Pourquoi dois-je croire ta prédiction ? Comment expliquer les résultats d'une classification automatique de sentiments à partir de textes

Waleed Ragheb[1,2], Jérôme Azé[1,2], Sandra Bringay[1,3], Maximilien Servajean[1,3]

[1] LIRMM UMR 5506, CNRS, University of Montpellier

[2] IUT DE BÉZIERS, University of Montpellier, Béziers

[3] AMIS, Paul Valéry University - Montpellier 3

Résumé : Dans le cadre d'un problème classique de classification de sentiments, nous proposons un modèle qui utilise 1) l'apprentissage par transfert plutôt que les méthodes classiques de word embedding et 2) des mécanismes d'attention permettant de se concentrer sur les parties du texte importantes pour la tâche de classification étudiée. Notre modèle a été évalué sur plusieurs jeux de données et montre des résultats très compétitifs. Or, si ces méthodes d'apprentissage en profondeur s'avèrent très efficaces, elles sont souvent considérées comme des boites noires, difficiles à interpréter. Dans cet article, nous évaluons l'impact des mécanismes d'attention traduits sous la forme de nuages de mots-clés pour aider les utilisateurs à interpréter les résultats de la classification. L'expérimentation d'une telle visualisation sur plus de 85 participants a permis de montrer son intérêt en terme d'interprétabilité.

Mots-clés : classification de sentiments, mécanisme d'attention et interprétabilité.

1 Introduction

L'analyse de sentiments à partir de textes est un domaine de recherche très actif dans la communauté de l'apprentissage automatique (Mäntylä *et al.*, 2018) avec de nombreuses applications comme le e-commerce, la gestion de la réputation, le support client, la politique, etc. De nombreux outils sont désormais disponibles pour détecter des sentiments subjectifs, tels que la polarité (positif ou négatif) et les émotions (peur, joie, etc.). Si ces outils sont de plus en plus efficaces, ils sont généralement critiqués pour deux raisons.

Tout d'abord, la plupart des approches existantes ne sont efficaces que pour des domaines d'application spécifiques ou des types de textes particuliers tels que l'analyse de sentiments des sites Web liés à la finance (Luo *et al.*, 2018) ou des avis clients relatifs à des ordinateurs portables ou des restaurants (Li *et al.*, 2018). Dans ce contexte, l'apprentissage par transfert et l'adaptation au domaine sont largement utilisés, en particulier combinés à des réseaux de neurones profonds, pour aider à réutiliser les modèles développés pour une tâche source vers une autre tâche cible. L'apprentissage par transfert fonctionne particulièrement bien quand les caractéristiques apprises pour la tâche source sont générales et peuvent être réutilisées pour les tâches cibles. Ce type d'approche a fait ses preuves dans le domaine de la vision par ordinateur où l'extraction des caractéristiques se fait à partir de modèles pré-entraînés de type AlexNet, ResNet, MS-COCO, etc. (Voulodimos *et al.*, 2018). Dans les modèles de traitement du langage naturel, cette approche n'a connu de réel succès que très récemment grâce au modèle de langage universel (ULMFiT) proposé par (Howard & Ruder, 2018) qui servira de baseline à ces travaux.

Par ailleurs, en traitement automatique de la langue, la plupart des modèles de transduction compétitifs ont une structure de type codeur-décodeur (Vaswani *et al.*, 2017). Une limite de ces architectures est qu'elle code la séquence d'entrée dans une représentation interne de longueur fixe. Cela entraîne une dégradation des résultats lorsque la longueur de la séquence

augmente. Dans ce contexte, les mécanismes d'attention (Young *et al.*, 2018) ont récemment été utilisés pour résoudre ce problème. Inspirés des mécanismes d'attention visuelle que l'on retrouve chez l'homme, ils focalisent l'analyse sur certaines régions d'une image avec une "haute résolution" tout en percevant le reste de l'image en "basse résolution". L'attention guide le réseau pour qu'il sache où accorder son attention sur la séquence d'entrée. Les premières applications des mécanismes d'attention ont été naturellement réalisées dans le domaine de la vision par ordinateur (Anderson *et al.*, 2017) et plus récemment sur les textes pour des application de traduction automatique (Bahdanau *et al.*, 2014) et d'analyse des sentiments (Ma *et al.*, 2018).

Dans ce travail, nous évaluerons la combinaison des mécanismes d'attention à une architecture de type ULMFiT sur des jeux de données réelles de la littérature.

Par ailleurs, nous nous poserons la question de l'interprétabilité de notre approche. Les modèles d'apprentissage sont souvent décriés car perçus comme des boites noires. Or, pour un utilisateur qui va baser des décisions et des actions à partir de prédictions, il peut être fondamental d'interpréter les raisons sous-jacentes aux prédictions pour ainsi faire confiance à ces prédictions. En effet, il y a des cas où l'erreur a peu d'importance, par exemple quand on recommande un film, mais il y a des cas où l'erreur porte à conséquence, par exemple pour le diagnostic médical ou la détection d'obstacles pour une voiture autonome. Selon l'importance de l'erreur pour l'activité de l'utilisateur, celui-ci aura besoin de comprendre comment le modèle fonctionne en général et comment il est arrivé à proposer une prédiction particulière.

Dans ce travail, nous nous intéressons plus particulièrement à la sortie des mécanismes d'attention qui est la visualisation des parties du texte ayant impacté la prédiction du sentiment. Cette visualisation donne des justifications complémentaires aux étiquettes prédites associées aux textes. À notre connaissance, aucune étude n'a montré l'apport de la visualisation des mécanismes d'attention sur l'interprétation des utilisateurs pour l'analyse de sentiments à partir de textes.

Finalement, l'objectif de cet article est double : 1) montrer l'impact en terme d'exactitude des mécanismes d'attention pour l'analyse des sentiments lorsqu'ils sont ajoutés à l'architecture de référence ULMFiT qui intègre un apprentissage par transfert ; 2) évaluer dans quelle mesure une visualisation basée sur le mécanisme d'attention facilite la prise de décision en apportant un élément explicatif supplémentaire à l'étiquette proposée.

Le reste de cet article est organisé comme suit. Dans la section 2, nous décrivons les travaux connexes de la littérature puis le modèle proposé dans la section 3. Dans la section 4, nous présentons les expérimentations sur les performances du modèle, les jeux de données et les résultats obtenus. Dans la section 5, nous détaillons les expérimentations réalisées sur l'interprétabilité du modèle. Nous concluons et donnons des perspectives dans la section 6.

2 État de l'art

Afin d'améliorer l'efficacité des algorithmes d'analyse de sentiments et leur explicabilité, nous proposons une nouvelle méthode basée sur l'apprentissage par transfert de modèles de langue ainsi que sur les mécanismes d'attention.

Les modèles de langue (LM) visent à prédire un mot à partir des mots le précédant. Ces modèles sont utilisés dans de nombreuses applications de traitement automatique de la langue naturelle car ils permettent de capturer des dépendances éloignées ainsi que la structure hiérarchique du texte. L'apprentissage des modèles de langue est non supervisé, car ne nécessitant pas de corpus de texte préalablement étiqueté. Or, ces modèles sont peu adaptés aux petits ensembles de données et peuvent donner un mauvais rappel pour certaines tâches de classification. Récemment, (Howard & Ruder, 2018) ont proposé la méthode ULMFiT, basée sur une représentation des mots peu profonde, qui combinée à un apprentissage par transfert s'est avérée être très efficace pour différentes tâches dont l'analyse de sentiments (Merity *et al.*, 2018). Cette méthode servira de baseline à nos travaux. Nous avons choisi d'utili-

ser ULMFiT pour ses performances en *fine tuning* et sa taille raisonnable en comparaison à d'autres approches très récentes comme BERT (Devlin *et al.*, 2018) et ELMO (Peters *et al.*, 2018). De plus, pour des tâches de classification de sentiments, ULMFiT est utilisé comme l'état de l'art sur la plupart des jeux de données de la littérature, dont ceux sur lesquels nous avons travaillé.

Dans le domaine du traitement automatique de la langue naturelle, les mécanismes d'attention ont été récemment étudiés. L'auto-attention met en relation différentes positions d'une séquence afin d'en calculer une représentation (Lin *et al.*, 2017). Chaque partie de la séquence d'entrée est associée à un score de probabilité. L'attention a été appliquée avec succès pour de nombreuses tâches, notamment la compréhension de la lecture, le résumé, la traduction automatique (Su *et al.*, 2018).

Outre le fait que les mécanismes d'attention améliorent l'efficacité de la classification, ce qui nous intéresse dans cet article est le fait que le résultat de la couche d'attention peut être utilisé en entrée d'une visualisation visant l'explication d'une prédiction (Wang *et al.*, 2018) (Lin *et al.*, 2017).

Une explication est une réponse à une question de type « Pourquoi ? ». Dans notre cas, pourquoi le système prédit une polarité pour un texte ? Il a été démontré par (Herlocker *et al.*, 2000) qu'associer des explications aux prédictions améliore l'acceptation de ces prédictions dans le cas de la recommandation de films. En effet, les explications influencent le ressenti de celui à qui l'on fournit les explications et donc les actions qu'il peut être amené à réaliser. Pour expliquer une prédiction, la plupart des méthodes utilisent des artefacts visuels qui fournissent une compréhension qualitative des liens existants entre les instances (des mots, des parties d'images...) et les prédictions.

Il existe différentes méthodes d'interprétabilité. On distingue les méthodes d'apprentissage intrinsèquement interprétables lorsque la sélection et l'entraînement du modèle d'apprentissage sont par eux-même interprétables (e.g. les arbres de décision) et les méthodes d'interprétabilité post-hoc qui donnent des explications *a posteriori* et qui s'appliquent sur des méthodes d'apprentissage de type boite noire après la sélection et l'entraînement du modèle, pour expliquer les prédictions réalisées. La plupart des modèles d'apprentissage de l'état de l'art dont les modèles de type réseau de neurones étudiés dans cet article, n'étant pas interprétables, nous nous sommes focalisés sur cette deuxième catégorie de méthodes d'interprétabilité.

À notre connaissance, il n'y a pas eu d'étude qualitative de l'impact de la visualisation de l'attention sur l'interprétation que les utilisateurs peuvent faire de cette nouvelle information dans le cas de la classification de données textuelles. Cette étude nous parait importante dans le contexte du besoin de confiance manifesté par les utilisateurs des méthodes d'apprentissage notamment dans des domaines comme la santé. Ces utilisateurs ne sont pas uniquement en attente de résultats performants mais ils recherchent des explications qui, associées aux prédictions, vont améliorer la prise de décision (Ribeiro *et al.*, 2016)(Vellido *et al.*, 2012).

3 Architecture proposée

Notre architecture est composée de quatre composants décrits dans les sections suivantes.

3.1 Auto-Attention basée sur un encodeur AWD-LSTM

Un LSTM traditionnel possède un portail d'entrée i_t, d'oubli f_t, de sortie o_t et une cellule mémoire c_t. Ce sont des vecteurs de $\mathbb{R}^d$ qui correspondent à la représentation vectorielle

d'une dimension d. Les équations de transition de LSTM sont les suivantes :

$$\begin{aligned} f_t &= \sigma(W_f x_t + U_f h_{t-1}) \\ i_t &= \sigma(W_i x_t + U_i h_{t-1}) \\ o_t &= \sigma(W_o x_t + U_o h_{t-1}) \\ \tilde{c}_t &= tanh(W_c x_t + U_c h_{t-1}) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\ h_t &= o_t \odot tanh(c_t) \end{aligned} \tag{1}$$

où x_t est l'entrée au pas de temps actuel, σ est la fonction sigmoïde ; $\odot$ l'opération de multiplication par élément, $W_{\{i,f,o,c\}}$, $U_{\{i,f,o,c\}}$ sont des ensembles de poids appris.

Dans notre modèle, nous utilisons le vecteur d'état caché à chaque pas de temps comme représentation du mot correspondant dans une phrase. Afin d'éviter le sur-apprentissage lors de l'entraînement du LSTM, (Merity *et al.*, 2018) ont proposé AWD-LSTM. basée sur DropConnect (Wan *et al.*, 2013) pour pondérer les matrices $U_{\{i,f,o,c\}}$. Nous avons utilisé les trois mêmes couches liées de LSTM et avons également appliqué une auto-attention sur les vecteurs d'état cachés à chaque pas de temps. La séquence d'états cachés en entrée $H^{i-1} = \{h_1^{i-1}, h_2^{i-1}, \ldots, h_N^{i-1}\}$, où N est la longueur de la séquence, est passée aux états de la couche LSTM. Les états de sortie sont de la forme de $H^i = \{h_1^i, h_2^i, \ldots, h_N^i\}$. La couche d'attention prend la séquence en entrée encodée et calcule les scores d'attention $S^i = \{s_1^i, s_2^i \ldots, s_N^i\}$. La couche d'attention est une couche linéaire sans biais.

$$\begin{aligned} \alpha^i &= \{V^i.H^i\} \\ S^i &= exp(\alpha^i) / \sum_{j=1}^{N} exp(\alpha_j^i) \end{aligned} \tag{2}$$

Où V^i est le poids de la couche d'attention i^{th} du Att-LSTM.

3.2 Agrégation sur plusieurs niveaux de l'auto-attention

L'architecture proposée utilise un empilement de trois Att-LSTM superposés exactement, de la même manière que l'architecture AWD-LSTM classique. Nous n'utilisons pas d'architecture Bi-LSTM ici, car notre modèle correspond à l'ensemble des directions avant et arrière. La figure 1 montre un aperçu du modèle. À chaque couche, les scores d'attention sont obtenus selon un niveau spécifique de codage de la séquence, puis agrégés pour obtenir les scores d'attention globaux $\overline{S}$. La fonction d'agrégation est la moyenne logarithmique des trois niveaux de scores d'attention.

$$\overline{S} = log \sum_{i=1}^{3} S^i / 3 \tag{3}$$

Les scores globaux d'attention $\overline{S}$ sont utilisés pour calculer la séquence $O = \{o_1, o_2, \ldots, o_N\}$ où o_i est le produit du score d'attention et de la sortie de la couche Att-LSTM, tel que :

$$o_i = \overline{s}_i \otimes h_i^3 \tag{4}$$

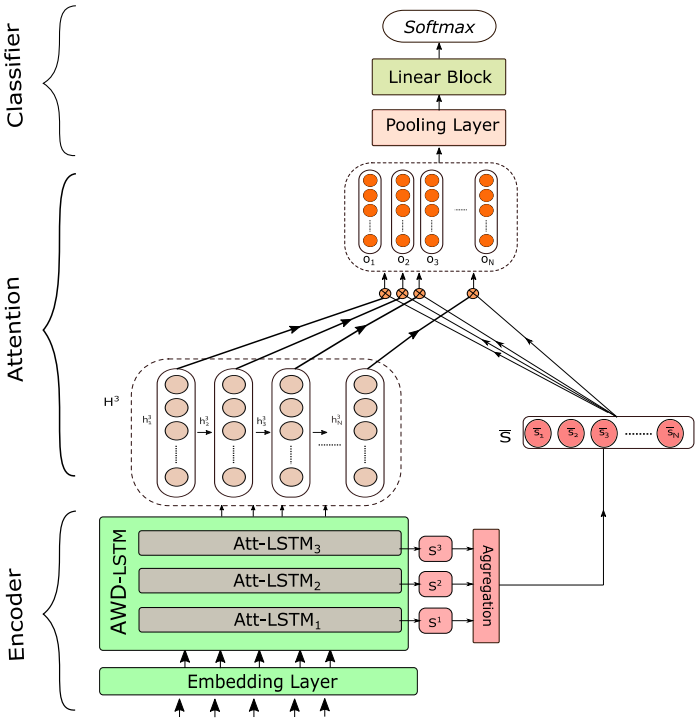

FIGURE *1 – Notre architecture*

3.3 Couches de classification

Après avoir agrégé les informations provenant de l'attention multi-niveaux et de la sortie de l'encodeur, nous convertissons les représentations résultantes de toutes les positions de O en un vecteur à longueur fixe avec une opération de pooling. Nous avons utilisé trois fonctions de pooling. Nous appliquons une attention X_{att} telle que :

$$X_{att} = \sum_{i=1}^{N} exp(\overline{s}_i) \otimes h_i^3 \tag{5}$$

Nous appliquons un pooling maximum X_{max} et moyen X_{avg} à O afin d'obtenir les représentations finales du texte saisi après encodage et application de l'attention, donné par :

$$X_{in} = [X_{att} \oplus X_{max} \oplus X_{avg}] \tag{6}$$

Puis, nous alimentons le bloc linéaire du classifieur. Ce bloc est constitué de deux couches entièrement connectées, de tailles différentes, suivies d'un indicateur *softmax* pour déterminer la classe de sentiment en sortie.

3.4 entraînement du modèle

L'entraînement se fait en trois étapes :

— Le modèle de langue est initialisé de manière aléatoire, puis entraîné en empilant un décodeur linéaire au-dessus de l'encodeur. Le modèle de langue est appris à partir d'un corpus du domaine général. Cela aide le modèle à apprendre des caractéristiques générales du langage.

— Le même modèle de langue après entraînement est utilisé pour l'initialisation puis ajusté à partir des données de la tâche cible, dans notre cas : différents jeux de données d'analyse de sentiments. Ici, nous limitons le vocabulaire du modèle de langue aux mots fréquents (répétés plus de deux fois).

— Nous conservons l'encodeur et remplaçons le décodeur par le classifieur et les paramètres de ces deux étapes sont réglés avec précision pour la tâche cible.

Lors de la première étape et pour l'apprentissage du modèle de langage, nous avons utilisé le jeu de données Wikitext-103 (Merity *et al.*, 2016). Avec plus de 28 000 articles Wikipédia et 103 millions de mots, le modèle détermine la structure principale et la hiérarchie du langage par modélisation séquence par séquence. Pour l'entraînement du modèle de classification, à partir du jeu étiqueté, nous optimisons tous les paramètres afin de minimiser autant que possible la fonction objectif (fonction de perte). Dans notre travail, nous prenons y_i la polarité du sentiment correcte et $\hat{y_i}$ qui désigne la polarité de sentiment prédite. Nous considérons l'entropie croisée comme la fonction de perte, dont la formule est la suivante :

$$loss = -\sum_{<T>} y_i \log(\hat{y_i}) + \lambda \|\theta\|^2 \tag{7}$$

Où λ est le facteur de régularisation, θ contient tous les paramètres du modèle et T est l'ensemble des exemples d'entraînement. L'entraînement de l'architecture est effectué à l'aide de taux d'apprentissage triangulaires inclinés (STLR) qui modifient le taux d'apprentissage pour chaque itération de manière triangulaire. Nous avons utilisé un seul cycle comme recommandé par (Howard & Ruder, 2018). Le modèle a été entraîné en utilisant un taux d'apprentissage différent pour chaque groupe de couches.

Nous entraînons le modèle sur les modèles de langue avant et arrière pour les ensembles de données du domaine général et spécifiques à la tâche. Les deux modèles de langue avant et arrière sont utilisés pour construire deux versions de la même architecture. La décision finale est l'ensemble des deux. Nous avons utilisé Pytorch [1] pour construire l'ensemble du modèle et les bibliothèques Fastai [2] pour l'apprentissage et les modèles de langue. Pour le prétraitement du texte, celui-ci a tout d'abord été normalisé et tokenisé. Des tokens spéciaux ont été ajoutés pour les mots en majuscules et les mots répétés. Nous conservons les symboles de ponctuation et de sentiments dans le texte. Nous avons utilisé Spacy [3] et FastText [4] pour ces prétraitements. Les modèles sont appris et testés sur 4 GPU Nvidia GEFORCE GTX 1080 ti.

4 Expérimentations sur les performances du modèles

Dans cette section, nous discutons plus en détail de l'efficacité de la méthode proposée.

4.1 Jeux de données

Nous appliquons le modèle à différents jeux de données de classification de sentiments. Le tableau 1 présente des statistiques sur ces jeux de données. Le jeu de données IMDB est un ensemble de données pour la classification des sentiments binaires de critiques de films (Maas *et al.*, 2011). Nous avons également utilisé les versions binaires et complètes des jeux de données d'avis d'utilisateurs de Yelp et d'Amazon (Zhang *et al.*, 2015). Pour les jeux de données binaires (IMDB, Yelp-bi et Amazon-bi), les classes à prédire sont positif ou négatif. Pour les autres, il s'agit d'un nombre d'étoiles : de négatif (1 étoile) à positif (5 étoiles). Ces jeux de données sont tous équilibrés.

4.2 Baselines et Résultats

Nous comparons notre modèle à plusieurs baselines compétitives de l'état de l'art qui utilisent le mécanisme d'attention pour la classification de sentiments :

1. https://pytorch.org/
2. http://www.fast.ai/
3. https://spacy.io/
4. https://fasttext.cc/

TABLE *1 – Jeux de données de sentiments et nombre d'exemples d'apprentissage et de test*

Dataset	#Exemples d'apprentissage	#Exemples de test	#classes
IMDB	25K	25K	2
Yelp-bi	560K	38K	2
Yelp-Full	650K	50K	5
Amazon-bi	3.6M	400K	2
Amazon-Full	3M	650K	5

TABLE *2 – Taux d'erreur (%) de notre modèle et des baselines*

Models	IMDB	Yelp-bi	Yelp-Full	Amazon-bi	Amazon-Full
HN-ATT	-	-	-	-	36.40
DCCNN-ATT	-	2.64	30.58	**3.32**	34.81
SANet	-	4.77	36.03	4.52	38.67
SA-Embedding	-	5.10	36.60	-	40.20
CSC	-	6.90	35.97	4.90	39.89
CRAN	7.90	-	-	-	-
IRAM	8.80	-	-	-	-
ULMFiT	4.60	**2.16**	29.98	-	-
Ours	**4.51**	2.25	**29.76**	3.43	**34.78**

— HN-ATT (Yang *et al.*, 2016) Le modèle reflète la structure hiérarchique des documents à travers deux niveaux d'attention dans les mots et les phrases.
— DCCNN-ATT (Wang *et al.*, 2018) Ce modèle est un réseau de neurones convolutifs avec des connexions denses et des fonctionnalités multi-échelles.
— SANet (Letarte *et al.*, 2018) Ce modèle utilise l'attention pour modéliser les interactions entre toutes les paires de mots d'entrée.
— SA-Embedding (Lin *et al.*, 2017) Ce modèle est basé sur l'extraction d'une représentation interprétable de la phrase donnée en entrée via un mécanisme d'attention.
— CSC (Mokhtari *et al.*, 2018) Ce modèle utilise un réseau de neurones hiérarchique basé sur l'attention qui intègre les préférences de l'utilisateur et les caractéristiques du produit dans les tâches de classification de sentiments.
— CRAN (Du *et al.*, 2017) Le modèle associe à la fois les attentions basées sur la convolution et les attentions basées sur la récurrence.
— IRAM (Tutek & Šnajder, 2018) Il s'agit d'un modèle d'attention, qui construit de manière récursive des représentations d'entrée des données par la réutilisation des résultats qui ont été précédemment calculés.
— ULMFiT (Howard & Ruder, 2018) Il s'agit de la méthode de référence actuelle .

Le tableau 2 montre l'erreur obtenue lors du test du modèle proposé et de toutes les baselines sur les jeux de tests. Nous présentons les résultats tels que rapportés dans la publication d'origine. Notre modèle surpasse tous les modèles basés sur l'attention avec une marge significative et reste compétitif par rapport à ULMFiT. Notre modèle proposé obtient de meilleurs résultats qu'ULMFiT pour le jeu IDBM et la version complète de Yelp.

Dans la suite, nous allons nous demander comment améliorer l'interprétabilité de ce modèle et montrer comment les mécanismes d'attention peuvent être utilisés pour expliquer les prédictions.

5 Expérimentations sur l'interprétabilité du modèle

Dans cette section, nous discutons plus en détail de l'impact de l'application de l'attention sur l'interprétabilité.

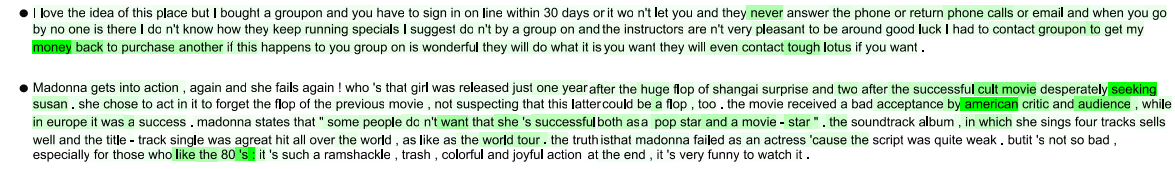

FIGURE *2 – Exemples de visualisation de l'attention sur une critique de restaurant et de film*

5.1 Visualisation de l'attention

L'utilisateur doit avoir confiance dans un modèle et pour cela, il voudra vérifier que celui-ci fonctionne bien sur des données réelles selon une métrique d'intérêt, comme le taux d'erreur utilisé dans la section 4. Ce type d'évaluation peut s'avérer inefficace notamment quand les données évoluent au cours du temps car elle revient à évaluer entre autre l'écart entre les données réelles et celles d'entraînement. Un humain peut alors élaborer différentes stratégies pour sélectionner un modèle parmi plusieurs modèles et notamment préférer un modèle explicable mais moins performant, qui lui laisse la responsabilité du choix final.

Dans ce contexte, l'un des résultats les plus intéressants du mécanisme d'attention est sa capacité à traiter toutes les séquences d'entrée avec différents poids d'attention. Le modèle accorde une plus grande attention aux éléments qui influencent la décision du réseau. La figure 2 montre des exemples d'avis positifs et négatifs correctement classés pour les avis de restaurants (Yelp-bi) et de critiques de film (IDBM). Les scores d'attention sont utilisés pour colorer le texte. Cette information peut être alors présentée à l'utilisateur avec la prédiction comme explication.

Dans notre contexte, une explication est un ensemble de mots-clés associés à un score d'attention. Dans la suite, nous allons chercher à évaluer la qualité des explications fournies via la visualisation de l'attention selon une approche centrée sur le système (le niveau d'interprétabilité est basé sur l'analyse des sorties du système) puis selon une approche centrée sur l'humain (le niveau d'interprétabilité est basé sur une tâche expérimentale d'inférence de la polarité).

5.2 Évaluation centrée système

Nous avons tout d'abord mis en place une évaluation centrée système, sans intervention humaine. Pour cela, nous avons mesuré l'intersection entre le nombre de mots appartenant à des lexiques ayant fait leur preuve pour les tâches de classification de sentiments et les mots repérés par les mécanismes d'attention. Nous avons utilisé pour cela EmoLex proposé par (Mohammad & Turney, 2013) (Mohammad, 2018) qui comporte plus de 10 000 entrées.

Le tableau 3 montre les résultats de la mise en correspondance des principaux mots repérés avec l'attention et retrouvés dans EmoLex pour les deux jeux de tests. Par exemple, nous pouvons dire que 88,11% des exemples du jeu de données IMDB contiennent des sentiments dans les 5% supérieurs des mots identifiés par le processus d'attention. Cela reflète la précision du mécanisme d'attention qui se concentre sur ces mots et expressions.

TABLE *3 – Sentiments et émotions dans les top 5, 10 ou 20% des scores d'attention calculés*

Dataset	Top 5%	Top 10%	Top 20%
IMDB	88.11%	97.30%	99.65%
Yelp-bi	51.86%	78.81%	94.37%
Yelp-Full	64.11%	84.65%	95.47%
Amazon-bi	55.71%	80.18%	94.65%
Amazon-Full	57.67%	81.02%	94.84%

FIGURE *3 – Exemple de nuage de mots-clés généré à partir des mécanismes d'attention*

5.3 Évaluation centrée humain

Même si les modèles d'attention se concentrent sur certains mots porteurs d'émotions, ils se concentrent aussi sur d'autres parties du texte, ayant un impact important sur la décision finale du réseau. Sur la base d'exemples tels que ceux de la figure 2, nous pourrions facilement trouver des parties du texte surlignées qui ne sont pas liées aux lexiques de sentiments.

Afin de déterminer dans quelle mesure une explication basée sur le simple fait de prêter attention aux mots importants, malgré une sélection imparfaite par l'algorithme, pourrait induire une décision de polarité meilleure qu'une explication basée sur les termes d'un lexique, nous avons mis en place une évaluation centrée humain impliquant des utilisateurs non experts. Pour cela, nous leur avons présenté les deux types d'explications générées sous la forme de nuages de mots comme présenté sur la figure 3.

Les nuages de mots sont générées de trois manières différentes : en utilisant 1) uniquement le lexique, 2) uniquement l'attention, ou 3) les deux. Dans le cas des nuages de mots générés à partir du lexique, la taille des mots est déterminée par les fréquences des mots et leur ordre dans le texte. Dans le cas des mécanismes d'attention, la taille et la position des mots sont déterminées par le score d'attention donné dans la section 3.1 . Pour les nuages de mots mixtes, nous combinons les deux informations.Les couleurs des mots sont choisies pour augmenter le contraste entre les mots de tailles différentes pour une meilleure visualisation. Ces couleurs ne représentent aucune information supplémentaire.

Chaque participant a répondu à 30 questions, 10 questions pour chaque modalité. La question posée a consisté à deviner la polarité des critiques (positives ou négatives) en se basant sur les nuages de mots. Nous informons le participant si le nuage de mots provient d'un commentaire de film, de restaurant ou d'un autre produit. La sélection des questions est randomisée pour chaque type de question et est différente pour chaque participant. Le nombre total de participants est de 85. Il s'agit d'étudiants inscrits dans des cursus en informatique. Pour chaque type de question, nous calculons l'index moyen de facilité F_i comme suit :

$$F_i = 100 * \frac{\bar{x_i} - x^i_{min}}{x^i_{max} - x^i_{min}} \tag{8}$$

où $\bar{x_i}$ est le score moyen par question de type i, x^i_{min} et x^i_{max} sont les scores minimum et maximum pour une question de type i.

Le score par question donne une indication sur la facilité avec laquelle la question est posée. Nous calculons également l'écart-type de tous les scores pour chaque question. De plus, nous calculons l'indice de discrimination D_i comme :

$$D_i = 100 * corr(x_i, T) \tag{9}$$

qui correspond au pourcentage de corrélation entre les scores de chaque type de question et le score total de chaque participant. Le tableau 4 présente les résultats du sondage sur les nuages de mots pour les trois types de questions.

TABLE *4 – Résultats de l'enquête sur les nuages de mots*

Types de Question	Index de facilité	Déviation Standard	Index de discrimination
Attention	**72.12%**	**43.90%**	**7.70%**
Lexicon	58.90%	49.86%	-16.10%
Mixte	68.27%	49.19%	6.70%

Ces résultats restent préliminaires. Ils montrent que deviner le sentiment sans le texte complet est une tâche difficile. Cependant, les questions basées sur les nuages de mots construits par les mécanismes d'attention sont plus faciles que les autres avec un indice de facilité moyen de 72.12% et un écart-type de 43.90%. Par ailleurs, nous notons une corrélation positive entre l'accord et l'indice de discrimination. Nous pouvons en conclure que la tâche devient plus difficile en utilisant uniquement le lexique. Toutefois, les nuages construits à partir des scores d'attention captent des arguments qui ne sont pas centraux à la tâche et ne mettent pas en avant les expressions de négation et adversatives qu'il serait intéressant de repérer et de mettre en relief dans cette visualisation. D'autres formats de visualisation pourraient également être explorés.

5.4 Discussions

La méthode d'interprétabilité proposée, basée sur les mécanisme d'attention, permet de visualiser un résumé des statistiques (taille et ordre) des attributs (les mots) et de leur impact sur les prédictions du modèle (intensité de la couleur). Cette méthode d'explication est spécifique à l'algorithme d'apprentissage utilisé qui fournit l'information de l'attention et ne rentre donc pas dans la catégorie des modèles d'interprétation *a posteriori*, qui peuvent être utilisés avec n'importe quel autre algorithme d'apprentissage. Un avantage de notre méthode est qu'elle est expressive car la représentation visuelle retenue permet de structurer les explications. Comme les auteurs de la méthode LIME (Hu *et al.*, 2018), nous avons pu remarquer que les utilisateurs apprécient les explications qu'ils sélectionnent par eux-mêmes. En effet, ils ne s'attendent pas à une liste finie d'explications de la prédiction mais plutôt au choix d'une ou plusieurs explications dans une liste de choix, ce qui est possible en piochant des mots-clés dans le nuage. Toutefois, une limite est que l'interprétation proposée est locale et se limite à une prédiction unique alors que d'autres approches dites globales vont concerner tout le modèle d'apprentissage et notamment le choix des paramètres.

(Molnar, 2019) a listé les propriétés que l'on peut rencontrer pour évaluer la qualité des explications individuelles. Il considère :

— la précision (*dans quelle mesure une explication permet-elle de prédire des données non encore rencontrées ?*) : contrairement aux approches par lexique, les explications liées à l'attention resteront efficaces même avec un vocabulaire non standard,
— la fidélité (*dans quelle mesure l'explication se rapproche-t-elle de la prédiction du modèle vu comme une boîte noire ?*) : l'expérimentation a montré que l'humain est capable de retrouver la prédiction à partir des nuages,
— la cohérence (*en quoi une explication diffère-t-elle entre les modèles entraînés à la même tâche et produisant des prédictions similaires ?*) : le nuage de mots permet par exemple d'identifier les raisons pour lesquelles deux critiques de films vont être considérées comme positives (en soulignant par exemple des termes liés à la qualité de la réalisation ou au jeu des acteurs),
— la stabilité (*dans quelle mesure les explications sont-elles similaires pour des instances similaires* ?) : pour des critiques similaires, les mêmes mots sont repérés par l'attention,

— la compréhensibilité (*dans quelle mesure les humains comprennent-ils les explications ?* : même si la tâche a été considérée comme difficile, les humains ont réussi à prendre une décision à partir des nuages et ont trouvé la bonne prédiction avec un index de facilité de 72.12%,
— la certitude (*l'explication reflète-t-elle la certitude du modèle d'apprentissage automatique ?*) : le nuage de mots ne retourne pas d'information sur la certitude que le modèle a en sa prédiction mais cette information pourrait être ajoutée sous la forme d'une nouvelle variable visuelle,
— le degré d'importance (*dans quelle mesure l'explication reflète-t-elle l'importance de caractéristiques ou de parties de l'explication ?*) : cette information est représentée par la taille des mots dans le nuage,
— la nouveauté (*L'explication indique-t-elle si une instance à expliquer provient d'une région très éloignée de la distribution des données d'entraînement ?*) : la visualisation actuelle ne donne pas cette information.

Notre modèle d'explication locale des prédictions basé sur les nuages de mots-clés repérés par l'attention possède donc des propriétés intéressantes au sens de (Molnar, 2019).

6 Conclusions et perspectives

Nous pouvons conclure de cette étude que l'idée de l'apprentissage par transfert est très efficace pour une application de classification de sentiments en terme de taux d'erreur. Elle fonctionne mieux que les modèles classiques d'apprentissage basés sur les word embeddings. De plus, l'ajout d'un mécanisme d'auto-attention a un impact direct sur les performances de ces modèles. Le modèle proposé a été évalué sur cinq jeux de données de la littérature. Nos expériences montrent des résultats compétitifs par rapport aux modèles basés sur l'attention à la pointe de l'état de l'art dont ULMFit. Pour finir, même si la visualisation proposée peut facilement être améliorée et que les expérimentations restent préliminaires, nous avons démontré que la visualisation des scores d'attention présentés sous la forme de nuages de mots, impacte positivement les perceptions liées à l'interprétabilité des utilisateurs.

Dans les travaux futurs, une nouvelle expérimentation, cette fois comparative, devra être menée pour savoir si l'ajout d'une information comme la visualisation proposée permet effectivement à un utilisateur d'avoir confiance dans le modèle et si cela impacte sa prise de décision. Nous allons chercher à améliorer l'interprétation locale d'une prédication. Par exemple, il a été démontré que les humains apprécient les explications contrastives, qui permettent de comprendre pourquoi une prédiction a été faite à la place d'une autre. Pour un texte pour lequel nous cherchons à déterminer la polarité, faire évoluer notre visualisation pour colorer d'une couleur les mots-clés ayant contribué au choix du sentiment prédit et d'une autre couleur les mots-clés ayant contribué au sentiment inverse. Nous pourrions également travailler sur les techniques de résumé automatique, apparié avec des méthodes d'analyse de sentiments basées sur les facettes afin d'améliorer la visualisation.

Nous prévoyons également de travailler sur une interprétation globale du modèle. Tout d'abord, nous pourrions définir la "représentativité" des explications. L'intuition est la suivante. Les bonnes explications sont souvent générales. La représentativité d'une explication correspondrait au nombre de textes couverts par cette explication. Nous pourrions également détecter les explications fréquemment associées via les mots-clés se retrouvant dans les nuages de mots. Ces deux informations pourrait aider l'utilisateur à comprendre sur quoi le classifieur s'appuie généralement pour prendre sa décision. Par opposition, nous pourrions repérer les explications "anormales" (peu fréquentes). Si l'une des caractéristiques d'entrée d'une prédiction est anormale (un groupe de mots-clés rares), elle devrait être présentée dans l'explication du modèle afin d'identifier les cas particuliers. Pour conclure, un outil permettant de naviguer dans les explications basé sur la recherche d'explications représentatives, fréquemment associées et anormales permettrait d'améliorer l'interprétabilité globale du modèle.

7 Remerciement

Nous tenons à remercier la Région Occitanie et la Communauté d'Agglomération de Béziers Méditerranée pour le financement de la thèse de Waleed Ragheb.

Références

ANDERSON P., HE X., BUEHLER C., TENEY D., JOHNSON M., GOULD S. & ZHANG L. (2017). Bottom-up and top-down attention for image captioning and VQA. *CoRR*, **abs/1707.07998**, 6077–6086.

BAHDANAU D., CHO K. & BENGIO Y. (2014). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*.

DEVLIN J., CHANG M., LEE K. & TOUTANOVA K. (2018). BERT : pre-training of deep bidirectional transformers for language understanding. *CoRR*, **abs/1810.04805**.

DU J., GUI L., HE Y. & XU R. (2017). A convolutional attentional neural network for sentiment classification. In *International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, p. 445–450.

HERLOCKER J. L., KONSTAN J. A. & RIEDL J. (2000). Explaining collaborative filtering recommendations. In *ACM Conference on Computer Supported Cooperative Work*, CSCW '00, p. 241–250, New York, NY, USA : ACM.

HOWARD J. & RUDER S. (2018). Universal language model fine-tuning for text classification. In *56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 328–339.

HU L., CHEN J., NAIR V. N. & SUDJIANTO A. (2018). Locally interpretable models and effects based on supervised partitioning (LIME-SUP). *CoRR*, **abs/1806.00663**.

LETARTE G., PARADIS F., GIGUÈRE P. & LAVIOLETTE F. (2018). Importance of self-attention for sentiment analysis. In *EMNLP Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*, p. 267–275.

LI X., BING L., LI P., LAM W. & YANG Z. (2018). Aspect term extraction with history attention and selective transformation. *CoRR*, **abs/1805.00760**.

LIN Z., FENG M., DOS SANTOS C. N., YU M., XIANG B., ZHOU B. & BENGIO Y. (2017). A structured self-attentive sentence embedding. In *International Conference on Learning Representations (ICLR)*.

LUO L., AO X., PAN F., WANG J., ZHAO T., YU N. & HE Q. (2018). Beyond polarity : Interpretable financial sentiment analysis with hierarchical query-driven attention. In *27th International Joint Conference on Artificial Intelligence, IJCAI*, p. 4244–4250.

MA Y., PENG H. & CAMBRIA E. (2018). Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm. In *AAAI Conference on Artificial Intelligence*, p. 5876–5883.

MAAS A. L., DALY R. E., PHAM P. T., HUANG D., NG A. Y. & POTTS C. (2011). Learning word vectors for sentiment analysis. In *49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies - Volume 1*, HLT '11, p. 142–150, Stroudsburg, PA, USA : Association for Computational Linguistics.

MÄNTYLÄ M. V., GRAZIOTIN D. & KUUTILA M. (2018). The evolution of sentiment analysis - a review of research topics, venues, and top cited papers. *Computer Science Review*, **27**, 16–32.

MERITY S., KESKAR N. S. & SOCHER R. (2018). Regularizing and optimizing LSTM language models. In *International Conference on Learning Representations (ICLR)*.

MERITY S., XIONG C., BRADBURY J. & SOCHER R. (2016). Pointer sentinel mixture models. *CoRR*, **abs/1609.07843**.

MOHAMMAD S. M. (2018). Word affect intensities. In *11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.

MOHAMMAD S. M. & TURNEY P. D. (2013). Crowdsourcing a word-emotion association lexicon. In *Computational Intelligence*, volume 29, p. 436–465.

MOKHTARI S., LI T. & XIE N. (2018). Context-sensitive neural sentiment classification. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, p. 293–299.

MOLNAR C. (2019). *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/.

PETERS M., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of*

the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers), p. 2227–2237, New Orleans, Louisiana : Association for Computational Linguistics.

RIBEIRO M. T., SINGH S. & GUESTRIN C. (2016). "why should i trust you ?" : Explaining the predictions of any classifier. In *22nd ACM International Conference on Knowledge Discovery and Data Mining SIGKDD*, p. 1135–1144, New York, NY, USA : ACM.

SU J., ZENG J., XIONG D., LIU Y., WANG M. & XIE J. (2018). A hierarchy-to-sequence attentional neural machine translation model. In *IEEE/ACM Trans. Audio, Speech & Language Processing*, volume 26, p. 623–632.

TUTEK M. & ŠNAJDER J. (2018). Iterative recursive attention model for interpretable sequence classification. In *EMNLP Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*, p. 249–257.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U. & POLOSUKHIN I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30*, p. 5998–6008. Curran Associates, Inc.

VELLIDO A., MARTÍN-GUERRERO J. D. & LISBOA P. J. G. (2012). Making machine learning models interpretable. In *European Symposium on Artificial Neural networks, computational intelligence and machine learning*.

VOULODIMOS A., DOULAMIS N., DOULAMIS A. & PROTOPAPADAKIS E. (2018). Deep learning for computer vision : A brief review. In *Computational Intelligence and Neuroscience*, volume 2018, p. 1–13.

WAN L., ZEILER M., ZHANG S., CUN Y. L. & FERGUS R. (2013). Regularization of neural networks using dropconnect. In *30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, p. 1058–1066, Atlanta, Georgia, USA : PMLR.

WANG S., HUANG M. & DENG Z. (2018). Densely connected cnn with multi-scale feature attention for text classification. In *IJCAI*.

YANG Z., YANG D., DYER C., HE X., SMOLA A. & HOVY E. (2016). Hierarchical attention networks for document classification. In *Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1480–1489 : Association for Computational Linguistics.

YOUNG T., HAZARIKA D., PORIA S. & CAMBRIA E. (2018). Recent trends in deep learning based natural language processing [review article]. In *IEEE Computational Intelligence Magazine*, volume 13, p. 55–75.

ZHANG X., ZHAO J. & LECUN Y. (2015). Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28*, p. 649–657.