Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4260–4269
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

Generating Artificial Texts as Substitution or Complement of Training Data

Vincent Claveau[*], **Antoine Chaffin**[*,†], **Ewa Kijak**[*]
[*] IRISA - CNRS - Univ. Rennes
IRISA, Campus de Beaulieu, 35042 Rennes, France
[†] IMATAG
13 Rue Dupont-des-Loges, 35000 Rennes, France
vincent.claveau@irisa.fr, antoine.chaffin@imatag.com, ewa.kijak@irisa.fr

Abstract

The quality of artificially generated texts has considerably improved with the advent of transformers. The question of using these models to generate learning data for supervised learning tasks naturally arises, especially when the original language resource cannot be distributed, or when it is small. In this article, this question is explored under 3 aspects: (i) are artificial data an efficient complement? (ii) can they replace the original data when those are not available or cannot be distributed for confidentiality reasons? (iii) can they improve the explainability of classifiers? Different experiments are carried out on classification tasks — namely sentiment analysis on product reviews, Fake News detection and news categorization — using artificially generated data by fine-tuned GPT-2 models. The results show that such artificial data can be used in a certain extent but require pre-processing to significantly improve performance. We also show that bag-of-words approaches benefit the most from such data augmentation.

Keywords: Text generation, Data augmentation, Classification, Categorization, Fake News, Sentiment Analysis

1. Introduction

Even if text generation is not a new technology, recent neural approaches based on transformers offers good enough performance to be used in various contexts (Vaswani et al., 2017). In this paper, we explore the use of artificially generated texts for supervised machine learning tasks within two different scenarios: the artificial data is used as a complement of the original training dataset (for instance, to yield better performance) or 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 (Amin-Nejad et al., 2020)). The generation of these artificial texts is performed with a neural language model trained on the original training texts. In this paper, we show the interest of these scenarios with several text classification tasks, handling well written or noisy language: fake news detection, opinion mining and news categorization.
Precisely, the main research questions studied in this paper are the following ones:

1. what is the interest of text generation to improve text classification (complement);
2. what is the interest of text generation to replace the original training data (substitution);
3. what is the interest of text generation for explainable classifiers, based on bag-of-words representation.

In the remaining of the paper, after a presentation of related work in Section 2, we detail our classification approaches based on artificial text generation (Sec. 3). The tasks and experimental data are described in Section 4. The experiments and their results for each of our research questions are reported in Section 5 for the neural classifiers and Section 6 for bag-of-words based classifiers.

2. Related work

Data augmentation for tasks of Natural Language Processing (NLP) has already been explored in several studies. Some researchers propose more or less complex automatic modifications of the original examples in order to create new examples that are differently worded but similar with respect to the NLP task (same class, same relation between words...). This can be done for instance by simply replacing some words by synonyms (Kobayashi, 2018; Wei and Zou, 2019; Mueller and Thyagarajan, 2016; Jungiewicz and Smywinski-Pohl, 2019). The synonyms can be found in external resources such as WordNet (Miller, 1995) or in distributional thesauri, or computed from static word embeddings (such as Glove (Pennington et al., 2014) or word2vec (Mikolov et al., 2013)).
In a similar vein since it only modifies the original examples locally, some neural techniques exploit masked language models (such as BERT (Devlin et al., 2019)), that is, context-sensitive word embedding. These approaches works by masking a word in an original examples with the [MASK] token and to condition its replacement by a word from the expected class (Wu et al., 2019). It allows to generate a new example by replacing a word with another semantically close word (ideally a synonym). It is worth noting that, contrary

to what we propose, the new example is not totally different (the syntactic structure of the new example is for instance identical to the original one).
Other approaches make the most of language models such as GPT-2 (Generative Pre-Trained Transformers (Radford et al., 2019)) in order to produce a large quantity of data (texts) that are similar to the original data distribution. In Information Retrieval, this principle has been exploited to expand users' queries (Claveau, 2020b). Even closer, text generation has been used for relation extraction (Papanikolaou and Pierleoni, 2020), sentiment analysis of critics and questions (Kumar et al., 2020) or for the prediction of hospital readmission and phenotype classification (Amin-Nejad et al., 2020). This paper is part of this line of work. Our interest here is to examine the gains and losses of our different scenarios of using artificial data, their preparation, and to examine their effects on different families of classifiers.

3. Generating artificial data

Let us assume to have a set of original texts $\mathcal{T}$ divided into n classes c_i, from which we wish to generate artificial texts $\mathcal{G}_{c_i}$ for each class c_i. As explained in the introduction, we want to examine different scenarios of usage of these generated data: complement or substitution. The scenarios, as well as the usual text classification framework, are examplified in Figure 1.
We use GPT models to generate the artificial texts. These models are built by stacking *transformers* layers (more precisely decoders), and trained on large corpora by auto-regression, i.e. on a task of predicting the next word (or *token*) knowing the previous ones. The second version, GPT-2 (Radford et al., 2019), contains 1.5G parameters for its largest model, trained on more than 8 million documents from Reddit (i.e. general domain language such as discussions on news articles, mostly in English).
A newer version, GPT-3, has been released in July 2020; it is much more larger (175 billion parameters) and outperforms GPT-2 on any tested task. Yet, the experiments reported below needs fine-tuning, which is not feasible with such a large model which rather rely on prompt engineering for task adaptation. Thus, we rely on the GPT-2 framework for the experiments presented below.

3.1. Fine-tuning the language model.

For this fine-tuning step, we start from the medium model (774M parameters) pre-trained for English and made available by OpenAI[1].
In the work presented in this paper, we fine-tune one language model per class with the original training data $\mathcal{T}$. Another training procedure available in the literature is to adapt a single model, but to condition it with a special *token* indicating the expected class at the beginning of the text sequence (i.e. at the beginning of each original example), as done by CTRL (Keskar et al., 2019). . Due to the limited amount of data available per class (compared to the number of parameters of the GPT-2 model), it is important to control the fine-tuning to avoid overfitting (see our setting in the Appendix). On a Tesla V100 GPU card, this fine-tuning step lasts about 1 hour for each dataset (see below).

3.2. Text generation.

For each class c_i of the dataset $\mathcal{T}$, we use the corresponding model to generate artificial texts $\mathcal{G}_{c_i}$ which hopefully will fall into the desired class. We provide prompts for these texts in the form of a start-of-text token followed by a word randomly drawn from the set of original texts. Several parameters can influence the generation; those used in our experiments are described in the Appendix.
The texts generated for the class c_i containing a sequence of 5 consecutive words appearing identically in a text of $\mathcal{T}_{c_i}$ are removed. This serves two purposes : on the one hand, it limits the risk of revealing an original document in the case where the $\mathcal{T}_{c_i}$ data are confidential, and on the other hand, it limits the duplicates which are harmful to the training of a classifier in the case where the $\mathcal{G}_{c_i}$ data are used in addition to $\mathcal{T}_{c_i}$. In practice, it concerns about 10% of the generated texts in our experiments. In the experiments reported below, 16,000 texts are generated for each c_i class (this number of texts has been fixed arbitrarily). Note that in the scenario where the data are confidential, providing the generator itself is not possible, since it may be used to find back the data it was fine-tuned on.

3.3. About confidentiality

In the scenario where the original data cannot be distributed, notably for confidentiality reasons, it is appropriate to ask whether sensitive information can be recovered with the proposed approach. As said earlier, if the whole generative model is made available, this risk has been studied (Carlini et al., 2020), and exists, at least from a theoretical point of view under particular conditions[2].
When only the generated data are made available, there is also a risk of finding confidential information in them. Without other safeguards, it is indeed possible that among the generated texts, some are paraphrases of sentences of the training corpus. However, in practice, the risk is very limited:

- first of all, because there is no way for the user to distinguish these paraphrases among all the generated sentences;
- secondly, because additional measures can be taken upstream (for example, de-identification of the training corpus) and downstream (deletion of

[1] https://github.com/openai/gpt-2

[2] See also the discussion on the Google AI blog: https://ai.googleblog.com/2020/12/privacy-considerations-in-large.html.

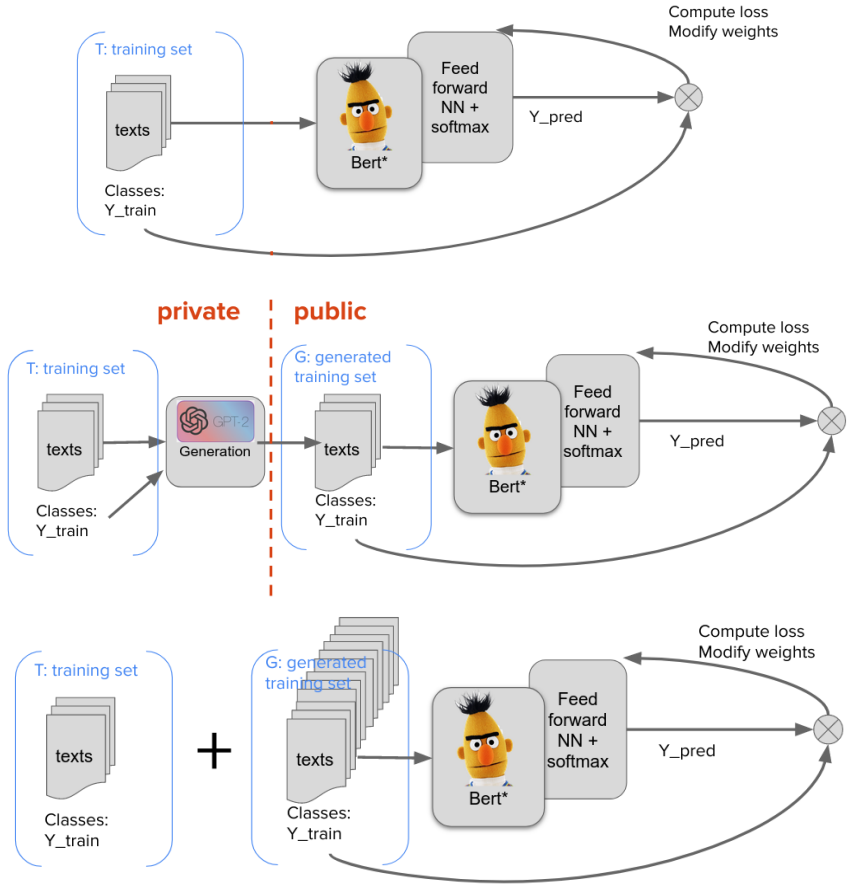

Figure 1: Different training scenarios: usual text classification framework (here with a standard BERT classifier), generated data used as substitution (especially useful when the original data cannot be shared), generated data as complement.

generated sentences containing specific or nominative information...);

- Finally, more complex systems to remove paraphrases, such as those developed for the *Semantic Textual Similarity* tasks (Jiang et al., 2020, *inter alia*), can even be considered.

These measures make it highly unlikely that any truly usable information can be extracted from the generated data.

4. Classification tasks and datasets

The experiments detailed in the next section are real classification tasks: fake news detection in tweets and sentiment analysis in reviews, news article categorization. They are classification tasks usually dealt with by machine learning. We test different languages: the fake news and news categorization datasets consists of tweets/texts in English while the sentiment analysis dataset is in French. They are presented hereafter.

4.1. FakeNews MediaEval 2020: English dataset of tweets

This dataset was developed for the detection of fake news within social networks as part of the MediaEval 2020 FakeNews challenge (Pogorelov et al., 2020). In this task, tweets about 5G or coronavirus were manually annotated according to three classes $c_i, i \in \{'5G', 'other', 'non'\}$ (Schroeder et al., 2019). $'5G'$ contains tweets propagating conspiracy theories associating 5G and coronavirus, $'other'$ are for tweets propagating other conspiracy theories (which may be about 5G or covid but not associated), and $'non'$ tweets not propagating any conspiracy theories.

It is worth noting that the classes are imbalanced; indeed, in the training dataset $\mathcal{T}$: $|\mathcal{T}_{5G}| = 1,076, |\mathcal{T}_{\text{other}}| = 620, |\mathcal{T}_{\text{non}}| = 4,173$.

The data augmentation (i.e., text generation) is performed as explained in the previous section. Figure 2 presents three examples of generated texts from the MediaEval 2020 '5G' class.

```
- 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.
- We don't know if it's the 5G or the O2 masks that are killing
  people.  It's the COVID19 5G towers that are killing people.  And
  it's the Chinese people that are being controlled by the NWO
```

Figure 2: Examples of tweets artificially generated with a GPT-2 model trained on the MediaEval examples with class $\mathcal{T}_{5G}$.

4.2. AG-news: news classification

AG-news is a large collection of news articles in English [3]. It has been used for different NLP tasks. Hereafter, we use it as a classification dataset, as proposed by (Zhang et al., 2015). In this setting, short snippets of texts (usually the article title and one or two sentences summarizing the article) are associated with 4 different category labels: "World", "Sports", "Business", "Sci/Tech". We use data and the train/test split as provided by HuggingFace Datasets[4] in which the classes are balanced. Compared with the previous dataset, the interest is that AG-news has more than 2 classes, which make the generation task a priori more complex. The generation is done similarly to the previous dataset, by fine-tuning a English GPT-2 medium model on the training examples.

4.3. FLUE CLS-FR: French dataset for sentiment analysis

The third dataset is taken from the FLUE evaluation suite for French (Le et al., 2020). It is the French part of the Cross Lingual Sentiment (CLS-FR) dataset (Prettenhofer and Stein, 2010), which consists of product reviews (books, DVD, music) from Amazon. The task is to predict whether the review is positive (rated more than 3 stars on the merchant site) or negative (less than 3 stars). The dataset is divided into balanced training and test sets. In our experiments, we do not distinguish between products : we have only two classes (positive, negative) with reviews of books, DVDs or music.

As with the MediaEval data, a language model is tuned for each class using the training data. Generation is then done as described in the previous section. Examples of generated negative reviews are given in Figure 3.

As can be seen from these examples (including the MediaEval examples in Figure 2), the generated texts seem to belong to the expected class (see Section 5.2 for a discussion of this point). However, they often have flaws that make the fact that they were generated detectable. This is particularly the case for French texts, which can be explained by the fact that we did not have, at the time of the experiments, a pre-trained model for French; the model, as well as the tokenizer, are therefore based on the English GPT model. GPT-2 models for French released very recently[5] could improve this aspect if distributed.

5. Experiments: neural classification approaches

In the experiments reported below, the performance is measured in terms of micro-F1 (equivalent to accuracy), and, to take into account the imbalance of the classes (in the MediaEval dataset), in terms of macro-F1 and MCC (Matthews Correlation Coefficient[6]), as implemented in the library scikit-learn (Pedregosa et al., 2011). The performance is measured on the respective official test sets of the MediaEval (Pogorelov et al., 2020) and CLS-FR (Le et al., 2020) tasks, of course disjoint from the training sets $\mathcal{T}$.

5.1. First results

For our first experiments, we use state-of-the-art neural classification models based on transformers. For the MediaEval data, in English, we opt for a RoBERTa (Liu et al., 2019) pre-trained model for English (*large* model with a classification layer). It is this type of transformer-based models that obtained the best results on these data during the MediaEval 2020 challenge (Cheema et al., 2020; Claveau, 2020a). Among the variants of BERT (Devlin et al., 2019), RoBERTa was preferred here for its tokenizer that is more adapted to the specifics of the very free form of writing found

[3] http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html

[4] https://huggingface.co/datasets/viewer/?dataset=ag_news

[5] For example, the Pagnol model of LightOn: https://lair.lighton.ai/pagnol/.

[6] Also called Φ coefficient; see the dedicated Wikipedia page.

- Déçue... J'ai eu je l'avoue du mal à lire ce livre arrivé au milieu de celui-ci. L'histoire ne paraît pas vraiment très réaliste. Le policier est plus guidé par de la chance que par son instinct. Que se serait-il passer s'il n'avait pas rencontré cette dame insolite ? Non ! Je ne crois pas que je lirais d'autres livres de fred vargas... Dommage je n'ai pas encore trouvé une source infaillible de bons polars.
  translation: Disappointed... I have to admit that I had a hard time reading this book until the middle of it. The story doesn't really seem very realistic. The policeman is guided more by luck than by his instinct. What would have happened if he had not met this unusual lady? No! I don't think I would read any more books by fred vargas... Too bad I haven't found an infallible source of good thrillers yet.

- De la daube. Cet homme ferait mieux de mettre son piano à la benne. Il n'y a pas de musicalité, ce disque irrite et agresse, ou au mieux il agresse et abuse son timbre et pénible accent amoureux. Musicalement, c'est de la musique de... chandler, on se dit... "c'mere irons up". Une chose est surement restée disponible sur cet album, mais attention aux maisons de disque !
  translation: Rubbish. This man would do better to put his piano in the garbage. There is no musicality, this record irritates and assaults, or at best he assaults and abuses his timbre and painful love accent. Musically, it is music of... chandler, we say to ourselves... "c'mere irons up". A thing surely remained available on this album, but attention to the record companies!

- Gros navet. Décor atrocement kitch, couleurs d'un mauvais goût abominable qui rendrait effleuré un ami en le dire... ça marche. Aucun suspense, tout est répétitif, les personnages sont inconséquents, ennuyeux. A eviter absolument.
  translation: Such a turkey. Atrociously kitsch decor, colors of an abominable bad taste that would make a friend shudder to say it. No suspense, everything is repetitive, the characters are inconsistent, boring. To avoid at all costs.

Figure 3: Examples of artificial reviews generated with a GPT-2 model trained on the CLS-FR examples with the class $\mathcal{T}_{negatif}$.

in tweets (mix of upper and lower case, absence or multiplication of punctuation, abbreviations...). For the CLS-FR data of FLUE, we use the *large-cased* FlauBERT model (Le et al., 2020). This allows us to compare with the results originally published on these data. See the appendix for more details about the implementation used.

We evaluate the performance according to our different training scenarios: on the original data $\mathcal{T}$ (which serves as a *baseline*), on the artificial data $\mathcal{G}$, and finally on both the artificial and original data. In this last case, we test two training strategies :

- the first, $\mathcal{T} + \mathcal{G}$, mixes the original and artificial examples,
- the second, $\mathcal{G}$ then $\mathcal{T}$, trains on the artificial data on the first epochs, then on the original data for the last epoch. This results in a kind of fine-tuning on the original data after a first training on the artificial data.

The results for the MediaEval and CLS-FR datasets are reported in Table 1. On the CLS-FR data, we observe very few differences between the different scenarios and the baseline (note that our baseline is similar to the published state-of-the-art results). The classification task, which is relatively simple, obviously makes it possible to generate artificial data of as good quality as the original data, leading to comparable results, even without training on the original texts. On this type of task, artificially generated data can therefore be used without loss of performance.

The MediaEval task is more difficult as can be seen with the results of the baseline (RoBERTA / $\mathcal{T}$). On these data, in a substitution scenario (i.e. when the generated data are used alone as training data), the results are strongly degraded compared to a system trained on

model	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
BERT* / $\mathcal{T}$	79.57	**62.66**	**55.71**	95.44	95.42	90.86	94.35	94.35	92.47
BERT* / $\mathcal{G}$	62.68	54.03	39.27	95.13	95.12	90.25	90.25	90.25	87.12
BERT* / $\mathcal{T} + \mathcal{G}$	75.01	58.81	46.37	95.43	95.42	90.89	92.44	92.44	88.51
BERT* / $\mathcal{G}$ then $\mathcal{T}$	**79.89**	60.64	52.02	**95.76**	**95.75**	**91.51**	**94.88**	**94.88**	**92.57**

Table 1: Performance (%) of neural classification techniques on data from MediaEval and CLS-FR according to the scenario of usage of the generated texts (without filtering) (cf. Sec. 5.1). The BERT* model are respectively RoBERTA and FlauBERT.

the original data. This is of course due to the fact that the data generated by each of the language models may not belong to the expected class, as the models do not fully capture the specificity of the *fine-tuning* data. In a scenario where we aim at augmenting the training data, the impact is less significant, especially if the artificial data is used only on the first few epochs.

5.2. Results with automatic filtering

As we have seen, the examples generated by our trained GPT-2 models $\mathcal{G}$ may contain texts that do not belong to the expected classes. Manually filtering or annotating these texts is of course possible but remains a costly task. To reduce the effect of these texts on the classification at a low cost, we propose to exclude them using a classifier learned on the original data $\mathcal{T}$. Its goal is to filter the generated examples: any text of $\mathcal{G}_{c_i}$ which is not classified c_i by the classifier is removed. In this way, we hope to eliminate, automatically, the most obvious cases of problematic artificial texts. In the following experiments, we use the RoBERTa classifier trained on $\mathcal{T}$ (evaluated in the first row of Tab. 1). In this way, 40% of the examples are removed. The resulting filtered artificial dataset is noted $\mathcal{G}^f$.

The results with these new filtered sets of artificial examples in the same training scenarios are presented in Table 2 for the MediaEval and CLS-FR data. It can be seen that this filtering strategy pays off, with improved performance on all metrics compared to no filtering. In the substitution scenario, the performance is now close to the baseline, and is even better on the macro-F1; this is explained by the fact that the artificial set $\mathcal{G}$ is much more balanced than $\mathcal{T}$ and thus performs better on the minority classes of the test set. In the complement scenario, we observe a significant improvement over the baseline, especially with the sequential strategy $\mathcal{G}^f then \mathcal{T}$.

5.3. Differences between classifiers

Beyond the global performance measures, it is interesting to check if the classifier trained on the artificial data allows to make the same decisions as a classifier trained on $\mathcal{T}$. To do so, we can look at the proportion of examples (from the test set) for which the decision of BERT* / $\mathcal{T}$ and BERT* / $\mathcal{G}^f$ differs. For the CLS-FR data, the classifiers agree on a large majority of examples. Figure 4 shows the confusion matrix of FlauBERT / $\mathcal{T}$ and FlauBERT / $\mathcal{G}^f$ on the CLS-FR data.

From this confusion matrix, we can see that the classifiers do agree on the majority of examples. The cases of disagreement are proportionally more important on the false positives and false negatives, but even for these categories, we still find a lot of common errors (42 and 77 examples respectively for the false positives and false negatives). The classifiers have therefore not only comparable performance, but very similar behaviors in detail since they give the same class on most examples.

6. Experiments: bag-of-words approaches

We also tested classifiers based on bag-of-words representations; we present only the results of the logistic regression (LR) which gave the best results. In general, these classifiers perform less well than the transformers-based approaches, but they allow for better explainability (Miller, 2018; Carvalho et al., 2019, for a definition and characterization of learning methods), for example by examining the regression weights associated with words. They are also way less expensive to train.

6.1. First results

The implementation used is scikit-learn (Pedregosa et al., 2011), the texts are vectorized with TF-IDF weighting (see Appendix for more details). Results for the same scenarios as above are presented for the MediaEval, CLS-FR and AG-news tasks in Table 3.

For this type of classifier, the interest of the generated data appears for both scenarios and on the two datasets. In the case of substitution, the classifiers are slightly better than those trained on the original data. This demonstrates the importance of having a larger amount of data to capture form variants in texts (synonyms, paraphrases...) that the bag-of-words representations cannot otherwise capture as easily as the pre-trained embedding-based representations of the BERT models. In the scenario where data is used as a complement, the performance increase is even more marked and thus gets closer to the neural baseline, while having the advantages of a classifier considered more interpretable.

6.2. Impact of the quality of the generated data

It is interesting to examine what is the influence of the quality of the generated data (even filtered) on the results of the final classifier (see Section 5.2). To study

model	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
BERT* / $\mathcal{T}$	79.57	62.66	55.71	95.44	95.42	90.86	94.35	94.35	92.47
BERT* / $\mathcal{G}^f$	76.22	64.18	52.75	95.76	95.75	91.51	93.49	93.49	91.35
BERT* / $\mathcal{T} + \mathcal{G}^f$	80.12	66.08	57.44	**95.99**	**95.98**	**91.97**	93.47	93.47	91.34
BERT* / $\mathcal{G}^f$ then $\mathcal{T}$	**83.55**	**67.90**	**60.05**	95.96	95.95	91.96	**95.10**	**95.10**	**92.89**

Table 2: Performance (%) or neural classification approaches on the MediaEval, CLS-FR, and AG-news tasks according to our scenarios of usage of the artificially generated texts after filtering (cf. Sec. 5.2). The BERT* model are respectively RoBERTA, FlauBERT and RoBERTA.

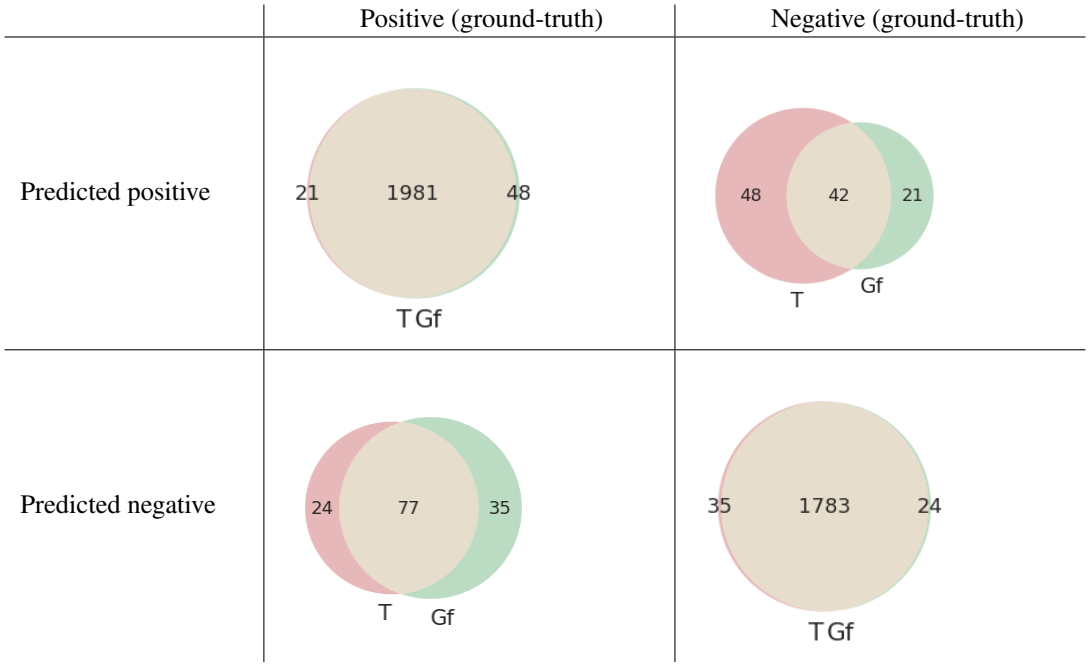

Figure 4: Confusion matrix of the FlauBERT / $\mathcal{T}$ and FlauBERT / $\mathcal{G}^f$ models on the CLS-FR data. The Venn diagrams shows the proportions of shared examples for each category.

this, we simulate filtering done with classifiers of varying quality (accuracy). This is done simply by replacing, for randomly drawn examples of $\mathcal{G}^f$, the predicted class (by the generator and by the filtering classifier) by a randomly drawn class. For instance, for a (randomly picked) text generated for the class 'a' (and classified as 'a' by the filtering classifier), we change its label to class 'b' (randomly picked among the classes but 'a'). The number of examples undergoing this treatment is computed so that the errors inserted make the accuracy of the dataset drops to 80,%, 70,%, etc. The effect of these errors in the generated examples on the final performance of the complement and substitution strategies are presented in Figure 5 (MediaEval data) with logistic regression as final classifier.

As can be seen in this figure, empirical results about the influence of filtering quality are unsurprising. In the substitution scenario, the final performance is strongly dependent on the quality of the filtering classifier; in this case, a performance level equivalent to the original dataset is achieved when the accuracy of the filter exceeds 70 %. In the case of the complement scenario, the gain is significant as soon as the filter has an accuracy higher than random.

7. Concluding remarks

In this work, we have explored the interest of text generation for three text classification tasks (news categorization, fake news detection in tweets and sentiment analysis on product reviews). In a scenario where the original language resource or training data cannot be distributed, we have shown that it is possible to generate artificial data for supervised learning purposes. For state-of-the-art classifiers based on transformers, using the generated data without additional precautions degrades the performance (compared to the one achieved with the original data) but in a contained proportion (-4% accuracy). Yet, if the generated texts are automatically filtered with a classifier trained on the original data (which can also be kept private), we have shown that it is possible to get equivalent or superior performance than with the original data. The generation model and the filtering model can be kept private and the (filtered) generated texts can be distributed, which

model	MediaEval			CLS-FR			AG-news		
	micro-F1	macro-F1	MCC	micro-F1	macro-F1	MCC	micro-F1	macro-F1	MCC
LR / $\mathcal{T}$	72.68	56.35	42.22	84.77	84.70	69.48	69.32	69.32	62.24
LR / $\mathcal{G}^f$	74.00	59.18	44.39	87.16	87.14	74.27	**83.82**	**83.82**	78.59
LR / $\mathcal{T} + \mathcal{G}^f$	**75.46**	**59.64**	**45.83**	**88.36**	**88.34**	**76.69**	83.47	83.47	**78.65**

Table 3: Performance (%) of the LR/bag-of-words approach on the MediaEval, CLS-FR and AG-news datasets according to our scenarios of usage of the artificially generated data after filtering: without, substitution, complement.

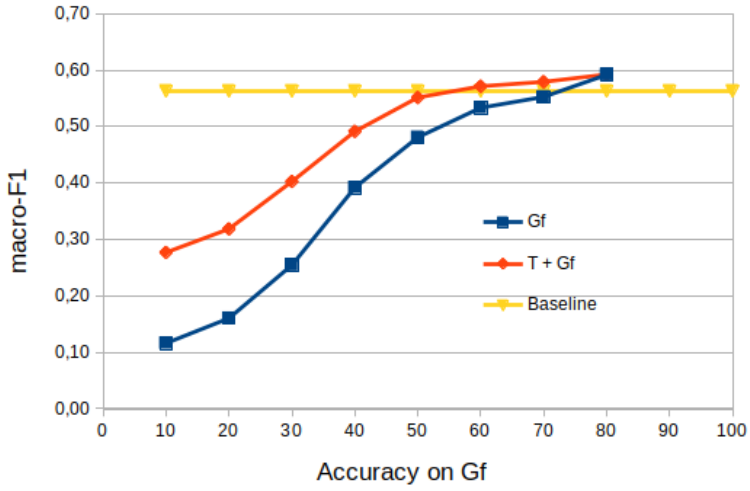

Figure 5: Performance (macro-F1) according to the quality (accuracy in %) of the classifier filtering the artificially generated data; MediaEval dataset with logistic regression.

complies with use cases in which the sensitive original data cannot be distributed.

For classifiers exploiting bag-of-words representations, we notice in every case an improvement due to the larger amount of training data available, which make it possible to get good results with more explainable ML methods if needed.

In a scenario where artificial data is added to the original data, we have shown that classifiers benefit from additional data, including neural networks. This result is particularly positive for the bag-of-words approaches, which are more sensitive to reformulations, and which clearly benefit from the addition of these artificial examples. We have seen that for some datasets, it even allows the performance to get close to a BeRT-based model. We thus have a good compromise between methods that are fast to train, more easily explainable, while having performance close to neural networks.

As we have seen, these best results are obtained provided that the generated data are filtered beforehand, which seems to contradict several studies cited in Sec. 2. In our experiments, this was done automatically; manual correction of the data (of their classes) is also possible and may allow better results, but with an additional annotation cost.

The use of these methods for other data and other NLP tasks than text classification remains a promising avenue. Among these NLP tasks, those based on word labeling (token classification) pose different problems and require adapted solutions. In the future, it would be interesting to verify the consistency of our results according to other generation approaches (Kumar et al., 2020). It also seems interesting to study more deeply the impact of the quality of the classifier used to filter the artificial data. Moreover, the integration of the filtering step as a constraint during the generation of artificial examples is a promising avenue.

8. Acknowledgements

This research was partially funded by a CIFRE grant from the Association Nationale Recherche Technologie and by the Agence de l'Innovation de Défense (contract 235483).

Appendix: Parameters and Reproducibility

Fine-tuning the generative models

All the generative models used are based on the large (774M parameters) GPT2 model, as provided by OpenAI. They are fine-tuned on the training data and we limit the number of steps to 2,000 in order to avoid overfitting. The other fine-tuning parameters are the default ones of the OpenAI GPT2 code that is used in our experiments.

Inference with the generative models

For the generation, we used the default values that we give here for reproducibility purposes, without detailing them (see the GPT-2 documentation): `temp.` = 0.7, `top_p` = 0.9, `top_k` = 40. The generation of examples relies on `https://github.com/minimaxir/gpt-2-simple`.

BERT-based classification

The classification models used either rely on RoBERTa, for English datasets, or FlauBERT for the French dataset. We use the implementation of the HuggingFace's transformer library (Wolf et al., 2020), with the ModelForSequenceClassification method.
The batch size is set to 16 and the number of *epochs* set to 3 in all scenarios (optimal number of epochs for the baseline), except for the last one (3 on $\mathcal{G}$ followed by 1 on $\mathcal{T}$).

BoW-based classification

For our experiments based on Logistic Regression, the implementation used is the one of scikit-learn (Pedregosa et al., 2011). The texts are vectorized with TF-IDF weighting and L2-normalized, and the LR parameters are the default ones except for the following: multiclass strategy *one-vs.-rest*. The number of iterations is set to a high value (2500), which ensures convergence for each of our experiments.

Replicability

For replicability purposes, the training scenarios presented in this article are available online:

- for the MediaEval dataset, at `https://colab.research.google.com/drive/1VDm-MZcgVJpMaVmmGa1Wvmyc71q6IYBJ`
- for the CLS-FR dataset, at `https://colab.research.google.com/drive/1i2IOBV5yEi2ID9atyMBn6PdX5xnschtK`

The data are available from their producers (see Section 4).

9. Bibliographical References

Amin-Nejad, A., Ive, J., and Velupillai, S. (2020). Exploring transformer text generation for medical dataset augmentation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4699–4708, Marseille, France, May. European Language Resources Association.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Raffel, C. (2020). Extracting training data from large language models. *arXiv*.

Carvalho, D. V., Pereira, E. M., and Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8).

Cheema, G. S., Hakimov, S., and Ewerth, R. (2020). TIB's Visual Analytics Group at MediaEval '20: Detecting Fake News on Corona Virus and 5G Conspiracy. In *MediaEval Benchmarking Initiative for Multimedia Evaluation (MediaEval 2020)*, online, United States, December.

Claveau, V. (2020a). Detecting fake news in tweets from text and propagation graph: IRISA's participation to the FakeNews task at MediaEval 2020. In *MediaEval Benchmarking Initiative for Multimedia Evaluation (MediaEval 2020)*, online, United States, December.

Claveau, V. (2020b). Query expansion with artificially generated texts. *CoRR*, abs/2012.08787.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Jiang, H., He, P., Chen, W., Liu, X., Gao, J., and Zhao, T. (2020). SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190, Online, July. Association for Computational Linguistics.

Jungiewicz, M. and Smywinski-Pohl, A. (2019). Towards textual data augmentation for neural networks: synonyms and maximum loss. *Computer Science*, 20(1).

Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., and Socher, R. (2019). CTRL: A conditional transformer language model for controllable generation. *CoRR*, abs/1909.05858.

Kobayashi, S. (2018). Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana, June. Association for Computational Linguistics.

Kumar, V., Choudhary, A., and Cho, E. (2020). Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China, December. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S.,

and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, et al., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Miller, T. (2018). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267.

Mueller, J. and Thyagarajan, A. (2016). Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2786–2792. AAAI Press.

Papanikolaou, Y. and Pierleoni, A. (2020). Dare: Data augmented relation extraction with gpt-2.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*.

Schroeder, D. T., Pogorelov, K., and Langguth, J. (2019). Fact: a framework for analysis and capture of twitter graphs. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 134–141. IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Wei, J. and Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China, November. Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Wu, X., Lv, S., Zang, L., Han, J., and Hu, S. (2019). Conditional BERT Contextual Augmentation. In João M. F. Rodrigues, et al., editors, *Computational Science – ICCS 2019*, pages 84–95, Cham. Springer International Publishing.

10. Language Resource References

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbe, B., Besacier, L., and Schwab, D. (2020). FlauBERT: Unsupervised Language Model Pre-training for French. In *LREC*, Marseille, France.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November.

Pogorelov, K., Schroeder, D. T., Burchard, L., Moe, J., Brenner, S., Filkukova, P., and Langguth, J. (2020). Fakenews: Corona virus and 5g conspiracy task at mediaeval 2020. In *MediaEval 2020 Workshop*.

Prettenhofer, P. and Stein, B. (2010). 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. Association for Computational Linguistics.

Zhang, X., Zhao, J., and LeCun, Y. (2015). 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. MIT Press.