Un modèle neurosymbolique fondé sur le genre, la thématique, et les techniques de persuasion pour améliorer la robustesse dans la détection de la propagande

RÉSUMÉ

La détection de la propagande constitue un enjeu démocratique de taille, compte tenu en particulier de la prolifération de sites d'ingérence étrangère diffusant des narratifs sous couvert de nouvelles ordinaires. Pour détecter la propagande, les approches actuelles fondées sur des modèles de langue tels que BERT sont prometteuses, mais tendent souvent à surapprendre leurs jeux de données d'entraînement, en raison de biais introduits lors de la collecte des données. Afin d'accroître la robustesse de la classification, nous proposons une approche neurosymbolique combinant des représentations vectorielles statiques du texte (embeddings fastText) et des caractéristiques conceptuelles symboliques, telles que le genre, le thème et les techniques de persuasion, de manière à améliorer la capacité de généralisation à de nouvelles sources. Les résultats mettent en évidence des gains par rapport à des méthodes équivalentes n'exploitant que le texte, tandis que les analyses d'explicabilité confirment l'apport des caractéristiques ajoutées.

ABSTRACT

A Neurosymbolic Model Using Genre, Topic, and Persuasion Techniques to Improve Robustness in Classification

Propaganda detection constitutes a major democratic challenge, particularly with the proliferation of pseudo-news sites that rewrite news articles to broadcast narratives. To detect propaganda, extant approaches based on Language Models such as BERT are promising but often overfit their training datasets, due to biases in data collection. To enhance classification robustness, we propose a neurosymbolic approach combining static fastText embeddings and symbolic conceptual features such as genre, topic, and persuasion techniques, to improve generalization to new sources. Results show improvements over equivalent text-only methods, and explainability analyses confirm the benefits of the added features.

MOTS-CLÉS : Désordre informationnel, Fausses nouvelles, Propagande, Classification, Modélisation thématique, Méthode hybride, Modèle neurosymbolique, Ablation, Robustesse, Explicabilité.

KEYWORDS: Information disorder, Fake news, Propaganda, Classification, Topic modeling, Hybrid method, Neurosymbolic model, Ablation, Robustness, Explainability.

1 Introduction

Ces dernières années ont été marquées par une forte augmentation des manipulations de l'information en ligne, alimentée par le regain de tensions internationales, comme l'ont documenté en Europe divers services de renseignement (VIGINUM[1] en France, ZEAM[2] en Allemagne) ainsi que des organisations de veille (notamment EU DisinfoLab[3]). Cette tendance, que nous pouvons qualifier de « désordre informationnel » (en adaptant la terminologie de Wardle & Derakhshan 2017), est souvent orchestrée via des sites web d'apparence journalistique, qui imitent les conventions de la presse et diffusent des récits ciblés afin d'influencer l'opinion (i.e. des « pseudo-actualités », voir Anonymisé 2024a). Ce phénomène est préoccupant, car ces contenus sont largement partagés sur les réseaux sociaux, pouvant atteindre des audiences importantes.[4]

La détection automatique des désordres informationnels a progressé selon diverses orientations, notamment l'étude des biais médiatiques (Hamborg *et al.*, 2019), la désinformation via la vérification automatique des faits (Thorne *et al.*, 2018), les fausses nouvelles (Zhou & Zafarani, 2020; Hu *et al.*, 2025) et les rumeurs (Shu *et al.*, 2020). Toutefois, ces méthodes se transfèrent souvent mal à d'autres types de désordre informationnel. En particulier, elles tendent à échouer face aux opérations d'influence et à la propagande, qui reposent sur un cadrage contextuel des genres et des thématiques, ainsi que sur un ensemble de techniques de persuasion spécifiques (Da San Martino *et al.*, 2019, 2020). Inversement, certaines approches neurosymboliques, qui utilisent des vecteurs spécifiques basés sur certains traits conceptuels, se sont avérées efficaces pour d'autres formes de désordre informationnel (voir Ruchansky *et al.*, 2017; Ma *et al.*, 2017; Baly *et al.*, 2018; Thorne *et al.*, 2018). Ce faisant, elles indiquent une piste également prometteuse pour renforcer l'explicabilité des méthodes.

Dans cet article, nous nous intéressons à l'analyse de la propagande d'ingérence étrangère, en particulier celle repérée en 2022 par les agences de renseignement Viginum et Newsguard. Pour motiver notre approche, nous comparons deux jeux de données collectés antérieurement sur la propagande (Anonymisé, 2024a) : le corpus `NOM-ANONYMISÉ`, qui rassemble des exemples de propagande russe sur le conflit ukrainien, et le corpus `MAINSTREAM`, un corpus d'articles de presse démocratique et reconnue fiable sur le même sujet.

Les analyses antérieures ont montré qu'un grand modèle de langage tel que `RoBERTa-base` était capable de distinguer les articles de `NOM-ANONYMISÉ` de ceux de `MAINSTREAM` avec une précision très élevée (99,7%). Une telle précision suggère un possible surapprentissage. Pour améliorer la fiabilité et la robustesse, nous choisissons d'utiliser des encodages plus simples et parcimonieux (fastText), mieux adaptés à de gros corpus, et d'enrichir ces représentations par des informations conceptuelles (genre, thématique, techniques de persuasion). Nous montrons ainsi qu'une approche neurosymbolique combinant plongements fastText et plongements conceptuels surpasse une méthode fondée uniquement sur des plongements textuels.

La Section 2 présente les deux jeux de données `NOM-ANONYMISÉ` et `MAINSTREAM`, et la section 3 identifie certaines différences saillantes entre les deux corpus du point de vue stylistique en particulier.

1. https://www.sgdsn.gouv.fr/notre-organisation/composantes/service-de-vigilance-et-protection-contre-les-ingerences-numeriques?
2. https://www.bmi.bund.de/SharedDocs/schwerpunkte/EN/disinformation-election/zeam-artikel-en.html?
3. https://www.disinfo.eu
4. En septembre 2025, le Pew Research Center estime que 53% des adultes aux États-Unis utilisaient les réseaux sociaux comme source d'actualités : https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/.

Sur la base de ces observations, la section 4 introduit un modèle neurosymbolique qui fusionne des représentations vectorielles denses du texte avec des caractéristiques conceptuelles interprétables, extraites automatiquement des textes, à savoir le genre, les topiques, et les techniques de persuasion. La section 5 compare les performances de notre méthode hybride à celles d'un système de référence n'exploitant que le texte ; elle évalue sa robustesse selon différents schémas de partitionnement des ensembles d'entraînement/validation/test, et présente une analyse d'explicabilité. Enfin, la section 6 discute les résultats et esquisse des pistes pour des travaux futurs.

2 Jeux de données utilisés

2.1 NOM-ANONYMISÉ et MAINSTREAM

Nous exploitons deux jeux de données récemment présentés dans Anonymisé, 2024a, qui forment respectivement un corpus de *pseudo-actualités* propagandistes (NOM-ANONYMISÉ), et un corpus d'articles fiables issus de la presse généraliste (MAINSTREAM).

— NOM-ANONYMISÉ[5] (Anonymisé, 2024a), pour *Propagandist Pseudo-News*, est une collection de 12 427 articles provenant de sources identifiées comme des organes de propagande par les organisations expertes NewsGuard et VIGINUM.[6] Ces organes ont été créés après l'invasion de l'Ukraine par la Russie en février 2022 et diffusent des contenus propagandistes en neuf langues (arabe, chinois, anglais, français, allemand, italien, russe, espagnol et ukrainien).
— MAINSTREAM est un corpus d'articles en français et en anglais, composé d'actualités provenant de journaux quotidiens reconnus, et utilisé comme contrôle pour l'analyse du corpus NOM-ANONYMISÉ. Les articles MAINSTREAM ont été sélectionnés en fonction des dates de publication et de mots-clés liés au conflit en Ukraine. MAINSTREAM comprend 1 004 articles en anglais et 1 367 articles en français, issus respectivement de 11 et de 5 sources.

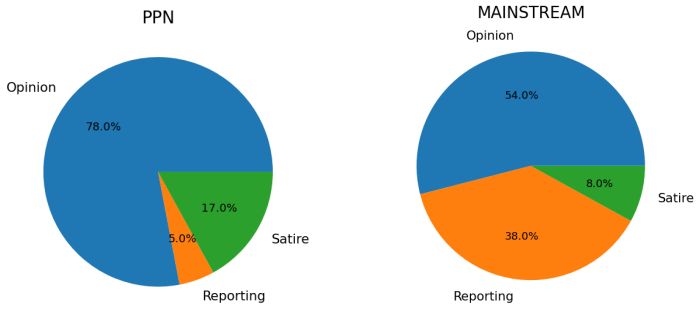

FIGURE 1 – Distribution des genres journalistiques dans les deux corpus.

Si ces jeux de données ont été introduits dans Anonymisé 2024a, seule une petite partie en a été analysée dans le cadre d'une expérience d'annotation (100 articles en français répartis également entre les deux corpus). Ici, nous analysons l'ensemble du corpus, de manière exhaustive et approfondie[7].

5. Lien url anonymisé.
6. https://www.sgdsn.gouv.fr/publications/maj-19062023-rrn-une-campagne-numeriqu

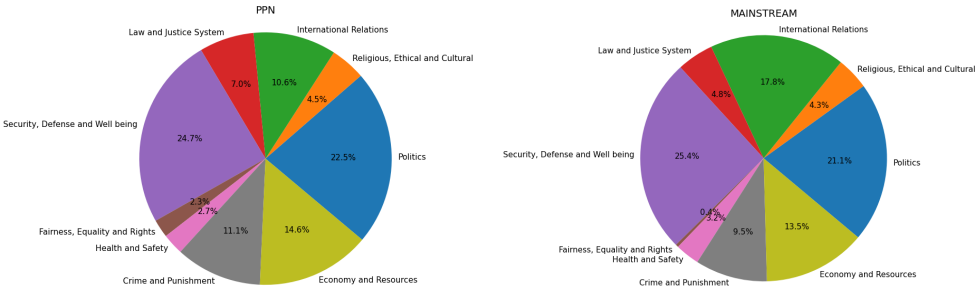

FIGURE 2 – Distribution thématique dans les deux corpus.

3 Genres, thématiques et techniques de persuasion

Afin de distinguer les deux types d'actualités (fiables, propagandistes), nous commençons par une analyse comparative des distributions de genres, de thématiques et de techniques de persuasion à travers les deux corpus. Ces annotations ont été obtenues au moyen des API publiques fournies par le classifieur d'actualités GATE Cloud[8], entraîné sur divers jeux de données académiques et de recherche.

Genres. Une distribution des genres en trois catégories (Reporting/Opinion/Satire) est présentée en Figure 1 pour les deux corpus. La comparaison met en évidence des différences marquées entre les deux corpus : MAINSTREAM se caractérise par une plus forte proportion d'articles de type Reporting, plus de six fois supérieure à celle de NOM-ANONYMISÉ. MAINSTREAM présente inversement une plus faible proportion de Satire. Enfin, si la catégorie Opinion est représentée à plus de 50% dans les deux corpus, elle constitue plus des trois quarts du corpus propagandiste NOM-ANONYMISÉ, ce qui confirme le lien entre propagande et persuasion (Da San Martino *et al.*, 2019), ainsi que la tendance de la propagande à brouiller la frontière entre comptes rendus factuels et tribunes d'opinion.

Thématiques et techniques de persuasion. Pour compléter cette comparaison, nous avons utilisé la même suite d'outils d'annotation afin d'obtenir la distribution thématique des articles. Une partition en neuf thématiques est présentée en Figure 2, montrant que les trois corpus ont des distributions relativement proches.

Enfin, les thématiques s'avérant peu discriminantes, nous avons mené une troisième analyse distributionnelle, cette fois relativement à l'ensemble des techniques de persuasion définies dans Piskorski *et al.* 2023, en utilisant de nouveau le classifieur multilingue de techniques de persuasion de Cloud. La distribution des techniques de persuasion par article est présentée en Figure 3.

Les graphiques indiquent que les articles de propagande du corpus NOM-ANONYMISÉ ont tendance à mobiliser un plus grand nombre de ces techniques de persuasion, ce qui est cohérent avec le fait qu'une large part de ce corpus est classée en Opinion ou en Satire. En particulier, on observe un

e-de-manipulation-de-linformation-complexe-et

7. En raison de contraintes de droits d'auteur, les textes ne sont pas redistribués, mais des liens directs vers les pages web correspondantes sont fournis avec les annotations associées.

8. https://cloud.gate.ac.uk/shopfront#tagged=Misinformation

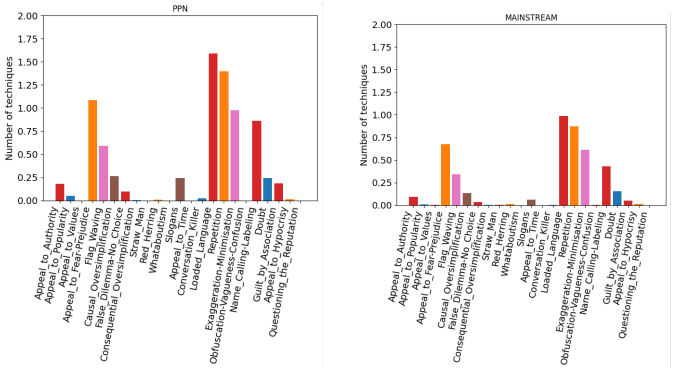

FIGURE 3 – Nombre moyen de techniques de persuasion par article, pour les deux corpus.

recours plus fréquent au *langage chargé* (loaded language), à la répétition, ainsi qu'aux techniques d'exagération et de minimisation dans le corpus propagandiste.

Dans l'ensemble, ces analyses montrent que, tant du point de vue des genres que des techniques de persuasion, les articles propagandistes sont plus facilement reconnaissables que d'autres types d'articles alors même qu'ils ont des couvertures thématiques comparables. Ce constat suggère la possibilité d'améliorer la détection de la propagande en tenant compte des genres et des techniques de persuasion, et potentiellement d'autres caractéristiques stylistiques (telles que celles rapportées dans Anonymisé 2024a). Pour explorer cette voie, la section qui suit propose une approche hybride intégrant des représentations neuronales et symboliques, en combinant des caractéristiques textuelles avec des concepts extraits du contenu.

4 Approche neurosymbolique pour la détection de la propagande

4.1 Architecture proposée

Dans un premier temps, les textes sont encodés à l'aide des *embeddings* pré-entraînés fastText (Bojanowski *et al.*, 2017). Ce choix vise un compromis favorable entre coût et robustesse, en s'appuyant sur des représentations parcimonieuses et stables sur gros volumes, complétées par des caractéristiques conceptuelles. Pour chaque article, ce processus produit un vecteur de 300 dimensions représentant les caractéristiques lexicales distributionnelles du texte. En complément de ce vecteur, nous ajoutons des informations relatives au genre, à la thématique et aux techniques de persuasion présentes dans les articles. Bien que les thématiques n'aient pas semblé particulièrement discriminantes dans la section précédente, elles sont néanmoins incluses ici afin de suivre d'éventuelles différences dans d'autres corpus.

— L'information de genre est encodée par *one-hot encoding* (OHE), produisant un vecteur ne contenant que des zéros à l'exception de la dimension correspondant à la modalité encodée, qui vaut un. Dans notre cas, cela ajoute 3 dimensions spécifiques (Reporting, Opinion et Satire).
— L'information thématique est également encodée en *one-hot*, ajoutant 9 dimensions (pour les

thématiques affichées en Figure 2).

— L'information relative aux techniques de persuasion est ajoutée sous la forme d'un vecteur comptant, pour chaque type, le nombre de techniques de persuasion présentes dans l'article. Les techniques de persuasion fines correspondent à 23 dimensions. Toutefois, ces techniques peuvent être regroupées en catégories plus grossières, ce qui conduit à seulement 6 dimensions (Piskorski *et al.*, 2023).

Au total, un vecteur de 35 dimensions ($= 3+9+23$) correspondant aux techniques de persuasion fines est ajouté à l'*embedding* fastText de 300 dimensions. Ce vecteur est ensuite traité par un perceptron à deux couches, composé d'une couche dense (de 335 dimensions vers 335 dimensions) avec une fonction d'activation ReLU, puis d'une couche dense (de 335 dimensions vers 1 dimension) avec une fonction d'activation sigmoïde, afin d'obtenir un score estimant la propagande entre 0 et 1. Le texte est ensuite classé à l'aide d'un seuil de 0,5. Une vue d'ensemble de l'architecture proposée est présentée en Figure 4. Lorsque l'ensemble des techniques de persuasion est inclus, nous appelons la méthode obtenue *Hybrid Method*. Une variante allégée est obtenue en utilisant un vecteur de 18 dimensions ($= 3 + 9 + 6$) n'intégrant que les techniques de persuasion regroupées en catégories grossières : nous appelons la méthode correspondante *Hybrid Lite*.

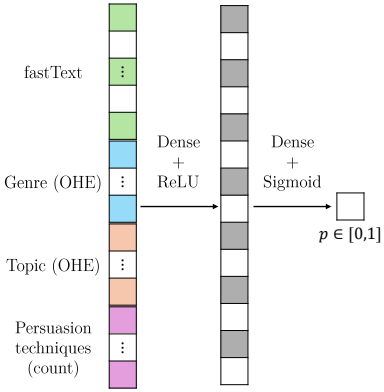

FIGURE 4 – Notre architecture hybride, combinant des caractéristiques neuronales et des concepts extraits.

Cette architecture est volontairement simple afin de permettre des analyses d'explicabilité, difficiles à mener avec des grands modèles de langue.

4.2 Méthodologie d'évaluation

Afin de limiter le surapprentissage, nous avons conçu une nouvelle méthodologie d'évaluation visant à garantir que l'approche proposée généralise mieux à de nouveaux événements et à de nouvelles sources.

L'idée générale consiste à partitionner les données en ensembles d'entraînement, de validation et de test selon différents critères, détaillés ci-dessous. Le modèle est entraîné sur l'ensemble d'entraînement et évalué sur l'ensemble de validation après chaque époque. Nous utilisons un arrêt anticipé (*early*

stopping) avec une patience de 20 afin de suivre le score F1 sur l'ensemble de validation, les articles de propagande étant la classe à détecter.

Les modèles sont entraînés avec une fonction de perte d'entropie croisée et l'optimiseur Adam (Kingma & Ba, 2015), avec un taux d'apprentissage de 10^{-4}, pour un maximum de 300 époques. Si aucune amélioration du score F1 sur l'ensemble de validation n'est observée pendant 20 époques, le meilleur modèle est restauré puis évalué sur l'ensemble de test, ce qui fournit les scores reportés dans les tableaux de résultats.

Sources	**Entraînement**	**Validation**	**Test**
MAINSTREAM	APNews The Guardian	CNN USA Today Forbes Fox News NBC News NYTimes Washington Post	CBSNews Daily Mail
NOM-ANONYMISÉ	RRN	TribunalUkraine War on Fakes	

Politique	**Entraînement**	**Validation**	**Test**
MAINSTREAM	APNews CNN USA Today NBC News NYTimes Washington Post CBSNews	Daily Mail Forbes Fox News	
NOM-ANONYMISÉ	Ensemble du corpus NOM-ANONYMISÉ (anglais)		

Crédibilité	**Entraînement**	**Validation**	**Test**
MAINSTREAM	APNews CNN USA Today Forbes The Guardian NYTimes Washington Post CBSNews	Daily Mail Fox News	
NOM-ANONYMISÉ	Ensemble du corpus NOM-ANONYMISÉ (anglais)		

TABLE 1 – Répartition des actualités dans les découpages **Sources**, **Politique** et **Crédibilité**.

La manière dont sont définis les ensembles d'entraînement/validation/test permet d'évaluer la robustesse, c'est-à-dire l'efficacité des caractéristiques apprises dans de nouveaux contextes. Dans cette optique, nous utilisons la partie anglaise des jeux de données NOM-ANONYMISÉ et MAINSTREAM, et définissons quatre types de découpage pour nos expérimentations :

— **Aléatoire** : les articles sont échantillonnés aléatoirement afin de produire des ensembles 80%–10%–10%.
— **Sources** : les ensembles ne contiennent que des articles provenant de sources spécifiques. Les sources ont été choisies de manière à obtenir une répartition approximative 80%–10%–10%. Le découpage par sources est donné dans le Tableau 1 (haut).
— **Politique** : les articles MAINSTREAM sont répartis selon leur annotation d'orientation politique fournie par MediaBiasFactCheck.[9] Comme la majorité des articles proviennent de sources plutôt situées à gauche, nous les utilisons comme sources d'entraînement, et répartissons aléatoirement les sources plutôt situées à droite entre validation et test. Les articles de propagande sont répartis aléatoirement entre les ensembles selon une distribution 80%–10%–10% (voir le Tableau 1, milieu).
— **Crédibilité** : de manière analogue à l'orientation politique, MediaBiasFactCheck propose des évaluations de crédibilité des sources, fondées sur la liberté de la presse, la factualité des articles, la propriété du média et des vérifications antérieures. À ce titre, tous les articles de propagande proviennent de sources à faible crédibilité. Comme une large majorité des articles ordinaires proviennent de sources à forte crédibilité, nous utilisons ces sources pour l'entraînement et les sources à faible crédibilité pour la validation et le test. Les articles de propagande sont répartis aléatoirement entre les ensembles selon une distribution 80%–10%–10% (voir le Tableau 1, bas).

9. Il est important de noter que MediaBiasFactCheck semble s'inscrire dans la fenêtre d'Overton américaine, de sorte que leurs annotations politiques peuvent différer de ce qui serait retenu dans d'autres pays.

Chaque découpage est conçu pour évaluer un type de biais potentiel du modèle, lié d'abord aux sources, susceptibles de combiner plusieurs biais non identifiés, puis à l'orientation politique et à la crédibilité.

5 Résultats

Cette section est organisée en trois parties. La première présente les résultats de l'approche proposée selon les différents découpages. La deuxième partie conduit des études d'ablation afin de mesurer l'apport de l'ajout d'*embeddings* conceptuels aux *embeddings* textuels. Enfin, une analyse d'explicabilité met en évidence les cas pour lesquels l'approche proposée est la plus bénéfique.

5.1 Résultats principaux

Les résultats obtenus pour les différents découpages sont présentés en ligne 1 du Tableau 2 (Hybrid), en termes d'exactitude (*Accuracy*, proportion de prédictions correctes) et de score F1. À noter que les ensembles de test diffèrent d'une colonne à l'autre. La colonne **Random** correspond à l'évaluation classique. La colonne **Sources** correspond au cas où le système est confronté à de nouvelles sources, la colonne **Political** au cas où il est confronté à de nouvelles orientations politiques, et la colonne **Credibility** au cas où il est confronté à des sources dont la crédibilité diffère de celle rencontrée à l'entraînement.

Les résultats obtenus sur le découpage **Random** sont modérés, mais restent satisfaisants au regard de la petite taille du modèle. Le système affiche des performances comparables pour les découpages **Sources** et **Credibility**, mais éprouve davantage de difficultés face à de nouvelles orientations **Political**. Par rapport à **Credibility**, le découpage **Political** n'inclut pas *Forbes* dans les ensembles de validation et de test, mais l'inclut dans l'ensemble d'entraînement aux côtés de *The Guardian*, et sans *CBSNews*. Ces variations suffisent à dégrader les performances, ce qui suggère que les ensembles d'entraînement devraient couvrir une plus grande diversité d'orientations politiques afin de construire des systèmes plus robustes.

5.2 Études d'ablation

L'approche proposée vise à améliorer la robustesse dans de nouveaux scénarios en combinant des caractéristiques conceptuelles avec des *embeddings* textuels, afin de réduire le surapprentissage. Pour évaluer les performances de notre méthode Hybrid, nous avons mené des études d'ablation. Tout d'abord, un modèle n'utilisant que les *embeddings* fastText est entraîné et évalué (Tableau 2, Text Only). Ensuite, les caractéristiques relatives aux techniques de persuasion sont modifiées afin de ne prendre en compte que des catégories de persuasion plus grossières (6 au lieu de 23 ; voir Tableau 2, Hybrid Lite).

Plusieurs observations peuvent être faites :

— Dans presque tous les cas, l'utilisation d'étiquettes fines pour les techniques de persuasion (Hybrid) améliore les performances par rapport à l'utilisation d'étiquettes grossières (Hybrid Lite). Une exception concerne le découpage **Random** ; toutefois, le gain est important lorsque

Méthode	Random		Sources		Political		Credibility	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Hybrid	78.08	87.5	**79.45**	**88.37**	69.86	81.66	**79.45**	**86.95**
Hybrid Lite	**79.45**	**88.54**	34.24	31.42	67.12	79.66	56.16	69.81
Text Only	**79.45**	**88.54**	20.54	0.0	**79.45**	**88.54**	20.54	0.0

TABLE 2 – Résultats pour la détection de la propagande selon différents découpages des données et différentes ablations.

la méthode Hybrid Lite est en difficulté (**Sources**, **Credibility**), et la baisse est faible dans les autres cas.

— En moyenne, la méthode Hybrid améliore les performances par rapport à la méthode Text Only (+26,89% d'exactitude et +41,85% de score F1 en moyenne), bien qu'elle soit moins performante sur **Random** et **Political**. Globalement, la méthode Hybrid est néanmoins la plus robuste sur les quatre découpages : elle apprend dans les quatre cas, et présente la meilleure performance moyenne avec la plus faible variance ($\mu_{F1} = 86.12, var_{F1} = 9.18$).

— À l'inverse, si la méthode Text Only surpasse les autres sur deux découpages (**Random** et **Political**), dans les découpages **Sources** et **Credibility**, les *embeddings* textuels ne suffisent pas à apprendre à discriminer les articles de propagande des articles de la presse généraliste. Cela indique que la méthode Text Only n'est pas robuste aux perturbations de l'ensemble d'entraînement, même si elle ne surapprend pas sur l'orientation politique.

En résumé, même si l'approche proposée n'obtient pas les meilleures performances selon le découpage aléatoire classique, elle présente une robustesse et une capacité de généralisation supérieures à l'approche équivalente fondée uniquement sur le texte, qui s'effondre dans deux des scénarios. Un autre avantage réside dans la simplicité de cette approche, qui permet l'application de méthodes d'explicabilité, auxquelles nous nous intéressons à présent.

5.3 Analyses d'explicabilité

Pour expliquer notre modèle hybride, nous avons utilisé SHAP (Lundberg & Lee, 2017). Pour obtenir une représentation globale de ce qu'un modèle a appris, nous moyennons les valeurs absolues des sommes des valeurs SHAP sur les différents découpages. Nous regroupons les valeurs SHAP par catégories afin d'en améliorer la lisibilité. Pour chaque exemple de l'ensemble associé à chaque découpage, nous calculons la valeur absolue de la somme de toutes les valeurs SHAP correspondant à l'encodage textuel, et procédons de même pour les *embeddings* de genre, de thématique et de techniques de persuasion. Les valeurs SHAP moyennes par groupe de découpage et par catégorie sont présentées en Figure 5.

La figure montre que, pour **Random**, les *embeddings* textuels contribuent le plus à la décision de classification. Pour **Credibility**, ce sont comparativement les *embeddings* de persuasion qui contribuent le plus. Pour **Sources** et **Political**, les résultats sont plus contrastés entre les différents groupes de caractéristiques.

Dans le cas **Random**, cela s'explique probablement par le fait qu'un échantillonnage aléatoire rend l'ensemble d'entraînement mieux aligné avec les ensembles de validation et de test, qui sont dès lors suffisamment informatifs pour la classification. Ce point est confirmé par les résultats, qui montrent de meilleures performances avec les seules caractéristiques textuelles.

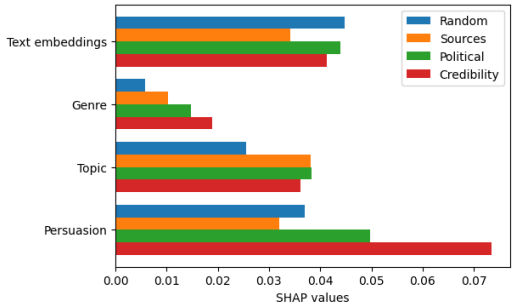

FIGURE 5 – Valeurs SHAP moyennes pour la méthode Hybrid selon le découpage.

En revanche, lorsque la distribution du test diffère de celle de l'entraînement, la méthode tend à mobiliser au moins autant d'*embeddings* conceptuels que d'*embeddings* textuels. En particulier, pour les découpages **Credibility** et **Sources**, c'est le recours à des caractéristiques conceptuelles qui permet au modèle de mieux généraliser à de nouvelles sources et à des articles dont la fiabilité est incertaine. Cette situation correspond au cas où un article est suspect et provient de sources inconnues, ce qui rend cette approche adaptée à la détection de la propagande.

6 Conclusion

Cet article a présenté une méthode de détection de la propagande qui intègre des plongements fastText et des caractéristiques conceptuelles extraites à partir d'une comparaison croisée des corpus NOM–ANONYMISÉ et MAINSTREAM. Les corpus diffèrent significativement du point de vue du vocabulaire et des stratégies de persuasion, ce qui suggère que des modèles entraînés sur une seule source peuvent manquer des signaux spécifiques à un corpus. Ces observations ont motivé l'inclusion de caractéristiques conceptuelles inter-corpus afin de mieux capturer les schémas de la propagande.

En concevant des découpages biaisés de nos jeux de données, qui correspondent à l'exposition du modèle à de nouveaux types d'articles, nous avons montré que l'ajout d'informations conceptuelles extraites des textes améliore les performances de détection, en particulier lorsque de nouvelles sources, associées à des niveaux de crédibilité variables, sont rencontrées.

Cependant, les expériences ont été menées sur un corpus centré sur un thème principal : le conflit Russie-Ukraine. Des expériences supplémentaires pourraient être réalisées sur d'autres thématiques récentes, telles que des élections récentes, ou d'autres conflits. Les corpus NOM–ANONYMISÉ et MAINSTREAM n'ont par ailleurs été traités qu'en anglais, et des expériences analogues devraient être conduites dans d'autres langues afin d'identifier d'éventuelles différences spécifiques aux langues.

Enfin, d'autres types de caractéristiques conceptuelles pourraient être mobilisées, sur la base d'autres systèmes de connaissances expertes, voire d'opérateurs humains. Dans d'autres travaux, un système expert d'estimation de l'imprécision (*vagueness*) a été combiné avec succès à un modèle de langue pour la tâche de détection de la subjectivité (Anonymisé, 2024b). Il pourrait être envisagé d'ajouter une estimation de la fiabilité de la source d'un document à un modèle de classification, afin d'améliorer les performances d'un classifieur fondé uniquement sur le texte.

Références

ANONYMISÉ (2024a). Anonymisé.

ANONYMISÉ (2024b). Anonymisé.

BALY R., KARADZHOV G., ALEXANDROV D., GLASS J. & NAKOV P. (2018). Predicting factuality of reporting and bias of news media sources. In E. RILOFF, D. CHIANG, J. HOCKENMAIER & J. TSUJII, Éds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 3528–3539, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1389.

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.

DA SAN MARTINO G., CRESCI S., BARRÓN-CEDEÑO A., YU S., DI PIETRO R. & NAKOV P. (2020). A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*, p. 4826–4832. Survey track, DOI : 10.24963/ijcai.2020/672.

DA SAN MARTINO G., YU S., BARRÓN-CEDEÑO A., PETROV R. & NAKOV P. (2019). Fine-grained analysis of propaganda in news articles. In K. INUI, J. JIANG, V. NG & X. WAN, Éds., *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. 5636–5646, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1565.

HAMBORG F., DONNAY K. & GIPP B. (2019). Automated identification of media bias in news articles : An interdisciplinary literature review. *International Journal on Digital Libraries*, **20**(2), 391–415. DOI : 10.1007/s00799-018-0261-y.

HU B., MAO Z. & ZHANG Y. (2025). An overview of fake news detection : From a new perspective. *Fundamental Research*, **5**(1), 332–346. DOI : 10.1016/j.fmre.2024.01.017.

KINGMA D. P. & BA J. (2015). Adam : A method for stochastic optimization. In Y. BENGIO & Y. LECUN, Éds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

LUNDBERG S. M. & LEE S.-I. (2017). A unified approach to interpreting model predictions. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems 30*, p. 4765–4774. Curran Associates, Inc.

MA J., GAO W. & WONG K.-F. (2017). Detect rumors in microblog posts using propagation structure via kernel learning. In R. BARZILAY & M.-Y. KAN, Éds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 708–717, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/P17-1066.

PISKORSKI J., STEFANOVITCH N., DA SAN MARTINO G. & NAKOV P. (2023). SemEval-2023 task 3 : Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup. In A. K. OJHA, A. S. DOĞRUÖZ, G. DA SAN MARTINO, H. TAYYAR MADABUSHI, R. KUMAR & E. SARTORI, Éds., *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, p. 2343–2361, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.semeval-1.317.

RUCHANSKY N., SEO S. & LIU Y. (2017). Csi : A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM'17, p. 797–806, New York, NY, USA : ACM. DOI : 10.1145/3132847.3132877.

SHU K., MAHUDESWARAN D., WANG S., LEE D. & LIU H. (2020). Fakenewsnet : A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, **8**(3), 171–188.

THORNE J., VLACHOS A., CHRISTODOULOPOULOS C. & MITTAL A. (2018). FEVER : a large-scale dataset for fact extraction and VERification. In M. WALKER, H. JI & A. STENT, Éds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 809–819, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1074.

WARDLE C. & DERAKHSHAN H. (2017). Information disorder : Toward an interdisciplinary framework for research and policymaking.

ZHOU X. & ZAFARANI R. (2020). A survey of fake news : Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, **53**(5). DOI : 10.1145/3395046.