

# Rapport de mini projet

Fake News Detection
Using Machine Learning



Réalisé par :

**Encadré par :** 

Ben Hamou Mehdi

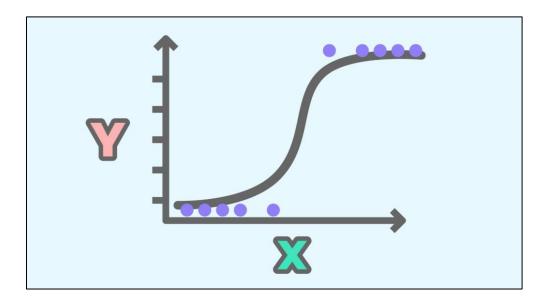
Mr. EL Aroussi EL Mehdi

# Sommaire

Cha	pitre 1 : Théorique					
	égression Logistique					
2. RI	FC (Random Forest Classifier)	5				
Cha	pitre 2 : Pratique					
1 A	nalyse exploratoire des données (AED)	7				
1.1	La recherche de l'ensemble de données	7				
1.2	Les bibliothèques utilisées	7				
1.3	Visualisation des données et Analyse de répartition	8				
1.4	Gestion des données manquantes	9				
1.5	Segmentation des données	10				
1.6	Détection des valeurs aberrantes	11				
1.7	Standardisation des données	12				
2 C	hoix de l'algorithme d'apprentissage automatique	12				
2.1	Sélection des algorithms	12				
2.2	Diviser les données	12				
2.3	Entraînement et évaluation	12				
2.4	Comparaison des performances	14				
2.5	Choix du meilleur modèle	15				
2.6	Évaluation sur l'ensemble de test	15				
3 Ir	nterface API	16				
4 Ir	ndex HTML	18				
5 J	5 Java Script19					
	ava Script	19				
	ava Script CSS					

# Chapitre 1: Théorique

# 1. Régression Logistique



Si vous vous intéressez un tant soit peu au Machine Learning et aux problèmes de classification, vous avez déjà dû avoir affaire au modèle de régression logistique. Et pour cause ! Il s'agit d'un des modèles de Machine Learning les plus simples et interprétables qui existe, prend des données à la fois continues ou discrètes, et les résultats obtenus avec sont loin d'être risibles. Mais que se cache-t'il derrière cette méthode miracle ? Et surtout comment l'utiliser sur Python ? La réponse dans cet article

Qu'est-ce que la régression logistique?

La régression logistique est un modèle statistique permettant d'étudier les relations entre un ensemble de variables qualitatives Xi et une variable qualitative Y. Il s'agit d'un modèle linéaire généralisé utilisant une fonction logistique comme fonction de lien.

Un modèle de régression logistique permet aussi de prédire la probabilité qu'un événement arrive (valeur de 1) ou non (valeur de 0) à partir de l'optimisation des coefficients de régression. Ce résultat varie toujours entre 0 et 1. Lorsque la valeur prédite est supérieure à un seuil, l'événement est susceptible de se produire, alors que lorsque cette valeur est inférieure au même seuil, il ne l'est pas.

Mathématiquement, comment ça se traduit / ça s'écrit ? Considérons une entrée  $X=x1\ x2\ x3\ ...\ xn$ , la régression logistique a pour objectif de trouver une fonction h telle que nous puissions calculer :

y= {1 si hX≥ seuil, 0 si hX< seuil}

On comprend donc qu'on attend de notre fonction h qu'elle soit une probabilité comprise entre 0 et 1, paramétrée par =1 2 3 n à optimiser, et que le seuil que nous définissons correspond à notre critère de classification, généralement il est pris comme valant 0.5.

La fonction qui remplit le mieux ces conditions est la fonction sigmoïde, définie sur R à valeurs dans [0,1]. Elle s'écrit de la manière suivante :

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

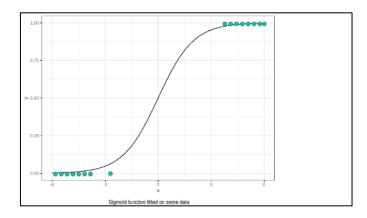
### Et notre classification dans tout ça?

La fonction h qui définit la régression logistique s'écrit alors :

$$\forall (X \in \mathbb{R}^n) \quad h(X) = \sigma(\Theta X)$$
 i.e. 
$$\forall (X \in \mathbb{R}^n) \quad h(X) = \frac{1}{1 + \varepsilon^{-\sum\limits_{i=1}^n \theta_i x_i}}$$

Tout le problème de classification par régression logistique apparaît alors comme un simple problème d'optimisation où, à partir de données, nous essayons d'obtenir le meilleur jeu de paramètre Θ permettant à notre courbe sigmoïde de coller au mieux aux données. C'est dans cette étape qu'intervient notre apprentissage automatique.

Une fois cette étape effectuée, voici un aperçu du résultat qu'on peut obtenir:



Il ne reste plus, à partir du seuil défini, qu'à classer les points en fonction de leurs positions par rapport à la régression et notre classification est faite!

#### La régression logistique en pratique

En Python c'est assez simple, on se sert de la classe LogisticRegression du module sklearn.linear\_model comme un classificateur normal et que l'on entraîne sur des données déjà nettoyées et séparées en ensembles d'entraînement et de test puis le tour est joué!

Niveau code, rien de plus basique :

```
from sklearn.linear_model import LogisticRegression

clf = LogisticRegression()

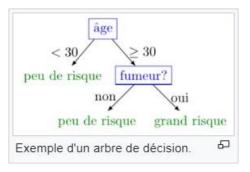
clf.fit(X_train,y_train)

y_pred = clf.predict(X_test)
```

# 2. RFC (Random Forest Classifier)

En intelligence artificielle, plus précisément en apprentissage automatique, les forêts d'arbres décisionnels1 (ou forêts aléatoires de l'anglais random forest classifier) forment une technique d'apprentissage à base d'arbres de décision. Elle apprend plusieurs arbres de décision et de ce fait, elle fait partie des méthodes d'apprentissage ensembliste, c'est-à-dire des méthodes qui utilisent la sagesse des foules2. Ils ont été premièrement proposées par Ho en 19953 et ont été formellement proposées en 2001 par Leo Breiman4 et Adele Cutler5. Cet algorithme combine les concepts de bagging (méthodes ensemblistes parallèles2) pour la phase de sélection des données, et de sous-espaces aléatoires. L'algorithme des forêts d'arbres décisionnels effectue un apprentissage sur de multiples arbres de décision entraînés sur des sous-ensembles de données légèrement différents

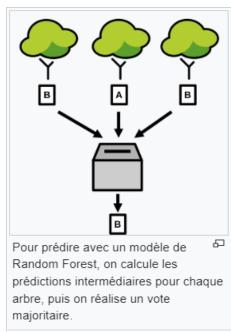
Un arbre décisionnel ou arbre de décision est une structure qui prend la forme d'un arbre, dans laquelle on pose des questions sur des attributs (dans certains ouvrages, on parle plutôt de variables2). La figure donne un exemple d'un arbre de décision. Il y a deux attributs : l'âge et le fait que la personne soit fumeuse. On pose d'abord la question de l'âge. Si on a moins de 30 ans, il y a peu de risque. Si on a plus de 30 ans, on pose la question si la personne est fumeuse. Si non, peu de risque. Si oui, le risque est grand.



Le modèle est constitué de B arbres de décision. Pour prédire la classe d'une observation (par exemple, prédire si une personne fumeuse de 35 ans a un grand risque ou non d'avoir un accident cardiovasculaire) :

On calcule le résultat pour chaque arbre. Si par exemple, on dispose de 3 arbres, on peut imaginer obtenir "peu de risque", "grand risque" et "grand risque".

La prédiction finale de la forêt aléatoire est un simple vote majoritaire (Ensemble learning). Dans l'exemple, il y a deux fois "grand risque" mais qu'un seul "peu de risque". La prédiction finale est donc "grand risque".



# Chapitre 2: Pratique

# 1 Analyse exploratoire des données (AED)

#### 1.1 La recherche de l'ensemble de données

```
train.csv: A full training dataset with the following attributes:

id: unique id for a news article

title: the title of a news article

author: author of the news article

text: the text of the article; could be incomplete

label: a label that marks the article as potentially unreliable

1: unreliable

0: reliable
```

Nous avons trouvé cette dataset sur kaggle elle contient 5 colonnes l'identifiant le titre de la news l'auteur le texte et le label (0 pour les vraies news et 1 pour les fausses) et 20800 lignes .

### 1.2 Les bibliothèques utilisées

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy score, confusion matrix, classification report, roc curve, auc
import re
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics import accuracy_score
from sklearn.model selection import train test split
from sklearn.model selection import cross val score
import nltk
nltk.download("stopwords")
```

Nous avons importé les bibliothèques nécessaires (numpy, pandas, sklearn, matplotlib) et d'autres comme :

"re" pour utiliser les expressions régulières.

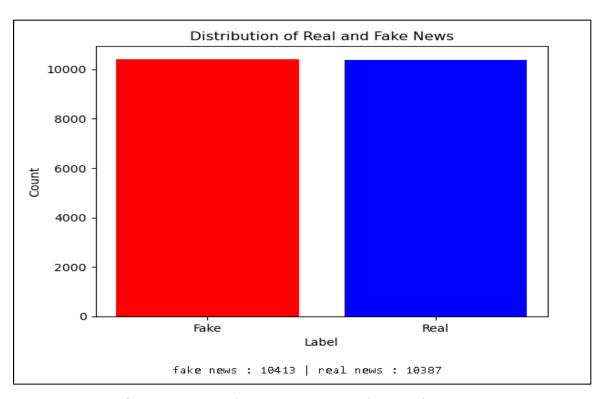
nltk.corpus.stopwords contient des mots qui n'affectent pas notre modèle lors de l'entrainement

Nous avons également Porterstemmer qui ramène le mot à sa forme racine.

tfidfvectorizer attribue une valeur numérique spécifique à chaque mot en fonction de son importance.

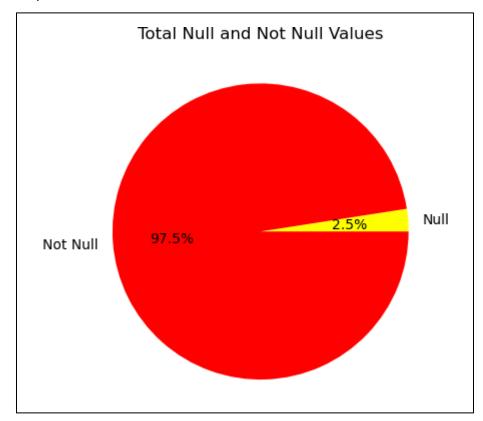
cross\_val\_score est utilisé pour évaluer le résultat de l'algorithme, et nous l'utiliserons pour déterminer quel algorithme est le meilleur pour notre ensemble de données.

### 1.3 Visualisation des données et Analyse de répartition



Nous avons compté le nombre de fausses nouvelles (label=1) et le nombre de vraies nouvelles (label=0) dans notre ensemble de données et attribué les valeurs aux variables count\_fake et count\_real afin de tracer une barre qui facilite la visualisation des données. nous pouvons clairement voir que nous avons presque la même quantité

de fausses (10413) et de vraies nouvelles (10387), donc notre dataset est équilibré, cela sera utile pour entraîner notre modèle.



Nous avons compté le nombre de valeurs nulles et non nulles pour chaque colonne et tracé un diagramme circulaire.

Nous pouvons voir qu'il existe des valeurs nulles 2.5%.

### 1.4 Gestion des données manquantes

```
In [9]: print("missing values :")
print(data.isna().sum())

missing values :
id 0
title 558
author 1957
text 39
label 0
dtype: int64
```

```
In [10]: data=data.fillna("")
In [11]: print("missing values after handling :")
    print(data.isna().sum())
    missing values after handling :
    id     0
    title     0
    author     0
    text     0
    label     0
    dtype: int64
```

Notre dataset contient près de 21 000 lignes, donc même supprimer les lignes qui ont des valeurs nulles n'affecterait pas notre modèle lors de l'entraînement, dans notre cas nous avons choisi d'attribuer une chaîne vide aux valeurs manquantes afin d'utiliser tout la dataset.

### 1.5 Segmentation des données

id	title	author	text	label	content
0	House Dem Aide: We Didn't Even See Comey's Let	Darrell Lucus	House Dem Aide: We Didn't Even See Comey's Let	1	Darrell Lucus House Dem Aide: We Didn't Even S
1	FLYNN: Hillary Clinton, Big Woman on Campus	Daniel J. Flynn	Ever get the feeling your life circles the rou	0	Daniel J. Flynn FLYNN: Hillary Clinton, Big Wo
2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29,	1	Consortiumnews.com Why the Truth Might Get You
3	15 Civilians Killed In Single US Airstrike Hav	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr	1	Jessica Purkiss 15 Civilians Killed In Single
4	Iranian woman jailed for fictional unpublished	Howard Portnoy	Print \nAn Iranian woman has been sentenced to	1	Howard Portnoy Iranian woman jailed for fictio

En raison de l'énorme quantité de mots que le texte peut contenir, nous ne pouvons pas utiliser cette colonne pour le traitement, cela prendra beaucoup de temps et de ressources. nous avons donc décidé de combiner à la fois la colonne auteur et le titre et nous l'avons nommé content.

```
In [16]: x= data["content"].values
y= data["label"].values
```

Puis nous avons attribué la variable cible à y et la variable prédictive à x.

#### 1.6 Détection des valeurs aberrantes

Nous avons créé une fonction qui traite vos données texte afin d'en obtenir uniquement les données essentielles

```
def stemming(content):
    stemmed_content = re.sub("[^a-zA-Z]"," ",content)
    stemmed_content = stemmed_content.lower()
    stemmed_content = stemmed_content.split()
    stemmed_content = [port_stem.stem(word) for word in stemmed_content if not word in stopwords.words("English")]
    stemmed_content = " ".join(stemmed_content)
    ineturn stemmed_content

    Nous ne gardons que les caractères, tous les caractères spéciaux et les chiffres
    seront supprimés
```

convertir le texte en minuscule

Diviser le texte en mots en utilisant le caractère espace comme séparateur

Ici nous supprimons le suffixe de chaque mot et ajoutons le mot s'il ne s'agit pas d'un stopword dans un tableau.

Enfin nous joignons les mots que contient le tableau et ramenons le séparateur

```
data["content"] = data["content"].apply(stemming)
x=data["content"].values
```

En utilisant la fonction apply, nous appelons notre fonction stemming pour chaque valeur de la colonne content.

#### 1.7 Standardisation des données

```
In [40]: vectorizer = TfidfVectorizer()
    vectorizer.fit(x)
    X = vectorizer.transform(x)
```

Nous avons utilisé le Tfidf\_vectorizer pour convertir nos données texte en valeurs numériques. il attribue une valeur numérique à chaque mot en fonction du nombre d'apparitions de ce mot. La valeur la plus élevée signifie que le mot est important.

# 2 Choix de l'algorithme d'apprentissage automatique

### 2.1 Sélection des algorithms

Nous avons choisi d'utiliser deux algorithmes pouvant être utilisés en classification la régression logistique et la forêt aléatoire.

#### 2.2 Diviser les données

```
In [42]: x_train,x_test,y_train,y_test = train_test_split(X,y,stratify=y,random_state=2)
```

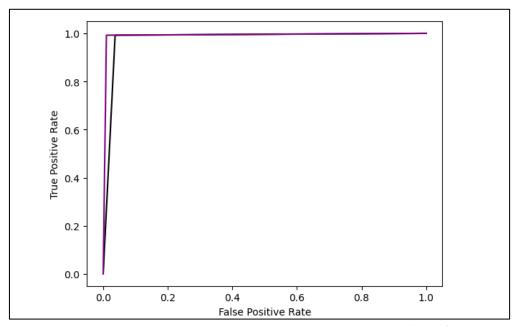
nous avons divisé nos données à l'aide de la fonction train\_test\_split et nous avons attribué label à stratify, ce qui garantit essentiellement que la distribution de label dans les données d'entraînement est similaire distribution de données de test.

### 2.3 Entraînement et évaluation

L'entrainement de model de régression logistique

L'entrainement de model de foret aléatoire

Tracer la courbe roc pour les deux algorithmes



Forêt aléatoire

Régression Logistique

Après avoir tracé les courbes, nous pouvons voir qu'elles ont presque le même taux de true positive mais l'algorithme de forêt aléatoire a obtenu moins de false positive.

### 2.4 Comparaison des performances

```
| score_lr_train =accuracy_score(y_train,y_pred_lr_train)
| score_lr_test =accuracy_score(y_test,y_pred_lr_test)
| score_lr_test =accuracy_score(y_test,y_pred_lr_test)
| score_rfc_train =accuracy_score(y_train,y_pred_rfc_train)
| score_rfc_test =accuracy_score(y_test,y_pred_rfc_test)
| In [49]: | print(f"acuracy_score using logitatic regression training : {score_lr_train:.3f} test :{score_lr_test:.3f}")
| acuracy_score using logitatic regression training : 0.986 test :0.978
| In [50]: | print(f"acuracy_score using random forest : {score_rfc_train:.3f} test :{score_rfc_test:.3f}")
| acuracy_score using random forest : 1.000 test :0.993
```

Nous avons utilisé accracy\_score pour comparer à la fois les données de test et les données prédites.

La forêt aléatoire a obtenu un meilleur score que la régression logistique.

```
In [51]: cm_lr=confusion_matrix(y_test,y_pred_lr_test)
    cm_rfc=confusion_matrix(y_test,y_pred_rfc_test)

In [52]: print("Logistic Regression")
    print(cm_lr)
    print("Random Forest Classifier")
    print(cm_rfc)

    Logistic Regression
    [[2501    96]
        [ 20    2583]]
    Random Forest Classifier
    [[2571    26]
        [ 10    2593]]
```

À partir des matrices de confusion, nous pouvons voir que la forêt aléatoire a obtenu moins de valeurs false négative et false positive.

Nous avons aussi utilisé le score de validation croisée pour évaluer les performances des deux algorithmes.

L'algorithme de foret aléatoire a obtenu le meilleur score.

#### 2.5 Choix du meilleur modèle

D'après la comparaison des performances, l'algorithme de forêt aléatoire est le meilleur algorithme pour notre objectif.

### 2.6 Évaluation sur l'ensemble de test

Nous avons créé une fonction nommée is\_real, elle renvoie la prédiction de notre modèle, nous lui avons donc donné la première valeur dans nos données de test et elle nous a indiqué qu'elle était vraie et après avoir afficher la sortie des données de test, nous avons découvert que c'était effectivement vraie.

### 3 L'interface API

```
from flask import Flask, request, jsonify
from flask_cors import CORS
import joblib
import signal
import sys
import re
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
```

- Flask, request et jsonify sont utilisés pour créer une application web Flask et manipuler les requêtes HTTP et les réponses JSON.
- CORS est utilisé pour activer CORS (Cross-Origin Resource Sharing), ce qui permet à l'API d'accepter les requêtes depuis n'importe quelle origine.
- joblib est utilisé pour charger le modèle de machine learning et le vectorizer préentraînés.
- signal et sys sont utilisés pour gérer les signaux système, ce qui permet de fermer correctement l'application lorsqu'elle est interrompue.
- re est utilisé pour effectuer des opérations de correspondance de motifs (dans ce cas, pour le prétraitement du texte).
- stopwords est utilisé pour obtenir une liste de mots vides à supprimer lors du prétraitement du texte.
- PorterStemmer est utilisé pour effectuer la racinisation (stemming) des mots.

```
port_stem = PorterStemmer()

def stemming(content):
    stemmed_content = re.sub("[^a-zA-Z]"," ",content)

stemmed_content = stemmed_content.lower()

stemmed_content = stemmed_content.split()

stemmed_content = [port_stem.stem(word) for word in stemmed_content if not word in stopwords.words("English")]

stemmed_content = " ".join(stemmed_content)

return stemmed_content
```

lci, nous initialisons un objet PorterStemmer et définissons une fonction stemming qui prend du texte en entrée, effectue plusieurs étapes de prétraitement (suppression de la ponctuation, mise en minuscules, racinisation des mots et suppression des mots vides) et renvoie le texte prétraité.

```
1 app = Flask(__name__)
2 CORS(app, resources={r"/*": {"origins": "*"}})
```

Nous créons une instance de l'application Flask et activons CORS pour permettre les requêtes depuis n'importe quelle origine.

Nous définissons une route /predict qui accepte les requêtes POST contenant les données des articles de presse. Les données sont prétraitées à l'aide de la fonction de stemming, puis transformées à l'aide du vectorizer. Enfin, le modèle est utilisé pour prédire la classe de l'article de presse et renvoyer la prédiction au client sous forme JSON. Si une erreur se produit, elle est renvoyée sous forme de réponse JSON.

```
1 signal.signal(signal.SIGINT, signal.SIG_DFL)
```

Nous configurons le gestionnaire de signaux pour capturer SIGINT (par exemple, CTRL+C) et utiliser la gestion par défaut de l'interruption.

```
1 if __name__ == '__main__':
2    app.run(host='localhost', port=5000, debug=True, use_reloader=False)
```

Nous exécutons l'application Flask sur le port 5000 en mode debug. Cette partie du code ne sera exécutée que si le script est exécuté directement (et non importé en tant que module).

### 4 Index Html

```
<!DOCTYPE html>
                                                                                          ¥14
<html lang="fr">
<head>
    <meta charset="UTF-8">
    <meta name="viewport" content="width=device-width, initial-scale=1.0">
    <title>Détecteur de Fausses Nouvelles</title>
    <link rel="stylesheet" href="style.css">
</head>
        <header>
            <h1>Détecteur de Fausses Nouvelles</h1>
        </header>
        <main>
            <form id="newsForm">
                <label for="authorInput">Auteur :</label>
                <input type="text" id="authorInput" placeholder="Entrez le nom de l'auteur">
```

Ce code HTML représente une page web qui met en œuvre un "Détecteur de Fausses Nouvelles". L'utilisateur peut saisir le nom de l'auteur ainsi que le titre de la nouvelle dans des champs dédiés. Ensuite, en cliquant sur le bouton "Vérifier les Nouvelles", une fonction JavaScript est déclenchée pour analyser les données fournies et évaluer la véracité de la nouvelle. Le résultat de cette analyse est affiché dans la section dédiée à cet effet sur la page.

## 5 Java Script

```
async function checkNews() {
    const authorInput = document.getElementById("authorInput").value || "";
    const newsInput = document.getElementById("newsInput").value;

if (newsInput.trim() === "") {
    alert("Veuillez saisir le nom de l'auteur et la nouvelle avant de vérifier.");
    return;
}

try {
    const response = await fetch("http://localhost:5000/predict", {
    method: "POST",
    headers: {
        "Content-Type": "application/json",
    },
    body: JSON.stringify({ author: authorInput, news: newsInput }),
});

const result = await response.json();
    displayResult(result);
} catch (error) {
    console.error("Erreur :", error);
    alert("Une erreur s'est produite lors du traitement de votre demande.");
}
}
```

```
function displayResult(result) {
   const resultElement = document.getElementById("result");

if (result.error) {
   resultElement.innerHTML = "Erreur : " + result.error;
} else {
   if (result.prediction === 1) {
      resultElement.innerHTML = "Cette nouvelle est fausse.";
} else {
      resultElement.innerHTML = "Cette nouvelle est vraie.";
} else {
      resultElement.innerHTML = "Cette nouvelle est vraie.";
}
}
```

Ce code JavaScript implémente une fonction asynchrone checkNews() qui envoie une requête HTTP POST à un serveur local, demandant une prédiction sur la véracité d'une nouvelle saisie par l'utilisateur. La réponse est ensuite affichée sur la page HTML. La fonction displayResult() est utilisée pour présenter le résultat de la prédiction à l'utilisateur.

## 6 CSS

```
body {
    font-family: 'Segoe UI', Tahoma, Geneva, Verdana, sans-serif;
    background-color: #eef5f9; /* Light blue-gray background */
    margin: 0;
    padding: 0;

    container {
        max-width: 800px;
        margin: 50px auto;
        background-color: #ffffff; /* White container background */
    padding: 30px;
    border-radius: 8px;
    box-shadow: 0 0 20px rgba(0, 0, 0, 0.1);
}

header {
    text-align: center;
    margin-bottom: 30px;
}
```

```
21
22  h1 {
23     color: #333; /* Dark gray text color */
     margin: 0;
25  }
26
27  main {
28     padding: 20px;
29  }
30
31  form {
32     max-width: 400px;
    margin: 0 auto;
34  }
35
36  label {
37     display: block;
    margin-bottom: 10px;
    color: #555; /* Medium gray text color */
40  }
```

```
input,
textarea {
    width: 100%;
    padding: 10px;
    border: 1px solid #ccc;
    border-radius: 4px;
    margin-bottom: 20px;
button {
    background-color: #00563F; /* Blue button color */
    color: #fff; /* White text color */
    padding: 15px;
    border: none;
    border-radius: 4px;
    cursor: pointer;
    font-size: 16px;
    width: 100%;
    transition: background-color 0.3s ease;
```

```
button:hover {
background-color: #0056b3; /* Darker blue on hover */

background-color: #0056b3; /* Darker blue on hover */

#result {
text-align: center;
font-size: 18px;
color: #333; /* Dark gray text color */

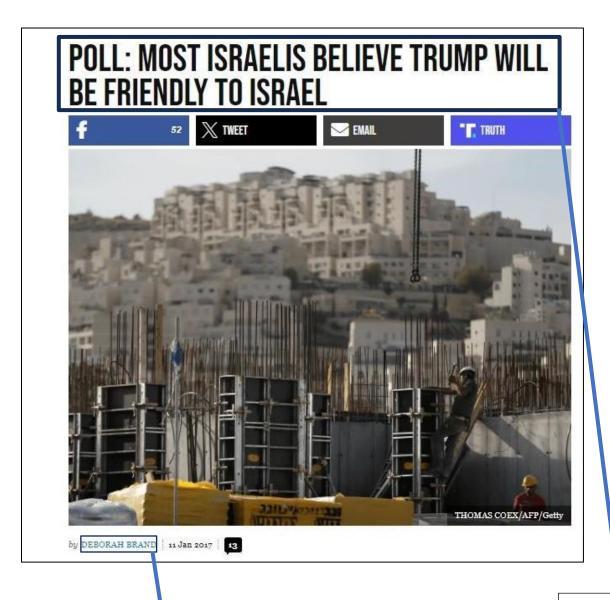
}
```

Ce code CSS fournit une mise en forme esthétique et conviviale pour une page web qui implémente un formulaire de saisie de données. Il utilise une palette de couleurs douces et une disposition équilibrée pour offrir une expérience visuelle agréable aux utilisateurs. Les styles appliqués assurent une présentation claire et lisible des éléments, facilitant ainsi la navigation et l'interaction avec le contenu.

## 7 Test

Cherchons les news:

## News 1:

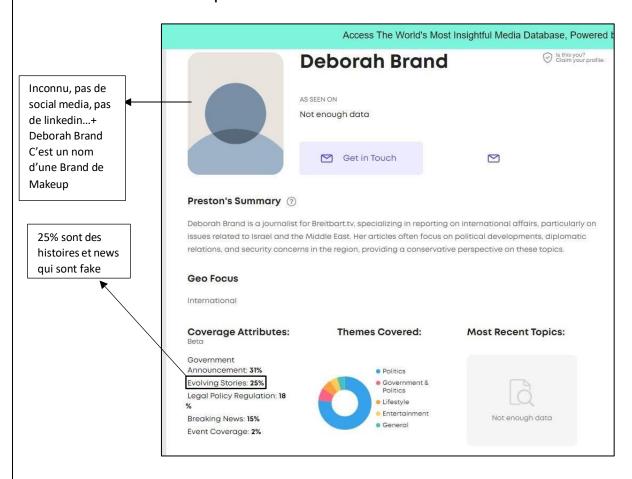


Titre de News

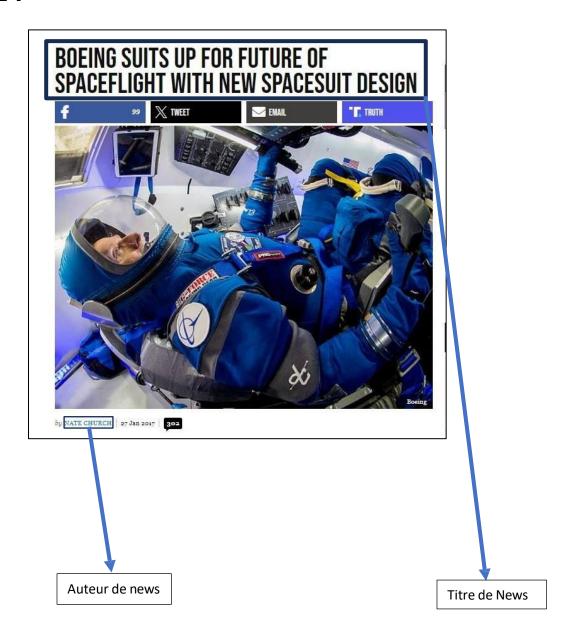
Auteur de news



Donc cette nouvelle est fausse, on a trouve que l'auteur de cette nouvelle a une mauvaise reputation



# News 2:





Donc cette nouvelle est vraie, on a trouve que l'auteur de cette nouvelle a une Bonne reputation

