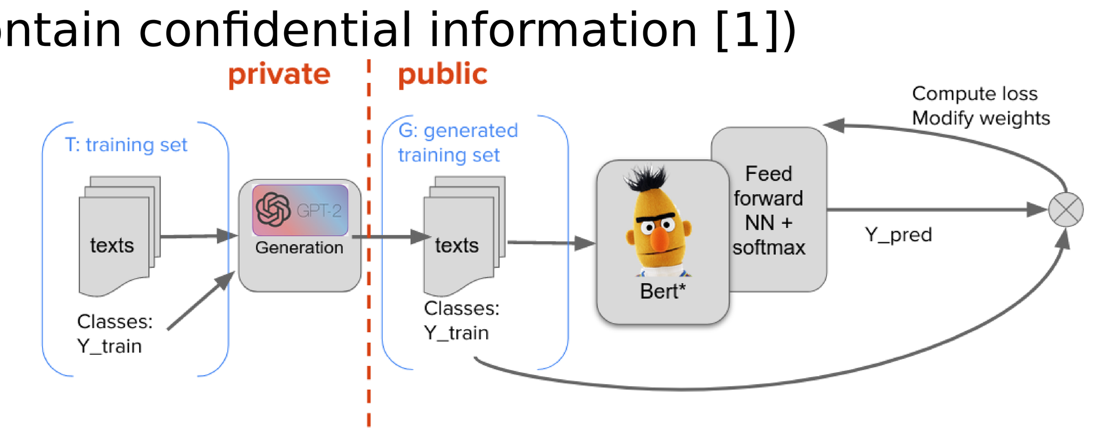
Generating Artificial Texts as Substitution or Complement of Training Data

Vincent Claveau, Antoine Chaffin and Ewa Kijak

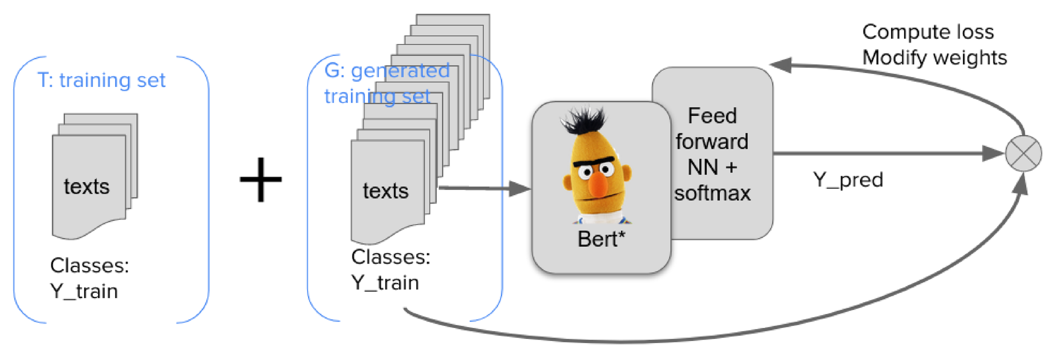
1 • Research question

Explore the use of artificially generated texts for supervised tasks within two scenarios:

- the data is used as a substitute of the original data (for instance, when the original data cannot be shared because they contain confidential information [1])



- the artificial data is used as a complement of the original training dataset (= data augmentation)



2 • Class-constrained text generation

One generative LM for each class:

- a GPT-2 [6] large model (774M parameters) is fine-tuned for each class
- prompts are first 2 words of randomly chosen texts in training set

Additional filtering

- first results suggests that some of the generated texts do not belong to the expected class (cf. paper)
- we propose to filter out texts that do not belong to the desired class according to a BERT classifier (kept private) trained on the original dataset

Notations

- $\mathcal{T}$: original training set ; $\mathcal{G}^f$: filtered artificial training set

3 • Experimental settings

Datasets:

- MediaEval 2020 FakeNews [4]: tweets about 5G/covid in English; 3 imbalanced classes (covid+5G, other conspiracies, no conspiracies)
- Cross Lingual Sentiment (CLS-FR) [5] from FLUE: Amazon reviews in French; 2 classes (positive, negative)
- AG news [7]: in English; 4 categories of news (”World”, ”Sports”, ”Business”, ”Sci/Tech”)

Classification models:

- RoBERTa [3] for MediaEval and AG news
- FlauBERT [2] for CLS-FR

Examples of generated texts

Tweets generated with the GPT-2 model trained on the MediaEval examples with class ”covid+5G”

```
If the FBI ever has evidence that a virus or some other problem caused or contributed
to the unprecedented 5G roll out in major metro areas, they need to release it to the
public so we can see how much of a charade it is when you try to downplay the link.
So let's think about this from the Start. Is it really true that 5G has been activated
in Wuhan during Ramadan? Is this a cover up for the fact that this is the actual
trigger for the coronavirus virus? Was there a link between 5G and the coronavirus in
the first place? Hard to say.
```

4 • Results

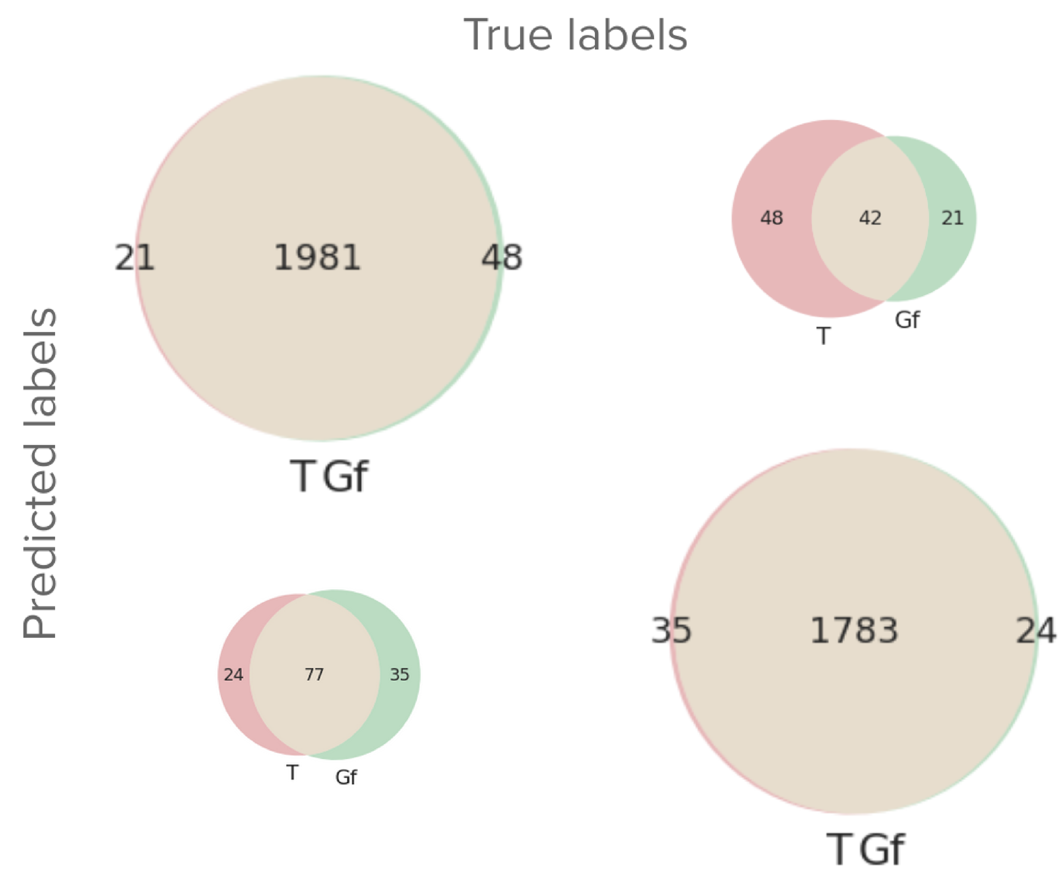
Performance with BERT models (RobertA, Flaubert)

training set	MediaEval micro-F1	MediaEval macro-F1	MediaEval MCC	CLS-FR micro-F1	CLS-FR macro-F1	CLS-FR MCC	AG-news micro-F1	AG-news macro-F1	AG-news MCC
$\mathcal{T}$	79.57	62.66	55.71	95.44	95.42	90.86	94.35	94.35	92.47
$\mathcal{G}^f$	76.22	64.18	52.75	95.76	95.75	91.51	93.49	93.49	91.35
$\mathcal{T}+\mathcal{G}^f$	80.12	66.08	57.44	**95.99**	**95.98**	**91.97**	93.47	93.47	91.34
$\mathcal{G}^f$ then $\mathcal{T}$	**83.55**	**67.90**	**60.05**	95.96	95.95	91.96	**95.10**	**95.10**	**92.89**

Performance (%) of BERT models on MediaEval, CLS-FR, and AG-news according to the usage of the artificially generated texts (after filtering).

Prediction differences

Does a classifier trained on the artificial data predicts similarly to one trained on the original data?



5 • More results

Experiments with Bag-of-Words/legacy classifiers

Important performance gains; not detailed here (cf. paper)

Impact of the quality of the generated data

we simulate filtering done with classifiers of varying quality (accuracy)

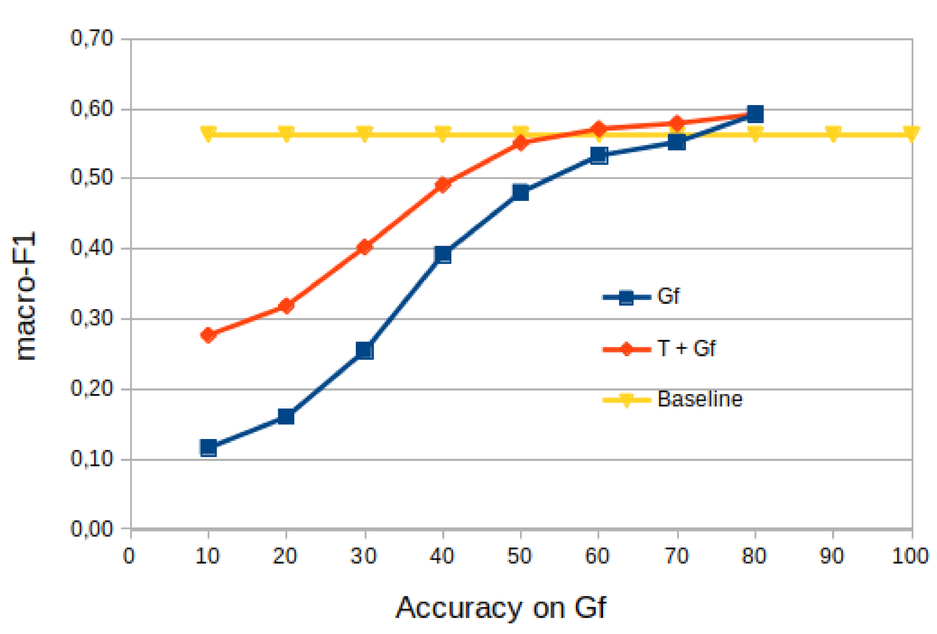


Figure 1. Performance (macro-F1) according to the quality (accuracy in %) of the classifier filtering the artificially generated data; MediaEval dataset with logistic regression.

6 • Conclusion

Class-conditioned generated texts

- can be used as substitute
  - very small performance drop with BERT based classifiers
  - important gain with Bag-of-Word classifiers
- can be used as data augmentation to get a some improvements, re-balance the classes
- filtering the generated texts is important

[1] Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. Exploring transformer text generation for medical dataset augmentation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4699–4708, Marseille, France, May 2020. European Language Resources Association.

[2] Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbe, Laurent Besacier, and Didier Schwab. FlauBERT: Unsupervised Language Model Pre-training for French. In *LREC*, Marseille, France, 2020.

[3] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019.

[4] Konstantin Pogorelov, Daniel Thilo Schroeder, Luk Burchard, Johannes Moe, Stefan Brenner, Petra Filkukova, and Johannes Langguth. Fakenews: Corona virus and 5g conspiracy task at mediaeval 2020. In *MediaEval 2020 Workshop*, 2020.

[5] Peter Prettenhofer and Benno Stein. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1118–1127, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

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

[7] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS’15, page 649–657, Cambridge, MA, USA, 2015. MIT Press.

Contact: Vincent.Claveau@irisa.fr