Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles
Lille, France, 28 juin au 2 juillet 2021

La génération de textes artificiels en substitution ou en complément de données d'apprentissage

Vincent Claveau[1] Antoine Chaffin[1,2] Ewa Kijak[1]
(1) IRISA - CNRS, Univ. Rennes, Campus de Beaulieu, 35042 Rennes, France
(2) IMATAG, Rennes, France
{vincent.claveau,ewa.kijak}@irisa.fr, antoine.chaffin@imatag.com

RÉSUMÉ

La qualité des textes générés artificiellement s'est considérablement améliorée avec l'apparition des *transformers*. La question d'utiliser ces modèles pour augmenter les données d'apprentissage pour des tâches d'apprentissage supervisé se pose naturellement. Dans cet article, cette question est explorée sous 3 aspects : (i) les données artificielles sont-elles un complément efficace ? (ii) peuvent-elles remplacer les données d'origines quand ces dernières ne peuvent pas être distribuées, par exemple pour des raisons de confidentialité ? (iii) peuvent-elles améliorer l'explicabilité des classifieurs ? Différentes expériences sont menées sur une tâche de classification en utilisant des données générées artificiellement en adaptant des modèles GPT-2. Les résultats montrent que les données artificielles ne sont pas encore suffisamment bonnes et nécessitent un pré-traitement pour améliorer significativement les performances. Nous montrons que les approches sac-de-mots bénéficient le plus de telles augmentations de données.

ABSTRACT

Generating artificial texts as substitution or complement of training data

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. 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 a classification task using artificially generated data by fine-tuned GPT-2 models. The results show that such artificial data are not yet good enough and require pre-processing to significantly improve performance. We show that bag-of-word approaches benefit the most from such data augmentation.

MOTS-CLÉS : Génération de textes, augmentation de données, classification.

KEYWORDS: Text generation, data augmentation, classification.

1 Introduction

Si la génération artificielle de texte n'est pas une tâche nouvelle, les approches récentes à base de *transformers* offrent des performances suffisamment bonnes pour être employées dans de nombreux contextes (Vaswani *et al.*, 2017). Dans cet article, nous explorons l'utilisation de données générées pour des tâches d'apprentissage supervisé dans différents contextes d'utilisation afin de compléter

les données d'entraînement originales (pour obtenir de meilleures performances) ou de se substituer (intégralement) aux données originales (par exemple, quand ces dernières ne peuvent pas être distribuées pour des raisons de confidentialité (Amin-Nejad *et al.*, 2020)). La génération de ces données est faite avec un modèle de langue neuronal appris sur les données d'origine.

Précisément, les principales questions de recherche abordées dans cet article sont les suivantes :

1. quel est l'intérêt de la génération pour améliorer les performances des approches à base d'apprentissage (complément) ;
2. quel est l'intérêt de la génération pour remplacer les données d'origines (substitution) ;
3. quel est l'intérêt de la génération pour des classifieurs neuronaux et ceux dits "explicables" reposant sur des représentations sac-de-mots.

Dans la suite de cet article, après une présentation des travaux connexes (sec. 2), nous détaillons le processus d'augmentation de données que nous mettons en œuvre en section 3. Les données expérimentales sont décrites en section 4. Les expériences et résultats pour nos différentes questions de recherche sont détaillés en section 5 pour les modèles neuronaux et section 6 pour les modèles expploitant des représentations sac-de-mots.

2 Travaux connexes

L'augmentation de données pour des tâches de TAL a déjà été explorée dans plusieurs travaux. Certains proposent des modifications automatiques plus ou moins complexes des exemples originaux afin de créer une version différente en surface mais identique au sens de la tâche (même classe, même relation entre les mots...) en remplaçant par exemple des mots par leur synonymes (Kobayashi, 2018; Wei & Zou, 2019; Mueller & Thyagarajan, 2016; Jungiewicz & Smywinski-Pohl, 2019). Les synonymes sont alors tirés de ressources langagières tels WordNet, de thésaurus distributionnels ou de plongements de mots (statiques).

Dans une veine similaire, puisque ne modifiant que localement les données originales, il existe des approches neuronales exploitant les modèles de langue à base de masque de type BERT et donc des plongements contextuels. Celles-ci fonctionnent en conditionnant le remplacement du jeton [MASK] par un mot en fonction de la classe attendue (Wu *et al.*, 2019). Cela produit un nouvel exemple avec un remplacement d'un mot par un mot sémantiquement proche (idéalement un synonyme), mais ce nouvel exemple n'est pas totalement différent (la structure du nouvel exemple est très similaire à l'exemple original) comme nous proposons de le faire.

D'autres approches exploitent les capacités des modèles de langue tels GPT-2 (Generative Pre-Trained Transformers) afin de produire des données proches de la distribution du jeu initial en grande quantité. En recherche d'information, ce principe a par exemple été utilisé pour augmenter les requêtes (Claveau, 2020b). Plus proche encore, la génération est exploitée pour l'extraction de relations (Papanikolaou & Pierleoni, 2020), la classification de sentiments de critiques et de questions (Kumar *et al.*, 2020) ou la prédiction de réadmission et la classification phénotypique (Amin-Nejad *et al.*, 2020). Notre article s'inscrit dans cette lignée de travaux. Notre intérêt est ici d'examiner les gains et pertes des différents scénarios d'emploi des données artificielles, de leur préparation, et d'examiner leurs effets sur différentes familles de classifieurs.

3 Génération de données artificielles

On suppose disposer d'un ensemble de textes (originaux) $\mathcal{T}$ divisé en n classes c_i, à partir desquels on souhaite générer des textes artificiels $\mathcal{G}_{c_i}$ pour chaque classe c_i. Nous employons les modèles GPT pour générer les textes artificiels. Ces modèles sont construits en empilant des *transformers* (plus précisément des décodeurs), entraînés sur de grands corpus par auto-régression, c'est-à-dire sur une tâche de prédiction du mot (ou *token*) suivant, sachant les précédents. La seconde version, GPT-2 (Radford *et al.*, 2019), contient 1,5G de paramètres pour son plus gros modèle, entraînés sur plus de 8 millions de documents issus de Reddit (i.e. du langage général comme des discussions sur des articles de presse, principalement en anglais).

3.1 Adaptation du modèle de langue.

Pour cette étape d'adaptation fine (*fine-tuning*), on part du modèle moyen (774M de paramètres) pré-entraîné pour l'anglais et mis à disposition par OpenAI[1].

Dans les travaux présentés dans cet article, nous adaptons un modèle de langue par classe. Une autre manière d'entraîner disponible dans la littérature consiste à adapter un unique modèle, mais de le conditionner par un *token* spécial indiquant la classe attendue en début de séquence. Du fait du peu de données disponibles par classe vis-à-vis du nombre de paramètres du modèle GPT-2, il est important de contrôler l'adaptation pour éviter le sur-apprentissage. Pour cela, nous limitons le nombre d'*epochs* à 2 000 ; les autres paramètres d'adaptation sont ceux par défaut. Sur une carte GPU Tesla V100, cette étape de *fine-tuning* dure environ 1h.

3.2 Génération.

Pour chacune des classes c_i du jeu de données $\mathcal{T}$, nous utilisons le modèle correspondant pour générer des textes artificiels $\mathcal{G}_{c_i}$ qui, nous l'espérons, relèveront bien de la même classe. Nous fournissons des amorces pour ces textes sous la forme d'une balise de début de texte suivi d'un mot tiré aléatoirement dans l'ensemble des textes originaux. Plusieurs paramètres peuvent influer sur la génération. Nous avons laissé les valeurs usuelles que nous redonnons ici, sans les détailler (voir la documentation GPT-2), à des fins de reproductibilité : `temp.` = 0,7, `top_p` = 0,9, `top_k` = 40.

Les textes générés pour la classe c_i contenant une séquence de 5 mots consécutifs apparaissant identiquement dans un texte de $\mathcal{T}_{c_i}$ sont supprimés. Cela sert deux objectifs : d'une part, cela limite le risque de dévoiler un document original dans le cas où les données $\mathcal{T}_{c_i}$ sont confidentielles, et d'autre part, cela limite les doublons néfastes à l'apprentissage d'un classifieur dans le cas où les données $\mathcal{G}_{c_i}$ sont utilisées en complément de $\mathcal{T}_{c_i}$. En pratique, cela concerne environ 10 % des textes générés dans nos expériences. Notons que dans le scénario où les données sont confidentielles, la mise à disposition du générateur lui-même n'est pas envisageable (Carlini *et al.*, 2020). Dans les expériences rapportées ci-dessous, ce sont 16 000 textes qui sont ainsi générés pour chaque classe c_i (ce nombre de textes a été fixé arbitrairement).

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

3.3 A propos de confidentialité

Dans le scénario où les données d'origine ne peuvent pas être distribuées notamment pour des questions de confidentialité, il convient de se demander si des informations sensibles peuvent être retrouvées avec l'approche proposée. Si tout le modèle génératif est mis à disposition, ce risque a été étudié (Carlini *et al.*, 2020), et existe, du moins d'un point de vue théorique dans des conditions particulières[2].

Quand seules les données générées sont mises à disposition, il y a également des risques d'y retrouver des informations confidentielles. Sans autre garde-fou, il est en effet possible que parmi les textes générés, certains soient des paraphrases de morceaux du corpus d'entraînement. Cependant, le risque est très limité :

- tout d'abord, parce qu'il n'y a pas moyen, pour l'utilisateur, de distinguer ces paraphrases parmi toutes les phrases générées ;
- d'autre part, parce qu'en pratique, des mesures supplémentaires peuvent être prises en amont (par exemple, dé-identification du corpus d'entraînement) et en aval (suppression des phrases générées contenant des informations spécifiques ou nominatives...) ;
- enfin, des systèmes plus complexes pour supprimer des paraphrases, tels ceux développés pour les tâches *Semantic Textual Similarity* (Jiang *et al.*, 2020, par exemple), peuvent même être envisagés.

Ces mesures rendent hautement improbable la possibilité d'extraire une information réellement exploitable des données générées.

4 Tâches et jeux de données

Les expériences rapportées dans la section suivante reposent sur deux jeux de données utilisés pour des tâches de classification. L'un est composé de tweets en anglais, l'autre de textes en français. Nous les présentons ci-dessous.

4.1 Classification de textes anglais avec les données MediaEval 2020

Ce jeu de données a été développé pour la détection de fausses informations au sein des réseaux sociaux dans le cadre du challenge FakeNews de MediaEval 2020 (Pogorelov *et al.*, 2020). Dans cette tâche, des tweets sur la 5G ou le coronavirus ont été manuellement annotés selon trois classes $c_i, i \in \{'5G', 'other', 'non'\}$ (Schroeder *et al.*, 2019). $'5G'$ contient les tweets propageant des théories complotistes associant 5G et coronavirus, $'other'$ des tweets propageant d'autres théories complotistes (pouvant porter sur la 5G ou le covid mais sans les associer), et $'non'$ des tweets ne propageant pas de théories complotistes. Il est important de noter que les classes sont déséquilibrées ; ainsi dans le jeu d'entraînement $\mathcal{T}$: $|\mathcal{T}_{5G}| = 1\,076, |\mathcal{T}_{\text{other}}| = 620, |\mathcal{T}_{\text{non}}| = 4\,173$.

[2] Voir également la discussion sur le blog de Google AI.

```
- 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 1 : Exemples de tweets générés artificiellement avec le modèle GPT-2 entraîné sur les données MediaEval avec la classe $\mathcal{T}_{5G}$.

L'augmentation de données est effectuée comme décrit ci-dessus. La figure 1 présente trois exemples de textes générés à partir du jeu de données MediaEval 2020 pour la classe '5G'.

4.2 Classification de textes français avec les données de FLUE

Le deuxième jeu de données que nous utilisons est tiré de la suite d'évaluation pour le français FLUE (Le *et al.*, 2020). Il s'agit de la partie française des données Cross Lingual Sentiment (CLS-FR) (Prettenhofer & Stein, 2010), qui consiste en des commentaires de produits (livres, DVD, musique) sur Amazon. La tâche est de prédire si le commentaire est positif (noté plus de 3 étoiles sur le site marchand) ou négatif (moins de 3 étoiles). Le jeu de données est divisé en ensembles d'entraînement et de test, équilibrés. Dans nos expériences, nous ne distinguons pas les produits : nous n'avons que deux classes (positif, négatif) avec des textes traitant de livres, de DVD ou de musique.

Comme pour les données MediaEval, un modèle de langue est appris pour chacune des nos deux classes à partir des données d'entraînement. La génération est ensuite faite comme décrit dans la section précédente. Des exemples pour la classe commentaires négatifs de CLS-FR sont donnés dans la figure 2.

Comme on peut le voir sur ces exemples (y compris ceux de MediaEval en figure 1), les textes générés semblent respecter la classe attendue (voir section 5.2 pour une discussion de ce point). Pour autant, ils comportent souvent des défauts qui rendent détectable le fait qu'ils aient été générés. C'est plus particulièrement le cas pour les textes en français, ce qui peut s'expliquer par le fait que l'on ne disposait pas, au moment des expériences, de modèle pré-entraîné pour le français ; le modèle, ainsi que le tokeniseur, sont donc basés sur le modèle GPT anglais. Des modèles GPT-2 pour le français sortis très récemment[3] pourraient améliorer cet aspect là.

[3]Par exemple, le modèle Pagnol de LightOn.

```
- 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.
- 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 !
- 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.
```

FIGURE 2 : Exemples de commentaires générés artificiellement avec le modèle GPT-2 entraîné sur les données CLS-FR avec la classe $\mathcal{T}_{negatif}$.

5 Expérimentations : approches neuronales

Dans les expériences rapportées ci-dessous, les performances sont mesurées en termes de micro-F1 (équivalent au taux de bonnes classifications), et, pour prendre en compte le déséquilibre des classes (notamment dans le jeu de données MediaEval), en termes de macro-F1 et de MCC (Matthews Correlation Coefficient[4]), tels qu'implémentés dans la bibliothèque scikit-learn. Ces performances sont mesurées sur les jeux de test officiels des tâches MediaEval (Pogorelov *et al.*, 2020) et CLS-FR (Le *et al.*, 2020), bien sûr disjoints des ensembles d'entraînement $\mathcal{T}$.

5.1 Premiers résultats

Pour nos premières expériences, nous utilisons des modèles neuronaux de classification état-de-l'art. Pour les données MediaEval, en anglais, nous optons pour RoBERTa (Liu *et al.*, 2019) pré-entraîné pour l'anglais (modèle *large* avec une couche de classification). C'est ce type de modèle à base de *transformers* qui a obtenu les meilleurs résultats sur ces données lors du challenge MediaEval 2020 (Cheema *et al.*, 2020; Claveau, 2020a). Parmi les variantes de BERT (Devlin *et al.*, 2019), RoBERTa a été ici préféré pour son tokeniseur plus adapté aux spécificités d'écriture très libre que l'on trouve dans les tweets (mélange de majuscules et minuscules, absence ou multiplication de ponctuations, abréviations...). Pour les données CLS-FR de FLUE, nous utilisons FlauBERT dans son modèle *large-cased* (Le *et al.*, 2020). Cela nous permet de nous comparer aux résultats publiés initialement sur ces données.

Nous mesurons les performances selon les divers scénarios d'entraînement : sur les données d'origine $\mathcal{T}$ (ce qui constitue notre *baseline*), sur les données artificielles $\mathcal{G}$, sur les données artificielles et originales. Dans ce dernier cas, nous testons deux stratégies d'entraînement :

- la première, $\mathcal{T} + \mathcal{G}$, mélange les exemples originaux et artificiels,

[4]Également appelé coefficient Φ ; voir la page Wikipedia dédiée.

modèle	MediaEval micro-F1	MediaEval macro-F1	MediaEval MCC	CLS-FR micro-F1	CLS-FR macro-F1	CLS-FR MCC
BERT* / $\mathcal{T}$	79,57	62,66	55,71	95,44	95,42	90,86
BERT* / $\mathcal{G}$	62,68	54,03	39,27	95,13	95,12	90,25
BERT* / $\mathcal{T} + \mathcal{G}$	75,01	58,81	46,37	95,43	95,42	90,89
BERT* / $\mathcal{G}$ puis $\mathcal{T}$	79,89	60,64	52,02	95,76	95,75	91,51

TABLE 1 : Performances (%) de l'approche neuronale sur les données MediaEval et CLS-FR selon les scénarios d'usage des données artificielles (sans filtrage) (cf. sec. 5.1). Les modèles BERT* utilisés sont respectivement RoBERTA et FlauBERT.

- la deuxième, $\mathcal{G}$ puis $\mathcal{T}$, entraîne sur les données artificielles sur les premières *epochs*, puis sur les données originales pour la dernière *epoch*. Cela implémente une sorte de *fine-tuning* sur les données originales après un premier entraînement sur les données artificielles.

L'implémentation que nous utilisons est celle d'HuggingFace (Wolf *et al.*, 2020) avec une taille du batch fixée à 16 et le nombre d'*epochs* fixé à 3 dans tous les scénarios (nombre d'*epochs* optimal pour la *baseline*), sauf le dernier (3 sur $\mathcal{G}$ puis 1 sur $\mathcal{T}$).

Les résultats pour les jeux de données MediaEval et CLS-FR sont reportés dans le tableau 1. Sur les donnés CLS-FR, on observe très peu de différences entre les différents scenarii et par rapport à la baseline (et notre baseline est tout à fait en ligne avec les résultats état-de-l'art (Le *et al.*, 2020)). La tâche de classification, relativement simple, permet visiblement de générer des données d'aussi bonne qualité que les données originales, menant à des résultats comparables. Sur ce type de tâche, l'utilisation de données générées artificiellement peut donc se faire sans perte de performances.

Les données MediaEval sont plus difficiles comme on peut le voir avec les résultats de la baseline (RoBERTA / $\mathcal{T}$). Sur ces données, dans un scénario de substitution (i.e. quand les données générées servent seules de données d'entraînement), les résultats sont fortement dégradés par rapport à un système entraîné sur les données originales. Cela s'explique bien sûr par le fait que les données générées par chacun des modèles de langue peuvent ne pas appartenir à la classe attendue, les modèles ne capturant pas complètement la spécificité des données de *fine-tuning*. Dans un scénario de complément des données d'apprentissage, l'impact est moins important, particulièrement si les données artificielles sont utilisées uniquement sur les premières *epochs*.

5.2 Résultats avec filtrage automatique

Comme nous l'avons vu, les exemples $\mathcal{G}$ générés par nos modèles GPT-2 peuvent contenir des textes ne relevant pas des classes attendues. Filtrer ou annoter manuellement ces textes est bien sûr possible mais reste une tâche coûteuse. Pour diminuer l'effet de ces textes sur la classification à moindre coût, nous proposons de les exclure à l'aide d'un premier classifieur appris sur les données originales $\mathcal{T}$: tout texte de $\mathcal{G}_{c_i}$ qui n'est pas classé c_i par le classifieur est exclu. On espère ainsi éliminer, automatiquement, les cas les plus évidents de texte artificiels problématiques. Dans les expériences suivantes, nous utilisons le classifieur RoBERTa entraîné sur $\mathcal{T}$ (évalué en première ligne de tab. 1). Ce sont ainsi 40 % des exemples qui sont supprimés. Les exemples artificiels gardés sont notés $\mathcal{G}^f$.

Les résultats avec ces nouveaux jeux épurés d'exemples artificiels dans les mêmes scénarios d'en-

modèle	MediaEval micro-F1	MediaEval macro-F1	MediaEval MCC	CLS-FR micro-F1	CLS-FR macro-F1	CLS-FR MCC
BERT* / $\mathcal{T}$	79,57	62,66	55,71	95,44	95,42	90,86
BERT* / $\mathcal{G}^f$	76,22	64,18	52,75	95,76	95,75	91,51
BERT* / $\mathcal{T} + \mathcal{G}^f$	80,12	66,08	57,44	95,99	95,98	91,97
BERT* / $\mathcal{G}^f$ puis $\mathcal{T}$	83,55	67,90	60,05	95,96	95,95	91,96

TABLE 2 : Performances (%) de l'approche neuronale sur les données MediaEval et CLS-FR selon les scénarios d'usage des données artificielles après filtrage (cf. sec. 5.2). Les modèles BERT* utilisés sont respectivement RoBERTA et FlauBERT.

traînement sont présentés dans le tableau 2 pour les données MediaEval et CLS-FR. On constate que cette stratégie de filtrage se révèle payante, les performances étant améliorées sur l'ensemble des métriques par rapport à l'absence de filtrage. Dans le scénario de substitution, les performances s'approchent désormais de la *baseline*, et sont même meilleures sur la macro-F1 ; cela s'explique par le fait que le jeu artificiel $\mathcal{G}$ est bien plus équilibré que $\mathcal{T}$ et donc plus performant sur les classes minoritaires du jeu de test. Dans le scénario de complément, on observe une amélioration significative par rapport à la *baseline*, notamment avec la stratégie séquentielle.

5.3 Différences entre les classifieurs

Au-delà des mesures de performances globales, il peut être intéressant de vérifier si le classifieur entraîné sur les données artificielles permet de prendre les mêmes décisions qu'un classifieur entraîné sur $\mathcal{T}$. Pour ce faire, on peut regarder la proportion d'exemples (du jeu de test) pour lesquels la décision entre BERT* / $\mathcal{T}$ et BERT* / $\mathcal{G}^f$ diffère. Pour les données CLS-FR, les classifieurs s'accordent sur une grande majorité d'exemples. La figure 3 présente la matrice de confusion des classifieurs FlauBERT / $\mathcal{T}$ et FlauBERT / $\mathcal{G}^f$ sur les données CLS-FR.

À partir de cette matrice de confusion, on peut remarquer que les classifieurs s'accordent effectivement sur la majorité des exemples. Les cas de désaccords sont proportionnellement plus importants sur les faux positifs et faux négatifs, mais même pour ces catégories, on constate tout de même beaucoup d'erreurs communes (resp. 42 et 77 exemples pour les faux positifs et faux négatifs). Les classifieurs ont donc non seulement des performances comparables, mais des comportements très similaires dans le détail puisqu'ils donnent la même classe sur la plupart des exemples.

6 Expérimentations : approches sac-de-mots

Nous testons également des classifieurs reposant sur des représentations sac-de-mots ; nous ne présentons que les résultats de la régression logistique (LR) qui a donné les meilleurs résultats. En général moins performants que les approches à base de *transformers*, ces classifieurs permettent cependant une meilleure explicabilité (Miller, 2018; Carvalho *et al.*, 2019, pour une définition et une caractérisation des méthodes d'apprentissage), par exemple en examinant les poids de régression associés aux mots. Ils sont aussi moins coûteux à entraîner.

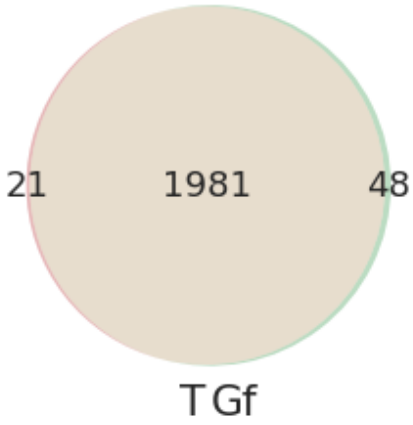
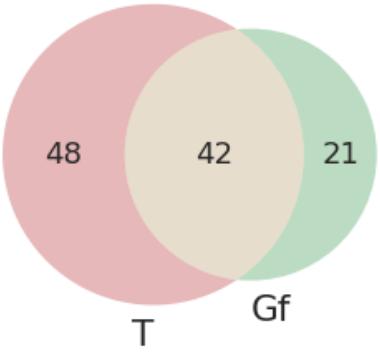
	Positif (vérité-terrain)	Négatif (vérité terrain)
Prédit positif		
Prédit négatif		

TABLE 3 : Matrice de confusion des modèles FlauBERT / $\mathcal{T}$ et FlauBERT / $\mathcal{G}^f$ sur les données CLS-FR. Les diagrammes de Venn font apparaître les proportions d'exemples en commun pour chacune des catégories.

6.1 Premiers résultats

L'implémentation utilisée est celle de scikit-learn (Pedregosa *et al.*, 2011), les textes sont vectorisés avec la pondération TF-IDF et normalisés L2, et les paramètres de la LR sont ceux par défaut sauf pour les suivants : stratégie multiclasse *one-vs.-rest*, nombre d'itérations = 2500. Les résultats des mêmes scénarios que précédemment sont présentés pour les tâches MediaEval et CLS-FR dans les tableaux 4 et 5.

Pour ce type de classifieur, l'intérêt des données générées apparaît pour les deux scénarios et sur nos deux jeux de données. Dans le cas de la substitution, les classifieurs sont légèrement meilleurs que ceux entraînés sur les données originales. Cela démontre l'intérêt de disposer d'une plus grande quantité de données permettant de capturer des variantes de forme dans les textes (synonymes, paraphrases...) que les représentations sac-de-mots ne peuvent sinon pas capturer aussi facilement que les représentations par plongements (pré-entraînées). Dans le scénario où les données sont utilisées en complément, l'augmentation de performances est encore plus marquée et s'approche ainsi de la *baseline* neuronale, tout en ayant les avantages d'un classifieur jugé plus interprétable.

modèle	micro-F1	macro-F1	MCC
LR / $\mathcal{T}$	72,68	56,35	42,22
LR / $\mathcal{G}^f$	74,00	59,18	44,39
LR / $\mathcal{T} + \mathcal{G}^f$	75,46	59,64	45,83

TABLE 4 : Performances (%) de l'approche LR/sac-de-mots sur les données MediaEval selon les scénarios d'usage des données artificielles filtrées : sans, par substitution, en complément.

modèle	micro-F1	macro-F1	MCC
LR / $\mathcal{T}$	84,77	84,70	69,48
LR / $\mathcal{G}^f$	87,16	87,14	74,27
LR / $\mathcal{T} + \mathcal{G}^f$	88,36	88,34	76,69

TABLE 5 : Performances (%) de l'approche LR/sac-de-mots sur les données CLS-FR selon les scénarios d'usage des données artificielles filtrées : sans, par substitution, en complément.

6.2 Effet de la qualité des donnés générées

On peut se demander quelle est l'influence de la qualité des données générées (mêmes filtrées) sur les résultats du classifieur final (cf. section 5.2). Pour étudier cela, nous injectons du bruit dans la classification pour simuler des filtrages faits avec des classifieurs de qualité variable. Cela est fait simplement en remplaçant, pour des exemples de $\mathcal{G}^f$ tirés au hasard, la classe prédite (par le générateur et par le classifieur filtrant) par une classe tirée aléatoirement. Le nombre d'exemples subissant ce traitement est calculé pour que la probabilité d'erreurs ainsi ajoutées fasse chuter le taux de bonnes de précision à 80 %, 70 %, etc. L'effet de ces filtrages sur les performances finales des stratégies complément et substitution sont présentés dans la figure 3 (données MediaEval) avec la régression logistique comme classifieur final.

Comme on peut le constater dans cette figure, ces résultats empiriques sur l'influence de la qualité du filtrage sont sans surprise. Dans le scénario substitution, la performance finale est fortement dépendante de la qualité du classifieur filtrant ; dans le cas présent, on atteint des performances équivalentes au jeu de données originales quand le taux de bonnes classifications du filtre dépasse 70 %. Dans le cas du scénario complément, le gain est sensible dès que le filtre a un taux de bonnes classifications supérieur au hasard.

Conclusion et perspectives

Dans un scénario où les données originales ne peuvent pas être distribuées, nous avons montré qu'il était possible de générer des données artificielles à des fins d'apprentissage supervisé. Pour les classifieurs état-de-l'art à base de *transformers*, cela dégrade les performances (par rapport à celles atteintes avec les données originales) mais dans une proportion contenue (-4 % de taux de bonnes classifications). En revanche, pour les classifieurs exploitant des représentations sac-de-mots, on constate une amélioration portée par la plus grande quantité de données d'apprentissage disponibles.

Dans un scénario où les données artificielles viennent en complément des données originales, nous avons montré que les classifieurs bénéficiaient de l'apport de données supplémentaires, y compris les

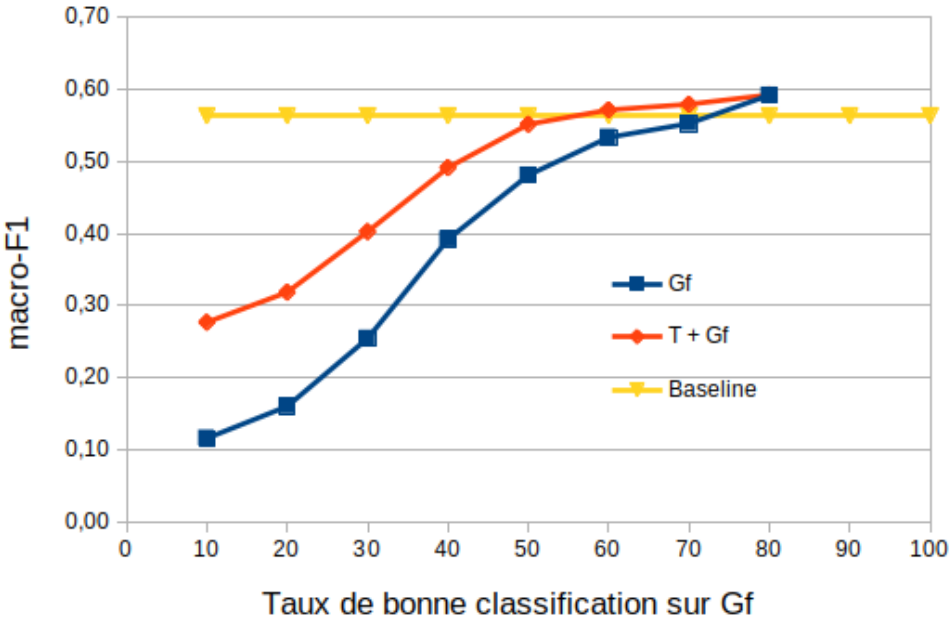


FIGURE 3 : Performances (macro-F1) en fonction de la qualité (taux de bonnes précisions en %) du classifieur servant au filtrage des données générées automatiquement ; données MediaEval avec la classifieur régression logistique.

réseaux de neurones. Ce résultat est particulièrement positif pour les approches sac-de-mots, plus sensibles aux reformulations, qui bénéficient clairement de l'ajout de ces exemples artificiels. On a ainsi un bon compromis entre méthodes rapides à entraîner, plus facilement interprétables, tout en ayant des performances proches des réseaux de neurones.

Comme nous l'avons vu, ces résultats sont obtenus à condition que les données générées soient filtrées au préalable, ce qui semble contredire plusieurs travaux cités en sec. 2. Dans nos expériences, elles l'ont été automatiquement ; une correction manuelle des données (de leurs classes) est aussi envisageable et permettra de meilleurs résultats, mais avec un coût d'annotation supplémentaire. L'emploi de ces méthodes à d'autres données et d'autres tâches de TAL que la classification de texte reste une piste prometteuse. Parmi ces tâches de TAL, celles reposant sur de l'étiquetage de mots posent des problèmes différents et nécessitent des solutions adaptées. Dans le futur, il serait intéressant de vérifier la consistance de nos résultats selon d'autres approches de génération (Kumar *et al.*, 2020). Il semble également intéressant d'étudier plus profondément l'impact de la qualité du classifieur servant à filtrer les données artificielles. De plus, l'intégration de cette étape de filtrage comme une contrainte lors de la génération des exemples artificiels est une piste prometteuse.

À des fins de réplicabilité, les scénarios d'entraînement présentées dans cet article sont accessibles en ligne pour les données MediaEval et CLS-FR. La génération des exemples repose sur `https://github.com/minimaxir/gpt-2-simple`. Les données sont accessibles auprès de leurs producteurs (voir section 4).

Références

AMIN-NEJAD A., IVE J. & VELUPILLAI S. (2020). Exploring transformer text generation for

medical dataset augmentation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, p. 4699–4708, Marseille, France : 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. & RAFFEL C. (2020). Extracting training data from large language models. *arXiv*.

CARVALHO D. V., PEREIRA E. M. & CARDOSO J. S. (2019). Machine learning interpretability : A survey on methods and metrics. *Electronics*, **8**(8). DOI : 10.3390/electronics8080832.

CHEEMA G. S., HAKIMOV S. & 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.

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. HAL : hal-03116027.

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

DEVLIN J., CHANG M.-W., LEE K. & 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)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

JIANG H., HE P., CHEN W., LIU X., GAO J. & 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*, p. 2177–2190, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.197.

JUNGIEWICZ M. & SMYWINSKI-POHL A. (2019). Towards textual data augmentation for neural networks : synonyms and maximum loss. *Computer Science*, **20**(1). DOI : 10.7494/csci.2019.20.1.3023.

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)*, p. 452–457, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-2072.

KUMAR V., CHOUDHARY A. & CHO E. (2020). Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, p. 18–26, Suzhou, China : Association for Computational Linguistics.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBE B., BESACIER L. & SCHWAB D. (2020). FlauBERT : Unsupervised Language Model Pre-training for French. In *LREC*, Marseille, France. HAL : hal-02890258.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach.

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

MUELLER J. & THYAGARAJAN A. (2016). Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, p. 2786–2792 : AAAI Press.

PAPANIKOLAOU Y. & 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. & DUCHESNAY E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

POGORELOV K., SCHROEDER D. T., BURCHARD L., MOE J., BRENNER S., FILKUKOVA P. & LANGGUTH J. (2020). Fakenews : Corona virus and 5g conspiracy task at mediaeval 2020. In *MediaEval 2020 Workshop*.

PRETTENHOFER P. & STEIN B. (2010). Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 1118–1127, Uppsala, Sweden : Association for Computational Linguistics.

RADFORD A., WU J., CHILD R., LUAN D., AMODEI D. & SUTSKEVER I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*.

SCHROEDER D. T., POGORELOV K. & 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)*, p. 134–141 : IEEE.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U. & POLOSUKHIN I. (2017). Attention is all you need. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems 30*, p. 5998–6008. Curran Associates, Inc.

WEI J. & 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)*, p. 6382–6388, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1670.

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. & 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*, p. 38–45, Online : Association for Computational Linguistics.

WU X., LV S., ZANG L., HAN J. & HU S. (2019). Conditional bert contextual augmentation. In J. M. F. RODRIGUES, P. J. S. CARDOSO, J. MONTEIRO, R. LAM, V. V. KRZHIZHANOVSKAYA, M. H. LEES, J. J. DONGARRA & P. M. SLOOT, Éds., *Computational Science – ICCS 2019*, p. 84–95, Cham : Springer International Publishing.