Une approche basée sur la méthode LRP pour l'explication des Réseaux de Neurones Convolutifs appliqués à la classification des textes

Florentin Jiechieu, Norbert Tsopze

Une approche basée sur la méthode LRP pour l'explication des Réseaux de Neurones Convolutifs appliqués à la classification des textes

Florentin Jiechieu[*1,2] **and Norbert Tsopze**[1,2]

[1]Département d'Informatique - Université de Yaoundé I, Yaoundé, Cameroun
[2]Sorbonne Université, IRD, UMMISCO, F-93143, Bondy, France

*E-mail : florentin.jiechieu@facsciences-uy1.cm, tsopze.norbert@gmail.com

Résumé

Jacovi et al. en 2018 ont proposé une méthode pour l'explication des Réseaux de Neurones Convolutifs conçus pour la classification des textes (Text-CNN). Le problème avec cette méthode est qu'elle ne s'applique que sur les Text-CNN qui n'ont pas de couche cachée dans la partie densément connectée. On se retrouvera donc limité en terme de performance si on veut utiliser cette méthode d'explication. Par ailleurs, la méthode LRP (Layer-wise Relevance Propagation) permet de calculer les contributions des entrées des réseaux de neurones profonds (plusieurs couches cachées), mais ne peut s'appliquer directement sur les Text-CNN car contrairement aux images où chaque pixel a un sens pris seul, chaque composante d'une représentation vectorielle (word embedding) d'un mot prise seule n'a pas de signification particulière. C'est ainsi que nous proposons dans cet article, d'étendre la méthode de Jacovi et al. en mettant à contribution la méthode LRP afin de pouvoir expliquer les Text-CNN dont la partie densément connectée est profonde. L'évaluation qualitative laisse percevoir que les explications fournies par la méthode proposée sont cohérentes sur les problèmes de classification des textes comme le Question-Answering (QA) ainsi que l'analyse des sentiments. Par ailleurs, l'évaluation quantitative montre que les explications sont fidèles à 100% au modèle.

Mots-Clés

Explicabilité, LRP, Classification des Textes, Text-CNN.

I INTRODUCTION

Les réseaux de neurones profonds en l'occurrence les réseaux de neurones convolutifs, ont démontré des performances exceptionnelles ces dernières décenies avec des résultats très intéressants dans plusieurs problèmes de classification d'images, de textes, et autres [1-3]. Malgré ces brillantes performances, ces modèles ont longtemps été considérés comme des boites noires qui produisent de bon résultats sans qu'on puisse expliquer pourquoi. Ce facteur en fait l'une des principales limites surtout dans des domaines sensibles comme la médecine où des vies humaines sont en jeu ou encore dans le domaine du trading, où un trader voudrait savoir pourquoi on lui demande d'investir dans telle action ou pas. En effet, confier des décisions importantes à des systèmes dont on ne peut déchiffrer le raisonnement présente un risque sérieux [4]. C'est la raison pour laquelle, ce domaine de recherche concernant l'explicabilité des réseaux de neurones a vu le jour et suscite énormément d'attention de la part des chercheurs.

Plusieurs travaux en rapport à l'explicabilité des réseaux de neurones convolutifs ont à ce jour été effectués et l'imense majorité de ces travaux concerne l'explication des modèles de classification des images [5-7]. Les travaux concernant l'explicabilité de ce type de réseau pour la classification des textes sont en proportions faibles [8-10]. Parmi les auteurs, Jacovi et al. en 2018 ont proposé une méthode d'explication des Text-CNN dont le principe est de déterminer les importances des *n-grammes* (groupe de n mots successifs) mais ne s'applique que sur les Text-CNN avec une seule couche neuronale dans la partie densément connectée [8]. D'autre part, la méthode LRP [11], pour le cas d'un CNN qui prend une matrice en entrée permettra de calculer les contributions de chaque élément de la matrice par rapport à la valeur prédite en sortie du CNN. Dans le cas de l'image, chaque composante de la matrice représente un pixel et LRP permettra d'évaluer la contribution de ce pixel ; pour ce qui est du texte, chaque composante du vecteur d'un mot (dans une représentation word embedding) n'a pas une signification particulière ; ce qui fait que la méthode LRP ne peut être appliquée directement sur les Text-CNN à cause de la nature discrète de ces derniers.

Fort de ces considérations, nous proposons une méthode d'explicabilité qui combine la méthode LRP avec celle de Jacovi et al. afin de pouvoir expliquer les Text-CNN disposant d'un nombre quelconque de couches cachées dans la partie densément connectée. Dans cette approche, LRP est utilisée pour calculer les contributions des unités situées en sortie de la couche *max-pooling* alors que le principe de la méthode de Jacovi et al. est utilisé pour assigner ces contributions aux mots correspondant et situés en entrée du Text-CNN.

Dans la suite, nous débuterons par une présentation du fonctionnement du Text-CNN ainsi que les méthodes de Jacovi et LRP avant de nous appesentir sur la méthode que nous avons proposée.

II ARCHITECTURE DU TEXT-CNN

En 2014, KIM a proposé une architecture de Text-CNN possédant un ensemble de couches, chacune jouant un rôle bien précis (figure 1) [12].

2.1 La couche d'entrée

Le Text-CNN prend en entrée une séquence de mots, chaque mot étant encodé par un vecteur de numérique en utilisant une technique de *word embedding*. Un mot w peut ainsi être modélisé comme un vecteur de dimension d i.e $w \in R^d$. Ainsi, un texte de longueur n qui est une séquence de n mots sera représenté par une matrice M de dimension $n \times d$ i.e $M = w_1, w_2, ..., w_n \in R^{n \times d}$; et un l-gramme (groupe de l mots successifs) par une matrice $u_i =< w_i, ..., w_{i+l-1} > (0 \leq i \leq n-l)$, $u_i \in R^{l \times d}$.

2.2 La couche de convolution

La couche de convolution est constituée d'un ensemble de filtres de noyau de taille l qui peuvent être représentés par une matrice $f_j \in R^{l \times d}$ où l représente le nombre de mots successifs que le filtre pourra détecter ; d est la dimension des vecteurs de représentation des mots. Ainsi, un filtre de noyau l détectera un l-gramme. La convolution réalise le produit scalaire $< u_i, f_j >$ entre un l-gramme u_i et un filtre f_j. Les convolutions d'un filtre f_j avec la matrice en entrée produisent un vecteur colonne $F,_j$ appelé *feature map* associé au filtre f_j, et pour l'ensemble des filtres, on aura une matrice $F \in R^{(n-l+1) \times m}$ où m est le nombre total de filtres, et les colonnes de la matrice représentent les *feature maps* associés aux différents filtres convolutifs. Les valeurs à l'intérieur des *feature maps* sont ensuite rectifiées

en utilisant la fonction ReLU (Rectified Linear Unit) qui opère en mettant toutes les valeurs négatives à 0 avant de les passer en entrée de la couche de *Max-Pooling*.

2.3 La couche de Max-Pooling (Global Max-Pooling)

Le filtre global max-pooling sélectionne dans chaque *feature map* la valeur maximale correspondante au *n-gramme* qui a eu le score de convolution le plus élevé avec le filtre associé à ce *feature map*. Le résultat du *global max-pooling* est un vecteur de dimension m (m étant le nombre de filtres).

2.4 La couche densément connectée

Cette couche prend en entrée le vecteur P issue du filtre *max-pooling* et produit un vecteur de sortie qui correspond au résultat de la classification du texte. Les composantes du vecteur de sortie représentent généralement les probabilités d'appartenance à chaque classe.

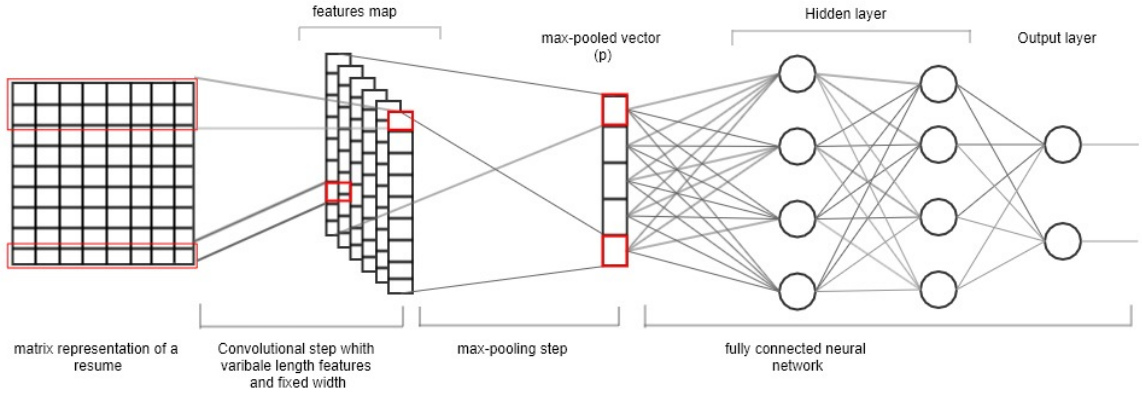


FIGURE 1 – Architecture d'un Text-CNN.

III EXPLICABILITÉ DU TEXT-CNN

Dans cette section nous présentons la méthode de Jacovi et al. ainsi que la méthode LRP dont dérive notre méthode d'explicabilité et présentons également la Méthode LIME (Local Interpratable Model-Agnostic Explanations), une méthode de reférence avec laquelle nous allons nous comparer.

3.1 La méthode de Jacovi

Le principe de la méthode de Jacovi et al. [8] pour l'explication est assez simple. Il s'agit de trouver la classe que chaque caractéristique (mot, *n-gramme*) explique. Pour ce faire, les auteurs proposent de calculer la contribution de chaque *n-gramme* par rapport à chaque classe et d'attribuer ce *n-gramme* à la classe pour laquelle la contribution est maximale. L'ensemble des *n-grammes* caractérisant une classe sera utilisé pour expliquer la prédiction de cette classe. Pour ce faire, il faut pouvoir déterminer les contributions de chaque *n-gramme* en entrée par rapport aux sorties du Text-CNN. Les auteurs constatent que si l'on dispose des contributions des entrées de la couche densément connectée (qui correspondent aux résultats du max-pooling), on peut facilement retrouver les *n-grammes* associés à chacune de ces entrées et leur associer ces contributions. En effet, il est aisé de retrouver le *n-gramme* qui a produit la valeur maximale lors de la convolution et qui a été sélectionnée par le filtre

max-pooling. Le problème de cette méthode réside surtout dans la difficulté de calculer les contributions des entrées de la partie densément connectée lorsqu'on y a plusieurs couches cachées. Les auteurs ont contourné ce problème en travaillant sur un *Text-CNN* avec une seule couche dans la partie densément connectée. Ainsi ils considérent l'importance d'une entrée x_i de cette couche par rapport à une classe y_j comme étant le poids de connection w_{ij} de x_i avec y_j. La classe caractérisée par l'entrée x_i est donc la classe pour laquelle w_{ij} est maximale.

La méthode de Jacovi et al. limite donc son utilisateur en terme de performance du modèle à expliquer puisqu'il faut nécessairement utiliser un modèle de Text-CNN dont la partie densément connectée est constituée d'une seule couche neuronale. Pour résoudre ce problème, nous avons pensé à utiliser la méthode LRP pour pouvoir calculer les contributions des entrées de la couche densément connectée en partant des sorties vers ces entrées. Cela nous permet d'expliquer des architectures de Text-CNN plus complexes et plus performantes.

3.2 La méthode LRP

La méthode LRP a été proposée pour la première fois dans [11] et est utilisée pour évaluer les contributions des caractéristiques en entrée d'un réseau de neurones par rapport à ses sorties ; et ce, quelque soit le nombre de couches qui séparent les entrées des sorties. Le principe de LRP est assez simple. Si on considère une sortie f_j d'un neurone j de la couche de sortie, la méthode LRP commence par calculer les contributions de chaque unité de la couche qui précède directement la couche de sortie en décomposant f_j et en attribuant à chaque unité de cette couche son apport dans la valeur f_j. De manière générale, la contribution d'une unité i de la couche l à une unité j de la couche $l+1$ notée C_{ij} est donnée par la formule 1 :

$$C_{ij} = \frac{a_i w_{ij}}{\sum_k a_k w_{kj}} \tag{1}$$

a_i représente l'activation du neuron i de la couche k, et w_{ij} le poids de connexion du neurone i de la couche l au neurone j de la couche $l+1$. Ayant obtenu les contributions de l'avant dernière couche à la couche de sortie par l'équation précédente, nous pouvons calculer les contributions des couches qui la précèdent en utilisant une formule transitive qui se traduit par l'équation récurrente (2) :

$$C^*_{ij} = \sum_k C_{ik} \times C^*_{kj} \tag{2}$$

De manière litérale, cette équation stipule que la contribution d'une unité i d'une couche l à une unité de sortie j notée C^*_{ij} est égale à la somme des contributions de cette unité aux unités de la couche suivante (les C_{ik}) multipliées par les contributions de ces dernières aux sorties (C^*_{kj}). En procédant ainsi de proche en proche on arrive à calculer les contributions des entrées du modèle par rapport aux sorties.

3.3 La méthode LIME

LIME est une méthode d'explicabilité gérénique applicable à tout type de modèle [13]. Pour expliquer, LIME n'a pas besoin de comprendre la structure ou le fonctionnement interne du modèle. Le principe de LIME est d'effectuer des petites variations du texte en entrée du modèle (suppression des mots) et d'en évaluer l'incidence sur la sortie du modèle ; cela permet d'évaluer en quelque sorte l'importance de chaque mot du texte en entrée du modèle

par rapport à la valeur prédite en sorite. L'inconvénient majeure avec cette méthode est que pour expliquer une prédiction donnée pour un texte en entrée, il faut faire plusieurs variations du même texte ; ce qui augmente la complexité de la méthode.

IV NOTRE MÉTHODE

Le principe de notre méthode consiste à utiliser LRP pour calculer les contributions des unités de la couche densément connectée et par la suite d'utiliser le principe de la méthode de Jacovi et al. pour retrouver les *n-grammes* associés à ces contributions. Nous proposons en plus une formule permettant d'aggréger les contributions au cas où plusieurs sorties du *max-pooling* correspondraient au même *n-gramme*.

La figure 2 illustre schématiquement le principe précédemment décrit.

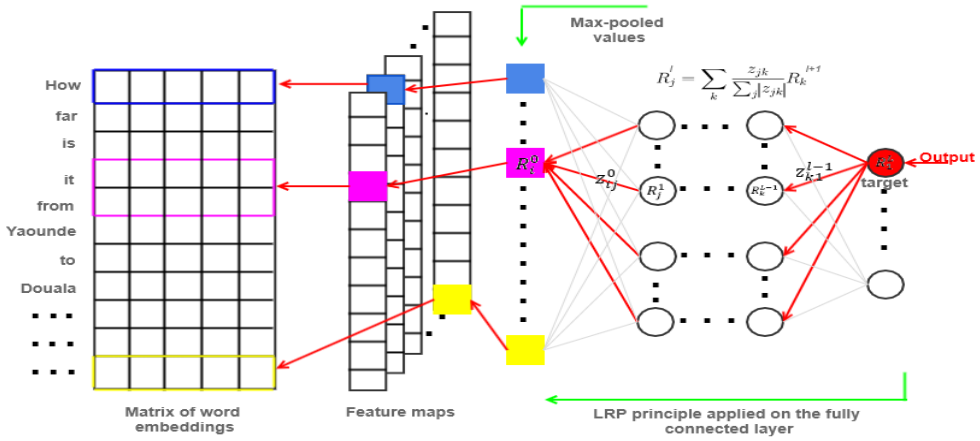


FIGURE 2 – Apperçu de la méthode d'explicabilité proposée.

Les étapes de notre méthode telles qu'illustrées dans la figure 2 sont les suivantes :

1. Premièrement, un texte est passé en entrée du Text-CNN et la sortie $f(x)$ est calculée : c'est la classification ;
2. Ensuite, la méthode LRP est utilisée pour calculer les contributions de chaque composante du vecteur issu du max-pooling par rapport à la sortie $f(x)$. Il s'agit du vecteur en entrée de la couche densément connectée ;
3. Enfin, les *n-grammes* associés sont déterminés et leurs importances calculées.

En effet, chaque *n-gramme* ayant produit la valeur maximale et sélectionné par le filtre *max-pooling* sera considéré comme la caractéristique détectée par le filtre convolutif. Seuls ces derniers influenceront le résultat par la suite [8]. Les contributions calculées par LRP correspondent en quelque sorte aux contributions de chaque filtre puisque chaque entrée de la couche densément connectée est issue d'un filtre. Notons donc par R_f le vecteur de contributions d'un filtre f par rapport au vecteur de sortie du modèle. la contribution d'un *n-gramme* u_i noté R_{u_i} sera calculée comme étant la somme des contributions des filtres qui détectent ce *n-gramme*. Ce qui se traduira par la formule suivante :

$$R_{u_i} = \sum_{f \in A_i} R_f \tag{3}$$

où A_i représente l'ensemble des filtres qui détectent le *n-gramme* u_i. R_{u_i} représente les contributions du *n-gramme* u_i par rapport aux sorties. Avec ce procédé, on obtient ainsi les contributions de chaque *n-gramme* détecté par les filtres convolutifs.

V RÉSULTATS

La méthode d'explicabilité proposée dans cet article a été testée sur deux problèmes de classification des textes à savoir le *Question-Answering* et l'analyse des sentiments. Les jeux de données utilisés sont les suivants :

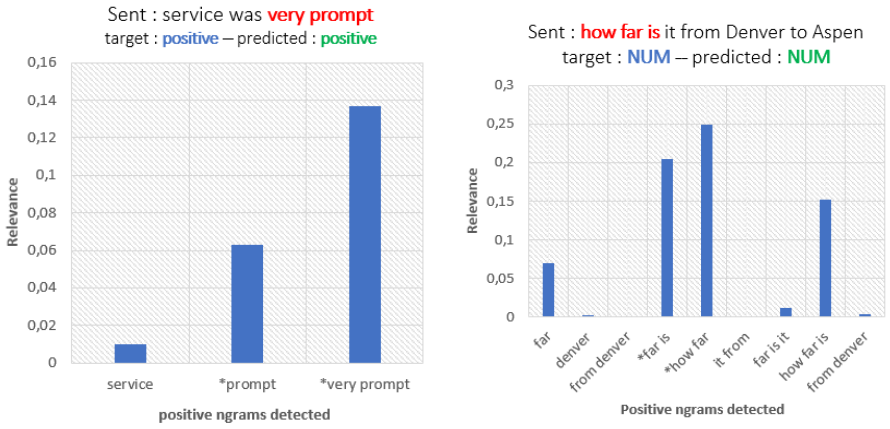
— IMDB : commentaires sur les vidéos pour la classification binaire des sentiments [14] ;
— TREC-QA_5500 : jeu de données pour la classification des questions-réponses avec 5500 questions [15].

5.1 Evaluation qualitative

Une évaluation qualitative a permis d'apprécier perceptuellement la qualité des résultats sur des cas concrets. Le tableau 1 représente la distribution des importances des mots par rapport aux classes sous forme de carte d'attention dans le cadre de la classification d'une question. Les classes en colonne représentent les types de réponse attendus pour une question (DESC=DESCRIPTION, ENTY=ENTITY, ABBR=ABBREVIATION, HUM=HUMAN, NUM=NUMBER, LOC= LOCALISATION). On observe à travers cette carte, l'importance de chaque mot par rapport aux différentes classes. Ainsi, le mot who par exemple, caractérise beaucoup la classe "HUM" et moins la classe "NUM".

TABLE 1 – Importance des *n-grammes* détectés pour la question *« who was the star witness at the senate watergate hearings ? »*

	DESC	ENTY	ABBR	HUM	NUM	LOC
senate	0.02	-0.02	-0.00	-0.00	0.01	-0.00
witeness	0.01	-0.04	-0.01	-0.00	0.04	-0.01
hearings	0.03	0.03	0.02	0.00	-0.03	-0.03
water-gate	0.01	0.01	0.01	0.00	-0.02	-0.03
who	-0.15	0.07	-0.17	0.37	-0.28	-0.16
at	-0.01	0.01	0.00	-0.00	-0.01	-0.00
the	0.00	-0.03	-0.07	0.08	-0.02	-0.02
was	-0.01	-0.02	-0.02	0.04	0.00	-0.01



(a) Cas de l'analyse des sentiments (b) Cas du "Question-Answering"

FIGURE 3 – Distribution des importances des *n-grammes*.

La figure 3 quant à elle présente les distributions des importances des *n-grammes* dans un problème d'analyse des sentiments (figure 3(a)) et un problème de *Question-Aswering* (figure 3(b)). La projection des termes les plus importants (en rouge) dans chaque cas dans la phrase en entrée permet de se rendre compte que ces termes expliquent effectivement le sentiment de la phrase dans le premier cas, ou le type de la réponse attendu dans le second.

5.2 Evaluation Quantitative

L'évaluation quantitative se fait en mesurant la fidélité [7] des explications c'est-à-dire à quel degré les explications fournies reflètent effectivement le raisonnement interne du modèle. La technique que nous utilisons pour évaluer la fidélité consiste à utiliser un modèle de substitution [16] qui calcule ses sorties en faisant la somme linéraire des contributions des *n-grammes* calculées par notre méthode d'explicabilité (formule 4).

$$y_j = \sum_i R_{u_i}^j \tag{4}$$

La classe prédite par ce modèle de substitution est la classe $j*$ pour laquelle la valeur y_j est maximale (ie $j* = \arg\max_j y_j$). R_{u_i} représente les contributions des *n-grammes* que notre méthode a calculées par rapport aux sorties. Evaluer la fidélité dans ce cas revient à calculer le taux de classification du modèle de substitution en considérant les classes attendues (vérité terrain) comme étant les classes prédites par le modèle à expliquer pour les mêmes phrases en entrée. Avec cette évaluation, nous obtenons les résultats suivants :

TABLE 2 – Comparaison de la fidélité de notre méthode avec celle de LIME sur les problèmes de Question-Answering (QA) et d'analyse de sentiment (SA).

	QA	SA
Our Method	100.0%	100.0%
LIME	96%	96,07%

Suivant cette méthode d'évaluation, notre méthode a une fidélité de 100% sur les deux problèmes, bien au dessus de celle de LIME qui maintien une fidélité de 96% environ sur ces mêmes problèmes. Cette performance peut se justifier par la propriété de conservation induite par l'utilisation de LRP qui veut que la somme des importances par rapport à une sortie soit égale à la valeur de cette sortie. On peut dire que l'utilisation de LRP induit une bonne redistribution des importances aux caractéristiques d'entrée par rapport à leur contribution réelle sur les sorties.

VI CONCLUSION ET PERSPECTIVES

Il était question dans le cadre de ce travail de combiner les méthodes LRP et Jacovi afin d'obtenir une méthode d'explicabilité des Text-CNN plus générique. Les expériementations que nous avons effectuées montrent que cette méthode fournit des explications cohérentes pour l'être humain et sont hautement fidèles au fonctionnement interne du modèle. par ailleurs les résultats obtenus montrent que notre méthode permet effectivement d'expliquer plusieurs architectures de Text-CNN en conservant les propriétés de la méthode LRP et ayant une fidélité supérieure à celle de LIME. Toutefois nous envisageons explorer davatange de méthodes de calcul de la fidélité pour une meilleure fiabilité de l'évaluation.

RÉFÉRENCES

[1] D. THECKEDATH et R. SEDAMKAR. « Detecting affect states using VGG16, ResNet50 and SE-ResNet50 networks ». In : *SN Computer Science* 1.2 (2020), pages 1-7.

[2] A. KRIZHEVSKY, I. SUTSKEVER et G. E. HINTON. « Imagenet classification with deep convolutional neural networks ». In : *Advances in neural information processing systems* 25 (2012).

[3] S. LAI, L. XU, K. LIU et J. ZHAO. « Recurrent convolutional neural networks for text classification ». In : *Twenty-ninth AAAI conference on artificial intelligence*. 2015.

[4] A. ADADI et M. BERRADA. « Peeking inside the black-box : A survey on Explainable Artificial Intelligence (XAI) ». In : *IEEE Access* 6 (2018), pages 52138-52160.

[5] X. LEI, H. PAN et X. HUANG. « A dilated CNN model for image classification ». In : *IEEE Access* 7 (2019), pages 124087-124095.

[6] Q.-s. ZHANG et S.-C. ZHU. « Visual interpretability for deep learning : a survey ». In : *Frontiers of Information Technology & Electronic Engineering* 19.1 (2018), pages 27-39.

[7] R. GUIDOTTI, A. MONREALE, S. RUGGIERI, F. TURINI, F. GIANNOTTI et D. PEDRESCHI. « A survey of methods for explaining black box models ». In : *ACM computing surveys (CSUR)* 51.5 (2018), pages 1-42.

[8] A. JACOVI, O. SAR SHALOM et Y. GOLDBERG. « Understanding Convolutional Neural Networks for Text Classification ». In : *Proceedings of the 2018 EMNLP Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium : Association for Computational Linguistics, 2018, pages 56-65.

[9] L. ARRAS, F. HORN, G. MONTAVON, K.-R. MÜLLER et W. SAMEK. « " What is relevant in a text document?" : An interpretable machine learning approach ». In : *PloS one* 12.8 (2017), e0181142.

[10] S. SERRANO et N. A. SMITH. « Is Attention Interpretable? » In : *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy : Association for Computational Linguistics, juill. 2019, pages 2931-2951.

[11] S. BACH, A. BINDER, G. MONTAVON, F. KLAUSCHEN, K.-R. MÜLLER et W. SAMEK. « On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation ». In : *PLOS ONE* 10.7 (juill. 2015), pages 1-46.

[12] Y. KIM. « Convolutional Neural Networks for Sentence Classification ». In : *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar : Association for Computational Linguistics, oct. 2014, pages 1746-1751.

[13] M. T. RIBEIRO, S. SINGH et C. GUESTRIN. « "Why Should I Trust You?": Explaining the Predictions of Any Classifier ». In : *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA : ACM, 2016, pages 1135-1144. ISBN : 978-1-4503-4232-2.

[14] A. L. MAAS, R. E. DALY, P. T. PHAM, D. HUANG, A. Y. NG et C. POTTS. « Learning Word Vectors for Sentiment Analysis ». In : *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*. Portland, Oregon, USA : Association for Computational Linguistics, juin 2011, pages 142-150.

[15] E. M. VOORHEES. « The TREC question answering track ». In : *Natural Language Engineering* 7.4 (2001), pages 361-378.

[16] O. LAMPRIDIS, R. GUIDOTTI et S. RUGGIERI. « Explaining sentiment classification with synthetic exemplars and counter-exemplars ». In : *International Conference on Discovery Science*. Springer. 2020, pages 357-373.