Same Author or Just Same Topic? Towards Topic-Independent Style Representations

Anonymous ACL submission

Abstract

Style is an integral component of language. Recent advances in the development of style representations have increasingly used training objectives from *authorship verification (AV)*: Do two texts have the same author? The assumption underlying the AV training task (same author approximates same writing style) enables self-supervised and, thus, extensive training. However, AV usually does not or only on a coarse-grained level control for topic. The resulting representations might therefore also encode topical information instead of style alone. We introduce a variation of the AV training task that controls for topic using conversation, domain or no topic control as a topic proxy. To evaluate whether trained representations prefer style over topic information, we propose an original variation to the recent `STEL` framework. We find that representations trained by controlling for conversation are better than representations trained with domain or no topic control at representing style independent from topic.

1 Introduction

Linguistic style (i.e., how something is said) is an integral part of natural language. Style is relevant for natural language understanding and generation (Hovy, 2015; Ficler and Goldberg, 2017) as well as the stylometric analysis of texts (El and Kassou, 2014; Goswami et al., 2009). Applications include author profiling (Rao et al., 2010) and style preservation in machine translation systems (Niu et al., 2017; Rabinovich et al., 2017).

While authors are theoretically able to talk about any topic and (un-)consciously choose to use many styles (e.g., designed to fit an audience in Bell (1984)), it is typically assumed that there are combinations of style features that are distinctive for an author (sometimes called an author's *idiolect*). Based on this assumption, the *authorship verification* task (AV) aims to predict whether two texts have been written by the same author (Coulthard, 2004; Neal et al., 2017; Martindale and McKenzie, 1995). Recently, training objectives based on the AV task have been used to train neural style representations (Boenninghoff et al., 2019b; Hay et al., 2020; Zhu and Jurgens, 2021). Training objectives on AV are especially promising because they do not require any additional labeling assuming author identifiers are available. Similar to the distributional hypothesis, the assumption underlying the AV training task (same author approximates same writing style) enables extensive self-supervised learning.

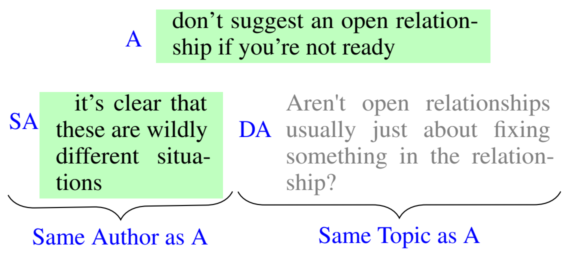

Figure 1: **Triple Authorship Verification (TAV) Task.** Similar to the traditional authorship verification task (AV), the TAV task is to match A with the utterance that was written by the same author (SA). This is complicated by another utterance that was written on the same topic but by a different author (DA). We test three topic proxies: conversation, domain and no topic control.

Style and topic are often correlated (Gero et al., 2019; Bischoff et al., 2020): For example, people might write more formally and about their professional career in a cover letter but more informally and about personal hobbies in an online chat with friends. As a result, style representations might encode spurious topic correlations (Poliak et al., 2018), especially when their AV training objective does not control for topic (Halvani et al., 2019; Sundararajan and Woodard, 2018). Current style representation learning methods either use no or only limited control for content (Hay et al., 2020) or use domain labels (Boenninghoff et al., 2019a; Zhu and Jurgens, 2021). For example, Zhu and Jurgens (2021) work with 24 domain labels for more than 100,000 `Amazon` reviews. However, using a small set of labels might be too coarse-grained to fully control for topic.

Approach. We introduce a training task for style representation learning that addresses topic correlation: The Triple Authorship Verification (TAV) task with the help of a topic proxy (Figure 1). We compare using no topic control and using conversation or domain as topic proxies. We train several siamese BERT-based neural networks to learn style representations (Reimers and Gurevych, 2019) by using the TAV and the more common binary AV as training tasks. We train on utterances from the platform `Reddit`. Our approach could be applied to any other conversation dataset as well. We propose a variation to the STyle EvaLuation framework (`STEL`) to tackle a lack in evaluation methods that can assess whether models prefer style over topic information in their representations.

Contribution. With this paper, we (a) contribute an extension of the AV task that inherently controls for topic with conversation labels, (b) compare style representation on various AV and TAV tasks that vary in their topic proxies, (c) introduce a variation of the `STEL` framework (Wegmann and Nguyen, 2021) to evaluate whether representations prefer content over style information and (d) demonstrate found stylistic features via agglomerative clustering. We find that representations trained on the conversation topic proxy are better than representations trained with a domain or no topic proxy at representing style independent from topic (Section 4). Additionally, combining the conversation topic proxy with the TAV training task leads to better results than combining it with the binary AV task. We show that our representations are sensitive to stylistic features like punctuation and apostrophe types such as ’ vs. ' using agglomerative clustering. We hope to further the development of content-controlled style representations. Our code and data will be publicly released on GitHub.

2 Related Work

Authorship Attribution (AA) is the task of determining who authored a particular document from a pool of possible authors. Texts are assumed to contain stylistic tendencies that help with classifying unattributed documents (Coulthard, 2004; Neal et al., 2017). A sub-field of authorship attribution is authorship verification (AV) (Koppel and Schler, 2004). There are several recent approaches in deep authorship attribution and verification (Shrestha et al., 2017; Litvak, 2019; Boenninghoff et al., 2019a; Saedi and Dras, 2021; Hay et al., 2020; Hu et al., 2020; Zhu and Jurgens, 2021). Training on transformer architectures like BERT has been shown to be competitive with other neural as well as non-neural approaches in AV and style representation (Zhu and Jurgens, 2021; Wegmann and Nguyen, 2021). As style and topic are often correlated (Gero et al., 2019; Bischoff et al., 2020), AV and AA methods have controlled for topic by restricting the feature space to contain “topic-independent” features like function words or character n-grams (Neal et al., 2017; Stamatatos, 2017; Sundararajan and Woodard, 2018). However, even these features have been shown to not necessarily be topic-independent (Litvinova, 2020).

Semantic Sentence Representations Semantic sentence embeddings are typically trained using supervised or self-supervised learning (Reimers and Gurevych, 2019). For supervised learning, models are often trained on manually labelled natural language inference datasets (Conneau et al., 2017). For self-supervised learning, *contrastive* learning objectives (Hadsell et al., 2006) have been increasingly used. Contrastive objectives push semantically distant sentence pairs apart and pull semantically close sentence pairs together. Different strategies for selecting positive and negative pairs have been used, e.g., slightly augmented and thus almost identical vs. randomly sampled other sentences (Giorgi et al., 2021; Gao et al., 2021). Reimers and Gurevych (2019) also experiment with a *triplet loss*, which pushes an anchor closer to a semantically close sentence and pulls the same anchor apart from a semantically distant sentence. Semantic representations are typically first evaluated on the task that they have been trained on, e.g., binary tasks for binary contrastive objectives and triplet tasks (similar to Figure 1) for triplet objectives (Reimers and Gurevych, 2019). Semantic representations are often also evaluated on the STS benchmark (Cer et al., 2017) or semantic downstream tasks like semantic search, NLI (Bowman et al., 2015; Williams et al., 2018) or SentEval (Conneau and Kiela, 2018).

Style Representations Recently, style representations of sentences have been trained using AV as training tasks. Typically, objective functions that are known from semantic embedding learning have been used (Hay et al., 2020; Zhu and Jurgens, 2021). As a result of the correlation of topic with AV, style representations trained on AV tasks might also encode spurious topic correlations (Bischoff et al., 2020). Zhu and Jurgens (2021) address this by sampling half of the different and same author utterances from the same and the other half from different domains (e.g., subreddits for `Reddit`). Style representations are often evaluated on the AV task (Boenninghoff et al., 2019a; Zhu and Jurgens, 2021; Bischoff et al., 2020).

3 Style Representation Model

We describe the new triple authorship verification task (TAV, Section 3.1), the generation of the TAV tasks (Section 3.2) and the models we train based on the TAV and the binary AV tasks (Section 3.3).

3.1 Triple Authorship Verification Task

The more common authorship verification (AV) task is the binary task of predicting whether two texts are written by the same (SA) or different authors (DA). Methods optimized for AV have been known to make use of topical cues (Sari et al., 2018; Sundararajan and Woodard, 2018; Potha and Stamatatos, 2018) and to perform badly in cross-topic settings (Halvani et al., 2019; Bischoff et al., 2020). Recent studies use AV tasks to train style representations and address possible topic-correlation by controlling for domain (Zhu and Jurgens, 2021; Boenninghoff et al., 2019b). However, using a (usually small set of) domain labels might be too coarse-grained to fully control for topic.

Topic proxy. We compare the effect of three different topic proxies by sampling different author utterances from the same *conversation*, from the same *domain* (i.e., subreddit for `Reddit` as in Zhu and Jurgens (2021)) or *randomly* (as a baseline, similar to Hay et al. (2020)). In semantic sentence embedding learning, conversations have also previously been used as a proxy for semantic information encoded in utterances (Yang et al., 2018; Liu et al., 2021). We expect two utterances that were sampled from the same conversation to usually be closer w.r.t. topic than two utterances sampled from the same domain. Similarly, we expect randomly sampled utterances to be more distant in topic than utterances sampled from the same domain. Conversation and random topic labels can easily be inferred from conversation datasets without requiring additional labeling. Many conversation datasets also include community or domain labels.

TAV task. We further introduce an adaption of the more common binary Authorship Verification task — the Triple Authorship Verification task (TAV, Figure 1): Given an anchor utterance A and two other utterances SA and DA, the task is to identify which of the two sentences is SA (i.e., written by the same author as A). Using a triple AV setup enables the use of learning objectives that require three input sentences and have been successful in semantic embedding learning (Reimers and Gurevych, 2019). It is also possible to adapt this setup to include one positive SA and several negative DA utterances (similar to Gao et al. (2021)). We experiment with both TAV and AV for style representation learning.

Topic-controlled AV task. One TAV task, which consists of 3 utterances (A, SA, DA), can be split up into two topic-controlled, binary AV tasks: (A, SA) and (A, DA). In comparison to the more common AV task, the TAV and the topic-controlled AV tasks select DA from utterances that have been written by a different author to be about a similar topic as A.

3.2 Task Generation

We use a 2018 `Reddit` sample with utterances from 100 active subreddits[1] extracted via `ConvoKit` (Chang et al., 2020)[2]. Per subreddit, we sample 600 conversations with at least 10 posts (which we call utterances). All subreddits are directed at an English audience, which we infer from the subreddit descriptions.

Generation. First, we removed all invalid utterances[3]. Then, we split the set of authors into a non-overlapping 70%, 15% and 15% train, dev, test author split. For each author split we generate a set of Triple Authorship Verification tasks for the three topic proxies (conversation, domain, random), i.e., nine sets in total. First, we generate the conversation topic proxy tasks for all author splits. (A, DA) are sampled to be written by different authors but in the same conversation. Then, utterance SA is

[1] https://zissou.infosci.cornell.edu/convokit/datasets/subreddit-corpus/subreddits_small_sample.txt

[2] MIT license

[3] utterance of only spaces, tabs, line breaks or of the form: `""`, `" [removed] "`, `"[ removed ]"`, `"[removed]"`, `"[ deleted ]"`, `"[deleted]"`, `" [deleted] "`

Topic Proxy	Data Split	Task # TAV	Task # AV	Uttterance #	Author #	Author ma	(A, SA) sc	(A, SA) sd	(A, DA) sc	(A, DA) sd
Conversation	train set	210,000	420,000	546,757	194,836	9	0.27	0.56	1.00	1.00
	dev set	45,000	90,000	116,451	41,848	8	0.26	0.55	1.00	1.00
	test set	45,000	90,000	116,621	41,902	8	0.27	0.55	1.00	1.00
Domain	train set	210,000	420,000	544,587	240,065	9	0.27	0.56	0.01	1.00
	dev set	45,000	90,000	116,490	50,939	8	0.26	0.55	0.02	1.00
	test set	45,000	90,000	116,586	51,182	8	0.27	0.55	0.02	1.00
Random	train set	210,000	420,000	548,082	270,079	9	0.27	0.56	0.00	0.01
	dev set	45,000	90,000	117,149	57,352	8	0.26	0.55	0.00	0.01
	test set	45,000	90,000	117,434	57,726	8	0.27	0.55	0.00	0.02

Table 1: **Data Split Statistics.** Per topic proxy, we display the number of tasks (# TAV, # AV), unique utterances and authors for each split. We also show the maximum number of times an author occurs as the anchor's author (ma) and the fraction of (A, SA) and (A, DA)-pairs that occur in the same conversation (sc) and domain (sd).

sampled from all utterances written by A's author that are different from utterance A. Then, to keep as many possibly correlating variables constant, we reuse the same (A, SA)-pairs for the domain and random topic proxy tasks. (A, DA) is sampled from the same domain or randomly respectively. There are no identical (A, SA) or (A, DA) pairs, and thus no repeating TAV or topic-controlled AV tasks (Section 3.1). However, it is possible that some utterances occur more than once across tasks. In total, we generate 210k train, 45k dev and 45k test tasks for each topic proxy (see Table 1), corresponding to a total of 420k, 90k and 90k topic-controlled AV-pairs when splitting the TAV task into (A, SA) and (A, DA) pairs (c.f. Section 3.1).

3.3 Models

We use the `Sentence-Transformers`[4] python library (Reimers and Gurevych, 2019)[5] to fine-tune several siamese networks based on (1) 'bert-base-uncased', (2) 'bert-base-cased' (Devlin et al., 2019) and (3) 'roberta-base' (Liu et al., 2019). We expect those to perform well based on experiments by Zhu and Jurgens (2021) and Wegmann and Nguyen (2021). To the best of our knowledge the performance of triplet loss (e.g., Reimers and Gurevych (2019)) vs. binary contrastive loss (Hadsell et al., 2006) has not been rigorously compared on the same set of examples for style representation learning. The binary contrastive loss function uses a pair of sentences as input while the triplet loss expects three input sentences. Thus, we compare them by using (a) contrastive loss (Hadsell et al., 2006) with our topic-controlled AV (Section 3.1) tasks and (b) triplet loss (Reimers and Gurevych, 2019) with our TAV tasks (Figure 1). For the loss functions, we experiment with three different values for the margin hyperparameter (i) 0.4, (ii) 0.5, (iii) 0.6. We train with a batch size of 8 over 4 epochs using 10% of the training data as warm-up steps. We use the Adam optimizer with the default learning rate (0.00002). We leave all other parameters as default. We use the BinaryClassificationEvaluator on the binary AV training task with contrastive loss and the TripletEvaluator on the TAV training task with triplet loss from `Sentence-Transformers` to select the best model out of the 4 epochs. The BinaryClassificationEvaluator calculates the accuracy of identifying similar and dissimilar sentences, while the TripletEvaluator checks if the distance between A and SA is smaller than the distance between A and DA. We use cosine similarity as the distance function.

[4] `https://sbert.net/`
[5] with Apache License 2.0

4 Evaluation

We evaluate the learned style representations w.r.t. the training task (i.e., the topic-controlled AV and TAV task) in Section 4.1. Then, we evaluate whether models learn to represent style via the performance on the `STEL` framework (Section 4.2). Last, we evaluate representations on their topic-independence with our adapted version of `STEL` (Section 4.3).

4.1 Authorship Verification

To compare AV performance, typically AUC is calculated, or a similarity threshold is chosen to calculate AV accuracy (Zhu and Jurgens, 2021; Kestemont et al., 2021). We use AUC as a performance measure for the binary AV task and accuracy for the TAV task. On the dev set, RoBERTa models

Model		Conversation		Domain		Random	
		AV	TAV	AV	TAV	AV	TAV
Topic Proxy	Training Task	AUC $\pm\sigma$	acc $\pm\sigma$	AUC $\pm\sigma$	acc $\pm\sigma$	AUC $\pm\sigma$	acc $\pm\sigma$
original RoBERTa		.53	.53	.57	.58	.61	.63
Conversation	**AV**	**.69** $\pm$ .02	**.68** $\pm$ .02	.70 $\pm$ .02	.69 $\pm$ .02	.71 $\pm$ .02	.70 $\pm$ .02
	TAV	**.69** $\pm$ .00	**.68** $\pm$ .00	.70 $\pm$ .00	.69 $\pm$ .00	.71 $\pm$ .00	.70 $\pm$ .00
Domain	**AV**	.68 $\pm$ .01	.67 $\pm$ .01	**.71** $\pm$ .01	**.70** $\pm$ .01	.73 $\pm$ .02	.73 $\pm$ .00
	TAV	.68 $\pm$ .00	**.68** $\pm$ .00	.70 $\pm$ .00	**.70** $\pm$ .00	.72 $\pm$ .00	.72 $\pm$ .01
Random	**AV**	.58 $\pm$ .01	.59 $\pm$ .01	.63 $\pm$ .02	.66 $\pm$ .01	**.79** $\pm$ .00	**.78** $\pm$ .00
	TAV	.58 $\pm$ .00	.59 $\pm$ .00	.63 $\pm$ .03	.65 $\pm$ .00	.77 $\pm$ .00	.77 $\pm$ .00

Table 2: **Test Results.** Results for 6 different fine-tuned RoBERTa models on the test sets. We display the accuracy of the models for the triple authorship verification task (TAV) and the AUC for the binary topic-controlled authorship verification task (AV). We display the standard deviation (σ). Best performance per column is boldfaced. Models generally outperform others on the topic proxy they have been trained on.

consistently outperformed the cased and uncased BERT models and different margin values only led to small performance differences (Appendix A). Consequently, we only display the performance of the six fine-tuned RoBERTa models for the binary AV (using contrastive loss) and the TAV training task (using triplet loss) with margin values of 0.5 on the test sets in Table 2. We aggregate performance with mean and standard deviation for three different random seeds per model parameter combination.[6] Generally, the fine-tuned models tested on the topic proxy they were trained on (diagonal) outperform other models that were not trained on that same topic proxy.

Tasks with the conversation topic proxy are hardest to solve. For all models the performance is lowest on the conversation test set and increases on the domain and further on the random test set. This is in line with our assumption that the conversation test set has semantically closer (A, DA)-pairs that make the AV task harder (Section 3.1).

Models trained with the conversation topic proxy perform similarly on all three test sets. Across the three test sets, the difference in performance is biggest for models trained with the random topic proxy and smallest for models trained with the conversation topic proxy. Representations trained with the random (or domain) topic proxy might latch on to topical features that are helpful in the random (and domain) test set but not the conversation test set. Models learned with the conversation topic proxy might in turn learn more topic-agnostic representations. We investigate this further in Section 4.3.

[6] We used seeds 103-105. A total of 5 out of 18 models did not learn. We re-trained those with different seeds.

The AV & TAV training task lead to similar performance on the test sets. Models trained on the TAV task generally have a smaller standard deviation than models trained on the binary AV task. For the same topic proxy used in training, the mean accuracy and AUC scores are similar.

4.2 STEL Framework

We calculate the performance of the representations on the `STEL` framework (Wegmann and Nguyen, 2021)[7]: Models are evaluated on whether they are able to measure differences in style across 4 dimensions (formal vs. informal style, complex vs. simple style, contraction usage and number substitution usage) in a content-controlled setup, i.e., while the content remains the same. Models have to match two sentences to the style of two given anchor sentences (e.g., Figure 2 before alterations). We display the `STEL` results for the RoBERTa models in Table 3. **`STEL` performance is comparable across the different topic proxies and AV & TAV tasks.** Surprisingly, the overall `STEL` performance for the fine-tuned models is lower than that of the original RoBERTa model (Liu et al., 2019). Thus, models might 'unlearn' some style information. Performance stays approximately the same or improves for the formal/informal and the contraction dimensions, but drops for the complex/simple and the nb3r substitution dimensions. Based on manual inspection, we notice nb3r substitution to regularly appear in specific conversations and for specific

[7] `https://github.com/nlpsoc/STEL`, with data from Rao and Tetreault (2018); Xu et al. (2016) and with permission from Yahoo for the "L6 - Yahoo! Answers Comprehensive Questions and Answers version 1.0 (multi part)": `https://webscope.sandbox.yahoo.com/catalog.php?datatype=l`. Data and code available with MIT License with exceptions for proprietary Yahoo data.

		all		formal, n = 815		complex, n = 815		nb3r, n = 100		c'tion, n = 100	
		o	t-a	o acc±σ	t-a acc±σ	o acc±σ	t-a acc±σ	o acc±σ	t-a acc±σ	o acc±σ	t-a acc±σ
org		**.80**	.05	.83	.09	**.73**	.01	**.94**	**.13**	**1.0**	.00
c	**b**	.71	**.35**	.83 ± .02	.64 ± .00	.57 ± .02	.13 ± .04	.61 ± .02	.04 ± .01	.91 ± .10	.00 ± .01
	t	.71	**.42**	.81 ± .02	**.69** ± .02	.59 ± .01	**.24** ± .02	.65 ± .09	.03 ± .01	.99 ± .02	**.04** ± .02
d	**b**	.73	.28	.84 ± .01	.56 ± .04	.69 ± .05	.05 ± .02	.61 ± .02	.03 ± .02	.98 ± .03	.00 ± .00
	t	.71	.32	.82 ± .01	.61 ± .02	.57 ± .01	.12 ± .01	.64 ± .05	.03 ± .01	.99 ± .01	.01 ± .01
r	**b**	.72	.22	**.85** ± .01	.46 ± .04	.57 ± .01	.03 ± .01	.62 ± .04	.05 ± .02	.98 ± .01	.00 ± .00
	t	.71	.24	**.85** ± .00	.50 ± .02	.56 ± .01	.04 ± .01	.59 ± .03	.06 ± .01	.98 ± .04	.00 ± .00

Table 3: **(Topic-adapted) `STEL` Results.** We display `STEL` accuracy across 4 style dimensions (n =number of instances) for the same RoBERTa models as in Table 2: Per topic proxy (conversation - c, domain - d, random - r), and training task (AV - b, TAV - t) the performance on the set of task instances with (t-a) and without topic-adaption (o) is displayed. Per column, the best performance is boldfaced. For the fine-tuned RoBERTa models, performance generally increases on the topic-adapted `STEL` task compared to the original RoBERTa model (org).

topics. Future work could investigate whether the use of nb3r substitution is less consistent for one author than other stylistic dimensions. As the nb3r dimension of `STEL` only consists of 100 instances, future work could increase the number of instances.

We perform an error analysis to further investigate the `STEL` performance drop in the complex/simple dimension. We manually look at consistently unlearned (i.e., wrongly predicted by the fine-tuned but correctly predicted by the original RoBERTa model) or learned (i.e., wrongly predicted by the RoBERTa model and correctly predicted by the fine-tuned model) `STEL` instances (Appendix B.1). The share of examples with problematic ambiguities (e.g., typos, errors in grammar, words that might actually increase and not decrease complexity) is higher for the unlearned (50/55) than for the newly learned `STEL` instances (29/41). We display two examples of ambiguous instances in Table 4. Generally, the number of complex/simple `STEL` instances with ambiguities is surprisingly high for both the learned as well as the unlearned instances, consistent with the lower performance of the models in this category. Several of the found ambiguities should be relatively easy to correct in the future (e.g., typos or punctuation differences).

4.3 Topic-Independence of Style Representations

We tested whether models are able to represent different authors (in Section 4.1) and styles when the topic remains the same (Section 4.2). However, we have not tested whether models learn to represent style independent from topic.

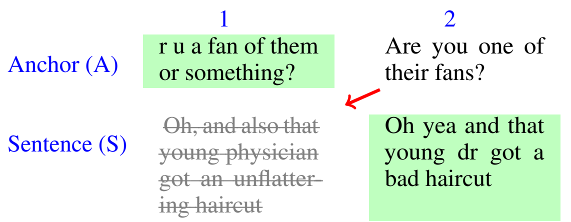

Figure 2: **Topic-adapted `STEL` Task.** We take the original `STEL` instances and move A2 to the sentence position with the different style (here: the more formal A2 replaces the more formal S1). These resulting triple tasks lose the topic-control property but can test if a model prefers style over content cues.

There have been a few different methods used to test whether style representations encode unwanted topical information by (a) comparing performance on the AV task across domain (Boenninghoff et al., 2019b; Zhu and Jurgens, 2021), (b) assessing performance on function vs. content words (Hay et al., 2020; Zhu and Jurgens, 2021) or (c) predicting domain labels from utterances using their style representations (Zhu and Jurgens, 2021). However, these evaluation methods remain incomplete: Domain labels usually come from a small set of coarse-grained labels and function words have been shown to not necessarily be topic-independent (Litvinova, 2020). Additionally, high AV performance might not be the same as a good style representation — as same author = same style is only an approximation.

To test if models learn to prefer style over topic, we introduce a variation to the `STEL` framework — the *topic-adapted `STEL` task*: From one original `STEL` instance (Wegmann and Nguyen, 2021), we take the sentence that has the same style as A2 and replace it with A2 (Figure 2). Thus, here S2 is

Agg.	GT	Anchor 1 (A1)	Anchor 2 (A2)	Sentence 1 (S1)	Sentence 2 (S2)	Ambiguity
un	✓	TDL Group announced in March 2006, in response to a request [...]	[...] storm names Alberto Helene Beryl Isaac Chris [...]	Palestinian voters in the Gaza Strip [...] were eligible to participate in the election.	1. Palestinian voters in the Gaza Strip [...] were eligible to participate in the election.	A1/A2 about different topics
l	✗	[...] 51 Phantom [...] received nominations in that same category.	[...] 1 phantom [...] received nominations in the same category.	[...] the Port Jackson District Commandant could exchange with all military land with buildings on the harbor.	[...] the Port Jackson District Commandant could communicate with all military installations on the harbour.	A2 spelling mistake, S1 sounds unnatural

Table 4: **STEL Error Analysis.** For the complex/simple STEL dimension, we display examples of ambiguous instances that were learned (l) or unlearned (un) the fine-tuned RoBERTa models. A ground truth (GT) of ✓ means that S1 matches with A1 and S2 with A2 in style, while ✗ means S1 matches with A2 and S2 with A1.

written in the same style as A1 but about a different topic and the new S1 is written in a different style but has the same topic. This setup is similar to the TAV task (Figure 1). The main difference to the TAV task is that we do not use same author as a proxy for same style but instead use the predefined style dimensions from the STEL framework. We display the topic-adapted STEL results in Table 3. The performance for the new task is low (< 0.5 which corresponds to a random baseline). However, the task is also very difficult as lexical overlap is usually high between the anchor and the false choice (i.e., the sentence that was written in a different style but has the same topic). Nevertheless, performance should only be considered in combination with other evaluation approaches (Sections 4.1 and 4.2) as on this task alone models might perform well because they punish same topic information.

Models trained on the TAV task with the conversation topic proxy are the best at representing style independent from topic. The performance increases from an accuracy of 0.05 for the original RoBERTa model to up to $0.42 \pm .01$ for the representation trained with the TAV task on the conversation topic proxy. This 'TAV conversation representation' did not just learn to punish same topic cues because of its performance on the AV task and the STEL framework: (1) On the AV task, the representation performed comparably on all three test sets. If the model had learned to just punish same topic cues, we would expect a clearer difference in performance as confounding same topic information should be more prevalent for the random than the conversation test set. (2) The representation performed comparably to the other representations on the STEL framework, where style information is needed to solve the task but topic information cannot be used.

C	Consistent	Example
3	no last punct.	I am living in china, they are experiencing an enormous baby boom
4	punctuation / casing	huh thats odd i'm in the 97% percentile on iq tests, the sat, and the act
5	' vs '	I assume it's the blind lady?
7	linebreaks	I admire what you're doing but [...] I know I'm [...]

Table 5: **Clusters for RoBERTa Trained on TAV with Conversation Topic Proxy.** We display one example for 4 out of 7 clusters. We mention noticeable consistencies within the cluster (Consistent).

5 Style Representation Analysis

We want to further understand what the learned style representations learn to be similar styles. We take the best-performing style representation (RoBERTa trained on the TAV task with the conversation topic proxy and seed 106) and perform agglomerative clustering on a sample of 5.000 TAV tasks of the conversation test set resulting in 14,756 unique utterances. We use 7 clusters based on an analysis of Silhouette scores (Appendix C). Out of all utterance pairs that have the same author, 46.2% appear in the same cluster. This is different from random assignments among 7 clusters[8] which corresponds to $20.1\% \pm .0$. As authors will have a certain variability to their style, a perfect clustering according to writing style would not assign all same author pairs to the same cluster.

In Table 5, we display examples for 4 out of 7 clusters. We manually looked at a few hundred examples per cluster to find consistencies. We mostly

[8]Calculated mean and standard deviation of 100 random assignments of utterances to the 7 clusters of the same size.

found consistent differences between clusters in the punctuation (e.g., 97% of utterances have no last punctuation mark in Cluster 3 vs. an average of 37% in the other clusters), casing (e.g., 67% of utterances that use *i* instead of *I* appear in Cluster 4), contraction spelling (e.g., 22 out of 27 utterances that use *didnt* instead of *didn't* appear in Cluster 4), the type of apostrophe used (e.g., 90% of utterances use ' vs ' in Cluster 5 vs. an average of 0% in the other clusters) and line breaks within an utterance (e.g., 72% of utterances in Cluster 7 include line breaks vs. an average of 22% in the other clusters). For comparison we also cluster with the original RoBERTa model (Liu et al., 2019). The only three interesting RoBERTa clusters (i.e., clusters that contain more than three elements and not as many as 86.7% of all utterances), seem to mostly differ in utterance length (average number of characters are 15 in Cluster 2 vs. in 1278 in Cluster 3) and in the presence of hyperlinks (84% of utterances contain 'https://' in Cluster 4 vs. an overall average of 2%) Average utterance lengths are not as clearly separated by the clusters of the trained representations. For more detail, refer to Appendix D.

6 Limitations and Future Work

We propose several directions for future research:

First, conversation labels are already inherently available in conversation corpora like `Reddit`. However, it remains a difficulty to transfer the conversation topic proxies to other than conversation datasets. With the recent advances in semantic sentence embeddings, it might be interesting to train style representations on TAV tasks with a new topic proxy: Two utterances could be labelled as having the same topic if their semantic embeddings are close to each other (e.g., when cosine similarity is above a constant threshold).

Second, even when using our topic proxies semantic information can still be useful for AV: If one person writes "my husband" in one utterance and another writes "my wife" in another utterance, it is highly unlikely that those have been generated by the same person. We expect this issue to only occur in a limited number of examples.

Third, for the topic-adapted `STEL` task, the so-called "triplet problem" (Wegmann and Nguyen, 2021) remains a potential problem. Consider the example in Figure 2. Here, the `STEL` framework only guarantees that A1 is more informal than A2 and S2 is more informal than S1. Thus, in some cases A2 can be stylistically closer to A1 than S2. However, we expect this case to be less prevalent: A2 would need to be already pretty close in style to A1 or both S2 and S1 substantially more informal or formal than A1. In the future, removing problematic instances could alleviate a possible maximum performance cap.

Fourth, the representation models may learn to represent individual stylistic variation as we use utterances from the same individual author as positive signals (c.f. Zhu and Jurgens (2021)). However, because the representation models learn with same author pairs that are generated from thousands of authors, it is likely that they also learn consistencies along groups of authors that use similar style features (e.g., demographic groups based on age or education level, or subreddit communities). Future work could explore how different topic proxies and training tasks influence the type of styles that are learned.

7 Conclusion

Recent advances in the development of style representations have increasingly used training objectives from authorship verification (Hay et al., 2020; Zhu and Jurgens, 2021). However, AV tasks — and style representations trained on them — often do not or only on a coarse-grained level control for topic (e.g., with domain labels). We train different style representations by controlling for topic using conversation or domain membership as a topic proxy. We also introduce the new Triple Authorship Verification task (TAV) and compare it to the more common binary AV task. We propose an original adaptation of the recent `STEL` framework (Wegmann and Nguyen, 2021) to test whether learned representations favor style over topic information. We find that representations that were trained on the TAV task with a conversation topic proxy represent style in a way that is more independent from topic than models using other topic proxies or the AV training task. We demonstrate some of the learned stylistic differences via agglomerative clustering — e.g., the use of a right single quotation mark vs. an apostrophe in contractions. We hope to contribute to increased efforts towards learning topic-controlled style representations.

Ethical Considerations

We use utterances taken from 100 subcommunities (i.e., subreddits) of the popular online platform `Reddit` to train style representations with different training tasks and compare their performance. With our work, we aim to contribute to the development of general style representations that are disentangled from content. Style representations have to potential to increase classification performance for diverse demographics and social groups (Hovy, 2015).

The user demographics on the selected 100 subreddits are likely skewed towards particular demographics. For example, locally based subreddits (e.g., canada, singapore) might be over-represented. Generally, the average `Reddit` user is typically more likely to be young and male.[9] Thus, our representations might not be representative of (English) language use across different social groups. However, experiments on the set of 100 distinct subreddits should still demonstrate the possibilities of the used approaches and methods. We hope the ethical impact of reusing the already published `Reddit` dataset (Baumgartner et al., 2020; Chang et al., 2020) to be small but acknowledge that reusing it will lead to increased visibility of data that is potentially privacy infringing. As we aggregate the styles of thousands of users to calculate style representations, we expect it to not be indicative of individual users.

We confirm to have read and that we abide by the ACL Code of Ethics.

References

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 830–839, Atlanta, USA. Association for the Advancement of Artificial Intelligence.

Allan Bell. 1984. Language style as audience design. *Language in Society*, 13(2):145–204.

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

[9]https://www.journalism.org/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

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

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

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60, 1st virtual meeting. Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Malcolm Coulthard. 2004. Author identification, idiolect, and linguistic uniqueness. *Applied linguistics*, 25(4):431–447.

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

Sara El Manar El and Ismail Kassou. 2014. Authorship analysis studies: A survey. *International Journal of Computer Applications*, 86(12).

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, Copenhagen, Denmark. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Katy Gero, Chris Kedzie, Jonathan Reeve, and Lydia Chilton. 2019. Low level linguistic controls for style transfer and content preservation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 208–218, Tokyo, Japan. Association for Computational Linguistics.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895, Online. Association for Computational Linguistics.

Sumit Goswami, Sudeshna Sarkar, and Mayur Rustagi. 2009. Stylometric analysis of bloggers' age and gender. In *Proceedings of the International AAAI Conference on Web and Social Media (Volume 3)*, pages 214–217.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*, pages 1735–1742.

Oren Halvani, Christian Winter, and Lukas Graner. 2019. Assessing the applicability of authorship verification methods. In *Proceedings of the 14th International Conference on Availability, Reliability and Security (ARES '19)*, New York, NY, USA. Association for Computing Machinery.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

Julien Hay, Bich-Lien Doan, Fabrice Popineau, and Ouassim Ait Elhara. 2020. Representation learning of writing style. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 232–243, Online. Association for Computational Linguistics.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.

Zhiqiang Hu, Roy Ka-Wei Lee, Lei Wang, Ee-peng Lim, and Bo Dai. 2020. Deepstyle: User style embedding for authorship attribution of short texts. In *Web and Big Data*, pages 221–229, Cham. Springer International Publishing.

Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Benno Stein, and Martin Potthast. 2021. Overview of the cross-domain authorship verification task at PAN 2021. In *Proceedings of the Working Notes of CLEF 2021*, pages 1743–1759, Bucharest, Romania.

Moshe Koppel and Jonathan Schler. 2004. Authorship verification as a one-class classification problem. In *Proceedings of the twenty-first international conference on Machine learning*, page 62, Banff, Canada.

Marina Litvak. 2019. Deep dive into authorship verification of email messages with convolutional neural network. In *5th International Conference on Information Management and Big Data*, pages 129–136, Lima, Peru. Springer International Publishing.

Tatiana Litvinova. 2020. Stylometrics features under domain shift: Do they really "context-independent"? In *22nd International Conference on Speech and Computer*, pages 279–290, Cham. Springer International Publishing.

Che Liu, Rui Wang, Jinghua Liu, Jian Sun, Fei Huang, and Luo Si. 2021. DialogueCSE: Dialogue-based contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2396–2406, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Colin Martindale and Dean McKenzie. 1995. On the utility of content analysis in author attribution: "the federalist". *Computers and the Humanities*, 29(4):259–270.

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

Xing Niu, Marianna Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2814–2819, Copenhagen, Denmark. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Nektaria Potha and Efstathios Stamatatos. 2018. Intrinsic author verification using topic modeling. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, SETN '18, New York, NY, USA. Association for Computing Machinery.

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084, Valencia, Spain. Association for Computational Linguistics.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*, SMUC '10, page 37–44, New York, NY, USA. Association for Computing Machinery.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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.

Efstathios Stamatatos. 2017. Masking topic-related information to enhance authorship attribution. *Journal of the Association for Information Science and Technology*, 69(3):461–473.

Kalaivani Sundararajan and Damon Woodard. 2018. What represents "style" in authorship attribution? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2814–2822, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Yinfei Yang, Steve Yuan, Daniel Cer, Sheng-yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning semantic textual similarity from conversations. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 164–174, Melbourne, Australia. Association for Computational Linguistics.

Jian Zhu and David Jurgens. 2021. Idiosyncratic but not arbitrary: Learning idiolects in online registers reveals distinctive yet consistent individual styles. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 279–297, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Results on the Development Set

A.1 Hyperparameter Tuning

We evaluated contrastive (on the binary AV training task), triple (on the TAV training task) and online contrastive loss (on the binary AV training task) using implementations from `Sentence-Transformers`. We experiment with the loss hyperparameter value margin 0.4, 0.5, 0.6 for the uncased BERT model (Devlin et al., 2019) on the domain training data. Results are displayed in Figure 6. Contrastive and triplet loss perform better than online contrastive loss. The margin value only has a small influence on the performance scores. Based on these results, we decided to run all further models only with the contrastive and triplet loss functions and a margin value of 0.5.

	conversation		domain		random	
	TAV acc	AV AUC	TAV acc	AV AUC	TAV acc	AV AUC
c 0.4	0.63	0.63	**0.68**	**0.68**	**0.71**	**0.71**
c 0.5	0.63	0.63	**0.68**	**0.68**	**0.71**	**0.71**
c 0.6	0.62	0.63	**0.68**	**0.68**	**0.71**	**0.71**
t 0.4	0.63	0.62	0.68	0.67	0.70	0.70
t 0.5	**0.64**	**0.64**	**0.68**	**0.68**	0.70	0.70
t 0.6	0.63	0.63	0.67	0.67	0.70	0.70
c-on 0.4	0.58	0.58	0.64	0.64	0.67	0.67
c-on 0.5	0.58	0.58	0.64	0.64	0.67	0.67
c-on 0.6	0.58	0.58	0.64	0.64	0.67	0.67

Table 6: **Hyperparameter-tuning Results on the dev TAV datasets with varying topic proxies.** Results for BERT uncased trained on the triple authorship verification tasks (TAV). With different loss functions (contrastive - c, triple - t, contrastive online - c-on) and margin values (0.4, 0.5, 0.6). For each dev set (conversation, domain and random), we display the accuracy of the models for the triple authorship verification task (TAV) and the AUC for the binary authorship verification task (AV). For each dev set and TAV/AV task, the best performance is boldfaced. Contrastive and Triple loss behave comparable. The margin value only has a small influence.

		conv		sub		rand	
		TAV	AV	TAV	AV	TAV	AV
		acc	AUC	acc	AUC	acc	AUC
-	bert	0.52	0.51	0.59	0.57	0.64	0.61
	BERT	0.53	0.52	0.59	0.57	0.63	0.60
	RoBERTa	0.53	0.53	0.58	0.57	0.63	0.61
c	bert c 0.5	0.65	0.66	0.66	0.67	0.68	0.68
	bert t 0.5	0.65	0.66	0.66	0.67	0.67	0.68
	BERT c 0.5	0.66	0.67	0.67	0.68	0.69	0.70
	BERT t 0.5	0.66	0.67	0.67	0.68	0.68	0.69
	RoBERTa c 0.5	**0.69**	**0.70**	0.70	0.71	0.70	0.72
	RoBERTa t 0.5	0.68	0.69	0.69	0.70	0.70	0.70
s	bert c 0.5	0.63	0.63	0.68	0.68	0.71	0.71
	bert t 0.5	0.64	0.64	0.68	0.68	0.70	0.70
	BERT t 0.5	0.65	0.65	0.68	0.68	0.71	0.71
	BERT c 0.5	0.64	0.65	0.69	0.69	0.71	0.72
	RoBERTa c 0.5	0.67	0.68	**0.71**	**0.72**	0.73	0.74
	RoBERTa t 0.5	0.68	0.68	0.70	0.70	0.72	0.73
r	bert c-0.5	0.55	0.54	0.63	0.62	0.76	0.76
	bert t-0.5	0.55	0.54	0.62	0.61	0.74	0.75
	BERT c 0.5	0.57	0.56	0.64	0.63	0.76	0.77
	BERT t 0.5	0.58	0.56	0.64	0.62	0.75	0.75
	RoBERTa c 0.5	0.59	0.58	0.65	0.64	**0.77**	**0.78**
	RoBERTa t 0.5	0.59	0.57	0.65	0.63	**0.77**	0.77

(a) TAV and AV Performance

conv		sub		rand	
AV		AV		AV	
thr	acc	thr	acc	thr	acc
0.82	0.51	0.70	0.55	0.69	0.58
0.86	0.51	0.85	0.55	0.85	0.58
0.96	0.52	0.97	0.55	0.97	0.58
0.72	0.61	0.73	0.62	0.73	0.63
0.27	0.61	0.27	0.62	0.29	0.63
0.24	0.62	0.28	0.63	0.26	0.64
0.72	0.62	0.73	0.63	0.73	0.64
0.72	**0.64**	0.72	0.64	0.73	0.65
0.30	0.63	0.31	0.64	0.32	0.64
0.73	0.59	0.73	0.63	0.73	0.65
0.16	0.60	0.19	0.63	0.19	0.64
0.20	0.61	0.27	0.63	0.23	0.65
0.74	0.60	0.74	0.64	0.72	0.66
0.72	0.63	0.72	**0.65**	0.72	0.67
0.22	0.63	0.24	**0.65**	0.19	0.66
0.76	0.53	0.77	0.58	0.74	0.69
0.14	0.53	0.37	0.57	0.24	0.68
0.40	0.54	0.35	0.59	0.23	0.69
0.74	0.54	0.76	0.59	0.74	0.69
0.80	0.56	0.77	0.60	0.74	**0.71**
0.38	0.55	0.34	0.59	0.19	0.66

(b) Details on the AV results

Table 7: **(Dev) Results on the triple task.** We display the accuracy of the models for the triple authorship verification task (TAV) and the AUC for binary authorship verification task (AV) on each dev set (conversation, domain and random). We show results for 18 fine-tuned models: BERT uncased (bert), RoBERTa and BERT cased trained with the conversation, domain and random topic proxy. With different loss functions (contrastive - c, triple - t, contrastive online - c-on) and margin values (0.4, 0.5, 0.6). For the AV task, we also display the optimal threshold according to AUC (thr) and its matching accuracy. Generally, RoBERTa models perform the best with increasing performance from conversation to domain to random. Accuracies for the TAV are higher than for AV. Models perform the best on the task they have been trained on. Contrastive and Triple loss seem to behave comparable. Best performance per dev set and TAV/AV task is boldfaced.

A.2 Detailed Dev Results

We display the performance of further fine-tuned models on the dev sets in Table 7. RoBERTa generally performs better than the uncased and cased BERT model (Devlin et al., 2019). Performance for the triplet and contrastive loss functions are comparable. We only use RoBERTa models in the main paper and both contrastive and triplet loss as a result.

train data	model	all		formal		complex		nb3r		c'tion	
		STEL	t-a	STEL	t-a	STEL	t-a	STEL	t-a	STEL	t-a
-	BERT uncased (bert)	0.75	0.03	0.76	0.05	0.70	0.00	**0.93**	0.09	1.00	0.00
	BERT cased (BERT)	**0.78**	0.05	0.80	0.10	**0.71**	0.00	0.92	**0.11**	1.00	0.00
conv.	bert c 0.5	0.68	0.21	0.72	0.40	0.59	0.07	0.73	0.06	1.00	0.01
	bert t 0.5	0.68	0.30	0.71	0.52	0.61	0.15	0.72	0.05	0.99	0.06
	BERT c 0.5	0.73	0.32	0.83	0.62	0.60	**0.19**	0.67	0.06	1.00	0.00
	BERT t 0.5	0.73	**0.37**	0.79	**0.66**	0.63	0.15	0.74	0.05	1.00	**0.15**
domain	bert c 0.4	0.70	0.12	0.76	0.26	0.61	0.01	0.72	0.02	1.00	0.00
	bert c 0.5	0.69	0.13	0.74	0.27	0.59	0.01	0.68	0.05	1.00	0.00
	bert c 0.6	0.70	0.13	0.76	0.26	0.61	0.01	0.72	0.04	1.00	0.00
	bert c-on 0.4	0.65	0.02	0.67	0.03	0.60	0.00	0.69	0.02	0.84	0.00
	bert c-on 0.5	0.65	0.02	0.67	0.03	0.60	0.00	0.69	0.02	0.84	0.00
	bert c-on 0.6	0.65	0.02	0.67	0.03	0.60	0.00	0.69	0.02	0.84	0.00
	bert t 0.4	0.71	0.15	0.78	0.31	0.59	0.01	0.78	0.05	1.00	0.00
	bert t 0.5	0.68	0.18	0.74	0.37	0.58	0.03	0.72	0.06	1.00	0.00
	bert t 0.6	0.69	0.22	0.76	0.44	0.58	0.04	0.69	0.06	1.00	0.00
	BERT c-0.5	0.73	0.23	0.82	0.48	0.61	0.02	0.77	0.03	1.00	0.00
	BERT t-0.5	0.71	0.28	0.81	0.56	0.57	0.06	0.80	0.04	1.00	0.00
random	bert c 0.5	0.69	0.09	0.77	0.20	0.58	0.01	0.68	0.02	0.98	0.00
	bert t 0.5	0.70	0.13	0.75	0.26	0.61	0.03	0.79	0.06	1.00	0.00
	BERT c-0.5	0.72	0.21	**0.84**	0.44	0.55	0.02	0.75	0.07	1.00	0.01
	BERT t-0.5	0.73	0.23	**0.84**	0.48	0.59	0.03	0.68	0.05	1.00	0.00

Table 8: **Results on STEL and topic-adapted STEL.** We display STEL accuracy for different language models and methods. The performance on the set of task instances with (t-a) and without topic-adaption (STEL) is displayed. The best performance is boldfaced. Performance for the trained models goes down for the original STEL framework in the complex/simple and nb3r substitution dimension. Performance generally increases for the topic-adapted STEL task.

B Details on STEL results

We display the STEL results on further trained models in Table 8. Interestingly, cased BERT seems to be the better choice for the contraction STEL dimension.

aggregate		unlearned f/i	unlearned c/s	learned f/i	learned c/s
topic	conversation	21	34	62	22
	domain	13	34	62	24
	random	21	44	67	24
loss	contrastive	8	9	61	11
	triplet	6	14	55	14
-	all	1	4	48	8

Table 9: **Error Analysis STEL Results.** For the formal/informal (f/i) and complex/simple (c/s) STEL dimension, we display the number of instances that were unlearned and learned by all RoBERTa models in an aggregate. We use three different aggregates: (i) all models trained with a given topic proxy, (ii) all models trained with a certain loss function and (iii) all models.

	unlearned	learned
no ambiguity	$\frac{5}{55} \approx 9\%$	$\frac{12}{41} \approx 29\%$
typo simple	$\frac{21}{55} \approx 38\%$	$\frac{13}{41} \approx 32\%$
typo complex	$\frac{11}{55} \approx 20\%$	$\frac{6}{41} \approx 15\%$
error grammar simple	$\frac{15}{55} \approx 27\%$	$\frac{9}{41} \approx 22\%$
error grammar complex	$\frac{5}{55} \approx 9\%$	$\frac{3}{41} \approx 7\%$
changed content	$\frac{5}{55} \approx 9\%$	$\frac{3}{41} \approx 7\%$
word as/more complex	$\frac{16}{55} \approx 29\%$	$\frac{11}{41} \approx 27\%$
naturalness	$\frac{7}{55} \approx 13\%$	$\frac{3}{41} \approx 7\%$

Table 10: **Categories Error Analysis STEL Results.** For the six fine-tuned RoBERTa models, we manually looked at the common learned as well as the unlearned simple/complex examples. We put the examples in the displayed ambiguity classes.

n	avg. silhouette
2	0.23
3	0.21
4	0.23
5	**0.27**
6	**0.27**
7	0.26
8	0.23
9	0.19
10	0.20
11	0.19
12	0.18
13	0.19
14	0.17
15	0.16
16	0.16
17	0.16
18	0.17
19	0.17
20	0.17
21	0.16
22	0.16
23	0.15
24	0.15
25	0.15
26	0.15
30	0.15
40	0.15
50	0.15
100	0.13
150	0.13
200	0.12

Table 11: **Silhouette values.** We experiment with different numbers of clusters for one fine-tuned RoBERTa model (R TAV CONV 106). It was on the TAV task with the conversation topic proxy. The highest Silhouette score is reached for a cluster sizes of 5-7.

B.1 Error Analysis RoBERTa STEL results

In Table 9, we display the number of learned and unlearned STEL instances across different aggregates for the RoBERTa models. We combine all such unique STEL instances across the aggregates and annotate if they contain ambiguities. In Table 10, we display the results. Overall, the learned STEL instances contain fewer ambiguities. However, they still show considerable amounts of ambiguities.

C Details on cluster parameters

We do agglomerative clustering for the RoBERTa model trained on the triplet loss with a margin of 0.5 and conversations as topic proxy with seed 106 (R TAV CONV 106) with different number of clusters. We display the results in Table 11. The highest Silhouette scores are reached for cluster sizes of 5, 6, 7. We select a cluster size of 7 for evaluation as we expect a difference of 0.01 in Silhouette scores not to make a too big difference.

D Details on the cluster analysis

We give more examples of the seven clusters in Table 12. Refer to our Github repository for the complete clustering. We did not find obvious consistencies for clusters 1, 2 and 6. That does, however, not mean that more nuanced stylistic consistencies are not present. We recommend using a higher number of clusters, possibly different clustering algorithms and testing out statistics for known style features to pinpoint more consistencies.

Out of all utterance pairs that have the same author, 46.2% appear in the same cluster for the style embedding model. This is different from a random distribution among 7 clusters[10] which corresponds

[10]Calculated mean and standard deviation of 100 random assignments of utterances to the 7 clusters, with the same number of elements in each cluster.

to $20.1\% \pm .0$. As authors will have a certain variability to their style as well (e.g., Zhu and Jurgens (2021)), a perfect clustering according to writing style would not assign all same author pairs to the same cluster. For the RoBERTa base model the fraction of same author pairs in the same cluster is closer to the random distribution (75.4% vs. 76.1% for the random distribution[11]). The fraction of utterance pairs that appear in the same domain are close to the random distribution for both the style embedding model (23.6% vs. 20.1%) and the RoBERTa base model (77.6% vs. 76.0%). The percentage for the RoBERTa base models is a lot higher as the first cluster contains almost 90% of all utterances. Random assignment of utterances across the 7 clusters, that keeps the clustering size would already lead to 76.0% same author pairs appearing in the same cluster (almost all of them in the first). Results are similar for utterance pairs that appear in the same conversation.

[11]The share is high for RoBERTa base because the first cluster already contains 86.7% of all utterances.

C	#	Consistency	Example 1	Example 2	Example 3
1	4065	citing previous comments, standard punctuation, URLs	Yes. Proportionally, this kid's feet are absolutely enormous.	> Please delete your account. Says the no life who always shits on anything Kanye or anti-Drake I can promise you that capitalism is very much alive in Norway.	[This should help.](YOUTUBE-LINK)
2	4016	short sentences?	Nice catch! Well done. cookies are in the back of this Grammar party. You can have two.	You can mute them we've been told!	Came here to post this only to find it's already the top voted comment. This is a good sub.
3	2165	no last punctuation mark	I am living in china, they are experiencing an enormous baby boom	Seems like sarcasm. But could also be Poe	[...] The earth probably has two or more degrees of symmetry, but less than infinite (like a sphere), but I'm honestly not too concerned about the minutiae of it
4	1794	punctuation / casing	huh thats odd i'm in the 97% percentile on iq tests, the sat, and the act	Its not a problem if you a got a full game. Whats the problem if a game didnt get expansions?	Fair point, I didnt know that. Just at glance I kind of went 'woah that doesnt seem right'
5	1555	' instead of ' apostrophe	I assume it's the blind lady?	Oh I wasn't really dismissing them. I'm saying Ford will try their own thing compared to Fiat	It's 4am in Brussels and I am still hyped
6	781	similar to 1?	Well, as your neighbors, I'd say Fuck you.. But we're not like that, see? We want to be part of the alliance, not part of the 'fuck you, we cant be competitive with jobs or innovate any more, so we're going to run massive tariffs against all our friendly nations	Hah, thus the one calf larger than the other issue. I have it too ;)	[So you are saying that current encryption falls apart as long as the quantum computer is large enough](URL). (for reference, the current highest qubit is 50)'
7	380	linebreaks	I admire what you're doing but [...] I know I'm in the minority. [...]	75% of the problems I run into are solved by [...] I work in live streaming.	All the suggestions others have given are excellent. RS7 makes the most sense to me. But [...] Meanwhile, [...]

Table 12: **Clustering - fined-tuned RoBERTa model.** We display examples for each cluster of the 7 clusters that resulted from the agglomerative clustering of 14,756 randomly sampled texts with the RoBERTa model fine-tuned on the TAV task with the conversation topic proxy. We mention noticeable consistencies (Consistency) within the cluster and give three examples each.

C	#	Consistency	Example 1	Example 2	Example 3
1	12798	wide variety	Just googled it, looks like a great device for the price! If I weren't so impatient I would have bought this online. Great battery life!	This is exactly why i believe iphone 5 body was perfect example of good balance with design(timeless) and utility	[...] The earth probably has two or more degrees of symmetry, but less than infinite (like a sphere), but I'm honestly not too concerned about the minutiae of it
2	1110	short utterances	here we go!!	And her good posture.	Not in California.
3	310	long utterances	I've never had the pleasure of seeing Neil live but I got on a big kick a few years ago after buying one of his live albums (can't remember which one) where I listened to all his live albums and then wanted to see as many of his live performance I could find on YouTube. [...]	> but the movie has the superior ending I think. [...] [...]	So .... heavily influenced by the social economics ... but still voluntary, got it. [...] Then how about this. [...] Everyone still keeps their child that way, you even promote child birth. No sterilization, no stigmatization of poor people, no poor people stuck with child with heavy needs requiring care that they can't pay for.
4	232	URLs	`https://youtu.be/GmULc5VANsw`	[This](`https://np.reddit.com/r/MakeupAddiction/comments/25hkqi/how_to_tell_if_your_foundationprimer_is_silicone/`) might help!	I thought there was 51 stars because of Puerto Rico `https://en.m.wikipedia.org/wiki/51st_state`

Table 13: **Clusters for RoBERTa base.** We display examples for 4 out of 7 clusters as a result of the agglomerative clustering of 14756 randomly sampled texts from the conversation test set. We mention noticeable consistencies (Notice) within the cluster and give three examples each.

E Computing Infrastructure

The training of 23 RoBERTa (Liu et al., 2019), 13 uncased BERT and 6 cased BERT models (Devlin et al., 2019) took about 846 GPU hours with one RTX6000 card with 24 GB RAM on a Linux computing cluster. Further analysis and clustering of two RoBERTa models took about 24 GPU hours. We used a machine with 32 GB RAM and 8 intel i7 CPUs using Ubuntu 20.04 LTS without GPU access to generate the training data.

We use `Sentence-Transformers` 2.1.0 (Reimers and Gurevych, 2019) and `numpy` 1.18.5 (Harris et al., 2020), `scipy` 1.5.2 (Virtanen et al., 2020) and `scikit-learn` 0.24.2 (Pedregosa et al., 2011).

We use previous work, including code and data, consistent with their specified or implied intended use (Reimers and Gurevych, 2019; Chang et al., 2020; Wegmann and Nguyen, 2021). The `ConvoKit` open-source Python framework invites NLP researchers and 'anyone with questions about conversations' to use it (Chang et al., 2020). The `SentenceTransformers` Python framework can be used to compute sentence / text embeddings.[12] We comply with asking permission for part of the dataset for `STEL` and citing the specified works (Wegmann and Nguyen, 2021). Wegmann and Nguyen (2021) state the intended use of developing improved style(-sensitive) measures.

F Intended Use

We hope our work will inform further research into style and its representations. We invite researchers to reuse any of our provided results, code and data for this purpose.

[12] https://sbert.net/