

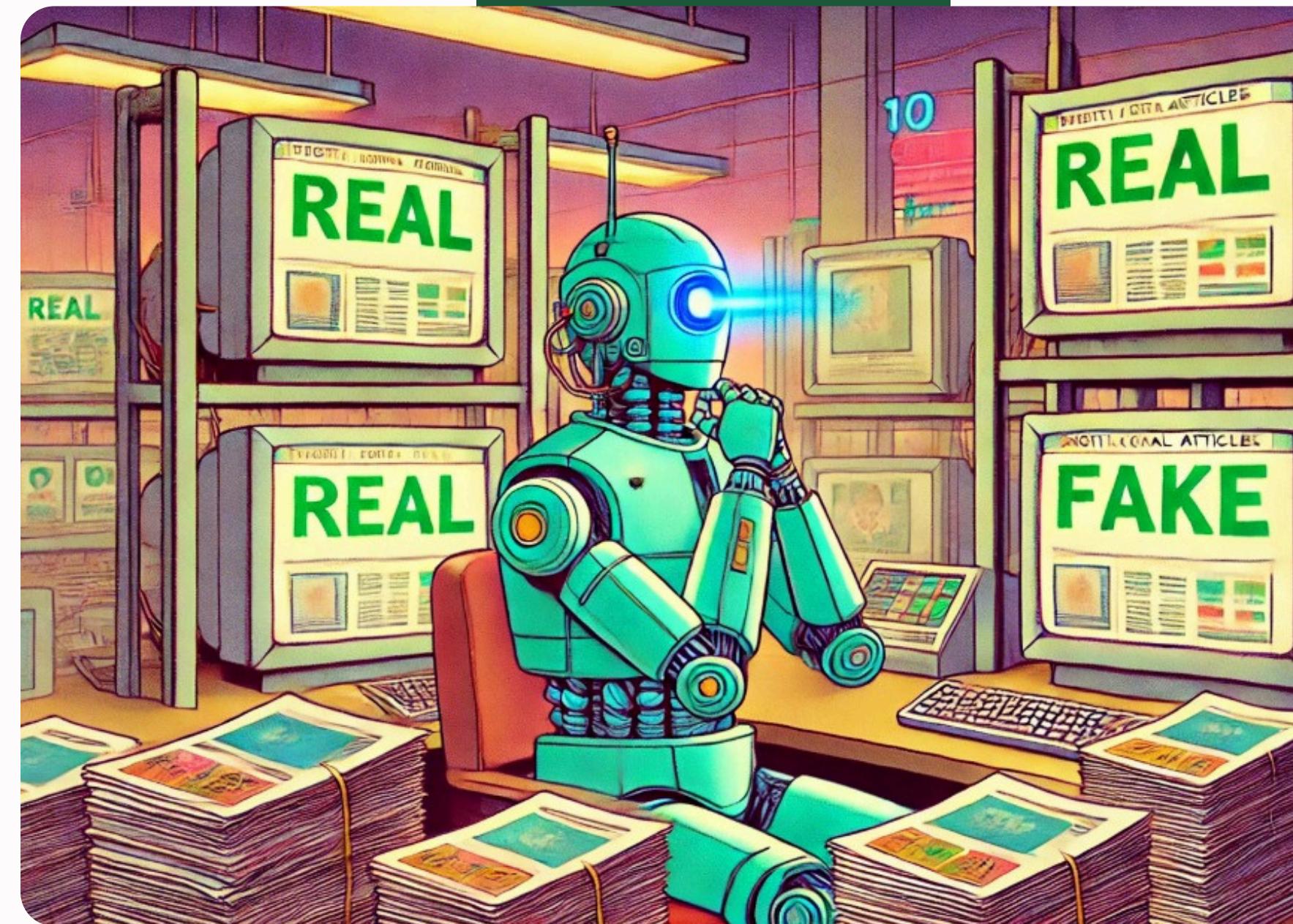
Fake news Detection

Présenté par :
Sanor Camara
Mamadou Saïdou Diallo
Papa Boubacar Diop
Ibrahima Fa Lo

Analystes statisticiens

Sous la supervision de :
Mme Mously Diaw

Senior Data Scientist /
ML Engineer



Plan

- 01  Introduction
- 02  Présentation
des données
- 03  Pre-processing
- 04  Modélisation
- 05  Conclusion

The background image shows a modern office space with a high ceiling featuring exposed pipes and ductwork. Large green plants are integrated throughout the room, hanging from the ceiling and growing in planters on the walls and desks. There are several wooden desks with black office chairs, and a large sofa area in the background. The overall atmosphere is bright and natural.

INTRODUCTION

Introduction

À l'ère de l'information instantanée

Nous vivons dans une époque où l'information circule à une vitesse fulgurante à travers internet, les réseaux sociaux et les messageries instantanées. Cette instantanéité facilite l'accès au savoir, mais augmente aussi les risques d'exposition à des contenus non vérifiés, voire mensongers.

Prolifération des fausses nouvelles

Ces contenus trompeurs ou Fake News sont souvent créés intentionnellement pour manipuler l'opinion publique, influencer des élections, ou simplement générer du clic. Leur impact peut être considérable, allant de la désinformation en période de crise à l'instabilité sociale et politique. Détecter et filtrer ces informations erronées est donc devenu un enjeu prioritaire pour les plateformes, les gouvernements, et les citoyens.

IA

Face à l'ampleur du phénomène, les solutions traditionnelles de vérification humaine ne suffisent plus. L'intelligence artificielle, avec sa capacité à traiter dénormes volumes de données en temps réel, offre une réponse pertinente à cette problématique. En particulier, les modèles d'apprentissage automatique peuvent reconnaître des schémas linguistiques, contextuels ou comportementaux associés aux Fake News. Grâce à eux, il devient possible d'automatiser une partie de la détection, tout en gagnant en rapidité et en efficacité.

NLP (Natural Language Processing)

Au cœur de cette démarche se trouve le traitement automatique du langage naturel, ou NLP. Cette branche de l'IA permet aux machines de comprendre, d'interpréter et de générer du langage humain. Dans notre contexte, le NLP est utilisé pour analyser le contenu textuel des articles ou publications, en extrayant les informations et en identifiant les signes caractéristiques d'une fausse nouvelle grâce à des techniques de traitement de texte, et l'usage de modèles de classification.

PRÉSENTATION DES DONNÉES

Présenter le jeu de données utilisé

Présentation des données

Dataset Kaggle

- ISOT Fake News Dataset
- Deux datasets, les fake news et les vraies news.
- Les articles véridiques ont été extraits de Reuters.com (site web d'actualités)
- Ceux qui sont faux proviennent de sites jugés non fiables, identifiés par Politifact (une organisation de vérification des faits aux États-Unis) et Wikipédia.
- Couvre divers sujets, avec une majorité d'articles liés à l'actualité politique et internationale.

Structure de l'ensemble de données

- L'ensemble de données comprend deux fichiers CSV : « True.csv » avec plus de 12600 articles véridiques de Reuters.com, et « Fake.csv » avec un nombre équivalent d'articles issus de sources de fausses nouvelles. Chaque article contient un titre, un texte, une catégorie et une date de publication. Les données couvrent principalement la période de 2016 à 2017, en cohérence avec celles collectées sur Kaggle.

Présentation des données

Structure de l'ensemble de données

News	Size (Number of articles)	Subjects		
		Type	Articles size	
Real-News	21417	<i>World-News</i>	10145	
		<i>Politics-News</i>	11272	
Fake-News	23481	<i>Type</i>	Articles size	
		<i>Government-News</i>	1570	
		<i>Middle-east</i>	778	
		<i>US News</i>	783	
		<i>left-news</i>	4459	
		<i>politics</i>	6841	
		<i>News</i>	9050	

Nom	Description
title	Titre de l'article
text	Le contenu de l'article
subject	Le sujet auquel l'article fait référence
date	La date à laquelle l'article a été publié
isFake	La variable cible qui dit si l'article est vérifié

PRE-PROCESSING

Méthodologie adoptée pour le pré-traitement des données



Nettoyage et préparation des données

Traitement des dates

- Extraction des jours, mois et années à partir de différents formats.
- vérification des dates mal formatées ou des chaînes non interprétables.

Suppression des stopwords

Les mots courants sans valeur significative (ex. : "le", "et", "de") ont été supprimés pour ne conserver que les mots importants

Suppression des caractères inutiles

Les caractères spéciaux, la ponctuation, et les espaces superflus ont été retirés pour uniformiser le texte.

Conversion en minuscules et Tokenisation

Toutes les lettres ont été transformées en minuscules pour éviter que la casse ne crée des doublons et puis le texte a été divisé en mots distincts.

Transformations finales

- Vectorisation avec le TF-IDF
- One-Hot Encoding pour la variable catégorielle source
- Standardisation de la matrice finale

MODÉLISATION

Présentation du modèle utilisé

CONCLUSION



Conclusion

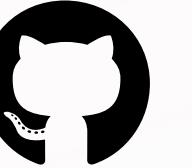
Inquiétude quant aux Fake News
surtout par rapport à la société

Les différents modèles affichaient
des résultats presque parfaits (1
partout)

Recommandation

Essayer le modèle sur un autre jeu de
données d'entraînement

THANK YOU



[Lien vers dépôt github](#)