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

Alessandra Zarcone and Fabian Kopf
Technische Hochschule Augsburg, Augsburg, Germany

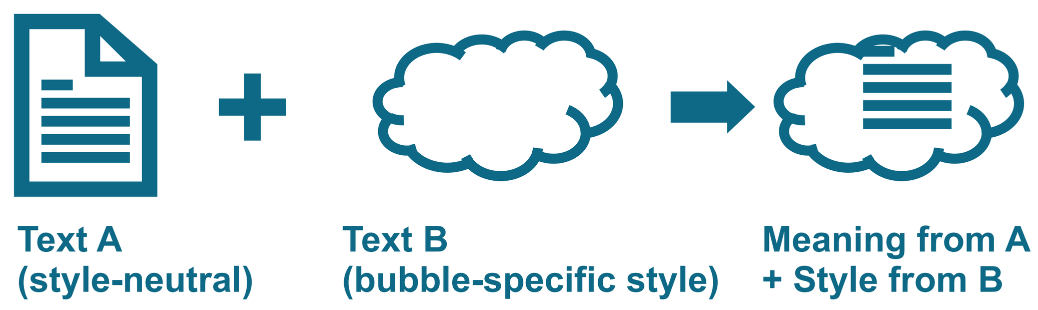
Technische Hochschule Augsburg

HIGHTECH Agenda Bayern

Style Transfers to Bubble-specific Language

Language on social media

- a way to create a **community identity** in social media bubbles (e.g *HODL* for "holding a share")
- can be difficult for outsiders to understand or to mimic
- style transfer between social media bubbles as a first step to analyze / detect style in social media



1) How do we translate text into the language of one bubble **without losing the original meaning**?
2) Can we successfully perform style transfer **in a resource-efficient way**, that is without resorting to very large Language Models (LMs) or large amounts of training data?

The Dataset

- Source: **20 Subreddits** (topic-specific forums) selected for variety of topics + wide between-Subreddit style variance + homogenous style within each Subreddit
- > 49K comments, 10 to 512 token long

Subset used for training (16 Subreddits) and evaluation (4)

- 150 most stylistically-marked comments (top GPT-2 perplexity)
- style-neutral version created synthetically with a large LM (GPT-3.5) adopting the zero-shot approach in Reif et al. (2022)
- high-perplexity comments were similar to their neutral versions, but different in style (comparison with GYAFC dataset)

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

Just saying, no brag or anything, but I make $35/hr off construction knowledge. I started low but got good at it

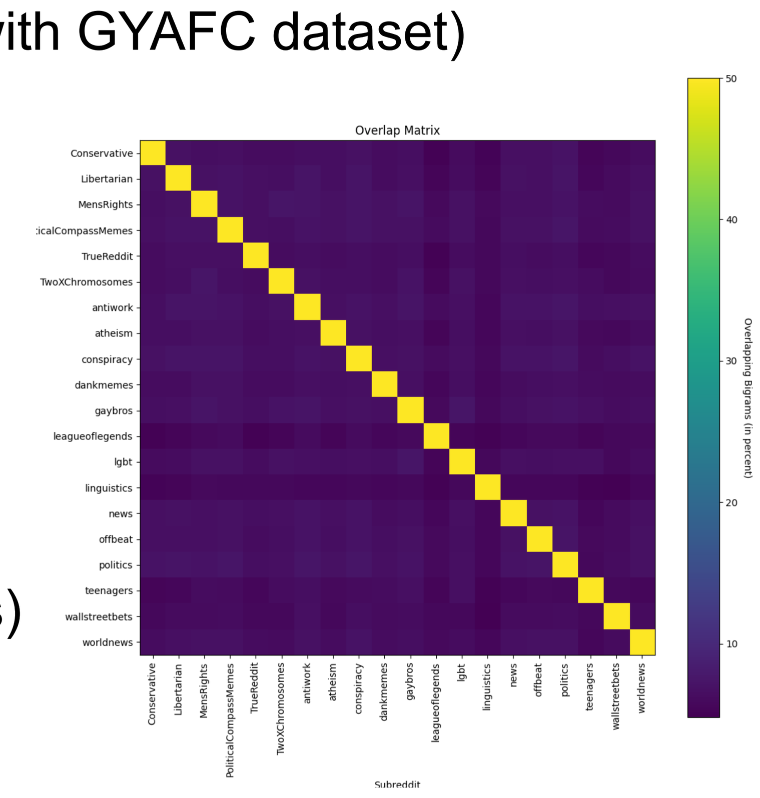
➡

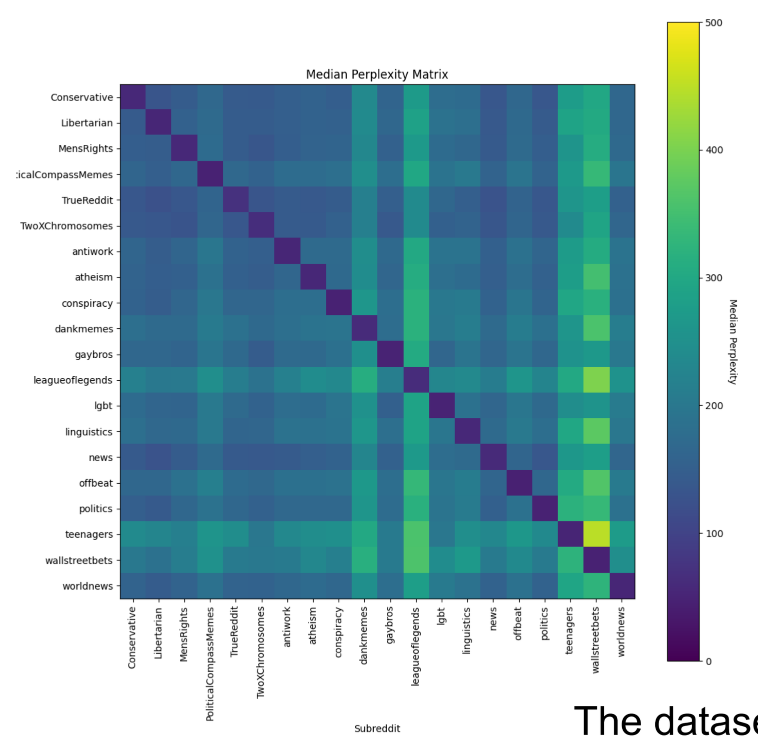
I have experience in construction and I make $35/hr. I started out with a lower rate, but I have become more skilled over time.

	BERTScore F1	Perplexity
Our data (machine-generated)	0.89	123.77
GYAFC (formal / informal, human-generated, Rao and Tetreault, 2018)	0.81	99.21

Lexical Overlap

high-perplexity comments in each Subreddit easily distinguishable from those in the other Subreddits (all possible pairs, % shared bigrams)





Perplexity of Subreddit-specific LMs

bubble-specific LMs were "surprised" when exposed to the style of a different bubble (fine-tuned Subreddit-specific GPT-2 models)

The dataset on Zenodo: https://doi.org/10.5281/zenodo.8023142

The Models

- **Baseline:** GPT-3.5 (text-davinci-003), zero-shot

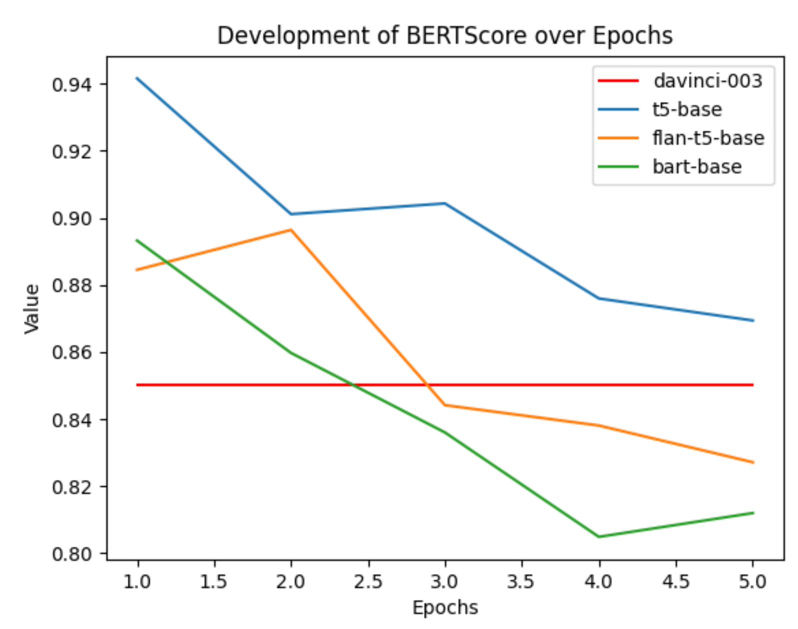
"Here are example sentences: {example1} {example2} {example3}
Here is a sentence: {neutral-style comment}
Here is a rewrite of this sentence according to the example sentences: {"

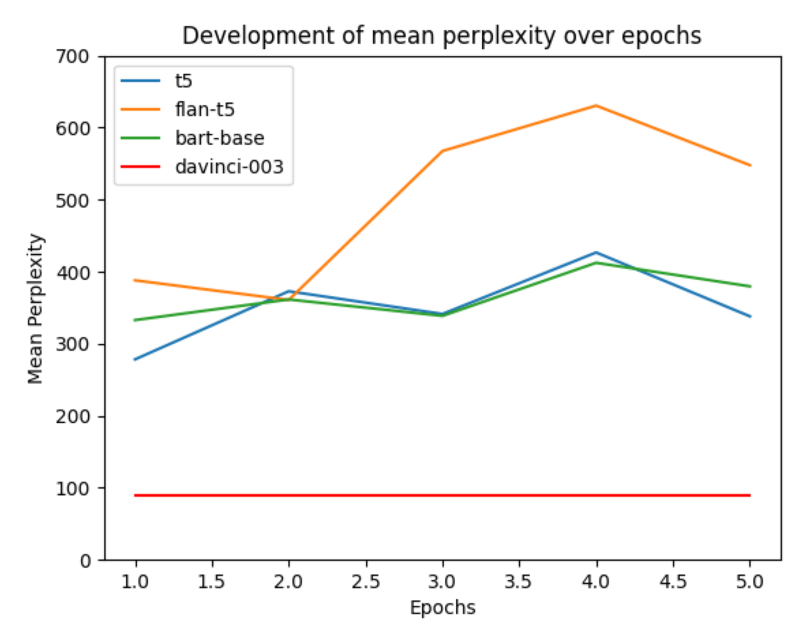
- **Fine-tuned models:** bart-base, t5-base, flan-t5-base
  - Fine-tuned using the training set (16 SubReddits)
  - Task: completing the prompt with the original version of the {neutral-style comment}

Evaluation

- **Evaluation set:** 4 SubReddits not included in fine-tuning
- **Meaning Equivalence**
  BERTScore (Zhang et al., 2019) between source and target
- **Style Transfer**
  - **GPT-2 Perplexity** as a proxy of deviation from standard use
  - **Subreddit-specific-GPT-2 Perplexity** as a proxy of deviation from other Subreddits

Results





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

Subreddit-specific perplexity scores for matching style-transferred outputs

- **Fine-tuning necessary** for smaller LMs
- Fine-tuned models yielded satisfactory results **compared to the larger baseline LM**
- The outputs of all fine-tuned models yielded the **lowest perplexity scores** for the **corresponding style-specific LM** (exception: flan-t5-base for *worldnews* and the *offbeat*-LM)

Contributions and Outlook

Dataset

- 150 high-perplexity comments for each of 20 Subreddits, each with a machine-generated neutral-style version
- 16 Subreddits used for finetuning, 4 for evaluation

Dataset evaluation

- different bubbles sufficiently distinguishable from one another
- quality of the machine-generated neutral version comparable to quality of similar, human generated datasets

Learning to style transfer, not a specific style

- Successful style transfer without using large LMs with a zero-short approach but finetuning with a small amount of data
- BERTscore improved after fine-tuning but decreased as we finetuned (as the models learn to "style transfer")

Bubble style and semantic content are difficult to disentangle

- This can affect semantic similarity and perplexity scores