Sensibilité des explications à l'aléa des grands modèles de langage : le cas de la classification de textes journalistiques

Jérémie Bogaert[*], **Marie-Catherine de Marneffe**[**], **Antonin Descampe**[**], **Louis Escouflaire**[**], **Cédrick Fairon**[**], **François-Xavier Standaert**[*]

[*] *Université catholique de Louvain, ICTEAM Institute, Louvain-la-Neuve, Belgium*
[**] *Université catholique de Louvain, ILC Institute, Louvain-la-Neuve, Belgium*

e-mails: firstname.lastname@uclouvain.be

RÉSUMÉ. *Les grands modèles de langage sont performants en traitement automatique du langage mais posent des défis d'explicabilité. Nous examinons l'effet des éléments aléatoires de leur entraînement sur l'explicabilité de leurs prédictions en nous focalisant sur une tâche de classification de textes journalistiques d'opinion en français. Utilisant un modèle CamemBERT peaufiné et une méthode d'explication basée sur la propagation de pertinence, nous constatons que des entraînements avec différentes graines aléatoires produisent des modèles aux performances similaires mais aux explications variables. Nous affirmons dès lors que caractériser la distribution statistique des explications est nécessaire pour une explicabilité satisfaisante de ce type de modèle. Nous explorons ensuite un modèle basé sur des traits textuels qui offre des explications stables mais une précision moindre. Celui-ci correspond donc à un compromis différent entre exactitude et explicabilité et nous montrons qu'il est possible de l'améliorer en intégrant des traits extraits des explications de CamemBERT. Nous discutons enfin de pistes de recherche que nos résultats suggèrent, en particulier sur l'origine de la sensibilité à l'aléa observée.*

MOTS-CLÉS : *explicabilité, modèles transformer, classification, discours journalistique.*

TITLE. *Sensitivity of Explanations to the Randomness of Large Language Models: a Case Study on Journalistic Text Classification*

ABSTRACT. *Large language models perform well in natural language processing but raise explainability challenges. We examine the effect of random elements in their training on the explainability of their predictions by focusing on a task of opinionated journalistic text classification in french. Using a fine-tuned CamemBERT model and an explanation method based on relevance propagation, we find that training with different random seeds produces models with similar accuracies but variable explanations. We therefore claim that characterizing the explanations' statistical distribution is needed for this type of model to be explainable. We then*

explore a simpler model based on textual features which offers stable explanations but is less accurate. Hence, this model corresponds to a different tradeoff between accuracy and explainability and we show that it can be improved by inserting features derived from CamemBERT's explanations. We finally discuss new research directions suggested by our results, in particular regarding the origin of the observed sensitivity to the training randomness.

KEYWORDS: *explainability, transformer models, classification, press discourse.*

1. Introduction

1.1. *Contexte général de la recherche*

Les grands modèles de langage de type *transformer*, tels que BERT (Devlin *et al.*, 2019) et GPT (Brown *et al.*, 2020) montrent des performances impressionnantes pour une variété de tâches en traitement automatique du langage (TAL), par exemple dans la classification automatique de textes (Acheampong *et al.*, 2021). Cependant, le manque d'explicabilité de ces modèles complexes (parfois dits « boîtes noires ») est une préoccupation majeure dans de nombreux contextes où ils sont exploités, en particulier lorsque les décisions de tels modèles peuvent avoir des implications importantes, par exemple dans le domaine juridique (Zini et Awad, 2023). En outre, la définition des conditions nécessaires d'explicabilité d'un modèle ne fait pas encore l'objet d'un consensus large (Murdoch *et al.*, 2019). Comme détaillé dans une étude récente (Lyu *et al.*, 2022), différents critères supposés désirables pour l'explicabilité d'un modèle ont été introduits dans la littérature, mais les liens entre ces critères ne sont pas formellement établis et leur évaluation rigoureuse est souvent difficile.

Les deux critères couramment mis en avant comme étant les plus fondamentaux pour l'explicabilité d'un modèle de langage sont la fidélité (*faithfulness*) et la plausibilité (*plausibility*). La fidélité se définit comme la capacité d'une explication à refléter avec précision le processus de raisonnement (algorithmique) qui a mené à une prédiction (Ribeiro *et al.*, 2016 ; Jacovi et Goldberg, 2020). La plausibilité se définit comme la capacité d'une explication à être compréhensible et convaincante pour un lecteur (Herman, 2017 ; Jacovi et Goldberg, 2020). Différentes méthodes d'explication ont été proposées dans la littérature, avec pour objectif de combiner ces deux critères. Un exemple récent, que nous utilisons dans l'article, est la méthode de propagation de pertinence couche par couche (Chefer *et al.*, 2021), ou *Layerwise Relevance Propagation* (LRP). Elle produit des explications dans le format des cartes d'attention, supposé facilement compréhensible par un lecteur humain (Sen *et al.*, 2020).

Un critère plus technique (et donc plus facile à quantifier) des explications d'un modèle de langage est leur sensibilité à différents types de variation (Sundararajan *et al.*, 2017 ; Adebayo *et al.*, 2018). Par exemple, la sensibilité aux données d'entrée implique qu'une explication doive (ou ne doive pas) dépendre des modifications des textes à traiter si ces derniers modifient (ou ne modifient pas) une prédiction pour un texte. En lien plus direct avec nos préoccupations, il a également été proposé qu'une explication doive être sensible aux modèles, c'est-à-dire qu'elle dépende (ou ne dépende pas) des changements du modèle qui influencent (ou n'influencent pas) les prédictions. Plus précisément, le concept d'invariance à l'implémentation formalise le fait que des modèles fonctionnellement équivalents[1] devraient avoir des explications identiques (Sundararajan *et al.*, 2017). Ce besoin est néanmoins relativisé par certains auteurs, rien n'excluant en effet que deux méthodes algorithmiques différentes mènent de façon déterministe à la même solution (Lyu *et al.*, 2022).

1. Des modèles fonctionnellement équivalents ont des prédictions identiques pour toute entrée.

À notre connaissance, la fidélité et la plausibilité des explications d'un modèle ont jusqu'à présent été étudiées pour des modèles fixés, résultant d'une exécution donnée de leur optimisation. En outre, leur sensibilité a été étudiée pour des modèles ayant des implémentations différentes, par exemple modifiées en retirant certaines caractéristiques (*features*). En revanche, la sensibilité des méthodes d'explications aux hyperparamètres utilisés lors de différentes exécutions de l'optimisation d'un modèle n'a pas été étudiée systématiquement. Ceci alors que la phase d'entraînement de nombreuses méthodes d'apprentissage utilise des hyperparamètres choisis aléatoirement ou, au mieux, heuristiquement. Dans cet article, nous nous intéressons dès lors aux méthodes d'apprentissage qui incluent des éléments aléatoires dans leur processus d'entraînement, pour lesquelles on ne peut pas déterminer *a priori* l'impact sur l'optimisation de l'exactitude (*accuracy*) des prédictions. C'est donc le processus d'apprentissage qui permet de déterminer *a posteriori* l'impact de ces éléments aléatoires.

Plus précisément, l'entraînement d'une méthode d'apprentissage nécessite généralement la sélection d'un certain nombre d'hyperparamètres, tant pour le modèle lui-même (par exemple, taille et topologie du réseau) que pour l'algorithme d'optimisation (par exemple, vitesse d'apprentissage et taille des lots ou *batch size*). En outre, les méthodes d'optimisation stochastiques exploitent une quantité d'aléa qui est souvent générée de façon déterministe à partir d'un paramètre supplémentaire, habituellement appelé graine (*seed*). La graine est alors utilisée pour initialiser un générateur de nombres pseudo-aléatoires afin de produire la quantité d'aléa requise par l'algorithme d'optimisation. Ce paramètre est généralement rendu public à des fins de reproductibilité des résultats mais, fondamentalement, rien n'empêche de l'utiliser comme hyperparamètre et de comparer la qualité des modèles obtenus avec différentes graines. Cette approche est rarement recommandée en pratique, car elle ne permet pas d'autre stratégie qu'une recherche exhaustive (rien ne permettant en effet de distinguer l'aléa généré avec une graine de l'aléa généré avec une autre graine). Elle nous intéresse néanmoins dans cet article car il s'agit d'un exemple d'élément parfaitement aléatoire utilisé lors de la phase d'entraînement d'un modèle. Dans la suite de cet article, c'est donc précisément l'impact des graines aléatoires utilisées comme des hyperparamètres de l'apprentissage que nous allons étudier. Rien n'empêcherait par ailleurs d'étendre cette réflexion à d'autres hyperparamètres qui, même s'ils ne sont pas choisis de façon parfaitement aléatoire, impliquent des éléments aléatoires dans leur sélection.

1.2. *Question de recherche et contributions*

Partant de ce contexte général, la question qui nous occupe dans cet article est la suivante : la sensibilité des grands modèles de langage aux éléments aléatoires de leur entraînement peut-elle être significative au point d'influencer leur explicabilité ?

De façon évidente, vu qu'un processus d'entraînement utilisant différents hyperparamètres choisis aléatoirement peut mener à des modèles différents, ces derniers peuvent également avoir des exactitudes (*accuracies*) différentes. En général, les valeurs des hyperparamètres menant au modèle le plus performant sont donc choisies.

Notre étude demande dès lors une première restriction : nous nous intéressons à des sous-ensembles d'hyperparamètres qui mènent à des modèles ayant une exactitude suffisamment proche. Nous définirons comme statistiquement équivalents des modèles dont les différences d'exactitude ne sont pas statistiquement significatives. En outre, même des modèles statistiquement équivalents ne sont pas forcément fonctionnellement équivalents : nous nous intéressons donc aux sous-ensembles d'entrées pour lesquelles ces modèles mènent à la même prédiction. Nous appellerons concordantes ces entrées de modèles équivalents qui donnent la même prédiction[2].

Informellement, ces définitions permettent de restreindre notre étude à des sous-ensembles d'entrées pour lesquelles rien ne permet de déterminer si un modèle est préférable à un autre. Dans ce contexte, nous affirmons que si une méthode d'apprentissage mène à un ensemble non-négligeable de modèles équivalents dont les explications diffèrent, alors se limiter à l'explication d'un seul modèle obtenu avec cette méthode d'apprentissage est insuffisant. Précisément, s'il est vrai que différentes méthodes algorithmiques peuvent mener à la même solution (auquel cas, l'explication d'une seule méthode peut suffire), la présence d'aléa dans le processus d'entraînement de modèles équivalents nécessite de s'assurer que la distribution statistique des explications issues de modèles équivalents diffère assez de la distribution uniforme. Une telle distribution impliquerait en effet que toutes les explications possibles sont équiprobables et le choix d'une explication relèverait alors de l'arbitraire.

Dans une première partie de l'article, nous démontrons expérimentalement que pour une combinaison raisonnable d'une méthode d'apprentissage et d'un outil d'explication, des ensembles non négligeables de modèles équivalents peuvent être observés en pratique. Pour ce faire, nous utilisons la méthode LRP mentionnée précédemment et cherchons à expliquer les résultats de différents modèles de type *transformer*, tous basés sur le modèle CamemBERT (Martin *et al.*, 2020), peaufiné (*fine-tuned*) avec le même ensemble d'apprentissage et avec différentes graines choisies aléatoirement. Cette combinaison d'un modèle de langage, d'une méthode d'apprentissage pour le peaufinage (*fine-tuning*), et d'un outil d'explication, est appliquée à une tâche de classification d'articles de presse en français, qui consiste à prédire si un article appartient au genre journalistique de l'opinion (éditoriaux, chroniques…) ou de l'information (dépêches, nouvelles…). Il s'agit d'une sous-tâche du champ de l'analyse d'opinions en TAL, considérée comme particulièrement complexe (Ravi et Ravi, 2015), qui devient de plus en plus cruciale face aux nouveaux modes de partage d'information sur le web et les réseaux sociaux, et au vu de la polarisation grandissante de la société.

Nous insistons (et reviendrons en conclusion) sur le fait que notre affirmation se limite à l'observation qu'en présence d'explications dépendant de facteurs aléatoires, il est donc nécessaire de caractériser cet aléa, et que cette caractérisation peut influencer certains critères désirables des explications d'un modèle. En guise de première étape dans cette direction, nous proposons une caractérisation visuelle se basant sur des

2. Ces notions d'équivalence et de concordance pourraient également être définies pour d'autres métriques de performance, sans modifier les conclusions générales de l'article.

boîtes à moustache (*box-plots*). Ces dernières mettent en évidence que l'aléa du processus d'apprentissage a un impact négatif sur la minimalité des explications, qui est parfois présentée comme un critère désirable supplémentaire (Miller, 2019). Suivant le principe du rasoir d'Ockham, ce critère suggère que parmi différentes explications, la plus simple est souvent la meilleure. Nous insistons en outre sur le fait que notre affirmation se limite à une combinaison raisonnable d'une méthode d'apprentissage et d'un outil d'explication appliquée à une tâche spécifique. Il est dès lors possible que d'autres combinaisons permettent de diminuer cette sensibilité à l'aléa ou que d'autres tâches y soient intrinsèquement moins sujettes. Enfin, si nous affirmons que caractériser la sensibilité à l'aléa des décisions de modèles équivalents est une condition nécessaire à leur explicabilité, nous n'affirmons pas que l'impact de cette sensibilité est positif ou négatif pour d'autres critères désirables des explications de ces décisions, comme leur plausibilité. Ces précisions seront aussi discutées en conclusion.

Constatant la sensibilité à l'aléa des explications du modèle CamemBERT, nous explorons ensuite une méthode plus traditionnelle de TAL basée sur des traits textuels, que nous utilisons pour entraîner un modèle de régression logistique. De telles méthodes de classification sont habituellement reconnues comme étant plus faciles à expliquer (Gémes *et al.*, 2021). Elles sont cependant limitées par des exactitudes plus faibles pour de nombreuses applications et ont de ce fait tendance à être délaissées en faveur de techniques utilisant l'apprentissage profond (*deep learning*) (Li *et al.*, 2020). Elles correspondent donc à un compromis très différent entre exactitude et explicabilité. Concrètement, nous utilisons des cartes d'attention linguistiques qui permettent de visualiser les explications d'un modèle basé sur des traits dans un format similaire à celui fourni par LRP, en attribuant une certaine pertinence à chaque *token* (mot ou signe de ponctuation) d'un texte classé par ce modèle. De façon peu surprenante, nous observons que ce type de modèle basé sur des traits textuels mène à des prédictions ayant une exactitude légèrement réduite, mais que son entraînement converge vers une solution unique qui mène à des explications identiques pour un texte donné.

À partir de cette observation, et bien que plusieurs travaux de recherche aient proposé d'insérer des traits théoriques dans les modèles *transformers* afin d'améliorer leur potentiel d'explicabilité (Koufakou *et al.*, 2020 ; Polignano *et al.*, 2022), nous proposons à l'inverse d'enrichir notre modèle basé sur des traits linguistiques au moyen d'une série de nouveaux traits extraits des explications dérivées d'un modèle *transformer*. En utilisant cette approche, nous montrons qu'il est possible d'améliorer l'exactitude du modèle linguistique (qui reste néanmoins inférieure à celle des modèles *transformers*), tout en conservant des résultats déterministes et donc une invariabilité des explications pour une prédiction donnée. Cette approche hybride suggère au minimum un intérêt des grands modèles de langage dans des tâches exploratoires, par exemple pour identifier des hypothèses de travail à confirmer par un travail d'analyse inductive. Elle laisse par contre ouvert le problème fondamental de l'explicabilité de ces grands modèles, nécessaire en vue d'une utilisation plus automatisée de ceux-ci.

Nous mentionnons enfin des travaux complémentaires qui étudient la sensibilité d'explications aux hyperparamètres (choisis aléatoirement ou heuristiquement) utili-

sés dans les méthodes d'explication elles-mêmes (Bansal *et al.*, 2020). Cette sensibilité est présentée comme préjudiciable de façon plus générale, car elle implique un caractère imprévisible des explications pour un modèle et une prévision donnés. Ces travaux se distinguent néanmoins du nôtre, qui s'intéresse au caractère aléatoire des hyperparamètres d'entraînement du modèle plutôt qu'à celui des méthodes d'explication. Nous mentionnons également la note récente de Bethard (2022) qui met en évidence différents types d'usage (justifiés ou risqués) de l'aléa d'entraînement. Cette discussion générale n'est néanmoins pas liée à la question de l'explicabilité.

2. État de l'art

2.1. *Méthodes d'explication*

Pour les modèles de classification de textes, les méthodes d'explication existantes se divisent en deux catégories, selon leur portée (Danilevsky *et al.*, 2020) : les méthodes d'explication globales, qui visent à expliquer le raisonnement du modèle pour classer n'importe quel document, et les méthodes d'explication locales, qui se concentrent sur le raisonnement du modèle pour une prédiction donnée. Les méthodes d'explication locales (qui sont les plus pertinentes pour nos investigations) se divisent en outre en différentes sous-catégories, suivant qu'elles se basent sur la similarité avec d'autres exemples, sur l'analyse de la structure interne des modèles, des mécanismes de rétropropagation ou une analyse contre-factuelle (Lyu *et al.*, 2022). Dans cet article, nous nous intéressons aux méthodes basées sur des mécanismes de rétropropagation cherchant à interpréter les couches d'attention utilisées par les modèles de type *transformer* (Kovaleva *et al.*, 2019 ; Clark *et al.*, 2019). Bien que le débat concernant le fait que l'attention seule puisse être utilisée comme source valide d'explication reste ouvert (Bibal *et al.*, 2022), des travaux récents ont montré que la combinaison de plusieurs couches d'attention (selon les gradients du modèle) permet de générer des explications convaincantes (Srinivas et Fleuret, 2019 ; Abnar et Zuidema, 2020). Parmi celles-ci, nous utilisons la méthode d'explication LRP, basée sur la propagation de pertinence couche par couche (Chefer *et al.*, 2021), qui, malgré des défauts inhérents aux méthodes d'explication basées sur la rétropropagation[3], apparaît comme une des plus fidèles (Arras *et al.*, 2017) et constitue donc un bon point de départ pour analyser la sensibilité à l'aléa des explications de modèles de type *transformer*.

Dans le cadre de la classification de textes, la méthode LRP explique les prédictions faites par les modèles basés sur l'apprentissage profond en évaluant l'importance de chaque *token* dans le texte. Cette importance est mesurée en suivant, couche par couche, la contribution du *token* évalué à la prédiction du modèle (Bach *et al.*, 2015). Ce processus permet d'assigner un score de pertinence à chaque *token* en partant de la valeur de sortie et en effectuant une rétropropagation via des contraintes de

3. Elles sont par exemple incapables d'expliquer l'influence d'information au-delà du niveau des *tokens*, comme de l'information syntaxique ou des dépendances à long terme.

conservation[4]. Différentes règles définissent comment la pertinence d'un *token* pour une couche du modèle doit être distribuée vers la précédente, avec la contrainte que les scores de pertinence doivent s'additionner à 1 à chaque couche pour aboutir à la prédiction finale. Étant donné que la contrainte de propagation est plus difficile à satisfaire pour certaines couches, cette méthode est constamment améliorée (Binder *et al.*, 2016 ; Voita *et al.*, 2019). Les explications qui en ressortent sont habituellement considérées comme plus fidèles au raisonnement du modèle que d'autres méthodes d'explication, comme celles basées sur la perturbation (Arras *et al.*, 2017).

2.2. *Visualisation des explications*

La visualisation des explications influence leur plausibilité vis-à-vis des lecteurs humains (Reif *et al.*, 2019). L'une des approches les plus populaires pour expliquer les prédictions d'un modèle de classification de textes est l'utilisation de cartes d'attention (Li *et al.*, 2016). Celles-ci consistent à mettre en évidence dans le texte, en utilisant différentes nuances de couleur, les *tokens* (mots ou signes de ponctuation) qui ont été les plus influents pour la décision du modèle. Les cartes d'attention sont donc limitées à l'explication locale au niveau des *tokens* et ne sont pas capables de mettre en évidence l'influence d'autres types de traits potentiellement déterminants dans la prédiction du modèle, comme les dépendances à long terme et les relations entre différents éléments d'explication. Le format des cartes d'attention, qui permet de visualiser l'importance de chaque *token* séparément (*token-level attention*) possède néanmoins l'avantage d'être *a priori* très compréhensible par un lecteur humain, ce qui contribue à la plausibilité des explications. Dans la suite de cet article, nous utilisons le terme « attention » pour renvoyer à l'importance attribuée à un *token* dans une explication produite par n'importe quelle méthode pour n'importe quel modèle, indépendamment du fait que celui-ci soit basé sur le mécanisme d'attention ou non.

2.3. *Subjectivité en journalisme*

Dans le champ journalistique, la notion d'objectivité se trouve historiquement au cœur de nombreux débats (Schudson, 2001). Depuis la fin du XX^{e} siècle, l'objectivité est considérée comme l'une des valeurs les plus importantes de la profession. Elle est souvent perçue par les journalistes comme un « idéal structurant » vers lequel ils doivent s'efforcer de tendre, même si beaucoup reconnaissent que l'objectivité journalistique totale est inatteignable (Lagneau, 2002). La subjectivité inhérente au processus journalistique est liée aux opérations inévitables de sélection et de prise de décision qui imprègnent chaque étape du processus éditorial de transmission de l'information : choisir une histoire, décider de son format, donner la priorité à certains

4. Pour obtenir une valeur par *token*, lisible en langage naturel, et pas par partie de *token* (ou *wordpiece*), comme initialement encodé par l'architecture de CamemBERT, nous concaténons les différentes parties et calculons la moyenne de la somme de leurs valeurs d'attention.

articles par rapport à d'autres, etc. (Tong et Zuo, 2021). De nombreuses décisions subjectives sont également prises lors de l'écriture de l'article, par exemple en ce qui concerne la manière de cadrer les sujets, l'ordre des citations ou le choix des mots. La présentation de faits est toujours influencée par l'interprétation personnelle de ces faits par l'auteur, guidé par son point de vue et ses expériences, ce qui complexifie la quête d'objectivité dans le reportage d'information (Muñoz-Torres, 2012).

Pour cette raison, les journalistes sont formés à exploiter divers outils stylistiques afin d'apparaître aussi objectifs que possible dans leurs articles, en suivant ce que Tuchman (1972) appelle le « rituel stratégique de l'objectivité ». Ils appliquent une variété de mécanismes de neutralisation de la subjectivité, qui atténuent ou dissimulent l'influence des opinions du journaliste sur le texte de l'article (Koren, 2004). Ces recommandations, enseignées dans les manuels de journalisme, imposées dans les salles de rédaction ou corrigées par les relecteurs, incluent la citation systématique des sources d'information, l'utilisation de phrases impersonnelles et d'un lexique neutre, et l'absence du langage figuré dans les textes (Charaudeau, 2006). Cependant, cette recherche d'objectivité textuelle s'applique uniquement aux articles appartenant aux genres de l'information, comme les dépêches d'agences de presse, les distinguant des genres de l'opinion, comme les éditoriaux ou les chroniques (Grosse, 2001).

2.4. *Classification de textes journalistiques d'opinion en TAL*

En TAL, les techniques d'écriture utilisées par les journalistes pour donner à leurs articles un « masque d'objectivité » peuvent être utilisées pour classer les textes dans les genres de l'information ou de l'opinion. Wiebe *et al.* (2004) considèrent la subjectivité linguistique comme un continuum, dans lequel « les phrases objectives sont des phrases sans expressions significatives de subjectivité », et cherchent à identifier les éléments potentiellement subjectifs dans des textes en anglais. Les textes contenant peu de ces éléments subjectifs sont considérés comme non subjectifs ou neutres (Riloff *et al.*, 2005). Au fil des années, plusieurs marqueurs de subjectivité dans les articles de presse dans différentes langues ont été analysés et évalués suivant différentes approches. Krüger *et al.* (2017) ont utilisé une série de vingt-huit traits linguistiques, comme la complexité lexicale ou le taux de chiffres présents dans le texte, pour classer les articles d'opinion et d'information publiés par des journaux américains, soulignant la force prédictive de certains traits pour cette tâche de classification. Une étude similaire (sur laquelle nous nous basons) a été menée sur un corpus d'articles en français, évaluant trente traits et en combinant dix-neuf d'entre eux pour construire un modèle de classification de textes d'opinion et d'information (Escouflaire, 2022).

De nos jours, ces approches traditionnelles basées sur des traits linguistiques sont largement délaissées au profit des grands modèles de langage de type *transformer* (Vaswani *et al.*, 2017). Fondé sur l'architecture du modèle RoBERTa (Liu *et al.*, 2019), CamemBERT (Martin *et al.*, 2020) est un modèle *transformer* entraîné sur un corpus de textes en français, qui surpasse les méthodes précédentes dans diverses tâches, notamment des tâches de classification de textes (Bailly *et al.*, 2021 ; Chenais

et al., 2021). Ces modèles nécessitent cependant beaucoup plus de ressources calculatoires que les modèles traditionnels basés sur des traits (Cunha *et al.*, 2021) et possèdent une plus grande complexité architecturale. Bien que les modèles *transformers* aient été utilisés avec succès pour différentes tâches liées au domaine du journalisme, telles que la détection de *fake news* (Vargo *et al.*, 2018 ; Zellers *et al.*, 2019), il n'existe pas à notre connaissance d'études concernant la classification d'articles d'opinion et d'information avec des modèles *transformers*, en particulier en français.

3. Méthodologie

3.1. *Corpus*

Nous utilisons le corpus RTBF-InfOpinion de Bogaert *et al.* (2023), qui contient 10 000 articles de presse français publiés entre 2012 et 2021 sur le site web de la RTBF (Radio-télévision belge francophone, `www.rtbf.be`), le média de service public belge francophone. Ce corpus a été constitué à partir du corpus RTBF en libre accès (Escouflaire *et al.*, 2023), en sélectionnant 5 000 articles identifiés comme des articles d'opinion par leurs auteurs ou par le média, et 5 000 articles d'information appartenant aux catégories « Belgique », « Monde » et « Société » du site web de la RTBF, traitant de sujets similaires à ceux discutés dans les articles d'opinion. Le corpus RTBF-InfOpinion est donc divisé en deux classes équilibrées : *information* et *opinion*. Il contient un total de 5 323 166 *tokens*. En moyenne, les articles d'opinion contiennent 705 *tokens*, contre 360 pour les articles d'information. Le corpus RTBF-InfOpinion a en outre été divisé en des ensembles d'entraînement (80 %), de validation (10 %) et de test (10 %), tous équilibrés entre les deux classes.

Afin d'évaluer la robustesse des modèles face au changement de données, nous avons constitué un second corpus, composé d'articles de presse publiés par un autre média, Le Soir (`www.lesoir.be`), qui est le quotidien le plus populaire en Belgique francophone. Ce corpus sert uniquement d'ensemble de test pour les modèles. Nous l'avons construit en suivant la méthodologie et le prétraitement présentés par Bogaert *et al.* (2023). Le corpus contient 1 000 articles, publiés en ligne entre 2015 et 2021. Tout comme le jeu de données RTBF-InfOpinion, le corpus LeSoir-InfOpinion est composé à 50 % d'articles d'opinion et à 50 % d'articles d'information, suivant les mêmes catégories éditoriales que celles choisies pour le corpus RTBF. Il contient 669 154 *tokens* au total, pour une moyenne de 859 *tokens* par article d'opinion contre 480 par article d'information et est disponible par simple demande aux auteurs.

3.2. *Modèle transformer peaufiné*

Le premier modèle utilisé dans nos expériences est le modèle CamemBERT (Martin *et al.*, 2020), dans sa version de base : il est insensible aux majuscules et contient 110 millions de paramètres. CamemBERT est basé sur l'architecture du modèle RoBERTa (Liu *et al.*, 2019) et a été préentraîné sur la partie française du corpus

OSCAR (138 Go de texte). CamemBERT a été préféré à FlauBERT (Le *et al.*, 2020), un autre modèle de langue française de grande taille, en raison de sa compatibilité architecturale avec la méthode d'explication LRP de Chefer *et al.* (2021) que nous utilisons pour produire des explications. Nous opérons nous-même un peaufinage de ce modèle préentraîné. Ce peaufinage est effectué en entraînant une tête de classification constituée d'une couche dense et d'une couche permettant de récupérer deux valeurs de sortie, associées à la prédiction du modèle. Ces couches ont une fonction d'activation en tangente hyperbolique et un mécanisme de décrochage (*dropout*).

Les hyperparamètres qui influencent le peaufinage du modèle sont la vitesse d'apprentissage (*learning rate*), la taille de lot (*batch size*) et le nombre d'époques (*epochs*). Nous avons empiriquement évalué plusieurs combinaisons de valeurs pour chacun de ces paramètres, sélectionnant ensuite les paramètres optimaux selon la précision obtenue par le modèle sur l'ensemble de validation et pour la graine aléatoire 0. Pour la vitesse d'apprentissage, nous avons testé des valeurs allant de 1×10^{-6} à 1×10^{-4} et avons sélectionné la valeur 2×10^{-5}. Pour la taille de lot, nous avons essayé des valeurs allant de 1 à 64 et avons sélectionné la valeur 4. Pour le nombre d'époques, nous avons évalué la précision obtenue en entraînant le modèle durant une à quatre époques, et avons décidé d'entraîner le modèle durant deux époques.

Les éléments aléatoires de l'entraînement du modèle CamemBERT que nous étudions concernent exclusivement le peaufinage. Ils sont régis par une graine aléatoire utilisée pour l'optimisation, qui influence (*i*) l'initialisation des poids de la tête de classification, (*ii*) l'ordre des textes dans l'ensemble d'entraînement, et (*iii*) les neurones qui sont visés par la technique de décrochage visant à limiter le surapprentissage (*overfitting*). Nous n'avons pas modifié ce dernier paramètre et l'avons laissé à sa valeur par défaut (10 %). Le peaufinage du modèle pour la tâche de classification a été effectué sur l'ensemble d'entraînement RTBF-InfOpinion pendant deux époques. L'exactitude du modèle est évaluée à chaque époque sur l'ensemble de validation, et à la fin du peaufinage sur les ensembles de test (RTBF- et LeSoir-InfOpinion).

3.3. *Modèle basé sur des traits*

Le second modèle utilisé dans nos expériences est un classifieur utilisant dix-neuf traits textuels issus de l'état de l'art sur la subjectivité linguistique, et identifiés comme des prédicteurs efficaces de l'opinion dans le discours de presse francophone (Escouflaire, 2022). La plupart de ces indicateurs reposent sur la présence ou la proportion de certains *tokens* ou types de *tokens* dans le texte à classer : adjectifs, verbes, pronoms de la première personne et déterminants, pronoms relatifs, le pronom indéfini « on », signes de ponctuation expressifs (points-virgules, points d'exclamation et points d'interrogation), guillemets, chiffres, mots de négation, mots de plus de sept caractères, mots apparaissant dans le lexique de New *et al.* (2004) ou le lexique NRC (Mohammad et Turney, 2013). Seuls deux traits ne sont pas liés aux *tokens* et s'appliquent à l'ensemble de l'article : le rapport lexique-occurrences (ratio *type-token*) corrigé de Carroll (Carroll, 1964) et la longueur moyenne des mots.

Le classifieur est basé sur une régression logistique. Nous avons utilisé une recherche par grille pour sélectionner la combinaison d'hyperparamètres optimale pour la régression. Nous utiliserons l'abréviation LING-LR pour ce modèle de régression logistique basé sur des traits linguistiques. Dans cet l'article, il servira d'exemple de modèle menant à des explications stables, que nous chercherons à enrichir en section 4.

3.4. *Méthodes d'explication*

Pour générer et visualiser les explications des modèles de langage utilisés, nous choisissons le format des cartes d'attention au niveau des *tokens*, qui permet de produire des explications plausibles pour des lecteurs humains (Sen *et al.*, 2020). Pour chaque modèle, nous avons utilisé une méthode d'explication pour produire des explications au niveau des *tokens* sous forme de cartes d'attention : la propagation de pertinence couche par couche (LRP), qui cherche à identifier l'attention déployée par le modèle *transformer* CamemBERT peaufiné, et une méthode de création de « cartes d'attention linguistique » (CAL) pour le modèle basé sur des traits (LING-LR). Des illustrations des deux types de cartes d'attention sont présentées dans la figure 1.

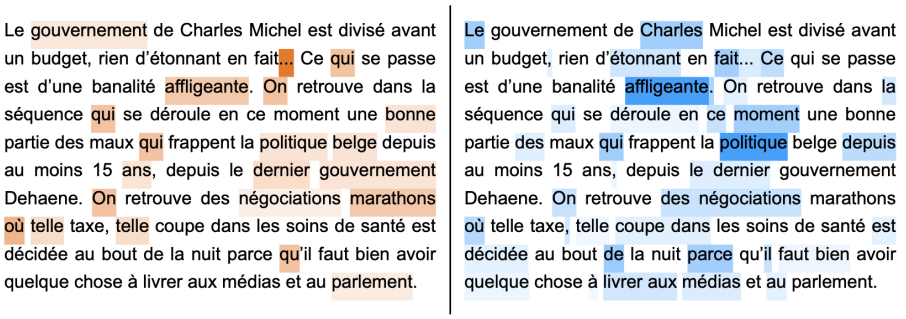


FIGURE 1. *Cartes d'attention issues du modèle linguistique (à gauche, en orange) et CamemBERT (à droite, en bleu) pour un même article du corpus de validation. Les deux modèles classent correctement l'article dans la catégorie opinion.*

3.4.1. *Propagation de pertinence couche par couche (LRP)*

En vue de produire les explications à partir des prédictions du modèle CamemBERT, nous utilisons la méthode LRP de propagation de pertinence par couche décrite en section 2.1. Elle permet d'associer à chaque *token* d'un texte une valeur d'attention selon son influence dans la prédiction faite par le modèle expliqué, et de visualiser l'explication sous forme de carte d'attention. Nous utilisons la version de LRP développée par Chefer *et al.* (2021) pour l'interface de BERT, adaptée pour la rendre compatible avec l'interface de RoBERTa (sur laquelle se base CamemBERT).

3.4.2. *Cartes d'attention linguistique (CAL)*

La cartographie d'attention linguistique est une méthode que nous avons conçue dans le cadre de ce travail, dans l'objectif de produire des explications, lisibles au niveau des *tokens*, des prédictions de notre modèle linguistique. Les CAL permettent de visualiser l'importance accordée à chaque *token* dans un texte classé par ce modèle (ou par n'importe quel modèle basé sur des traits mesurés à partir du texte). Selon la classe prédite par le modèle pour un article, cette méthode produit une carte d'attention qui met en évidence les *tokens* qui contribuent le plus aux traits les plus déterminants pour la classe prédite. Pour le modèle LING-LR, l'importance de chaque *token* est calculée à partir des coefficients de régression accordés à chaque trait auquel le *token* est associé. Un *token* est surligné dans une couleur plus foncée s'il est associé à un ou plusieurs traits linguistiques et si les poids de ces traits dans le modèle sont élevés. Ainsi, l'attention linguistique attribuée à un mot *i* pour un trait *j* peut s'écrire :

$$A_{ij} = \begin{cases} \frac{w_{ij}}{\sum_i w_{ij}} \times T_j & \text{si signe}(T_j) = \text{signe(prediction)}, \\ 0 & \text{dans le cas contraire,} \end{cases} \quad [1]$$

où le premier terme représente l'importance relative du mot *i* au sein du trait *j* et le second représente le coefficient associé au trait *j* dans la régression[5]. Enfin, l'attention attribuée à un mot *i* pour tous les traits combinés se moyenne comme $A_i = \sum_j A_{ij}$.

Les CAL génèrent des explications comparables à celles générées par LRP (section 3.4.1). Vu qu'elles représentent l'importance au niveau des *tokens*, elles ne peuvent pas refléter l'importance des traits linguistiques globaux (dont la mesure ne dépend pas de certains *tokens* spécifiques). Le modèle LING-LR contenant deux traits globaux (le rapport lexique-occurrences et la longueur moyenne des mots), ce défaut nuit dès lors à la fidélité des explications. Nous supposons néanmoins que cette fidélité est au moins aussi bonne que dans le cas des explications de la méthode LRP appliquée au modèle CamemBERT où le même problème (l'explication d'un modèle avec une méthode qui ne peut pas refléter toute sa complexité) se pose de façon accrue.

4. Sensibilité des explications à l'aléa de l'entraînement

Dans cette section, nous étudions la sensibilité des explications du modèle CamemBERT peaufiné aux éléments aléatoires de son entraînement. Pour ce faire, nous commençons par mettre en évidence l'existence d'ensembles non négligeables de modèles équivalents en section 4.1. Nous étudions ensuite la corrélation entre les explications obtenues pour un même texte avec des modèles qui ne diffèrent que par l'aléa utilisé pour les optimiser en section 4.2. Nous poursuivons avec une caractérisation visuelle de l'impact de cet aléa sur les explications en section 4.3. Nous proposons enfin une analyse qualitative de quelques cartes d'attention choisies afin d'illustrer les

5. Pour les traits catégoriques (adjectifs, verbes …), ce terme vaut 0 ou $\frac{1}{n}$, mais il peut différer de ces cas limites lorsqu'on considère des variables continues (imageabilité, concrétude …).

évaluations qui précèdent. Nous comparons en outre les résultats obtenus avec ceux du modèle LING-LR lorsque cette comparaison peut éclairer nos discussions.

4.1. *Génération de modèles statistiquement équivalents*

Afin de générer des modèles équivalents, nous avons d'abord répété le peaufinage du modèle CamemBERT 200 fois (avec 200 graines aléatoires différentes), chaque entraînement se basant sur le même ensemble de 8 000 textes. Nous avons ensuite estimé l'exactitude sur l'ensemble de test pour la totalité des modèles et pour des sous-ensembles de modèles, en choisissant les modèles menant aux exactitudes les plus proches ou les plus élevées. Nous avons enfin calculé un paramètre ϵ qui correspond à la différence entre l'exactitude du modèle le plus exact (a) et celle du modèle le moins exact (b) des ensembles étudiés, afin de déterminer si cette différence est significative. Pour ce faire, nous avons estimé la statistique z (Lehmann et Romano, 2008), qui permet de déterminer si deux proportions (ici, les exactitudes) diffèrent :

$$z = \left| \frac{a - b}{\sqrt{\frac{\frac{a+b}{2} * (1 - \frac{a+b}{2})}{n}}} \right| . \quad [2]$$

Nous avons supposé que les différences d'exactitude sont significatives pour des z supérieurs à 1,96, ce qui correspond à une valeur-p inférieure à 0,025. Pour des z inférieurs (et des valeurs-p supérieures), nous concluons par contre que les exactitudes des modèles ne diffèrent pas de façon significative. Nous définissons dès lors les modèles dans les ensembles correspondants comme statistiquement équivalents, car leur exactitude ne permet pas de préférer un de ces modèles à un autre.

	Exact. min.	Exact. max.	ϵ	Valeur-p
CamemBERT(200 modèles)	93,1	96,6	3,50	$2,8 \times 10^{-7}$
CamemBERT(150 plus proches)	94,5	95,9	1,40	0,0191
CamemBERT(150 plus exacts)	95,0	96,6	1,60	0,0860
CamemBERT(100 plus proches)	95,0	95,7	0,70	0,1466
CamemBERT(100 plus exacts)	95,4	96,6	1,20	0,3227
CamemBERT(50 plus proches)	95,3	95,6	0,30	0,3245
CamemBERT(50 plus exacts)	95,7	96,6	0,90	0,5616
LING-LR	88,9	88,9	0	/

TABLEAU 1. *Exactitude minimale et maximale, paramètre ϵ et valeur-p*

Les résultats de ces estimations sont rapportés dans le tableau 1[6]. Ils montrent qu'à partir des 200 modèles générés, nous avons pu générer des sous-ensembles de 100 modèles équivalents. La diminution du paramètre ϵ avec le nombre de modèles conservés

6. Celui-ci donne en outre l'exactitude du modèle LING-LR estimé et testé avec les mêmes textes. L'entraînement de ce modèle converge vers une solution unique et le paramètre ϵ est nul.

étant mécanique, elle ne sert qu'à illustrer la complexité de la tâche d'identification de modèles équivalents. À cet égard, il faut noter que cette définition de modèles équivalents dépend de la taille de l'ensemble de test : si cette dernière augmente, les exactitudes des modèles seront estimées avec plus de précision et des différences plus petites seront considérées comme significatives. Ceci implique qu'il faudra générer plus de modèles pour atteindre un ϵ qui n'est pas significatif. La quantité d'aléa utilisée pour l'optimisation d'un modèle étant pratiquement illimitée, cette observation ne change pas fondamentalement le problème de sensibilité dont nous discutons. Elle augmente uniquement le coût calculatoire pour le mettre en évidence. Les temps de calcul du modèle CamemBERT étant de façon générale significativement supérieurs à ceux du modèle LING-LR (tableau 2), ils peuvent dès lors devenir un problème concret, en particulier si l'explicabilité d'un tel (grand) modèle demande la génération de grands ensembles de modèles équivalents, ce que nous discutons en section 4.5.

Modèle	LING-LR	CamemBERT
Pré-entraînement	15 min.	13 jours*
Entraînement	0.5 sec./fit	4 min./epoch
Prédiction	1.3 sec.	9.6 sec.
Méthode	CAL	*LRP*
Explication	4 sec.	4 min.

TABLEAU 2. *Temps de calcul des modèles LING-LR et CamemBERT estimés sur les 1 000 articles de l'ensemble de test RTBF, sur un serveur GPU NVIDIA RTX A6000*

4.2. *Corrélation des explications*

Ayant identifié des ensembles de modèles équivalents, nous évaluons maintenant dans quelle mesure les explications de ces modèles diffèrent. Pour ce faire, et à titre de première analyse informelle permettant de mettre en évidence de telles différences, nous estimons la corrélation entre les explications obtenues pour les prédictions de modèles équivalents sur des entrées concordantes[7]. Précisément, nous avons choisi deux textes classés comme textes d'information de longueur similaire (49 *tokens* pour le texte 1, 51 *tokens* pour le texte 2), généré 100 explications équivalentes pour chaque texte et construit deux vecteurs correspondant chacun à la concaténation de 50 explications. Nous avons ensuite calculé la corrélation entre ces deux vecteurs. Dans le cas d'explications identiques (ou aléatoires), cette corrélation serait de 1 (ou 0). Nous avons enfin répété cette opération pour 10 000 différents choix de vecteurs, afin de calculer un intervalle de confiance de type *bootstrap* (Efron et Tibshirani, 1993).

Les résultats de ces estimations sont rapportés en tableau 3. On y observe que les explications diffèrent significativement, indépendamment du fait que les sous-ensembles de modèles soient choisis en fonction de la valeur ou de la proximité de

7. L'étude d'entrées presque concordantes serait possible également, et demanderait simplement d'étudier les corrélations pour les décisions *information* et *opinion* séparément.

	Modèle	Corrélation de Pearson	
		Estimation	Intervalle de confiance bootstrap
Texte 1	CamemBERT (200 modèles)	0,1523	[0,0939 ; 0,2098]
	CamemBERT (150 plus proches)	0,1448	[0,0734 ; 0,2162]
	CamemBERT (150 plus exactes)	0,1401	[0,0738 ; 0,2065]
	CamemBERT (100 plus proches)	0,1259	[0,0401 ; 0,2139]
	CamemBERT (100 plus exactes)	0,1492	[0,0707 ; 0,2268]
	CamemBERT (50 plus proches)	0,1176	[0,0116 ; 0,2494]
	CamemBERT (50 plus exactes)	0,1835	[0,0912 ; 0,2803]
Texte 2	CamemBERT (200 modèles)	0,3947	[0,3424 ; 0,4470]
	CamemBERT (150 plus proches)	0,3978	[0,3364 ; 0,4589]
	CamemBERT (150 plus exactes)	0,4086	[0,3505 ; 0,4678]
	CamemBERT (100 plus proches)	0,3801	[0,3055 ; 0,4536]
	CamemBERT (100 plus exactes)	0,4225	[0,3518 ; 0,4946]
	CamemBERT (50 plus proches)	0,3443	[0,2370 ; 0,4519]
	CamemBERT (50 plus exactes)	0,4661	[0,3781 ; 0,5532]

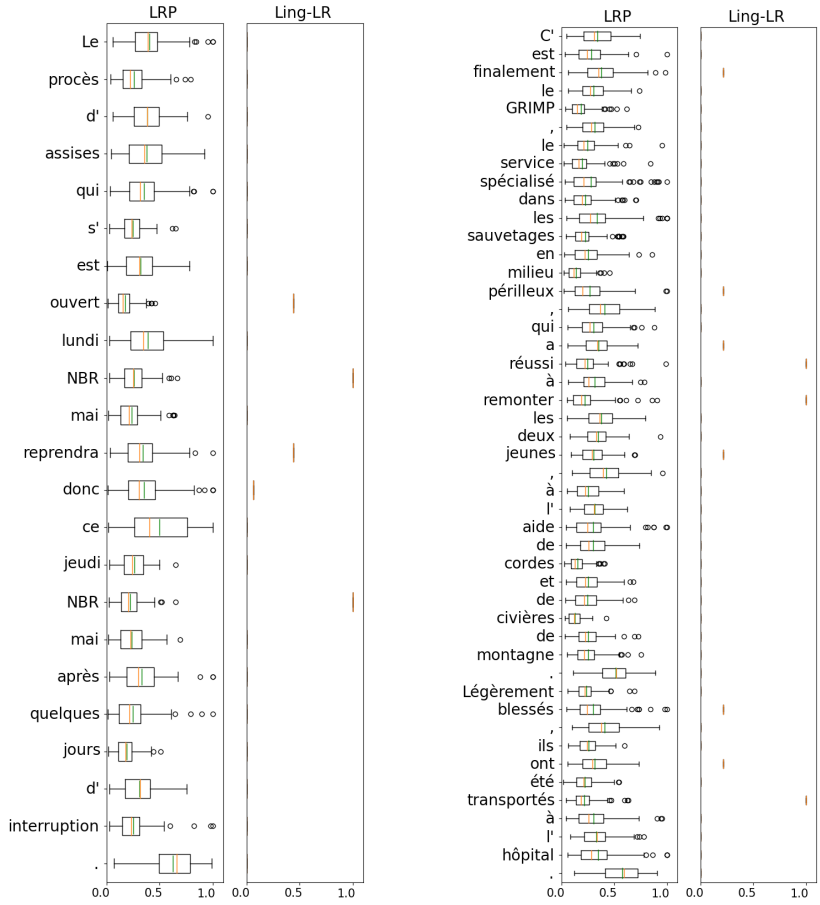
TABLEAU 3. *Corrélation entre les explications de modèles équivalents pour des entrées concordantes et intervalles de confiance de type bootstrap*

leurs exactitudes. De façon intéressante, nous observons aussi que ces corrélations diffèrent selon le texte choisi, suggérant une sensibilité aux données de la dépendance à l'aléa des explications qui confirme l'intérêt de caractériser cette dépendance.

4.3. *Caractérisation visuelle de la sensibilité des explications à l'aléa*

Afin d'illustrer la sensibilité à l'aléa du modèle CamemBERT, nous avons ensuite généré des boîtes à moustache correspondant aux explications de deux textes courts de notre ensemble de test. Elles donnent une intuition sur la fréquence à laquelle les explications de modèles équivalents accordent de l'importance à chaque *token*.

Les résultats de la figure 2 montrent qu'en augmentant le nombre d'explications, la totalité des *tokens* du texte finit par avoir une attention non nulle dans les boîtes à moustache des modèles CamemBERT. Bien que la distribution des *tokens* considérés indépendamment ne soit pas uniforme (et qu'une distribution uniforme des *tokens* considérés indépendamment n'implique pas une distribution uniforme de toutes les explications possibles), ces résultats questionnent la plausibilité des explications obtenues. Par ailleurs, et de façon assez triviale, ils confirment aussi une réduction de la minimalité des explications par rapport aux explications (déterministes) du modèle LING-LR. Cette dernière est reflétée par la distribution des explications plus riche, et potentiellement plus complexe à interpréter, des modèles CamemBERT.



(a) *Explications pour le texte 1* **(b)** *Explications pour le texte 2*

FIGURE 2. *Caractérisation visuelle des explications de 100 modèles équivalents. L'axe des X correspond à la distribution de l'importance des mots.*

4.4. *Analyse qualitative d'exemples choisis*

Enfin, et afin de confirmer que les différences d'explications quantifiées en section 4.2 et caractérisées visuellement en section 4.3 ne se résument pas à une combinaison d'explications concordantes et d'explications aberrantes faciles à filtrer, nous concluons cette section avec quelques exemples de cartes d'attention (figure 3).

Nous observons d'abord que les explications des figures 3a et 3b sont visuellement très similaires pour un lecteur humain. Malgré des différences minimes quant à l'attention attribuée à certains *tokens*, les deux modèles semblent se concentrer sur les mêmes éléments, qui sont principalement des éléments structurants du texte : les

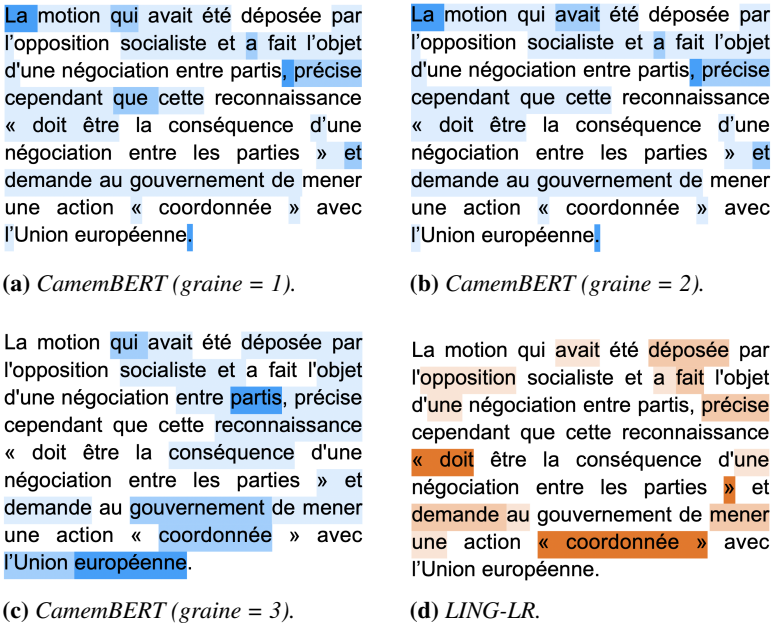


FIGURE 3. *Cartes d'attention de quatre différents modèles pour un article de la classe information (correctement classé par les quatre modèles)*

signes de ponctuation forte (« . », « , »), les conjonctions (« que », « et »), les guillemets et les débuts de propositions. Au contraire, la carte de la figure 3c, produite à partir d'une troisième graine, apparaît comme très différente des deux premières. L'importance y est plutôt accordée à des *tokens* chargés sur le plan sémantique, en particulier les noms représentant les acteurs politiques dont il est question dans le texte (« partis », « gouvernement », « l'Union européenne »). Ces cartes d'attention suggèrent donc que des explications très différentes, dès lors difficiles à filtrer sans une analyse plus approfondie, peuvent être extraites de modèles équivalents. La carte 3d montre enfin que les *tokens* qui influencent le plus le modèle LING-LR sont la présence de guillemets et de verbes. Elle diffère également des explications des modèles CamemBERT.

4.5. *Discussion*

Les résultats de cette section démontrent clairement que les explications de grands modèles de langage peuvent être sensibles à l'aléa utilisé pour leur entraînement. D'une part, nous en concluons que caractériser cette sensibilité est nécessaire, ne fût-ce que pour se convaincre que la distribution des explications diffère suffisamment de l'uniforme. Le cas échéant, le choix d'une explication serait en effet complètement arbitraire. Au mieux de notre connaissance, la caractérisation de cette sensibilité ne fait pas encore l'objet d'études systématiques dans la littérature. D'autre part, ces résultats

posent la question essentielle de déterminer dans quelle mesure cette sensibilité est un problème. Prenant l'exemple de la justice, on pourrait par exemple arriver à une situation où un jugement automatique propose à un justiciable un grand nombre d'explications, assez différentes du point de vue de l'entendement humain mais indistinguables du point de vue algorithmique. Cela semble peu compatible avec l'exigence de compréhensibilité d'un jugement (et le fait de ne montrer qu'une explication au justiciable, même si par hasard elle est plausible, relèverait au moins en partie de l'arbitraire). On pourrait également arriver à une situation où l'on retrouve quelques groupes (*clusters*) d'explications qui correspondent à des interprétations différentes mais plausibles de textes juridiques, ce qui serait moins pathologique. Mais même dans ce cas, il semble difficile de rendre le processus d'explication complètement automatique.

Par ailleurs, la combinaison de ce besoin de caractérisation avec la complexité calculatoire des grands modèles de langage (tableau 2) suggère que des modèles plus simples pourraient devenir des alternatives intéressantes dès lors que l'explicabilité et le coût de calcul des classifications sont considérés comme importants pour une application donnée. Cette observation motive la tentative d'enrichissement de ce type de modèle dans la section qui suit. À ce sujet, nous notons également que le coût de la caractérisation de la sensibilité à l'aléa dépend de la distribution des explications. Par exemple, plus la variance de l'attention accordée à un mot de la figure 2 augmente, plus il faudra générer des explications pour bien estimer son attention moyenne.

5. Enrichissement du modèle basé sur les traits

Nous proposons ici d'améliorer l'exactitude d'un modèle dont l'entraînement converge vers une solution unique (et donc sans variabilité des explications extraites pour un texte donné) au moyen d'éléments dérivés des explications de modèles *transformers*. Nous insérons dans le modèle basé sur des traits linguistiques (LING-LR) de nouveaux traits à partir des explications des modèles CamemBERT peaufinés. Ce modèle hybride sera appelé « modèle linguistique enrichi » (LING-LR-E).

La première étape consiste à mesurer l'attention moyenne accordée à chacun des *tokens* présents au moins dix fois dans l'ensemble des mille cartes d'attention produites en appliquant la méthode LRP à partir des prédictions de modèles CamemBERT équivalents sur les articles de l'ensemble de test RTBF-InfOpinion. Afin d'illustrer la possibilité d'enrichissement avec un temps de calcul raisonnable, nous limitons le nombre de modèles équivalents utilisés à dix. Nous classons ensuite les *tokens* selon leur attention moyenne, en ordre décroissant. Cette opération est répétée pour les dix modèles étudiés. Ensuite, seuls les *tokens* apparaissant parmi les cent premiers *tokens* pour au moins cinq modèles équivalents sur dix sont conservés. Il s'agit enfin d'analyser qualitativement la liste des *tokens* afin d'identifier des motifs linguistiques qui peuvent être convertis en traits. Cette méthode est similaire à l'approche présentée par Zhou *et al.* (2022), qui consiste à dériver de l'information sur le raisonnement interne d'un modèle complexe à partir d'explications locales des prédictions de ce modèle.

Nous limitons notre analyse aux cinquante *tokens* bénéficiant du plus d'attention pour chaque classe. Les deux listes résultantes sont présentées dans le tableau 4.

En examinant qualitativement les cinquante *tokens* de la liste *opinion*, on peut observer plusieurs motifs récurrents : des mots marqués axiologiquement (*désastre*, *pauvre*), des signes de ponctuation expressive (« ... », « ! »), des verbes de pensée (*imaginez*, *oublier*), des mots renvoyant à des concepts abstraits (*idéologie*, *humour*), ou encore des marqueurs de discours (*bref*, *certes*). Pour la liste *information*, on peut plutôt repérer des mots renvoyant à des entités temporelles non déictiques (*lundi*, *GMT*), des verbes de citation (*précise*, *affirmé*), des mots avec une fréquence subjective élevée (*ordinateur*, *aéroport*), à savoir des mots perçus comme fréquents dans le langage quotidien (Balota *et al.*, 2001), et des mots renvoyant à des sources d'information (*selon*, *AFP*). Certains de ces motifs recouvrent des traits déjà présents dans le modèle LING-LR, comme la présence d'adjectifs et de signes de ponctuation expressive (pour la classe *opinion*), d'autres constituent des découvertes originales, comme le nombre de marqueurs de discours et la concrétude (Bonin *et al.*, 2018) moyenne des mots du texte. Enfin, nous notons que certains *tokens* peuvent être considérés comme des artéfacts (Gururangan *et al.*, 2018) liés aux données utilisées (*parking*, *flamandes*).

De cette analyse, nous extrayons neuf nouveaux traits linguistiques à partir des motifs identifiés dans les listes d'attention du tableau 4 : les taux relevés dans le texte de marqueurs temporels déictiques, de marqueurs temporels non déictiques, de verbes de pensée, de verbes de citation, de verbes au passif, et de marqueurs de discours, ainsi que la concrétude, l'imageabilité, et la fréquence subjective moyenne des mots du texte. La concrétude est mesurée à partir du lexique de Bonin *et al.* (2018), tandis que l'imageabilité et la fréquence subjective sont mesurées à partir des lexiques de Desrochers et Thompson (2009). L'enrichissement avec ces neuf nouveaux traits (ajoutés aux dix-neuf traits du modèle LING-LR original) permet d'atteindre avec LING-LR-E une exactitude de 89,6 % sur l'ensemble de test RTBF-InfOpinion, soit une augmentation de 0,8 % par rapport à LING-LR (qui n'est pas statistiquement significative). Sur l'ensemble de test LeSoir-InfOpinion, LING-LR-E atteint une exactitude de 80,6 %, contre 76,8 % pour LING-LR, soit une augmentation de 3,8 % (significative selon la même statistique z et la même valeur-p de 1,96 que dans les sections précédentes) qui montre que les éléments extraits des explications des modèles CamemBERT contribuent à rendre le modèle LING-LR-E plus généralisable. En comparaison, la meilleure exactitude atteinte par un modèle CamemBERT sur l'ensemble de test LeSoir-InfOpinion est de 90,5 % (96,6 % sur l'ensemble de test RTBF).

6. Limitations

La contribution principale de cet article est la mise en évidence d'une question (la sensibilité des explications à l'aléa des grands modèles de langage est-elle significative ?) qui nous semble trop peu étudiée à ce stade alors qu'elle cristallise la difficulté d'expliquer les prédictions de ces modèles. Nous montrons que cette question peut se poser en pratique pour une certaine combinaison d'une méthode d'apprentissage et

Information		Opinion	
précise	indiqué	révélations	fed
jeudi	mardi	chômeurs	imaginez
indique	expliqué	bref	désigne
mercredi	explique	ressemble	pension
lundi	vendredi	fout	extension
précisé	a-t-elle	...	latin
poursuit	souligne	formateur	aiment
adaptation	souligné	Hitler	inverse
ajouté	dimanche	tort	disons
parking	poursuivi	aujourd'hui	ombre
conclu	AFP	accords	bulle
suspect	correctionnel	démontrer	mélange
ajoute	aéroport	politiquement	libéral
samedi	affirmé	désastre	dépit
trafic	assuré	flamandes	chômage
locales	affirme	suffisamment	correction
chanteuse	selon	pire	illustre
priorités	températures	retraite	idéologie
déclaré	a-t-il	médiatique	immobilier
disponible	rappelé	voyons	;
février	ordinateur	!	oublier
communiqué	organisateurs	pauvre	monétaire
blessé	km	certes	utilise
GMT	pourront	mauvais	impôt
dépistage	visiteurs	calcul	inutile

TABLEAU 4. *Liste des cinquante tokens (de gauche à droite puis de haut en bas) avec le plus d'attention en moyenne dans les vecteurs d'explications des prédictions des modèles CamemBERT sur l'ensemble de test RTBF-InfOpinion.*

d'un outil d'explication appliquée à une tâche spécifique. Il en découle que la généralité de nos observations mériterait d'être étendue. D'une part, l'étude d'autres corpus (notamment dans d'autres langues que le français) serait intéressante, tout comme l'étude d'autres tâches comme la détection de *fake news*, qui est également reconnue comme particulièrement complexe (Vargo *et al.*, 2018 ; Zellers *et al.*, 2019). D'autre part, et au niveau technique, l'évaluation d'autres modèles de langage et d'autres méthodes d'explications serait nécessaire aussi. Nous donnons des motivations supplémentaires pour ces différentes extensions dans la conclusion qui suit.

7. Conclusion et problèmes ouverts

Dans cet article, nous avons étudié la sensibilité des explications de grands modèles de langage aux éléments aléatoires de leur entraînement. Plus précisément, nous

avons montré que des modèles optimisés avec le même ensemble d'entraînement mais avec différents hyperparamètres aléatoires, et qui offrent une exactitude similaire, peuvent donner des explications différentes pour des textes sur lesquels ils donnent la même prédiction. En d'autres termes, nous avons observé des explications qui dépendent de la structure des modèles générés à partir de différents paramètres aléatoires plutôt que du résultat de leurs prédictions. Rien ne permettant de préférer un modèle à un autre dans ce contexte, nous en concluons que se limiter à l'explication d'un seul modèle est alors insuffisant et peut relever de l'arbitraire, et affirmons que l'explication des décisions de ce type de modèles nécessite de caractériser leur part aléatoire.

Observant qu'une première caractérisation de la dépendance à l'aléa des explications de grands modèles de langage, utilisant des boîtes à moustache, diminue leur minimalité (Miller, 2019), nous avons ensuite évalué dans quelle mesure un modèle plus simple permet des explications plus compactes et plus stables. Dans cette optique, nous avons d'abord constaté que, de façon peu surprenante, un modèle basé sur des traits linguistiques combiné avec une régression logistique produit des explications qui ne présentent effectivement pas la sensibilité à l'aléa du modèle CamemBERT peaufiné, au prix d'une exactitude réduite. Observant en outre, grâce aux cartes d'attention des deux modèles, qu'ils ne semblent pas s'appuyer sur les mêmes *tokens* pour prédire les mêmes classes, nous avons ensuite essayé d'enrichir le modèle linguistique. Pour ce faire, nous avons cherché à extraire de nouveaux traits à partir des cartes d'attention calculées sur le modèle CamemBERT peaufiné (le modèle le plus précis), et de les intégrer au modèle de régression logistique basé sur des traits linguistiques (le modèle le plus stable). Cette méthode a permis d'améliorer l'exactitude de classification de ce modèle sur deux ensembles de test, sans diminuer la stabilité de ses explications.

Le modèle linguistique amélioré conservant une exactitude inférieure au modèle CamemBERT peaufiné, nous en concluons néanmoins aussi que les grands modèles de langage caractérisent de façon utile des caractéristiques des textes à classer qui ne sont pas intégrées (et probablement pas intégrables) au modèle basé sur des traits. Le problème fondamental de l'explicabilité des décisions des grands modèles de langage reste donc ouvert, amplifié par sa dépendance à l'aléa d'entraînement mise en évidence dans cet article. Nos résultats suggèrent dès lors de nouvelles pistes de recherche tournées vers l'identification de l'origine de cette sensibilité à l'aléa et sa diminution.

Une première piste de recherche serait d'évaluer l'impact de la sensibilité à l'aléa mise en évidence sur la plausibilité des explications. En effet, affirmer que la dépendance à l'aléa des explications d'un modèle doit être caractérisée n'implique pas forcément une diminution de la plausibilité. En particulier pour la tâche de détection d'opinion étudiée, il se pourrait que la variabilité observée reflète la variabilité des explications que donneraient des annotateurs humains. La mise en place d'une telle expérience d'annotation humaine serait donc intéressante. Étudier dans quelle mesure les explications de modèles équivalents peuvent être groupées en *clusters*, comme mentionné en section 4.5, serait particulièrement pertinent dans cette optique.

Une autre piste de recherche serait d'évaluer dans quelle mesure cette sensibilité à l'aléa provient d'un manque de fidélité des explications. On pourrait notamment supposer que des méthodes fournissant des explications simples, par exemple basées sur des *tokens* comme dans cet article, ne sont pas adaptées à la multiplicité des caractéristiques exploitées par les grands modèles de langage, et que cette inadéquation entre un modèle complexe et des explications simples augmente la sensibilité des explications à l'aléa. Pour tester cette hypothèse, il faudrait évaluer dans quelle mesure un modèle plus simple que CamemBERT (par exemple avec moins de paramètres) est moins sensible à l'aléa. Améliorer la fidélité des explications en adaptant les méthodes et formats utilisés à la complexité des grands modèles de langage et au processus aléatoire de leur entraînement serait alors utile, tout en reposant la question du compromis avec leur plausibilité. En parallèle, diminuer la dépendance de l'entraînement des grands modèles de langage à des éléments aléatoires pourrait simplifier ce problème.

Remerciements. Louis Escouflaire, Marie-Catherine de Marneffe et François-Xavier Standaert sont respectivement aspirant, chercheuse qualifiée et maître de recherche du Fond National de Recherche Scientifique (FNRS-F.R.S.). Jérémie Bogaert est financé par le Service Public de Wallonie Recherche, via le projet 2010235-ARIAC.

8. Bibliographie

Abnar S., Zuidema W. H., « Quantifying Attention Flow in Transformers », *ACL*, p. 4190-4197, 2020.

Acheampong F. A., Nunoo-Mensah H., Chen W., « Transformer Models for Text-Based Emotion Detection : a Review of BERT-Based Approaches », *Artif. Intell. Rev.*, vol. 54, n° 8, p. 5789-5829, 2021.

Adebayo J., Gilmer J., Muelly M., Goodfellow I. J., Hardt M., Kim B., « Sanity Checks for Saliency Maps », *NeurIPS*, p. 9525-9536, 2018.

Arras L., Montavon G., Müller K., Samek W., « Explaining Recurrent Neural Network Predictions in Sentiment Analysis », *WASSA@EMNLP*, ACL, p. 159-168, 2017.

Bach S., Binder A., Montavon G., Klauschen F., Müller K.-R., Samek W., « On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation », *PloS one*, vol. 10, n° 7, p. e0130140, 2015.

Bailly A., Blanc C., Guillotin T., « Classification Multi-Label de Cas Cliniques avec CamemBERT (Multi-Label Classification of Clinical Cases with CamemBERT) », *TALN (DEFT)*, ATALA, p. 14-20, 2021.

Balota D., Pilotti M., Cortese M., « Subjective Frequency Estimates for 2,938 Monosyllabic Words », *Memory & cognition*, vol. 29, p. 639-647, 2001.

Bansal N., Agarwal C., Nguyen A., « SAM : The Sensitivity of Attribution Methods to Hyperparameters », *CVPR Workshops*, Computer Vision Foundation / IEEE, p. 11-21, 2020.

Bethard S., « We Need to Talk about Random Seeds », *CoRR*, 2022.

Bibal A., Cardon R., Alfter D., Wilkens R., Wang X., François T., Watrin P., « Is Attention Explanation ? An Introduction to the Debate », *ACL (1)*, p. 3889-3900, 2022.

Binder A., Montavon G., Lapuschkin S., Müller K., Samek W., « Layer-Wise Relevance Propagation for Neural Networks with Local Renormalization Layers », *ICANN (2)*, vol. 9887 of *Lecture Notes in Computer Science*, Springer, p. 63-71, 2016.

Bogaert J., Escouflaire L., de Marneffe M.-C., Descampe A., Standaert F.-X., Fairon C., « TIPECS : A Corpus Cleaning Method using Machine Learning and Qualitative Analysis », *International Conference on Corpus Linguistics (JLC)*, 2023.

Bonin P., Méot A., Bugaiska A., « Concreteness Norms for 1,659 French Words : Relationships with Other Psycholinguistic Variables and Word Recognition Times », *Behavior research methods*, vol. 50, p. 2366-2387, 2018.

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., « Language Models are Few-Shot Learners », *NeurIPS*, 2020.

Carroll J. B., « Language and Thought », *Reading Improvement*, vol. 2, n° 1, p. 80, 1964.

Charaudeau P., « Discours Journalistique et Positionnements Énonciatifs. Frontières et Dérives », *Revue de Sémio-Linguistique des Textes et Discours*, 2006.

Chefer H., Gur S., Wolf L., « Transformer Interpretability Beyond Attention Visualization », *CVPR*, Computer Vision Foundation / IEEE, p. 782-791, 2021.

Chenais G., Touchais H., Avalos M., Bourdois L., Revel P., Gil-Jardiné C., Lagarde E., « Performance en Classification de Données Textuelles des Passages aux Urgences des Modèles BERT pour le Français », *Santé & IA, PFIA*, 2021.

Clark K., Khandelwal U., Levy O., Manning C. D., « What Does BERT Look at ? An Analysis of BERT's Attention », *BlackboxNLP@ACL*, ACL, p. 276-286, 2019.

Cunha W., Mangaravite V., Gomes C., Canuto S. D., Resende E., Nascimento C., Viegas F., França C., Martins W. S., Almeida J. M., Rosa T., Rocha L., Gonçalves M. A., « On the Cost-Effectiveness of Neural and Non-Neural Approaches and Representations for Text Classification : a Comprehensive Comparative Study », *Inf. Process. Manag.*, vol. 58, n° 3, p. 102481, 2021.

Danilevsky M., Qian K., Aharonov R., Katsis Y., Kawas B., Sen P., « A Survey of the State of Explainable AI for Natural Language Processing », *AACL/IJCNLP*, ACL, p. 447-459, 2020.

Desrochers A., Thompson G., « Subjective Frequency and Imageability Ratings for 3,600 French Nouns », *Behavior research methods*, vol. 41, n° 2, p. 546-557, 2009.

Devlin J., Chang M., Lee K., Toutanova K., « BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding », *NAACL-HLT (1)*, ACL, p. 4171-4186, 2019.

Efron B., Tibshirani R., *An Introduction to the Bootstrap*, Springer, 1993.

Escouflaire L., « Identification des Indicateurs Linguistiques de la Subjectivité les Plus Efficaces pour la Classification d'Articles de Presse en Français », *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 2 : 24e Rencontres Etudiants Chercheurs en Informatique pour le TAL (RECITAL)*, p. 69-82, 2022.

Escouflaire L., Bogaert J., Descampe A., Fairon C., « The RTBF Corpus : a Dataset of 750,000 Belgian French News Articles Published between 2008 and 2021 », *International Conference on Corpus Linguistics (JLC)*, 2023.

Gémes K., Kovács Á., Reichel M., Recski G., « Offensive Text Detection on English Twitter with Deep Learning Models and Rule-Based Systems », *FIRE (Working Notes)*, vol. 3159 of *CEUR Workshop Proceedings*, CEUR-WS.org, p. 283-296, 2021.

Grosse E.-U., « Évolution et Typologie des Genres Journalistiques. Essai d'une Vue d'Ensemble », *Revue de Sémio-Linguistique des Textes et Discours*, 2001.

Gururangan S., Swayamdipta S., Levy O., Schwartz R., Bowman S. R., Smith N. A., « Annotation Artifacts in Natural Language Inference Data », *NAACL-HLT (2)*, Association for Computational Linguistics, p. 107-112, 2018.

Herman B., « The Promise and Peril of Human Evaluation for Model Interpretability », *CoRR*, 2017.

Jacovi A., Goldberg Y., « Towards Faithfully Interpretable NLP Systems : How Should We Define and Evaluate Faithfulness ? », *ACL*, p. 4198-4205, 2020.

Koren R., « Argumentation, Enjeux et Pratique de l'Engagement Neutre : le Cas de l'Écriture de Presse », *Revue de Sémio-Linguistique des Textes et Discours*, 2004.

Koufakou A., Pamungkas E. W., Basile V., Patti V., « HurtBERT : Incorporating Lexical Features with BERT for the Detection of Abusive Language », *WOAH*, ACL, p. 34-43, 2020.

Kovaleva O., Romanov A., Rogers A., Rumshisky A., « Revealing the Dark Secrets of BERT », *EMNLP/IJCNLP (1)*, ACL, p. 4364-4373, 2019.

Krüger K. R., Lukowiak A., Sonntag J., Warzecha S., Stede M., « Classifying News versus Opinions in Newspapers : Linguistic Features for Domain Independence », *Nat. Lang. Eng.*, vol. 23, n° 5, p. 687-707, 2017.

Lagneau E., « Le Style Agencier et ses Déclinaisons Thématiques : l'Exemple des Journalistes de l'Agence France Presse », *Réseaux*, vol. 1, p. 58-100, 2002.

Le H., Vial L., Frej J., Segonne V., Coavoux M., Lecouteux B., Allauzen A., Crabbé B., Besacier L., Schwab D., « FlauBERT : Unsupervised Language Model Pre-training for French », *LREC*, p. 2479-2490, 2020.

Lehmann E. L., Romano J. P., *Testing Statistical Hypotheses, Third Edition*, Springer texts in statistics, Springer, 2008.

Li J., Chen X., Hovy E. H., Jurafsky D., « Visualizing and Understanding Neural Models in NLP », *HLT-NAACL*, ACL, p. 681-691, 2016.

Li Q., Peng H., Li J., Xia C., Yang R., Sun L., Yu P. S., He L., « A Survey on Text Classification : From Shallow to Deep Learning », *CoRR*, 2020.

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

Lyu Q., Apidianaki M., Callison-Burch C., « Towards Faithful Model Explanation in NLP : A Survey », *CoRR*, 2022.

Martin L., Müller B., Suárez P. J. O., Dupont Y., Romary L., de la Clergerie É., Seddah D., Sagot B., « CamemBERT : a Tasty French Language Model », *ACL*, p. 7203-7219, 2020.

Miller T., « Explanation in Artificial Intelligence : Insights from the Social Sciences », *Artif. Intell.*, vol. 267, p. 1-38, 2019.

Mohammad S. M., Turney P. D., « Crowdsourcing a Word-Emotion Association Lexicon », *Comput. Intell.*, vol. 29, n° 3, p. 436-465, 2013.

Muñoz-Torres J.-R., « Truth and Objectivity in Journalism : Anatomy of an Endless Misunderstanding », *Journalism studies*, vol. 13, n° 4, p. 566-582, 2012.

Murdoch W. J., Singh C., Kumbier K., Abbasi-Asl R., Yu B., « Interpretable Machine Learning : Definitions, Methods, and Applications », *CoRR*, 2019.

New B., Pallier C., Brysbaert M., Ferrand L., « Lexique 2 : A New French Lexical Database », *Behavior Research Methods, Instruments, & Computers*, vol. 36, n° 3, p. 516-524, 2004.

Polignano M., Basile V., Basile P., Gabrieli G., Vassallo M., Bosco C., « A Hybrid Lexicon-Based and Neural Approach for Explainable Polarity Detection », *Inf. Process. Manag.*, vol. 59, n° 5, p. 103058, 2022.

Ravi K., Ravi V., « A Survey on Opinion Mining and Sentiment Analysis : Tasks, Approaches and Applications », *Knowl. Based Syst.*, vol. 89, p. 14-46, 2015.

Reif E., Yuan A., Wattenberg M., Viégas F. B., Coenen A., Pearce A., Kim B., « Visualizing and Measuring the Geometry of BERT », *NeurIPS*, p. 8592-8600, 2019.

Ribeiro M. T., Singh S., Guestrin C., « "Why Should I Trust You ?" : Explaining the Predictions of Any Classifier », *KDD*, ACM, p. 1135-1144, 2016.

Riloff E., Wiebe J., Phillips W., « Exploiting Subjectivity Classification to Improve Information Extraction », *AAAI*, AAAI Press / The MIT Press, p. 1106-1111, 2005.

Schudson M., « The Objectivity Norm in American Journalism », *Journalism*, vol. 2, n° 2, p. 149-170, 2001.

Sen C., Hartvigsen T., Yin B., Kong X., Rundensteiner E. A., « Human Attention Maps for Text Classification : Do Humans and Neural Networks Focus on the Same Words ? », *ACL*, p. 4596-4608, 2020.

Srinivas S., Fleuret F., « Full-Gradient Representation for Neural Network Visualization », *NeurIPS*, p. 4126-4135, 2019.

Sundararajan M., Taly A., Yan Q., « Axiomatic Attribution for Deep Networks », *ICML*, vol. 70 of *Proceedings of Machine Learning Research*, PMLR, p. 3319-3328, 2017.

Tong J., Zuo L., « The Inapplicability of Objectivity : Understanding the Work of Data Journalism », *Journalism Practice*, vol. 15, n° 2, p. 153-169, 2021.

Tuchman G., « Objectivity as Strategic Ritual : an Examination of Newsmen's Notions of Objectivity », *American Journal of sociology*, vol. 77, n° 4, p. 660-679, 1972.

Vargo C. J., Guo L., Amazeen M. A., « The Agenda-Setting Power of Fake News : a Big Data Analysis of the Online Media Landscape from 2014 to 2016 », *New Media Soc.*, vol. 20, n° 5, p. 2028-2049, 2018.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I., « Attention is All you Need », *NIPS*, p. 5998-6008, 2017.

Voita E., Talbot D., Moiseev F., Sennrich R., Titov I., « Analyzing Multi-Head Self-Attention : Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned », *ACL (1)*, Association for Computational Linguistics, p. 5797-5808, 2019.

Wiebe J., Wilson T., Bruce R. F., Bell M., Martin M., « Learning Subjective Language », *Comput. Linguistics*, vol. 30, n° 3, p. 277-308, 2004.

Zellers R., Holtzman A., Rashkin H., Bisk Y., Farhadi A., Roesner F., Choi Y., « Defending Against Neural Fake News », *NeurIPS*, p. 9051-9062, 2019.

Zhou Y., Ribeiro M. T., Shah J., « ExSum : From Local Explanations to Model Understanding », *NAACL-HLT*, Association for Computational Linguistics, p. 5359-5378, 2022.

Zini J. E., Awad M., « On the Explainability of Natural Language Processing Deep Models », *ACM Comput. Surv.*, vol. 55, n° 5, p. 103 :1-103 :31, 2023.