"Honey, Tell Me What's Wrong", Explicabilité Globale des Modèles de TAL par la Génération Coopérative

Antoine Chaffin, Julien Delaunay

“Honey, Tell Me What’s Wrong”, Explicabilité Globale des Modèles de TAL par la Génération Coopérative

Antoine Chaffin[1, 2*] Julien Delaunay[3*]
(1) IRISA, 263 Avenue du Général Leclerc, 35000 Rennes, France
(2) IMATAG, 13 Rue Dupont-des-Loges, 35000 Rennes, France
(3) Inria, 263 Avenue du Général Leclerc, 35000 Rennes, France
*Désigne une contribution équivalente
antoine.chaffin@irisa.fr, julien.delaunay@inria.fr

RÉSUMÉ

L’omniprésence de l’apprentissage automatique a mis en lumière l’importance des algorithmes d’explicabilité. Parmi ces algorithmes, les méthodes agnostiques au type de modèle génèrent des exemples artificiels en modifiant légèrement les données originales. Elles observent ensuite les changements de décision du modèle sur ces exemples artificiels. Cependant, de telles méthodes nécessitent d’avoir des exemples initiaux et fournissent des explications uniquement sur la décision pour ces derniers. Pour répondre à ces problématiques, nous proposons *Therapy*, la première méthode d’explicabilité modèle-agnostique pour les modèles de langue qui ne nécessite pas de données en entrée. Cette méthode génère des textes qui suivent la distribution apprise par le classifieur à expliquer grâce à la génération coopérative. Ne pas dépendre d’exemples initiaux permet, en plus d’être applicable lorsqu’aucune donnée n’est disponible (e.g, pour des raisons de confidentialité), de fournir des explications sur le fonctionnement global du modèle au lieu de plusieurs explications locales, offrant ainsi une vue d’ensemble du fonctionnement du modèle. Nos expériences montrent que, même sans données en entrée, Therapy fournit des informations instructives sur les caractéristiques des textes utilisées par le classifieur qui sont compétitives avec celles fournies par les méthodes utilisant des données.

ABSTRACT

“Honey, Tell Me What’s Wrong”, Global Explainability and Diagnosing of NLP Models through Cooperative Generation

The ubiquity of complex machine learning has raised the importance of model-agnostic explanation algorithms. These methods sample artificial instances by slightly perturbing target instances and observing the variations in the model decision. However, such methods require access to initial samples and only provide explanations of the decision for these. To tackle these problems, we propose Therapy, the first model-agnostic explanation method adapted to text which requires no input dataset. This method generates texts following the distribution learned by a classifier through cooperative generation. Not relying on initial samples, in addition to allowing use in cases where no data is available (e.g, for confidentiality reasons), provides global explanations of the model rather than multiple local ones, offering an overview of the model behavior. Our experiments show that although no input data is used to generate samples, Therapy provides insightful information about features used by the classifier that are competitive with the ones from methods relying on input samples.

MOTS-CLÉS : explicabilité, génération coopérative, traitement automatique des langues.

KEYWORDS: explainability, cooperative generation, natural language processing.

1 Introduction

L'émergence des modèles basés sur l'apprentissage automatique à permis leur adoption dans des domaines variés, allant de la simple recommandation à des secteurs critiques tels que la santé (Buch *et al.*, 2018; Esteva *et al.*, 2017; Karatza *et al.*, 2021) et le droit (Araszkiewicz *et al.*, 2022; Tagarelli & Simeri, 2022). Le besoin grandissant de précision induit une augmentation de la complexité de ces modèles, accentuant leur dénomination de boite noire. Ce manque de transparence limite (et empêche) leur déploiement dans différents domaines, en raison par exemple de l'augmentation significative de modèles souffrant de biais. Entre autres, certains *chatbots* ont été déployés bien que comportant des biais liés aux minorités religieuses (Abid *et al.*, 2021) ou de genre (Lucy & Bamman, 2021) et expliquer leur fonctionnement reste un problème ouvert.

Parmi les méthodes proposées pour répondre à ces problèmes, celles qui sont modèle-agnostiques sont préférées car applicables à tout type de modèle d'apprentissage automatique. Parmi ces dernières, les explications locales ont obtenu un fort succès car elles fournissent un bon compromis entre précision et utilité des explications. Les explications locales sont générées en perturbant une entrée afin de construire un voisinage autour de cette entrée et en étudiant comment le modèle réagit à ces petites différences. Cela permet de mettre en évidence les caractéristiques importantes pour le modèle et de fournir des éléments d'explication sur la décision du modèle pour cette entrée (e.g., les mots les plus importants de chaque classe). Selon une étude récente sur les tendances dans le domaine de l'explicabilité (Jacovi, 2023), les méthodes d'explication locale agnostique au modèle, telles que LIME (Ribeiro *et al.*, 2016) sont les plus utilisées.

Cependant, l'explication d'un modèle à partir d'un exemple précis présente trois défauts. Premièrement, il faut évidemment disposer d'entrées à expliquer ; ce qui peut être impossible pour des raisons de confidentialité ou de respect de la vie privée par exemple. Deuxièmement, choisir des entrées qui sont représentatives du modèle et/ou des données sur lesquelles le modèle sera utilisé est une tâche complexe. Troisièmement, cela donne une explication de la décision **pour cette entrée** et pour cette entrée uniquement. Ainsi, cela ne fournit que des informations très locales sur le comportement du modèle, qui ne représentent qu'une petite partie du domaine d'entrée du modèle. Pour cette raison, certaines méthodes d'explications ont proposé d'agréger plusieurs explications locales afin de fournir une explication plus globale. Néanmoins, ces explications restent fortement liées aux données d'entrée et ne fournissent des informations que sur le voisinage de ces échantillons. Ces méthodes nécessitent donc que les différentes données d'entrée couvrent la partie de l'espace la plus large possible.

Avec l'objectif de supprimer cette dépendance aux données d'entrée et de générer une explication globale du modèle, nous proposons ***Therapy***, une méthode qui utilise la génération coopérative (Holtzman *et al.*, 2018; Scialom *et al.*, 2020; Chen *et al.*, 2020; Bakhtin *et al.*, 2021; Chaffin *et al.*, 2022) pour générer des textes suivant la distribution d'un classifieur. La distribution des textes générés permet ensuite d'étudier les caractéristiques importantes du modèle, fournissant ainsi une explication globale agnostique au modèle. Dans ce papier, nous présentons d'abord les travaux connexes dans la Section 2, la Section 3 introduit ensuite des notions sur la génération de textes et plus particulièrement la génération coopérative. La Section 4 détaille quant à elle le fonctionnement de la méthode tandis que la Section 5 présente les expériences réalisées pour comparer les performances de Therapy avec les méthodes d'explication usuelles.

2 Travaux connexes

Générer une explication pour des données textuelles est une tâche ardue qui nécessite de prendre en compte à la fois la sémantique du texte mais également le domaine de la tâche (e.g. analyse de sentiment, détection de spam). De plus, il est fréquent de devoir évaluer un modèle déjà déployé pour des raisons d'équité ou de détection de biais par exemple mais que les données ne soient plus accessibles pour des raisons de sécurité ou de confidentialité. Afin de résoudre ce problème, les chercheurs se sont concentrés sur les méthodes d'explications post-hoc (Jacovi, 2023). Suivant la catégorisation de Bodria *et al.* (2021), nous différencions les explications sous forme d'exemples de celles par attribution de caractéristiques.

2.1 Explications sous forme d'exemples

Les méthodes d'explications sous forme d'exemples tirent leur racine des sciences sociales (Miller, 2019) et montrent des contrefactuels qui indiquent le changement minimum requis pour modifier une prédiction, ou des prototypes, des exemples représentatif d'une classe. Les méthodes contrefactuelles perturbent le document cible jusqu'à trouver le document le plus proche qui soit classé différemment par le modèle complexe. À l'inverse, les méthodes qui génèrent des prototypes sélectionnent les instances qui représentent le plus une classe cible. Parmi les méthodes *post-hoc*, certaines proposent des codes de contrôle permettant de surveiller la perturbation du texte en entrée, tandis que d'autres entraînent des mécanismes complexes pour générer des phrases réalistes en perturbant une instance dans un espace latent. Polyjuice (Wu *et al.*, 2021) et GYC (Madaan *et al.*, 2021) feront partie de la première catégorie et proposent des codes de contrôle allant du changement de sentiment ou de temps jusqu'à l'ajout ou le retrait de mots. xSPELLS (S. Punla *et al.*, 2022) et CounterfactualGAN (Robeer *et al.*, 2021), sont des méthodes qui entraînent respectivement un auto-encodeur variationnel et un réseau adversarial génératif pour convertir les textes en entrée dans un espace latent. Des modifications y sont ensuite réalisées afin de retourner des contrefactuels réalistes proches de l'exemple original.

2.2 Explications par attribution

Parmi les méthodes post-hoc, les explications par attribution associent un poids aux termes en entrée afin d'indiquer leur impact positif ou négatif sur la prédiction finale. Les méthodes telles que SHAP (Lundberg & Lee, 2017), LIME (Ribeiro *et al.*, 2016) et ses variantes (Gaudel *et al.*, 2022; Shankaranarayana & Runje, 2019; Zafar & Khan, 2019; Visani *et al.*, 2020; ElShawi *et al.*, 2019; Bramhall *et al.*, 2020) restent les plus utilisées pour générer des explications (Jacovi, 2023). Les explications sont dites locales puisque ces méthodes perturbent un document en entrée en modifiant légèrement les valeurs et en observant le comportement du modèle complexe dans cette localité. Pour les données texte, LIME masque aléatoirement les mots du document en entrée afin d'en créer diverses variations et entraîne un modèle linéaire sur ces exemples. Les coefficients du modèle linéaire, associés aux différents mots, sont ensuite retournés et utilisés comme explication. Bien que la majorité des études (Arrieta *et al.*, 2020; Bodria *et al.*, 2021) fasse une différence entre les explications locales et globales, LIME introduit LIME-SP (pour sélection sous modulaire), une méthode qui génère une explication globale à partir de n explications locales. Ces n explications sont choisies parmi un plus grand ensemble afin de couvrir le plus possible l'espace d'entrée tout en réduisant la redondance.

3 Génération de textes

3.1 Génération coopérative

Les modèles de langue génératifs (LM) tels que la famille des GPT (Radford *et al.*, 2018, 2019; Brown *et al.*, 2020) apprennent la distribution de probabilité de séquences de symboles $x_1, x_2, \cdots, x_T$ (souvent appelés *tokens*) appartenant à des séquences de taille variable T sur un vocabulaire $\mathcal{V}$. La probabilité d'une séquence x (aussi appelée vraisemblance) est définie comme la probabilité jointe de chacun de ses tokens. Cette probabilité peut être factorisée en utilisant la formule des probabilités composées : $p(x_{1:T}) = \prod_{t=1}^{T} p(x_t \mid x_{1:t-1})$. Le LM est entraîné à produire une distribution de probabilité sur le dictionnaire pour le prochain token sachant ceux en entrée, i.e, $p_\theta(x_t \mid x_{1:t-1})$ à un pas de temps donné t. Cela permet d'obtenir un modèle de langue auto-régressif qui génère des séquences itérativement. Le modèle utilise les distributions apprises pour émettre un token x_t et l'ajouter au contexte $x_{1:t-1}$, qui sera utilisé pour la prochaine itération. Le processus de génération –ou décodage– démarre généralement en utilisant une petite séquence initiale appelée l'amorce. Les grands modèles de langue apprennent une très bonne approximation de la distribution de leurs données d'apprentissage, ce qui permet de générer des textes plausibles en maximisant la vraisemblance du modèle $p(x)$. Cependant, cette approche offre très peu de contrôle sur le texte généré à l'exception de l'amorce.

Les approches de génération coopérative (Holtzman *et al.*, 2018; Scialom *et al.*, 2020; Chen *et al.*, 2020; Bakhtin *et al.*, 2021; Chaffin *et al.*, 2022) permettent de résoudre ce problème en utilisant des modèles discriminatifs pour guider le modèle de langue durant la génération. Elles utilisent l'information du modèle externe pour guider le modèle de langue afin de générer des textes qui possèdent une propriété reconnue par le modèle discriminatif. Dans le cas où le modèle est un classifieur qui apprend à prédire la probabilité $D(c \mid x)$ qu'une séquence x d'appartenir à la classe c, le but est de générer un texte qui maximise la probabilité d'appartenir à la classe cible. En raison de la taille de l'espace de toutes les séquences possibles ($|\mathcal{V}|^n$ pour une séquence de longueur n) ; il est infaisable d'évaluer $D(c \mid x)$ pour toutes les séquences possibles. Ainsi, les méthodes coopératives utilisent la distribution du modèle de langue génératif pour restreindre l'exploration aux séquences plausibles uniquement. De cette manière, la séquence produite maximise $p(x) * D(c \mid x) \propto p(x \mid c)$, résultant en une séquence qui est à la fois bien écrite et qui appartient à la classe cible.

3.2 Décodage guidé par *Monte Carlo Tree Seach*

Parmi ces approches coopératives, celles qui utilisent le *Monte Carlo Tree Search* (MCTS) pour guider le décodage des modèles de langue ont obtenu de très bons résultats (Scialom *et al.*, 2021a; Chaffin *et al.*, 2022; Leblond *et al.*, 2021; Lamprier *et al.*, 2022). Le MCTS est un algorithme itératif qui cherche une solution dans un arbre trop grand pour être parcouru de façon exhaustive. Il est adapté à la génération de texte car l'espace de recherche créé durant le décodage correspond à un arbre : la racine correspond à l'amorce et les enfants d'un nœud correspondent aux séquences construites en augmentant le préfixe de leur parent d'un token supplémentaire. La boucle du MCTS est composée de 4 étapes : sélection, expansion, simulation et rétro-propagation.

1. **Sélection.** Une exploration de la racine de l'arbre à une feuille non explorée. Le chemin vers la feuille est défini en sélectionnant, à chaque nœud, l'enfant qui maximise la *Polynomial Upper*

Confidence Trees (PUCT) (Rosin, 2011; Silver *et al.*, 2017) qui est définie, pour un nœud i par :

$$PUCT(i) = \frac{s_i}{n_i} + c_{puct}\, p(x_i \mid x_{1:t-1}) \frac{\sqrt{N_i}}{1+n_i} \quad (1)$$

avec n_i le nombre de simulations jouées après le nœud i, s_i le score agrégé du nœud, N_i le nombre de simulations jouées après son parent et c_{puct} une constante qui définit le compromis entre exploitation (se focaliser sur des nœuds avec des bons scores) et exploration (explorer des nœuds prometteurs).

2. **Expansion.** La création des enfants du nœud sélectionné s'il n'est pas terminal (i.e, correspondant au token de fin de séquence).
3. **Simulation (roll-out).** Le tirage de tokens additionnels (en utilisant la distribution du modèle de langue) jusqu'à un nœud terminal.
4. **Rétro-propagation.** L'évaluation de la séquence x associée au nœud terminal et l'agrégation de son score à tous les parents jusqu'à la racine. Pour guider la génération vers des textes qui appartiennent à une classe donnée, le score d'une séquence x associée à une feuille peut être défini par la probabilité $D(c \mid x)$ donnée par le classifieur. Différentes méthodes d'agrégation peuvent être utilisées, Chaffin *et al.* (2022) calcule la moyenne du score actuel et de celui du nœud terminal alors que (Scialom *et al.*, 2021b; Lamprier *et al.*, 2022) prennent le maximum des deux.

Cette boucle est répétée un certain nombre de fois, puis le token à ajouter à l'amorce est choisi grâce à l'arbre construit. Parmi les nœuds enfants de la racine, deux choix sont possibles : soit celui ayant été sélectionné le plus de fois, soit celui ayant un score agrégé le plus élevé. Comme nous souhaitons obtenir des séquences aussi stéréotypiques des classes du modèle discriminatif que possible, nous choisissons celui ayant le score le plus élevé. Ce nœud devient ensuite la nouvelle racine, et le processus est répété jusqu'à produire la séquence finale.

Contrairement aux méthodes traditionnelles qui décodent de gauche à droite et peuvent manquer des séquences qui deviennent meilleures après quelques étapes de génération ou se retrouver bloquer dans des séquences sous optimales, le MCTS brise le décodage myope en définissant le score d'un token sur la base des continuations possibles de la séquence. En plus d'être *plug-and-play*, c'est à dire que n'importe quel modèle de langue génératif (auto-régressif) peut être guidé durant le décodage par n'importe quel classifieur, cette approche a obtenu des résultats à l'état de l'art dans la tâche de génération contrainte, qui consiste à générer des textes qui maximisent $D(c \mid x)$ tout en maintenant une qualité d'écriture élevée.

4 Méthode

Dans cet article, nous présentons ***Therapy***, une méthode d'explication agnostique au modèle qui n'utilise pas de données en entrée. Therapy utilise un modèle de langue guidé par le classifieur à expliquer pour générer des textes représentatifs des classes apprises par le classifieur. Pour ce faire, Therapy extrait les mots importants pour le classifieur en l'utilisant pour guider le modèle de langue via la génération coopérative. Puisque les textes générés coopérativement suivent la distribution $p(x)D(c \mid x)$, leur distribution peut ensuite être utilisée pour étudier le classifieur D : les mots avec des fréquences élevées sont susceptibles d'être importants pour le classifieur. Ainsi, Therapy entraîne un modèle de régression logistique sur les représentations tf-idf des textes générés et retourne les

coefficients les plus importants et leurs termes associés comme explication. Comme $p(x)$ est le même pour toutes les classes, l'utilisation des tf-idf sur le corpus entier (i.e, les textes générés pour toutes les classes) filtre les mots qui sont fréquents uniquement à cause de $p(x)$ ou pour plusieurs classes. Les termes résultants sont donc dûs à la partie de la distribution venant du classifieur. L'entraînement d'un modèle de régression logistique sur les tf-idf permet d'extraire les termes les plus importants et d'étudier leur importance relative pour chaque classe. Therapy offre donc le niveau d'explicabilité d'une régression logistique basée sur des n-grams. Enfin, grâce à l'aspect plug-and-play de la génération guidée par MCTS, Therapy est une méthode agnostique au modèle qui peut expliquer tout type de modèle via n'importe quel modèle de langue et retourner une explication du fonctionnement global du classifieur.

En substance, la méthode est similaire à l'utilisation de LIME combinée à un modèle de langue qui remplace des tokens masqués lorsque le nombre de tokens remplacés tend vers l'infini mais avec deux avantages. Premièrement, la méthode ne repose pas sur des exemples en entrée mais génère ces exemples à partir de rien en utilisant le modèle de langue auto-régressif. Ceci est particulièrement utile pour les cas où les données ne peuvent pas être partagées pour des raisons de confidentialité (Amin-Nejad *et al.*, 2020). De plus, au lieu d'explorer le voisinage de ces exemples (et donc de conditionner les explications au contexte de ces exemples), le domaine d'exploration est défini par le modèle de langue. Ce modèle de langue peut être générique ou spécifique à un domaine sur lequel sera utilisé le classifieur pour s'assurer qu'il fonctionne correctement sur ce type de données précis.

Deuxièmement, la méthode ne génère pas **avant** de classifier mais utilise le classifieur **durant** la génération. Ainsi, au lieu de générer des textes "au hasard" et espérer que les caractéristiques importantes apparaissent, la méthode interroge explicitement le modèle pour des caractéristiques discriminantes via la maximisation de $D(c \mid x)$. Cela rend la méthode plus efficace et réduit la probabilité de générer (a) des mots rares mais qui ne sont pas importants pour le modèle, et (b) des textes "entre-deux" qui possèdent les caractéristiques de plusieurs classes et peuvent être perturbants. Par ailleurs, notre méthode s'appuie directement sur la distribution apprise par le classifieur étudié pour guider la génération, contrairement aux méthodes comme Polyjuice et GYC, qui en plus d'utiliser des exemples en entrée, s'appuient sur une distribution apprise par le modèle de langue pour biaiser la génération vers certaines caractéristiques (via les codes de contrôle).

Nous avons décidé d'appeler cette approche Therapy puisque nous associons son fonctionnement à celui d'un thérapeute (le LM). Ce thérapeute interroge son patient (le classifieur) afin de comprendre son comportement et éventuellement découvrir des comportements pathologiques (des biais).

5 Expériences

Dans cette section, nous présentons d'abord les détails expérimentaux de l'évaluation de Therapy (Section 5.1). Nous évaluons ensuite Therapy au travers de 3 expériences. La première expérience (Section 5.2), mesure la corrélation de Spearman des explications avec les poids de la boite transparente et étudie l'impact de la quantité de textes générés sur la qualité de l'explication retournée par le modèle linéaire. Nous comparons ensuite la capacité de Therapy à identifier les mots les plus importants pour le classifieur avec celles de LIME et SHAP dans la Section 5.3. Enfin, nous observons si les termes retournées par les différentes approches sont suffisants pour modifier la prédiction du classifieur. Le code de Therapy ainsi que celui des expériences sont disponibles publiquement.

5.1 Configuration expérimentale

Explication de boite transparente. Puisqu'il n'y a pas de vérité terrain disponible pouvant être utilisée comme objectif pour les méthodes d'explication évaluées, nous utilisons un modèle boite transparente. Un modèle est dit boite transparente lorsque ses paramètres utilisés pour faire une prédiction sont connus, on dit aussi que le modèle est explicable par conception. Tout au long de nos expérimentations, nous utilisons comme modèle boite transparente une régression logistique implémentée en utilisant sklearn (Pedregosa *et al.*, 2011). Les poids de ce modèle représentent le score d'importance associé à chacun des termes du vocabulaire.

Implémentation de Therapy. Lors de l'évaluation de la méthode proposée, nous utilisons l'implémentation de PPL-MCTS (Chaffin *et al.*, 2022) disponible sur GitHub. L'utilisation de la boite transparente dans PPL-MCTS se fait simplement en définissant la fonction qui prend en entrée une séquence et retourne son score. Le choix du modèle de langue génératif définit le domaine sur lequel nous voulons des explications pour le comportement du classifieur. Pour montrer que la méthode fonctionne bien avec un modèle de langue général (sans domaine particulier), nous utilisons OPT-125m (Zhang *et al.*, 2022). Une régression logistique est ensuite apprise sur la représentation tf-idf des textes générés et les coefficients de la régression logistique sont finalement retournés comme scores d'importance des différents tokens.

Jeu de données utilisés. Toutes les expériences ont été réalisées sur deux jeux de données différents issus de (Zhang *et al.*, 2015). Le premier, amazon_polarity : est un jeu de données de classification binaire composé de commentaires de produits Amazon labelisés comme positifs ou négatifs. Les textes qui le composent sont relativement courts et possèdent des champs lexicaux très caricaturaux. Le second, AG_news est un jeu de classification thématique à 4 classes (`world`, `sport`, `business` et `sci/tech`). Les textes de ce jeu sont plus longs et plus variés, mais ils possèdent différents indicateurs caractéristiques liés au fait qu'ils sont extraits d'articles de presse en ligne. Des exemples de textes générés par Therapy pour chacun des jeux de données ainsi que les premiers top-mots retournés sont disponibles en Annexe A.

Méthodes comparées. Nous comparons dans les Section 5.3 et 5.4, les résultats de Therapy avec les deux approches par attribution de caractéristiques les plus répandues : LIME (Ribeiro *et al.*, 2016) et SHAP (Lundberg & Lee, 2017) en utilisant les implémentations publiques. La différence principale entre LIME et SHAP est que la première génère un échantillon en perturbant une instance puis entraîne localement un modèle de régression linéaire tandis que la seconde utilise la théorie des jeux pour calculer l'importance de chaque élément. Nous avons utilisé les versions globales de ces méthodes sur 500 textes du jeu de test. Pour SHAP, nous avons décidé de garder les 10000 mots les plus importants pour chacun des jeux de données tandis que pour LIME, nous avons retourné les 500 explications locales avec les 35 mots les plus importants et regroupé les couples mots-valeurs dans un dictionnaire composé de 4592 mots pour amazon_polarity et 5770 pour ag_news.

5.2 Corrélation de Spearman

Une bonne explication de boite transparente est une liste de mots qui contient à la fois ses mots importants (i.e, a une bonne couverture) et les relie à des scores d'importance similaires. Ainsi, nous calculons la corrélation de Spearman entre les top mots de la boite transparente (ceux ayant un poids > 1) et leur score attribué par la méthode d'explication. La corrélation de Spearman à été préférée à celle de Pearson car les scores retournés par LIME et SHAP peuvent être très différents des poids de

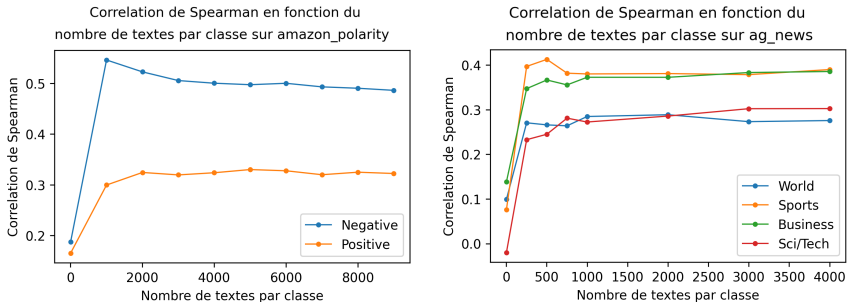

FIGURE 1 – Corrélation de Spearman en fonction du nombre de textes générés par classe pour amazon_polarity et ag_news

la boite transparente et une corrélation de rang donne donc une comparaison plus juste.

5.2.1 Influence du nombre de textes générés

Un des paramètres critiques de la méthode proposée est le nombre de textes à générer. En effet, un nombre élevé de tokens permet une plus grande couverture mais nécessite plus de calculs. Nous reportons la corrélation de Spearman en fonction du nombre de textes générées par classe dans la Figure 1. Nous pouvons voir que la corrélation augmente rapidement jusqu'à atteindre un plateau, ce qui signifie qu'une petite quantité de textes permet d'obtenir un bon aperçu du comportement du modèle et que la méthode ne nécessite pas énormément de calculs pour fonctionner. Pour la suite des expérimentations, nous fixons le nombre de textes générés par Therapy à 3000 par classe.

Jeu de données	AMAZON_POLARITY		AG_NEWS			
Classe	Positive	Negative	World	Sports	Business	Sci/Tech
LIME	0.64 (5.0e-7)	0.44 (1.5e-3)	0.09 (0.53)	0.16 (0.27)	0.20 (0.16)	0.19 (0.19)
LIME-other	0.21 (0.14)	0.18 (0.21)	-0.03 (0.85)	0.23 (0.12)	0.09 (0.52)	0.29 (0.04)
SHAP	0.71 (7.6e-9)	0.76 (1.6e-10)	0.47 (6.2e-4)	0.62 (1.7e-06)	0.53 (8.0e-5)	0.61 (2.4e-6)
SHAP-other	0.02 (0.87)	0.26 (0.06)	-0.05 (0.71)	0.04 (0.77)	0.15 (0.31)	0.12 (0.41)
Therapy	0.49 (3.3e-08)	0.31 (1.0e-4)	0.27 (1.6e-07)	0.37 (4.0e-12)	0.38 (5.6e-13)	0.3 (8.9e-09)

TABLE 1 – Corrélation de Spearman (p-valeur) entre les top mots de la boite transparente et les différentes méthodes d'explications. Les résultats sont donnés par classe et par jeu de données. Le suffixe 'other' indique que les explications sont générées à partir de l'autre jeu de données.

5.2.2 Comparaison avec les autres méthodes

Les corrélations de Spearman pour chacune des approches évaluées sont disponibles dans la Table 1. Les résultats obtenus par Therapy sont meilleurs que ceux de LIME sur ag_news mais moins bon sur amazon_polarity tandis que SHAP donne des résultats meilleurs que les autres méthodes sur les deux jeux de données. Ces résultats sont bons pour Therapy puisque LIME et SHAP génèrent des explications à partir du jeu de données de test, garantissant ainsi que les caractéristiques cibles se trouvent dans les exemples en entrée. Or, lorsque cette hypothèse ne tient plus, par exemple en

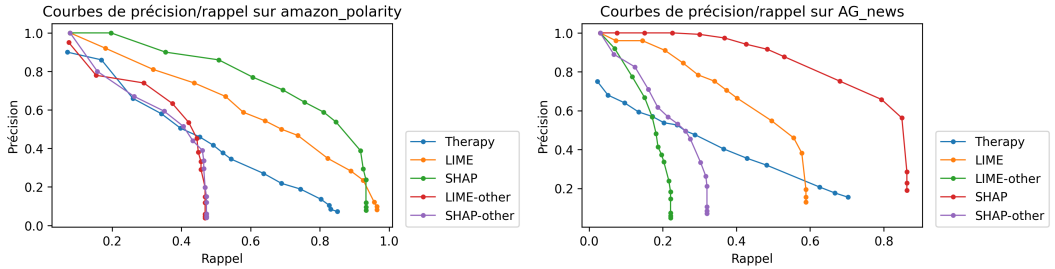

FIGURE 2 – Courbes de précision/rappel sur les top mots de la boite transparente (régression logistique) pour les différentes méthodes d'explications

utilisant comme données d'entrée le jeu de test de l'autre jeu de données (dénoté par *other* dans nos résultats), les méthodes ne parviennent plus à trouver les caractéristiques importantes et les corrélations chutent drastiquement, bien en dessous des résultats obtenus par Therapy.

5.3 Précision et rappel

Outre l'attribution correcte de scores aux caractéristiques importantes du modèle, il est également nécessaire que l'explication fournisse une sortie informative en pratique. Il faut donc s'assurer que les mots retournés dans l'explication (i.e, les mots avec les scores les plus élevés) soient effectivement des mots importants pour le modèle étudié et que ses mots les plus importants soient trouvés. Ainsi, la Figure 2 montre pour différents nombres de mots importants retournés, les valeurs de précision et de rappel moyennées sur toutes les classes. Le nombre k de mots retournés varie de 10 à 1500. La précision est obtenue en mesurant la proportion de mots retournés dans l'explication qui appartiennent au top mots de la boite transparente tandis que le rappel est la proportion des top mots de la boite transparente qui sont présents dans l'explication.

On observe que Therapy obtient de moins bons résultats que LIME (bien qu'obtenant un meilleur rappel sur ag_news) tandis que SHAP est meilleur que les deux méthodes sur les deux jeux de données. À nouveau, lorsque les données ne contiennent plus nécessairement les caractéristiques importantes pour le modèle (-other), les résultats s'écroulent et Therapy surpasse les deux approches. Cette limitation est visible par le plateau au niveau des scores de rappels pour ces méthodes : elles trouvent effectivement bien les caractéristiques importantes **présentes dans les données**, mais sont limitées à celles-ci uniquement, fixant la limite supérieure des caractéristiques pouvant être trouvées. En pratique, les biais contenus dans le modèle peuvent être suffisamment subtils pour ne pas être présents dans le jeu de données à disposition, auquel cas, LIME et SHAP ne peuvent pas les détecter. Therapy, au contraire, obtient de bons résultats en utilisant le même LM générique pour les deux jeux de données, sans utiliser d'apriori. La méthode permet donc d'obtenir un très bon aperçu du comportement du modèle lorsqu'aucune donnée, ou plus largement, lorsqu'aucune donnée représentative des caractéristiques importantes du modèle n'est disponible. Dans ce dernier cas de figure, Therapy permet d'offrir une recherche plus exhaustive que celles se basant sur des textes existants, obtenant un score de rappel important.

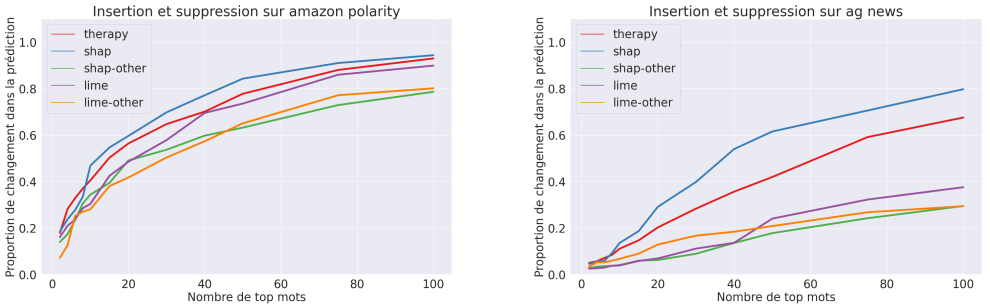

FIGURE 3 – Proportion de textes dont la classification change en fonction du nombre de mots importants utilisés pour effectuer le remplacement.

5.4 Insertion/suppression de mots importants

Une manière de valider l'exactitude des explications est de retirer les mots que l'explication indique comme étant importants pour une classe et d'observer comment les prédictions du modèle évoluent. L'intuition derrière cette approche est que supprimer la "cause" forcera le modèle à changer sa décision (Petsiuk *et al.*, 2018). De la même manière, rajouter un mot indiqué par l'explication comme étant important pour une autre classe devrait réduire la confiance du modèle. Ainsi, nous calculons une métrique d'insertion/suppression en mesurant la proportion de textes dont la classification par la boite transparente est modifiée lorsque l'on retire un mot étant indiqué comme important pour cette classe et que nous le remplaçons par un mot important pour une autre classe. La Figure 3 montre les résultats agrégés sur chacune des classes des deux jeux de données pour Therapy, LIME, SHAP et leurs versions utilisant l'autre jeu de données (-other). On observe que SHAP et Therapy sont les plus efficaces et obtiennent des résultats similaires, parvenant à modifier la prédiction du classifieur plus souvent que LIME à mesure que le nombre de mots échangés augmente. On remarque également que sur le jeu de données ag_news, où les mots importants sont moins communs que pour amazon_polarity, Therapy est capable de modifier la prédiction près de 70% du temps, tandis que SHAP et LIME n'arrivent à modifier la prédiction que dans 30% de cas en utilisant l'autre jeu de données. Ce qui montre encore une fois que ces méthodes nécessitent des données très spécifiques alors que Therapy est capable de trouver les mots importants pour chaque classe sans avoir accès à aucune donnée et sans utiliser d'apriori sur le modèle.

6 Conclusion

Les méthodes d'explicabilité usuelles s'appuient fortement sur des données en entrée, qui ne sont pas forcément disponibles et peuvent ne pas contenir les biais et caractéristiques importantes du modèle. Nous proposons Therapy, une méthode qui emploie la génération de texte coopérative afin de générer des données synthétiques qui suivent la distribution apprise par le modèle étudié. Ainsi, la recherche est dirigée par un modèle de langue génératif pré-entraîné et permet une exploration plus large que celle restreinte au voisinage des données d'entrée. Cela permet de relaxer la plupart des contraintes et aprioris induits par les méthodes qui se basent sur des exemples. Dans le cas extrême

où des données très représentatives (comme le jeu de test d'un jeu de données) des caractéristiques importantes du modèle sont disponibles, Therapy obtient des résultats légèrement moins bon que la méthode état-de-l'art SHAP, tout en restant compétitif avec LIME. Cependant, si on considère des cas d'usage plus réalistes dans lesquels les caractéristiques importantes ne sont pas explicitement données en entrée de la méthode d'explication, alors les performances des autres méthodes s'effondrent et Therapy devient la meilleure approche. Ainsi, Therapy est un outil pertinent pour explorer le comportement d'un modèle lorsque la collecte des données sur lesquelles le modèle sera utilisée n'est pas possible ou très complexe.

Remerciements

Ces travaux sont financés par l'Agence Nationale de la Recherche (ANR) dans le cadre de la convention de subvention ANR-19-CE23-0019-01 et par le réseau TAILOR (EU Horizon 2020 programme d'innovation et de recherche avec la convention de subvention 952215).

Références

ABID A., FAROOQI M. & ZOU J. (2021). Persistent anti-muslim bias in large language models. In *AIES '21 : AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, p. 298–306 : ACM.

AMIN-NEJAD A., IVE J. & VELUPILLAI S. (2020). Exploring transformer text generation for medical dataset augmentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 4699–4708, Marseille, France : European Language Resources Association.

ARASZKIEWICZ M., BENCH-CAPON T., FRANCESCONI E., LAURITSEN M. & ROTOLO A. (2022). Thirty years of artificial intelligence and law : overviews. *Artificial Intelligence and Law*.

ARRIETA A. B., RODRÍGUEZ N. D., SER J. D., BENNETOT A., TABIK S., BARBADO A., GARCÍA S., GIL-LOPEZ S., MOLINA D., BENJAMINS R., CHATILA R. & HERRERA F. (2020). Explainable artificial intelligence (XAI) : concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion*, **58**, 82–115.

BAKHTIN A., DENG Y., GROSS S., OTT M., RANZATO M. & SZLAM A. (2021). Residual energy-based models for text. *Journal of Machine Learning Research*, **22**(40), 1–41.

BODRIA F., GIANNOTTI F., GUIDOTTI R., NARETTO F., PEDRESCHI D. & RINZIVILLO S. (2021). Benchmarking and survey of explanation methods for black box models. *CoRR*.

BRAMHALL S., HORN H., TIEU M. & LOHIA N. (2020). QLIME-A : Quadratic Local Interpretable Model-Agnostic Explanation Approach. *SMU Data Science Rev*, **3**.

BROWN T. B., MANN B., RYDER N., SUBBIAH M., KAPLAN J., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A., AGARWAL S., HERBERT-VOSS A., KRUEGER G., HENIGHAN T., CHILD R., RAMESH A., ZIEGLER D. M., WU J., WINTER C., HESSE C., CHEN M., SIGLER E., LITWIN M., GRAY S., CHESS B., CLARK J., BERNER C., MCCANDLISH S., RADFORD A., SUTSKEVER I. & AMODEI D. (2020). Language models are few-shot learners. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN & H. LIN, Éds., *Advances in Neural Information Processing Systems 33 : Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

BUCH V. H., AHMED I. & MARUTHAPPU M. (2018). Artificial intelligence in medicine : current trends and future possibilities. *Br. J. Gen. Pract.*, **68**(668), 143–144.

CHAFFIN A., CLAVEAU V. & KIJAK E. (2022). PPL-MCTS : constrained textual generation through discriminator-guided MCTS decoding. In M. CARPUAT, M. DE MARNEFFE & I. V. M. RUÍZ, Éds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, p. 2953–2967 : Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.215.

CHEN X., CAI P., JIN P., WANG H., DAI X. & CHEN J. (2020). Adding a filter based on the discriminator to improve unconditional text generation. *arXiv preprint arXiv :2004.02135*.

ELSHAWI R., SHERIF Y., AL-MALLAH M. & SAKR S. (2019). ILIME : Local and Global Interpretable Model-Agnostic Explainer of Black-Box Decision. In *ADBIS*.

ESTEVA A., KUPREL B., NOVOA R. A., KO J., SWETTER S. M., BLAU H. M. & THRUN S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *nature*, **542**(7639), 115–118.

GAUDEL R., GALÁRRAGA L., DELAUNAY J., ROZÉ L. & BHARGAVA V. (2022). s-LIME : Reconciling locality and fidelity in linear explanations. In *Advances in Intelligent Data Analysis XX - 20th International Symposium on Intelligent Data Analysis, IDA 2022, Rennes, France, April 20-22, 2022, Proceedings*, volume 13205 de *Lecture Notes in Computer Science*, p. 102–114 : Springer.

HOLTZMAN A., BUYS J., FORBES M., BOSSELUT A., GOLUB D. & CHOI Y. (2018). Learning to write with cooperative discriminators. In I. GUREVYCH & Y. MIYAO, Éds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1 : Long Papers*, p. 1638–1649 : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1152.

JACOVI A. (2023). Trends in explainable AI (XAI) literature. *CoRR*, **abs/2301.05433**. DOI : 10.48550/arXiv.2301.05433.

KARATZA P., DALAKLEIDI K., ATHANASIOU M. & NIKITA K. (2021). Interpretability methods of machine learning algorithms with applications in breast cancer diagnosis. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, p. 2310–2313. DOI : 10.1109/EMBC46164.2021.9630556.

LAMPRIER S., SCIALOM T., CHAFFIN A., CLAVEAU V., KIJAK E., STAIANO J. & PIWOWARSKI B. (2022). Generative cooperative networks for natural language generation. In K. CHAUDHURI, S. JEGELKA, L. SONG, C. SZEPESVÁRI, G. NIU & S. SABATO, Éds., *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 de *Proceedings of Machine Learning Research*, p. 11891–11905 : PMLR.

LEBLOND R., ALAYRAC J., SIFRE L., PISLAR M., LESPIAU J., ANTONOGLOU I., SIMONYAN K. & VINYALS O. (2021). Machine translation decoding beyond beam search. In M. MOENS, X. HUANG, L. SPECIA & S. W. YIH, Éds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, p. 8410–8434 : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.662.

LUCY L. & BAMMAN D. (2021). Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, p. 48–55, Virtual : Association for Computational Linguistics. DOI : 10.18653/v1/2021.nuse-1.5.

LUNDBERG S. M. & LEE S. (2017). A unified approach to interpreting model predictions. In I. GUYON, U. VON LUXBURG, S. BENGIO, H. M. WALLACH, R. FERGUS, S. V. N. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems 30 : Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, p. 4765–4774.

MADAAN N., PADHI I., PANWAR N. & SAHA D. (2021). Generate your counterfactuals : Towards controlled counterfactual generation for text. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, p. 13516–13524 : AAAI Press.

MILLER T. (2019). Explanation in artificial intelligence : Insights from the social sciences. *Artif. Intell.*, **267**, 1–38. DOI : 10.1016/j.artint.2018.07.007.

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.

PETSIUK V., DAS A. & SAENKO K. (2018). RISE : randomized input sampling for explanation of black-box models. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, p. 151 : BMVA Press.

RADFORD A., NARASIMHAN K., SALIMANS T. & SUTSKEVE I. (2018). Improving language understanding by generative pre-training.

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

RIBEIRO M. T., SINGH S. & GUESTRIN C. (2016). "why should I trust you ?" : Explaining the predictions of any classifier. In B. KRISHNAPURAM, M. SHAH, A. J. SMOLA, C. C. AGGARWAL, D. SHEN & R. RASTOGI, Éds., *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, p. 1135–1144 : ACM. DOI : 10.1145/2939672.2939778.

ROBEER M., BEX F. & FEELDERS A. (2021). Generating realistic natural language counterfactuals. In *Findings of the Association for Computational Linguistics : EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, p. 3611–3625 : Association for Computational Linguistics.

ROSIN C. D. (2011). Multi-armed bandits with episode context. *Ann. Math. Artif. Intell.*, **61**(3), 203–230. DOI : 10.1007/s10472-011-9258-6.

S. PUNLA C., HTTPS ://ORCID.ORG/ 0000-0002-1094-0018, CSPUNLA@BPSU.EDU.PH, C. FARRO R., HTTPS ://ORCID.ORG/0000-0002-3571-2716, RCFARRO@BPSU.EDU.PH & BATAAN PENINSULA STATE UNIVERSITY DINALUPIHAN, BATAAN, PHILIPPINES (2022). Are we there yet ? : An analysis of the competencies of BEED graduates of BPSU-DC. *International Multidisciplinary Research Journal*, **4**(3), 50–59.

SCIALOM T., DRAY P., LAMPRIER S., PIWOWARSKI B. & STAIANO J. (2020). Discriminative adversarial search for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 de *Proceedings of Machine Learning Research*, p. 8555–8564 : PMLR.

SCIALOM T., DRAY P., LAMPRIER S., PIWOWARSKI B. & STAIANO J. (2021a). To beam or not to beam : That is a question of cooperation for language gans. *Advances in neural information processing systems*.

SCIALOM T., DRAY P., STAIANO J., LAMPRIER S. & PIWOWARSKI B. (2021b). To beam or not to beam : That is a question of cooperation for language gans. In M. RANZATO, A. BEYGELZIMER, Y. N. DAUPHIN, P. LIANG & J. W. VAUGHAN, Éds., *Advances in Neural Information Processing Systems 34 : Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, p. 26585–26597.

SHANKARANARAYANA S. M. & RUNJE D. (2019). ALIME : Autoencoder Based Approach for Local Interpretability. *CoRR*, **abs/1909.02437**.

SILVER D., SCHRITTWIESER J., SIMONYAN K., ANTONOGLOU I., HUANG A., GUEZ A., HUBERT T., BAKER L., LAI M., BOLTON A., CHEN Y., LILLICRAP T. P., HUI F., SIFRE L., VAN DEN DRIESSCHE G., GRAEPEL T. & HASSABIS D. (2017). Mastering the game of go without human knowledge. *Nat.*, **550**(7676), 354–359. DOI : 10.1038/nature24270.

TAGARELLI A. & SIMERI A. (2022). Unsupervised law article mining based on deep pre-trained language representation models with application to the italian civil code. *Artificial Intelligence and Law*, **30**(3), 417–473.

VISANI G., BAGLI E. & CHESANI F. (2020). Optilime : Optimized LIME explanations for diagnostic computer algorithms. In *Proceedings of the CIKM 2020 Workshops co-located with 29th ACM International Conference on Information and Knowledge Management (CIKM 2020), Galway, Ireland, October 19-23, 2020*, volume 2699 de *CEUR Workshop Proceedings* : CEUR-WS.org.

WU T., RIBEIRO M. T., HEER J. & WELD D. S. (2021). Polyjuice : Generating counterfactuals for explaining, evaluating, and improving models. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1 : Long Papers), Virtual Event, August 1-6, 2021*, p. 6707–6723 : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.523.

ZAFAR M. R. & KHAN N. M. (2019). DLIME : A Deterministic Local Interpretable Model-Agnostic Explanations Approach for Computer-Aided Diagnosis Systems. *CoRR*, **abs/1906.10263**.

ZHANG S., ROLLER S., GOYAL N., ARTETXE M., CHEN M., CHEN S., DEWAN C., DIAB M. T., LI X., LIN X. V., MIHAYLOV T., OTT M., SHLEIFER S., SHUSTER K., SIMIG D., KOURA P. S., SRIDHAR A., WANG T. & ZETTLEMOYER L. (2022). OPT : open pre-trained transformer language models. *CoRR*, **abs/2205.01068**. DOI : 10.48550/arXiv.2205.01068.

ZHANG X., ZHAO J. J. & LECUN Y. (2015). Character-level convolutional networks for text classification. In C. CORTES, N. D. LAWRENCE, D. D. LEE, M. SUGIYAMA & R. GARNETT, Éds., *Advances in Neural Information Processing Systems 28 : Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, p. 649–657.

A Résultats qualitatifs

Nous donnons dans cette annexe, quelques exemples d'échantillons générés par Therapy pour les différentes classes des deux jeux de données étudiés ainsi que les 20 premiers top-mots retournés par la méthode. Veuillez noter que certains "mots" correspondent à des "sous-mots", dû à la décomposition en unigrammes (ve, ll, ...). L'approche proposée permet d'étudier l'impact de n-grammes, mais ceci n'étant pas possible avec LIME et SHAP (avec le code disponible), nous avons limité l'étude aux unigrammes.

A.1 amazon_polarity, classe "positive"

Échantillons :

— "The amount of support they gave was amazing. I'm so happy with the game and it's a great way for me to play my favorite role! Also, I love that they're giving us a new character every week, which is awesome :)I think they've been pretty good lately too. It's nice when we see someone who really loves their team and has fun doing things"

— "The whole point of this post is to show people what a great job you've put in. You are doing an excellent job, I love this sub. Keep going. :)I'm so happy for all those people that have helped! It's amazing to be here and it was fun being the one that got the most attention on the other day :). Thank you so much <3You're very welcome :) And yes... I am glad you enjoyed this subreddit - it is a great place"

Top-mots : *great, love, good, ve, years, people, lot, friends, fun, life, world, works, easy, things, happy, heard, including, awesome, nice, family*

A.2 amazon_polarity, classe "negative"

Échantillons :

— "I guess you could say it has to be the most boring thing ever.It is a very boring experienceI don't think they have any idea how to make money off this, so they will probably just do what you say they will if it doesn't work. They may try to sell you on the fact that it isn't working and then try and get you to buy back into it when they are done making money. If I was them, would they not want to pay me?"

— "It might be a bug, but it doesn't work on my phoneYeah that sucks! I was hoping the same thing happened.I don't think you can get a refund without having to buy a new one...You have no idea what they would be like without the warranty. They wouldn't know how much money is left if they didn't do it

properly or at all and they would probably just give it back as soon as they found out it was defective (or something)"

Top-mots : *don, money, bad, doesn, didn, idea, work, device, isn, thing, guess, wrong, back, buy, fact, time, phone, point, problem, thought*

A.3 ag_news, classe "world"

Échantillons :

— The world's most popular and sought after luxury home in Dubai, the Al Khaleel International Airport (ALI), is a city that is renowned as the hub of the Middle East and North Africa region. It is also the capital city of Abu Dhabi, United Arab Emirates with a population over 2 billion people and a GDP per person.Al-Khaliel airport is an important gateway into Europe to the north east and west of Europe, Asia, and Australia

— In the last month, the US government has been accused of using a "secret" military intelligence agency in Iraq and Afghanistan for spying on its allies, including Pakistan, Turkey, Iran, Russia, Syria, Sudan, Lebanon, Libya, Egypt, Nigeria...and even Saudi Arabia!According to a new report published today, the Pentagon is now investigating the alleged use by the US military Intelligence Community (IUC), which was created by President Bush after the 9/11 terror attack that killed over

Top-mots : *people, man, country, city, party, killed, family, agree, wrong, general, children, sex, president, police, working, military, dead, missing, woman, days*

A.4 ag_news, classe "sport"

— I think it's pretty safe to assume that the guy has a lot of experience with the game. He played in the NHL, and he was a top 10 player on the team for most games last season (he had a goal in his final three playoff series), and he won a Stanley Cup as a rookie this past season (he finished third in the league in points scored, which was good for second in the league) and is still one of the best players in hockey at this stage in the year

— We've got to keep playing this game. This team needs to win games and we need to play the best basketball that's been in our league all season, every night. And it is time to get out there and do that.""The Warriors have won three straight games at home and are 2-0 against the NBA's best team on both sides of the ball this season with a record of 21-1 (13.7 points per game)

Top-mots : *time, game, back, season, play, didn, team, guy, field, night, games, left, 12, title, won, saturday, playing, great, day, wasn*

A.5 ag_news, classe "business"

Échantillons :

— I am still in shock after hearing of that.It’s a pretty big deal. It happened last month. They are trying to get the money out of the company by selling their stock for profit so they can sell more shares and buy more shares at higher prices (which I think would have helped with the stock market) and it was reported as an ""investment fraud"" by the SEC which has been going on all over this subreddit for months, but no one ever seems to care much

— Biden is planning to spend millions of dollars to buy a new home, but the real estate market in America is still struggling with the housing shortage. The average house sale cost $1 billion and was up by nearly 50 percent from the previous year’s price of about $800 million -- according to the Real Estate Board of New York (RBE).The RBE estimates that the average house sales prices are expected to rise 1,000 per month this fiscal year as the economy continues its rebound

Top-mots : *money, buy, care, doesn, things, deal, pay, worth, business, car, biggest, interested, month, trade, don, compagny, happened, store, kind, price*

A.6 ag_news, classe "sci/tech"

Échantillons :

— 2K Games’ Dark Souls 3 is coming to PC, Mac & Linux in the near future.The new game will launch for free on PC, Mac & Linux and Xbox One, PlayStation 5 and Microsoft Windows, as well. It’ll come out sometime during this week, with an official release expected soon thereafter, though we don’t yet know what it will be called or where exactly you’re getting the title. We also have some news from Sony that’s not quite so surprisingetc...

— In this new age of technology, the world needs more people. We have a lot in our hands. The internet can help us connect to others through video chat and online games.""The company will launch a mobile game called ’Gangster’, where it plans to offer ""an interactive experience"" with its users, according to the company. The game has been developed for the Apple iPad and Android phones that use Apple TV, which also uses Google Chromecast, according to a release.

Top-mots : *ve, ll, idea, phone, internet, make, system, video, online, life, understand, version, pc, found, 13, thing, computer, lot, hard, issue, people, work, information, future*