Utiliser l'explicabilité des modèles pour mettre en évidence les expressions genrées dans la parole

François Buet, Camille Guinaudeau, Cyril Grouin, Sahar Ghannay, Shin'Ichi Satoh

Utiliser l'explicabilité des modèles pour mettre en évidence les expressions genrées dans la parole

François Buet[1] Camille Guinaudeau[2]
Cyril Grouin[1] Sahar Ghannay[1] Shin'ichi Satoh[3]
(1) Université Paris-Saclay, CNRS, LISN, 91405 Orsay, France
(2) Université Paris-Saclay, CNRS, JFLI, 101-0003 Tokyo, Japon
(3) Nationat Institue of Informatics, 101-0003 Tokyo, Japon
prenom.nom@lisn.upsaclay.fr, satoh@nii.ac.jp

RÉSUMÉ

Dans de nombreux pays, des études ont souligné la sous-représentation des femmes dans les médias. Mais au-delà du déséquilibre quantitatif se pose la question de l'asymétrie qualitative des représentations des hommes et des femmes. Comment automatiser l'évaluation des contenus et des traits saillants spécifiques aux discours masculins et féminins ? Nous proposons dans cette étude d'exploiter les connaissances acquises par un modèle de classification entraîné à la détection du genre sur des transcriptions automatiques, afin de mettre en évidence des motifs distinctifs du discours masculin ou féminin. Notre approche est basée sur l'utilisation de méthodes développées pour l'intelligence artificielle explicable (IAX), afin de calculer des scores d'attribution au niveau des unités.

ABSTRACT

Using model explainability to highlight gendered expressions in speech

In many countries, studies have highlighted the under-representation of women in the media. But beyond quantitative imbalance is the question of the qualitative asymmetry of men's and women's representations. How to automate the evaluation of content and salient features specific to male and female discourse ? We propose in this study to leverage the knowledge acquired by a classification model trained for gender detection on automatic transcripts, in order to highlight patterns distinctive of male or female speech. Our approach is based on the use of methods developed for explainable artificial intelligence (XAI), to compute token-level attribution scores.

MOTS-CLÉS : Détection du genre, explicabilité, médias.

KEYWORDS : Gender Detection, Explainability, Media.

1 Introduction

La représentation du genre dans les médias est une question de société qui a été suivie à l'échelle mondiale (MediaWatch, 1995), afin de garantir l'égalité dans la participation à la vie publique. Dans de nombreux pays, les études ont souligné la moindre présence des femmes dans les organes d'information et les médias en général (GMMP, 2021). Parallèlement à l'analyse manuelle, le développement et la diffusion des chaînes d'outils et des ressources de traitement automatique des langues (TAL) ont permis d'accroître l'ampleur des observations (Ash *et al.*, 2022). Si les outils modernes permettent de détecter le genre du locuteur avec une certaine exactitude (Doukhan *et al.*, 2018), l'éva-

luation qualitative et automatique du contenu des discours masculins et féminins reste une question ouverte. Notre but dans cette étude est d'exploiter les connaissances acquises par un modèle de classification entraîné pour la détection du genre à partir d'une transcription de parole, et de l'associer à l'IA explicable (IAX), afin de mettre en évidence des motifs lexicaux distinctifs du discours masculin ou féminin (Figure 1), et ainsi d'introduire une part d'automatisation dans l'analyse de la parole masculine et féminine dans les médias.

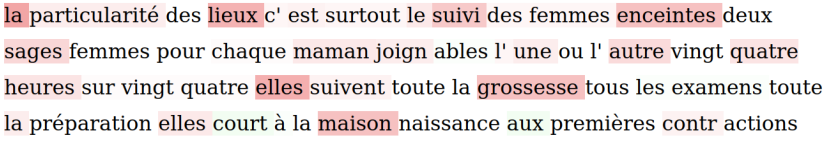


FIGURE 1 – Exemple de visualisation d'une explication issue de nos expériences. La teinte de rouge indique l'attribution à la classe féminine. L'énoncé (transcrit automatiquement) provient d'un journal radiophonique.

Les méthodes d'explicabilité sont conçues pour donner un aperçu des raisons qui sous-tendent les prédictions des modèles (Marcinkevics & Vogt, 2020). Elles peuvent s'avérer utiles soit aux utilisateurs de l'IA qui doivent s'assurer de leur confiance dans les prédictions du modèle, soit aux utilisateurs finaux qui veulent comprendre les décisions qui les concernent, soit aux développeurs et aux scientifiques des données qui doivent vérifier la robustesse de leur système. Nous proposons dans cette étude d'utiliser des techniques d'explicabilité répandues (Zeiler & Fergus, 2014 ; Ribeiro *et al.*, 2016 ; Sundararajan *et al.*, 2017) afin de calculer des explications locales pour la prédiction du genre du locuteur, sous la forme d'attributions au niveau des unités dans chaque segment de la transcription de parole. Il convient de souligner que notre problème principal n'est pas la classification du genre du locuteur (une tâche pour laquelle la modalité audio est plus adaptée), mais d'apporter une assistance pour l'analyse des discours masculins et féminins dans les médias. En masquant les informations acoustiques (par l'utilisation d'une transcription automatique pour entrée), nous cherchons à nous concentrer sur des indicateurs textuels permettant de discriminer les genres. Nos expériences se fondent sur des données provenant de programmes de télévision et de radio français et japonais. Nos principales contributions sont les suivantes : une première tentative (à notre connaissance) de détection du genre à partir de transcriptions automatiques de la parole (Section 3), et a fortiori l'application de l'IAX à cette détection (Section 4).

2 Représentation du genre dans les médias

2.1 Contexte et motivations

Les Nations unies ont reconnu l'importance de la participation et de la représentation des femmes dans les médias lors de la *Quatrième conférence mondiale sur les femmes* qui s'est tenue en septembre 1995 à Pékin (section J de la déclaration officielle) (UN, 1995). En conséquence, des initiatives telles que le *Global Media Monitoring Project* [1] (GMMP) ont été lancées pour mesurer l'état et l'évolution de la présence des femmes dans les sources d'information traditionnelles (journaux,

1. https://whomakesthenews.org/

Français	Durée (h)	Exemples	Femmes %
informations	14.8 / 1.5 / 1.8	5258 / 518 / 630	50 / 50 / 27
thématiques	27.0 / 2.9 / 3.5	9114 / 968 / 1168	50 / 50 / 27
téléréalité	12.1 / 0.8 / 1.0	6250 / 406 / 525	50 / 50 / 60
Japonais	**Durée (h)**	**Exemples**	**Femmes %**
train_5k	30,4	5k	50
train_100k	599	100k	50
Val / Test	9,1 / 8,4	1365 / 1184	50 / 50

TABLE 1 – Informations sur les ensembles de données. Pour la partie française, les valeurs des divisions sont rapportées dans l'ordre Train / Val / Test.

télévision, radio), ainsi que dans les médias numériques (sites web et tweets de la presse en ligne). Depuis 1995, les études successives du GMMP (une tous les cinq ans) ont révélé, entre autres, un déséquilibre entre les genres en ce qui concerne les sujets et les sources d'information (seulement 45 % de femmes en 2020), ainsi que pour les journalistes effectuant des reportages (40 % de femmes en 2020) (GMMP, 2021). De même, dans le contexte français, l'Arcom a noté dans son rapport annuel de veille 2022 (Arcom, 2023) que les femmes ne représentaient que 36 % du temps de parole global dans les émissions de télévision et de radio. Au-delà des analyses quantitatives, reste la question de la représentation [2] du genre, qui peut véhiculer et entretenir les stéréotypes et le sexisme. Ceux-ci existent, notamment, à travers la façon dont les hommes et les femmes parlent (p. ex., un style d'élocution caricatural), et à travers la préférence différenciée pour certains thèmes. Les exemples de ces phénomènes peuvent être relativement rares et subtils à détecter, ce qui demande la réalisation d'analyses fines à grande échelle.

Dans cet article, nous soutenons que l'utilisation de techniques de l'IA explicable peut aider à entreprendre une analyse qualitative fine sur des ensembles de données importants, en particulier dans le cas de la représentation des genres dans les médias audiovisuels.

2.2 Ensembles de données

Les données utilisées pour nos expériences sont des programmes de télévision et de radio diffusés en France et au Japon [3]. Cette dualité nous permet notamment de vérifier l'applicabilité de notre approche pour des langues éloignées, et nous donne une opportunité de chercher des points de comparaison entre des contextes culturels différents. Notons qu'en contrepartie de leur authenticité, ces données ne sont pas publiquement accessibles (elles ne peuvent être redistribuées que par les entreprises qui les ont produites et en détiennent les droits). Comme précédemment indiqué, nous ne traitons que les transcriptions préalablement engendrées par des systèmes de reconnaissance automatique de la parole (RAP), puisque nous nous limitons à l'étude de divergences lexicales dans le discours. La composition des ensembles de données français et japonais est résumée dans le tableau 1. Nous avons équilibré les classes de genre, sauf pour les ensembles de test en français, afin de ne pas affecter leur significativité statistique.

2. Nous entendons ici « représentation » au sens d'image renvoyée et non de présence ou de distribution.
3. Nous avons combiné plusieurs ensembles de données auxquels nous avions accès dans le cadre de contrats de projets de financement.

Données françaises Le corpus français est une combinaison de différents types d'émissions : (i) des programmes liés aux informations (des journaux télévisés locaux et nationaux, d'une durée de 20 à 30 minutes, ainsi que des matinales radio axées sur l'actualité, d'une durée de 2 à 4 heures), diffusés en 2021 par un ensemble de chaînes privées et publiques, (ii) des programmes radiophoniques thématiques (des magazines composés d'interviews, 50-60 min), diffusés en 2018 et centrés sur des thèmes tels que l'économie, le sport, la cuisine et les questions sociales, (iii) et des émissions de téléréalité (d'une durée de 45 minutes chacune) diffusées en 2021. La transcription automatique est réalisée avec une variante du système LIUM ASR décrit dans Tomashenko *et al.* (2016). L'outil InaSpeechSegmenter (Doukhan *et al.*, 2018) est utilisé pour effectuer la détection du genre et attribuer une étiquette (homme ou femme [4]) à chaque segment de parole reconnu par LIUM ASR. Ces étiquettes serviront de référence [5] lors de l'entraînement de nos classificateurs (Section 3). Certains programmes contiennent des publicités que nous avons manuellement retirées des ensembles de développement et de test.

Données japonaises Le corpus japonais comprend des transcriptions automatiques du programme de nouvelles télévisées *NHK News 7* du diffuseur public japonais NHK, retransmis entre 2001 et 2022. Ces programmes de 30 minutes ont été transcrits à l'aide de l'outil de RAP Whisper (Radford *et al.*, 2023). L'ensemble d'entraînement comprend des programmes d'information quotidiens de 2005 à 2022, tandis que les sections de test et de développement correspondent à des programmes diffusés en 2001, 2002 et 2004, respectivement. Les étiquettes de genre sont automatiquement associées à chaque énoncé avec l'outil InaSpeechSegmenter [6] (comme pour la partie française) et corrigées manuellement pour les ensembles de test et de développement, ainsi que pour 5000 exemples d'entraînement (train_5k).

3 Détection de genre fondée sur BERT

Pour effectuer la classification du genre du locuteur, nous avons opté pour une architecture fondée sur BERT (Devlin *et al.*, 2019), comme réalisé pour divers types de classification de textes (Sun *et al.*, 2019 ; González-Carvajal & Garrido-Merchán, 2020). BERT s'appuie sur le préentraînement d'un large modèle conçu pour être facilement affiné par la suite pour les tâches en aval, avec un minimum de modifications architecturales. Dans le cas de la classification de textes, cette modification consiste en une couche linéaire supplémentaire pour traiter la représentation agrégée de la séquence d'entrée (l'encodage de l'unité `[CLS]`).

3.1 Modèles

Modèles français Les modèles les plus couramment utilisés pour le français sont CamemBERT (Martin *et al.*, 2020) et FlauBERT (Le *et al.*, 2020). Comme nos données sont constituées de transcriptions automatiques de la parole, nous utilisons dans nos expériences des modèles FlauBERT-

4. L'attribution des étiquettes est basée sur le chevauchement des segments ; les énoncés qui chevauchent principalement les étiquettes de musique ou de bruit sont éliminés.

5. Doukhan *et al.* (2018) indiquent une F-mesure de détection du genre au niveau de la trame de 96,52 sur le corpus REPERE, qui contient des flux télévisés de chaînes françaises, similaires à nos données.

6. Une évaluation manuelle a posteriori sur 10 heures montre une exactitude de 94,98 % sur ces étiquettes automatiques.

Oral (Hervé *et al.*, 2022), basés sur FlauBERT, qui sont partiellement ou entièrement préentraînés sur des sorties de RAP (19 Go générés à partir d'émissions d'actualités françaises diffusées entre 2013 et 2020). Plus précisément : `FlauBERT-O-mixed` est un modèle préentraîné sur un mélange de données écrites (13 Go provenant de Wikipédia et d'articles de presse) et de transcriptions, et `FlauBERT-O-asr_nb` est un modèle préentraîné sur des données de transcriptions uniquement.

Modèles japonais Pour le japonais, nous utilisons les modèles bert-japanese [7], qui ont été préentraînés sur des articles Wikipédia (2,6 Go). Plus précisément, nous utilisons deux versions qui diffèrent par la méthode de segmentation : soit WordPiece (`bert-base-jp`), soit au niveau des caractères (`bert-base-jp-char`).

Enfin, nous utilisons le BERT multilingue original (Devlin *et al.*, 2019) (`mBERT`), qui a été préentraîné sur 100 langues correspondant aux versions de Wikipédia les plus importantes (47 Go au total), à la fois pour le français et le japonais [8].

3.2 Implémentation

Nous avons utilisé la bibliothèque Transformers de HuggingFace (Wolf *et al.*, 2020) pour mettre en uvre les modèles BERT. Nous avons suivi l'usage d'hyperparamètres largement acceptés : Adam en tant qu'optimiseur (β_1 = 0,9, β_2 = 0,999, ϵ = 1e-08), taux d'apprentissage = 2e-5, taille de batch = 32. L'entraînement est poursuivi jusqu'à ce qu'aucune amélioration de la fonction perte de validation ne soit constatée pendant 3 époques consécutives (le meilleur modèle est conservé). Les expériences ont été réalisées sur un seul GPU Tesla V100-SXM2. L'affinage d'un seul modèle prenait 3 à 6 minutes (environ 80 min en utilisant train_100k), tandis que l'inférence ne nécessitait approximativement qu'une minute. Les tailles des modèles sont les suivantes : `mBERT`, `CamemBERT`, `bert-base-jp`—110M de paramètres, `FlauBERT-O-mixed`, `FlauBERT-O-asr_nb`—138M, `bert-base-jp-char`—90M.

3.3 Résultats

Les résultats de la classification du genre du locuteur sont présentés dans le tableau 2. Comme certains ensembles de test ne sont pas équilibrés entre les classes masculine et féminine, nous utilisons P4 (Sitarz, 2023), la moyenne harmonique des scores F1 des deux classes (c.-à-d. la moyenne harmonique de la précision et du rappel mesurés pour chaque classe), comme principale mesure de classification binaire. Des intervalles de confiance ont été calculés selon l'approche de rééchantillonage *bootstrap* [9] (bootstraps = 5000, = 5, condition = instance d'émission).

En ce qui concerne la partie française, de façon générale, les modèles `FlauBERT-O` sont plus performants que `mBERT` et `CamemBERT`, ce qui montre l'avantage, dans notre cas, d'un préentraînement sur les sorties de RAP. Pour les programmes d'information, `FlauBERT-O-asr_nb` est le meilleur modèle. Il s'agit probablement d'une conséquence logique du fait que ce modèle a

7. https://github.com/cl-tohoku/bert-japanese/
8. Notons que la segmentation de `mBERT` est fondée sur WordPiece, comme pour `bert-base-jp`.
9. Ferrer et Riera,"Confidence Intervals for evaluation in machine learning." [Logiciel]. https://github.com/luferrer/ConfidenceIntervals

Modèle	P4	F1♂	F1♀	Ex.
Français (*informations*)				
mBERT	46,9 (31,1-50,9)	61,3	38,0	52,4
CamemBERT	49,9 (27,6-53,6)	56,5	44,7	51,3
FBO-mixed	51,0 (30,9-55,2)	58,4	45,2	52,7
FBO-asr_nb	55,0 (36,0-58,6)	58,9	51,6	55,6
Français (*thématiques*)				
mBERT	51,7 (30,7-63,7)	81,0	38,0	70,9
CamemBERT	55,3 (39,0-62,2)	76,2	43,4	66,4
FBO-mixed	58,8 (37,0-70,1)	82,8	45,6	73,9
FBO-asr_nb	57,9 (39,7-65,4)	78,0	46,1	68,8
Français (*téléréalité*)				
mBERT	60,5 (52,4-65,3)	55,6	66,3	61,7
CamemBERT	59,9 (52,5-65,4)	54,0	67,2	61,7
FBO-mixed	59,9 (52,9-65,3)	53,7	67,6	61,9
FBO-asr_nb	57,2 (49,9-62,3)	52,5	62,8	58,3

Modèle	P4	F1♂	F1♀	Ex.
Japonais (`train_5k`)				
mBERT	59,9 (56,9-62,7)	64,5	55,9	60,6
BB-jp	60,3 (57,5-63,0)	58,4	62,3	60,5
BB-jp-char	57,5 (54,6-60,2)	57,3	57,8	57,5
Japonais (`bert-base-jp`)				
train_5k	60,3 (57,5-63,0)	58,4	62,3	60,5
train_100k	61,8 (58,6-64,6)	70,5	54,9	64,4

TABLE 2 – Scores P4, F1 (pour les classes masculine-♂ et féminine-♀), et exactitude pour les différents modèles évalués sur l'ensemble de test. FBO : FlauBERT-O, BB : bert-base.

été entièrement préentraîné sur le même type de données. Pour les émissions thématiques, les modèles FlauBERT-O obtiennent les meilleurs résultats, alors que pour les émissions de téléréalité, FlauBERT-O-asr_nb obtient les plus mauvais scores (les autres modèles étant comparables). Encore une fois, nous pouvons supposer l'influence de la correspondance des domaines entre le préentraînement et le test (les transcriptions des émissions de télé-réalité sont assez bruitées, les gens parlant très spontanément dans ce type de programme, tandis que les informations télévisées et radiophoniques préparent une partie du discours à l'avance).

Concernant la partie japonaise, nous avons d'abord évalué des versions des modèles BERT affinées sur un ensemble de 5000 exemples vérifiés manuellement (train_5k). Les meilleurs résultats sont obtenus avec mBERT et bert-base-jp (qui utilisent tous deux la segmentation WordPiece). Nous avons ensuite comparé les performances de ce modèle en utilisant un plus grand jeu d'affinage (annoté automatiquement) : l'augmentation substantielle de la quantité de données n'entraîne qu'un léger gain.

Il apparaît que la plupart des classificateurs basés sur BERT peuvent, dans une mesure limitée, détecter le genre du locuteur sur la base d'une transcription d'énoncé (P4 $>$ 50). Ces résultats sont toutefois dans l'absolu plutôt modestes. Cela doit être mis en relation avec la difficulté intrinsèque de la tâche : les énoncés à classer ne sont composés que de 30-40 mots en moyenne dans la partie française, et seulement d'une vingtaine de mots dans la partie japonaise. Une part significative des exemples ne contient probablement pas d'indice clair. À titre de comparaison, les systèmes soumis à la tâche de profilage de genre PAN 2019 devaient prédire le genre de l'auteur à partir d'un ensemble de 100 tweets (pour l'anglais, la meilleure équipe a obtenu une exactitude de 84,17 %). Dans la section suivante, nous expliquons comment nous utilisons la confiance exprimée par le classificateur et les techniques d'explicabilité pour identifier des mots porteurs d'information sur le genre.

4 Analyse qualitative par les techniques d'explicabilité

4.1 Méthodes

Nous avons utilisé trois méthodes bien connues, représentatives de différents groupes de techniques de l'IAX, afin de calculer des valeurs de contribution aux classes masculine et féminine au niveau des mots dans les énoncés transcrits (comme illustré dans la Figure 1).

Occlusion Initialement proposée par Zeiler & Fergus (2014) dans le contexte du traitement des images, Occlusion est une approche *fondée sur les perturbations* qui, dans son principe, est l'une des techniques d'explicabilité les plus simples. Elle mesure l'effet sur la sortie du système de la suppression, ou du remplacement par une valeur neutre, d'une partie de l'entrée. Dans notre cas, il s'agit d'utiliser l'unité de masquage utilisée dans le préentraînement de BERT. Pour ce qui est de la quantification du changement causé par la perturbation, nous envisageons deux options : (a) tenir compte du changement (potentiel) de la classe prédite (ou *changement d'étiquette*), (b) tenir compte de la variation de la distribution de probabilité binaire (ou *changement de probabilité*). La première peut être considérée comme moins flexible, car toutes les perturbations n'impliquent pas un changement de la classe prédite. Pour s'assurer que l'effet de perturbation n'est pas contourné en raison de la répétition dans la séquence, nous masquons toutes les occurrences d'une unité en une seule fois. En outre, comme nous pensons que le changement d'étiquette ou de probabilité devrait être moins important dans le cas d'une séquence courte, nous pondérons chaque perturbation par le nombre d'unités uniques dans la séquence.

LIME Proposée par Ribeiro *et al.* (2016), LIME (*Local Interpretable Model-Agnostic Explanations*) est une technique *fondée sur la simplification* qui entraîne un modèle linéaire [10] de substitution à approximer, autour d'un exemple donné, la limite de décision locale du modèle complexe d'origine. Dans la classification des textes, LIME masque aléatoirement les unités de la séquence exemple. Elle ajuste alors un modèle linéaire pour faire correspondre la sortie (c.-à-d. la probabilité d'une certaine classe) du modèle complexe pour ces variantes masquées. Ce modèle linéaire prenant en entrée une représentation simplifiée, sous la forme d'un vecteur binaire indiquant la présence ou l'absence de chaque unité dans l'échantillon perturbé. En conséquence, les coefficients appris fournissent des valeurs au niveau des unités pour la contribution à la classe cible.

LIG Proposée par Mudrakarta *et al.* (2018), LIG (*Layer Integrated Gradients*) est une approche *fondée sur les gradients*, directement inspirée des gradients intégrés de Sundararajan *et al.* (2017). Intuitivement, on peut considérer que si le produit d'un modèle change considérablement en fonction de la variation d'une dimension d'entrée (c.-à-d. que le gradient de la sortie par rapport à la dimension d'entrée est élevé en valeur absolue), cela signifie que la valeur d'entrée de cette dimension particulière est importante pour la décision du modèle. Cependant, Sundararajan *et al.* (2017) remarquent qu'au lieu de la variation locale autour de la valeur d'entrée, il faudrait considérer la variation agrégée entre la valeur d'entrée et une valeur de base non informative représentant l'absence de la caractéristique (ce qui renvoie à l'idée de mesurer l'effet d'une perturbation). Mudrakarta *et al.*

10. Ce cas est le plus courant pour la mise en uvre de LIME, mais la description générale donnée par Ribeiro *et al.*, 2016 permet d'utiliser d'autres types de modèles interprétables (p. ex., des arbres de décision).

(2018) ont appliqué ce principe au traitement des textes, en définissant le neutre comme une séquence d'unités de remplissage, et en intégrant le gradient de la sortie du modèle sur les dimensions de la couche de plongement (qui correspond à un espace continu, par opposition aux unités de la séquence).

4.2 Implémentation

La mise en uvre des méthodes LIME et LIG a été réalisée à l'aide de la bibliothèque Captum (Kokhlikyan *et al.*, 2020). Pour LIME, nous calculons la mesure de proximité (notée π_x dans Ribeiro *et al.*, 2016) par le biais de la distance cosinus entre les encodages de `CLS` au sein des échantillons d'origine et perturbé, en utilisant le modèle *lasso linéaire* de scikit-learn (α = 0.001), et en échantillonnant 200 perturbations par exemple [11]. Les expériences ont été réalisées sur un seul GPU Tesla V100-SXM2, et sur un processeur Intel Cascade Lake 6248 (10 curs à 2,5 GHz). Les temps d'exécution des méthodes d'IAX, lors de l'analyse des prédictions de `FlauBERT-O-mixed` pour les émissions de téléréalité françaises de l'ensemble de test (525 exemples), sont les suivants : Occlusion—1 min, LIME—1 h, LIG—30 min.

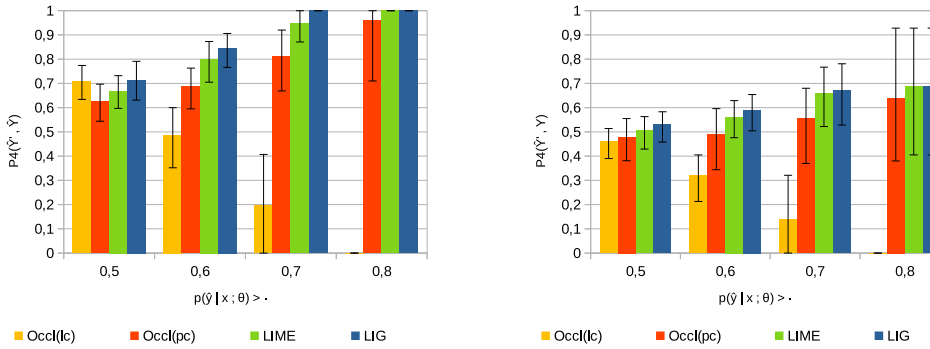
4.3 Identification des expressions genrées

Nous avons appliqué les trois méthodes d'explicabilité afin de produire des explications pour les prédictions de nos modèles sur les ensembles de tests. Nous avons choisi de calculer les scores d'attribution au niveau des unités en fonction de leur contribution à la classe cible masculine (les scores positifs indiquent une orientation masculine et les scores négatifs une orientation féminine). Afin de vérifier la **cohérence** des attributions avec, d'une part, les prédictions effectives du modèle, et d'autre part, les étiquettes de références, nous définissons une procédure de calcul de « pseudo-prédictions ». Plus précisément, soient θ un modèle de classification, ψ une méthode d'explication, x une séquence, et $(\psi(x, \theta, ♂))_t$ la valeur de contribution à la classe homme (♂) attribuée par ψ à la t-ième unité de x. Alors la pseudo-prédiction dérivée de l'explication $\psi(x, \theta, ♂)$ est : $\hat{y}' =$ ♂ si $\sum_t (\psi(x, \theta, ♂))_t > 0$ sinon ♀

Du calcul des pseudo-prédictions $\hat{Y}'$ nous inférons $\mathrm{P4}(\hat{Y}', \hat{Y})$, une évaluation vis-à-vis des prédictions effectives du modèle, ainsi que $\mathrm{P4}(\hat{Y}', Y)$, une évaluation vis-à-vis des références. La figure 2a compare les scores $\mathrm{P4}(\hat{Y}', \hat{Y})$ obtenus en fonction des techniques d'explicabilité employées, sur différents sous-ensembles d'exemples de test filtrés selon des seuils croissants de confiance du classificateur ($p(\hat{y}|x; \theta) > \cdot$). Nous pouvons notamment voir que les pseudo-prédictions sont d'autant plus proches des vraies prédictions quand la confiance du modèle est élevée (p. ex. pour LIG, à partir de $p(\hat{y}|x; \theta) > 0{,}7$, correspondant à 117 instances, $\hat{Y}' = \hat{Y}$). Nous observons également que LIG est la technique affichant la plus grande cohérence entre ses attributions et les prédictions. Notons que pour Occlusion(changement d'étiquette) $\mathrm{P4}(\hat{Y}', \hat{Y})$ diminue quand le niveau de filtrage augmente : de façon logique, puisque que la croissance de la confiance s'oppose naturellement au changement d'étiquette prédite pour l'échantillon perturbé. La figure 2b compare les scores $\mathrm{P4}(\hat{Y}', Y)$: les mêmes tendances se retrouvent dans ce cas, à ceci près que les scores sont globalement plus faibles (P4 ne dépasse pas $0{,}7$, même avec le seuil le plus élevé).

Afin de fournir une vue d'ensemble, nous avons calculé la moyenne des scores associés à chaque

11. Notre code est disponible à cette adresse : https://github.com/Cyossnarf/XSpeakerGender



(a) (b)

FIGURE 2 – Évaluation de la cohérence des explications. Le modèle analysé est `FlauBERT-O-mixed`, appliqué sur les émissions françaises de téléréalité (les exemples sont filtrés selon la confiance du classificateur). Les intervalles de confiance ont été calculés selon l'approche *bootstrap* (bootstraps = 5000, = 5, condition = instance de programme).

occurrence d'unité pour produire un lexique genré agrégé. La figure 3 présente des exemples de lexiques correspondant aux scores calculés les plus élevés (♂) et les plus bas (♀) parmi les vocabulaires [12]. Pour les données japonaises, certains thèmes sont davantage associés aux hommes (politique) et aux femmes (météo), reflétant une tendance existant effectivement dans les programmes. Pour la partie française, nous avons pu noter un usage plus marqué des pronoms personnels par les femmes dans les émissions de téléréalité. Cela coïncide avec les observations d'études antérieures (Pennebaker, 2011 ; Kocher & Savoy, 2016).

	Français (*téléréalité*), `FlauBERT-O-mixed`, LIME
♂	`soirée`, `famille`, `temps`, `soir`, `aime`, `amour`, `vie`, `faut`, `demain`, `chez`
♀	`lui`, `clairement`, `cas`, `euh`, `cela`, `tu`, `son`, `avoir`, `moins`, `avait`
	Japonais (*informations*, `train_100k`), LIG
♂	市場 (marché), 国会 (régime), まし (meilleur), 選挙 (élection), 政府 (gouvernement), 議員 (membre du parlement), 党 (parti), 側 (côté), ます (masu), です (est), ね (hé)
♀	雪 (neige), 朝 (matin), 北海道 (Hokkaido), 晴れ (ensoleillé), 東北 (Tohoku), 雨 (pluie), 夜 (nuit), そう (oui), にかけて (dessus), 日 (jour)

FIGURE 3 – Lexique pour les classes masculine-♂ et féminine-♀, extrait par le biais des techniques d'explicabilité (automatiquement en français dans le cas du japonais).

12. Nous n'avons conservé que les unités qui apparaissaient 9 fois ou plus, afin de calculer des valeurs moyennes fiables.

5 Travaux connexes

Ces dernières années, l'IAX a été appliquée à une variété de sous-domaines afin de fournir une assistance dans la prise de décision par un agent : par exemple pour la détection de fausses nouvelles (Yang *et al.*, 2019), l'intervention d'un instructeur dans un MOOC (Alrajhi *et al.*, 2022), ou encore la détection de discours haineux (Kim *et al.*, 2022). La détection du genre fait partie du domaine plus vaste du profilage des auteurs, qui vise à déduire des traits sociaux ou de personnalité sur la base des messages produits par une personne (Stajner & Yenikent, 2020). Sánchez *et al.* (2022) mentionnent trois contextes en particulier dans lesquels elle peut être utilisée : la linguistique judiciaire—afin d'identifier les auteurs de violence et de harcèlement sur internet, le marketing—afin de mener des stratégies publicitaires personnalisées, et la sociolinguistique—afin de mettre en relation des motifs linguistiques avec des variables sociales comme le genre. Les algorithmes utilisés pour le profilage de genre à partir de textes vont des approches classiques, telles que les SVM et la classification naïve bayésienne (Burger *et al.*, 2011), aux approches modernes basées sur les neurones (Bartle & Zheng, 2015). Depuis 2013, l'association PAN [13], dans le cadre du *Conference and Labs of the Evaluation Forum* (CLEF), a organisé une série de tâches partagées de profilage des auteurs, parmi lesquelles la détection du genre a été abordée à plusieurs reprises. Enfin, plusieurs études ont lié la détection du genre avec les analyses stylistiques (Savoy, 2022). Par exemple, en analysant des tweets, Ikae & Savoy (2022) constatent que certains termes ou catégories de mots—tels que les articles, les pronoms personnels, les négations, les émotions, les nombres, la ponctuation, les émojis—peuvent être davantage liés à un genre qu'à l'autre.

6 Conclusion

Cette étude présente notre méthodologie de détection du genre à partir de transcriptions automatiques de la parole. Nous avons expliqué comment utiliser les techniques développées pour l'explicabilité des modèles de façon à mettre en évidence des informations lexicales genrées, en prenant comme cas d'application des programmes télévisés et radiophoniques en français et en japonais. Nos expériences montrent que les classificateurs basés sur BERT peuvent, dans une mesure limitée, prédire le genre du locuteur sur la base d'un énoncé. Par conséquent, l'IAX devrait être associée à l'exploitation de la confiance de la prédiction du classificateur afin de potentiellement localiser des indicateurs spécifiques au genre. Les orientations de recherche future pourraient inclure : l'utilisation des méthodes d'IAX afin de produire des attributions plus complexes, et la récupération des annotations sur le genre de l'interlocuteur pour analyser les différences selon ce critère.

Remerciements

Ce travail a été financé par l'ANR (projet *Gender Equality Monitor* – ANR-19-CE38-0012). Il a en outre bénéficié dun accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2023-AD011014209 attribuée par GENCI.

13. https://pan.webis.de/

Références

ALRAJHI L., PEREIRA F. D., CRISTEA A. I. & ALJOHANI T. (2022). A good classifier is not enough : A XAI approach for urgent instructor-intervention models in moocs. In M. M. T. RODRIGO, N. MATSUDA, A. I. CRISTEA & V. DIMITROVA, Éds., *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium - 23rd International Conference, AIED 2022, Durham, UK, July 27-31, 2022, Proceedings, Part II*, volume 13356 de *Lecture Notes in Computer Science*, p. 424–427 : Springer. DOI : 10.1007/978-3-031-11647-6_84.

ARCOM (2023). Women representation on television and radio. Available online : https://www.arcom.fr/sites/default/files/2023-06/Representation_des_femmes_a_la_television_et_a_%20la_radio-Rapport_sur_exercice_2022-Arcom.pdf. In French. Last accessed : 16/10/2023.

ASH E., DURANTE R., GREBENSHCHIKOVA M. & SCHWARZ C. (2022). *Visual Representation and Stereotypes in News Media*. CESifo Working Paper Series 9686, CESifo.

BARTLE A. & ZHENG J. (2015). Gender classification with deep learning. *Stanfordcs, 224d Course Project Report*, p. 1–7.

BURGER J. D., HENDERSON J., KIM G. & ZARRELLA G. (2011). Discriminating gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, p. 1301–1309, Edinburgh, Scotland, UK. : Association for Computational Linguistics.

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.

DOUKHAN D., CARRIVE J., VALLET F., LARCHER A. & MEIGNIER S. (2018). AN OPEN-SOURCE SPEAKER GENDER DETECTION FRAMEWORK FOR MONITORING GENDER EQUALITY. In *IEEE International Conference on Acoustic Speech and Signal Processing*, Calgary, Canada. HAL : hal-01927560.

GMMP (2021). 6th global media monitoring project highlight of findings. Available online : https://whomakesthenews.org/wp-content/uploads/2021/08/GMMP-2020.Highlights_FINAL.pdf. Last accessed : 16/10/2023.

GONZÁLEZ-CARVAJAL S. & GARRIDO-MERCHÁN E. C. (2020). Comparing BERT against traditional machine learning text classification. *CoRR*, **abs/2005.13012**.

HERVÉ N., PELLOIN V., FAVRE B., DARY F., LAURENT A., MEIGNIER S. & BESACIER L. (2022). Using ASR-Generated Text for Spoken Language Modeling. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, p. 17–25, virtual+Dublin, France : Association for Computational Linguistics. DOI : 10.18653/v1/2022.bigscience-1.2, HAL : hal-03770460.

IKAE C. & SAVOY J. (2022). Gender identification on twitter. *J. Assoc. Inf. Sci. Technol.*, **73**(1), 58–69. DOI : 10.1002/asi.24541.

KIM J., LEE B. & SOHN K. (2022). Why is it hate speech ? masked rationale prediction for explainable hate speech detection. In N. CALZOLARI, C. HUANG, H. KIM, J. PUSTEJOVSKY, L. WANNER, K. CHOI, P. RYU, H. CHEN, L. DONATELLI, H. JI, S. KUROHASHI, P. PAGGIO, N. XUE, S. KIM, Y. HAHM, Z. HE, T. K. LEE, E. SANTUS, F. BOND & S. NA, Éds., *Proceedings*

of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022, p. 6644–6655 : International Committee on Computational Linguistics.

KOCHER M. & SAVOY J. (2016). Unine at CLEF 2016 : Author profiling. In K. BALOG, L. CAPPELLATO, N. FERRO & C. MACDONALD, Éds., *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016*, volume 1609 de *CEUR Workshop Proceedings*, p. 903–911 : CEUR-WS.org.

KOKHLIKYAN N., MIGLANI V., MARTIN M., WANG E., ALSALLAKH B., REYNOLDS J., MELNIKOV A., KLIUSHKINA N., ARAYA C., YAN S. & REBLITZ-RICHARDSON O. (2020). Captum : A unified and generic model interpretability library for pytorch. *CoRR*, **abs/2009.07896**.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBE B., BESACIER L. & SCHWAB D. (2020). FlauBERT : Unsupervised Language Model Pre-training for French. In *LREC*, Marseille, France. HAL : hal-02890258.

MARCINKEVICS R. & VOGT J. E. (2020). Interpretability and explainability : A machine learning zoo mini-tour. *CoRR*, **abs/2012.01805**.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., VILLEMONTE DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a Tasty French Language Model. In *ACL 2020 - 58th Annual Meeting of the Association for Computational Linguistics*, Seattle / Virtual, United States. DOI : 10.18653/v1/2020.acl-main.645, HAL : hal-02889805.

MEDIAWATCH (1995). Global media monitoring project : Women's participation in the news.

MUDRAKARTA P. K., TALY A., SUNDARARAJAN M. & DHAMDHERE K. (2018). Did the model understand the question ? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1896–1906, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1176.

PENNEBAKER J. W. (2011). Your use of pronouns reveals your personality. *Harvard business review*, **89**(12), 32–33.

RADFORD A., KIM J. W., XU T., BROCKMAN G., MCLEAVEY C. & SUTSKEVER I. (2023). Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*.

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

SÁNCHEZ D. M., MORENO A. & JIMÉNEZ-LÓPEZ M. D. (2022). Machine learning methods for automatic gender detection. *Int. J. Artif. Intell. Tools*, **31**(3), 2241002 :1–2241002 :8. DOI : 10.1142/S0218213022410020.

SAVOY J. (2022). Stylometric analysis of characters in Shakespeares plays. *Digital Scholarship in the Humanities*, **38**(3), 1238–1246. DOI : 10.1093/llc/fqac092.

SITARZ M. (2023). Extending F1 metric, probabilistic approach. *Adv. Artif. Intell. Mach. Learn.*, **3**(2), 1025–1038. DOI : 10.54364/aaiml.2023.1161.

STAJNER S. & YENIKENT S. (2020). A survey of automatic personality detection from texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 6284–6295, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.553.

SUN C., QIU X., XU Y. & HUANG X. (2019). How to fine-tune BERT for text classification ? In M. SUN, X. HUANG, H. JI, Z. LIU & Y. LIU, Éds., *Chinese Computational Linguistics - 18th China National Conference, CCL 2019, Kunming, China, October 18-20, 2019, Proceedings*, volume 11856 de *Lecture Notes in Computer Science*, p. 194–206 : Springer. DOI : 10.1007/978-3-030-32381-3_16.

SUNDARARAJAN M., TALY A. & YAN Q. (2017). Axiomatic attribution for deep networks. In D. PRECUP & Y. W. TEH, Éds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 de *Proceedings of Machine Learning Research*, p. 3319–3328 : PMLR.

TOMASHENKO N., VYTHELINGUM K., ROUSSEAU A. & ESTÈVE Y. (2016). LIUM ASR systems for the 2016 Multi-Genre Broadcast Arabic Challenge. In *IEEE Workshop on Spoken Language Technology*, San Diego, CA, USA, United States. DOI : 10.1109/SLT.2016.7846278, HAL : hal-01433188.

UN U. N. (1995). Beijing declaration and platform for action. Available online : https://www.un.org/womenwatch/daw/beijing/pdf/BDPfA%20E.pdf. Last accessed : 16/10/2023.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M., DAVISON J., SHLEIFER S., VON PLATEN P., MA C., JERNITE Y., PLU J., XU C., SCAO T. L., GUGGER S., DRAME M., LHOEST Q. & RUSH A. M. (2020). Transformers : State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 38–45, Online : Association for Computational Linguistics.

YANG F., PENTYALA S. K., MOHSENI S., DU M., YUAN H., LINDER R., RAGAN E. D., JI S. & HU X. B. (2019). Xfake : Explainable fake news detector with visualizations. In *The World Wide Web Conference*, WWW '19, p. 3600‑3604, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3308558.3314119.

ZEILER M. D. & FERGUS R. (2014). Visualizing and understanding convolutional networks. In D. FLEET, T. PAJDLA, B. SCHIELE & T. TUYTELAARS, Éds., *Computer Vision – ECCV 2014*, p. 818–833, Cham : Springer International Publishing.