Abstract

Face à l'instabilité croissante causée par la propagation massive de fausses informations aussi communément appelées *Fake News*. Nous souhaitons proposer une méthode capable de détecter ces fake news et les différencier des informations légitimes. En particulier, nous nous intéressons aux "fake news" dont la thématique est d'ordre politique. En effet, de nombreux articles sur le web prennent la forme d'articles de presse à première vue fiable et peuvent alors influencer à tort l'opinion publique et avoir des impacts considérables dans des moments politiques décisifs comme lors d'élections présidentielle ou législative.

Une littérature relativement abondante et mature ayant déjà été développée sur le sujet, nous reprendrons certaines de ces méthodes sur nos propres jeux de données et nous testerons également leur généralisabilité en les confrontant à autre jeu de données que celui vu en entraînement.

1 Introduction

De l'origine du Covid, aux élections américaines jusqu'aux conflits en Ukraine et au Proche-Orient, ce sont des milliers voire des millions d'informations en tout genre (tweet, commentaires, blogs, article de presse...) qui circulent en temps réel. Il peut être difficile de démêler le vrai du faux surtout lorsque des conflits d'intérêts entre en jeu ou que certains cherchent délibérement à détourner la vérité. Il devient alors crucial de pouvoir détecter ces *fake news* en particulier dans des domaines sensibles qui touche par exemple à la politique des États. De nombreuses approches combinant des techniques de Natural Language Processing (NLP) et Machine Learning (ML) ont pu être développées et ont su prouver leur efficacité pour détecter ces "fake news". En nous appuyant sur cette littérature, nous commencerons par faire un bref état de l'art des techniques existantes [1], puis nous présenterons et analyserons nos jeux de données [2]. Finalement après avoir justifié le choix des méthodes de NLP et de Machine Learning retenues, nous commenterons nos résultats et leur généralisabilité. [3]

2 État de l'art

Les Fake News étant destinées à être lus, diffusées et compris par des humains sont construits à partir du langage. Ainsi ce sont surtout des modèles basés sur de la feature engineering à partir de techniques de NLP qui sont employés pour réaliser la tache de detection de fake news. On peut distinguer principalement deux types de feature en NLP selon l'approche retenue :

- Représentation à l'échelle des mots
- Représentation de marqueurs linguistiques

Plus précisément, la représentation à l'échelle des mots comprend entre autres les techniques suivantes :

- Bag of Words (BoW) qui permet de créer un vecteur de comptage des mots présents. Cette technique ne prend néanmoins pas en compte ni l'ordre des mots ni leur polysémie.
- TF-IDF, sensiblement similaire à (BoW) mais qui ajoute une pondération aux mots rares pour leur donner plus de sens et d'importance face aux mots fréquents mais non significatifs comme "le", la", "et" etc.
- Word2Vec permet d'encoder les mots dans un espace vectoriel de telle sorte que les mots proches en sens soient également proches en distance dans cet espace. C'est cette propriété d'analogies vectorielles qui est notamment à l'origine d'un fait intéressant : la relation vectorielle entre oncle et tante est souvent proche de celle par exemple entre homme et femme . Néanmoins Word2Vec ne permet pas de représenter la polysémie des mots et par exemple donnera le même encoding à "orange" quel que soit le sens : le fruit ou la couleur.
- BERT repose sur un principe similaire à Word2Vec, à savoir la représentation vectorielle des mots. Néanmoins, à la différence de Word2Vec, BERT intègre aussi la polysémie des mots et permet de prendre en compte un contexte élargi plus riche en incluant l'ordre et le positionnement des mots dans la phrase.

D'autre part, la représentation basée sur des marqueurs linguistiques se concentre sur la création de statistiques pertinentes, à l'échelle du texte dans sa globalité. Ces indicateurs souvent définies par des travaux linguistiques cherchent notamment à extraire les inforamtions suivantes, moyenne de la longueur des mots, la fréquence des pronoms personnels et/ou impersonnels, la complexité du vocabulaire ou encore la fréquence de la ponctuation etc.

Enfin des techniques plus récentes et basées sur du deep learning comme les Convolutional Neural Networks (CNN) permettent d'analyser les images présentes en complément du contenu textuel dans le but d'améliorer les prédictions. Dans un cadre encore plus multimodal, on pourrait également intégrer l'analyse des vidéos présentes ou analyser l'historique de propagation de l'information sur les réseaux sociaux pour d'affiner les prédictions.

3 Analyse descriptive et statistiques des données

Sur la base de l'article de référence Exploring the Generalisability of Fake News Detection Models de Nathaniel Hoy et Theodora Koulouri (2022), nous avons repris deux bases de données largement exploitées dans la tache de détection de fake news à savoir :

- ISOT Dataset
- Kaggle Fake or Real

3.1 ISOT Dataset

Ce dataset combine 44 898 articles de thème politique équilibrément répartis en *fake news* (23 481 arcticles) et *real news* (21417 articles).

Il comporte cinq types d'informations et ne présente aucune valeurs manquantes : Nombre de valeurs manquantes par colonne :

1. title : 0
2. text : 0
3. subject : 0
4. date : 0
5. label : 0

La première étape d'analyse exploratoire des données a pour objectif de mieux visualiser et comprendre les données à disposition. Son analyse peret également d'orienter le choix des features à conserver pendant la phase de pré-processing.

Dans toute la partie graphique : 0/bleu = Fake news et 1/orange = Real News

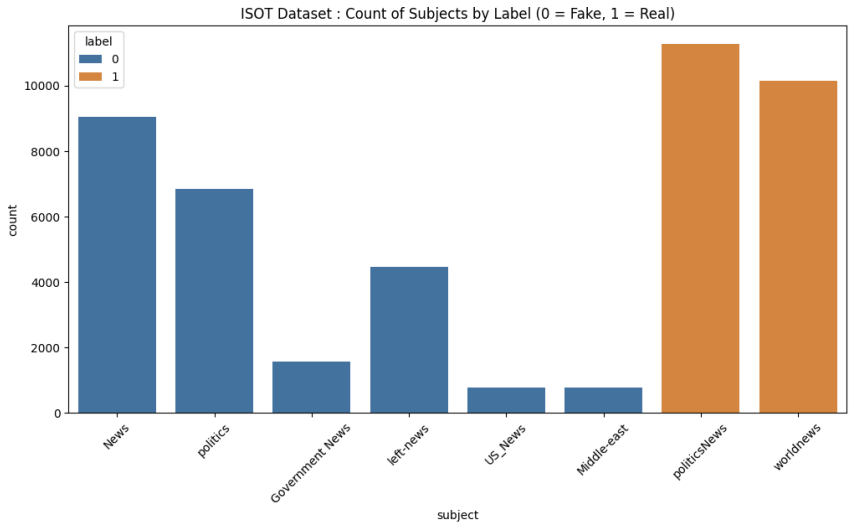

Figure 1: ISOT Dataset : thématique de l'article 0/bleu = Fake News, 1/orange = Real News

Sur la Figure 1, on peut voir comme prévu que, les articles abordent tous une thématique politique. Néanmoins une différence notable et qui sera confirmer sur l'analyse en Word Cloud ci-dessous, est que les Real News abordent beaucoup plus les faits politiques internationaux comparé aux Fake News dont la portée est davantage centrée sur la politique interne aux États-Unis.

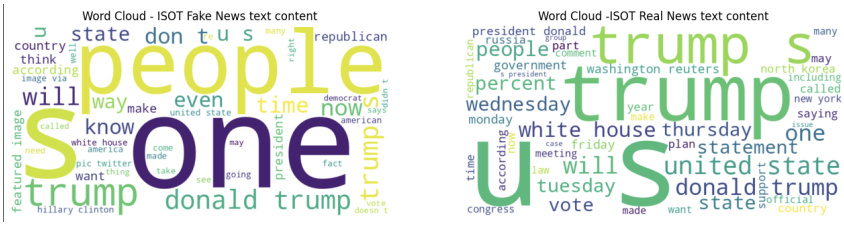

Figure 2: IDOT Dataset : Word Clouds Fake (gauche) vs Real News (droite)

La Figure 2 présentent deux Word Clouds (WC) : à gauche le WC associé au mots les plus fréquents dans les Fake News et à droite ceux présents les plus fréquemments dans les Real News. Plusieurs différences sont notables :

- Le nuage associé au Fake News est beaucoup moins dense alors qu'ils contiennent le même nombre de mots : les mots très fréquemment utilisés dans les Fake News ont tendance à être plus courts et plus simples.
- Le nuage associé au Real News présente beaucoup plus de dates et de termes servant à quantifier ou à être précis; par exemple on peut voir "percent", "time", "plan", "wednesday", "monday"
- Le nuage associé au Real News présente également le nom du journal sur lequel il est publié à savoir Washington Reuters alors que la source est souvent indéfinie sur les Fake News.

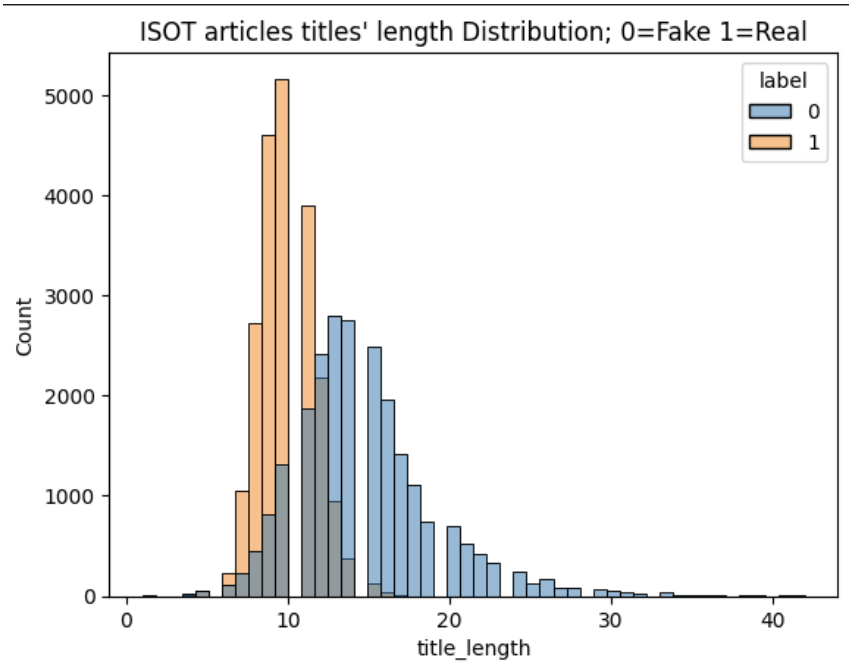

Figure 3: ISOT Dataset : Distribution de la longueur du titre des articles

Sur la Figure 3 nous pouvons voir que la longueur du titre des Fake News est souvent plus longues que celles des Real News

3.2 Kaggle Fake or Real

Le dataset Kaggle Fake or Real combine 6335 articles de thème politique répartis en fake news (3164 articles) et real news (3171 articles).

Il comporte trois types d'informations et ne présente aucune valeurs manquantes : Nombre de valeurs manquantes par colonne :

- 1. title : 0
- 2. text : 0
- 3. label : 0

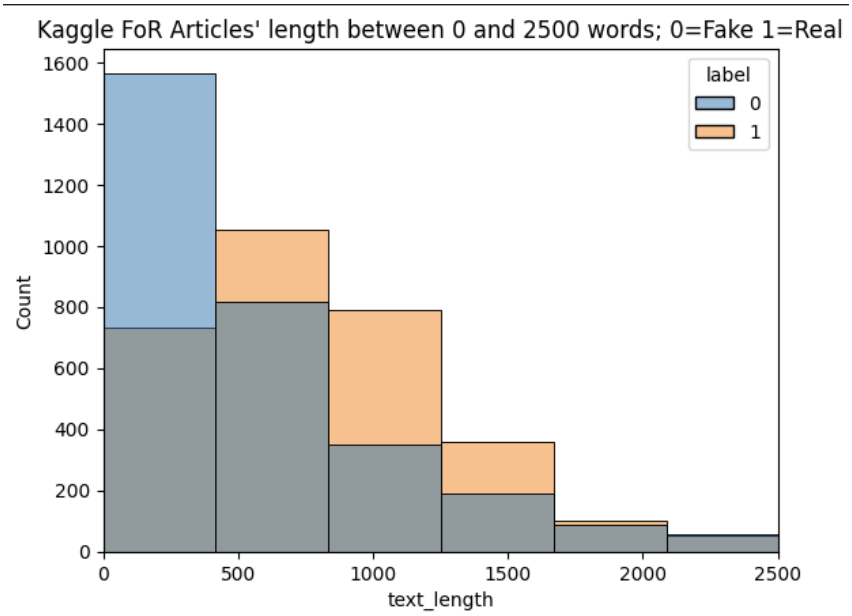

Figure 4: Kaggle Dataset : Distribution de la longueur des articles

Sur la Figure 4, on peut voir que la longueur des faux articles (en bleu) est souvent plus petites que celles des vraies articles dont la longueur dépasse souvent les 500 mots.

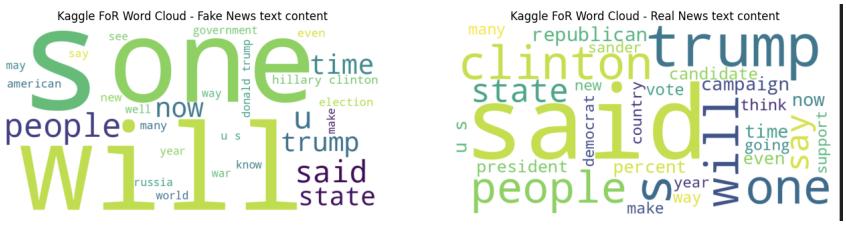

Figure 5: Kaggle Dataset : Word Cloud Fake (gauche) vs Real (droite) News

De la même manière que sur le WC des données ISOT, on peut voir sur la Figure 5 que le WC à droite associé aux Real News est plus dense et est représenté par des mots plus longs et plus variés.

4 Pré-Processing et résultats

4.1 Pré-Processing

Nous avons choisi de ré-implémenter deux features de NLP qui ont obtenu de très bons résultats dans l'article de Hoy, Nathaniel Koulouri, Theodora. (2022). Exploring the Generalisability of Fake News Detection Models et qui sont basées sur la représentation à l'échelle des mots : TF-IDT et Word2Vec.

Le Pré-Processing consiste à enlever les bruits perturbateurs au modèle, concrètement :

- - Conversion des mots en minuscule : pour éviter que le modèle différencie "Lundi" de 'lundi"
- - Retrait de la ponctuation, des liens URLs et des pseudos Twitter
- Pour la méthode basée sur le TF-IDT on retire également les espaces et les *stop words*.

4.2 Résultats - Régression Logistique

Tous les modèles ont été classés via une régression logistique et les résultats ont été évalués après k-fold cross validation avec une répartition train/test split de 80/20.

Nom du Dataset	**Feature utilisé**	**Précision**	**Recall**	**Interprétation précision / recall**
ISOT	TF-IDF	0,99	0,99	Excellents score de précision et rappel : 99% sont correctes et parmi les vrais "0", 99% ont été bien détectés.
ISOT	Word2Vec	0,97	0,97	Très bonnes performances,mais légèrement moindre qu'avec TF-IDT
Kaggle FoR	Word2Vec	0,88	0,88	Performances correctes, précision et rappel équilibrés, mais plus d'erreurs.
Kaggle FoR	TF-IDF	0,92	0,91	Meilleures performances qu'avec Word2Vec.

Table 1: Résumé des performances après cross validation

4.3 Généralisabilité du modèle

On s'intéresse maintenant à la généralisabilité du modèle c'est-à-dire à sa capacité à conserver de bonnes performances de prédictions lorsqu'il est intraîné sur un dataset 1 et testé sur un dataset 2 inconnu.

La Table 2 résume les résultats obtenus et montre bien que de manière générale les modèles ont du mal à se généraliser.

Dans le cadre de cette expérience, nous n'avons pas utiliser de train/test split mais avons entrainé sur l'ensemble du dataset 1 et testé sur l'ensemble du dataset 2.

Data Train	Data test	Feature	Précision	Recall	Interprétation précision / recall
ISOT	Kaggle FoR	TF-IDF	0,70	0,64	Forte baisse par rapport à la Table 1 (99%)
Kaggle FoR	ISOT	TF-IDF	0,69	0,69	Moyennes correctes avec une meilleure généralisation que dans l'autre sens, mais en baisse par rapport à la Table 1 également
Kaggle FoR	ISOT	Word2Vec	0,76	0,75	Amélioration par rapport à TF-IDF ; précision meilleure sur classe 0 (0,81), rappel meilleur sur classe 1 (0,82).
ISOT	Kaggle FoR	Word2Vec	0,67	0,64	Performances assez faibles ; précision en classe 1 correcte (0,75) mais rappel très faible (0,42) ce qui traduit un problème de généralisation inter-dataset.

Table 2: Résumé des performances de la généralisation des modèles

5 Conclusion

L'objectif était de concevoir un modèle capable de détecter les faux articles (Fake News) des vrais. En se basant sur les travaux menés par Hoy, Nathaniel Koulouri, Theodora, Exploring the Generalisability of Fake News Detection Models. Nous avons pu obtenir sur deux jeux de données d'articles datant de 2015 à nos jours des résultats très satisfaisant : jusqu'à 99 pourcent d'accuracy sur le test set du dataset ISOT après cross validation en combinant la méthode du TF-IDT à une régression logistique. Néanmoins, sur la question de la généralisabilité des modèles, c'est-à-dire leur capacité à garder une excellente performance sur des données extérieures au jeu de données initiale d'entrainement. Tous les modèles connaissent une perte d'efficacité. Il serait donc intéressant de savoir si l'entrainement sur une base de données plus importante et variée (thèmes abordées, années de publication...) pourrait améliorer ces résultats.

References

[1] Hoy, Nathaniel Koulouri, Theodora. (2022). Exploring the Generalisability of Fake News Detection Models. 5731-5740. 10.1109/BigData55660.2022.10020583.