Does It Capture STEL? A Modular, Similarity-based Linguistic Style Evaluation Framework

Anna Wegmann and Dong Nguyen
Department of Information and Computing Sciences
Utrecht University
Utrecht, the Netherlands
a.m.wegmann@uu.nl, d.p.nguyen@uu.nl

Abstract

Style is an integral part of natural language. However, evaluation methods for style measures are rare, often task-specific and usually do not control for content. We propose the modular, fine-grained and content-controlled *similarity-based STyle EvaLuation framework* (STEL) to test the performance of any model that can compare two sentences on style. We illustrate STEL with two general *dimensions* of style (formal/informal and simple/complex) as well as two specific *characteristics* of style (contrac'tion and numb3r substitution). We find that BERT-based methods outperform simple versions of commonly used style measures like 3-grams, punctuation frequency and LIWC-based approaches. We invite the addition of further tasks and task instances to STEL and hope to facilitate the improvement of style-sensitive measures.

1 Introduction

Natural language is not only about what is said (i.e., content), but also about how it is said (i.e., *linguistic style*). Linguistic style and social context are highly interrelated (Coupland, 2007; Bell, 2013). For example, people can accommodate their linguistic style to each other based on social power differences (Danescu-Niculescu-Mizil et al., 2012). Furthermore, linguistic style can influence perception, e.g., the persuasiveness of news (El Baff et al., 2020) or the success of pitches on crowdsourcing platforms (Parhankangas and Renko, 2017). As a result, style is relevant for natural language understanding, e.g., in author profiling (Rao et al., 2010), abuse detection (Markov et al., 2021) or understanding conversational interactions (Danescu-Niculescu-Mizil and Lee, 2011). Additionally, style can be important to address in natural language generation (Ficler and Goldberg, 2017), including identity modeling in dialogue systems (Li et al., 2016) and style preservation in machine translation (Niu et al., 2017; Rabinovich et al., 2017).

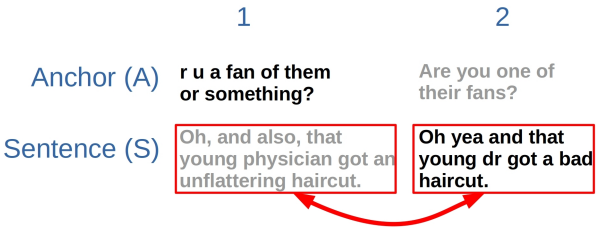

Figure 1: **STEL Task Instance.** Anchor 1 (A1) and anchor 2 (A2) and the alternative sentences 1 (S1) and 2 (S2) are split along the same style dimension (here: formal/informal). The sentences and anchors are paraphrases of each other. The STEL task is to order S1 and S2 to match A1-A2. Here, the correct order is S2-S1.

There are several general evaluation benchmarks for different linguistic phenomena (e.g., Wang et al. (2018, 2019)) but less emphasis has been put on linguistic style. Nevertheless, natural language processing literature shows a variety of approaches for the evaluation of style measuring methods: They have been tested on whether they group texts by the same authors together (Hay et al., 2020; Bevendorff et al., 2020), whether they can correctly classify the style for ground truth datasets (Niu and Carpuat, 2017; Kang and Hovy, 2021) and whether 'similar style words' are similarly represented (Akama et al., 2018). However, these evaluation approaches are (i) often application-specific, (ii) rarely used to compare different style methods, (iii) usually do not control for content and (iv) often do not test for fine-grained style differences.

These shortcomings (i)-(iv) might be the result of the following challenges for the construction of style evaluation methods: 1. Style is a highly ambiguous and elusive term (Biber and Conrad, 2009; Crystal and Davy, 1969; Labov, 2006; Xu, 2017). We propose a *modular* framework where components can be removed or added to fit an application or specific understanding of style. 2. Variation in style can be very small. Our proposed evaluation framework can be used to test for *fine-grained* style

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7109–7130
November 7–11, 2021.

differences. 3. Style is hard to disentangle from content as the two are often correlated (e.g., Gero et al. (2019); Bischoff et al. (2020)). For example, people might speak more formally in a job interview than in a bar with friends. Thus, language models and methods might pick up on spurious content correlations (similar to Poliak et al. (2018)) in a benchmark that does not *control* for *content*.

To this end, we propose the **modular, fine-grained and content-controlled *similarity-based STyle EvaLuation framework* (STEL)**. We demonstrate it for the English language. An example task is shown in Figure 1. The task is to order sentence 1 (S1) and sentence 2 (S2) to match the style order of anchor 1 (A1) and anchor 2 (A2). Our STEL framework encompasses two general *dimensions* of style (formal/informal and simple/complex) as well as two specific *characteristics* of style (contraction and number substitution). By design, the style characteristics are easy to identify. Thus, the STEL characteristic tasks are easier to solve than the STEL dimension tasks. STEL contains 815 task instances per dimension and 100 task instances per characteristic (see Table 1). To be evaluated on STEL, style measuring methods need not be able to classify styles directly. Instead, any method that can calculate the style *similarity* between two sentences can be evaluated: (1) Style (measuring) methods that calculate similarity values directly (e.g., edit distance or cross-encoders Reimers and Gurevych (2019)) and (2) vector representations of a sentence's style (e.g., Hay et al. (2020); Ding et al. (2019)) by using a distance or similarity measure between them (e.g., cosine similarity). This similarity-based setup also simplifies task extension (c.f. modularity). STEL components can easily be generated from parallel sets of paraphrases which differ along a style dimension (§3), e.g., sets of paraphrases that vary along the formal/informal dimension (Rao and Tetreault, 2018).

Contribution. With this paper, we contribute **(a)** the modular, fine-grained and content-controlled STEL framework (§3), **(b)** 1830 validated task instances for the considered style components (§4) and **(c)** baseline results of STEL on 18 style measuring methods (§5). We find that the BERT base model outperforms simple versions of commonly used style measuring approaches like LIWC, punctuation frequency or character 3-grams. We invite the addition of complementary tasks and hope that this framework will facilitate the development of improved style-sensitive models and methods. Our data and code are available on GitHub.[1]

2 Related Work

Linguistic style has been analyzed from different perspectives and along different dimensions. A speaker's style can, for example, be influenced by the situation, the speaker's choices, and/or the speaker's identity (Preoţiuc-Pietro et al., 2016; Flekova et al., 2016; Nguyen et al., 2016; Bell, 1984). In NLP, previously analyzed style dimensions include formal/informal, simple/complex, abstract/concrete and polite/impolite (Pavlick and Nenkova, 2015; Pavlick and Tetreault, 2016; Paetzold and Specia, 2016; Brooke and Hirst, 2013; Madaan et al., 2020).

Linguistic style is usually defined to be distinct from content. However, style is often found to be correlated with content (e.g., Gero et al. (2019)). To address this, some control for content with word-level paraphrases (Pavlick and Nenkova, 2015; Preoţiuc-Pietro et al., 2016; Niu and Carpuat, 2017), topic labels (e.g., Boenninghoff et al. (2019)) or by avoiding the use of content-specific features (c.f. Neal et al. (2017); Stamatatos (2017)), others choose no or only limited control for content (e.g., Zangerle et al. (2020); Kang and Hovy (2021)). There has been considerable work in creating parallel datasets of (sentence level) paraphrases with shifting style, often using human annotations (Xu et al., 2012, 2016; Rao and Tetreault, 2018; Krishna et al., 2020). The task of generating paraphrases of text fragments with different style properties is sometimes also called *style transfer*.

There is little work on general evaluation benchmarks for style measuring methods. Kang and Hovy (2021) use style classification tasks to compare 5 language models. Only models that classify style into the given 15 dimensions can be evaluated. They do not control for content. Individually fine-tuned RoBERTa (Liu et al., 2019) and BERT (Devlin et al., 2019) classifiers for one style were outperformed by a fine-tuned T5 model (Raffel et al., 2020) that was jointly trained on multiple style labels. BERT/RoBERTa outperformed the T5 model on some styles (e.g., 'sarcasm' and 'methaphor'). Other related tasks are the PAN *Authorship Verification* (Kestemont et al., 2020) and *Style Change Detection* (Zangerle et al., 2020) tasks which aim at identifying whether two documents or consec-

[1] https://github.com/nlpsoc/STEL

Comp	Size	Order	Anchor 1 (A1)	Anchor 2 (A2)	Sentence 1 (S1)	Sentence 2 (S2)
formal/-informal	815	S1-S2	Are you one of their fans?	r u a fan of them or something?	Oh, and also, that young physician got an unflattering haircut.	Oh yea and that young dr got a bad haircut.
simple/-complex	815	S1-S2	These rock formations are made of sandstone with layers of quartz.	These rock formations are characteristically composed of sandstone with layers of quartz.	The Odyssey is an ancient Greek epic poem attributed to Homer.	The Odyssey is an old Greek epic poem written by Homer.
number substitution	100	S2-S1	<3 friends forever	<3 friends 4ever	D00d $30 is heaps cheap, that must work out to just a couple of bucks an hour	Dude $30 is heaps cheap, that must work out to just a couple of bucks an hour
contraction	100	S1-S2	In that time, it's become one of the world's most significant financial and cultural capital cities.	In that time, it has become one of the world's most significant financial and cultural capital cities.	Will doesn't refer to any particular desire, but rather to the mechanism for choosing from among one's desires.	Will does not refer to any particular desire, but rather to the mechanism for choosing from among one's desires.

Table 1: **STEL Examples.** We give an example for each component (Comp) of STEL: Formal/informal and simple/complex for the more complex style dimensions as well as number substitution and contraction for the simpler style characteristics. The task is to order sentence 1 (S1) and sentence 2 (S2) to match the style order of anchor 1 (A1) and anchor 2 (A2). The correct order is given in the 'Order' column.

utive paragraphs have been written by the same author. In their current version both tasks do not control for topic. However, Kestemont et al. (2020) controls for domain (here: 'fandom' of the considered 'fanfictions'). The best performing model for Kestemont et al. (2020) was a neural LSTM-based siamese network (Boenninghoff et al., 2020), which is conceptually similar to some variants of sentence BERT (Reimers and Gurevych, 2019). The PAN setup assumes that authors tend to write in a relatively consistent style. Based on similar assumptions, the field of *authorship attribution* wants to determine which author wrote a given document.

Especially in authorship attribution, recurring style features include character n-grams, punctuation, average word length or function word frequency (Neal et al., 2017; Grieve, 2007; Stamatatos, 2009). Other recurring methods for style measurement include LIWC (Pennebaker et al., 2015; Danescu-Niculescu-Mizil et al., 2011; El Baff et al., 2020), and learned vector representations of words and sentences (Akama et al., 2018; Ding et al., 2019; Hay et al., 2020). Niu and Carpuat (2017) suggests that style variations are already represented in commonly used neural embeddings.

Binary and more fine-grained style classification has been employed on word, text fragment as well as document level (Danescu-Niculescu-Mizil et al., 2013; Pavlick and Nenkova, 2015; Preoţiuc-Pietro et al., 2016; Pavlick and Tetreault, 2016; Kang et al., 2019). Traditionally, considered documents in authorship attribution were longer than 1,000 words (e.g., Eder (2013)), but recently there has been increased interest in text fragments with fewer than 300 words (e.g., Brocardo et al. (2013); Boenninghoff et al. (2019)).

3 Style Evaluation Framework

We introduce the modular, fine-grained, and content-controlled similarity-based STyle EvaLuation framework (STEL). STEL tests a (language) model's ability to capture the style of a sentence.

Modular Operationalization of Style. Style has previously been conceptualized in many different ways. From being defined as purely aesthetic in Biber and Conrad (2009) to encompassing all forms of language variation, e.g., in Crystal and Davy (1969). We refrain from meddling in the style definition debate and instead use the broad notion of "how vs. what", i.e., how something is said as opposed to what is said. Inspired by Campbell-Kibler et al. (2006), we use different *characteristics* (i.e., more specific linguistic choices) as well as more general *dimensions* of style (i.e., more complex combinations of style features). By not only using complex style dimensions, but also small scale and simpler characteristics, STEL allows for very controlled and **fine-grained**

testing. We can easily make sure that only the characteristics and no other aspects change (c.f. Table 1). Depending on one's goal and understanding of style, some *components* (i.e., dimensions or characteristics) should be excluded and others should be added to this modular framework. We exemplify the framework's more complex dimensions with the formal/informal distinction as this has been one of the most agreed upon dimensions of style (Heylighen et al., 1999; Labov, 2006). Additionally, we use the simple/complex dimension which has been used in connection to linguistic-stylistic choices as well (Haaften and Leeuwen, 2021; Pavlick and Nenkova, 2015). We exemplify the framework's simpler style dimensions (i.e., characteristics) with numb3r substitutions and contraction usage. See Table 1 for examples for each component.

Controlling for Content. It is difficult to clearly separate style from content (Stamatatos, 2017; Gero et al., 2019). Specific scenarios might correlate with both style and content. For example, in a job interview applicants might use a more formal style and talk more about their profession than in a more informal setting at a bar. Then, a model that generally rates texts about jobs as formal and texts about beverage choices as informal might perform well at style prediction. In other words, models that correctly use style features could sometimes be indistinguishable from those that use topical features. To control for content, we use parallel paraphrase datasets (§4.1), which consist of a set of sentences written in one style and a parallel set of sentences written in another.

Task Setup. We test a method's style measuring capability with tasks of the setup shown in Figure 1. The sentences (S1 and S2) have to be ordered to match the order of the anchor sentences (A1 and A2). Here, 'r u' (A1) and 'Oh yea' (S2) are written in a more informal style than their respective paraphrases A2 ('Are you') and S1 ('Oh, and also'). Thus, the correct order is S2, then S1. We call this setup the *quadruple setup*. Additionally, we explore a second task setup, the *triple setup*, which leaves out anchor 2 (A2). There, the task is to decide which of the two sentences matches the style of anchor 1 (A1) the most. The two different setups are similar to the triple and quadruple training instances in the field of metric learning, e.g., (de Vazelhes et al., 2020; Law et al., 2016; Kaya and Bilge, 2019).

4 Task Generation

We describe the task instances of STEL: First, we generate potential task instances (§4.1). Second, we describe problems with the generated instances (i.e., ambiguity in §4.2). Third, we filter out the problematic instances via crowd-sourcing (§4.3).

4.1 Potential Task Instances

We generate potential task instances on the basis of parallel paraphrase datasets written in style 1 and style 2 respectively. For each style 1/style 2 paraphrase pair (anchors in Table 1), we randomly select another sentence pair (sentences in Table 1). Again randomly, we decide which of the anchor pair is anchor 1 (A1) and which is anchor 2 (A2) and fix that ordering for all future considerations. We do the same for the sentence pair. The answer to the STEL task (Figure 1) is labeled as S1-S2 if A1 was taken from the same style set as S1, e.g., both from style 1. Otherwise the order is reversed.

Formal/Informal Dimension. We use the test and tune split of the Entertainment_Music GYAFC subcorpus (Rao and Tetreault, 2018) as the parallel paraphrase dataset. It consists of a set of informally phrased sentences and a parallel set of crowd-sourced formal paraphrases. We generate 918 potential STEL formal/informal task instances.

Simple/Complex Dimension. We use the test and tune split from Xu et al. (2016). It consists of English Wikipedia sentences and 8 crowd-sourced simplifications per sentence. For each Wikipedia sentence, we randomly draw the parallel paraphrase out of the 8 simplifications. We discard sentences that are too close to the original via the character edit distance of 3 or lower. From this parallel paraphrase dataset, we generate 1195 potential STEL simple/complex task instances.

Contraction Characteristic. We generated the parallel contraction dataset from the December 2018 abstract dump of English Wikipedia[2]. The Wikipedia style guide discourages contraction usage and provides a dictionary with contractions that should be avoided.[3] We use an adapted version[4] to select 100 sentences where an apostrophe is present and a contraction is possible. Such a sentence could

[2] https://archive.org/details/enwiki-20181220

[3] https://en.wikipedia.org/wiki/Wikipedia:List_of_English_contractions

[4] See the Appendix for the filtered contraction list.

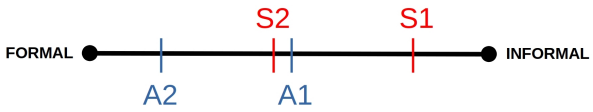

Figure 2: **Triple Problem.** Tasks are generated from sentence pairs (A1, A2) and (S1, S2) that are split along the same style dimension (e.g., formal/informal). For each pair, only the order on the axis (e.g., S2 < S1) but not the absolute localization is known. This might lead to a wrong generated label for the triple setup. Here, removing A2 leads to S2 being stylistically closer to A1, whereas the generated label would be S1.

be "It is near Thomas's car". For each sentence, we generate a parallel sentence with a contraction, e.g., "It's near Thomas's car", c.f. Table 1.

Number Substitution Characteristic. The character by number substitution task instances were semi-automatically generated from the Reddit comment corpus of the months May 2007-September 2007, June 2012, June 2016 and June 2017 taken from the Pushshift dataset (Baumgartner et al., 2020). We selected a pool of potential sentences where words contained character substitution symbols (4,3,1,!,0,7,5) or are part of a manually selected list of number substitution words (see Appendix). Then, we manually filtered out sentences without number substitutions (e.g., common measuring units or product numbers). We selected 100 sentences, 50 of which were selected to contain at least one additional number that is not part of a number substitution word (e.g., Anchor 1 in Table 1). This setup ensures that the task is not as simple as checking whether there are numbers present in the sentence. To generate the parallel phrases, we manually translated the sentences to contain no number substitutions. As we looked for naturally occurring number substitution words, we decided to keep word pairs that contain additional changes besides number substitution. For example, generally different spelling (e.g., 'd00d', 'dude') or phonetic spelling (e.g., 'str8', 'straight'). We decided to replace the number substitution symbols with characters only — e.g., not with punctuation marks as seen in 's1de!!!!!1' -> 'side!!!!!1'.

4.2 Ambiguity

Manual inspection shows that the generated potential task instances of the formal/informal and simple/complex dimension contain ambiguities: (i) Some are a result of unclear or very fine distinctions between the two parallel styles in the original data. For example, *"Each band member chose an individual number as their alias towards the end of 1997."* and *"Towards the end of 1997, each band member chose an individual number as their alias."*. The first is labelled as written in a simpler style in Xu et al. (2016). However, putting "towards the end of 1997" at the beginning of the sentence could also be understood as structuring the sentence more clearly, and thus as simpler. After manual inspection, ambiguities seem to be more prevalent for the simple/complex than the formal/informal dimension. (ii) Other ambiguities are the result of entangled additional linguistic components. For example, consider the potential task instance (A1) *"He's supposed to be in jail!"*, (A2) *"I understood he was still supposed to be incarcerated."* and (S1) *"green day is the best i think"*, (S2) *"I think Green Day is the best."*. The sentences are clearly split along the formal/informal dimension leading to the label S1-S2. Still, (A1) and (S2) could also be understood as being written in a more decisive tone than (A2) and (S1) leading to the order S2-S1.

We find that the triple setup has additional theoretical limitations that can lead to ambiguity: Consider the 'Triple Problem' in Figure 2 where S2 is labelled as formal and A1 is labelled as informal. Removing A2, to get from the quadruple to a triple setup, will leave A1 closer to S2, contrary to the original labelling (see also the previous example in (ii)). Additionally, having fewer sentences in the triple setup increases the chance of a random correlation with a different linguistic component (similar to the 'decisive tone' in example (ii)).

4.3 Removing Ambiguity

Using crowd-sourced annotations, we filter the previously discussed ambiguity out of the potential formal/informal and simple/complex task instances. The simpler STEL tasks (contraction and number substitution) mostly differ in the amount of apostrophes and numbers (e.g., Table 1). As a result, we expect the style characteristics to contain little to no ambiguity and do not filter those further.

Annotation Tasks. For both the triple and quadruple setup we collected annotations on a subsample of all generated task instances (301 simple/complex and formal/informal instances respectively). Then, we annotated a larger set of task instances on the quadruple setup alone. Based on performance results from the subsample ('Anno-

	Sample					Total		
	Triple		Quadruple		n	Quadruple		n
Dim	κ	acc.	κ	acc.		κ	acc.	
all	0.29	0.62	0.35	0.78	602	0.30	0.77	2113
c	0.19	0.51	0.16	0.68	301	0.17	0.68	918
f	0.39	0.74	0.51	0.89	301	0.48	0.90	1195

(a) Results on the sample and total of task instances

Triple	Quadruple	Dimension	Share
✗	✓	formal	0.196
		complex	0.312
✓	✗	formal	0.047
		complex	0.140
✗	✗	formal	0.066
		complex	0.179
✓	✓	formal	**0.691**
		complex	**0.369**

(b) Subsample analysis

Table 2: **Annotation Results.** We filter out ambiguous task instances via annotations. In (a), we display inter-annotator agreement (Fleiss's κ) and annotation accuracy (acc.) for the sample and total of potential task instances on the quadruple and triple setup for the simple/complex (c) and the formal/informal (f) dimensions. We also display the number of task instances per dimension (n). In (b), we display the share of all combinations of correct (✓) and wrong (✗) annotations per dimension and task setup. The union of ✓✓and ✗✗cases make up a majority.

tation Results'), we had 617 and 894 more task instances annotated for the formal/informal and simple/complex dimension respectively.

Annotation Setup. We used annotations from 839 different *Prolific*[5] crowdworkers with 5 distinct annotators per potential task instance. We paid participants 10.21£/h[6] on average. All annotators were native English speakers as we assume them to have a better intuition about their language. See the Appendix for further detail.

Annotator Agreement. In Table 2a, we report inter-annotator agreement with Fleiss's κ (Fleiss, 1971) as κ allows different items to be rated by different sets of raters. Inter-annotator agreement is only moderate. This does not mean that the annotations are of poor quality. As discussed in §4.2, our generated data contains ambiguous, noisy or faulty task instances. Manual inspection confirms that low annotator agreement is a sign of ambiguity (see also 'Annotation Analysis' Table in Appendix). This problem is more pronounced for the simple/complex than the formal/informal dimension. We ensured annotator quality with screening questions (appendix Table 1) and by selecting annotators with the highest platform-internal rating.

Annotation Results. Results are reported in Table 2a. Annotation accuracy is the share of correctly annotated task instances (by a majority of at least 3) out of all potential task instances.

The accuracy and the inter-annotator agreement are considerably higher for the formal/informal dimension (Table 2a) than for the simple/complex dimension. This aligns with our expectation of more ambiguity in the simple/complex task instances (c.f. §4.2(i)). Similarly, our expectations regarding theoretical problems with the triple setup (§4.2) are confirmed: Accuracy for the sample is generally higher for the quadruple than the triple setting. There are more examples where the quadruple setup was correctly annotated but the triple setup was not (✗✓in Table 2b), than there are for the opposite kind (✓✗).

As a consequence, the annotation of the bigger set of task instances was only done on the quadruple setup. On the total set of potential task instances (which includes the sample) we obtained similar accuracy and annotator agreement as on the sample (see Table 2a). We filter the potential task instances by only keeping those that were correctly annotated by a majority (i.e., at least 3/5). This leaves 822 task instances for the formal/informal and 815 for the simple/complex dimension. We randomly remove 7 task instances from the formal/informal dimension for equal representation of the two style dimensions. In the following, and under the name STEL, we will only consider the quadruple setup on the 1830 filtered task instances (i.e., 815, 815, 100 and 100 for simple/complex, formal/informal, number substitution and contraction respectively).

5 Evaluation

We use our STEL framework to test several models and methods that could be expected to capture style information (§5.1). We describe how the models decide the STEL tasks (§5.2) and discuss their performance on STEL (§5.3).

[5] https://www.prolific.co/

[6] above UK minimum wage of 8.91£/h at the time of the study (April 2021), see https://www.gov.uk/national-minimum-wage-rates

	all	formal		complex		nb3r	c'tion	random	
		filter	full	filter	full			filter	full
BERT uncased	**0.74**	**0.79**	0.77	**0.65**	0.63	0.90	0.90	**0**	0
BERT cased	**0.77**	**0.82**	0.81	**0.68**	0.64	**0.92**	**1.0**	**0**	0
RoBERTa	0.61	0.63	0.62	0.54	0.53	0.62	0.98	**0**	0
SBERT mpnet	0.61	0.64	0.62	0.53	0.52	0.71	0.84	**0**	0
SBERT para-mpnet	0.68	0.73	0.72	0.55	0.54	**0.95**	**1.0**	**0**	0
USE	0.59	0.59	0.58	0.55	0.52	0.58	0.85	**0.00**	0.00
BERT uncased NSP	0.66	0.72	0.71	0.59	0.57	0.67	0.70	0.10	0.12
BERT cased NSP	0.71	**0.79**	0.77	0.60	0.58	0.77	0.96	0.02	0.02
char 3-gram	0.55	0.58	0.57	0.52	0.50	0.50	0.64	0.05	0.05
word length	0.58	0.53	0.53	0.59	0.57	0.50	0.94	0.08	0.08
punctuation	0.56	0.58	0.58	0.50	0.49	0.50	0.92	0.38	0.39
LIWC	0.55	0.52	0.52	0.52	0.52	0.50	**0.99**	0.09	0.09
LIWC (style)	0.50	0.52	0.52	0.50	0.50	0.50	0.50	0.62	0.64
LIWC (function)	0.53	0.48	0.48	0.52	0.51	0.50	**1.0**	0.28	0.28
deepstyle	0.66	0.71	0.70	0.55	0.52	0.84	0.96	**0**	0
POS Tag	0.52	0.53	0.53	0.52	0.52	0.50	0.50	0.20	0.20
share cased	0.56	0.55	0.54	0.53	0.51	0.50	**1.0**	0.08	0.08
edit dist	0.54	0.56	0.56	0.52	0.51	0.50	0.39	0.08	0.07

Table 3: **STEL Results.** We display STEL accuracy for different language models and methods. Random performance is at 0.5. The share of task instances for which a method decides randomly as it can not decide between the two options ('=' in Equation 1) is given in the 'random' column. Both the performance on the set of task instances before (full) and after crowd-sourced filtering (filter) is displayed. The two best accuracies are boldfaced. The BERT-based models perform the best, followed by the "deepstyle" style sentence embedding method. On average, methods perform best for the c'tion and worst for the simple/complex dimension.

5.1 Style Measuring Methods

We describe methods and models that can be used to calculate a (style) similarity. Given two sentences, the methods return a similarity value between 0 and 1 or -1 and 1 (when using cosine similarity), where 1 represents the highest similarity.

Language Models. We use the base *BERT uncased* and base *BERT cased* model (Devlin et al., 2019). We calculate the mean over the subwords in the last hidden layer to generate two sentence embeddings. Then, we use cosine similarity to compare the sentences. We do the same with the cased *RoBERTa* base model (Liu et al., 2019). Additionally, we compare to the sentence BERT 'all-mpnet-base-v2' (*SBERT mpnet*)[7] and 'paraphrase-multilingual-mpnet-base-v2' (*SBERT para-mpnet*)[8] models (Reimers and Gurevych, 2019, 2020). Like BERT, MPNet uses a transformer architecture, but with a permuted instead of a masked language modeling pre-training task (Song et al., 2020). Furthermore, we experiment with the universal sentence encoder (*USE*) from Cer et al. (2018).

[7]best performing pre-trained sentence embedding in September 2021, see https://www.sbert.net

[8]best performing embedding trained on paraphrase data

Authorship Attribution Methods. The following methods are inspired by successful or commonly used approaches in authorship attribution (Neal et al., 2017; Sari et al., 2018). We use character 3-gram similarity by calculating the cosine similarity between the frequencies of all *character 3-grams*. We calculate the *word length* similarity via the average word lengths a and b of two sentences: $1 - |a - b|/\max(a, b)$. We calculate the *punctuation* similarity by using the cosine similarity between the frequencies of punctuation marks {',:,';',_,!,?,;,.,",(,),-}, taken from Sari et al. (2018).

LIWC-based Style Methods. LIWC categories have previously been used as style features (Niederhoffer and Pennebaker, 2002). We use LIWC 2015 (Pennebaker et al., 2015) for (a) *LIWC* similarity by taking the cosine similarity between the complete LIWC frequency vectors, (b) *LIWC (style)* similarity by taking the cosine similarity between the 8 dimensional binary LIWC style vectors (1 if a word of the category is present in the sentence, 0 otherwise) proposed in Danescu-Niculescu-Mizil et al. (2012), (c) *LIWC (function)* similarity by taking $1-$ the difference between the relative frequencies of function words. Function words have previously been used as a proxy for style (Neal et al., 2017).

Other Methods. We also experiment with the "*deepstyle*" model (Hay et al., 2020) by taking the cosine similarity between the style vector representations. Additionally, we consider the following sentence features: NLTK *POS Tags* (Bird et al., 2009) and *share of cased* characters (e.g., Sari et al. (2018)) via the cosine similarity between the frequency vectors and 1 - the difference between the proportion of cased characters respectively. We also include the *edit dist*ance as a simple baseline.

5.2 Similarity-based Decision

To determine an answer for a STEL task in the quadruple setup, the methods need to order two sentences (Figure 1). We do this by calculating the similarities (sim) between Anchor 1 (A1), Anchor 2 (A2), Sentence 1 (S1) and Sentence 2 (S2). We decide for the order S1-S2 if

$$\begin{aligned}&(1-\text{sim}(\text{A1},\text{S1}))^2+(1-\text{sim}(\text{A2},\text{S2}))^2 < \\ &(1-\text{sim}(\text{A1},\text{S2}))^2+(1-\text{sim}(\text{A2},\text{S1}))^2\end{aligned} \quad (1)$$

For the '$>$' case we use the order S2-S1, for '=' ordering is settled randomly (c.f., 'random' in Table 3). See the Appendix for a proof sketch after transforming similarities to distances.

5.3 Results

Performance results are shown in Table 3. The accuracy is a weighted mean of 0.5 (proportional to the share of undecided instances, c.f. 'random' in Table 3) and the accuracy in the decided cases. Random guessing would show an accuracy of 0.5 exactly. Stylistic differences can be subtle for the STEL dimensions and we expect this to be a hard task to solve. In contrast, the STEL characteristics (i.e., contraction and number substitution) should be easier to solve (via detecting an additional apostrophe or number) and are especially interesting for model error analysis. Note: We do not make general quality judgements because models were not trained on the components of STEL and were often not even meant to measure style directly.

The BERT base model encodes style information. The best performing cased BERT base model has an overall accuracy of 0.77. RoBERTa, a successor of BERT, includes BERT's training data (Liu et al., 2019). However, the cased RoBERTa base model does less well (0.61, $p < 0.001$ with McNemar's test (McNemar, 1947)). A possible explanation of RoBERTa's reduced performance might be the removal of the next sentence prediction (NSP) task. Closer sentences could generally be more similar in style than a different random sentence — possibly making the NSP a valuable learning objective for style similarity learning. To further look at this, we experiment with the *BERT NSP* head on the cased and uncased base model. For the quadruple setup, we calculate the four 'similarity' values as described in Equation 1 by using the predicted softmax probability that A1 is followed by S1 for sim(A1, S1). The other similarities are calculated equivalently. Interestingly, BERT's cased NSP head (accuracy of 0.71) performs better than RoBERTa ($p < 0.001$) across the STEL tasks. The effect of training objectives on learning style information could be explored in future work.

(Semantic) Sentence Embedding Methods perform well. SBERT para-mpnet (0.68) trained on the paraphrase data performs better than SBERT mpnet (0.61, $p < 0.001$) and USE (0.59, $p < 0.001$). Overall, SBERT para-mpnet is the third best performing model after the base BERT models and the best performing model in the nb3r dimension. In future work, it could be interesting to explore the effect of different training data on the performance of embedding models.

LIWC alone does not perform well. On the style dimensions LIWC performs similar to the random baseline. Possibly because the LIWC methods often find no difference between the two possible orderings (10%, 71% and 32% of tasks). The difference between the three LIWC-based methods is not significant ($p > 0.05$). Future work could explore models that consider more fine-grained differences between LIWC categories.

Authorship attribution methods perform better than random. Character 3-grams and punctuation perform at 0.58 accuracy on the formal/informal dimension. Considering some of the informal examples, punctuation seems to be one of the most prominent visible changes from a formal to an informal style (see Appendix). Interestingly, word length is the method that most clearly performs better on the simple/complex than the formal/informal dimension. This aligns with the intuition that shorter words are a sign of a simpler style as found in Paetzold and Specia (2016).

Casing encodes style information. The uncased performs worse than the cased BERT model (0.74 vs. 0.77, p=0.008). Additionally, the cased letter ratio performs slightly better than random for the

formal/informal dimension (0.55) and perfect for the contraction characteristic (1.0): When the sentence consists of fewer lower cased characters (as a result of removing them when using contractions), the share of upper cased characters increases.

Style embedding yields promising results. The method "deepstyle" (Hay et al., 2020) performs well across STEL components (0.66). It performs the worst on the simple/complex dimension (0.55). The method embeds sentences in a vector space where texts by "similar" authors are similarly embedded. In the training data (blog and news articles), authors might not consistently use one style over the other. The difference between same author and same style could be explored in future work.

Less ambiguous task instances reach higher accuracy values. Table 3 (c.f. 'full') shows the accuracy of the style measuring methods for the complete set of potential task instances before filtering out ambiguity (§4.1). The accuracies are the same or lower than the crowd-validated task instances in STEL. The differences are more pronounced for the simple/complex than the formal/informal dimension. This aligns with the higher (expected) ambiguity in the simple/complex dimension (§4.2 and §4.3). In general, we recommend to use the filtered STEL task with less ambiguity for testing.

6 Limitations and Future Work

Our illustrative set of task instances does not cover all possibilities of style variation. Future work could extend STEL to cover additional style dimensions or more fine-grained task instances using several sources of data.

The STEL task instances for one style component can contain correlations with unconsidered (style) components. Consider the following task instance (shortened for readability): (A1) *"Forty-nine species of pipefish [...] have been recorded."*, (A2) *"Forty-nine type of pipefish [...] have been found"*, (S1) *"Patients [...] must have their liver checked for damage and other side effects."* and (S2) *"[...] patients [...] must be monitored for liver damage and other possible side effects."*. (A2) and (S1) are the simpler version of (A1) and (S2) (Xu et al., 2016). Additionally, the sentences vary along other aspects: (A2) is missing the punctuation mark and includes a misspelling. (S1) is different in content from (S2) as (S1) is only considering effects on the liver while (S2) also includes other side effects. However, those aspects did not change the label given by the annotators (S2-S1) and should mostly be secondary to the considered style dimension.

With STEL, language models and methods are tested only on whether they capture clear differences in style when content is approximately the same. When there are also content differences, such models might put more emphasis on content than stylistic aspects. Our framework could be extended to allow testing for whether a model prefers style over content (e.g., with a new task format where sentence 1 is closer in content to anchor 1 but closer in style to anchor 2, c.f. Figure 1).

STEL could also be extended to test for individual author styles and style variation related to the social or regional background of authors (e.g., different age groups). For example, by including sentence pairs with the same content but written by different authors. Current and future dimensions could also be extended by a train/dev/test split to enable training on the task directly. Further, STEL could be enriched by including longer texts (e.g., paragraphs or documents) as anchor and alternative sentences.

7 Conclusion

Style is an integral part of language. However, there are only few benchmarks for linguistic style. In this work, we introduce STEL, a modular, content controlled and fine-grained similarity-based style evaluation framework. Out of the evaluated language models and methods, the cased BERT base model performs the best on STEL. Simpler sentence features perform close to the random baseline. STEL includes two general style dimensions and two specific style characteristics. We hope that this framework will grow to include an even more exhaustive representation of linguistic style and will facilitate the development of improved style(-sensitive) measures.

Task Usage

When using this task, please also cite the original datasets the tasks were generated from: (1) Rao and Tetreault (2018) for the formal/informal component and (2) Xu et al. (2016) for the simple/complex component. (1) also needs the permission for usage of the "L6 - Yahoo! Answers Comprehensive Questions and Answers version 1.0 (multi part)"[9].

[9] https://webscope.sandbox.yahoo.com/catalog.php?datatype=l

Ethical Considerations

The STEL tasks are based on datasets (Rao and Tetreault, 2018; Baumgartner et al., 2020; Xu et al., 2016) from popular online forums and web pages (Yahoo! Answers, Reddit, Wikipedia). However, the user demographics on these platforms are often skewed towards particular demographics. For example, Reddit users are more likely to be young and male.[10] Thus, our dataset might not be representative of (English) language use across different social groups. Further, the usage of posts from online platforms without explicit consent from users might lead to (among others) privacy concerns. The Wikipedia simplifications and formal Yahoo! Answers paraphrases were generated by consenting crowdworkers (Xu et al., 2016; Rao and Tetreault, 2018). We expect the sentences that were extracted from Wikipedia for the contraction dimension and for the complex/simple dimension to lead to minimal privacy concerns as they were meant to be read and copied by a broader public.[11] Rao and Tetreault (2018) and the nb3r dimension do not include user names. However, we acknowledge that users might be identifiable from the exact wording of posts. We removed nb3r substitution instances that included Reddit user names. We hope the ethical impact of reusing the already published Rao and Tetreault (2018) dataset to be small.

Acknowledgements

We thank the anonymous EMNLP reviewers for their helpful feedback. We thank Yupei Du and Qixiang Fang for the productive discussions and their equally helpful feedback. This research was supported by the "Digital Society - The Informed Citizen" research programme, which is (partly) financed by the Dutch Research Council (NWO), project 410.19.007. Dong Nguyen was supported by the research programme Veni with project number VI.Veni.192.130, which is (partly) financed by the Dutch Research Council (NWO).

References

Reina Akama, Kento Watanabe, Sho Yokoi, Sosuke Kobayashi, and Kentaro Inui. 2018. Unsupervised learning of style-sensitive word vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 572–578, Melbourne, Australia. Association for Computational Linguistics.

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.

Allan Bell. 2013. *The Guidebook to Sociolinguistics*, chapter 11. John Wiley & Sons.

Janek Bevendorff, Bilal Ghanem, Anastasia Giachanou, Mike Kestemont, Enrique Manjavacas, Ilia Markov, Maximilian Mayerl, Martin Potthast, Francisco Rangel, Paolo Rosso, Günther Specht, Efstathios Stamatatos, Benno Stein, Matti Wiegmann, and Eva Zangerle. 2020. Overview of PAN 2020: Authorship verification, celebrity profiling, profiling fake news spreaders on Twitter, and style change detection. In *CLEF 2020: Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 372–383, Thessaloniki, Greece. Springer International Publishing.

Douglas Biber and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge University Press.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly.

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*.

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, Los Angeles, USA. IEEE.

Benedikt Boenninghoff, Julian Rupp, Robert M. Nickel, and Dorothea Kolossa. 2020. Deep Bayes factor scoring for authorship verification—notebook for PAN at CLEF 2020. In *CLEF 2020 Labs and Workshops, Notebook Papers*, Thessaloniki, Greece. CEUR-WS.

Marcelo Luiz Brocardo, Issa Traore, Sherif Saad, and Isaac Woungang. 2013. Authorship verification for short messages using stylometry. In *2013 International Conference on Computer, Information and Telecommunication Systems (CITS)*, pages 1–6, Athens, Greece. IEEE.

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

[11] https://en.wikipedia.org/wiki/Wikipedia:Copyrights

Julian Brooke and Graeme Hirst. 2013. Hybrid models for lexical acquisition of correlated styles. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 82–90, Nagoya, Japan. Asian Federation of Natural Language Processing.

Kathryn Campbell-Kibler, Penelope Eckert, Norma Mendoza-Denton, and Emma Moore. 2006. The elements of style. In *Poster Session at New Ways of Analyzing Variation*, Columbus, USA. NWAV.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Nikolas Coupland. 2007. *Style: Language Variation and Identity*. Cambridge University Press.

David Crystal and Derek Davy. 1969. *Investigating English Style*. English Language Series. Routledge.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words! Linguistic style accommodation in social media. In *WWW '11: Proceedings of the 20th International Conference on World Wide Web*, page 745–754, Hyderabad, India. Association for Computing Machinery.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, USA. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *WWW '12: Proceedings of the 21st International Conference on World Wide Web*, page 699–708, Lyon, France. Association for Computing Machinery.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

William de Vazelhes, CJ Carey, Yuan Tang, Nathalie Vauquier, and Aurélien Bellet. 2020. metric-learn: Metric learning algorithms in Python. *Journal of Machine Learning Research*, 21(138):1–6.

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, USA. Association for Computational Linguistics.

Steven H. H. Ding, Benjamin C. M. Fung, Farkhund Iqbal, and William K. Cheung. 2019. Learning stylometric representations for authorship analysis. *IEEE Transactions on Cybernetics*, 49(1):107–121.

Maciej Eder. 2013. Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, 30(2):167–182.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2020. Analyzing the persuasive effect of style in news editorial argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3160, Online. Association for Computational Linguistics.

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.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378.

Lucie Flekova, Daniel Preoţiuc-Pietro, and Lyle Ungar. 2016. Exploring stylistic variation with age and income on Twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 313–319, Berlin, Germany. Association for Computational Linguistics.

Lesya Ganushchak, Andrea Krott, and Antje Meyer. 2012. From gr8 to great: Lexical access to sms shortcuts. *Frontiers in Psychology*, 3:150.

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.

Jack Grieve. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3):251–270.

Ton van Haaften and Maarten van Leeuwen. 2021. On the relation between argumentative style and linguistic style: Integrating linguistic-stylistic analysis systematically into the analysis of argumentative style. *Journal of Argumentation in Context*, 10(1):97–120.

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.

Francis Heylighen, Jean-Marc Dewaele, and Leo Apostel. 1999. Formality of language: definition, measurement and behavioral determinants. *Internal Report*.

Dongyeop Kang, Varun Gangal, and Eduard Hovy. 2019. (male, bachelor) and (female, Ph.D) have different connotations: Parallelly annotated stylistic language dataset with multiple personas. 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 1696–1706, Hong Kong, China. Association for Computational Linguistics.

Dongyeop Kang and Eduard Hovy. 2021. Style is NOT a single variable: Case studies for cross-stylistic language understanding. 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 2376–2387, Online. Association for Computational Linguistics.

Mahmut Kaya and Hasan Şakir Bilge. 2019. Deep metric learning: A survey. *Symmetry*, 11(9):1066.

Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Martin Potthast, and Benno Stein. 2020. Overview of the cross-domain authorship verification task at PAN 2020. In *CLEF 2020 Labs and Workshops, Notebook Papers*, Thessaloniki, Greece. CEUR-WS.

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, Online. Association for Computational Linguistics.

William Labov. 2006. *The Social Stratification of English in New York City*, 2 edition. Cambridge University Press.

Marc T. Law, Nicolas Thome, and Matthieu Cord. 2016. Learning a distance metric from relative comparisons between quadruplets of images. *International Journal of Computer Vision*, 121:1–30.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. 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*.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, 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, Online. Association for Computational Linguistics.

Ilia Markov, Nikola Ljubešić, Darja Fišer, and Walter Daelemans. 2021. Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 149–159, Online. Association for Computational Linguistics.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12:153–157.

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

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A survey. *Computational Linguistics*, 42(3):537–593.

Kate G. Niederhoffer and James W. Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360.

Xing Niu and Marine Carpuat. 2017. Discovering stylistic variations in distributional vector space models via lexical paraphrases. In *Proceedings of the Workshop on Stylistic Variation*, pages 20–27, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, USA. Association for Computational Linguistics.

Annaleena Parhankangas and Maija Renko. 2017. Linguistic style and crowdfunding success among social and commercial entrepreneurs. *Journal of Business Venturing*, 32(2):215–236.

Ellie Pavlick and Ani Nenkova. 2015. Inducing lexical style properties for paraphrase and genre differentiation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 218–224, Denver, USA. Association for Computational Linguistics.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

James Pennebaker, Ryan Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015. *University of Texas at Austin*.

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, USA. Association for Computational Linguistics.

Daniel Preoţiuc-Pietro, Wei Xu, and Lyle Ungar. 2016. Discovering user attribute stylistic differences via paraphrasing. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI)*, Phoenix, USA. Association for the Advancement of Artificial Intelligence.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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*, page 37–44, Toronto, Canada. 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.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

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

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pre-training for language understanding. *CoRR*, abs/2004.09297.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

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

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, page 3261–3275, Vancouver, Canada. Neural Information Processing Systems Foundation.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Wei Xu. 2017. From Shakespeare to Twitter: What are language styles all about? In *Proceedings*

of the Workshop on Stylistic Variation, pages 1–9, Copenhagen, Denmark. 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.

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of COLING 2012*, pages 2899–2914, Mumbai, India. The COLING 2012 Organizing Committee.

Eva Zangerle, Maximilian Mayerl, Günther Specht, Martin Potthast, and Benno Stein. 2020. Overview of the style change detection task at PAN 2020. In *CLEF 2020 Labs and Workshops, Notebook Papers*. CEUR-WS.

A Removing Ambiguity

Annotation Setup Information. Prolific[12] crowdworkers could participate up to 5 times in annotating different generated tasks from the formal/informal and simple/complex style dimensions. Each time a participant was asked to annotate 14 potential task instances as well as 2 additional screening questions. The screening questions were randomly sampled from a list of 10 screening questions (see Table 5). The screening questions were manually created and then unanimously and correctly answered by 3 lab-internal annotators in the triple setting. We filtered out all crowdworkers that wrongly answered any of the screening questions. We display the task description (Figure 5 and 6) as well as the phrasing of the questions (Figure 4 and 3). Participants were payed 10,21£/hour on average (above 8.91£ UK minimum wage) and gave consent to the publication of their annotations. We required annotators to be native speakers as we assume them to have a better intuition about their language than non-native speakers: During study design, we conducted a pilot study with 8 different non-native annotators. Several felt their English-speaking abilities were insufficient for the task. The study projected a higher perceived and measured difficulty of the simple/complex dimension. As a result we required annotators to be native speakers and generated more potential simple/complex than formal/informal tasks.

	Sample Acc.	
Dim	Triple	Quadruple
all	**0.68** (+.053)	**0.85** (+.067)
complex	0.54 (+.033)	0.73 (+.044)
formal	0.78 (+.042)	0.94 (+.050)

Table 4: **Subsample Accuracy with Opposite Filtering.** The table displays the accuracy results on the STEL tasks that were correctly annotated in the opposite setup, corresponding to 472 and 375 task instances for the triple and quadruple setup respectively. We display the increase over the accuracy on the full sample in brackets (c.f. 'Sample' results in Table 2a). Compared to the overall accuracy on the sample, the accuracy is higher with opposite filtering for all dimensions (i.e., complex and simple) and setups (i.e., triple and quadruple). The increase is higher for the quadruple than for the triple setup.

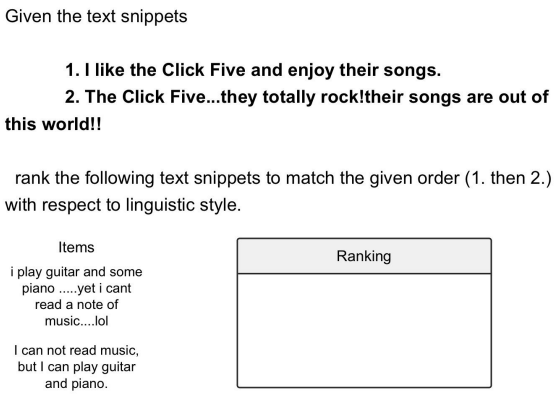

Figure 3: **Example survey question for the quadruple setup.** This is an example of what was shown to the crowdworkers.

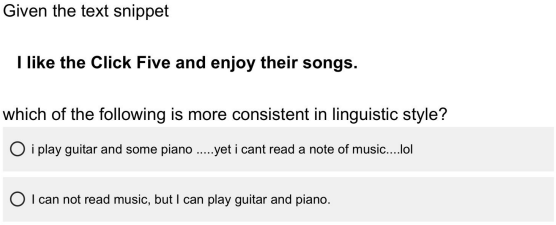

Figure 4: **Example survey question for the triple setup.** This is an example of what was shown to the crowdworkers.

Additional Annotation Results. To further look at the difference between the quadruple and triple setup, we display additional results in Table 4. Here, we only consider the potential task instances that were correctly annotated in the opposite setup. For example, we take the task instances that were correctly annotated in the quadruple setup and see how many of them were also correctly annotated in the triple setup (in this case 0.68). One goal of this analysis was to see whether we can use annotations from the quadruple setup to also remove the ambiguities from the triple setup (or the other way around). However, in both cases accuracy is somewhat low (i.e., below 0.9) and we decided against such an approach. See Table 6 for examples of every combination of (in)correctly annotated triple and quadruple setup of a potential task instance.

[12] https://www.prolific.co/

Comp	Order	Anchor 1 (A1)	Anchor 2 (A2)	Sentence 1 (S1)	Sentence 2 (S2)
formal/-informal	S1-S2	They were engaging in intercourse.	They were having sex.	You do not have the perspective.	It's cause ya got no sense.
formal/-informal	S2-S1	OH, REALLY?	Oh, is that so?	Girlfriends is one of my favorite shows on television.	GIRLFRIENDS IS ONE OF MY FAVORITE SHOWS.
simple/-complex	S1-S2	Many species had vanished by the end of the nineteenth century.	Many animals had disappeared by the end of the 1800s.	They are culturally akin.	Their culture is like the other.
simple/-complex	S1-S2	This stamp remained the standard letter stamp for the remainder of Victoria's reign, and vast quantities were printed.	This stamp stayed the standard letter stamp for the remainder of Victoria's reign, and a lot of them were printed.	Both names became defunct in 2007 when they were merged into The National Museum of Scotland.	Both names stopped being used in 2007 when they became a part of The National Museum of Scotland.
number substitution	S2-S1	You are a n00b.	You are a noob.	This is cool.	This is c00l.
number substitution	S1-S2	|-|0w n3rdy d0 y0u 1!k3 !7¿ ! d0 h4v3 4 107 0f !7 ;-)	How nerdy do you like it? I do have a lot of it ;-)	lol iM N0t CH3At1ng!	lol iM Not CHeating!
Shakespeare	S1-S2	Why, uncle tis a shame.	It's a shame, uncle.	O, wilt thou leave me so unsatisfied?	Oh, you're gonna leave me unsatisfied, right?
formal/-informal	S2-S1	i got limewire if i download songs on it will i get a ticket???	Will I get a ticket if I download songs?	The original song is very good.	The original song is like too good.....
formal/-informal	S2-S1	I like the Click Five and enjoy their songs.	The Click Five...they totally rock!their songs are out of this world!!	i play guitar and some piano .....yet i cant read a note of music....lol	I can not read music, but I can play guitar and piano.
formal/-informal	S2-S1	It reminds me of an old song from the Beatles.	Reminds me of an old beatles song... cant remember which one tho.	KEVIN n nfnhfnigbubjbni.....I dunt really watch American Idol.....	Kevin, I am not exactly an 'American Idol' viewer.

Table 5: **Screening Questions.** List of manually created screening questions to test annotator quality. Anchor 2 is only used in the quadruple setup. The task is to match sentence 1 or sentence 2 with anchor 1 and anchor 2 to the respective other sentence. The correct matching is given in the Order column. The Shakespeare example was taken from Krishna et al. (2020). The rest were either inspired or taken from Xu et al. (2016), Rao and Tetreault (2018) and Baumgartner et al. (2020).

T	Q	Dim	Share	Order	Anchor 1	Anchor 2	Sentence 1	Sentence 2
✗	✓	**formal**	$59 \approx 0.196$	S1-S2	List your best April Fools Pranks here	Please compile a list on here of your best April Fool pranks.	becuase in one of her songs she talks about saying no to sex pressure from her boyfriend	In one of her songs ,she addresses the issue of not letting her boyfriend pressure her into having sexual intercourse.
✗	✓	**complex**	$94 \approx 0.312$	S1-S2	The Book of Nehemiah is a book of the Hebrew Bible, historically seen as a follow-up to the Book of Ezra, and is sometimes called the second book of Ezra.	The Book of Nehemiah is a book of the Hebrew Bible, historically regarded as a continuation of the Book of Ezra, and is sometimes called the second book of Ezra.	All the bats look up to him, and he says he caught two tiger moths which everyone in the colony knows to be a difficult feat for such a young bat	All the bats admire him, and he claims to have caught two tiger moths which are known by all the others in the colony to be an extraordinary achievement by such a young bat.
✓	✗	**formal**	$14 \approx 0.047$	S2-S1	pointsreaper is lame he cannot sue Yahoo for him cheating, what a cry baby	He is not smart. You can not sue a website because you cheated.	A woman did not perform the vocals.	A girl did not sing it.
✓	✗	**complex**	$42 \approx 0.14$	S1-S2	Meanwhile the KLI has about 20 of those former Beginners' Grammarians.	Meanwhile, the KLI has about 20 of those past Beginner's Grammarians.	N-Dubz are a MOBO award winning hip hop group from London, based around Camden Town.	N-Dubz is a MOBO award winning hip hop group, based around Camden Town in London.
✗	✗	**formal**	$20 \approx 0.066$	S1-S2	Gentleman, and I thank God everyday for the one that I have!	I thank God for each day that I have.	GIRLFRIENDS IS ONE OF MY FAVORITE SHOWS.	Girlfriends is one of my favorite shows on television.
✗	✗	**complex**	$54 \approx 0.179$	S1-S2	Among the casualties were two fishers who were reported missing.	Two fisherman are missing among the people who may have been hurt or killed.	Baduhennna is solely attested by Tacitus' Annals where Tacitus records that a grove in Frisia was dedicated to her, and that near this grove 900 Roman prisoners were killed in 28 CE.	In Tacitus' Annals by Tacitus, it is recorded that a grove in Frisia was dedicated to Baduhennna, and near to this grove 900 Roman prisoners were killed in 28 CE.
✓	✓	**formal**	$208 \approx 0.691$	S1-S2	im pretty sure that it was kiss	I am fairly certain it was a kiss.	Law and Order... it just has a clunk clunk	I like Law and Order, although it is a bit clunky lately.
✓	✓	**complex**	$111 \approx 0.369$	S1-S2	Mifepristone is a synthetic steroid compound used as a pharmaceutical.	Mifepristone is a synthetic steroid compound which is used as a medicine.	The video was released on 7/14/06.	The video was premiered on MTV2 on July 14, 2006.

Table 6: **Annotation analysis.** For the simple/complex and the formal/informal dimensions, we give the number of occurrences of each combination of correct (✓) and wrong (✗) annotations in the triple (T) and quadruple (Q) setting. For every combination and style dimension an example is given. The share is calculated out of 301 examples. In total 602 examples were annotated for both Q and T settings with 301 per style dimension. The most common cases are ✓✓and the ✗✓combination for both style dimensions totaling 68.1% and 88.7% of the cases for the simple/complex and formal/informal dimensions respectively. There are ambiguous examples, where one could argue for both possible orders. After manual inspection, this seems to be more prevalent for the simple/complex dimension but it also happens for the formal/informal style dimension. E.g., for row (✗✗, formal), Anchor 1 could be understood as more formal (e.g., 'gentleman') or more informal (e.g., '!' and an unusual grammatical structure). Row (✗✓, formal) is an example of the 'triple problem'.

In this study, you will be expected to complete the following task:

Compare the linguistic style of text snippets.

Opposed to content, **style is not about "what" is said but about "how" it is said.**

The survey will take place in the form of multiple choice questions of the same setup. An example could be:

Question: Given the text snippet

It reminds me of an old song from the Beatles.

which of the following is more consistent in linguistic style?

Alternative A: KEVIN n nfnhfnigbubjbni.....I dunt really watch American Idol.....
Alternative B: Kevin, I am not exactly an 'American Idol' viewer.

Here, alternative B is more consistent in style as alternative A is noticeably more informal (e.g., 'nfnhf' or 'dunt really') than the other two text snippets.

Another example could be:

Question: Given the text snippet

This stamp became the standard for the remnant of Victoria's reign, and vast quantities were printed.

which of the following is more consistent in linguistic style?

Alternative A: Both names became defunct in 2007 when they were merged into The National Museum of Scotland.
Alternative B: Both names stopped being used in 2007 when they became a part of The National Museum of Scotland.

Here, alternative A is more consistent in style as alternative B is noticeably less complex (e.g., 'stopped being used' instead of 'became defunct') than the other two text snippets.

The examples in the survey might be quite hard. In case you can not find a good reasoning for which alternative is more consistent in style, **try to compare and find the differences between alternative A and alternative B**.

Figure 5: **Survey task description for the triple setup.** This is a copy of what was shown to the crowdworkers.

In this study, you will be expected to complete the following task:

Compare the linguistic style of text snippets.

Opposed to content, **style is not about "what" is said but about "how" it is said**.

The survey will take place in the form of ranking questions of the same setup. An example could be:

Question: Given the text snippets

1. It reminds me of an old song from the Beatles.
2. Reminds me of an old beatles song... cant remember which one tho.

rank the following text snippets to match the given order (1. then 2.) with respect to linguistic style.

Alternative A: KEVIN n nfnhfnigbubjbni.....I dunt really watch American Idol.....
Alternative B: Kevin, I am not exactly an 'American Idol' viewer.

Here, alternative B is more formal than alternative A (e.g., 'nfnhf' or 'dunt really' in A). We can also see that text snippet 1 is more formal than text snippet 2 (e.g., snippet 2 contains 'tho' and 'cant'). As a result, the ordering that is most consistent with the text snippets is alternative B then alternative A.

Another example could be:

Question: Given the text snippets

1. This stamp remained the standard letter stamp for the remainder of Victoria's reign, and vast quantities were printed.
2. This stamp stayed the standard letter stamp for the remainder of Victoria's reign, and a lot of them were printed.

rank the following text snippets to match the given order (1. then 2.) with respect to linguistic style.

Alternative A: Both names became defunct in 2007 when they were merged into The National Museum of Scotland.
Alternative B: Both names stopped being used in 2007 when they became a part of The National Museum of Scotland.

Here, alternative A is phrased in a more complex style than alternative B (e.g., 'became defunct' instead of 'stopped being used'). We can also see that text snippet 1 is more complex than 2 (e.g., 'vast quantities' instead of 'a lot of them'). As a result, the ordering that is most consistent with the text snippets is alternative A then alternative B.

The examples in the survey might be quite hard. In case you can not find a good reasoning for which ordering is more consistent in style, **try to compare and find the differences between alternative A and alternative B and match them to differences in 1. and 2.**

Figure 6: **Survey task description for the quadruple setup.** This is a copy of what was shown to the crowdworkers.

B Additional STEL Results

In Table 7, we display the share of task instances where models and methods could not decide between the two possible answers. This is adding more detail to the 'random' column of Table 3. The share of random decisions is lower for the more complex style dimensions (formal/informal: 0.05 and simple/complex: 0.13) and higher for the simpler style characteristics (nb3r substitution: 0.38 and contrac'tion usage: 0.15). This aligns with the intuition that the difference between the sentence pairs in the nb3r and contrac'tion dimension is smaller. The neural methods have a lower share of random decisions overall.

	all		formal		complex		nb3r	c'tion
	filter	full	filter	full	filter	full		
BERT uncased	0	0	0	0	0	0	0	0
BERT cased	0	0	0	0	0	0	0	0
RoBERTa	0	0	0	0	0	0	0	0
SBERT mpnet	0	0	0	0	0	0	0	0
SBERT para-mpnet	0	0	0	0	0	0	0	0
USE	0.00	0.00	0.00	0.00	0.00	0.01	0	0
BERT uncased NSP	0.10	0.12	0	0	0.21	0.21	0	0.19
BERT cased NSP	0.02	0.02	0	0	0.04	0.04	0	0.03
char 3-gram	0.05	0.05	0.03	0.04	0.01	0.01	0.57	0.02
word length	0.08	0.08	0.04	0.05	0.04	0.04	0.91	0
punctuation	**0.38**	**0.39**	**0.31**	**0.31**	**0.42**	**0.42**	**0.97**	0.06
LIWC	0.09	0.09	0.01	0.01	0.12	0.13	0.53	0
LIWC (style)	**0.62**	**0.64**	**0.37**	**0.38**	**0.80**	**0.79**	**0.94**	**1.0**
LIWC (function)	0.28	0.28	0.14	0.14	0.38	0.38	0.81	0
deepstyle	0	0	0	0	0	0	0	0
POS Tag	0.20	0.20	0.02	0.02	0.24	0.24	0.64	**1.0**
share cased	0.08	0.08	0.02	0.02	0.05	0.05	0.91	0
edit dist	0.08	0.07	0.01	0.01	0.05	0.05	0.52	**0.33**
Average	0.11	0.11	0.05	0.05	0.13	0.13	0.38	0.15

Table 7: **Share of Random Decisions.** The share of task instances for which a method can not decide between the two options and decides randomly is given per dimension. The performance on the set of task instances before (full) and after crowd-sourced filtering (filter) is displayed. The two highest shares of random decisions is boldfaced. The share of random decisions is highest for the nb3r and lowest for the formal dimension. LIWC (style) and punctuation similarity have the overall highest share of random decisions.

C Task Generation

C.1 Contraction dictionary

The Wikipedia style guide discourages contraction usage and provides a dictionary with contractions that should be avoided.[13] Some of those contractions are more colloquial (e.g., 'twas or ain't). We use an adapted version removing colloquial and less common contractions: { "aren't": "are not", "can't": "cannot / can not", "could've": "could have", "couldn't": "could not", "didn't": "did not", "doesn't": "does not", "don't": "do not", "everybody's": "everybody is", "everyone's": "everyone is", "hadn't": "had not", "hasn't": "has not", "haven't": "have not", "he'd": "he had / he would", "he'll": "he will", "he's": "he has / he is", "here's": "here is", "how'd": "how did / how would", "how'll": "how will", "how's": "how has / how is", "I'd": "I had / I would / I should", "I'll": "I shall / I will", "I'm": "I am", "I've": "I have", "isn't": "is not", "it'd": "it would / it had", "it'll": "it shall / it will", "it's": "it has / it is", "mightn't": "might not", "mustn't": "must not", "must've": "must have", "needn't": "need not", "oughtn't": "ought not", "shan't": "shall not", "she'd": "she had / she would", "she'll": "she shall / she will", "she's": "she has / she is", "should've": "should have", "shouldn't": "should not", "somebody's": "somebody has / somebody is", "somebody'd": "somebody would / somebody had", "somebody'll": "somebody will", "someone's": "someone has / someone is", "someone'd": "someone would / someone had", "someone'll": "someone will", "something's": "something has / something is", "something'd": "something would / something had", "something'll": "something will", "that'll": "that will", "that's": "that has / that is", "that'd": "that would / that had", "there'd": "there had / there would", "there'll": "there shall / there will", "there's": "there has / there is", "there've": "there have", "these're": "these are", "they'd": "they had / they would", "they'll": "they shall / they will", "they're": "they are", "they've": "they have", "wasn't": "was not", "we'd": "we had / we would / we should", "we'll": "we shall / we will", "we're": "we are", "we've": "we have", "weren't": "were not", "what's": "what has / what is / what does", "when's": "when has / when is", "who'd": "who would / who had", "who'll": "who will", "who's": "who has / who is", "won't": "will not", "would've": "would have", "wouldn't": "would not", "you'd": "you had / you would", "you'll": "you shall / you will", "you're": "you are", "you've": "you have"}

C.2 Number substitutions

We selected a pool of potential sentences where words contained character substitution symbols (4,3,1,!,0,7,5) or are part of a manually selected "seed list" of number substitution words[14]:
{ "2morrow": "tomorrow", "c00l": "cool", "n!ce":"nice", "l0ve":"love", "sw33t":"sweet", "l00k":"look", "4ever":"forever", "l33t":"leet", "1337":"leet", "sk8r":"skater" "n00b":"noob", "d00d":"dude", "ph34r":"fear", "w00t":"woot", "b4":"before", "gr8":"great", "2day":"today", "t3h":"teh", "m4d":"mad", "j00":"joo", "0wn":"own", "h8":"hate", "w8":"wait" }

Then, we manually filtered out sentences without number substitutions (e.g., common measuring units or product numbers). Our resulting list of 100 sentences pairs contains more substitution words than the above "seed list" (e.g., "d4rk", "appreci8", "h1m").

[13] https://en.wikipedia.org/wiki/Wikipedia:List_of_English_contractions

[14] Inspired by https://www.gamehouse.com/blog/leet-speak-cheat-sheet/, https://simple.wikipedia.org/wiki/Leet, Ganushchak et al. (2012), https://h2g2.com/edited_entry/A787917 and manually looking at a few Reddit posts

D Similarity-based Decision

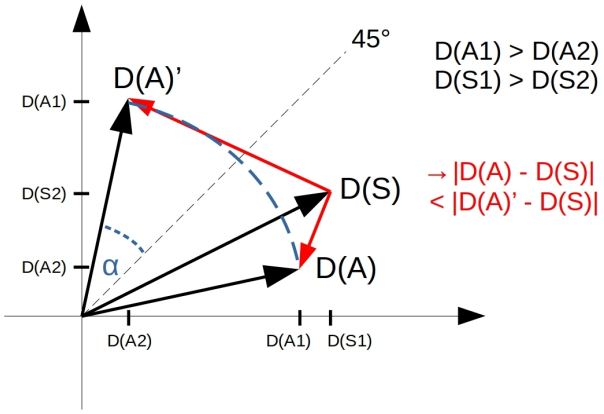

Figure 7: **Proof Sketch.** Let D be the considered style component (e.g., formal/informal) and $D(\text{A1})$, $D(\text{A2})$, $D(\text{S1})$, $D(\text{S2})$ be the localization of $\text{A1}, \text{A2}, \text{S1}, \text{S2}$ along that component. W.l.o.g., let the correct ordering be S1-S2 and $D(\text{A1}) > D(\text{A2})$. Let us assume that for all other style and content aspects $\widetilde{D}$ (e.g., simple/complex), $\widetilde{D}(\text{A1}) = \widetilde{D}(\text{A2})$ and $\widetilde{D}(\text{S1}) = \widetilde{D}(\text{S2})$ hold. We define $D(\text{A}) := \big(D(\text{A1}) \quad D(\text{A2})\big)^\intercal$ and $D(\text{S}) := \big(D(\text{S1}) \quad D(\text{S2})\big)^\intercal$ as the style vectors of the combined anchor (A1 and A2) and alternative sentences (S1 and S2). Then, with the correct ordering being S1-S2 and $D(\text{A1}) > D(\text{A2})$, $D(\text{S1}) > D(\text{S2})$ holds. Thus, both $D(\text{A})$ and $D(\text{S})$ point to a coordinate below the $45°$-axis when the first component of the respective vectors corresponds to the x-axis and the second to the y-axis (see sketch). Let $D(\text{A})'$ be the reflected vector of $D(\text{A})$ along the $45°$-axis, i.e., $\big(D(\text{A2}) \quad D(\text{A1})\big)^\intercal$. Then, as shown in the sketch, the length of the vector $D(\text{A}) - D(\text{S})$ is smaller than $D(\text{A})' - D(\text{S})$ as the angle between $D(\text{A})$ and $D(\text{S})$ will always be smaller than the one between $D(\text{A})'$ and $D(\text{S})$. This corresponds to equation (1), when replacing similarities (i.e., $1 - \text{sim}(x, y)$) with distances (i.e., $|x - y|$). Thus, equation (1) holds when working with style-sensitive similarity functions that can be translated to distances. Note: As only cosine 'angular distance' is a distance metric, this would need to be the angular cosine similarity. However, angular cosine similarity can be replaced by cosine similarity in inequality (1) as relative ordering is the same for the two similarity metrics.

E Computing Infrastructure

The evaluation of the 18 (language) models and methods took 14 hours in total on a machine with 32 GB RAM and 8 intel i7 CPUs using Ubuntu 20.04 LTS. No GPU was used.