MSTYLEDISTANCE: Multilingual Style Embeddings and their Evaluation

Justin Qiu* **Jiacheng Zhu*** **Ajay Patel**
Marianna Apidianaki **Chris Callison-Burch**
University of Pennsylvania
{jsq, jiachzhu, ajayp, marapi, ccb}@seas.upenn.edu

Abstract

Style embeddings are useful for stylistic analysis and style transfer; however, only English style embeddings have been made available. We introduce Multilingual STYLEDISTANCE (MSTYLEDISTANCE), a multilingual style embedding model trained using synthetic data and contrastive learning. We train the model on data from nine languages and create a multilingual STEL-or-Content benchmark (Wegmann et al., 2022) that serves to assess the embeddings' quality. We also employ our embeddings in an authorship verification task involving different languages. Our results show that MSTYLEDISTANCE embeddings outperform existing models on these multilingual style benchmarks and generalize well to unseen features and languages. We make our model publicly available at https://huggingface.co/StyleDistance/mstyledistance.

1 Introduction

Style embedding models seek to embed texts with similar style closer in the embedding space regardless of their content. Style embeddings are useful for tasks like style transfer and authorship attribution, but only exist for English (Wegmann et al., 2022; Patel et al., 2024b). Multilingual style embeddings could also serve to automatically evaluate style preservation in machine translation. Models like XLM-RoBERTa (Conneau et al., 2019) and E5 (Wang et al., 2024) create multilingual representations for semantic tasks, but have not addressed style mainly due to the scarcity of style datasets.

We propose a procedure, called multilingual STYLEDISTANCE (MSTYLEDISTANCE), to train style embeddings using contrastive learning with synthetic data in multiple languages. Early work on style representations learning often involved unlabeled social media data (Hay et al., 2020; Wegmann et al., 2022; Rivera-Soto et al., 2021; Patel et al., 2023), but Patel et al. (2024b) showed that a contrastive learning objective with synthetic examples (sentence pairs with similar content and different style) can generate high quality style representations for English. We create MSYNTHSTEL, a synthetic dataset of paraphrases addressing various style features in nine languages, and use it to create our multilingual style embeddings.

In order to evaluate their quality, we contribute a new multilingual and cross-lingual STEL-or-Content (SoC) evaluation benchmark which, following the original SoC evaluation task (Wegmann et al., 2022), measures the ability of a model to embed sentences with the same style closer in the embedding space than sentences with the same

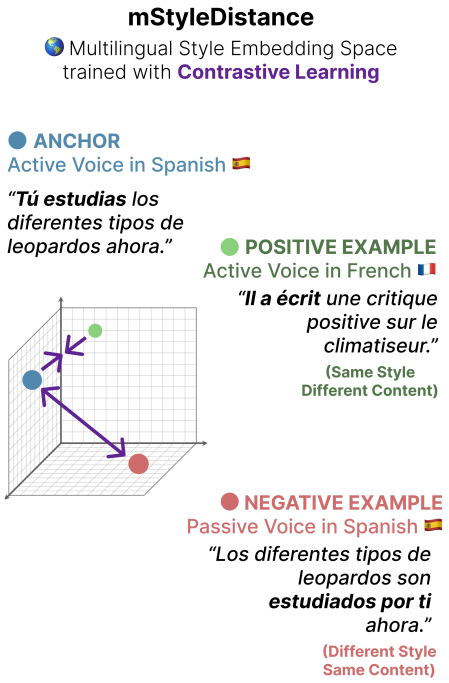

Figure 1: MSTYLEDISTANCE is trained using contrastive learning with synthetic examples created for ~40 style features in 9 languages. These positive and negative examples serve to form multilingual and cross-lingual training triplets.

*Denotes equal contribution.

Findings of the Association for Computational Linguistics: ACL 2025, pages 16917–16931

content. We show that MSTYLEDISTANCE embeddings outperform other representations in these evaluations, and demonstrate their usefulness in a downstream setting addressing a multilingual authorship verification task. We publicly release our model, data, and evaluation benchmarks.

2 Multilingual Synthetic Data

We extend the Patel et al. (2024b) dataset to nine languages (L): Arabic, German, Spanish, French, Hindi, Japanese, Korean, Russian, and Chinese.[1]

Style Feature Selection We use the set of 40 features addressed in Patel et al. (2024b), leaving out features not applicable to specific languages. For example, articles are not relevant for Chinese and Japanese so the corresponding features have been discarded. Our set of features (F) includes syntactic features (e.g., active/passive voice, contractions, frequent use of function words), emotional and cognitive features (e.g., words indicating sentiment or cognitive processes), stylistic and aesthetic features (e.g., metaphors, formal tone), social and interpersonal features (e.g., polite or offensive tone), graphical and digital features (e.g., capitalization, emojis, numerical digits), temporal and aspectual features (e.g., focus on present or future). A full list of the features that were considered for each language and details about the features that were omitted are given in Appendix A.

Data Generation For each retained feature $f \in F$ for a language $l \in L$, we generate 100 pairs of positive (`pos`) and negative (`neg`) examples (paraphrases). In each pair, `pos` is a sentence that contains the style feature (e.g., a formal sentence or a metaphorical one) while `neg` does not. This is illustrated in Figure 1 for the feature "Active Voice". Features that cannot be removed completely (e.g., "usage of articles") are present with higher frequency in `pos` than in `neg` examples.

Using the same prompting workflow as Patel et al. (2024b), we generate sentence pairs by prompting GPT-4 with the DataDreamer library and an attributed prompt (Yu et al., 2023) illustrated in Figure 2 (OpenAI et al., 2024; Patel et al., 2024a). For diversity, a "Topic" for each generation is sampled by extracting a random sentence from a random document in the C4 corpus (Raffel et al., 2020), and prompting GPT-4 to identify the topic. Our training dataset also includes English-only data from SYNTHSTEL (Patel et al., 2024b). For further details on all prompts, see Appendix B.

> Generate a pair of active and passive Russian sentences with the following attributes:
>
> 1. Topic: Brake parts and components
> 2. Length: 15-20 words
> 3. Point of view: second-person
> 4. Tense: past
> 5. Type of Sentence: Exclamation

Figure 2: Example prompt for generating a pair of sentences in Russian.

We experiment with two approaches to multilingual data generation. In our first approach, sentence pairs are directly generated in each $l \in L$ using a language-specific instruction in the prompt, as illustrated in Figure 2. In our second approach, English sentence pairs are generated using the prompting workflow and then translated into each target language $l \in L$. We use DeepL for all languages except for Hindi which is not supported, where we use Google Translate instead. We generate data using both methods and conduct human validation on a random 10% split of the training data in order to determine the best approach for each language.[2]

Human validation Each sentence pair for a style feature creates two annotation instances, one for the positive and one for the negative sentence. For each instance, we asked the annotators to judge whether the style feature is present in the sentence by selecting an answer among "Yes", "Possibly", and "No". We also provided a definition for each style feature to help annotators in their decision. We then asked the annotators to rate the fluency of the sentence by selecting one among "Fluent", "Mostly Fluent", "Mostly Disfluent", "Disfluent". We collected a total of 13,651 annotations for the 9 languages in L. When considering only instances annotated by at least three annotators, the inter-annotator agreement was $\alpha = 0.4247$ (Krippendorff, 2011). More details on the annotation task and the exact number of collected annotations per language can be found in Appendix C.

We calculate an aggregate "feature presence" accuracy score for each $l \in L$ based on whether the average feature presence score over all annotations

[1] These languages were chosen because they align with the linguistic background of our annotators.

[2] Our annotators were undergraduate and graduate students native speakers of a language, who were offered extra credit for participation.

is higher for the positive sentence than for the negative sentence in a pair. We assign a score of 1 to "Yes", 0.5 to "Possibly", and 0 to "No". We also calculate an aggregate fluency score by taking an average of the fluency scores that each annotator gave to each text. We assign a score of 1 to "Fluent", 0.67 to "Mostly Fluent", 0.33 to "Mostly Disfluent", and 0 to "Disfluent". We selected the best approach (direct generation v.s. English → MT) for each $l \in L$ as the one that produced the most fluent sentences, or the highest feature presence score if both methods produced similarly fluent sentences. The direct approach was selected for all languages in L except for Japanese and Hindi, where the translation approach produced higher quality data. Our final average feature presence and fluency scores over all $l \in L$, with the best generation approach selected for each l, are 0.79 and 0.93, respectively, both above random chance (0.5). Detailed results by language are given in Appendix D.

Automatic Data Validation Following Patel et al. (2024b), we also perform automatic validations of the generated data. We examine whether our parallel examples are indeed paraphrases by computing their average cosine similarity[3] (Reimers and Gurevych, 2019). For comparison, we calculate the similarity of 100 gold-standard paraphrases sampled from the multilingual dataset compiled by Scherrer (2020) for each language. The average similarity of our parallel examples is 0.91 which is comparable to that calculated on the Scherrer (2020) natural data (0.88), indicating that our pairs are reasonable paraphrases.

We also measure topic diversity across generated sentences for a $l \in L$ using the metric proposed by Yang et al. (2024) which relies on cosine distance. In this case, we only use the `pos` sentence which contains the style feature for each pair. For comparison, we also compute the diversity score for texts from Scherrer (2020). Again, the two scores are comparable (0.83 vs. 0.85), showing that our examples are similar to natural data in terms of diversity. Detailed results of the automatic evaluations are given in Appendix D.

3 Training mSTYLEDISTANCE

Following the contrastive training approach of Patel et al. (2024b), we construct feature-specific triples for each language $l \in L$ which contain: an anchor text (a); a text with the same style as a but different content (`pos`); a distractor text (`neg`) which is a paraphrase of a or `pos`, but different in style from a. We use the multilingual `xlm-roberta-base` as our base model and train with a triplet loss (Conneau et al., 2019; Schroff et al., 2015). We ensure half of our triplets are *cross-lingual*, i.e. the `pos` and `neg` texts are randomly sampled from a different language than the anchor text. Our triplet creation process ensures equal coverage of the languages we support. Full training details can be found in Appendix E.

4 Evaluation

STEL-or-Content (SoC) Benchmark In order to evaluate our style embeddings, we construct a multilingual version of the SoC benchmark (Wegmann et al., 2022).[4] SoC measures the ability of a model to embed sentences with the same style closer in the embedding space than sentences with the same content. We construct our **multilingual SoC benchmark** by sampling 100 pairs of parallel `pos`-`neg` examples for each language from four ground-truth datasets covering four style features and 22 languages: simplicity (Ryan et al., 2023), formality (Briakou et al., 2021), toxicity (Dementieva et al., 2024), and positivity (Mukherjee et al., 2024).[5] Each instance in our multilingual SoC benchmark consists of a triplet (a, `pos`, `neg`) constructed as explained in Section 3. However, following Wegmann et al. (2022), the distractor text in our SoC benchmark is always a paraphrase of `pos`. A model tested on this benchmark is expected to embed a and `pos` closer than a and `neg`. We rate a model by computing the percentage of instances it achieves this goal for across all instances. We form test instances for each feature $f \in F$ available for a language corresponding to all possible triplets, resulting in 4,950 instances for each language-feature style combination.

We also construct a **cross-lingual SoC benchmark** that addresses embeddings' ability to capture style similarity *across languages*. This can be useful, for example, to evaluate style preservation in translations. We construct the benchmark with the XFormal dataset (Briakou et al., 2021),

[3] We use `paraphrase-multilingual-mpnet-base-v2`.

[4] The English SoC benchmark covered formality, complexity, number usage, contraction usage, and emoji usage.

[5] Combined, these datasets cover the following languages: Amharic, Arabic, Bengali, German, English, Spanish, French, Hindi, Italian, Japanese, Magahi, Malayalam, Marathi, Odia, Punjabi, Portuguese (Brazil), Russian, Slovenian, Telugu, Ukrainian, Urdu, and Chinese.

Model	Test Set	Simplicity	Formality	Toxicity	Positivity	Formality (cross-lingual)
Wegmann et al. (2022)	Original	0.23	0.63	0.19	0.23	0.45
StyleDistance	Original	0.21	0.67	0.15	0.18	0.49
xlm-roberta-base	Original	0.12	0.16	0.09	0.07	0.19
LISA	Original	0.15	0.09	0.09	0.21	0.27
mStyleDistance	Original	**0.36**	**0.71**	0.37	**0.30**	**0.53**
Wegmann et al. (2022)	Translated	0.25	0.50	0.21	0.19	0.36
StyleDistance	Translated	0.25	0.44	0.23	0.27	0.35
xlm-roberta-base	Translated	0.12	0.07	0.08	0.07	0.08
LISA	Translated	0.04	0.02	0.07	0.13	0.04
mStyleDistance	Translated	0.33	0.51	**0.41**	0.28	0.43

Table 1: Performance of multilingual and English embeddings on the original multilingual and cross-lingual SoC benchmarks (rows 1-5), averaged across languages for each style feature. mStyleDistance leads in cross-lingual and overall performance. For comparison, we also report results obtained on the test set translated into English (rows 6-10).

	PAN 2013		PAN 2014			PAN 2015			Average			
Model	**Greek**	**Spanish**	**Greek**	**Spanish**	**Dutch**	**Greek**	**Spanish**	**Dutch**	**Greek**	**Spanish**	**Dutch**	**Overall**
Wegmann et al. (2022)	0.66	0.87	0.56	0.54	0.59	0.47	0.61	0.59	0.56	0.67	0.59	0.61
StyleDistance	0.61	0.62	0.48	0.51	0.65	0.47	0.73	0.59	0.52	0.62	**0.62**	0.59
LISA	0.51	0.64	0.46	0.56	0.62	0.48	0.66	0.48	0.48	0.62	0.55	0.55
mStyleDistance	0.48	0.78	0.71	0.68	0.72	0.73	0.74	0.48	**0.64**	**0.73**	0.60	**0.66**

Table 2: Results on the PAN 2013-2015 Authorship Verification shared task for Greek, Spanish, and Dutch. We report performance separately on each PAN dataset and average performance across datasets for the same language. We use the standard ROC-AUC metric for authorship verification.

which includes parallel data in French, Italian and Portuguese. We again create triplets as described above, but instead of using `pos` and `neg` texts from the same language as the anchor (a), we sample them from a different language than a. We end up with 19,800 instances for each style in each language. Appendix F contains illustrative examples from each benchmark.

SoC Evaluation Results The results obtained by mStyleDistance on the multilingual and cross-lingual SoC benchmarks are presented in Table 1. Since no general multilingual style embeddings are currently available, we compare with a base multilingual encoder model `xlm-roberta-base` (Conneau et al., 2019) as well as with a number of English-trained style embedding models applied in zero-shot fashion to multilingual text: the Wegmann et al. (2022) style embeddings, StyleDistance embeddings (Patel et al., 2024b), and `LISA` embeddings (Patel et al., 2023). The results of the models on the original cross-lingual and multilingual test set are given in the top part of Table 1 (rows 1-5). mStyleDistance embeddings outperform the other models on this dataset, indicating their suitability for multilingual applications. The other models perform slightly better than the untrained `xlm-roberta-base` but still worse than mStyleDistance.

We also compare mStyleDistance with the English-only models on test data translated into English.[6] The results on this translated test set, which are given in the lower part of Table 1 (rows 6 to 10), are often lower than those obtained on the original multilingual test set. This might look surprising but can be explained by the inherent limitations of machine translation, which can cause certain stylistic nuances to be diminished or lost during the translation process. This is illustrated in the example below where the Chinese formal (您) and informal (你) pronouns are both translated as "you" in English, and the original formality nuances are not present in the translation.

> Formal: 您是否对自己过去对于膳食补充剂立场的批评有所回应?
> Translation: *Have you responded to criticisms of your past position on dietary supplements?*
>
> Informal: 你回应过对你过去膳食补充剂立场的批评了吗?
> Translation: *Have you responded to the criticisms of your past stance on dietary supplements?*

Breakdown of results by feature and language in

[6] We use DeepL for languages it supports and Google Translate for the other languages.

Features Tested	m avg	c avg	Retained Perf (%)	
			m	c
In-Domain	0.38	0.53	100%	100%
Out of Domain	0.31	0.44	75%	74%
Out of Distribution	0.31	0.40	75%	62%
No Language Overlap	0.35	0.52	89%	97%

Table 3: MSTYLEDISTANCE embeddings under three generalization conditions on the multilingual (m avg) and cross-lingual (c avg) STEL-or-Content tasks.

the multilingual and cross-lingual STEL-or-content benchmarks can be found in Appendix H.

Ablation Experiments Following (Patel et al., 2024b), we run several ablation experiments to evaluate how well our model generalizes to unseen style features and languages. In the **In-Domain** condition, all style features are included in the training data for every language. To test generalization to unseen style features, in the **Out of Domain** condition, any style features directly equivalent to those features tested in the Multilingual and Cross-lingual SoC benchmarks are excluded from the training data. **Out of Distribution** further removes any style features indirectly similar or related to those tested in the benchmarks. Finally, **No Language Overlap** removes the languages present in the benchmark from the training data, in order to test generalization to new languages. Our results are given in Table 3 where we measure how much of the performance increase on SoC benchmarks over the base model is retained, despite ablating training data. The results indicate that our method generalizes reasonably well to both "out of domain" and "out of distribution" style features, and very well to languages not in the training data. Further details on features and languages ablated and full results are provided in Appendix G.

Downstream Task Following Patel et al. (2024b), we also evaluate our MSTYLEDISTANCE embeddings in the authorship verification task, where the goal is to determine if two documents were written by the same author using stylistic features (Koppel and Winter, 2014). We use the datasets released by PAN[7] between 2013 and 2015 in Greek, Spanish, and Dutch. Our results are given in Table 2. MSTYLEDISTANCE vectors outperform existing English style embedding models on Spanish and Greek, while Dutch shows similar performance to English STYLEDISTANCE. We hypothesize that the linguistic proximity (West Germanic roots) of the two languages helps STYLEDISTANCE to generalize to Dutch.

[7] https://pan.webis.de

5 Conclusion

We introduced a novel approach to learning multilingual style embeddings from synthetic examples, and contribute a two benchmarks for evaluating the quality of multilingual style representations. We show that MSTYLEDISTANCE embeddings are able to distinguish style from content better than other English and multilingual embeddings, and generalize well to unseen features and languages. The authorship verification evaluation shows that MSTYLEDISTANCE embeddings also offer strong performance on multilingual downstream tasks.

Limitations

Our synthetic data generation approaches rely on direct generation or machine translation techniques, both of which have limitations for languages other than English. Most of the languages included in our multilingual and cross-lingual STEL-or-Content and authorship evaluations are not really low-resource, so our evaluations may not reflect performance in languages with less resources. Furthermore, our approach only targets the 33-40 style features (depending on the language) we generated data for, and cannot account for the wide range of possible style variations. We generated data for features that are applicable to some extent to a given language, even if their expression is relatively weaker than in another language. Language comparability in style space is a question worth exploring in depth in future work. While these constraints may limit our approach, our ablation experiments show strong generalization capabilities to unseen languages and style features indicating promising generalized performance.

Ethical Considerations

This work demonstrates the potential of using synthetic data for creating style embeddings in languages lacking such resources, increasing access to broader communities. However, it is important to recognize that the synthetic data generated by large language models may reflect and reinforce existing biases inherent in these models (Patel et al., 2023). While our approach shows significant promise, ongoing efforts should ensure that such synthetic

datasets are evaluated for fairness and bias to promote more equitable outcomes.

Contribution Statement

Justin Qiu created the SoC benchmarks and performed most of the evaluations. Jiacheng Zhu created MSYNTHSTEL by carrying out data generation and human annotation collection. Ajay Patel trained MSTYLEDISTANCE. Marianna Apidianaki helped greatly with advising and the final version of the manuscript. Chris Callison-Burch also gave us valuable advice as our advisor. All authors contributed to the final manuscript.

Acknowledgements

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021. Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. In *Annual Meeting of the Association for Computational Linguistics*.

Daryna Dementieva, Daniil Moskovskiy, Nikolay Babakov, Abinew Ali Ayele, Naquee Rizwan, Frolian Schneider, Xintog Wang, Seid Muhie Yimam, Dmitry Ustalov, Elisei Stakovskii, Alisa Smirnova, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2024. Overview of the multilingual text detoxification task at pan 2024. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*. CEUR-WS.org.

Julien Hay, Bich-Liên Doan, Fabrice Popineau, and Ouassim Ait Elhara. 2020. Representation learning of writing style. In *Proceedings of the 6th Workshop on Noisy User-generated Text (W-NUT 2020)*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Moshe Koppel and Yaron Winter. 2014. Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, 65(1):178–187.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Sourabrata Mukherjee, Atul Kr. Ojha, Akanksha Bansal, Deepak Alok, John P. McCrae, and Ondřej Dušek. 2024. Multilingual text style transfer: Datasets i& models for indian languages. *Preprint*, arXiv:2405.20805.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li,

Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ajay Patel, Colin Raffel, and Chris Callison-Burch. 2024a. DataDreamer: A tool for synthetic data generation and reproducible LLM workflows. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3781–3799, Bangkok, Thailand. Association for Computational Linguistics.

Ajay Patel, Delip Rao, Ansh Kothary, Kathleen McKeown, and Chris Callison-Burch. 2023. Learning interpretable style embeddings via prompting LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15270–15290, Singapore. Association for Computational Linguistics.

Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. 2024b. Styledistance: Stronger content-independent style embeddings with synthetic parallel examples. *Preprint*, arXiv:2410.12757.

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*, 21:1–6.

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.

Rafael A Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. Learning universal authorship representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 913–919.

Michael Ryan, Tarek Naous, and Wei Xu. 2023. Revisiting non-English text simplification: A unified multilingual benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4898–4927, Toronto, Canada. Association for Computational Linguistics.

Yves Scherrer. 2020. TaPaCo: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6868–6873, Marseille, France. European Language Resources Association.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Anna Wegmann, Marijn Schraagen, Dong Nguyen, et al. 2022. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, page 249. Association for Computational Linguistics.

Yue Yang, Mona Gandhi, Yufei Wang, Yifan Wu, Michael S Yao, Chris Callison-Burch, James C Gee, and Mark Yatskar. 2024. A textbook remedy for domain shifts: Knowledge priors for medical image analysis. *arXiv preprint arXiv:2405.14839*.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as

attributed training data generator: A tale of diversity and bias. In *Advances in Neural Information Processing Systems*, volume 36, pages 55734–55784. Curran Associates, Inc.

A Style Features and Definitions

The style features addressed in our experiments included most of the 40 style features in the Patel et al. (2024b) dataset. In Table 9, we list the 40 style features with an 'Excluded in' column indicating the languages where each feature is not applicable and was therefore omitted from our dataset.

B Generation Prompts and Details

Below we detail the structure of the prompts and the inference parameters used for our two multilingual synthetic data generation methods.

B.1 Topic Extraction from C4

We use the same topic extraction method as Patel et al. (2024b), which is derived from the C4 dataset (Raffel et al., 2020), to identify 50,000 topics through zero-shot prompting with GPT-4 (OpenAI et al., 2024). These 50,000 fine-grained, unique topics ensure that each sentence pair has a distinct context across various features and languages. We perform topic sampling with a `temperature` setting of 1.0 and `top_p` = 0.0.

```
What is the fine-grained topic of the following
text:
 {sentence} Only return the topic.
```

The fine-grained topic is used as part of the attributed prompt in Section B.2 to ensure diverse generations.

B.2 Generation of Positive and Negative Example Sentences

We use the same prompt as Patel et al. (2024b) to generate positive and negative example sentences in English. We then translate these sentence pairs into the target languages using the DeepL API. The only exception is Hindi, which we translate using Google Translate API due to DeepL's limited language support.

For our second method, where example sentences are directly generated in the target language, we use the following prompt with `temperature` setting of 1.0 and `top_p` = 1.0

```
Generate a pair of {target language} sentences with
and without sarcasm with the following attributes:
 1. Topic: {topic}
 2. Length: {sentence_length}
 3. Point of view: {point_of_view}
 4. Tense: {tense}
 5. Type of Sentence: {sentence_type}

Ensure that the generated sentences meet the
following conditions:
 1. There is no extra information in one sentence
 that is not in the other.
 2. The difference between the two sentences is
 subtle.
 3. The two sentences have the same length.
 {special_conditions_for_style_feature}

Use Format:
 With sarcasm: [sentence in {target language}]
 Without sarcasm: [sentence in {target language}]

Your response should only consist of the two
sentences, without quotation marks.
```

C Human Annotation Details

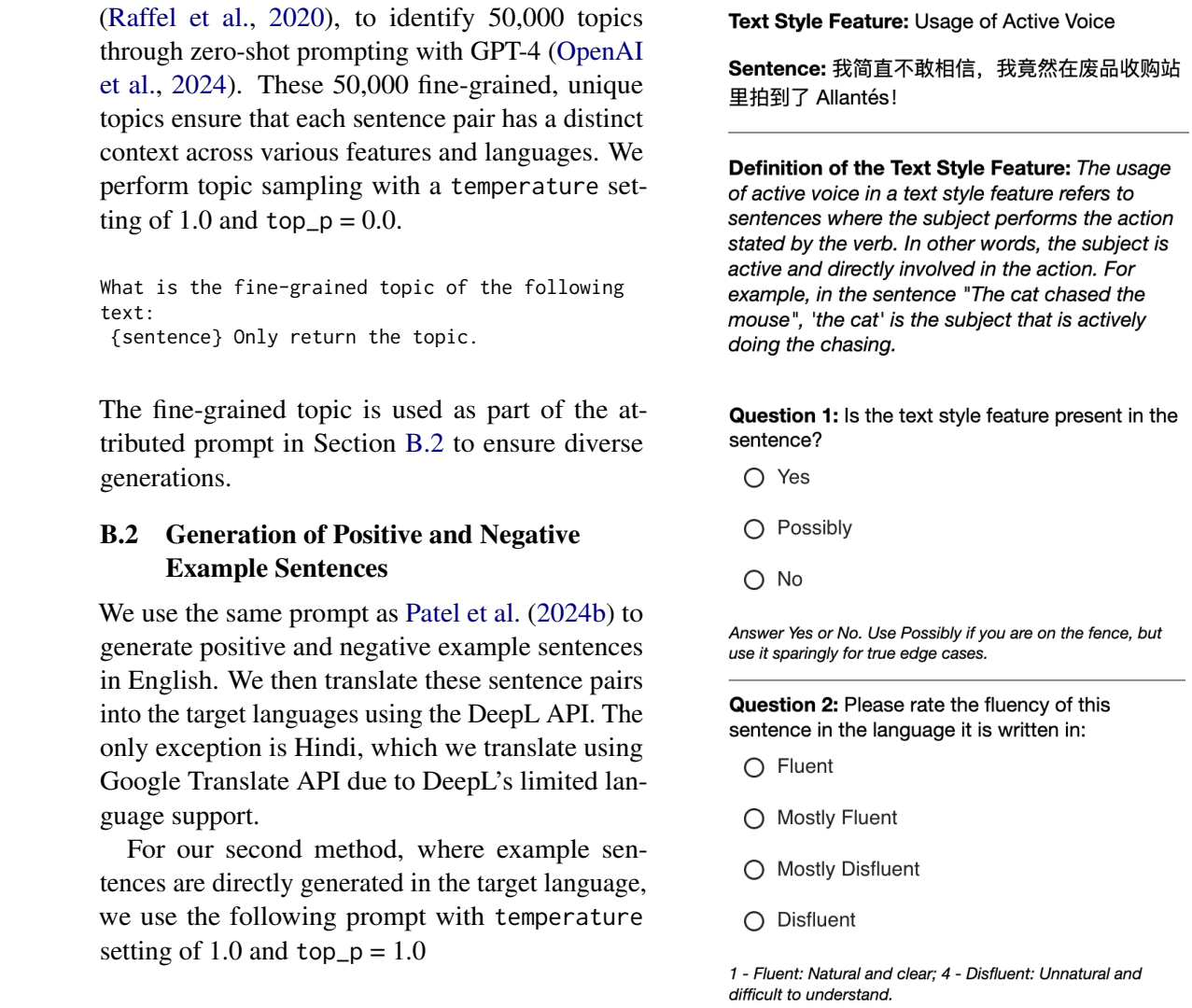

Figure 3: Instances from the annotation interface.

In Figure 3, we show an instance from the human annotation interface. We first asked the annotators

Language	Baseline Feature Presence	Feature Presence	Baseline Fluency	Fluency	Baseline Diversity	Diversity	Baseline Similarity	Similarity
ar	0.5	0.7475	0.5	0.9526	0.8278	0.8245	0.9232	0.9156
de	0.5	1.0000	0.5	0.7708	0.8345	0.8341	0.8799	0.9171
es	0.5	0.8125	0.5	0.9853	0.8449	0.8478	0.8567	0.9298
fr	0.5	0.7391	0.5	0.9855	0.8483	0.8404	0.8573	0.9224
hi	0.5	0.7595	0.5	0.9958	0.8588	0.8253	0.9468	0.8903
ja	0.5	0.6667	0.5	0.8889	0.8528	0.8321	0.8514	0.8761
ko	0.5	0.8000	0.5	0.8972	0.8540	0.8214	0.8652	0.9286
ru	0.5	0.8000	0.5	0.8972	0.8542	0.8097	0.8713	0.9171
zh-hans	0.5	0.7475	0.5	0.9526	0.8571	0.8220	0.8729	0.9322
Average	**0.5**	**0.7859**	**0.5**	**0.9251**	**0.8480**	**0.8286**	**0.8805**	**0.9144**

Table 4: Human and automatic evaluations on our synthetic dataset.

whether a given style feature was present in a sentence in their chosen language. We also provided a definition for each style feature to help the annotators in their decision. The annotators then had to rate the fluency of the sentence by selecting one answer among: "Fluent", "Mostly Fluent", "Mostly Disfluent", "Disfluent".

Our annotators were undergraduate and graduate students from a NLP class and were offered extra credit for their participation in the study. Each instance was annotated by at least three annotators: three for languages with fewer native speakers such as Arabic and Russian; over ten for languages with a large number of native speakers, such as Chinese. We used Krippendorff's Alpha (Krippendorff, 2011) to measure inter-annotator agreement, which indicated moderate agreement of 0.4247 ± 0.1719.

A breakdown of the 13,651 annotated instances by language:

- Chinese: 6,553
- Spanish: 1,420
- Hindi: 1,186
- French: 1,003
- Korean: 982
- Japanese: 524
- German: 357
- Arabic: 344
- Russian: 282

D Dataset Evaluation

In Table 4, we show the per language results of the human and automatic evaluations for our synthetic dataset. Our synthetic dataset is comparable to a reference dataset compiled by Scherrer (2020) in terms of feature presence, fluency, diversity, and similarity. Note that baselines shown for feature presence and fluency are just 0.5 to represent random guessing.

E Training Details

Table 5 contains details about the hyperparameters used for training. More exact training details can be found in the source code provided in the supplementary materials for this work.

Hyperparameter	Value
Model	`xlm-roberta-base`
Hardware	4x or 8x NVIDIA RTX A6000
Distributed Protocol	PyTorch FSDP
Data Type	`torch.bfloat16`
Loss Function	`TripletLoss` (Schroff et al., 2015)
Triplet Loss Margin	0.1
LoRA (Hu et al., 2021)	`all-linear, r=8, lora_alpha=8, lora_dropout=0.0`
Optimizer	`adamw_torch`
Learning Rate	1e-4
Weight Decay	0.01
Learning Rate Scheduler	`linear`
Warmup Steps	0
Batch Size	384
Train-Validation Split	90/10%
Early Stopping Threshold	0.0
Early Stopping Patience	1 epoch

Table 5: Hyperparameters selected for contrastive learning training experiments.

F Instances from the Multilingual and Cross-lingual SoC Benchmarks

In our multilingual SoC benchmark, anchor (a) has the same style and different content from a positive

Ablation Condition	Ablated Features and Languages	
Out-of-Domain	**Ablated Style Features:**	Usage of Formal Tone, Usage of Contractions, Usage of Numerical Substitution, Complex Sentence Structure, Usage of Positive Tone, Usage of Offensive Tone, Usage of Polite Tone
Out-of-Distribution	**Ablated Style Features:**	Usage of Formal Tone, Usage of Polite Tone, Fluency in Sentence Construction, Usage of Only Uppercase Letters, Usage of Only Lowercase Letters, Incorporation of Humor, Usage of Sarcasm, Usage of Contractions, Usage of Numerical Substitution, Usage of Text Emojis, Usage of Emojis, Presence of Misspelled Words, Complex Sentence Structure, Usage of Long Words, Usage of Polite Tone, Usage of Offensive Tone
No Language Overlap	**Ablated Languages:**	ar (Arabic), de (German), es (Spanish), fr (French), hi (Hindi), ja (Japanese), ru (Russian)

Table 6: Style features and languages ablated for **Out-of-Domain**, **Out-of-Distribution**, and **No Language Overlap**, the three ablation conditions in our ablation study.

	Multi-lingual SoC				Cross-lingual SoC		
Features Tested	**Simplicity**	**Formality**	**Toxicity**	**Positivity**	**Formality**	**Retained Perf (%)** m	c
In-Domain	0.36	0.71	0.37	0.30	0.53	100%	100%
Out of Domain	0.29	0.63	0.31	0.23	0.44	75%	74%
Out of Distribution	0.33	0.39	0.26	0.32	0.40	75%	62%
No Language Overlap	0.27	0.51	0.41	0.32	0.52	89%	97%

Table 7: Results of the ablation study for MSTYLEDISTANCE embeddings on the SoC benchmarks.

example (pos), and the same content but different style from a negative example (neg). The anchor and the pos and neg sentences are in the same language. The tested model needs to successfully pair a with pos (rather than a and neg). Cross-lingual SoC has the same setup as multilingual SoC, except that the pos and neg examples are in a different language than the anchor. Figure 4 contains instances of each benchmark.

G Ablation Details and Results

Details about the ablated features and languages can be found in Table 6. Table 7 contains the results of the ablation study for MSTYLEDISTANCE embeddings on the SoC benchmarks under three generalization conditions: Out of Domain, Out of Distribution, and No Language Overlap. For multilingual SoC, we use all four style features: simplicity, formality, toxicity, and positivity. For the cross-lingual variant, we only use formality.

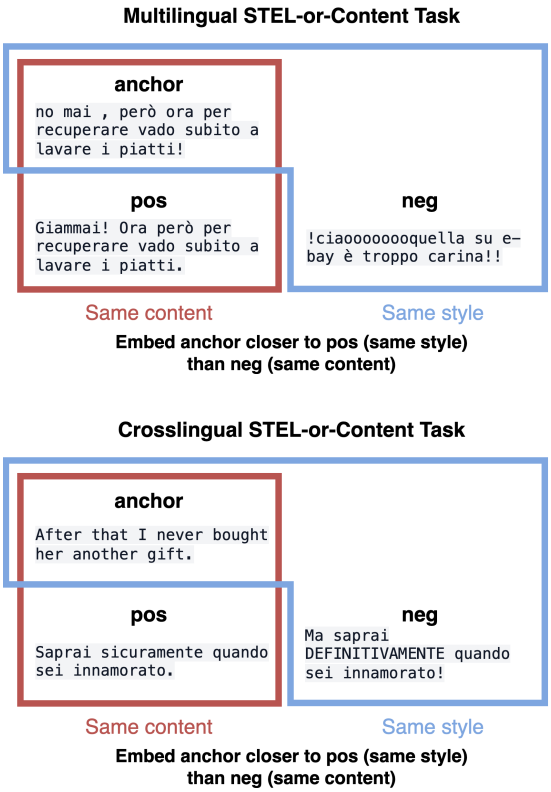

Figure 4: Instances from our multilingual and cross-lingual SoC benchmarks. For multilingual SoC, the anchor is in the same language as the pos and neg sentences. For cross-lingual SoC, the anchor is in a different language from the pos and neg sentences.

H Full SoC Results

Language	Wegmann et al. (2022)	STYLEDISTANCE	xlm-roberta-base	LISA	mSTYLEDISTANCE
Simplicity					
de	0.23	0.06	0.00	0.00	**0.24**
en	0.26	0.32	0.05	0.00	**0.36**
fr	0.29	0.33	0.22	0.12	**0.46**
it	0.21	0.15	0.08	0.03	**0.48**
ja	0.09	0.05	0.01	**0.48**	0.14
pt-br	0.10	0.07	0.04	0.03	**0.15**
ru	0.26	0.24	0.07	0.15	**0.38**
sl	0.43	0.43	0.46	0.39	**0.69**
average	0.23	0.21	0.12	0.15	**0.36**
Formality					
fr	0.70	0.81	0.16	0.06	**0.82**
it	0.64	0.63	0.18	0.10	**0.69**
pt-br	0.56	0.57	0.15	0.11	**0.62**
average	0.63	0.67	0.16	0.09	**0.71**
Toxicity					
am	0.35	0.29	0.24	0.21	**0.53**
ar	0.05	0.04	0.02	0.10	**0.18**
de	0.01	0.02	0.01	0.00	**0.28**
en	**0.56**	0.48	0.09	0.08	0.51
es	0.26	0.20	0.13	0.05	**0.35**
hi	0.15	0.09	0.09	0.15	**0.37**
ru	0.18	0.16	0.13	0.09	**0.61**
uk	0.07	0.05	0.04	0.02	**0.25**
zh	0.05	0.02	0.04	0.07	**0.23**
average	0.19	0.15	0.09	0.09	**0.37**
Positivity					
bn	0.27	0.13	0.04	0.23	**0.32**
en	**0.21**	0.20	0.03	0.19	0.18
hi	0.11	0.10	0.04	0.14	**0.22**
mag	0.09	0.08	0.08	0.13	**0.41**
ml	0.32	0.28	0.10	0.27	**0.39**
mr	0.19	0.18	0.03	0.17	**0.22**
or	0.27	0.19	0.08	0.24	**0.35**
pa	0.18	0.15	0.06	0.17	**0.23**
te	0.39	0.34	0.20	0.29	**0.40**
ur	0.24	0.20	0.08	0.28	**0.26**
average	0.23	0.18	0.07	0.21	**0.30**
Formality (cross-lingual)					
fr-it	0.47	0.51	0.22	0.28	**0.53**
fr-pt	0.45	0.48	0.19	0.29	**0.52**
it-fr	0.48	**0.53**	0.18	0.26	**0.53**
it-pt	0.41	0.45	0.19	0.27	**0.52**
pt-fr	0.46	**0.53**	0.17	0.27	**0.53**
pt-it	0.42	0.47	0.21	0.27	**0.52**
average	0.45	0.49	0.19	0.27	**0.53**

Table 8: Performance obtained for different features and languages in the multilingual and cross-lingual STEL-or-content benchmarks. For the cross-lingual SoC evaluation, "a-b" means that the anchor sentences were all in language a and alternative sentences were all in language b. MSTYLEDISTANCE leads in cross-lingual and overall performance.

Style Feature	Positive and Negative Prompts	Style Feature Definition	Excluded In
Usage of Conjunctions	Positive: With conjunctions Negative: Less frequent conjunctions	The "Usage of Conjunctions" text style feature refers to the use of words that connect clauses or sentences. Conjunctions are words like "and", "but", "or", "so", "because", etc. They are used to make sentences longer, more complex, or to show the relationship between different parts of a sentence.	
Usage of Numerical Substitution	Positive: With number substitution Negative: Without number substitution	Numerical substitution refers to the practice of replacing certain letters in words with numbers that visually resemble those letters. For example, replacing the letter 'e' with the number '3' in the word 'hello' to make it 'h3llo'. This is a common feature in internet slang and informal digital communication.	Arabic, Hindi, Japanese, Korean, Chinese
Usage of Words Indicating Affective Processes	Positive: Affective processes Negative: Cognitive processes	The text style feature "Usage of Words Indicating Affective Processes" refers to the use of words that express emotions, feelings, or attitudes. These could be words that show happiness, sadness, anger, fear, surprise, or any other emotional state. The presence of such words in a text indicates that the writer is expressing some form of emotional reaction or sentiment.	
Usage of Metaphors	Positive: With metaphor Negative: Without metaphor	The "Usage of Metaphors" text style feature refers to the presence of phrases or sentences in the text that describe something by comparing it indirectly to something else. This is often done to make a description more vivid or to explain complex ideas in a more understandable way. For example, saying "time is a thief" is a metaphor because it's not literally true but it helps to convey the idea that time passes quickly and can't be regained.	
Usage of Long Words	Positive: Long average word length Negative: Short average word length	The "Usage of Long Words" text style feature refers to the frequency or prevalence of long words, typically those with more than six or seven letters, in a given text. This style feature is often used to measure the complexity or sophistication of the text. If a text has many long words, it is said to have a high usage of long words.	Arabic, Japanese, Korean, Chinese
Usage of Uppercase Letters	Positive: With uppercase letters Negative: Without uppercase letters	The usage of uppercase letters as a text style feature refers to the frequency or manner in which capital letters are used in a text. This could be for emphasis, to denote shouting or strong emotions, or to highlight specific words or phrases. It's not just about the start of sentences or proper nouns, but also about other uses of capital letters in the text.	Arabic, Hindi, Japanese, Korean, Chinese
Usage of Articles	Positive: With articles Negative: Less frequent articles	The "Usage of Articles" text style feature refers to how often a text uses words like "a", "an", and "the". These words are called articles and they are used before nouns. This feature measures the frequency of these articles in a given text.	Arabic, Hindi, Japanese, Korean, Russian, Chinese
Usage of Text Emojis	Positive: Text Emojis Negative: No Emojis	The text style feature "Usage of Text Emojis" refers to the inclusion of emoticons or smileys in the text. These are combinations of keyboard characters that represent facial expressions or emotions, such as :-D for a big grin or happy face. The presence of these symbols in a text indicates the use of this style feature.	
Usage of Nominalizations	Positive: With nominalizations Negative: Without nominalizations	Nominalizations refer to the use of verbs, adjectives, or adverbs as nouns in a sentence. This style feature is often used to make sentences more concise or formal. For example, "the investigation of the crime" is a nominalization of "investigate the crime".	
Frequent Usage of Function Words	Positive: With function words Negative: Less frequent function words	The text style feature "Frequent Usage of Function Words" refers to the regular use of words that have little meaning on their own but work in combination with other words to express grammatical relationships. These words include prepositions (like 'in', 'at', 'on'), conjunctions (like 'and', 'but', 'or'), articles (like 'a', 'an', 'the'), and pronouns (like 'he', 'they', 'it').	
Usage of Self-Focused Perspective or Words	Positive: Self-focused Negative: Third-person singular	The "Usage of Self-Focused Perspective or Words" text style feature refers to the use of words or phrases that focus on the speaker or writer themselves. This includes the use of first-person pronouns like "I", "me", "my", "mine", and "myself", or statements that express the speaker's personal thoughts, feelings, or experiences.	
Usage of Formal Tone	Positive: Formal Negative: Informal	The "Usage of Formal Tone" text style feature refers to the use of language that is polite, impersonal and adheres to established conventions in grammar and syntax. It avoids slang, contractions, colloquialisms, and often uses more complex sentence structures. This style is typically used in professional, academic, or official communications.	
Usage of Emojis	Positive: With Emojis Negative: No Emojis	The "Usage of Emojis" text style feature refers to the inclusion of emojis, or digital icons, in a text. Emojis are often used to express emotions, ideas, or objects without using words. If a text contains emojis, it has this style feature.	
Usage of Offensive Language	Positive: Offensive Negative: Non-Offensive	The "Usage of Offensive Language" text style feature refers to the presence of words or phrases in the text that are considered rude, disrespectful, or inappropriate. These can include swear words, slurs, or any language that could be seen as insulting or derogatory.	
Usage of Present Tense and Present-Focused Words	Positive: Present-focused Negative: Future-focused	The text style feature "Usage of Present Tense and Present-Focused Words" refers to the use of verbs in the present tense and words that focus on the current moment or situation. This means the text is primarily discussing events, actions, or states that are happening now or general truths. It's like the text is talking about what is happening in the present time.	
Presence of Misspelled Words	Positive: Sentence With a Few Misspelled Words Negative: Normal Sentence	The text style feature "Presence of Misspelled Words" refers to the occurrence of words in a text that are not spelled correctly according to standard dictionary spelling. This could be due to typing errors, lack of knowledge about the correct spelling, or intentional for stylistic or informal communication purposes.	
Incorporation of Humor	Positive: With Humor Negative: Without Humor	The "Incorporation of Humor" text style feature refers to the use of language, phrases, or expressions in a text that are intended to make the reader laugh or feel amused. This could include jokes, puns, funny anecdotes, or witty remarks. It's all about adding a touch of comedy or light-heartedness to the text.	

Style Feature Name	Positive and Negative Prompts	Style Feature Definition	Excluded In
Usage of Personal Pronouns	Positive: With personal pronouns Negative: Less frequent pronouns	The "Usage of Personal Pronouns" text style feature refers to the use of words in a text that refer to a specific person or group of people. These words include "I", "you", "he", "she", "it", "we", and "they". The presence of these words in a text can indicate a more personal or direct style of communication.	
Fluency in Sentence Construction	Positive: Fluent sentence Negative: Disfluent sentence	"Fluency in Sentence Construction" refers to the smoothness and ease with which sentences are formed and flow together. It involves using correct grammar, appropriate vocabulary, and logical connections between ideas. A text with this feature would read smoothly, without abrupt changes or awkward phrasing.	
Usage of Only Uppercase Letters	Positive: All Upper Case Negative: Proper Capitalization	The usage of only uppercase letters style feature refers to the practice of writing all the letters in a text in capital letters. This means that every single letter in the text, whether at the beginning, middle, or end of a sentence, is capitalized. It's like the 'Caps Lock' key on your keyboard is always turned on while typing the text.	Arabic, Hindi, Japanese, Korean, Chinese
Usage of Self-Focused Perspective or Words	Positive: Self-focused Negative: Inclusive-focused	The "Usage of Self-Focused Perspective or Words" text style feature refers to the use of words or phrases that focus on the speaker or writer themselves. This includes the use of first-person pronouns like "I", "me", "my", "mine", and "myself", or statements that express the speaker's personal thoughts, feelings, or experiences.	
Usage of Pronouns	Positive: With pronouns Negative: Less frequent pronouns	The "Usage of Pronouns" text style feature refers to the frequency and types of pronouns used in a text. Pronouns are words like 'he', 'she', 'it', 'they', 'we', 'you', 'I', etc., that stand in place of names or nouns in sentences. This feature can indicate the level of personalization, formality, or perspective in a text.	
Usage of Words Indicating Cognitive Processes	Positive: Cognitive process Negative: Perceptual process	The text style feature "Usage of Words Indicating Cognitive Processes" refers to the use of words that show thinking or mental processes. These words can express understanding, knowledge, belief or doubt. For example, words like 'think', 'know', 'believe', 'understand' are used to indicate cognitive processes.	
Complex Sentence Structure	Positive: Complex Negative: Simple	The "Complex Sentence Structure" text style feature refers to sentences that contain multiple ideas or points, often connected by conjunctions (like 'and', 'but', 'or') or punctuation (like commas, semicolons). These sentences often include dependent clauses, which are parts of the sentence that can't stand alone as a complete thought, alongside independent clauses, which can stand alone. In simpler terms, if a sentence has more than one part and these parts are linked together in a way that they give more detailed information or express multiple thoughts, it has a complex sentence structure.	
Positive Sentiment Expression	Positive: Positive Negative: Negative	Positive Sentiment Expression is a text style feature that refers to the use of words, phrases, or expressions that convey a positive or optimistic viewpoint or emotion. This could include expressions of happiness, joy, excitement, love, or any other positive feelings. The text is considered to have this feature if it makes the reader feel good or positive after reading it.	
Usage of Numerical Digits	Positive: With digits Negative: Less frequent digits	The "Usage of Numerical Digits" text style feature refers to the presence and use of numbers in a text. This includes any digit from 0-9 used alone or in combination to represent quantities, dates, times, or any other numerical information.	
Usage of Words Indicating Affective Process	Positive: Affective process Negative: Perceptual process	The "Usage of Words Indicating Affective Process" text style feature refers to the use of words that express emotions, feelings, or attitudes. These words can show positive or negative sentiments, like happiness, anger, love, or hate. If a text uses a lot of these words, it means the writer is expressing a lot of emotion or personal feelings.	
Usage of Active Voice	Positive: Active Negative: Passive	The usage of active voice in a text style feature refers to sentences where the subject performs the action stated by the verb. In other words, the subject is active and directly involved in the action. For example, in the sentence "The cat chased the mouse", 'the cat' is the subject that is actively doing the chasing.	
Usage of Only Lowercase Letters	Positive: All Lower Case Negative: Proper Capitalization	The style feature "usage of only lowercase letters" refers to the practice of writing all words in a text with small letters only, without using any capital letters. This means that even the first word of a sentence, proper nouns, or the pronoun 'I' are not capitalized. It's like writing a whole text without ever pressing the shift key on your keyboard.	Arabic, Hindi, Japanese, Korean, Chinese
Frequent Usage of Common Verbs	Positive: With common verbs Negative: Less frequent common verbs	The text style feature "Frequent Usage of Common Verbs" refers to the regular use of basic action words in a text. These are often simple, everyday verbs that are widely used in language, such as 'is', 'have', 'do', 'say', 'go', etc. If a text frequently uses these common verbs, it has this style feature.	
Usage of Prepositions	Positive: With prepositions Negative: Less frequent prepositions	The "Usage of Prepositions" text style feature refers to the use of words that link nouns, pronouns, or phrases to other words within a sentence. These words often indicate location, direction, time, or manner. Examples of prepositions include words like "in", "at", "on", "over", "under", "after", and "before".	
Usage of Self-Focused Language	Positive: Self-focused Negative: Audience-focused	The "Usage of Self-Focused Language" text style feature refers to the use of words or phrases that focus on the speaker or writer themselves. This includes the use of first-person pronouns like "I", "me", "my", "mine", and "myself". It's a way of writing or speaking where the person is often referring to their own thoughts, feelings, or experiences.	
Usage of Certain Tone	Positive: Certain Negative: Uncertain	This text style feature refers to the use of a confident tone in writing, where the author avoids using uncertain words or phrases such as 'I think', 'might', or 'seems'. This results in a text that appears more assertive and sure of the information being presented.	
Usage of Present-Focused Tense and Words	Positive: Present-focused Negative: Past-focused	The "Usage of Present-Focused Tense and Words" text style feature refers to the use of verbs in the present tense and words that focus on the current moment or situation. This means the text is primarily discussing events, actions, or states that are happening right now or generally true.	

Style Feature Name	Positive and Negative Prompts	Style Feature Definition	Excluded In
Usage of Sarcasm	Positive: With sarcasm Negative: Without sarcasm	The "Usage of Sarcasm" text style feature refers to the presence of statements or expressions in the text that mean the opposite of what they literally say, often used to mock or show irritation. This style is often characterized by irony, ridicule, or mockery, and is used to express contempt or to criticize something or someone in a humorous way.	
Usage of Self-Focused Perspective or Words	Positive: Self-focused Negative: You-focused	The "Usage of Self-Focused Perspective or Words" text style feature refers to the use of words or phrases that focus on the speaker or writer themselves. This includes the use of first-person pronouns like "I", "me", "my", "mine", and "myself", or statements that express the speaker's personal thoughts, feelings, or experiences.	
Frequent Usage of Punctuation	Positive: With frequent punctuation Negative: Less Frequent punctuation	The text style feature "Frequent Usage of Punctuation" refers to the regular and abundant use of punctuation marks such as commas, periods, exclamation points, question marks, etc., in a piece of text. This style feature is present when the writer often uses these symbols to structure their sentences, express emotions, or emphasize certain points.	
Usage of Polite Tone	Positive: Polite Negative: Impolite	The "Usage of Polite Tone" text style feature refers to the use of respectful and considerate language in a text. This can include using words like 'please', 'thank you', or phrases that show deference or respect to the reader. It's about making the text sound courteous and respectful, rather than demanding or rude.	
Usage of Contractions	Positive: With contractions Negative: Without contractions	The "Usage of Contractions" text style feature refers to the use of shortened forms of words or phrases in a text. These are typically formed by omitting certain letters or sounds and replacing them with an apostrophe, such as "don't" for "do not" or "I'm" for "I am". If a text frequently uses such shortened forms, it has this style feature.	Arabic, Hindi, Japanese, Korean, Russian, Chinese
Frequent Usage of Determiners	Positive: With determiners Negative: Less frequent determiners	The text style feature "Frequent Usage of Determiners" refers to the regular use of words that introduce a noun and give information about its quantity, proximity, definiteness, etc. These words include 'the', 'a', 'an', 'this', 'that', 'these', 'those', 'my', 'your', 'his', 'her', 'its', 'our', 'their'. If a text often uses such words, it has this style feature.	

Table 9: The 40 style features addressed in our experiments. The 'Excluded in' column indicates that a particular feature was omitted from our dataset due to its inapplicability to a specific language.