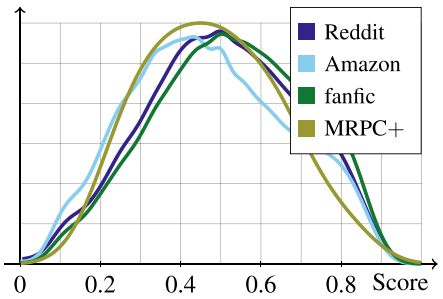

Figure 5: Distribution of edit distances between documents and their paraphrases.

paraphrases, which is labeled by MRPC−. Therefore we may rule out the possibility that the drop in ranking performance in §6.2 might be due to low-quality paraphrases.

6.4 Impact of Content Overlap

As a final illustration, in Figure 6 we plot the BERTScore b_i of q_i with q'_i against the change in rank $\Delta r_i = r'_i - r_i$ for all $1 \leq i \leq M$. If authorship representations were significantly influenced by content, then we might expect to see a strong negative relationship between b_i and Δr_i. Instead, we observe little correlation, with Kendall's τ values of -0.092, -0.019, and -0.015 for the fanfic, Reddit, and Amazon domains, respectively, suggesting that the ranking performance degradation in §6.2 cannot be well-explained by content overlap between q_i and q'_i.

6.5 A Further Application

Although beyond the scope of this paper, we remark that a broader research problem is to determine whether the capacity of an authorship representation to encode style is *correlated* with its performance on the authorship attribution task. For example, if a new representation were introduced that performed better than UAR on attribution, would it necessarily encode style to a greater degree than UAR? Conversely, if a new approach were proposed to learn representations that encode style to a greater degree than UAR, would such representations perform better on attribution?

Addressing these questions will require assessing the degree to which a representation encodes style. We submit that the experiments presented in this paper are well-suited to making such assessments. As an illustration, we repeat the primary experiment described in §6.2 using two further instances of the UAR architecture introduced in §3, but trained on the Reddit histories of around 100K and 5M authors, respectively, in contrast with the

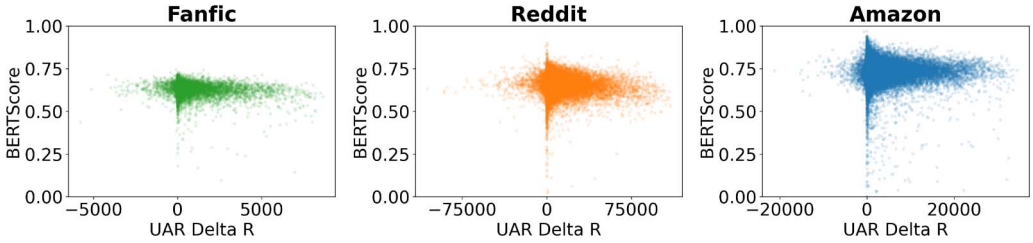

Figure 6: BERTScore against change in rank ΔR. BERTScore is minimally correlated with ΔR, suggesting that ΔR is not a function of content overlap.

Model	Training Examples	Orig	Para	Δ
UAR23	5M	0.293	0.032	0.261
UAR	1M	0.263	0.026	0.237
UAR19	100K	0.188	0.019	0.169

Table 2: Impact of paraphrasing on attribution performance for authorship representations trained on varying numbers of Reddit users.

version used throughout this paper, which was trained on the histories of around 1M authors.[3]

We report the results of these experiments in Table 2. A paired t test shows that the difference in rank induced by paraphrasing is significant with $p < 0.01$ for all three models. These differences are positively correlated with the MRR scores of the corresponding models, which are shown in footnote 3, suggesting that improved attribution performance may be attributed at least in part to increased sensitivity to stylistic features.

7 Generalization to Novel Tasks

Our experiments have thus far focused on authorship prediction, a task which is presumably best addressed with a model capable of distinguishing writing styles. We now use authorship representations to *directly* distinguish writing styles using a corpus of documents furnished with style annotations, namely, the CDS dataset (Krishna et al., 2020). This consists of writings in disparate styles, including writings by two classical authors (Shakespeare and Joyce), historical American English from different eras, social media messages, lyrics, poetry, transcribed speech, and the Bible. With the notable exception of the two classical authors, most styles in CDS are not author-specific, but rather represent broad stylistic categories. This means that identifying CDS styles is not the same problem as authorship prediction, an important observation we revisit below.

In addition, we repeat the experiment with a corpus furnished instead with *topic* annotations, namely, the Reuters21578 dataset (Lewis, 1997). This is a popular benchmark in text classification consisting of financial news articles, each annotated by one or more topics. We note that certain topics may be associated with particular authors and editors, and therefore style could be a spurious correlate, although we nevertheless expect the authorship representation to perform worse *relative* to the semantic baseline described below.

For each corpus, the experiment consists of simply applying an authorship representation trained on the Reddit dataset to two randomly chosen documents from the corpus. We used Reddit because it has been shown to yield representations that generalize well to other domains (Rivera-Soto et al., 2021). We record the dot product of the resulting vectors and pair this score with a binary indicator specifying whether the two documents carry the same labels. Noting that predicting the binary indicator from the dot product is a highly imbalanced problem, with most document pairs bearing *non-matching* rather than *matching* labels, we simply construct the corresponding receiver operating characteristic (ROC) curve, an illustrative device intended to explore the tradeoffs in making that prediction by thesholding. We report the equal error rate (EER), a simple summary statistic of the ROC curve. Smaller values of this metric are preferable. For good measure, we also report the

[3] We trained the smaller model with the dataset released by Andrews and Bishop (2019). For the larger model we queried Baumgartner et al. (2020) for comments published between January 2015 and October 2019 by authors publishing at least 100 comments during that period. All three models were trained using the default hyperparameter settings of `https://github.com/LLNL/LUAR`. The MRRs of UAR19, UAR, and UAR23 evaluated on a test set composed of comments published future to those constituting the three training corpora are 0.482, 0.592, and 0.682, respectively.

area under that curve (AUC) in §A.4, another summary statistic of the ROC curve.

Finally, because writing style may be difficult to assess without sufficient text content, we also vary the amount of text contributing to the dot products mentioned above. Specifically, rather than predicting whether the label of a *single* document matches that of another on the basis of the dot product of their representations, we more generally predict whether a *group* of randomly chosen documents of the same label shares that label with another group of randomly chosen documents sharing another label on the basis of the dot product of the means over each group of the representations of their constituents.

As a baseline we repeat both experiments using the general purpose document embedding SBERT in place of the authorship representation. SBERT is commonly regarded as a semantic embedding, but is not typically used to discriminate writing styles without further training.

The rationale for the experiment is the following. If the authorship representation primarily encodes stylistic features, then we would expect *poor performance* relative to SBERT on the topic discrimination task since the task presumably does not involve stylistic distinction. However, we would expect *better performance* from the authorship representation than SBERT on the style discrimination task.

These expectations are borne out in the results reported in Figures 7 and 8, which show EER against the number of input documents for each task and each model. The generalization performance of UAR on these novel tasks relative to SBERT is consistent with a representation that is sensitive to stylistic information. Namely, SBERT consistently outperforms UAR on topic classification, while UAR consistently outperforms SBERT on style classification. We present 95% confidence intervals for each curve as lighter regions of the same color surrounding the curve. Although these were calculated using a bootstrap approach, the confidence intervals of the corresponding AUC results shown in Figures 10 and 11 of §A.4 were calculated using a bootstrap-free calculation.

Also shown in Figures 7 to 11 are the results of the same experiments using the two variations of UAR introduced in §6.5. These additional models were included to support an auxiliary argument raised in §8, but also afford

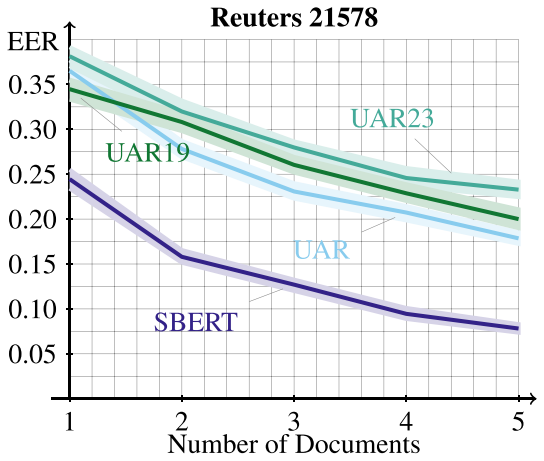

Figure 7: Equal error rate (EER) for UAR and SBERT on topic distinction as the size of the writing sample is varied. Smaller values of EER correspond with better performance.

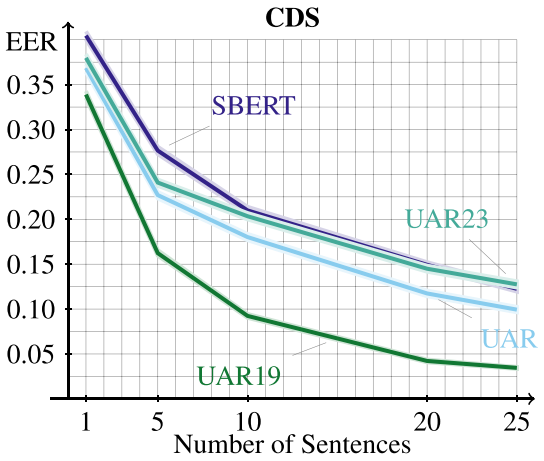

Figure 8: Equal error rate (EER) for UAR and SBERT on style distinction as the size of the writing sample is varied. Smaller values of EER correspond with better performance.

an interesting but subtle insight about the current task. Namely, although UAR19 performs *strictly worse* and UAR23 *strictly better* than UAR on the authorship attribution task, classifying style into broad categories is a different problem than authorship attribution, the latter dealing with fine-grained stylometric features. This accounts for the seemingly contradictory results in Figure 8, in which UAR19 performs *better* than UAR, which in turn performs better than UAR23. Indeed, training UAR on *more* authors produces representations that are more discriminative of individual authors, something which is at odds

with identifying broad stylistic categories for the simple reason that being exposed to more authors affords more opportunities for UAR to discover stylistic features that distinguish authors.

Notwithstanding these observations, we remark that within the CDS dataset, certain styles are likely to be correlated with particular topics, while in the Reuters dataset, certain authors are likely to often write about particular topics. This would suggest that SBERT might perform better on CDS and UAR better on Reuters than one might expect, so the absolute performance on both tasks is not particularly informative.

8 Discussion

Findings We have examined properties of an exemplary authorship representation construction, finding consistent evidence that the success of the representations it engenders at distinguishing authors may be attributed in large part to their sensitivity to style. First, the masking experiments of §5 show that for sufficiently large training corpora, masking a large fraction of content words at training time does not significantly affect ranking performance on held-out data, suggesting that these representations are largely invariant to the presence of content words. On the other hand, the paraphrasing experiments of §6, which seek to alter writing style while preserving content, confirm that paraphrasing drastically impairs ranking performance. Taken together, these experiments suggest that the authorship representations considered are indeed sensitive to stylistic features. This conclusion is corroborated in §7 where we see poor generalization of one of these representations to a topic discrimination task, but better generalization to a style discrimination task, both assessments relative to a semantic baseline.

Limitations All of the experiments in this paper involve instances of the UAR construction. Since our primary research question involves testing the capacity of representations trained for the authorship prediction task to capture stylistic features, we select this construction because there is prior evidence that the representations it engenders perform well at zero-shot cross-domain transfer, for certain combinations of source and target domains, which likely requires some degree of stylistic sensitivity (Rivera-Soto et al., 2021).

By design, our analysis is focused on aggregate model behavior. While this addresses the high-level research questions we pose in the introduction, such *global* analysis does not enable predictions about which specific *local* features are involved in model predictions. As such, an important avenue for future work lies in developing methods that can faithfully identify local authorship features. To this end, frameworks for evaluating the quality of explanations, such as Pruthi et al. (2022), are essential. Beyond the usual challenges of explaining the decisions of deep neural networks, explaining author style may pose further challenges, such as the need to identify groups of features that in combination predict authorship.

We emphasize that completely disentangling style from content may not be attainable, since certain aspects of writing likely blur the line between style and content (Jafaritazehjani et al., 2020). For example, we notice degradation in ranking performance of the SBERT model in Table 1, suggesting that to some extent, SBERT features are also stylistic. Nonetheless, UAR exhibits a markedly larger degradation in performance, suggesting a greater degree of sensitivity to writing style.

Broader Impact This work contributes to the broader goal of formulating notions of *content* and *style* that constitute mutually exclusive and collectively exhaustive aspects of writing that may be embedded in orthogonal subspaces of a Euclidean space. Not knowing whether this ambition is fully realizable, but hopeful others will explore the question in future research, we resign ourselves in this paper to exhibiting two embeddings that accomplish the objective to a limited extent. Specifically, we focus on UAR, which we show to mostly encode style, and SBERT, which is widely assumed to encode content. This being an imperfect decomposition, the primary goal of this paper is to qualitatively assess the *degree* to which UAR encodes style rather than content.

Authorship attribution is likely to be a task that benefits from a representation that is relatively stable over time, specifically an encoding capturing primarily writing style. To this end, another open question is whether a representation may be constructed that encodes style to a *greater* degree than UAR, and if so, whether the representation improves performance on the attribution task. If

such a representation were proposed in the future, the experiments we propose in this paper could be used to validate the assertion that it encodes style to a greater degree than UAR.

On the other hand, content features constitute perfectly legitimate discriminators of authorship in some cases. For example, an author who discusses only a narrow range of topics on a particular forum may easily be distinguished from other authors on the basis of topic features. Not knowing under which circumstances and to what degree content plays a role in authorship attribution, we maintain that the relationship between the performance of a representation on the attribution task and the degree to which the representation encodes content should be explored fully, something that will again require an estimate of the degree to which a representation encodes style.

Another promising application of authorship representations is *style transfer*, where one hopes to rephrase a given statement in the style of a given author. This has been analogously accomplished in the domain of speech, resulting in the ability to have a given statement recited in the voice of a given speaker (see, e.g., Kim et al., 2021). The primary ingredient in this task is a speaker embedding, which is analogous to an authorship representation. However, by construction, a speaker embedding encodes almost exclusively *acoustic* features, but encodes content features, namely the specific words spoken, to a very limited degree. The fact that this observation might be the primary reason for the success of speech transfer portends possible difficulties for the style transfer task. However, as with authorship attribution, the relationship between the success of using a representation for style transfer and the degree to which the representation encodes style should also be fully explored, and again, the experiments proposed in this paper would constitute a natural assessment of that degree.

Acknowledgments

We thank the TACL reviewers and action editors for their insightful comments. We also thank Carina Kauf for the initial masking idea and Hope McGovern for early discussions on PertLE. Part of this work was performed under the auspices of the US Department of Energy by Lawrence Livermore National Laboratory under contract DEAC52-07NA27344.

References

Nicholas Andrews and Marcus Bishop. 2019. Learning invariant representations of social media users. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1684–1695. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1178

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit dataset. In *Proceedings of the 14th International AAAI Conference on Web and Social Media (ICWSM)*, volume 14, pages 830–839. https://doi.org/10.1609/icwsm.v14i1.7347

Janek Bevendorff, Bilal Ghanem, Anastasia Giachanou, Mike Kestemont, Enrique Manjavacas, Ilia Markov, Maximilian Mayerl, Martin Potthast, Francisco Rangel Pardo, Paolo Rosso, Guenther Specht, Efstathios Stamatatos, Benno Stein, Matti Wiegmann, and Eva Zangerle. 2020. Overview of PAN 2020: Authorship verification, celebrity profiling, profiling fake news spreaders on Twitter, and style change detection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 372–383. Springer International Publishing. https://doi.org/10.1007/978-3-030-58219-7_25

Benedikt Boenninghoff, Steffen Hessler, Dorothea Kolossa, and Robert M. Nickel. 2019. Explainable authorship verification in social media via attention-based similarity learning. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 36–45. https://doi.org/10.1109/BigData47090.2019.9005650

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514. https://doi.org/10.1109/TASLP.2021.3124365

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*. Asia Federation of Natural Language Processing.

Julien Hay, Bich-Lien Doan, Fabrice Popineau, and Ouassim Ait Elhara. 2020. Representation learning of writing style. In *Proceedings of the 6th Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 232–243. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.wnut-1.30

Javier Huertas-Tato, Alvaro Huertas-Garcia, Alejandro Martin, and David Camacho. 2022. PART: Pre-trained Authorship Representation Transformer. cs.CL/2209.15373v1. https://doi.org/10.48550/arXiv.2209.15373

Somayeh Jafaritazehjani, Gwénolé Lecorvé, Damien Lolive, and John D. Kelleher. 2020. Style versus content: A distinction without a (learnable) difference? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2169–2180. International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.197

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205. https://doi.org/10.1162/coli_a_00426

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5530–5540.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.55

David D. Lewis. 1997. Reuters-21578 text categorization test collection, Distribution 1.0. AT&T Labs-Research.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: A simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1169

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabás Poczós, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W. Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.169

Frederick Mosteller and David L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, MA.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1018

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.396

Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C. Lipton, Graham Neubig, and William W. Cohen. 2022. Evaluating explanations: How much do explanations from the teacher aid students? *Transactions of the Association for Computational Linguistics*, 10:359–375. https://doi.org/10.1162/tacl_a_00465

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing

toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.14

Sudha Rao and Joel Tetreault. 2018. Dear Sir or Madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1012

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. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.70

Chakaveh Saedi and Mark Dras. 2021. Siamese networks for large-scale author identification. *Computer Speech & Language*, 70:101241. https://doi.org/10.1016/j.csl.2021.101241

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*, pages 343–353. Association for Computational Linguistics.

Efstathios Stamatatos. 2018. Masking topic-related information to enhance authorship attribution. *Journal of the Association for Information Science and Technology*, 69(3):461–473. https://doi.org/10.1002/asi.23968

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (Volume 70)*, pages 3319–3328.

Anna Wegmann and Dong Nguyen. 2021. Does it capture STEL? A modular, similarity-based linguistic style evaluation framework. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7109–7130. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.569

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? Towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.repl4nlp-1.26

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Volume 1 (Long Papers)*, pages 451–462. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1042

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. cs.CL/1904.09675v3.

A Appendix

A.1 Further Dataset Details

The experiments in this paper involve the same datasets used by Rivera-Soto et al. (2021). These datasets consist of Reddit comments (Andrews and Bishop, 2019; Baumgartner et al., 2020), Amazon product reviews (Ni et al., 2019), and fanfiction short stories (Bevendorff et al., 2020), all organized according to author and sorted by publication time. Table 3 presents some statistics of each dataset. Because these are all *anonymous* domains, we use account names as a stand-in for author labels, as proposed in the papers cited above. We recognize that this recourse may introduce a small amount of label noise, since an author may operate multiple accounts, or an account may contain contributions of multiple authors, both of which we assume to be relatively rare.

Each domain is composed of two disjoint splits used independently to train and evaluate models. In the case of Amazon and Reddit, the documents comprising the evaluation split were published in

Domain	Train Authors	Test Authors	Documents per Author
Amazon	100K	35K	≥ 100
fanfic	41K	16K	≥ 2
Reddit	120K	121K	≥ 100
Weibo	94K	90K	≥ 50

Table 3: Dataset statistics.

the *future* relative to those comprising the training split. In addition, in the fanfic domain, the *authors* contributing to the evaluation split are disjoint from those contributing to the training split.

We use a dataset derived from the Weibo social network in §A.3. This dataset contains posts published in the year 2012, primarily in Chinese. We use the first 26 weeks for training, the next 13 weeks for validation, and the final 13 weeks for evaluation. Restricting to authors who have posted a minimum of 50 times and a maximum of 1,000 times results in 94,292 authors in the training split and 90,489 in the evaluation split.

A.2 More on Evaluation

For each training domain we train an authorship representation f_θ, which maps episodes to the unit sphere in $\mathbb{R}^{512}$. In fact, we independently train *three* such representations for each training domain in an effort to reduce variance, a detail we revisit below after discussing the calculation of *mean reciprocal rank* (MRR) for a single representation f_θ.

We compare the authorship of two episodes through the dot product of their images under f_θ, which range from $+1$ to -1, with $+1$ (respectively, -1) corresponding to the strongest prediction that the two input episodes were (respectively, were *not*) composed by the same author.

To calculate the MRR, we evaluate f_θ on all the episodes of each evaluation corpus. Each evaluation corpus consists of episodes $q_1, q_2, \ldots, q_M, t_1, t_2, \ldots, t_N$ for some $M \leq N$, where $t_1, t_2, \ldots, t_N$ were each composed by a distinct author and where q_i and t_i have the same author for all $1 \leq i \leq M$.

For each $1 \leq i \leq M$ we sort the vectors $f_\theta(t_1), f_\theta(t_2), \ldots, f_\theta(t_N)$ according to their dot products with $f_\theta(q_i)$, with those $f_\theta(t_j)$ with the *greatest* dot products having the *lowest* numbered positions in the ranked list. We denote the position of $f_\theta(t_i)$ in this list by $r_i(\theta)$ and define $\text{MRR}(\theta) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{r_i(\theta)}$.

Finally, as mentioned above, to reduce variance, we independently train *three* representations $f_{\theta_1}, f_{\theta_2}, f_{\theta_3}$ for each domain and report the mean $\frac{1}{3}\sum_{k=1}^{3}\text{MRR}(\theta_k)$ for each evaluation domain.

		Level	
		G	L
Domain	**Reddit**	0.438	0.177
	Amazon	0.445	0.186
	Fanfic	0.425	0.147
	Weibo	0.436	0.170

Table 4: Proportions of tokens masked by the Grande (G) or Lite (L) levels of both PertLE and TertLE. For Weibo each proportion is the mean of the proportions at the same level in the other three domains.

A.3 The TertLE Schema

In addition to being expensive to compute, POS tags are unavailable in many low-resource languages. A more quantitative observation about the distinction between content and function words is that function words, such as *the*, tend to be very frequent in a given language, while content words, such as *wallpaper*, tend to be infrequent overall, but may be relatively frequent in documents dealing with those topics. In other words, content words may have higher Term Frequency-Inverse Document Frequency (TF-IDF) scores than function words. One may interpret the highest-scoring words in a document as the most unique or relevant to that document and thus the most likely to be content words. Based on this observation, we explore the possibility of masking words according to their TF-IDF scores rather than their POS tags, an approach we call *TertLE*.

For each domain we fit a TF-IDF model to the training split and use it to index all the documents in the corpus. We introduce the *TertLE Grande* and *TertLE Lite* levels, in which we mask the top-scoring proportion p of words in each document, where p is the proportion of words masked in the same domain by the corresponding PertLE Grande or PertLE Lite schema respectively. These values of p are shown in Table 4.

The experiment proceeds exactly as in §5.1 with results shown in Figure 9. Each number reported is

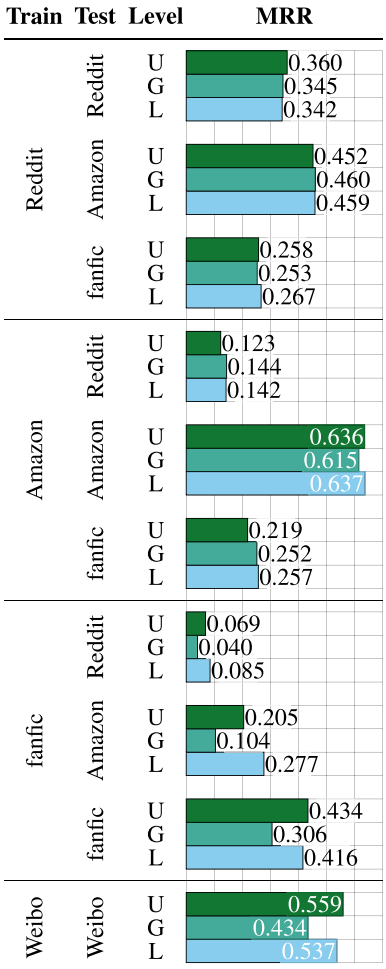

Figure 9: TertLE MRR results for models trained on either unmasked data (U) or data masked according to the TertLE Grande (G) or the TertLE Lite (L) schema.

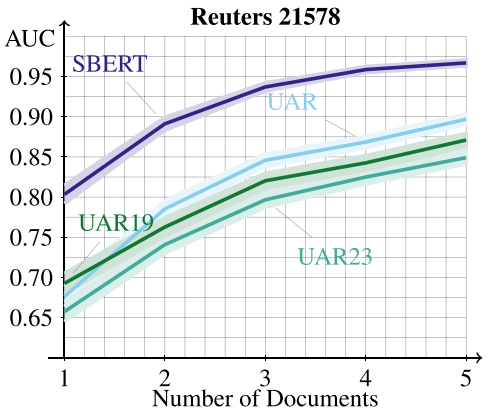

Figure 10: Area under the ROC curve (AUC) for UAR and SBERT on topic distinction as the size of the writing sample is varied. Larger values of AUC correspond with better performance.

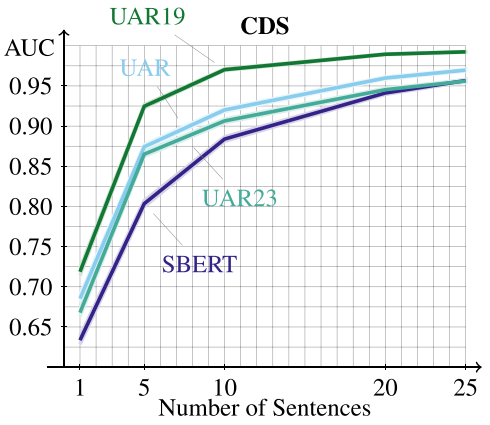

Figure 11: Area under the ROC curve (AUC) for UAR and SBERT on style distinction as the size of the writing sample is varied. Larger values of AUC correspond with better performance.

the mean MRR computed by three independently trained models, where 0.011 is the maximum sample standard deviation over all experiments reported in the table.

We observe that the Lite model outperforms the unmasked model in most cases and is also generally better than the Grande model, especially in the fanfic domain. Once again, the Grande model is generally worse than the unmasked model, but only slightly, and even improves on the unmasked model in some settings. If words with high TF-IDF scores are indeed primarily content words, then this experiment again suggests that authorship representations rely little on content and more heavily on writing style.

As mentioned above, one advantage of the TertLE schema is that it obviates POS tagging. To illustrate this potential, and to determine whether similar patterns hold in another language, we repeat our TertLE experiment with a Chinese dataset scraped from the Weibo social network. See §A.1 for more details on this dataset. This requires replacing the SBERT component of the UAR architecture with a Chinese BERT pre-trained using whole word masking (Cui et al., 2021). The results of the experiment with the Weibo dataset, displayed in Figure 9, show an overall pattern similar to that of the English-focused experiments.

A.4 Further AUC Results

Figures 10 and 11 report the area under the receiver operating characteristic curves (AUC) of the experiments in §7. AUC is a further summary statistic of the ROC that, in contrast to EER, admits a bootstrap-free confidence estimation.