Analyse des données textuelles d'articles de presse

Présentation Stage Première Année

Adib Habbou

Haut-Commissariat au Plan du Maroc

15 juillet 2022

Haut-Commissariat au Plan

المندوبية السامية للتخطيط
HAUT-COMMISSARIAT AU PLAN

- **Structure ministérielle** marocaine crée en septembre 2003
- Principal producteur de l'**information statistique** au Maroc
- Locaux dans les quartiers d'*Agdal* et de *Hay Riad* à Rabat
- Admet une **Direction des Systèmes d'Information Statistique**

Direction des Systèmes d'Information Statistique

المندوبية السامية للتخطيط
HAUT-COMMISSARIAT AU PLAN

- **Collecte des données** à travers différentes sources
- **Nettoyage**, **classification** et **tri** des données
- **Stockage** des données dans les **Bases de Données Statistiques**
- **Analyse des données** pour extraire des **indicateurs**
- Publication de **rapports** *(trimestriel, semestriel et annuel)*

Environnement

- Encadré par **Oumaima Hourrane**, doctorante et ingénieur en ML
- Notebook Python sur **Google Colaboratory**
- Script Python sur **PyCharm**

- Trois **sites de presse en ligne** marocaine étudié :
  1. Morocco World News → Anglais
  2. Le Matin → Français
  3. Hespress → Arabe

Problématique

Besoin de données sur la réalité économique, sociale et culturelle du pays

- → Études ministérielles et gouvernementales
- → Études de marché pour le secteur privé
- → Études académiques et universitaires

Problématique

Besoin de données sur la réalité économique, sociale et culturelle du pays

→ Études ministérielles et gouvernementales

→ Études de marché pour le secteur privé

→ Études académiques et universitaires

Existant

- Observatoire des conditions de vie des ménages
- Centre d'études et de recherches démographiques
- Recensement générale de la population et de l'habitat

Problématique

Besoin de données sur la réalité économique, sociale et culturelle du pays

→ Études ministérielles et gouvernementales

→ Études de marché pour le secteur privé

→ Études académiques et universitaires

Existant

- Observatoire des conditions de vie des ménages
- Centre d'études et de recherches démographiques
- Recensement générale de la population et de l'habitat

Solutions

- Web Scraping
- Machine Learning

Web Scraping

Technique de récupération de données à partir d'un site Web dynamique

Web Scraping

Technique de récupération de données à partir d'un site Web dynamique

- Interaction avec les navigateurs via un driver
- Navigation sur une page web dynamique (formulaires, boutons...)
- Localisation des éléments (id, name, xpath, class name...)

Textual Data Visualization

Représentation visuelle du contenu d'un document textuel

Textual Data Visualization

Représentation visuelle du contenu d'un document textuel

- Répartition des catégories, auteurs, dates avec **Plotly, Matplotlib**
- Synthèse des mots clés d'un texte avec **WordCloud**

Topic Modeling

Modèle non supervisé permettant d'extraire les sujets de documents

Topic Modeling

Modèle non supervisé permettant d'extraire les sujets de documents

- Preprocessing des données avec **NLTK**
- Application du modèle avec **Gensim**

Text Classification

Modèle supervisé nécessitant un entraînement sur des données labélisées

Text Classification

Modèle supervisé nécessitant un entraînement sur des données labélisées

- Preprocessing des données avec **NLTK**
- Application de différents modèles avec **Scikit-Learn**
- Évaluation des modèles de Machine Learning avec **Scikit-Learn**

Dashboard

Streamlit

Application Web pour présenter visuellement les résultats obtenus

Dashboard

Streamlit

Application Web pour présenter visuellement les résultats obtenus

- Transformation de script Python en application Web avec **Streamlit**
- Intégration de code HTML généré par **pyLDAvis**
- Intégration de graphiques crée avec **Plotly**

Web Scraping

Méthodologie adoptée

1. Parcours des catégories : https ://www.site.com/category
2. Parcours des pages : https ://www.site.com/category/page_number
3. Extraction des liens de chaque article
4. Parcours des liens d'articles et récupération des données

Topic Modeling - LDA

Latent Dirichlet Allocation

- Chaque document noté M est un mélange θ d'un petit nombre de sujets α
- Attribue topic à chaque mot selon une distribution de Dirichlet :

$$\theta_i \sim Dir(\alpha) \ \ pour \ \ 1 \leq i \leq M \ \ avec \ \ \alpha < 1$$

- Mise à jour du topic lié à chaque mot en fonction de la probabilité :

$$\mathcal{P}(t|d) \ x \ \mathcal{P}(w|t)$$

Topic Modeling - LDA

Latent Dirichlet Allocation

- Chaque document noté M est un mélange θ d'un petit nombre de sujets α
- Attribue topic à chaque mot selon une distribution de Dirichlet :

$$\theta_i \sim Dir(\alpha) \ pour \ 1 \leq i \leq M \ avec \ \alpha < 1$$

- Mise à jour du topic lié à chaque mot en fonction de la probabilité :

$$\mathcal{P}(t|d) \ x \ \mathcal{P}(w|t)$$

Calcul de probabilité

- $\mathcal{P}(t|d)$: la probabilité que le document d soit assigné au topic t
- $\mathcal{P}(w|t)$: la probabilité que le thème t soit assigné au mot w

Topic Modeling - LDA

Preprocessing

- Tokenisation : tout en minuscule, suppression ponctuation
- Lemmatisation : tout au présent et 1ère personne
- Racinisation : réduction à la forme radicale
- Bag of Words : dictionnaires (mots/nombre d'occurrences)

Topic Modeling - LDA

Preprocessing

- Tokenisation : tout en minuscule, suppression ponctuation
- Lemmatisation : tout au présent et 1ère personne
- Racinisation : réduction à la forme radicale
- Bag of Words : dictionnaires (mots/nombre d'occurrences)

Visualisation **pyLDAvis**

- Extraction des informations du topic model
- Réalisation d'une visualisation Web interactive

Text Classification

Modèles de Machine Learning

- K Nearest Neighbour
- Logistic Regression
- Random Forest
- Support Vector Machine
- Stochastic Gradient Descent
- Multi-Layer Perceptron

Text Classification

Modèles de Machine Learning

- K Nearest Neighbour
- Logistic Regression
- Random Forest
- Support Vector Machine
- Stochastic Gradient Descent
- Multi-Layer Perceptron

Évaluation des modèles

- Précision : taux de prédictions positives correctes
- Recall : taux de positifs correctement prédits
- F1-score : capacité d'un modèle à bien prédire les individus positifs
- Matrice Confusion : ligne catégorie réelle, colonne catégorie estimée

Dashboard

Contenu du Dashboard

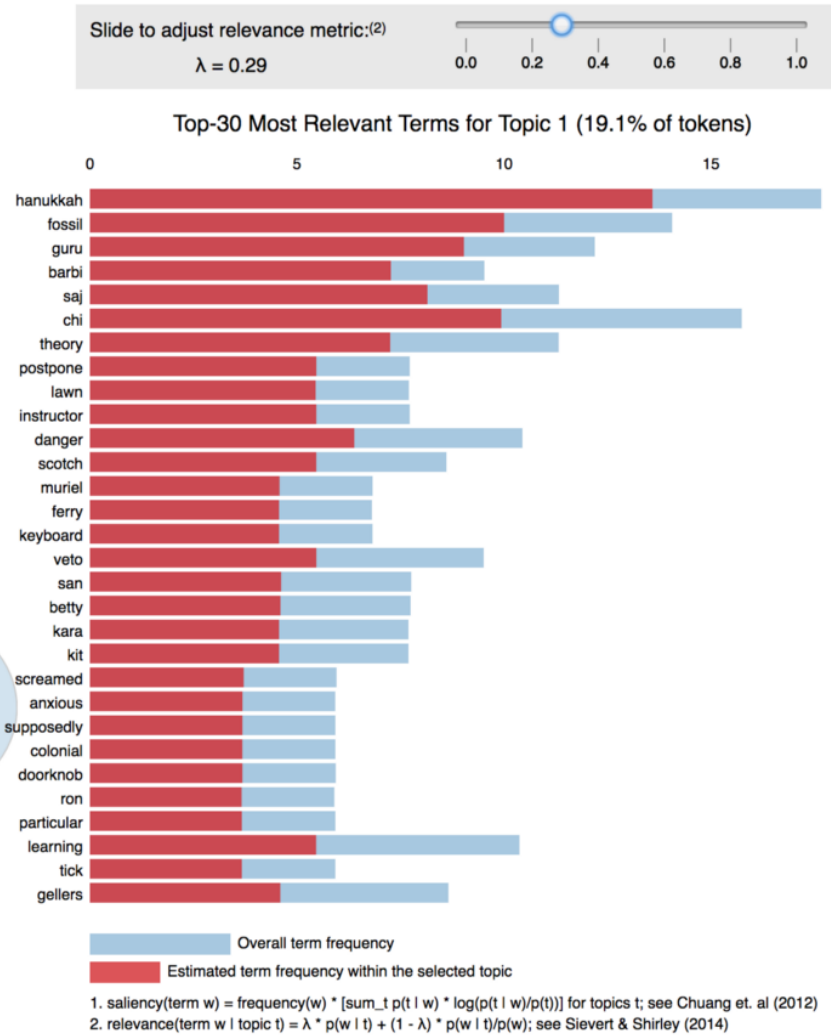
- Échantillon du data set
- Résultat du topic modeling
- Répartition des catégories et des auteurs
- Précisions des modèles de classification
- Matrices de confusion des modèles

Résultat Web Scraping

	title	lead	author	date	content
0	Spanish FM: Spain Works on 'Constructive, Firm...	Algeria decided to suspend its friendship trea...	Safaa Kasraoui	June 09, 2022 10:53 a.m.	Rabat - Spain regrets Algeria's decision to su...
1	Afro-Atlantic Treaty: Morocco Says Only Unity,...	Morocco's Foreign Affairs Minister insists tha...	Aya Benazizi	June 08, 2022 5:08 p.m.	Casablanca - With the Moroccan capital of Raba...
2	Cape Verde to Open Consulate General in Morocc...	Togo also announced the forthcoming opening of...	Safaa Kasraoui	June 08, 2022 4:40 p.m.	Rabat - Cape Verde's Foreign Affairs Minister ...
3	Mexican Delegation To Visit Morocco, Keen To C...	Migration management, the fight against climat...	Souad Anouar	June 08, 2022 4:20 p.m.	Rabat - A Mexican parliamentary delegation is ...
4	Spanish PM Renews Support for Morocco's Autono...	A number of marginal opposition political part...	Safaa Kasraoui	June 08, 2022 2:02 p.m.	Rabat - Spain's Prime Minister Pedro Sanchez h...

Figure – Data Frame de Morocco World News

Résultat Topic Modeling



Précisions des modèles

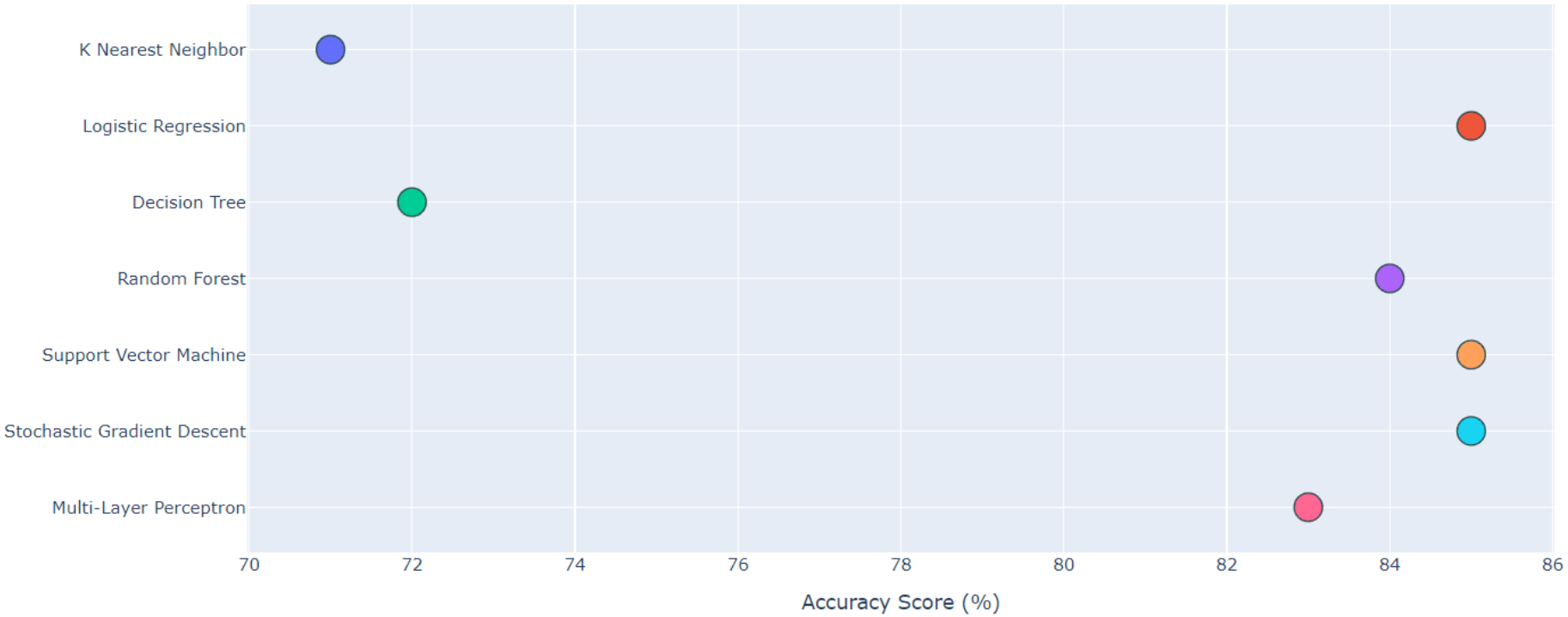
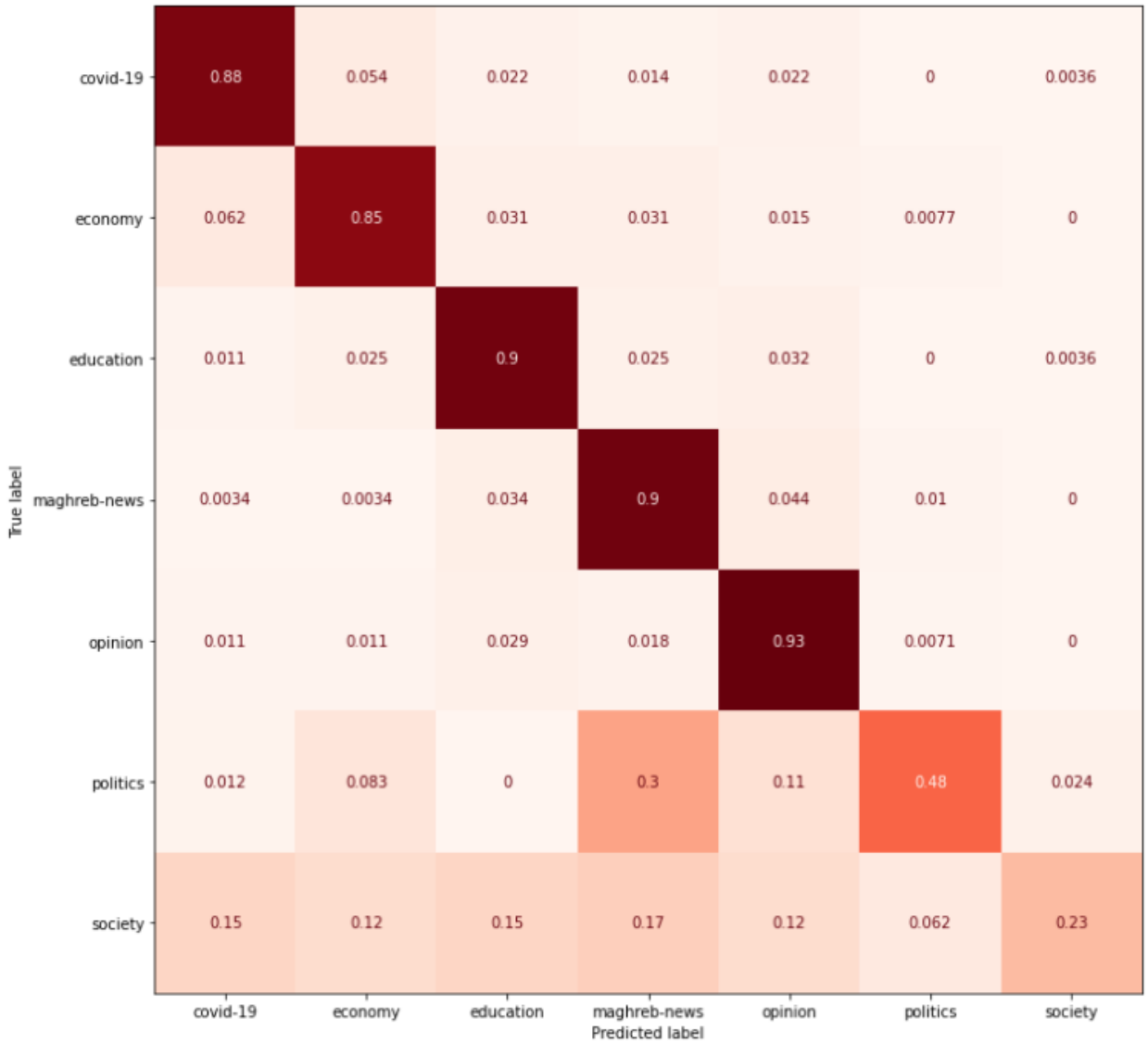


Figure – Précision des différents modèles

Matrice de confusion



Résultat Dashboard

	title	lead	author	date
0	Spanish FM: Spain Works on 'Constructive, Firm' Response to Algeria's Decision	Algeria decided to suspend its friendship treaty with Spain to further protest the country's endorseme	Safaa Kasraoui	June 09, 2022 10:53 a.
1	Afro-Atlantic Treaty: Morocco Says Only Unity, Solidarity Can Save Africa	Morocco's Foreign Affairs Minister insists that only continental solidarity can save Africa from being a p	Aya Benazizi	June 08, 2022 5:08 p.m
2	Cape Verde to Open Consulate General in Morocco's Dakhla	Togo also announced the forthcoming opening of a consulate in Dakhla.	Safaa Kasraoui	June 08, 2022 4:40 p.m
3	Mexican Delegation To Visit Morocco, Keen To Consolidate Relations	Migration management, the fight against climate change, and the Western Sahara dispute are expecte	Souad Anouar	June 08, 2022 4:20 p.m
4	Spanish PM Renews Support for Morocco's Autonomy Plan Amid Hostile Campaign	A number of marginal opposition political parties in Spain are taking issue with the Sanchez governme	Safaa Kasraoui	June 08, 2022 2:02 p.m
5	Moroccan Ambassador: Polisario's Sultana Khaya is Not a Human Rights Defender	While presenting herself as a pacifist champion of human rights, Sultana Khaya has been recorded anc	Safaa Kasraoui	June 08, 2022 11:10 a.
6	Minister: Spain Wants to Preserve Ties with 'Reliable, Fraternal' Morocco	Since endorsing Morocco's Autonomy Plan for Western Sahara, Madrid has repeatedly renewed its reac	Safaa Kasraoui	June 07, 2022 5:02 p.m
7	Saudi Arabia Engages in 'Serious Talks' With Israel For Potential Normalization	Many have suggested normalization of ties between Saudi Arabia and Israel is just a matter of time.	Safaa Kasraoui	June 07, 2022 2:14 p.m

Conclusion

Apport

- Étudier la faisabilité de l'ensemble du processus
- Identifier l'ensemble des technologies nécessaires
- Sélectionner les modèles de *Machine Learning* les plus performants

Conclusion

Apport

- Étudier la faisabilité de l'ensemble du processus
- Identifier l'ensemble des technologies nécessaires
- Sélectionner les modèles de *Machine Learning* les plus performants

Perspective

- Infrastructure d'extraction et de nettoyage des données
- Pipeline d'application du Topic Modeling et de la Text Classification
- Développement d'une application Web de monitoring des résultats

Merci pour votre attention