Bubble up – A Fine-tuning Approach for Style Transfer to Community-specific Subreddit Language

Alessandra Zarcone
Technische Hochschule Augsburg
Fakultät für Informatik
Augsburg, Germany
alessandra.zarcone@hs-augsburg.de

Fabian Kopf
Technische Hochschule Augsburg
Fakultät für Informatik
Augsburg, Germany
fabian.kopf@hs-augsburg.de

Abstract

Different online communities (social media bubbles) can be identified with their use of language. We looked at different social media bubbles and explored the task of translating between the language of one bubble into another while maintaining the intended meaning. We collected a dataset of Reddit comments from 20 different Subreddits and for a smaller subset of them we obtained style-neutral versions generated by a large language model. Then we used the dataset to fine-tune different (smaller) language models to learn style transfers between social media bubbles. We evaluated the models on unseen data from four unseen social media bubbles to assess to what extent they had learned the style transfer task and compared their performance with the zero-shot performance of a larger, non-fine tuned, language model. We show that with a small amount of fine-tuning the smaller models achieve satisfactory performance, making them more attractive than a larger, more resource-intensive model.

1 Introduction

Language on social media is not just a way to exchange information but it is a mean to effectively create a community identity, with each virtual community having their own, identifiable language (Baym, 2003; Gnach, 2017; Rheingold, 2000). For example, in some communities of investors it is common to use the term "HODL" to mean *hold* to describe holding a share (Duggan, 2023). Beyond the vocabulary, also the use of emojis and hashtags or the type of grammar can help identify the language of a social media community or bubble and can at the same time make it difficult for outsiders to understand what it is being said (Smith and Sturges, 1969).

We explore the task of translating between the language of one bubble into the language of another while maintaining the intended meaning of the original sentence (see Hovy, 1987 for a discussion of semantics vs. style). We define the task in the following way: given sentence A and sentence B, the task is to transfer the style of sentence B to sentence A while maintaining the original meaning of sentence A.

We demonstrate how to perform style transfers from neutral, non-style-marked English to a community-specific style (namely, the style or community-specific language of a Subreddit discussion forum). Our fine-tuning approach does not use a large amount of parallel data to specialize to a specific type of style transfer, but rather aims at working with a small amount of resources to learn the general task of style transfer from one social media bubble to another.

The task of transferring between the styles of different social media bubbles can provide interesting insights into what it means to perform a style transfer in the social media domain, where the style itself carries information about membership within a certain community. At the same time, looking at how style transfers can be automatically performed can contribute to the future detection of automatically-translated posts, which in turn can be used with malicious intents, for example to spread rumours or fake news.

We provide the dataset we collected and employed in this study as well as an evaluation of the datasets. The dataset itself provides a resource for future studies on the different styles used by different social media bubbles. We present our style transfer models, which we evaluate with regard to their ability to effectively perform the style transfer as well as their ability to maintain the original sentence meaning. The fine-tuned models can achieve satisfactory performance with a small amount of fine-tuning, which makes them more attractive than using a larger, more resource-intensive zero-shot approach.

Subreddit	Category	Participants
antiwork	politics	2.5M
atheism	religion	2.8M
Conservative	politics	1.0M
conspiracy	conspiracy theories	1.9M
dankmemes	memes	5.9M
gaybros	LGBT	380 000
leagueoflegends	computer games	6.3M
lgbt	LGBT	1.0M
Libertarian	politics	511 800
linguistics	science	297 800
MensRights	politics	348 200
news	news	26M
offbeat	news	690 500
politicalcompassmemes	memes	572 800
politics	politics	8.3M
teenager	memes	5.9M
TrueReddit	news	519 900
TwoXChromosomes	gender	13.5M
wallstreetbets	finance	13.8M
world news	news	31.5M

Table 1: The 20 Subreddits considered for our data collection.

2 Previous Work

Style transfer with parallel corpora The task of style transfer can be addressed by using parallel corpora, where to each sentence in the source style corresponds a sentence in the target style with the same meaning. Parallel corpora are employed for example to train sequence-to-sequence models to transfer an informal style to a more formal style (Rao and Tetreault, 2018), or to make a text more polite (Danescu-Niculescu-Mizil et al., 2013), or to transfer from Shakespeare's English to modern English (Xu et al., 2012).

Style transfer without parallel corpora It is not always possible or practical to collect a parallel corpus to train a style transfer model, which in the end would be specialized mostly on one specific style transfer. Thus more recent approaches have attempted at performing style transfer without resorting to parallel corpora, addressing the need to keep style and meaning separated (Shen et al., 2017; Bao et al., 2019; John et al., 2019), for example approximating the text content using bag-of-word vectors and aiming at predicting it (John et al., 2019), or training transformer models which could be fine-tuned to produce a network for each specific style transfer (Goyal et al., 2021). Luo et al. (2019) used pseudoparallel datasets with different styles and unrelated content and employed two different models to optimize the semantic similarity of source and target content and the style similarity between source and target styles. Generative models work particularly well for this: Riley et al.'s (2021) TextSETTR for example extracts a style vector using the T5 sequence-to-sequence model (Raffel et al., 2020) and then use it to condition the decoder during style transfer. For more style transfer approaches employing generative models see also Li et al. (2018); Lample et al. (2018, 2019); Krishna et al. (2020); Reid and Zhong (2021).

Prompt-based style transfer Large generative transformer models such as GPT3.5 (Brown et al., 2020) or GPTNeoX (Black et al., 2022) allow for zero-shot text style transfer. Reif et al. (2022) frame style transfer as sentence rewriting with natural language instruction, using prompts such as *"Here is some text: That is an ugly dress. Here is a rewrite of the text, which is more positive:"*. The text-davinci-003 GPT-3.5 model would for example rewrite it as *"That dress has an interesting style"*. Reif et al. (2022) also propose to provide several examples of style transfer as part of the prompt to obtain better results. Suzgun et al. (2022) additionally suggest generating multiple target candidates and ranking them regarding similarity to target content, strength of target style and fluency, showing that this approach is more suitable to smaller pre-trained language models and thus a more resource-effective approach.

3 The Reddit Comments Dataset

3.1 Data collection

We collected community-specific language data from Reddit. Reddit is a social network which is used by its users to discuss a wide range of topic. Users can post text, links, images or videos, which can be commented and / or rated by the other users. The discussions on Reddit are organized in the so-called Subreddits, which specialize in different topics and interests and arguably constitute some sort of social media bubble. We observed that the stylistic homogeneity within each Subreddit may vary: Subreddits dealing with more general topics, the writing style of the user is typically not marked, whereas Subreddits that deal with special topics and have a specific, delimited circle of users (in particular, Subreddits on political topics), the style is more homogeneous and more easily identifiable.

We chose 20 Subreddits of varying degree of popularity - the list of Subreddits along with their topic and number of participants is provided in Table 1. The rationale we followed was to select Subreddits with a variety of different topics, showing a wide style variance between each other but a homogeneous style within each other - that is, showing a clearly-identifiable "language". This was based on our own impression, which we validated with the dataset evalauation (see below, section 3.5). We also aimed at providing some sort of balance between Subreddits of opposite positions (e.g. *TwoXChromosomes* for *MensRights*, *Libertarian* for *Conservative*).

The text in the Subreddits is easily accessible thanks to the Reddit API[1] as well as the Pushshift API provided by Baumgartner et al. (2020). We collected comments to the posts using the Subreddit Comment Downloader[2]. We selected comments which were between 10 and 512 tokens long, which were not `[removed]` or `[deleted]` and which did not contain any url.

3.2 Perplexity-based selection

In order to ensure that the crawled data was showing enough style-specific features to be meaningful for a style-transfer task, we aimed at selecting comments with easily-identifiable style features. We assumed that a large language model such as GPT-2 would yield higher perplexity values for more stylistically-marked comments, which deviate the most from standard language. For our training and evaluation set, we selected the 150 comments with highest GPT-2 perplexity for each Subreddit.

3.3 Creation of style-neutral versions

For our training and evaluation, we also needed style-neutral versions of the Reddit comments. Comparable available datasets were proposed by Rao and Tetreault (2018), Wu et al. (2020), Atwell et al. (2022), Logacheva et al. (2022), Liu et al. (2020) included human-generated style-transferred sentences . Due to time and resource limitations, we used synthetic data generation in order to obtain style-neutral versions of the comments. Arguably, a translation from style-marked to style-neutral is an easier task than a translation in the opposite direction or between different styles, because it requires generating text of a more standard variety, which can be more easily produced by a generic language model. We thus expect that a large language model (LLM) would be particularly effective at this when the neutral style is used as a target style (see also Josifoski et al., 2023 for an example of synthetic data use leveraging the asymmmetry in task difficulty).

We follow the zero-shot approach in Reif et al. (2022). The style-neutral versions were generated using the text-davinci-003 version of the GPT-3.5 Model.

We made a first attempt with a prompt similar to the one used by Reif et al. (2022) for their style transfer to formal style:

"Here is some text {[...]} Here is a rewrite of the text, which is more formal: {"

With the text-davinci-003 model, however, the results obtained with these prompts were not really style-neutral but rather too pompous, as in the following example:

"Here is some text: {Just saying, no brag or anything, but I make $35/hr off construction knowledge. I started low but got good at it.} Here is a rewrite of the text, which is more formal: {"

⇒ *"I humbly state that I am able to earn $35 per hour through my expertise in construction. Initially, I began at a lower rate, but I have since developed a high level of proficiency}".*

We thus changed the prompt to require a "more neutral" continuation, leading to better style-neutral versions:

⇒ *"I have experience in construction and I make $35/hr. I started out with a lower rate, but I have become more skilled over time.}".*

3.4 Dataset description

The dataset includes more than 49K comments from the 20 Subreddits, their token length and perplexity scores, along with the name of the Subreddit, a timestamp and a link to the original comment. For two smaller subsets of the dataset (training set and evaluation set), neutral-style versions were provided for 150 comments in each Subreddit (the 150 high-perplexity ones). This was in line with our goal of limiting the use of the largest language model, which we used to obtain the style-neutral versions and to create a small dataset for fine-tuning. The training set contains comments from 16 Subreddits, the evaluation set from 4 Subreddits. The dataset is publicly available on Zenodo: `https://doi.org/10.5281/zenodo.8023142` (Kopf, 2023).

[1] `https://www.reddit.com/dev/api/`

[2] `https://github.com/pistocop/subreddit-comments-dl`

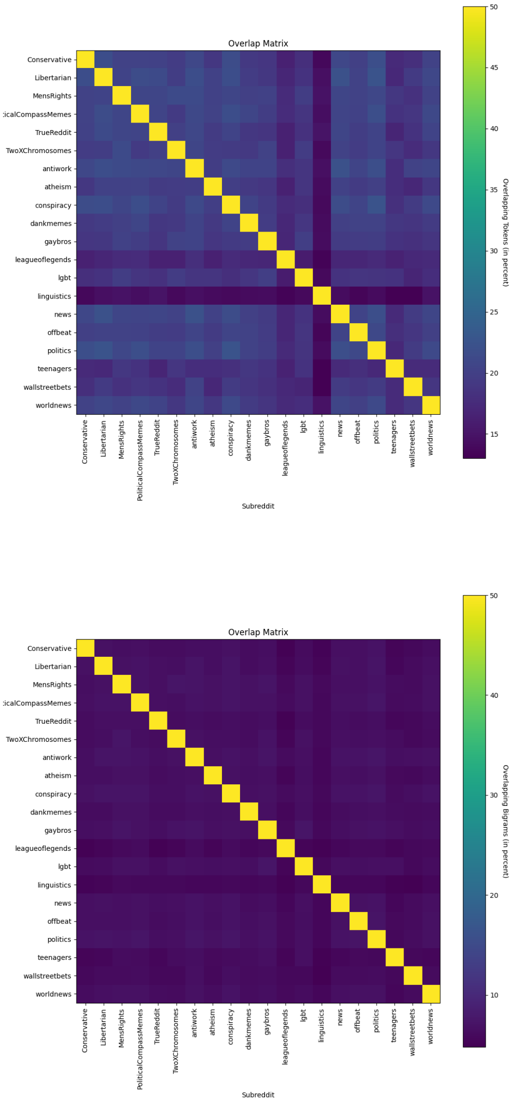

Figure 1: Lexical overlap (unigrams, top, and bigrams, bottom) between the high-perplexity comments for different Subreddit pairs.

3.5 Dataset Evaluation

Lexical Overlap In order to evaluate if the different Subreddits differ with regard to their lexical choices, we compared the percentage of shared lemmas and shared lemmatized bigrams between all possible Subreddit pairs for the high-perplexity comments (perplexity > 100). The lexical overlap scores (visualized in Figure 1) show that in particular the high-perplexity comments are not only clearly different from the standard language, but are also easily distinguishable from the other Subreddits, making them particularly suitable for our training and evaluation.

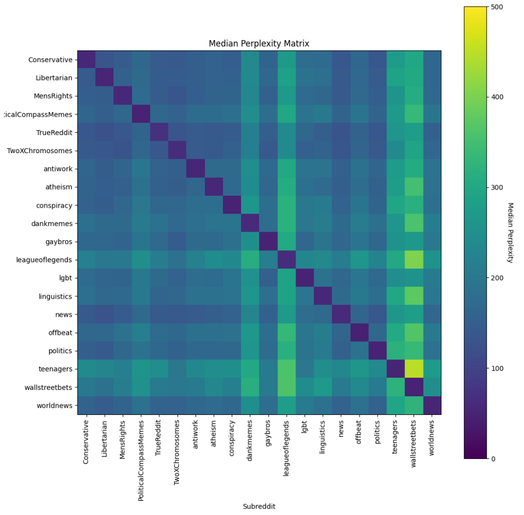

Figure 2: Median perplexity from fine-tuned GPT-2 language models between different Subreddits for the high-perplexity comments (high perplexity for the generic language model). On the y-axis are the fine-tuned language models, on the x-axis the comments of the Subreddit, for which the scores were computed.

Perplexity Perplexity scores were also employed to evaluate differences between the Subreddits beyond lexical overlap. We thus fine-tuned a GPT-2 model for each Subreddit and used it to compute perplexity scores for the comments in the other Subreddits. We expect the model for a specific Subreddit to be "surprised" when exposed to the style of a different Subreddit.

The perplexity scores are rather homogeneous with the exception of the Subreddits "leagoflegend", "teenagers" and "wallstreetbets", whose comments yield high perplexity scores in all style-specific language models - with the obvious exception of the style-specific model fine-tuned on the comments from this Subreddit.

Neutral-style versions We compared our dataset with the GYAFC dataset (Rao and Tetreault, 2018), a large human-labelled parallel datasets often used to evaluate formal/informal style transfer systems, in order do evaluate how the quality of our LLM-generated neutral-style versions compared with the quality of human-generated data. We used BERTScore (Zhang et al., 2019) and chrF++ (Popović, 2015, 2017) to compare our dataset with GYAFC with regard to the semantic similarity between style-specific and neutral version. While the BERTScore and chrF++ for the neutral-versions generated with the "more formal" and "more neutral" prompt are marginally lower than the human-

Data	F1-Score	Precision	Recall	chrF++	Perplexity
more formal	0.79	0.79	0.78	44.97	36.73
more neutral	0.79	0.79	0.78	45.15	34.63
more neutral, high perplexity	0.89	0.90	0.88	37.16	123.77
GYAFC (Rao and Tetreault, 2018)	0.81	0.82	0.81	45.74	99.21

Table 2: Comparison between the two prompting techniques and with the neutral-prompt versions of the high-perplexity comments. We compared our LLM-generated data to the GYAFC dataset (Rao and Tetreault, 2018) - whose formal versions were generated by human annotators - using the same evaluation metrics for better comparison.

Model	Version	Parameters
BART	bart-base (Lewis et al., 2020)	110M
T5	t5-base, flan-t5-base, (Raffel et al., 2020)	250M
GPT-3.5	text-davinci-003 (Brown et al., 2020)	175B

Table 3: The Language Models employed for style transfer.

generated versions in GYAFC, the picture is a bit differenw when we only look at the neutral versions of the high-perplexity subset, which in comparison yielded better BERTScore and chrF++ values as well as a higher perplexity (which is more in line with the perplexity of the human-generated versions in GYAFC). Overall, the machine-generated neutral versions seem comparably good with the human-generated versions in GYAFC. Examples are provided in Appendix A in Table 6.

4 Model description

4.1 Baseline model

As comparison we carried out style-transfer experiments using a very large language model (the text-davinci-003 version of GPT-3.5, with 175B parameters) without fine-tuning, using a zero-shot approach. We used the following prompt:

"Here are example sentences: {example1} {example2} {example3}
Here is a sentence: {neutral-style comment}
Here is a rewrite of this sentence according to the example sentences: {"

The model performs a style transfer by completing the prompt.

4.2 Fine-tuned models

We fine-tuned a BART models (bart-base) as well as two T5 models (t5-base and flan-t5-base). The models were fine-tuned using the training set, using the task of generating the style-transferred output by completing the prompts:

Input:

"Here are example sentences:
{example1} {example2} {example3}
Here is a sentence: {neutral-style comment}
Here is a rewrite of this sentence according to the example sentences: {"

Output:

"{original version of the neutral comment} }"

We used the style-neutral, LLM-generated versions in the input and the original Reddit versions of the comments in the output.

4.3 Model evaluation

We evaluate the models' performance on the evaluation set, which does not contain comments from the same Subreddits as the training set. In this way we evaluated how the models perform on unseen data and unseen styles.

Meaning equivalence We compute BERTScore (Zhang et al., 2019) and chrF++ (Popović, 2015, 2017) to assess the meaning equivalence between the neutral input and the style-transferred output.

BERTScore measures embedding similarity between tokens in the source text and in the target text. The the similarity are used to computes *recall* by matching each token x in the source to a token in the target $\hat{x}$, and *precision* by matching each token $\hat{x}$ in the target to a token x in the source, with greedy matching (Zhang et al., 2019). Precision and recall are used to compute the $F-$score.

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^\top \hat{x}_j \qquad (1)$$

Model		F1-Score	Precision	Recall	chrF++	Perplexity
BART	bart-base, zero-shot	0.48	0.48	0.49	9.75	650.59
	bart-base, 5 epochs	0.81	0.82	0.80	49.09	379.61
T5	t5-base, zero-shot	0.31	0.30	0.33	0.79	16140.79
	t5-base, 5 epochs	0.86	0.88	0.85	56.02	337.97
	flan-t5-base, zero-shot	0.93	0.95	0.92	88.68	829.89
	flan-t5-base, 5 epochs	0.82	0.83	0.82	50.01	547.81
GPT-3.5	text-davinci-003, zero-shot	0.85	0.83	0.87	59.43	151.72

Table 4: BERTScore, chrF++ and perplexity results for the baseline model and the fine-tuned models (after 5 epochs).

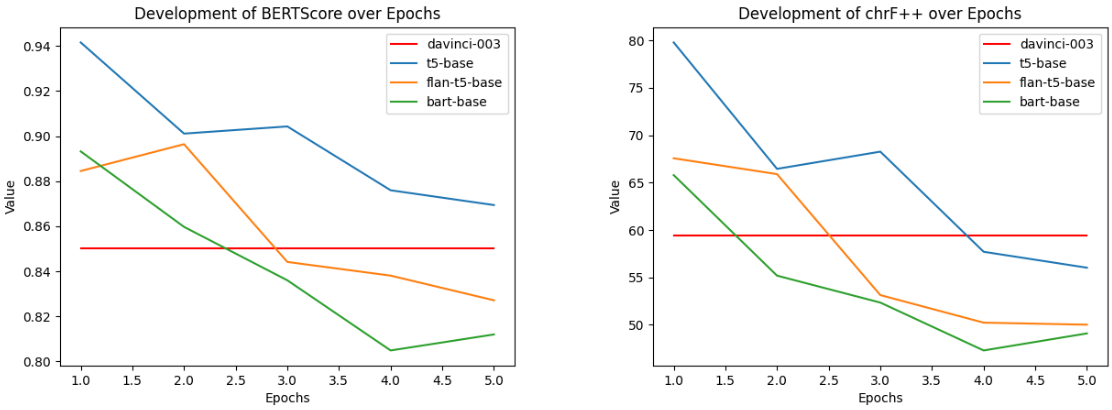

Figure 3: Changes in BERTScore and chrF++ over different epochs.

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^{\top} \hat{x}_j \tag{2}$$

$$F_{\text{ BERT}} = 2\frac{P_{\text{BERT}} * R_{\text{BERT}}}{R_{\text{BERT}} + R_{\text{BERT}}} \tag{3}$$

The chrF score (Popović, 2015, 2017) computes *precision* as the percentage of n-grams in the target which have a counterpart in the source, and *recall* as the percentage of n-grams in the source which have a counterpart in the target. For the chrF++ score , the word n-grams are added to the character n-grams and then averaged.

Style transfer In order to evaluate to what extent the style transfer was successful, we compute a general perplexity score using the non-fine-tuned GPT-2. This perplexity indicates how much the style-transferred output differs from the standard language use. We then compute perplexity values for Subreddit-specific fine-tuned language models as described in 3.5, to evaluate to what extent the obtained style for the target Subreddit differed from the style of the other Subreddits.

5 Results

The results of the model evaluation are summarized in Table 4. The fine-tuned models are compared with their own performance before fine-tuning (zero-shot) as well as with the larger baseline model, which is not fine-tuned either. We provide examples of the generated style-transferred comments in Appendix C.

5.1 Meaning equivalence

The results of this evaluation are summarized in Table 4 and Figure 3. The BERTscore and the chrF++ scores on the smaller non-finetuned models show that fine-tuning is indeed necessary for these models. The BERTscore and the chrF++ scores actually worsened with further fine-tuning on the training set, both as compared to the earlier epochs of the same models and to the baseline. This was probably a consequence of the style adaptation as well, as the models progressively differentiated themselves from the standard language use. However, after 5 epochs the fine-tuned models still yielded satisfactory measures of semantic similarity to the neutral input and considerably better scores compared to

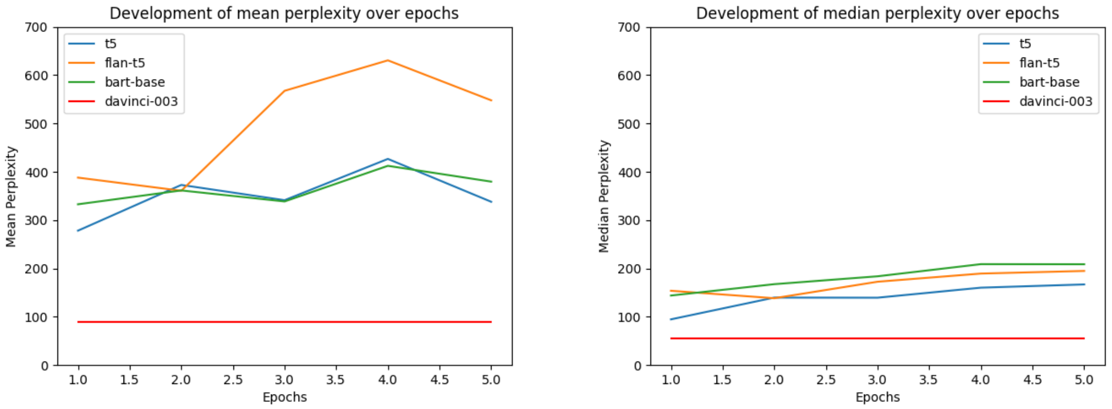

Figure 4: Changes in mean and median GPT-2 perplexity over different epochs.

Model		TrueReddit	TwoXChromosomes	wallstreetbets	worldnews
BART	bart-base	154.82	128.04	115.90	176.27
T5	t5-base	177.43	74.15	118.92	507.51
	flan-t5-base	74.50	107.30	105.29	487.38
GPT-3.5	text-davinci-003	68.14	56.43	92.18	89.69

Table 5: Within-style subbredit-specific perplexity for the four styles in the evaluation set, for the baseline and the fine-tuned models after 5 epochs.

their non-fine-tuned versions. As an exception, it is worth noting that flan-t5-base yielded better scores in the zero-shot version. This happened because the model tended to simply copy the source text.

5.2 General perplexity

The results of this evaluation are summarized in Table 4 and Figure 4. A high perplexity here shows a style differentiation from standard use. All fine-tuned models yield higher perplexity values compared to their zero-shot versions as the fine-tuning progresses – and higher than the baseline.

5.3 Subreddit-specific perplexity

Subreddit-specific perplexity scores were computed for style-transferred outputs, in order to evaluate the match between output style and target style.

Perplexity scores for target-style language models For the four Subreddits in the evaluation set, Table 5 shows the perplexity scores for the matching fine-tuned Subreddit-specific language model, obtained on the style-transferred outputs from the fine-tuned models and the baseline model. The outputs of all fine-tuned models yield particularly high perplexity when the target style is "worldnews" - but this is not the case for the outputs generated by the baseline model for the same target style. Note that the "worldnews" Subreddit did not seem to be a particularly dishomogeneous one during the dataset evaluation.

Comparison between target vs. other styles All style-transferred outputs of the fine-tuned style transfer models yielded the lowest perplexity scores for the Subreddit-specific language models of the corresponding target style compared to other Subreddit-specific language models. The only exception was the model flan-t5-base, whose outputs for the target style "worldnews" yielded the lowest perplexity scores for language model corresponding to the style "offbeat" instead. It is worth mentioning here again that the styles used in the evaluation were not the same styles using during fine-tuning of the style-transfer models. Figure 5 in Appendix C compares different style-specific perplexities for TrueReddit-style comments generated by the different models.

6 Discussion

The dataset evaluation showed that the different bubbles / Subreddits are sufficiently distinguishable from one another and that the quality of our machine-generated neutral-style translations is comparable to that achieved with similar, human-generated datasets.

We left 4 Subreddits aside for the evaluation, only using 16 for fine tuning, in order to evaluate if the fine-tuning improved the style transfer task itself and not a transfer to a particular style.

The style transfer capability of the fine-tuned models was explored using measures of semantic similarity / meaning equivalence between texts such as BERTscore and chrF++ as well as perplexity as a measure of style similarity. Our results show that scores such as BERTscore and chrF++, are improved after fine-tuning compared to the zero-shot scenario, but then decrease as we fine-tune for style transfer. It probably comes with the task of style transfer that, as the model learn to specialize for a specific social media bubble, the semantic similarity decreases. While we argue that BERTscore and chrF++ are more suitable than token-based (n-gram based) measures such as BLEU to assess meaning equivalence in style transfer and paraphrasing tasks, we also observe that the differences between the Subreddits do not only pertain to the style but also to the semantic content, which is probably also causing the semantic similarity scores to decrease with fine tuning. Similarly, topic differences between the Subreddits may also influence the perplexity scores, as a language model will be more "surprised" when encountering text with a very specific style and topic content which differs from those of the average texts it was trained and fine-tuned on.

7 Conclusion

For many downstream tasks it is tempting to use a LLM and to go for a zero-shot approach, in particular for a task such as style transfer, where style itself is a concept which is difficult to pinpoint, let alone finding specific style categories to be applied. Working with examples as prompts has the advantage of sidestepping the issue of defining what a particular style should look like.

However, we show that some fine-tuning of smaller models such as BART and T5 models is also a viable option. These models, when fine-tuned with a small amount of data to learn the style transfer from one social medial bubble to another, despite being much smaller than GPT-3.5, can achieve comparable or better results in performing new, arbitrary style transfers in the Subreddit domain.

For the fine-tuning itself we provide a dataset of different Subreddits under the assumption that to each Subreddit / social media bubble corresponds a characteristic, identifiable style. Just because a comment comes from one specific Subreddit however does not imply that the comment itself will have an identifiable style, some may be less marked. Thus we use perplexity as computed by a LLM (GPT-2) as a proxy to evaluate how stylistically charged a comment is and select 150 high-perplexity comments for each of 20 Subreddits. For the selected comments, we create a neutral-style version for each comment using a LLM (GPT-2). The neutral-style versions are used to create prompts which help the models learn the task of style transfer during fine-tuning. Note that steps requiring the use of a LLM are only involved in the database creation - once the database is created, it is enough to fine-tune smaller models for the task.

Four Subreddits were kept aside for evaluation purposes. Note that the fine-tuning is performed on different styles than the ones used in the evaluation. The semantic overlap between neutral versions and target-style versions was evaluated using BERTscore and chrF++, while the style match was evaluated using perplexity scores of language models. GPT-2 was used as a generic LLM to measure the match with a nonmarked use of language. Then it was fine-tuned to obtain style-specific language models to evaluate the match between the generated outputs and the different styles. The evaluation showed satisfactory results for the smaller, fine-tuned models (BART and T5) when compared to the outputs generated by a LLM (GPT-3.5).

Of course GPT-3.5, a much larger model, can already achieve very good results with a zero-shot approach, without fine-tuning - but we argue that it makes more sense to employ the relatively small resources required to fine-tune a smaller model for the style transfer task rather than following a zero-shot approach.

Limitations

Our goal is to teach the models a general task of style transfer, which is why we use different styles in the training and testing phases. However, we acknowledge that the style of Reddit posts, however different between different Subreddits, may still be rather homogeneous.

This work is limited to English and to social media language - in particular, we looked at comments of a maximum length of 512 tokens. We make a few assumptions during our work, probably the biggest assumption is that the language use learned by a LLM such as GPT-2 reflects the non-marked, standard use of the English language. We also assumed that, if a language model learns the style of a (collection of) texts, then the perplexity of that language model can be used as a proxy for the style match between a text and a target style.

We also assume that perplexity on the one hand and BERTscore and chrF++ on the other hand are optimal measures for style match and semantic content match respectively. However, what characterizes a particular style is not just the vocabulary use or the type of grammar but also the topics discussed, in particular when it comes to social bubbles such as the ones described in this work. The difference between topics may of course also influence perplexity values.

Ethics Statement

The scraped Reddit comments have not been filtered for explicit content or assessed for bias and may contain offensive or triggering languages that could upset the reader.

Sustainability The training and use of large language models requires a high amount of energy and CO2 emissions. We employed a large language model to generate our neutral-style sentences as well as for our baseline. In our experiments we showed that fine-tuning a smaller model may thus be preferrable to using a larger language model.

Harmful Language Generation Language models can be used for harmful language generation. For example, technology which may make text coherent and recognizable by a social media bubble, for example a group of conspiracy theorist, may favor the spreading of a large number of machine-generated contributions in those social media bubbles, with the risk of amplifying bias and misinformation. The presented technology can also be used to impersonate a certain author or group of authors.

References

Katherine Atwell, Sabit Hassan, and Malihe Alikhani. 2022. APPDIA: A discourse-aware transformer-based style transfer model for offensive social media conversations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6063–6074, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yu Bao, Hao Zhou, Shujian Huang, Lei Li, Lili Mou, Olga Vechtomova, Xin-yu Dai, and Jiajun Chen. 2019. Generating sentences from disentangled syntactic and semantic spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6008–6019, Florence, Italy. 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*, volume 14, pages 830–839.

Nancy K Baym. 2003. Communication in online communities. In *Encyclopedia of Community*, volume 3, pages 1015–1017. Sage Thousand Oaks, CA.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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.

Wayne Duggan. 2023. Was bedeutet HODL? — forbes.com. `https://www.forbes.com/advisor/de/geldanlage/krypto/was-ist-hodl/`. [Accessed 20-Jun-2023].

Aleksandra Gnach. 2017. Social media and community building: Creating social realities through linguistic interaction. In *The Routledge handbook of language and media*, pages 190–205. Routledge.

Navita Goyal, Balaji Vasan Srinivasan, Anandhavelu N, and Abhilasha Sancheti. 2021. Multi-style transfer with discriminative feedback on disjoint corpus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3500–3510, Online. Association for Computational Linguistics.

Eduard Hovy. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics*, 11(6):689–719.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Florence, Italy. Association for Computational Linguistics.

Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. 2023. Exploiting asymmetry for synthetic training data generation: Synthie and the case of information extraction. *arXiv preprint arXiv:2303.04132*.

Fabian Kopf. 2023. Reddit Comments Dataset for Text Style Transfer Tasks.

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.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.

Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You impress me: Dialogue generation via mutual persona perception. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1417–1427, Online. Association for Computational Linguistics.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. ParaDetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. A dual reinforcement learning framework for unsupervised text style transfer. *arXiv preprint arXiv:1905.10060*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. 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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

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.

Machel Reid and Victor Zhong. 2021. LEWIS: Levenshtein editing for unsupervised text style transfer. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3932–3944, Online. Association for Computational Linguistics.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 837–848, Dublin, Ireland. Association for Computational Linguistics.

Howard Rheingold. 2000. *The virtual community, revised edition: Homesteading on the electronic frontier*. MIT press.

Parker Riley, Noah Constant, Mandy Guo, Girish Kumar, David Uthus, and Zarana Parekh. 2021. TextSETTR: Few-shot text style extraction and tunable targeted restyling. 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 3786–3800, Online. Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems*, 30.

David E Smith and Clark S Sturges. 1969. The semantics of the san francisco drug scene. *ETC: A Review of General Semantics*, pages 168–175.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. *arXiv preprint arXiv:2205.11503*.

Yu Wu, Yunli Wang, and Shujie Liu. 2020. A dataset for low-resource stylized sequence-to-sequence generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9290–9297.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

A More details about the data collection

original comment	neutral version
It was World Pride in Sydney this weekend so HEAPS of gay dudery all round!	This past weekend, Sydney hosted an event celebrating pride, which saw many members of the LGBTQ+ community come together.
so you support OPs laziness and failure to confirm information	Are you in agreement with the idea of not verifying information and taking on a more relaxed attitude?
Literally weaponized the 2nd Amendment. Lol	The Second Amendment has been used to support various arguments.
Just saying, no brag or anything, but I make $35$hr off construction knowledge. I started low but got good at it.	I have experience in construction and I make $35$hr. I started out with a lower rate, but I have become more skilled over time.

Table 6: Examples for machine-generated neutral versions, generated with the "*more neutral*" prompt.

B Computational Details

We used a GPU Cluster with the following specifications:

- CPUs: 2x Intel® Xeon® Gold Prozessor 5315Y
- RAM: 512 GB
- GPUs: 2x Nvidia RTX A6000

C Results

neutral comment input	You don’t know their financial situation, so it’s best to move on.
style example 1	Or you could actually know what tf is going on first
style example 2	Please get the police involved I beg you
style example 3	I understand now
output bart-base	you don’t know his ex best to move on
output t5-base	You don’t know their fucking situation so move on
output flan-T5-base	You don’t know her financial situation so move on
output text-davinci-003	It’s wise to move on since you don’t know their financial situation
neutral comment input	The best way to trade this market is to consider buying calls on dips.
style example 1	it will be up just wait for the liquidity of trapped traders in the fake bull
style example 2	Bers so desperate to break 400
style example 3	HOT DAMN SHE BALD
output bart-base	Best way to trade this market is to buy calls on dips
output t5-base	The best way to trade this market is to buy calls on dips
output flan-T5-base	Best way to trade this market is to buy calls on dips
output text-davinci-003	What’s the best strategy for trading this market? Think about buying calls when the price dips.
neutral comment input	This post has received a significant amount of downvotes from Cyberi bots.
style example 1	We literally tried for 20 years to get the women in schools. If only the Afghan government hadn’t folded like a lawn chair
style example 2	And yet our gas prices are still way high! Dang our administration sucks.
style example 3	are there little ones for their * to?
output bart-base	This post has received 100+ downvotes from Cyberi bots
output t5-base	This post got a ton of downvotes from Cyberi bots.
output flan-T5-base	This post got a lot of downvotes from Cyberi bots.
output text-davinci-003	This post has been met with a considerable amount of disapproval from Cyberi robots.

Table 7: Examples for target sentences, generated with the baseline and the fine-tuned models after 5 training epochs.

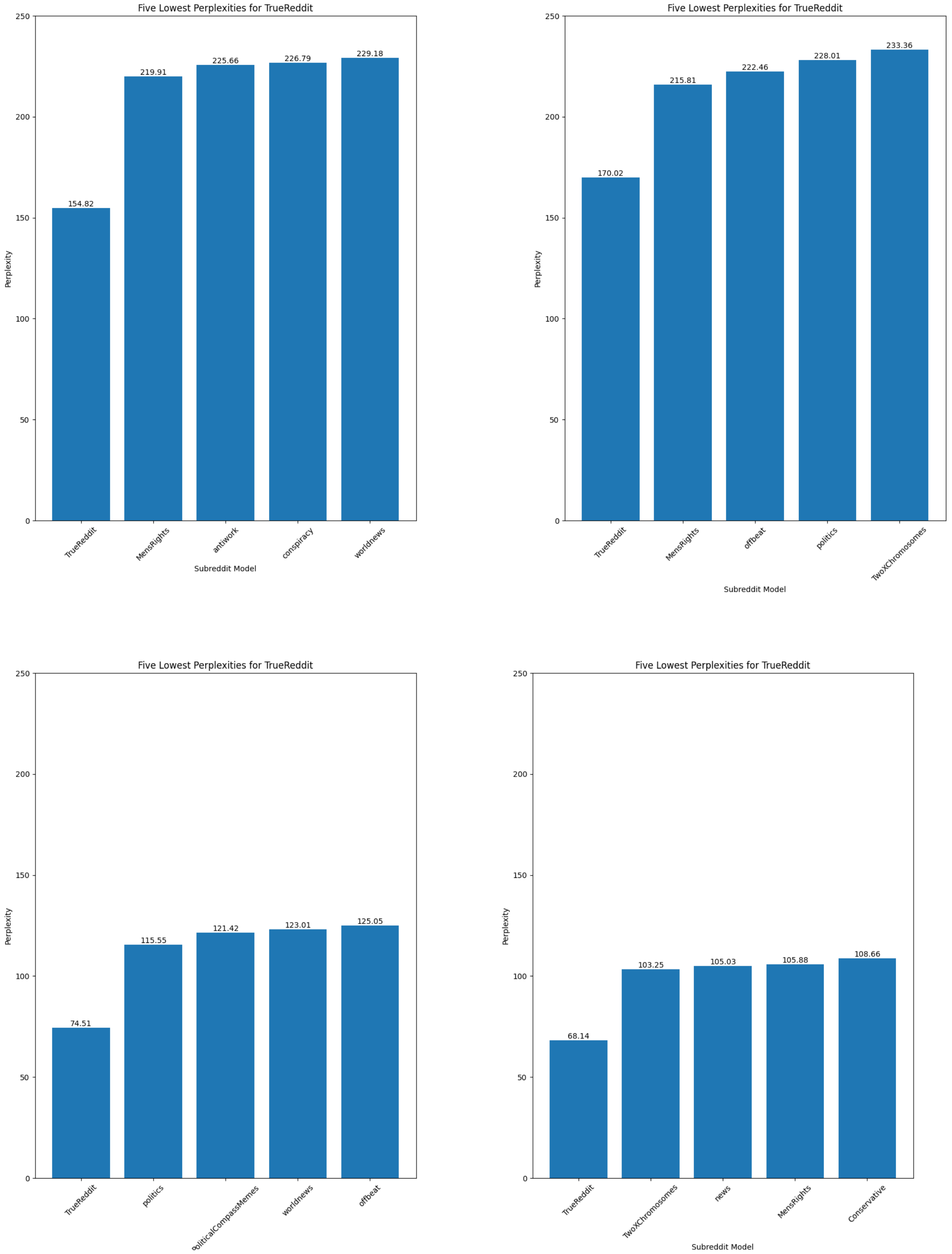

Figure 5: Comparison between different style-specific perplexities for TrueReddit-style comments generated by the different models.