Plongements Interprétables pour la Détection de Biais Cachés

Tom Bourgeade, Philippe Muller, Tim van de Cruys

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

Plongements Interprétables pour la Détection de Biais Cachés

Tom Bourgeade[1] Philippe Muller[1] Tim Van de Cruys[2]
(1) IRIT, Université Toulouse 3, 31062 Toulouse, France
(2) Leuven.AI institute, KU Leuven, 3000 Louvain, Belgique
tom.bourgeade@irit.fr, muller@irit.fr, tim.vandecruys@kuleuven.be

RÉSUMÉ

De nombreuses tâches sémantiques en TAL font usage de données collectées de manière semi-automatique, ce qui est souvent source d'artefacts indésirables qui peuvent affecter négativement les modèles entraînés sur celles-ci. Avec l'évolution plus récente vers des modèles à usage générique pré-entraînés plus complexes, et moins interprétables, ces biais peuvent conduire à l'intégration de corrélations indésirables dans des applications utilisateurs. Récemment, quelques méthodes ont été proposées pour entraîner des plongements de mots avec une meilleure interprétabilité. Nous proposons une méthode simple qui exploite ces représentations pour détecter de manière préventive des corrélations lexicales faciles à apprendre, dans divers jeux de données. Nous évaluons à cette fin quelques modèles de plongements interprétables populaires pour l'anglais, en utilisant à la fois une évaluation intrinsèque, et un ensemble de tâches sémantiques en aval, et nous utilisons la qualité interprétable des plongements afin de diagnostiquer des biais potentiels dans les jeux de données associés.

ABSTRACT

Interpretable Embeddings for Hidden Biases Detection

A lot of current semantic NLP tasks use semi-automatically collected data, that are often prone to unwanted artifacts, which may negatively affect models trained on them. With the more recent shift towards more complex, and less interpretable, pre-trained general purpose models, these biases may lead to undesirable correlations getting integrated into end-user applications. Recently a few methods have been proposed to train word embeddings with better interpretability. We propose a simple setup which exploits these representations to preemptively detect easy-to-learn lexical correlations in various datasets. We evaluate a few popular interpretable embedding models for English for this purpose, using both an intrinsic evaluation, and a large set of downstream semantic tasks, and we make use of the embeddings' interpretable quality in order to diagnose potential biases in the associated datasets.

MOTS-CLÉS : Interprétabilité, Plongements lexicaux, Biais.

KEYWORDS: Interpretability, Word embeddings, Bias.

1 Introduction

Les modèles de plongements de mots sont une méthode populaire et efficace pour l'association de tokens linguistiques à des représentations vectorielles, qui peuvent ensuite être exploitées par des architectures de réseaux de neurones dans le cadre de tâches diverses en traitement automatique des langues (TAL). Les modèles de plongements denses, tels que word2vec (Mikolov *et al.*, 2013), GloVe (Pennington *et al.*, 2014), ou fastText (Bojanowski *et al.*, 2017), font correspondre les mots de leur vocabulaire à des vecteurs denses de quelques centaines de dimensions (généralement 300 ou 500),

dérivés de manière non supervisée (ou auto-supervisée) de statistiques de cooccurrences extraites de grands corpus textuels, et existent pour de nombreuses langues. Ce type de modèles permettent d'obtenir de très bons résultats sur de nombreuses tâches de TAL en aval. Souvent, un simple calcul de moyenne (Arora *et al.*, 2017) ou d'addition matricielle (Kober *et al.*, 2017) des représentations de plusieurs mots peut donner des représentations efficaces des phrases, qui peuvent être directement exploitées par des modèles de classification tout aussi simples - fréquemment avec des performances étonnamment bonnes. Ces résultats indiquent que les modèles de plongements de mots denses ont tendance à capturer les informations sémantiques dans les énoncés en langage naturel. Cependant, le manque d'interprétabilité est un problème important pour la majorité de ces modèles, car il est pratiquement impossible de caractériser qualitativement la sémantique des différentes dimensions de la représentation d'un mot (voir par exemple fastText dans la Table 1).

Les notions d'interprétabilité et d'explicabilité sont difficiles à définir en tant que telles en règle général, mais ici nous nous intéressons en particulier à ces notions comme outils de détection et de caractérisation de biais dans des jeux de données. Cette notion de biais peut être associée à la notion concrète de décalage distributionnel des labels (Torralba & Efros, 2011; He *et al.*, 2019) entre un ensemble d'entraînement et un ensemble de test (à condition que ce dernier soit bien choisi, en particulier pour mettre en exergue cette notion, ce qui n'est malheureusement pas toujours le cas).

Dans le cas des modèles de plongements denses, la difficulté de comprendre comment les valeurs de chaque dimension se traduisent en informations sémantiques encodées se propage aux modèles de TAL exploitant ces plongements, excluant donc la possibilité d'expliquer directement, de manière numérique, le comportement d'un modèle, même linéaire, en fonction de ces dimensions – les méthodes d'explicabilité *a posteriori* en TAL emploient de ce fait le plus souvent une analyse en fonction de la présence ou non de vecteur-mots entiers, comme par exemple dans Li *et al.* (2016) ou Ribeiro *et al.* (2016) – ce qui rend difficile la détection de biais lexicaux cachés. Une alternative existe cependant, sous la forme de modèles de plongements interprétables, dont les dimensions, par construction, sont plus aptes à fournir des indications sur les champs lexicaux associés aux mots encodés.

Ces représentations auraient pour avantages de permettre des analyses de corrélations lexicales à des niveaux plus abstraits (liés aux aspects sémantiques encodés dans ces plongements) et donc plus facilement exploitables de manière générale, que la simple présence ou non de mots particuliers dans les entrées d'un modèle par exemple.

Une façon d'utiliser ces plongements interprétables dans la pratique serait de détecter et de réparer les biais ou les artefacts d'annotation/construction éventuellement présents dans les ensembles de données, appris de façon indésirable par les modèles de prédiction formés sur eux. Bien que beaucoup de travaux ont été réalisés pour isoler et corriger ces problèmes, les méthodes existantes exigent presque toujours une connaissance préalable de leur présence et de leur nature, et celle-ci est souvent acquise après une analyse qualitative du comportement de modèles suspects sur des tâches en aval, parfois des années après les faits (par exemple, l'ensemble de données SNLI, voir la section 2). Notre objectif est de détecter ces problèmes avant qu'ils ne s'infiltrent et n'entachent les résultats des recherches ultérieures ou les applications utilisateurs finales.

Nous proposons donc une approche simple pour diagnostiquer qualitativement les problèmes potentiels dans les ensembles de données de TAL. L'idée clé est d'exploiter les caractéristiques des modèles de plongements de mots interprétables existants comme indices de biais éventuels présents dans les données : nous entraînons un classifieur CBOW (*Continuous Bag-Of-Words*) intentionnellement élémentaire et de ce fait fondamentalement interprétable (par analyse direct des paramètres appris),

qui utilise simplement la moyenne des représentations des mots d'un texte comme caractéristiques d'entrée, qui sont ensuite introduites dans une couche de régression *Softmax*. En analysant les performances et les paramètres appris par cette couche de classification, qui correspondent chacun à une dimension interprétable dans la matrice de plongements choisie, nous sommes en mesure d'obtenir des indications sur d'éventuels biais lexicaux faciles à apprendre et à exploiter, dû à la nature élémentaire du modèle en question, et de déterminer qualitativement s'ils sont attendus ou non.

Nos principales contributions sont les suivantes : (1) une nouvelle variante d'une méthode d'évaluation intrinsèques de détection d'intrusion de mots (*word intrusion detection*), appliquée à divers modèles de plongements interprétables populaires en anglais, suivie par, (2) une évaluation extrinsèque de ces mêmes modèles par rapport à un ensemble de tâches de TAL en aval qui pourraient potentiellement mettre en évidence des artefacts d'annotation intéressants, et (3) une analyse de certains classifieurs produits qui démontrent le potentiel de cette approche pour analyser des ensembles de données.

2 Travaux Connexes

Un certain nombre d'approches ont été proposées pour l'induction de plongements interprétables, qui peuvent être divisés en deux catégories différentes : les modèles de plongements fondés sur des contraintes, et les modèles enrichis avec des informations *a priori*. La majorité des modèles de la première catégorie se concentrent sur deux types de contraintes qui améliorent l'interprétabilité des vecteurs de plongements des mots : la parcimonie et la non-négativité. Un large éventail de contributions (Lee & Seung, 1999; Fyshe *et al.*, 2014; Faruqui *et al.*, 2015; Dahiya *et al.*, 2016; Trifonov *et al.*, 2018; Subramanian *et al.*, 2018) ont montré que ces deux propriétés améliorent considérablement la capacité à comprendre à quoi correspond chaque dimension dans une représentation de mot en termes de sémantique abstraite. La parcimonie signifie que le nombre de dimensions différentes par lesquels un mot peut être encodé est limité, ce qui permet de les répertorier et de les analyser de manière relativement exhaustive. La non-négativité quant à elle signifie que les valeurs de chaque dimension peuvent être interprétées comme une "participation" relative de la sémantique associé dans la représentation. Les plongements peuvent être créés en imposant des contraintes à la méthode de construction (Murphy *et al.*, 2012; Panigrahi *et al.*, 2019), ou les contraintes peuvent être imposées comme étape supplémentaire *a posteriori*, en transformant les vecteurs de plongements de mots denses standards en représentations de facto plus éparses et moins bruyantes, par exemple par l'utilisation d'algorithmes de rotation de base connus de l'analyse en composantes principales et de l'analyse factorielle (Park *et al.*, 2017; Dufter & Schütze, 2019), ou par la factorisation en matrices non négatives (Faruqui *et al.*, 2015; Subramanian *et al.*, 2018).

Une seconde avenue de recherche tente d'injecter des informations sémantiques *a priori* dans les modèles de plongements afin d'améliorer leur interprétabilité : Hurtado Bodell *et al.* (2019) utilisent des informations préalables, sous la forme de paires de mots censés être discriminés par une dimension particulière (*homme-femme* pour une dimension de genre par exemple), afin de guider les plongements appris vers des formes plus facilement interprétables ; Fyshe *et al.* (2014) incorporent des données d'activation cérébrale – recueillies auprès des participants pendant qu'ils lisent les mots associés – dans un modèle de plongements de mots interprétables basé sur des contraintes, *Non-Negative Sparse Embedding* (NNSE) (Murphy *et al.*, 2012). Nous nous concentrons ici sur le premier type d'approches de représentations interprétables, les plongements construits sous contraintes, car elles peuvent être employées plus facilement dans des contextes similaires à ceux pour lesquels des modèles de plongements denses sont plus habituellement employés.

Ces dernières années, un certain nombre d'artefacts statistiques indésirables ont été découverts

fastText	NMF300
tortricidae, baronetage, poaceae, prószyński	desktop, server, linux, microsoft, firmware
eum, cydia, inj, papaya, honeydew	leaved, eucalyptus, trees, planted, juniper
kapamilya, inkigayo, noosa, pvo, puso	flavored, dessert, drinks, chocolate, drink
NNSE	**SPOWV**
mango, raspberry, peach, lemon, pear	onion, sauce, pradesh, streak, salad
strawberries, peaches, oranges, pears, apples	scout, scouts, fellows, dry, cub
fir, birch, pine, willow, spruce	malayalam, leopard, grape, karnataka, raft
SPINE	**Word2Sense**
grape, wine, wines, vineyards, winery	ipod, iphone, apple, mini, lansing
linux, windows, macintosh, playstation, xbox	macintosh, intel, apple, mac, dell
bread, lime, dessert, 1/4, apples	apples, citrus, fruits, ripe, berries

TABLE 1 – Comparatif qualitatif des dimensions des différents modèles de plongements de mots utilisés ici : sont listés ici les 5 mots associés avec la valeur la plus grande dans les 3 dimensions les plus actives pour le mot “apple”. On remarque que le modèle dense fastText n’est manifestement pas interprétable dans ce sens, tandis que différentes sémantiques du mot (y compris celles liées à la société informatique homonyme) ont été capturées par les modèles interprétables.

dans des ensembles de données de TAL bien connus et largement utilisés, par exemple dans la tâche d’inférence en langage naturel (NLI), en particulier dans le jeu de données SNLI (Stanford Natural Language Inference) introduit par Bowman *et al.* (2015) ainsi que dans sa variante améliorée multi-genres MNLI (Williams *et al.*, 2018), pour lesquelles Gururangan *et al.* (2018) et Poliak *et al.* (2018) ont par exemple découvert que des modèles “hypothèse-seule”, qui ne reçoivent qu’une partie de l’entrée, peuvent correctement prédire les étiquettes de grandes parties de ces corpus, ce qui indique la présence de corrélations indésirables, causées en partie par l’annotation de données par myriadisation. McCoy *et al.* (2019) montrent également que ces modèles ont tendance à reposer sur des heuristiques ad hoc “faciles” (quantité de chevauchements de mots, par exemple), elles aussi indicatives de problèmes dans les données d’apprentissage pour cette tâche. Afin de surmonter ces problèmes, He *et al.* (2019) conçoivent une méthode pour entraîner des modèles débiaisés sur ces corpus ; pour ce faire, ils entraînent d’abord un modèle biaisé qui exploite principalement les artefacts indésirables (en ne lui fournissant que des informations incomplètes, comme dans le cas des modèles à hypothèse-seule), et entraînent ensuite un nouveau modèle théoriquement débiaisé sur les résidus (dans le sens des instances avec une faible erreur) du classifieur biaisé obtenu précédemment.

3 Plongements Interprétables

Nous nous sommes concentrés ici sur quatre différents modèles de plongements interprétables non-négatifs et parcimonieux que l’on peut trouver dans la littérature pertinente, avec des niveaux de complexité variables :

- NNSE [1] (Murphy *et al.*, 2012), construit par reconstruction avec erreur quadratique modifiée de statistiques de concurrences (inspiré de la méthode *Non-Negative Sparse Coding* de Hoyer (2002)) collectées sur le jeu de données web anglais ClueWeb09 ;

1. http://www.cs.cmu.edu/~bmurphy/NNSE/

- SPOWV [2] (Faruqui *et al.*, 2015) et SPINE [2] (Subramanian *et al.*, 2018), tout deux construits directement sur un modèle de plongements denses existant (ici, nous utilisons les versions calculées sur GloVe), le premier par factorisation matricielle, et le second à l'aide d'un auto-encodeur k-éparse ;
- Word2Sense [3] (Panigrahi *et al.*, 2019), construit à l'aide d'une approche fondée sur l'allocation de Dirichlet latente, sur une combinaison des jeux de données UKWAC et Wackypedia ;

A ces modèles existants, nous avons ajouté nos propres plongements interprétables non-négatifs parcimonieux, NMF300, construits simplement par factorisation en matrices non-négatives (avec une largeur de 300 dimensions) à l'aide des règles de mise-à-jour multiplicatives définies par Lee & Seung (2001) pour la métrique de divergence de Kullback-Leibler, sur des statistiques de cooccurrences issues d'articles Wikipedia. Des cinq modèles étudiés ici, celui-ci est le plus simple, dans le sens où la seule contrainte présente lors de la factorisation est la non-négativité, nous l'employons comme une sorte de modèle étalon comparatif.

Tous les modèles de plongements utilisés ici sont non-contextualisés et statiques par nature, cependant, comme le montre Kober *et al.* (2017), l'utilisation d'une simple opération de composition (addition matricielle ou moyennage, par exemple) sur la séquence de représentations vectorielles correspondant à une phrase permet en pratique de désambiguïser contextuellement les différents sens encodés dans les plongements de mots. De plus, tous ces modèles ayant été originalement construits dans des contextes très différents (taille des plongements : 300 pour NNSE, 1000 pour SPOWV et SPINE, et 2250 pour Word2Sense ; tailles des vocabulaires ; corpus sources ; etc.), et en accord avec la tendance actuelle de mener des expériences avec des modèles pré-entraînés, nous avons fait le choix d'utiliser les versions fournies par leurs auteurs respectifs, même si de ce fait les paramètres de leur construction sont assez hétérogènes. Cette hétérogénéité est difficile à contrôler puisqu'un réglage fin de ces paramètres aurait été nécessaire à réaliser indépendamment pour chaque approche même sur un jeu de données d'entraînement unifié.

4 Jeux de Données Utilisés

Nous présentons ici rapidement les jeux de données annotées ou collectées que nous avons analysés pour détecter d'éventuels biais indésirables. Ils représentent diverses tâches de classification simple ou de relations, couvrant différents aspects sémantiques et genres textuels :

- **IMDB** (Maas *et al.*, 2011) est une collection de critiques de films (petits paragraphes) recueillies sur le site web IMDB, annotées avec des étiquettes de sentiments binaires (positif/négatif), dérivées des scores des critiques.
- **BoolQ** (Clark *et al.*, 2019) est un jeu de données de questions oui/non "d'origine naturelle", issues de requêtes sur un moteur de recherche, associées à des passages textuels issues d'articles Wikipedia pertinents permettant normalement de répondre à la question.
- **Sarcasm** (Oraby *et al.*, 2016) est un recueil d'extraits de débats de forums en ligne, composé d'une déclaration et d'une réponse, qualifiées de sarcastiques ou non.
- **UR-FUNNY** (Hasan *et al.*, 2019) est un ensemble de données multimodales créé pour soutenir l'analyse de l'humour, en intégrant le langage naturel, la parole et la vidéo. Nous utilisons ici la partie textuelle seule, composée d'instances de paires d'un texte contexte et d'une phrase de chute, étiquetées comme humoristiques ou non.

2. https://github.com/harsh19/SPINE#word-embeddings
3. https://github.com/abhishekpanigrahi1996/Word2Sense#pretrained-vectors

- **SST** (Socher *et al.*, 2013) (Stanford Sentiment Treebank) est un recueil de phrases tirées de critiques de films, annotées pour la polarité des sentiments au niveau des phrases des arbres de syntaxe. Nous utilisons ici l'annotation en cinq classes de haut niveau fournie, ramenée aux trois classes "positive", "négative" et "neutre" (en fusionnant les classes originales "très positive/négative" avec leur équivalent correspondant).
- **SNLI** (Bowman *et al.*, 2015) est une collection de paires de phrases ou de descriptions textuelles, conçue pour tester la capacité des modèles à prédire les relations inférentielles. Les étiquettes possibles pour la relation sont "inférence" (*entailment*), "contradiction" ou "neutre".
- **Emergent** (Ferreira & Vlachos, 2016) est un jeux de données pour la classification du positionnement, où chaque instance est composée d'une affirmation et de titres d'articles de journaux portant sur cette affirmation, avec trois labels de positionnement possible : "pour" (*for*) si les titres supportent l'affirmation, "observe" (*observing*) si les titres n'affichent pas de prise de position clair, ou "contre" (*against*) s'il contredisent l'affirmation.
- **PDTB** (Prasad *et al.*, 2008) (Penn Discourse TreeBank) est une partie du corpus Penn TreeBank, annoté de relations *rhétoriques*, soit entre les clauses d'une phrase, soit entre des phrases voisines. Nous utilisons ici seulement les 11 classes présentes dans le jeu de test (le jeu d'entraînement en contient normalement 16), ce qui est la pratique courante.

Nous répertorions également les caractéristiques de ces jeux de données dans la Table 2.

Corpus	E	T	C	Equilibrage/Classes principles
IMDB	25000	22500	2	eq.
BoolQ	9427	2943	2	true=62.3%, false=37.7%
Sarcasm	3754	469	2	eq.
UR-FUNNY	8074	1058	2	eq.
SST	8544	1989	3	positive=42.0%, negative=39.2%, neutral=18.8%
SNLI	549367	9824	3	entailment=33.4%, contradiction=33.3%, neutral=33.3%
Emergent	2076	259	3	for=47.7%, observing=37.0%, against=15.3%
PDTB	12907	1085	11	cause=26.5%, conjunction=22.1%, restatement=19.1%, contrast=12.4%, *reste*=19.9%

TABLE 2 – Statistiques des différents corpus utilisés : E/T = Taille des jeux d'Entraînement/Test respectivement ; C = Nombre de Classes. Pour PDTB, seules les 4 classes majoritaires sont répertoriées (ce corpus présente un fort déséquilibre entre les classes, notamment dans l'ensemble de test, où certaines classes ne sont pas du tout représentées).

5 Expériences et Résultats

5.1 Détection d'Intrusion de Mots

Afin d'évaluer qualitativement l'interprétabilité des différents modèles de plongements sous contraintes explorés ici, nous modifions la méthode d'évaluation de la détection d'intrusion de mots introduite dans Chang *et al.* (2009), qui semble être devenue la norme de facto à cet effet au fil des ans (Murphy *et al.*, 2012; Fyshe *et al.*, 2014; Faruqui *et al.*, 2015; Subramanian *et al.*, 2018). On peut la résumer ainsi : étant donné un échantillon mélangé de mots (généralement 4 ou 5) choisis

parmi les mots les plus “actifs” pour une dimension d’une matrice de plongements interprétables (dans le cas d’une matrice non négative, les mots ayant les plus grandes valeurs dans cette dimension particulière), auquel est ajouté un mot “intrus”, choisi parmi les mots les moins actifs pour cette dimension, un évaluateur humain peut-il trouver l’intrus ?

Nous avons ici employé une variante plus difficile de cette méthode, compte-tenu de la variété de modèles explorés, sur un échantillon de 50 dimensions pour chaque matrice de plongements (ces dimensions correspondant à la dimension la plus active pour 50 mots tirés dans l’intersection des vocabulaires des 5 modèles évalués ici), avec 3 évaluateurs (auteurs de cet article, en aveugle). Dans la variante “classique” de cette tâche, le mot intrus est généralement choisi parmi les mots les moins actifs de la dimension évaluée, et parmi les mots plus actifs d’une autre dimension (le plus souvent choisi aléatoirement). Après des essais sur de petits échantillons d’instances produites par cette première approche, nous avons remarquer des difficultés à différencier les différents modèles selon les performances obtenues, car la tâche semble trop “facile” pour la plupart des modèles interprétables (à l’exception de `SPOWV`). De plus, conceptuellement, choisir la dimension fortement active du mot intrus aléatoirement ne permet pas vraiment de distinguer la caractéristiques discriminante de la dimensions étudiée : idéalement, il faudrait pouvoir évaluer la dimension cible indépendamment des autres dimensions actives potentiellement communes aux “vrais” mots. Pour ce faire, nous avons modifier le processus de sélection de l’intrus : celui-ci est similairement choisi parmi les 10% des mots les moins actifs de la dimension cible, mais avec comme second critère le fait d’avoir pour seconde dimension la plus active la seconde dimension la plus active “commune” aux vrais mots (après expérimentation, nous sélectionnons celle avec la plus grande médiane parmi les vrais mots). Qualitativement, cela a pour effet d’augmenter significativement la difficulté à repérer l’intrus, notamment quand la dimension cible est plus ou moins associée à un champ lexical peu spécifique.

Voici un exemple d’application de cette méthode pour la dimension cible n°110 du modèle `SPINE` pré-entraîné utilisé ici : les 10 mots les plus actifs de cette dimension sont en ordre décroissant *“pius, pope, diocese, bishops, basilica, archdiocese, benedict, vatican, catholic, bishop”*, les 4 premiers mots les plus actifs sont donc sélectionnés comme vrais mots. Leur seconde dimension commune la plus active (au sens de la seconde médiane la plus élevée des valeurs de leurs dimensions) est la dimension n°178, avec pour mots les plus actifs *“baptist, jesus, christians, holy, lutheran, religious, judaism, believers, prayers, baptism”*, qui visiblement semble couvrir des champs lexicaux proches de ceux de la dimension cible. Dans la version “classique” de la tâche, on sélectionnerait ici un intrus aléatoirement parmi les 10% des mots les moins actifs de la dimension cible, dont *“baseline, sculptures, feedback, armoured, modeled”* serait un échantillon de 5 mots. Dans la variante plus difficile présentée ici, un même échantillon de 5 mots serait de plus tiré parmi les mots les plus actifs dans la dimension n°178 dans cette proportion, ici *“judaism, mormon, preacher, buddhism, meditation”*. On peut qualitativement observer que ce second échantillon est plus proche des 4 vrais mots *“pius, pope, diocese, bishops”* que l’échantillon aléatoire, mais en même temps semble démontrer les spécificités uniques de ces deux dimensions (la dimension n°110 semble ici plus spécifique aux termes associés à la religion catholique, tandis que la n°178 semble correspondre à des termes religieux plus généraux).

Nous n’avons pas effectué cette évaluation complète sur un modèle de plongements denses car après essai sur un échantillon, la précision des évaluateurs était équivalente à un choix aléatoire (traduisant la non-interprétabilité de ces modèles), ce qui apparaît également dans les résultats de Subramanian *et al.* (2018) par exemple.

Exemple d'instance la tâche : *"dermatologist, columnist, veterinarian, psychiatrist, pathologist"* est un ensemble de mots pour une dimension particulière de `NNSE`, où *"columnist"* est l'intrus ici.

Modèle	Précision moyenne évaluateurs	Accord Inter-évaluateurs	Kappa de Fleiss
`NMF300`	76%	**94%** ; 72%	0.74
`NNSE`	**79%**	90% ; **74%**	**0.76**
`SPOWV`	38%	84% ; 34%	0.43
`SPINE`	**79%**	92% ; 60%	0.63
`Word2Sense`	65%	88% ; 56%	0.61

TABLE 3 – Résultats de l'évaluation de la Détection d'Intrusion de Mots effectuée sur les 5 modèles considérés (L'accord est donné sous la forme : `majorité ; unanimité`).

Nous notons tout d'abord que nos résultats (Table 3) sont à peu près similaires à ceux de Subramanian *et al.* (2018) et Panigrahi *et al.* (2019) pour les modèles `SPOWV`, `SPINE` et `Word2Sense`, compte tenu des légères différences dans le dispositif expérimental, et de la nature intrinsèquement subjective de l'évaluation. Si les performances globales de tous les modèles sont assez bonnes, nous constatons après une analyse qualitative des dimensions des modèles que leur interprétabilité est très hétérogène : si la majorité est relativement facile à associer à des champs lexicaux précis, certaines semblent capter des phénomènes pseudo-lexicaux, selon les corpus dont ils sont issus. Par exemple, les modèles entraînés sur les articles de Wikipedia peuvent être la proie de certains artefacts de fréquences causés par des données tabulaires très répétitives, ou, plus subtilement, par des ensembles d'articles appartenant aux mêmes domaines. Par exemple, des articles sur un sport particulier conduisent à des articles sur des équipes sportives particulières, qui conduisent à des articles sur des joueurs sportifs particuliers, etc, occupant une place dans le corpus disproportionnée.

5.2 Évaluation sur les Tâches en Aval

Nous avons également choisi d'évaluer l'interprétabilité et l'utilité des plongements en les employant dans des tâches de classification de texte, et en vérifiant si certaines dimensions jouaient un rôle important dans les prédictions, mettant éventuellement au jour des corrélations lexicales indésirables faciles à apprendre. Pour chaque tâche, nous fabriquons donc un classifieur de régression *Softmax* élémentaire (avec une matrice de paramètres de taille $|dimensions\ des\ plongements| \times |classes|$, ainsi que les biais), prenant en entrée la moyenne des plongements interprétables des mots de la phrase, ou, dans le cas des tâches avec deux phrases en entrée, la concaténation vectorielle suivante (inspirée de Conneau *et al.* (2017)) : $< u, v, |u - v|, u * v >$ (u/v : moyenne des vecteurs première/deuxième phrase ; $|x|$: valeur absolue terme à terme du vecteur x ; $*$: opérateur de produit terme-à-terme). Pour chaque couple $(\text{modèle}, \text{corpus})$, nous entraînons le classifieur approprié pour un maximum de 200 époques en utilisant l'optimiseur Adam, avec 50 époques antérieures de réglages fins automatiques des hyperparamètres, en utilisant l'algorithme *Tree-structured Parzen Estimator* (Bergstra *et al.*, 2013, 2011), via son implémentation par la bibliothèque `optuna` (Akiba *et al.*, 2019).

Nous évaluons ensuite chaque classifieur produit sur son jeu de test respectif, et affichons les performances globales pour le corpus dans la Table 4. En plus des cinq modèles de plongements interprétables, nous avons également entraîné de la même manière un ensemble de classifieurs en utilisant les plongements en anglais dense `fastText` (Bojanowski *et al.*, 2017) (sans information

sur les sous-mots)[4], pour comparer l'efficacité des modèles interprétables contraints par rapport aux modèles denses, dans cette configuration de classification élémentaire. Nous présentons également les résultats du classifieur de base "factice" (Dummy) équivalent, qui génère des prédictions au hasard, en suivant la distribution des classes sur l'ensemble de test pour chaque corpus.

Modèle \ Corpus	IMDB	BoolQ	Sarcasm	UR-FUNNY	SST	SNLI	Emergent	PDTB
NMF300	67.8	62.6	60.5	57.7	54.6	58.6	50.9	33.2
NNSE	78.7	63.6	63.9	59.9	60.6	56.3	66.8	31.1
SPOWV	81.9	**66.9**	**70.5**	**65.0**	62.9	62.9	**72.2**	**36.6**
SPINE	81.3	65.9	67.8	63.6	59.9	64.1	**72.2**	34.5
Word2Sense	**82.2**	66.2	67.3	63.9	61.4	**65.5**	69.8	34.2
Dummy *(baseline)*	50.5	53.5	53.0	52.5	39.5	33.6	41.3	19.3
fastText	82.0	63.7	70.1	64.5	**64.4**	61.3	69.5	33.4
*Modèles Dédiés**	96.8	76.9	74†	64.4	96	91.5	73	48

TABLE 4 – Résultats de l'approche sur les tâches de classification en aval évalués. Les scores de justesse pour chaque paire modèle-corpus sont indiqués en pourcentages (les meilleurs scores de l'expérience sont en **gras**). *Nous énumérons également les résultats des modèles populaires dédiés à ces tâches, que l'on retrouve dans la littérature et qui atteignent (ou s'approchent) des performances de l'état de l'art, à titre de comparaison (IMDB, SST : Yang *et al.* (2019) ; BoolQ : Clark *et al.* (2019) ; Sarcasm : Oraby *et al.* (2016) ; UR-FUNNY : Hasan *et al.* (2019) ; SNLI : Liu *et al.* (2019) ; Emergent : Ferreira & Vlachos (2016) ; PDTB : Dai & Huang (2019)). †Mesure F1 pour la classe positive (justesse non disponible).

Nous constatons que, de manière assez surprenante, les classifieurs élémentaires entraînés avec des plongements interprétables semblent aussi performants, voire légèrement meilleurs, que leurs équivalents entraînés avec un modèle dense comme fastText. Il semble également que, pour la plupart des tâches étudiées ici, l'approche fonctionne assez bien, compte tenu de la nature simpliste des modèles de classification. Cela semble indiquer que de nombreuses tâches de TAL ont une composante purement lexicale plus ou moins forte, qui peut expliquer des sous-ensembles parfois importants des corpus correspondants, ce qui nous semble à un certain degré être problématique. En effet, s'il est cohérent que certains termes et champs lexicaux soient associés par nature, par exemple, à une classe de sentiment positive ou négative, la non-nécessité de prendre en compte les phénomènes structurels importants comme la négation (ce qui est le cas pour les modèles élémentaires utilisés ici, mais aussi pour une grande variété de modèles même plus complexes, comme le montre par exemple Naik *et al.* (2018), Kassner & Schütze (2020), ou Hossain *et al.* (2020), par construction d'instances où la compréhension de la négation est essentielle pour la classification, sur lesquelles la plupart des modèles de langue de l'état de l'art échouent) pour classifier une partie significative des données semblent indiquer un problème d'adéquation entre celles-ci et la tâche de TAL associée. Bien que ces résultats ne puissent à eux seuls indiquer l'existence de corrélations nécessairement erronées dans les données, les performances relativement importantes obtenus avec ces approches simples pourraient être le signe que certains biais indésirables entachent les ensembles de données.

Globalement, les modèles SPOWV, SPINE et Word2Sense semble être les plus performants, contrai-

4. wiki-news-300d-1M — https://fasttext.cc/docs/en/english-vectors.html

rement aux modèles plus simples, NMF300 et NNSE, avec quelques instances (BoolQ et SST pour le premier) de classes mal voir non-apprises (voir résultats des expériences [5]).

5.3 Diagnostics de Corpus

Dans cette section, nous présentons quelques éléments d'analyse qualitative des modèles entraînés, des fichiers contenant les résultats des évaluations sur les tâches en aval pour chaque modèle, ainsi que les dimensions les plus prédictives pour chaque classe de ces tâches étant à disposition [5] pour plus de détail.

D	M	C	I	P	Mots les plus actifs dans la dimension
IMDB	NNSE	*pos*	192	1.0	*utmost, sheer, immense, tremendous, newfound, unparalleled, ...*
IMDB	NNSE	*neg*	217	1.0	*debris, trash, garbage, lint, rubbish, sludge, dust, dirt, manure, ...*
IMDB	NMF300	*pos*	100	1.0	*imaginative, vivid, lyrical, poetic, realistic, imagery, subtle, ...*
IMDB	NMF300	*pos*	131	0.76	*shakira, lauper, mcentire, yearwood, parton, estefan, streisand, ...*
BoolQ	SPINE	*false*	575	1.0	*leaked, confidential, libby, fbi, classified, memo, leak, intelligence, ...*
BoolQ	SPINE	*false*	841	0.79	*astronaut, soyuz, spacecraft, iss, nasa, astronauts, shuttle, mir, ...*
BoolQ	SPOWV	*true*	758	1.0	*cyclone, katrina, hurricane, disaster, ike, flooded, shear, dolly, ...*
BoolQ	SPOWV	*true*	173	0.83	*tong, lumpur, myanmar, singaporean, kuala, chung, penang, ...*

TABLE 5 – Exemples de paramètres de prédiction appris par les modèles élémentaires : D = Jeu de Données ; M = Modèle de plongements ; C = Classe correspondant au paramètre ; I = Indice de la dimension correspondant au paramètre dans le modèle de plongements ; P = Poids du paramètre, divisé par la valeur du plus grand paramètre (pour cette classe).

IMDB : Il s'agit de l'un des jeux de données pour lequel les performances des modèles élémentaires entraînés sont les plus élevées. Sans trop de surprises, une partie significative des dimensions les plus actives pour les classes "positive" et "négative" semblent correspondre à des champs lexicaux contenant des marqueurs de sentiment appropriés, pour la plupart des modèles. Cependant, pour le modèle NMF300 en particulier, nous avons remarqué plusieurs dimensions associées à un grand nombre de noms de famille et de prénoms (par exemple, 4ème ligne de la Table 5) et qui apparaissent comme prédicteurs forts de la classe "positive". Pour analyser ce biais a priori peu pertinent dans l'absolu pour une tâche d'analyse de sentiment, nous avons utilisé le module de reconnaissance d'entités nommées de la bibliothèque spaCy pour compter les entités nommées de type "personne" dans les critiques du jeu de données, et nous avons constaté une corrélation linéaire faible (coefficient de Pearson $r = 0.124$) entre ces comptages et les classes des instances. Une analyse plus poussée serait nécessaire pour prouver si un modèle non-linéaire exploite en effet cet aspect, ou un aspect plus spécifique encore : il semble en effet que plusieurs dimensions créées par NMF300 sont caractérisées par des noms d'artistes célèbres. On peut noter que 80.68% des critiques du jeu de données contiennent au moins une entité nommée de ce type, ce qui est cohérent avec la pondération élevée du paramètre correspondant à cette dimension.

BoolQ : Sur ce jeu de données, les poids les plus forts portent essentiellement sur des thèmes particuliers : pour les réponses "fausses" ce sont des questions très débattues (de régime, de lois, etc.),

5. https://github.com/TomBourgeade/InterpEmbsForBiasDetection/tree/main/experiments_results

souvent sujette au conspirationnisme (services de renseignements, conquête/exploration spatiale, etc.), ou marquées émotionnellement (avec des adjectifs *dignified* ou des adverbes *dramatically*). Pour les réponses "vraies", ce sont des dimensions plus liées à la science, l'histoire, la géographie, ou la politique, ou des valeurs numériques. Tout ceci peut indiquer un léger biais dans la collecte qui repose sur les requêtes faites par les utilisateurs sur un moteur de recherche, qui est peut-être exploité par les modèles sans qu'ils aient besoin d'analyser la réponse. Néanmoins, comme il s'agit d'une tâche à deux entrées (question et passage), nous pouvons également observer dans quelle partie du vecteur de composition $< u, v, |u - v|, u * v >$ (voir Section 5.2) se trouve les paramètres correspondants les plus importants : pour la plupart des modèles (à l'exception notable de `NMF300` et `Word2Sense`), nous observons que ceux-ci se trouvent dans la partie produit terme-à-terme de la composition, ce qui indique que les modèles élémentaires se base principalement sur des interactions entre la question et le passage dans l'entrée, ce qui est attendu vis-à-vis de la tâche. Les paramètres importants restants sont en revanche le plus souvent situé dans la partie de la composition qui correspond à la question seule, ce qui pourrait indiquer la présence de questions plus ou moins "rhétoriquement" biaisées.

Emergent : On trouve ici également des thématiques cohérentes qui correspondent à des sujets peu controversés. Par exemple une dimension reliée aux animaux est fortement corrélée avec le label "pour" : une revue rapide des articles de journaux du jeu de données confirme que ceux-ci sont pratiquement toujours positifs sur ces sujets. Ce genre de biais semble inhérent à la façon dont le jeu de données est construit à partir de titres de journaux.

Sarcasm et UR-FUNNY : Les données `Sarcasm` révèlent quelques sujets populaires en regardant les résultats avec `NMF300`, les dimensions positives étant associées aux artistes musicaux, avec par exemple les mots les plus actifs suivants *"burnin, dreamin, rmx, blowin, movin"*, ou bien *"lil, ludacris, rapper, dogg, snoop"*, les dimensions négatives étant liées à des sujets médicaux, par exemple la dimension dont les mots actifs sont *"neurology, ophthalmology, oncology"*, ou légaux (*"plaintiffs, plaintiff, court, appeals"*, ou de manière général techniques, ce qui peut indiquer un manque de diversité. On observe des éléments similaires sur les données `UR-FUNNY` notamment avec le modèle `NMF300`, mais en se focalisant plus sur la chute du sketch. Le modèle `NNSE` montre plus de variétés, et des poids mieux répartis, se focalisant plus sur les traits composés des deux représentations en entrée. Les prédicteurs négatifs incluent toujours des dimensions à l'évidence techniques.

PDTB : La prédiction de relations implicites est intéressante car elle mixe des relations sémantiques et pragmatiques difficiles. Tous les modèles ne prédisent que 4-5 des relations, les plus fréquentes : *Cause*, *Contrast*, *Conjunction*, *Restatement* et *Instantiation*. L'analyse des instances semble montrer quelques particularités liés au type de textes (journalistiques). Par exemple, si l'on regarde plus en détail la relation d'Instantiation, elle semble dépendre beaucoup d'une dimension où les termes les plus actifs sont *"educator, historian, lecturer, researcher, scientist, essayist, journalist, curator, critic, playwright"* dans le second argument de la relation. En regardant les exemples de l'instance d'entraînement, nous avons observé que les seuls mots de cette liste qui apparaissent avec cette relation sont *"critic, journalist, scientist"*, dans une vingtaine d'instances. Cela semble indiquer que ce sont principalement des citations qui illustrent un point mentionné dans le premier argument de la relation. Cela est confirmé quand on analyse les autres signaux de citations, qui constituent un tiers de toutes les instances d'Instantiation, ce qui semble très typique de textes journalistiques mais sans doute peu représentatif au-delà. De la même façon, certaines dimensions importantes pour la prédiction des autres classes de relations semblent assez spécifique pour justifier une étude détaillée des exemples traités.

SNLI : Ce jeu de données semble être un cas particulier, avec de nombreuses dimensions différentes, interprétables mais sans liens évidents entre elles (pour tous les modèles), ce qui pourrait s'expliquer en partie par la taille conséquente et donc la variété des données. SNLI a des biais connus (voir section 2), qui sont en partie associés à la syntaxe (négations, phrases prépositionnelles supplémentaires). Ceux-ci sont bien sûr plus difficiles à découvrir avec les plongements utilisés ici, qui sont principalement lexicaux par nature.

Nous constatons que cette méthode d'analyse semble relativement intéressante pour rapidement détecter des corrélations lexicales faciles à apprendre, mais ne suffit pas seule à confirmer leurs natures exactes, leurs magnitudes, ou si elles ont effectivement pour cause la présence de biais erronés dans les données. La création de protocoles de diagnostic plus poussés, guidés par les corrélations relevées ici a priori, serait nécessaire pour confirmer ces aspects, mais ceux-ci pourraient requérir de lourds moyens humains, en particulier s'ils nécessitent des analyses qualitatives d'un grand nombre d'instances.

6 Conclusion et perspectives

Nous avons montré ici comment une méthode simple peut être utilisée pour identifier des biais non voulus dans des données de TAL, en exploitant des plongements lexicaux interprétables. Cela peut permettre de repérer des problèmes avant de mettre en jeu des modèles plus complexes, ou en amont de tâches différentes. Les évaluations intrinsèques et extrinsèques montrent que des plongements interprétables plus récents ont de meilleures performances sur certaines tâches, mais sans pour autant que l'interprétabilité de leurs dimensions soient meilleures.

Parmi les améliorations envisagées, la principale serait de pouvoir déterminer comment évaluer par l'humain les explications produite par une approche, en s'inspirant par exemple des évaluations d'explications de Strout *et al.* (2019).

Une limite de l'approche présentée est de n'avoir accès qu'à des phénomènes lexicaux, ce qui empêcherait de repérer des biais plus structurels (syntaxiques ou discursifs par exemple). Une avenue prometteuse serait de garder des modèles permettant d'encoder des informations de ce type, comme le modèle Transformer BERT (Devlin *et al.*, 2019) et ses descendants, en les entraînant d'une façon interprétable.

Enfin il serait utile de combiner l'approche avec les méthodes qui se focalisent sur l'interprétation d'instances : en plus de repérer des biais au niveau global d'un jeu de données, pouvoir identifier les instances spécifiques qui en sont responsables permettrait de concevoir des protocoles de diagnostic plus poussés, ciblant spécifiquement une ou plusieurs parties problématiques dans les données, afin d'en déterminer les causes et potentiellement les réparer. Une méthode comme la Layer-wise Relevance Propagation (Montavon *et al.*, 2019) par exemple peut diagnostiquer un modèle "suspect" sur les instances identifiées comme source potentielle de biais, permettant de corriger le problème dans les données, ou dans le modèle directement.

Références

AKIBA T., SANO S., YANASE T., OHTA T. & KOYAMA M. (2019). Optuna : A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

ARORA S., LIANG Y. & MA T. (2017). A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In *5th International Conference on Learning Representations, ICLR 2017, Conference Track Proceedings*.

BERGSTRA J., BARDENET R., BENGIO Y. & KÉGL B. (2011). Algorithms for hyper-parameter optimization. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, p. 2546–2554.

BERGSTRA J., YAMINS D. & COX D. D. (2013). Making a science of model search : Hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, p. I–115–I–123.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, **5**, 135–146. DOI : 10.1162/tacl_a_00051.

BOWMAN S. R., ANGELI G., POTTS C. & MANNING C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 632–642, Lisbon, Portugal : Association for Computational Linguistics. DOI : 10.18653/v1/D15-1075.

CHANG J., GERRISH S., WANG C., BOYD-GRABER J. L. & BLEI D. M. (2009). Reading Tea Leaves : How Humans Interpret Topic Models. In Y. BENGIO, D. SCHUURMANS, J. D. LAFFERTY, C. K. I. WILLIAMS & A. CULOTTA, Éds., *Advances in Neural Information Processing Systems 22*, p. 288–296. Curran Associates, Inc.

CLARK C., LEE K., CHANG M.-W., KWIATKOWSKI T., COLLINS M. & TOUTANOVA K. (2019). BoolQ : Exploring the Surprising Difficulty of Natural Yes/No Questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 2924–2936, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1300.

CONNEAU A., KIELA D., SCHWENK H., BARRAULT L. & BORDES A. (2017). Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 670–680, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1070.

DAHIYA Y., TALUKDAR P. & OTHERS (2016). Discovering response-eliciting factors in social question answering : A reddit inspired study. In *Tenth International AAAI Conference on Web and Social Media*.

DAI Z. & HUANG R. (2019). A regularization approach for incorporating event knowledge and coreference relations into neural discourse parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 2976–2987, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1295.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference*

of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers), p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DUFTER P. & SCHÜTZE H. (2019). Analytical Methods for Interpretable Ultradense Word Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 1185–1191, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1111.

FARUQUI M., TSVETKOV Y., YOGATAMA D., DYER C. & SMITH N. A. (2015). Sparse Overcomplete Word Vector Representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 1491–1500, Beijing, China : Association for Computational Linguistics. DOI : 10.3115/v1/P15-1144.

FERREIRA W. & VLACHOS A. (2016). Emergent : a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1163–1168, San Diego, California : Association for Computational Linguistics. DOI : 10.18653/v1/N16-1138.

FYSHE A., TALUKDAR P. P., MURPHY B. & MITCHELL T. M. (2014). Interpretable Semantic Vectors from a Joint Model of Brain- and Text- Based Meaning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 489–499, Baltimore, Maryland : Association for Computational Linguistics. DOI : 10.3115/v1/P14-1046.

GURURANGAN S., SWAYAMDIPTA S., LEVY O., SCHWARTZ R., BOWMAN S. & SMITH N. A. (2018). Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, p. 107–112, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-2017.

HASAN M. K., RAHMAN W., BAGHER ZADEH A., ZHONG J., TANVEER M. I., MORENCY L.-P. & HOQUE M. E. (2019). UR-FUNNY : A Multimodal Language Dataset for Understanding Humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 2046–2056, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1211.

HE H., ZHA S. & WANG H. (2019). Unlearn Dataset Bias in Natural Language Inference by Fitting the Residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, p. 132–142, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-6115.

HOSSAIN M. M., KOVATCHEV V., DUTTA P., KAO T., WEI E. & BLANCO E. (2020). An Analysis of Natural Language Inference Benchmarks through the Lens of Negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 9106–9118, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.732.

HOYER P. (2002). Non-negative sparse coding. In *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing*, p. 557–565. DOI : 10.1109/NNSP.2002.1030067.

HURTADO BODELL M., ARVIDSSON M. & MAGNUSSON M. (2019). Interpretable Word Embeddings via Informative Priors. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language*

Processing (EMNLP-IJCNLP), p. 6324–6330, Hong Kong, China : Association for Computational Linguistics.

KASSNER N. & SCHÜTZE H. (2020). Negated and Misprimed Probes for Pretrained Language Models : Birds Can Talk, But Cannot Fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7811–7818, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.698.

KOBER T., WEEDS J., WILKIE J., REFFIN J. & WEIR D. (2017). One Representation per Word - Does it make Sense for Composition ? In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, p. 79–90, Valencia, Spain : Association for Computational Linguistics. DOI : 10.18653/v1/W17-1910.

LEE D. D. & SEUNG H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**(6755), 788–791. DOI : 10.1038/44565.

LEE D. D. & SEUNG H. S. (2001). Algorithms for Non-negative Matrix Factorization. In T. K. LEEN, T. G. DIETTERICH & V. TRESP, Éds., *Advances in Neural Information Processing Systems 13*, p. 556–562. MIT Press.

LI J., CHEN X., HOVY E. & JURAFSKY D. (2016). Visualizing and Understanding Neural Models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 681–691, San Diego, California : Association for Computational Linguistics.

LIU X., HE P., CHEN W. & GAO J. (2019). Multi-Task Deep Neural Networks for Natural Language Understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 4487–4496, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1441.

MAAS A. L., DALY R. E., PHAM P. T., HUANG D., NG A. Y. & POTTS C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, p. 142–150, Portland, Oregon, USA : Association for Computational Linguistics.

MCCOY T., PAVLICK E. & LINZEN T. (2019). Right for the Wrong Reasons : Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 3428–3448, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1334.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. BURGES, L. BOTTOU, M. WELLING, Z. GHAHRAMANI & K. Q. WEINBERGER, Éds., *Advances in Neural Information Processing Systems 26*, p. 3111–3119. Curran Associates, Inc.

MONTAVON G., BINDER A., LAPUSCHKIN S., SAMEK W. & MÜLLER K.-R. (2019). Layer-Wise Relevance Propagation : An Overview. In W. SAMEK, G. MONTAVON, A. VEDALDI, L. K. HANSEN & K.-R. MÜLLER, Éds., *Explainable AI : Interpreting, Explaining and Visualizing Deep Learning*, p. 193–209. Cham : Springer International Publishing. DOI : 10.1007/978-3-030-28954-6_10.

MURPHY B., TALUKDAR P. & MITCHELL T. (2012). Learning Effective and Interpretable Semantic Models using Non-Negative Sparse Embedding. In *Proceedings of COLING 2012*, p. 1933–1950, Mumbai, India : The COLING 2012 Organizing Committee.

NAIK A., RAVICHANDER A., SADEH N., ROSE C. & NEUBIG G. (2018). Stress Test Evaluation for Natural Language Inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, p. 2340–2353, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

ORABY S., HARRISON V., REED L., HERNANDEZ E., RILOFF E. & WALKER M. (2016). Creating and Characterizing a Diverse Corpus of Sarcasm in Dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 31–41, Los Angeles : Association for Computational Linguistics. DOI : 10.18653/v1/W16-3604.

PANIGRAHI A., SIMHADRI H. V. & BHATTACHARYYA C. (2019). Word2Sense : Sparse Interpretable Word Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 5692–5705, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1570.

PARK S., BAK J. & OH A. (2017). Rotated Word Vector Representations and their Interpretability. In M. PALMER, R. HWA & S. RIEDEL, Éds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, p. 401–411 : Association for Computational Linguistics. DOI : 10.18653/v1/d17-1041.

PENNINGTON J., SOCHER R. & MANNING C. (2014). Glove : Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1532–1543, Doha, Qatar : Association for Computational Linguistics. DOI : 10.3115/v1/D14-1162.

POLIAK A., NARADOWSKY J., HALDAR A., RUDINGER R. & VAN DURME B. (2018). Hypothesis Only Baselines in Natural Language Inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, p. 180–191, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/S18-2023.

PRASAD R., DINESH N., LEE A., MILTSAKAKI E., ROBALDO L., JOSHI A. & WEBBER B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of LREC*.

RIBEIRO M. T., SINGH S. & GUESTRIN C. (2016). "Why Should I Trust You ?" : Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, p. 1135–1144, San Francisco, California, USA : ACM Press. DOI : 10.1145/2939672.2939778.

SOCHER R., PERELYGIN A., WU J., CHUANG J., MANNING C. D., NG A. & POTTS C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, p. 1631–1642, Seattle, Washington, USA : Association for Computational Linguistics.

STROUT J., ZHANG Y. & MOONEY R. (2019). Do Human Rationales Improve Machine Explanations ? In *Proceedings of the 2019 ACL Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*, p. 56–62, Florence, Italy : Association for Computational Linguistics.

SUBRAMANIAN A., PRUTHI D., JHAMTANI H., BERG-KIRKPATRICK T. & HOVY E. H. (2018). SPINE : SParse Interpretable Neural Embeddings. In S. A. MCILRAITH & K. Q. WEINBERGER, Éds., *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, p. 4921–4928 : AAAI Press.

TORRALBA A. & EFROS A. A. (2011). Unbiased look at dataset bias. In *CVPR 2011*, p. 1521–1528, Colorado Springs, CO, USA : IEEE. DOI : 10.1109/CVPR.2011.5995347.

TRIFONOV V., GANEA O.-E., POTAPENKO A. & HOFMANN T. (2018). Learning and Evaluating Sparse Interpretable Sentence Embeddings. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*, p. 200–210, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-5422.

WILLIAMS A., NANGIA N. & BOWMAN S. (2018). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. 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. 1112–1122, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1101.

YANG Z., DAI Z., YANG Y., CARBONELL J., SALAKHUTDINOV R. R. & LE Q. V. (2019). XLNet : Generalized Autoregressive Pretraining for Language Understanding. In H. WALLACH, H. LAROCHELLE, A. BEYGELZIMER, F. D'ALCHÉ BUC, E. FOX & R. GARNETT, Éds., *Advances in Neural Information Processing Systems 32*, p. 5753–5763. Curran Associates, Inc.