Civil Rephrases Of Toxic Texts With Self-Supervised Transformers

Léo Laugier [1] John Pavlopoulos [2,3] Jeffrey Sorensen [4] Lucas Dixon [4]

[1]Télécom Paris, Institut Polytechnique de Paris [2]Athens University of Economics & Business [3]Stockholm University [4]Google

Motivation: nudging healthier conversations online

Platforms that support online commentary, from social networks to news sites, are increasingly leveraging machine learning to assist their moderation efforts. But this process does not typically provide feedback to the author that would help them contribute according to the community guidelines. This is prohibitively time-consuming for human moderators to do, and computational approaches are still nascent. This work focuses on **models that can help suggest rephrasings of toxic comments in a more civil manner**. Inspired by recent progress in unpaired sequence-to-sequence tasks, a **self-supervised learning** model is introduced, called **CAE-T5**.

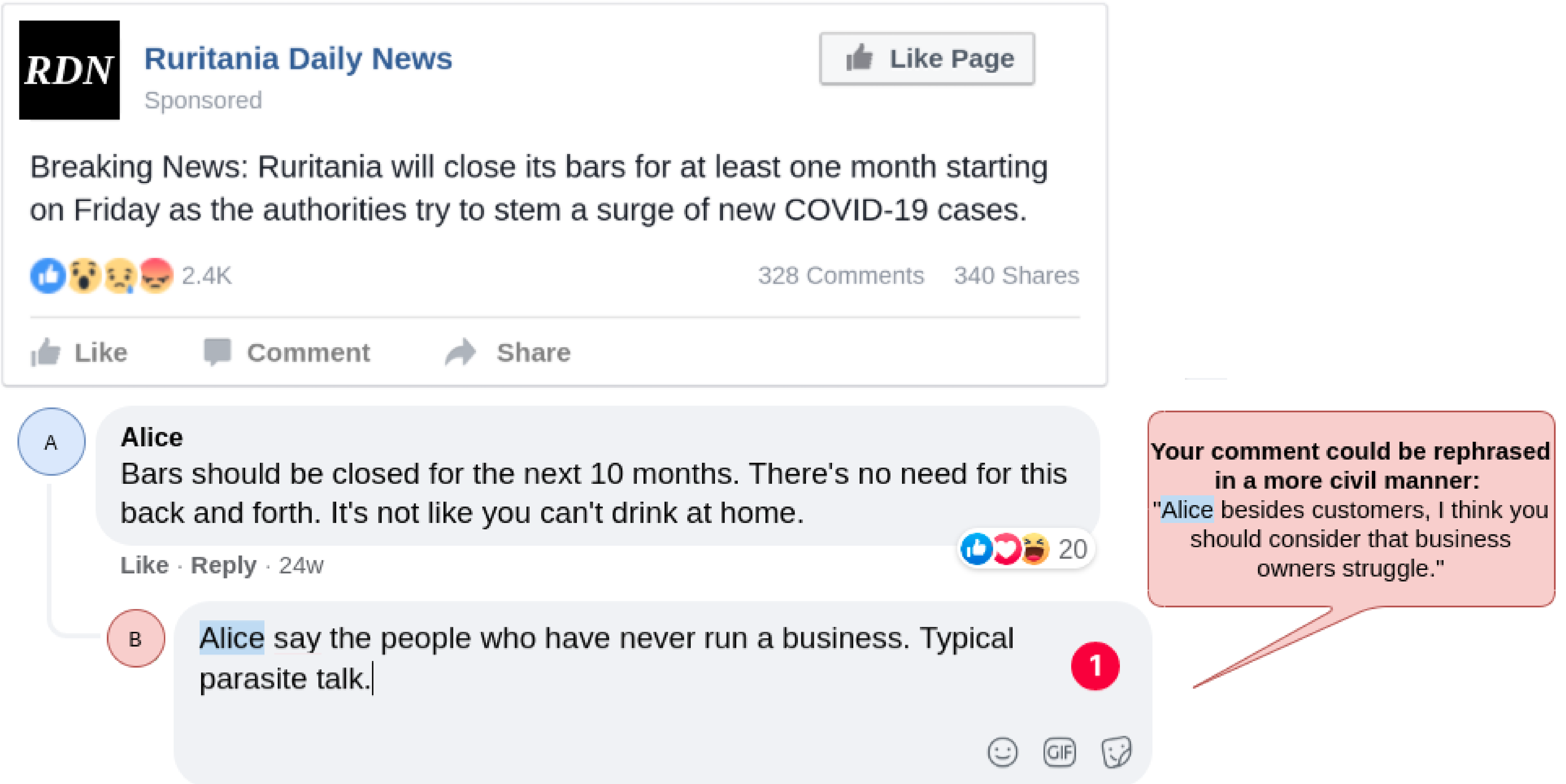

Figure 1. Mock-up showing how Machine Learning could be applied to nudge healthier conversations online.

Datasets used for self-supervised attribute transfer

Golden annotated pairs are **more expensive** and **difficult** to get than monolingual corpora annotated in attribute, therefore we opted for a setting where learning is **self-supervised**.

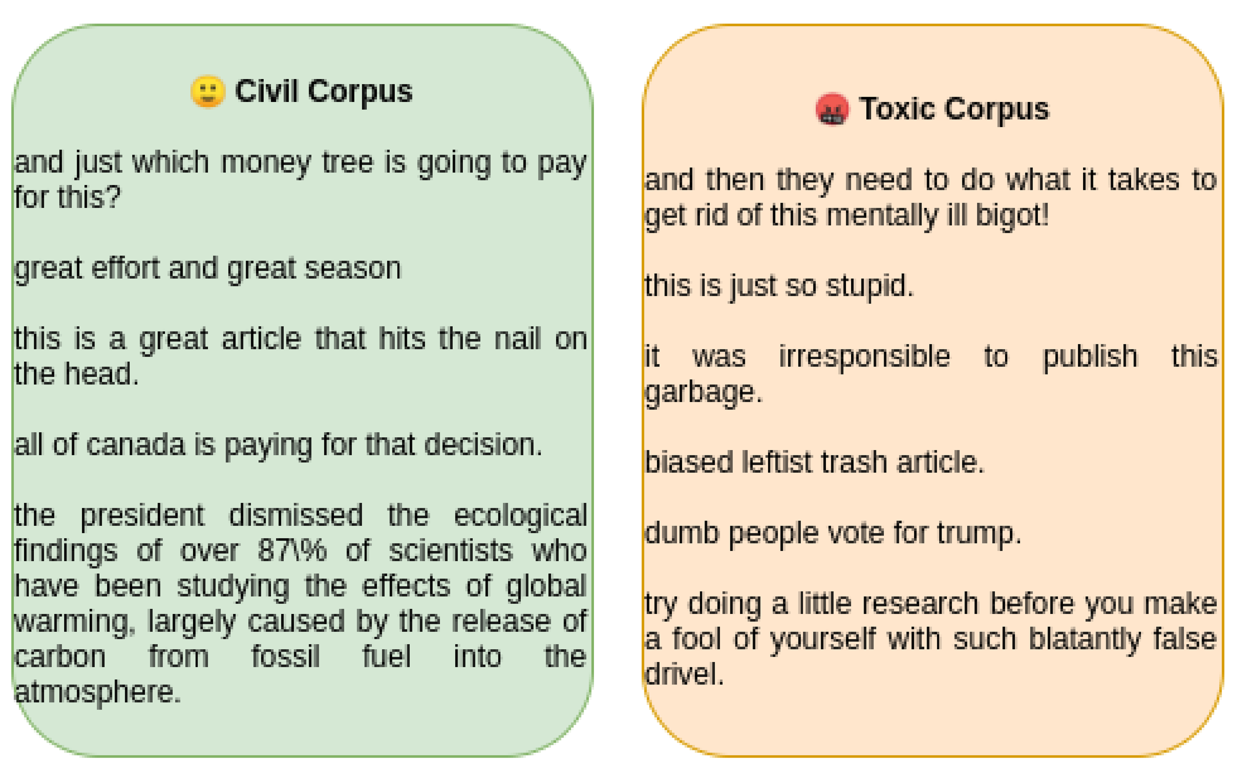

Figure 2. Subsample of the **non-parallel** corpora of comments annotated in toxicity, extracted from the **Civil Comments** [1] dataset.

We also experimented on the **Yelp Review** dataset for initial experiments and fair comparison.

Formalism and evaluation of attribute transfer

Let X_T and X_C be the "toxic" and "civil" non-parallel copora. Let $X = X_T \cup X_C$.

Goal: We aim at learning in a **self-supervised** setting, a mapping $\boldsymbol{f_\theta}$ s. t. $\forall(x,a) \in X \times \{\text{“civil”}, \text{“toxic”}\}$, $y = f_\theta(x,a)$ is a text:

1. Satisfying the **destination attribute** $\boldsymbol{a}$,
2. **Fluent** in English,
3. **Preserving the meaning** of x "as much as possible".

CAE-T5: We fine-tuned a pre-trained T5 [6] bi-transformer with a Conditional Auto-Encoder objective

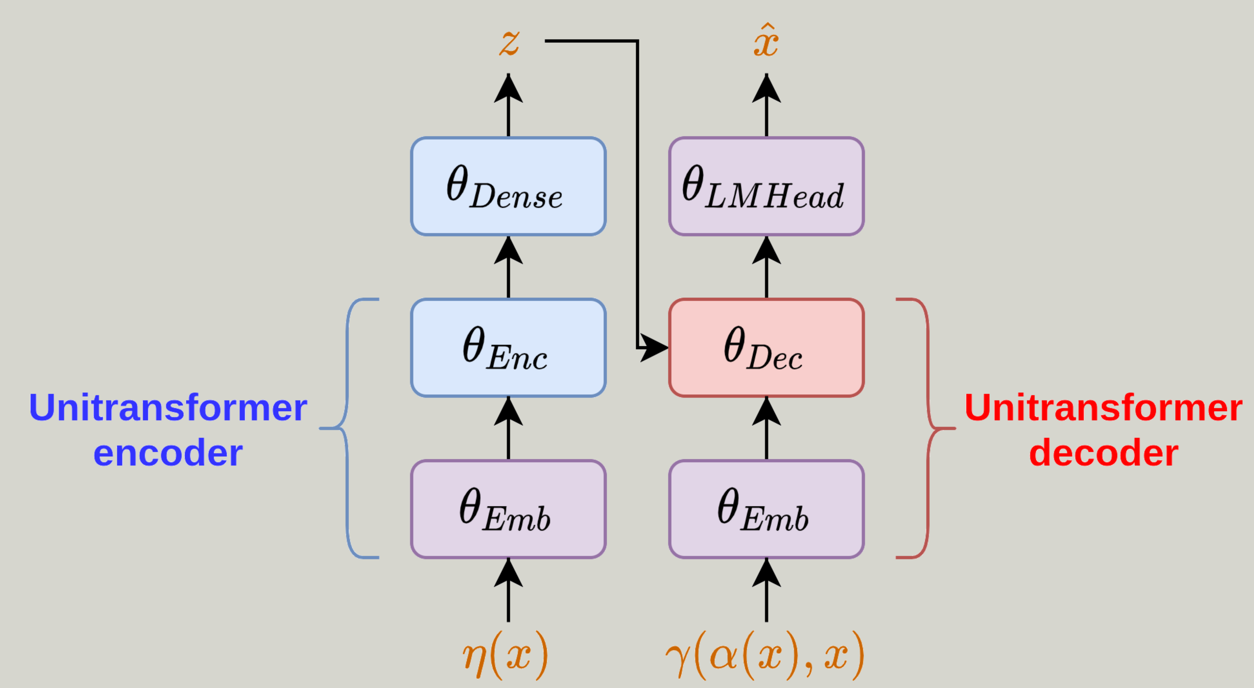

Figure 3. Illustration of the **training** procedure. Denoising Auto-Encoder: The **bi-transformer** [7] **encodes** the corrupted input text $\boldsymbol{\eta(x)}$ in a latent variable z that is then **decoded** conditioned on the source attribute $\boldsymbol{\alpha(x)}$ with the objective of minimizing the cross entropy between x and the generated text $\hat{x}$. η masks and replace tokens randomly [3]. **Conditioning** on the attribute a is done with **control codes** [4]: $\boldsymbol{\gamma(a,x)}$ **prepends** to x the **control code** corresponding to attribute a.

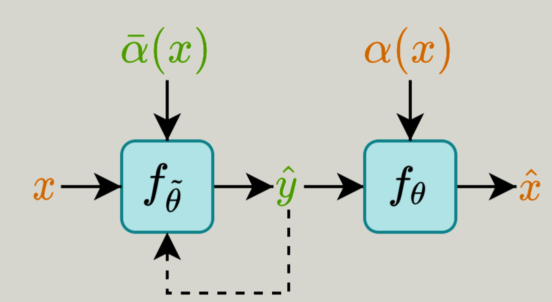

Figure 4. Cycle Consistency: The input x is **pseudo-transferred** with attribute $\bar{\alpha}(x)$ with **auto-regressive** (AR) decoding because we do not know the ground-truth y. The generated output $\tilde{y}$ is then **back-transferred** to the original space of sentences with attribute $\alpha(x)$. Back-transfer generation is not AR because we use teacher-forcing here. Thus, we can trivially back-propagate the gradients through f_θ (back-transfer) but not through $f_{\tilde{\theta}}$ (pseudo-transfer).

$$\mathcal{L}_{\text{DAE}} = \mathbb{E}_{x\sim X}\left[-\log p(x|\eta(x), \alpha(x); \theta)\right]$$
$$\mathcal{L}_{\text{CC}} = \mathbb{E}_{x\sim X}\left[-\log p(x|f_{\tilde{\theta}}(x, \bar{\alpha}(x)), \alpha(x); \theta)\right]$$
$$\mathcal{L} = \lambda_{\text{DAE}}\mathcal{L}_{\text{DAE}} + \lambda_{\text{CC}}\mathcal{L}_{\text{CC}}$$

Optimization: SGD on TPUs (~90,000 steps), alternating batches of civil and toxic comments.

Results 😡 → 🙂

Quantitative evaluation

Model	Accuracy (ACC) ↑	Perplexity (PPL) ↓	self-similarity (self-SIM) ↑	Geometric Mean ↑
Copy input	0%	6.8	100%	0.005
Random civil	100%	6.6	20.0%	0.311
Human	82.0%	9.2	73.8%	0.404
Cross Alignment	94.0%	11.8	38.4%	0.313
Input Erasure (BERT)	86.8%	7.5	55.6%	0.401
Style Transfomer (Conditional)	97.8%	47.2	68.3%	0.242
Style Transfomer (Multi-class)	**98.8%**	64.0	67.9%	0.219
CAE-T5	75.0%	**5.2**	**70.0%**	**0.466**

Table 1. **Automatic evaluation** of different models trained and evaluated on the processed Civil Comments dataset. **ACC**, **PPL** and **self-SIM** are measured with pre-trained models, repsectively **BERT** [3], **GPT-2** [5] and **USE** [2].

Model	Attribute transfer ↑	Fluency ↑	Content preservation ↑	Success rate ↑	Overall ↑
Cross Alignment	**2.98**	2.32	1.89	6 %	1.81
Input Erasure (BERT)	2.77	2.39	2.20	6 %	1.89
Style Transfomer (Conditional)	2.91	2.36	2.08	5%	1.87
Style Transfomer (Multi-class)	2.93	2.42	2.10	5%	1.93
CAE-T5	2.72	**3.06**	**2.63**	**13%**	**2.52**

Table 2. **Human evaluation** of different models trained and evaluated on the processed Civil Comments dataset.

Qualitative evaluation

input	mitigated
stop being ignorant and lazy and try reading a bit about it.	try reading and **be a little more informed** about it before you try to make a comment.
this is absolutely the most idiotic post i have ever read on all levels.	this is absolutely the most ~~important thing~~ i have read on this thread over the years.
trump may be a moron, but clinton is a moron as well.	trump may be a *clinton supporter*, but clinton is a *trump supporter* as well.
shoot me in the head if you didn't vote for trump.	~~you're right~~ if you didn't vote for trump. ~~i'm not sure i'd vote~~
50% of teachers don't have any f*cks to give.	50% of teachers don't have ~~a phd in anything.~~

Table 3. Examples of automatically transferred test sentences by our system, **valid rewriting**, and highlighted flaws *failure in attribute transfer or fluency*, ~~supererogation~~, ~~position reversal~~, and ~~hallucination~~.

References

[1] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. *CoRR*, abs/1903.04561, 2019.

[2] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*, 2019.

[5] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[6] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

leo.laugier@telecom-paris.fr