Peut-on prédire efficacement des "Fake News" à partir du dataset ISOT ?

Alexis Barrau*
ENSAE
91120 Palaiseau
alexis.barrau@ensae.fr

Abstract

Ce travail s'inspire de l'article de Hoy and Koulouri [2022] pour étudier la possibilité de construire un modèle de détection de "Fake News" à partir du dataset ISOT qui pourrait être généralisé à d'autres datasets. Il se concentre sur deux méthodes principales pour cela, une factorisation TF-IDF suivie d'une régression logistique ou l'utilisation du modèle BERT pour extraire les caractéristiques des textes. Les différents modèles testés se révèlent extrêmement précis lorsque testés sur une portion (non utilisée) du dataset ISOT mais voient leurs performances chuter drastiquement dès lors qu'ils sont appliqués à de nouveaux datasets. D'une manière générale, la performance du modèle dépends avant tout du choix de l'extracteur de caractéristiques utilisée alors que le modèle de prédiction n'a que peu d'influence sur les résultats. Plus spécifiquement, les conditions différentes de collecte et d'agrégation des différents datasets limite fortement la possibilité de généraliser un modèle sur un nouveau dataset, rendant difficile la possibilité de construire un prédicteur qui pourrait être appliqué à la détection de "Fake News" dans différents contextes. Seul une augmentation de la taille des datasets disponibles mais plus encore une diversification de leurs sources semblent à même de permettre la réalisation d'un tel modèle généralisable.

1 Introduction

La notion de "Fake News" s'est imposée dans le débat publique, puis la recherche scientifique, depuis une dizaine d'année, à la suite notamment de la campagne présidentielle américaine de 2016 durant laquelle le candidat Donald Trump a fait usage de cette notion pour décrédibiliser une partie des critiques le visant tout en étant lui-même accusé d'y recourir de façon massive. De fait, ces "Fake News" sont perçues comme une menace pour la démocratie et la cohésion sociale, à même de détruire l'intégrité de l'espace informationnel (OCDE [2024]). De ce point de vue, le développement d'outils capables de les détecter automatiquement apparaît comme une piste prometteuse afin de limiter leur influence.

La notion de "Fake News" pose néanmoins un certains nombre de problèmes conceptuels, qu'il s'agisse de sa délimitation par rapport à des concepts proches (rumeur, propagande, théorie du complot...), de ses caractéristiques (circulation, support...) ou des intentions de son auteur et des personnes qui la relaient (Giry [2020]). S'il ne s'agit pas ici de revenir sur ces débats, ils doivent être pris en compte dès lors que l'on se propose, comme c'est le cas dans ce travail, de construire un prédicteur capable de reconnaître, à partir du texte d'un article, une "Fake News". Cette démarche implique en effet que la notion s'applique au texte lui-même, indépendamment de son contexte d'énonciation, de la façon dont elle circule, ou même des intentions de l'auteur dès lors qu'elles

*Le dépôt GitHub associé à ce travail se situe à l'adresse https://github.com/Alexis-Barrau/MLforNLP

Preprint. Under review.

n'apparaissent pas dans le texte via des marqueurs linguistiques.[2] Le choix qui est fait ici est donc de se reposer sur des datasets (en langue anglaise) déjà labellisés, disponibles sur Kaggle (voir section suivante), et donc de considérer comme "Fake News" tout texte labellisé ainsi dans nos jeux de données.[3]

Dans leur article, Hoy and Koulouri [2022] cherchent à tester la possibilité de construire un prédicteur efficace de "Fake News" en mobilisant plusieurs datasets et plusieurs techniques d'analyses du langage naturel (extraction de caractéristiques) et plusieurs modèles prédicteurs. La conclusion principale de leur article est que, bien que la majorité des modèles testés atteignent des performances particulièrement élevées sur les ensembles de test, donc sur une partie non vue d'un dataset sur lequel a été réalisé l'entraînement, ceux-ci obtiennent des performances particulièrement dégradées lorsque appliquées à d'autres datasets. Plus encore, ils ne trouvent que peu de différences entre les extractions de caractéristiques les plus simples, de type *Bag-of-words* ou *TF-IDF*, et une extraction reposant sur le modèle BERT.[4] L'explication proposée repose sur la présence de biais dans la construction des datasets, qui mobilisent des sources privilégiées mais fortement spécifiques, avec la présence de mots qui caractérisent dans un dataset des vraies nouvelles alors qu'ils caractérisent plutôt les "Fake News" d'un autre.

Ce travail se propose de creuser cette question, en se concentrant plus spécifiquement sur le plus gros dataset de leur comparaison, le dataset ISOT, et en essayant de construire à partir de celui-ci un prédicteur qui est ensuite appliqué deux autres datasets. Différents modèles sont envisagés mais deux principales méthodes sont appliqués : 1) La construction d'un modèle "baseline" reposant sur une analyse de fréquence des mots dans les articles (TF-IDF) puis une régression logistique, 2) l'utilisation du modèle BERT afin de permettre un encodage contextuel des mots des articles, dans l'espoir que cette prise en compte du contexte améliore la généralisabilité à d'autres datasets. En effet, au regard des resultats de Hoy and Koulouri [2022] ce travail fait le choix d'appronfondir l'analyse de ces modèles plutôt que de multiplier des approches dont les résultats se sont révélés particulièrement proches.

2 Données

Les données mobilisées ici sont les mêmes que celles de l'article dont ce travail s'inspire, à l'exclusion d'un des datasets mobilisés, *FakeNewsNet*, qui a été exclu en raison de sa taille beaucoup plus modeste (422 observations), ne permettant pas un entraînement du même ordre que les autres. Le choix de s'appuyer sur les même données permet ainsi de s'assurer que les datasets portent sur des domaines proches, et contiennent des "Fake News" intentionnelles et non des articles satiriques. Les trois datasets mobilisés sont les suivants :[5].

ISOT [6] Le dataset ISOT (Information Security and Object Technology, du nom de la structure à l'origine de sa création au sein de l'université de Virginie) contient principalement des informations politiques américaines et mondiales. Les 23 481 articles "Fake News" ont été collectés sur des sites considérés comme non fiables (d'après Wikipedia et Politifact.com), alors que les 21 417 vraies nouvelles sont toutes issues de l'agence de presse Reuters.com. En plus du texte de l'article et du label, le dataset contient le titre de l'article, ainsi que la date de publication et le sujet (non mobilisés dans ce travail).

Kaggle Fake News Competition [7] Le dataset *Kaggle Fake News Competition* (abrégé ci-après *fake_news*) continent le titre, le texte et le label (ainsi que l'auteur, non mobilisé) des 10 413 vrais articles et des 10 387 "Fake News". Celui-ci est construit en combinant plusieurs autres datasets disponibles sur *Kaggle*. Le dataset contient de plus 39 observations dont le texte est manquant,

[2] De ce point de vue, l'utilisation de formules ironiques et satiriques peuvent traduire une intention humoristique qui distinguerait le texte d'une "Fake News" à proprement parler.

[3] Ceci explique le choix fait de ne pas traduire "Fake News", afin de coller au plus prêt à ces labels.

[4] Notons qu'ils utilisent la version pré-entrainé de BERT pour extraire les caractéristiques avant construction d'un prédicteurs sur les *features* ainsi obtenues, sans opérer de fine-tuning complet.

[5] la description des datasets repose largement sur celle présente dans l'article de Hoy et Koulouri

[6] https://www.uvic.ca/ecs/ece/isot/datasets/fake-news/index.php

[7] https://www.kaggle.com/c/fake-news

		Prédiction	
		Vraie	**Fausse**
Label	**Vraie**	4284	46
	Fausse	77	4573

Table 1: Performance du modèle baseline sur l'ensemble de test

exclues de l'analyse (de même que les 558 observations dont le titre est manquant lorsque l'analyse porte sur les titres).

Kaggle Fake or Real [8] Le dataset *Kaggle Fake or Real* (abrégé ci-après *fake_real*) contient le titre, le label et le texte de 6 060 articles, répartie de façon égale entre "Fake News" et vraies nouvelles. Les conditions de construction du dataset ne sont cependant pas indiquées clairement, de même que les sources.

En complément de ces trois datasets, ce travail mobilise aussi différents panachages entre ISOT et *fake_news* dans le but de tester si une diversification des sources permet d'améliorer la généralisabilité du modèle (testé donc uniquement sur *fake_real*). Ce panachage est opéré en sélectionnant tout d'abord uniquement la moitié du dataset ISOT, afin d'avoir un dataset de taille similaire à celle de *fake_news*. Deux types de panachages sont ensuite testés : le regroupement des vraies nouvelles d'un dataset avec les fausses de l'autres ("New data 1" et "New data 2"), ou encore un mélange à parité des vraies nouvelles et fausses nouvelles provenant de chaque dataset ("New data 3"). Les datasets ainsi construits restent donc équilibrés entre les labels.[9]

3 Modèles reposant sur une extraction de caractéristiques TF-IDF

3.1 Modèle baseline sur ISOT

Le premier modèle utilisé repose sur le principe d'une extraction de caractéristiques TF-IDF (pour *Term Frequency - Inverse Document Frequency*. Il s'agit d'une méthode relativement simple, qui consiste à compter le nombre d'occurrences d'un mot dans un document, divisé par le nombre de mots du document (ce qui donne donc sa fréquence) pondéré par (le logarithme de) l'inverse de la fréquence des documents contenant ce mot dans le corpus. L'avantage de cette représentation par rapport à un simple *Bag-of-words* est qu'elle permet d'accorder plus d'importance aux mots spécifiques, n'apparaissant que dans un nombre plus limité de documents. Cette méthode permet donc de construire une représentation de chaque article sous forme d'un vecteur de ces fréquences inversées, sur laquelle il est ensuite possible de construire un prédicteur.

Le choix a été fait de construire cette représentation à l'aide de la fonction intégrée à la bibliothèque Python `scikit-learn`, `TfidfVectorizer`. Celle-ci effectue automatiquement un certain nombre d'opérations de nettoyage (suppression de la casse, de la ponctuation...) puis la vectorisation des textes du corpus, et contient de plus une liste de "stop words" qui seront ignorés car très commun en Anglais, qui a été intégrée à la vectorisation. De plus, le choix a été fait d'exclure de la vectorisation les mots qui apparaissent dans plus de 70 % des documents du corpus. La calibration du modèle (pour les poids *IDF*) a été opéré sur l'ensemble d'entraînement, représentant 80 % du dataset *ISOT*, de façon à ce que la pondération ne dépende pas de l'ensemble de test, afin de donner à celui-ci le même statut que les autres datasets qui seront ensuite testés. L'analyse s'est de plus concentrée sur le texte des articles, sans tenir compte des titres. La prédiction à proprement parlé a elle été faite via une régression logistique.

		Prédiction	
		Vraie	**Fausse**
Label	**Vraie**	1896	8491
	Fausse	858	9516

Table 2: Performance du modèle baseline sur "Fake News Competition"

		Prédiction	
		Vraie	**Fausse**
Label	**Vraie**	727	2444
	Fausse	291	2873

Table 3: Performance du modèle baseline sur "Fake or Real"

[8] https://www.kaggle.com/datasets/jillanisofttech/fake-or-real-news

[9] Pour plus de détail sur ces constructions se reporter au dépôt GitHub.

Le modèle baseline ainsi obtenu se révèle particulièrement efficace pour reconnaître les "Fake News" de l'ensemble de test puisqu'il atteint une précision globale[10] de 98,63 % sur l'ensemble de test (voir table 3.1). Ce résultat est cohérent avec le papier de Hoy and Koulouri [2022] qui montre les très bonnes performances de modèles même simples. Néanmoins, le même modèle, (vectorisation avec les mêmes poids et régression logistique ajusté sur le train) se révèle beaucoup moins efficace lorsqu'il est appliqué à nos deux autres datasets, *fake_news* et *fake_real*, comme en témoignent les tables 3.1 et 3.1). En effet, la précision globale chute alors à 54,97 % sur *fake_news* et 56,83 % sur *fake_real*, soit à peine plus que ce que ferait un prédicteur aléatoire en présence de datasets équilibrés, comme le sont les notres.

Surtout, de façon plus intéressante, comme le montrent les matrices de confusion présentées, le modèle s'avère particulièrement biaisé en faveur de la détection de "Fake News". De fait, sur le dataset *fake_news* le rappel des vraies nouvelles, c'est à dire la proportion de vraies nouvelles correctement identifiées par le modèle, s'élève à seulement 18 %, alors que ce rappel monte à 23 % sur *fake_real*. Tout se passe donc comme si le modèle se montrait incapable de reconnaître les vraies nouvelles provenant d'un autre dataset que celles de ISOT. Cette situation peut s'interpréter comme **une forme particulière de suraprentissage**,[11] dans laquelle le modèle n'apprends pas spécifiquement les données d'entraînement (ce qui conduirait à des résultats mauvais sur l'ensemble de test) mais des caractéristiques particulières du dataset, liées à ses conditions de collectes, ses sources etc. Il s'agit d'ailleurs de la piste d'explication suggérée à la fin de l'article ayant servi de base à ce travail.

3.2 Variations du modèle baseline

Face aux limites du modèle baseline proposé ci-dessus, plusieurs pistes ont été explorées afin de chercher à améliorer la généralisabilité de notre prédicteur, ou à défaut de préciser les raisons de cette chute drastique de performance en changeant de dataset.

3.2.1 Modification de stop words

L'article de Hoy and Koulouri [2022] relève dans sa dernière section la présence presque systématique du terme "Reuters" dans les vraies articles de ISOT, lié à la source des articles. De fait, ce mot apparaît en cumulé 28 976 fois dans les vraies nouvelles, contre seulement 449 fois dans les fausses. Or, ceci est une spécificité de ce dataset qui ne signifie aucunement que ce mot doive apparaître de façon systématique dans une vraie nouvelle. Ainsi, le modèle baseline a été reproduit au détail près que le mot "Reuters" a été ajouté à la liste des mots ignorées de la fonction `TfidfVectorizer`.

Certains résultats de ce modèle sont présentés dans la table 4. La précision globale sur l'ensemble de test chute d'environ un point par rapport à notre modèle baseline, mais ceci était attendu car nous ignorons désormais un mot particulièrement discriminant. Par contre, celle-ci *augmente* sur les deux autres datasets, de respectivement environ 1 et 2 points. Surtout, si nous constatons à nouveau que le rappel des vraies nouvelles est particulièrement faible sur nos autres datasets, celle-ci croît cependant de 5 points sur *fake_news* et de 7 points sur *fake_real*. Ces résultats confortent donc largement notre hypothèse, à savoir que le prédicteur construit apprends des caractéristiques particulièrement spécifiques du dataset ISOT, et probablement en particulier de ses vraies articles.[12]

Dataset	**ISOT (test)**	**Fake News Competition**	**Fake or Real**
Précision globale	0.98	0.56	0.59
Rappel vraies nouvelles	0.98	0.23	0.30
Rappel "Fake News"	0.98	0.89	0.88

Table 4: Résultats partiels du modèle baseline avec "Reuters" dans les mots ignorés

[10] "accuracy" en Anglais

[11] Le terme étant ici un abus de langage par analogie avec une situation bien connue.

[12] une recherche d'autres mots pouvant jouer un rôle similaire a été effectuée afin de les exclure mais aucun n'a pu être mis en évidence de façon aussi évidente que "Reuters". En effet, il faudrait pour cela un mot qui trahisse une source particulière des articles, qui serait spécifique au dataset.

3.2.2 Prédiction à partir des titres

Si la reconnaissance de "Fake News" s'appuyant sur les textes des articles semble conduire à un suraprentissage des caractéristiques des (vraies) articles de ISOT, une possibilité de contourner ce problème pourrait être de ne s'intéresser non pas aux textes des articles mais à leurs titres, moins susceptibles de trahir une source particulière, ou une confition de collecte. Un modèle similaire au modèle baseline a donc été réalisé à partir de ceux-ci.

Les résultats partiels sont présentés dans la table 5. Ceux-ci montrent que les résultats globaux restent proches de ceux obtenus pour notre modèle baseline, malgré une chute de précision globale de 5 points sur l'ensemble de test, qui se maintient tout de même à 94,51 %. Cette performance est notable au regard de la forte diminution des informations données au modèle. Plus encore, nous constatons un maintien de la précision globale en généralisation, qui diminue certes d'un point sur *fake_real* mais augmente de de 5 points sur *fake_news*. Surtout, nous ne constatons plus de difficulté spécifique de notre modèle à détecter les vraies nouvelles des autres datasets, avec des rappels des vrais nouvelles et des "Fake News" qui se repprochent considérablement. Ainsi, cela supporte à nouveau notre hypothèse selon laquelle des caractéristiques spécifiques à la façon dont sont rédigés les articles de Reuters, en termes de mots ou de champ lexical, caractéristiques ne se retrouvant pas forcément dans les titres des articles, expliquent la très mauvaise généralisabilité de notre modèle baseline. La limitation des informations données au modèle contribue à limiter le surapprentissage.

Dataset	**ISOT (test)**	**Fake News Competition**	**Fake or Real**
Précision globale	0.94	0.59	0.55
Rappel vraies nouvelles	0.95	0.57	0.46
Rappel "Fake News"	0.94	0.62	0.64

Table 5: Résultats partiels du modèle baseline sur les titres des articles

3.2.3 Extraction de caractéristiques TF-IDF et perceptron multi-couche

Si la régression logistique permet d'atteindre des résultats particulièrement bons sur l'ensemble de test, d'autres modèles peuvent être envisagés pour générer la prédiction, qui pourraient potentiellement se généraliser de façon plus efficace à d'autres dataset. L'article de Hoy et Koulouri indique peu de différences selon le modèle utilisé, mais un perceptron multi couche (MLP) a néanmoins été implémenté. Le choix des paramètres du modèle a été réalisé via `GridSearchCV` avec une 5-folds cross-validation permettant de choisir notamment la fonction d'activation, la taille de la couche cachée et le niveau de réduction de dimension à apporter aux données en entré de MLP. Le modèle ainsi obtenu a des résultats très proches sur l'ensemble de test, avec un précision globale identique (même nombre d'erreurs) mais ne permet pas de gagner en précision sur les autres datasets. Surtout, le modèle présente le même biais de sous-reconnaissance des vraies nouvelles, avec un rappel des vraies nouvelles de 18 % sur *fake_news* et de 26 % sur *fake_real*. Ainsi, ce biais n'est pas propre au modèle logistique mais semble provenir des données elles-mêmes, ou *a minima* de l'extraction de caractéristiques TF-IDF.

3.3 Panachage entre les datasets ISOT et "Fake News Competition"

Afin de tenter de contourner les difficultés liées aux caractéristiques particulières des données de ISOT, deux modèles ont été entraînés, d'une part sur un dataset avec les vraies articles de *fake_news* et la moitié des fausses de ISOT ("new data 1"), et sur la moitié des vrais de ISOT et les fausses de *fake_news* ("new data 2"). De même, un dernier dataset ("new data 3") a été construit. L'idée sous-jacente est que la multiplication des sources dans l'ensemble d'entraînement devrait favoriser la reconnaissance de caractéristiques communes aux "Kake News", indépendamment de leurs sources. Les résultats, visibles dans la table 6, sont surprenants puisque le meilleur modèle sur *fake_real* est celui construit en utilisant les vraies nouvelles de ISOT, avec une précision globale de 71,08 % et, surtout, un rappel des vraies nouvelles qui, bien que toujours inférieur à 50 %, est 15 points supérieur à celui du modèle construit à partir des "Fake News" de ISOT. Ainsi, il semble que l'incapacité de notre modèle à reconnaître les vraies nouvelles soit bien plus une conséquence des caractéristiques spécifiques des "Fake News" de ISOT que des vraies nouvelles, contrairement à ce que les analyses précédentes semblaient suggérer. De fait, le dataset "new data 3", qui mélange à parité vraies et fausses nouvelles de ISOT et *fake_news* ne permet pas de gagner en généralisabilité sur *fake_real*

par rapport à "new data 2". L'augmentation par rapport à notre modèle baseline est néanmoins conséquente, justifiant l'intérêt de multiplier les sources de donnés.

Il apparaît donc au terme de cette discussion que ce sont bien les caractéristiques du dataset ISOT qui limitent la possibilité de généraliser un modèle de détection de "Fake News" à partir de ses données, dès lors qu'elles sont appréhendées par une vectorisation TF-IDF. De fait, en ne prenant pas en compte les informations contextuelles et les associations de mots au sein des phrases, il est probable qu'une telle approche soit particulièrement sensible au champ lexical et aux thématiques abordées dans les datasets, dépendant fortement des conditions de collecte, en laissant de côté des éléments liés à des associations spécifiques d'idées, donc de mots, voir une forme de style, qui seraient propres au style des "Fake News". De ce point de vue, il est donc nécessaire de se pencher vers des outils capables d'intégrer ces informations contextuelles.

	Test			**Fake or Real**		
Dataset	**Acc**	**Rec True**	**Rec Fake**	**Acc**	**Rec True**	**Rec Fake**
New Data 1	0.94	0.93	0.95	0.53	0.29	0.77
New Data 2	0.98	0.99	0.98	0.71	0.44	0.98
New Data 3	0.93	0.93	0.93	0.70	0.46	0.95

Table 6: Résultats partiels, modèles entraînés sur un panachage de données

4 Utilisation de BERT

Le modèle BERT (*Bidirectional Encoder Representations from Transformers*) est un modèle pré-entraîné disponible librement proposé en 2019 qui vise à produire des encodages contextuels des token (qui repose sur la technique WordPiece). Celui-ci repose sur une architecture de transformers bi-directionnel, c'est à dire que contrairement aux modèles utilisés pour la génération de textes (comme GPT) celui-ci a accès lors de sa phase d'entraînement et de codage au contexte antérieur mais aussi postérieur au mot (ou au token) considéré. De ce fait, s'il ne peut conduire à de la génération de texte celui-ci peut produire un encodage ("embedding") dépendant de *l'ensemble* du contexte (Devlin et al. [2019]).

Il existe pour notre problème deux façons d'utiliser un tel modèle. La première, qui correspond à la façon dont il est utilisé par Hoy and Koulouri [2022],consiste à utiliser le modèle pré-entraîné pour effectuer une vectorisation des articles sans modifier cet embedding, puis à entraîner des prédicteurs sur ces embeddings.[13] Si cela permet donc de gagner une information contextuelle, la façon de calculer ces embeddings n'est pas pas optimisée pour la tâche spécifique. Une autre façon de l'utiliser consiste donc à effectuer une étape de *fine-tuning* (réglage fin) afin de l'adapter à notre objectif précis. Les deux approches ont été envisagées pour ce travail.

4.1 Extraction de caractéristique avec base-BERT

L'utilisation de l'encoder de base de BERT conduit à une méthode très proche de ce qui a été évoqué dans la partie 3 avec la vectorisation TF-IDF, en ce sens que les vecteurs obtenus peuvent être traités de façon totalement identique. Là encore, le choix s'est donc porté sur une régression logistique comme premier prédicteur. Comme le montre la table7 le modèle se révèle bien plus efficace en généralisation que nos modèles de la partie précédente, avec un gain de précision globale de 7 points sur *fake_real* et de 19 points sur *fake_news*. Ce résultat était attendu, indiquant que la prise en compte du contexte permet au modèle de discriminer entre des utilisations différentes de même mots.[14] En particulier, le biais en faveur de la détection de "Fake News" se réduit considérablement, disparaissant presque totalement sur *fake_news*.

Plusieurs autres spécifications ont été explorées, par analogie avec ce qui a été présenté sur la vectorisation TF-IDF, mais ceux-ci ont abouti à des résultats globalement décevants. Ainsi, l'entraînement

[13] Nous avons ici utilisé plus spécifiquement `bert-base-uncased` via la librairie `Trasformers` de HuggingFace

[14] Ce résultat est cependant différent de ce qui était attendu au regard des résultats de Hoy et Koulouri, qui ne parvenaient pas à mettre en évidence un gain significatif via l'utilisation de BERT. La raison de cette différence n'a pas pu être mise en évidence. Une hypothèse pourrait être que le modèle BERT a été amélioré depuis la publication de leur article.

Dataset	ISOT (test)	Fake News Competition	Fake or Real
Précision globale	0.99	0.75	0.66
Rappel vraies nouvelles	0.99	0.74	0.59
Rappel "Fake News"	0.99	0.75	0.72

Table 7: Résultats partiels du modèle logistique sur embeddings (base) BERT

sur les mélanges de datasets (New Data 1, 2 et 3) aboutit à des résultats sur *fake_real* dégradés dans les cas de New Data 1 et 2 (même si dans une moindre mesure pour ce dernier, sur lequel nous obtenions les meilleures résultats dans la partie précédente), mais légèrement meilleurs pour New Data 3 (hausse de la précision globale de 7 points), ce qui indique à la fois que le mélange de sources améliore la généralisabilité du prédicteur mais aussi que cela suppose de diversifier les exemples pour les deux labels. Par ailleurs, le remplacement de la régression logistique par un perceptron multi-couche optimisé abouti à une précision globale sur l'ensemble de test de ISOT légèrement plus faible, de 98,57 % (contre 99,29 % pour la régression logistique), et des résultats très proches sur nos deux autres datasets, mais qui témoigne du retour d'un biais à la sous-détection des vraies nouvelles, même s'il est de moindre ampleur que dans la partie 3 (cf table 8). Cela s'explique sans doute par le fait que le MLP, qui implique une phase de fine-tuning, s'avère plus susceptible d'apprendre des caractéristiques particulières des données qui lui sont présentées, favorisant la forme particulière de suraprentissage précédemment évoquée.

Dataset	ISOT (test)	Fake News Competition	Fake or Real
Précision globale	0.99	0.73	0.66
Rappel vraies nouvelles	0.99	0.65	0.55
Rappel "Fake News"	0.98	0.80	0.77

Table 8: Résultats partiels du MLP sur embeddings (base) BERT

4.2 Fine-tuning d'un modèle basé sur BERT

Il existe une façon différente, potentiellement plus efficace, d'utiliser le modèle BERT pré-entrainé, qui consiste à ajouter, aux 12 couches de type transformers du modèle de base BERT (comme nous l'avons utilisé dans le 4.1), une dernière couche dense, de deux neurones, afin de prédire l'appartenance à nos deux classes (dans notre cas). Ensuite, on utilise nos données d'entraînement (ici, le sous-échantillon d'entraînement de ISOT) pour poursuivre l'entraînement du modèle afin de le spécialiser sur notre problème. Ainsi, en plus d'optimiser les poids de la dernière couche dense que nous rajoutons, nous modifions aussi l'ensemble des poids des 12 couches du modèle BERT de base. Cette démarche permet ainsi de réaliser dans une même action à la fois l'embedding et la prédiction, les deux tâches étant réalisées simultanément via le réseau de neurones.

Cette approche permet d'atteindre une précision absolument considérable sur le sous-échantillon de test de ISOT, comme le montre la matrice de confusion de la table 4.2, puis que nous n'avons plus que 5 erreurs de classifications (sur 8 980 observations). Malheureusement, le modèle ainsi construit est celui dont les performances en généralisabilité sont les plus mauvaises parmi ceux que nous avons vu, avec une précision globale de 61,29 % sur *fake_news* et de seulement 40,25 %sur *fake_real*, bien inférieure à l'espérance d'un prédicteur aléatoire. De plus, comme le montrent les tables 4.2 et 4.2 les prévisions semblent, à nouveau, particulièrement biaisées, même si cette fois-ci le modèle semble avoir énormément de mal à reconnaître les "Fake News", avec un rappel de seulement 23 % et 17 % respectivement.

Ainsi, l'entraînement spécifique d'un modèle BERT pour la détection de "Fake News" conduit à des résultats détériorés, et particulièrement biaisés vers la non détection de "Fake News", par rapport à

		Prédiction	
		Vraie	**Fausse**
Label	**Vraie**	4284	0
	Fausse	5	4691

Table 9: Performance du modèle avec fine-tuning de BERT sur l'ensemble de test

		Prédiction	
		Vraie	**Fausse**
Label	**Vraie**	10337	50
	Fausse	8002	2411

Table 10: Performance du modèle avec fine-tuning de BERT sur "Fake News Competition"

		Prédiction	
		Vraie	**Fausse**
Label	**Vraie**	2019	1152
	Fausse	2633	531

Table 11: Performance du modèle avec fine-tuning de BERT sur "Fake or Real"

		Prédiction	
		Vraie	**Fausse**
Label	**Vraie**	4266	18
	Fausse	100	4596

Table 12: Performance du modèle sur titre avec fine-tuning de BERT sur l'ensemble de test

une simple extraction de features via le modèle de base de BERT. Ceci s'explique probablement par le fait que cet entraînement modifie de très nombreux poids, augmentant considérablement le risque que le modèle apprenne parfaitement des caractéristiques très spécifiques aux données du dataset ISOT, mais non transposables à d'autre vraies informations, ou d'autres "Fake News". Cette interprétation est confortée par le dernier modèle ayant été testé, à savoir un fine-tuning du modèle BERT sur les titres des articles uniquement (voir matrices de confusions tables 4.2, 4.2 et 4.2). Celui-ci n'atteint certes qu'une précision globale de 63 % environ sur nos datasets de généralisation, mais il ne présente pas de biais particulier de non détection ni de vraies nouvelles ni de "Fake News", ce qui lui permet d'être plus efficace que le modèle précédent sur l'ensemble des articles : moins d'information est donnée au modèle afin d'apprendre, et il n'apprends donc pas les caractéristiques particulières des articles du dataset ISOT. Le surapprentissage de ces caractéristiques est donc bien plus limité.

		Prédiction	
		Vraie	**Fausse**
Label	**Vraie**	6607	3780
	Fausse	3989	6424

Table 13: Performance du modèle sur titres avec fine-tuning de BERT sur "Fake News Competition"

		Prédiction	
		Vraie	**Fausse**
Label	**Vraie**	1928	1243
	Fausse	1074	2090

Table 14: Performance du modèle sur titre avec fine-tuning de BERT sur "Fake or Real"

5 Discussion et conclusion

Le travail présenté ici constitue une exploration sur la possibilité de construire un prédicteur de "Fake News" généralisable à partir des données de ISOT. Si aucune des stratégies développées ici n'a permis d'atteindre cet objectif de façon satisfaisante, le meilleur modèle parmi ceux présentés est l'implémentation d'une régression logistique sur un embedding reposant sur le modèle BERT de base, permettant d'atteindre une précision de l'ordre de 70 % en généralisation (sur nos datasets de généralisation).

Au delà de ce résultat, il a surtout permit de relevé plusieurs éléments pouvant expliquer la difficulté de la tâche, éléments qui vont dans le sens de ceux mis en avant par Hoy and Koulouri [2022] dans la conclusion de leur article. Premièrement, le choix du modèle de prédicteur mis en place semble avoir un impact très secondaire par rapport à la technique d'Embedding utilisée. L'étape cruciale consiste donc avant tout à trouver une façon pertinente de résumer l'information contenue dans nos textes, via une extraction de caractéristique, de préférence en mesure de tenir compte du contexte. De plus, la limite principale à la généralisation semble se trouver du côté de la façon dont sont construits les datasets, et, dans le cas qui nous intéresse, ISOT, qui malgré leurs tailles regroupent des exemples partageant des caractéristiques très spécifiques, difficilement généralisables. Ce fait peut donc aboutir à une forme particulière de suraprentissage, qui se manifeste par des prédicteurs biaisés, incapable de reconnaître tantôt les "Fake News", tantôt les vraies nouvelles, dès lors qu'elles proviennent de sources différentes de celles de leur ensemble d'entraînement. En ce sens, limiter l'information disponible, en se limitant par exemple aux titres, ne détériore pas les performances en généralisabilité, voir les augmente.

La prise en compte d'informations contextuelles permise par le modèle BERT permet néanmoins de limiter ce biais, mais le fine-tuning du modèle pour cette tâche spécifique s'avère de son côté contre-productif, poussant à l'extrême la sur-spécialisation du modèle sur les données similaires à son ensemble d'entraînement. A l'inverse, le modèle BERT de base, produisant des embeddings contextuels, mais construit sur un corpus bien plus large que nos datasets, se montre plus à même de rendre compte de la diversité des formes de "Fake News", évitant que le processus d'embeddings n'élimine dès ce stade ce qui n'est pas propre à l'ensemble d'entraînement. Ainsi, au terme de cette discussion, il semble que la difficulté à construire un prédicteur de "Fake News" généralisable puisse s'expliquer avant tout par les difficultés de définition de la notion évoquées en introduction. Celles-ci peuvent en effet expliquer pourquoi les exemples contenues dans les datasets peuvent différer fortement, au delà même des choix pratiques réalisés pour leurs collectes. Or, ce sont bien ces différences ne permettent pas l'application d'un prédicteur construit sur un dataset à un autre dataset. Néanmoins, l'amélioration des performances dans ce domaine passe très probablement par la constitution de nouvelles bases de données, labélisées, permettant un entraînement de ces modèles. L'enjeu autours de la construction de ces bases de données ne consiste pas nécessairement à augmenter leur taille (en témoigne les performances quasi parfaites obtenues sur les échantillons de test de ISOT) mais bien en une diversification de leurs provenances, de leurs types, de leurs sources.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*, page 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL http://aclweb.org/anthology/N19-1423.

Julien Giry. Les fake news comme concept de sciences sociales:essai de cadrage à partir de notions connexes: rumeurs, théories du complot, propagande et désinformation. *Questions de communication*, 38(2):371–394, 2020. doi: 10.4000/questionsdecommunication.24263. URL https://shs.cairn.info/revue-questions-de-communication-2020-2-page-371.

Nathaniel Hoy and Theodora Koulouri. Exploring the generalisability of fake news detection models. In *2022 IEEE International Conference on Big Data (Big Data)*, page 5731–5740, Osaka, Japan, December 2022. IEEE. ISBN 9781665480451. doi: 10.1109/BigData55660.2022.10020583. URL https://ieeexplore.ieee.org/document/10020583/.

OCDE. *Les faits sans le faux : Lutter contre la désinformation, renforcer l'intégrité de l'information*. OECD, 2024. ISBN 9789264445123 9789264520431 9789264349889. doi: 10.1787/4078bb32-fr. URL https://www.oecd.org/fr/publications/les-faits-sans-le-faux-lutter-contre-la-desinformation-renforcer-l-integrite-de-l-information_4078bb32-fr.html.