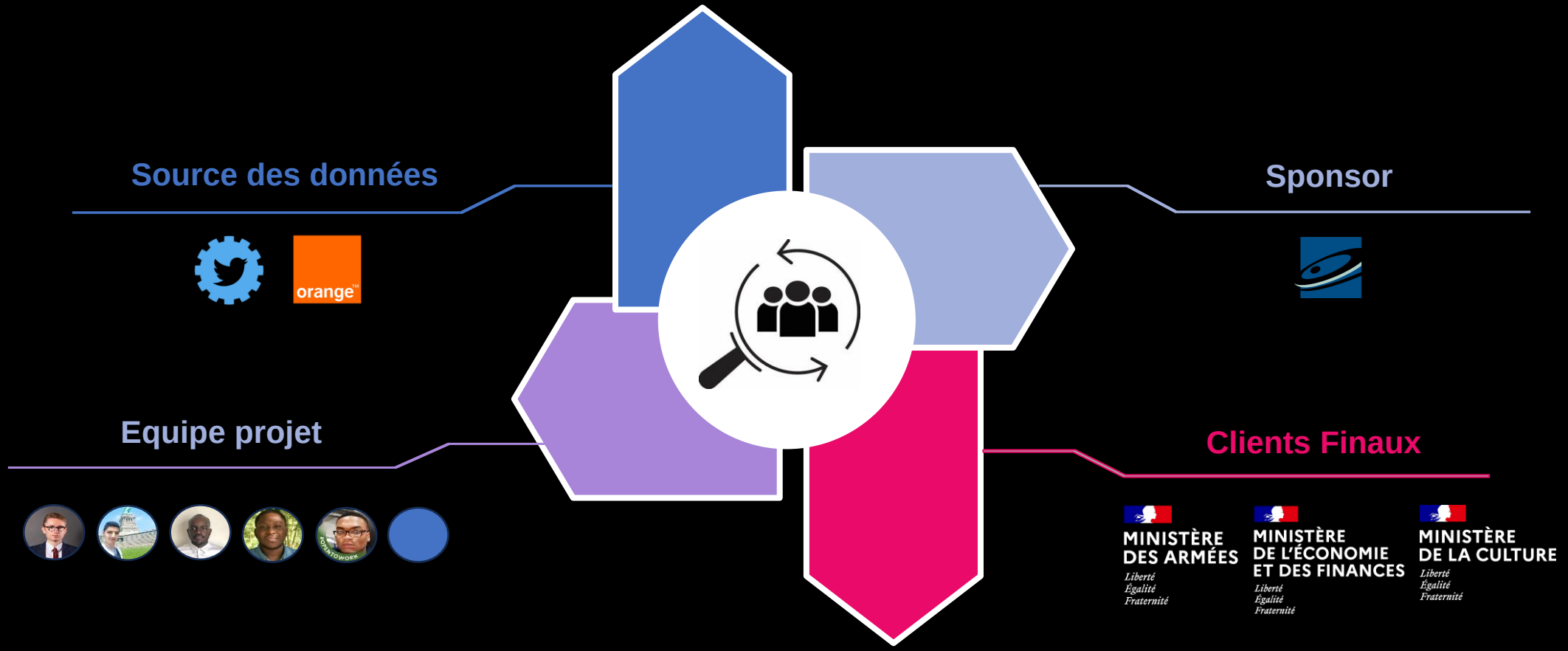
Détection d’anomalie sur un signal d’activité du réseau mobile

UV9
Paris
04/04/23

Parties prenantes du projet



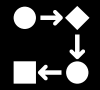
CRISP-DM Canva

Business Understanding

- Quels événements participent à une multiplication des anomalies sur Twitter ?
- Quels Tweets participent à cette multiplication ?
- Quelles zones géographiques sont touchées par ces anomalies ?

Data Understanding

Data Preparation

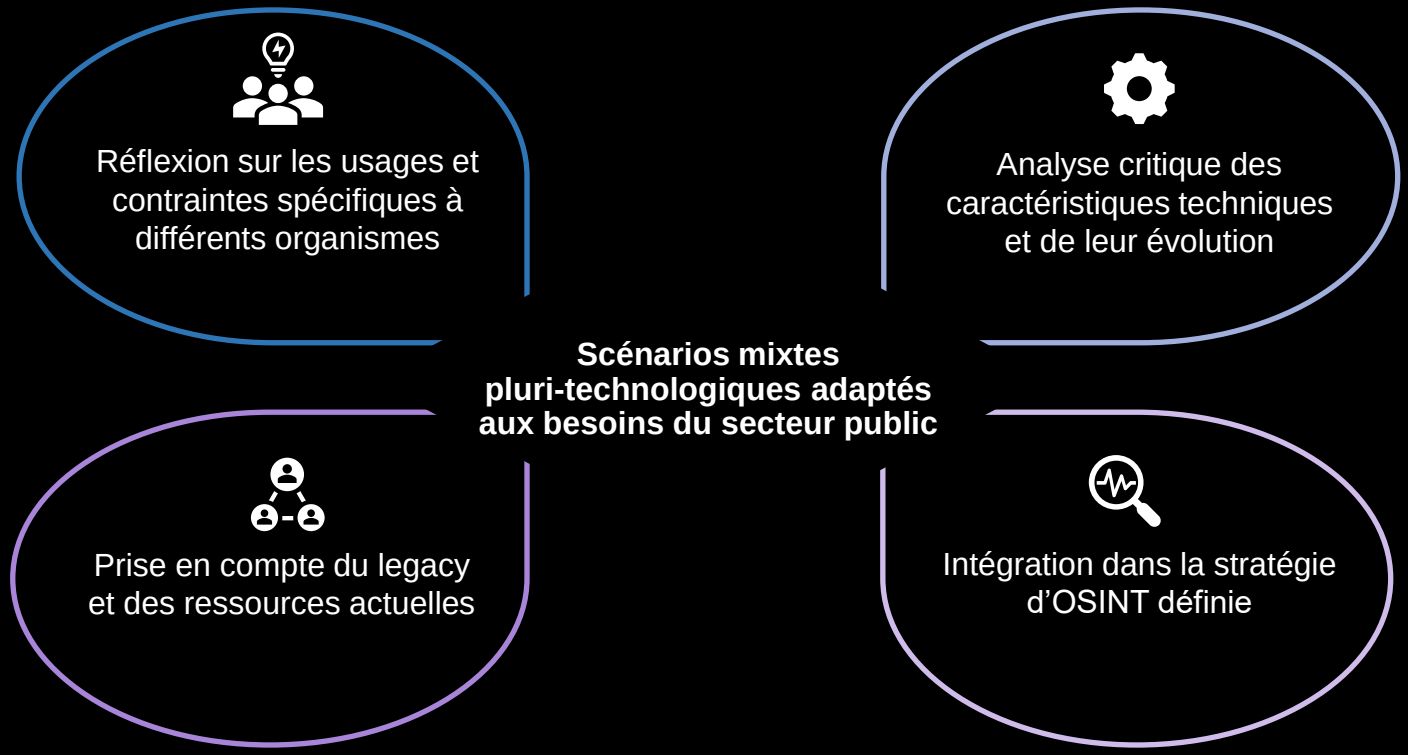


Modelling

Evaluation

Deployment

Enjeux du secteur public



Dimensionnement sur 3 axes : temporalité, distinction, localisation

CRISP-DM Canva

Business Understanding

- Quels événements participent à une multiplication des anomalies sur Twitter ?
- Quels Tweets participent à cette multiplication ?
- Quelles zones géographiques sont touchées par ces anomalies ?

Data Understanding

- Signal de l'activité de l'opérateur téléphonique Orange pendant une plage de temps
- Requête à l'API de Twitter (JSON, plusieurs Go)
- Récupérer les identifiants de l'émetteur du tweet, les hashtags utilisés, les mentions associées et la date de création du Tweet

- Réalisation d’un dictionnaire de variables

Data Preparation

Création de 2 datasets supplémentaires :

- ✓ Un dataset concentré sur les hashtags
- ✓ Un dataset concentré sur les liaisons entre les tweets (les retweets et les quoted tweets)

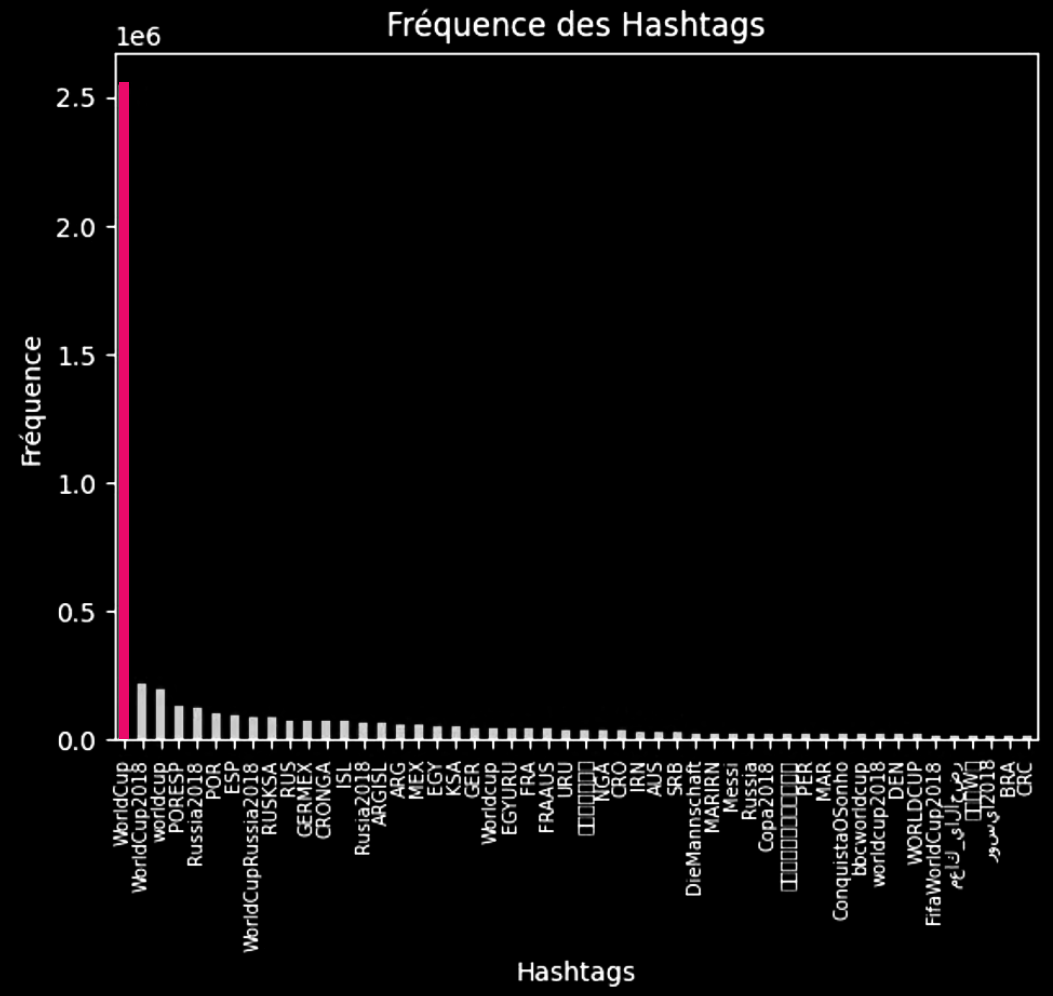
- Listage des dimensions intéressantes
- Récupération des dimensions sous Python
- Nettoyage par suppression des manquants

Modelling

Evaluation

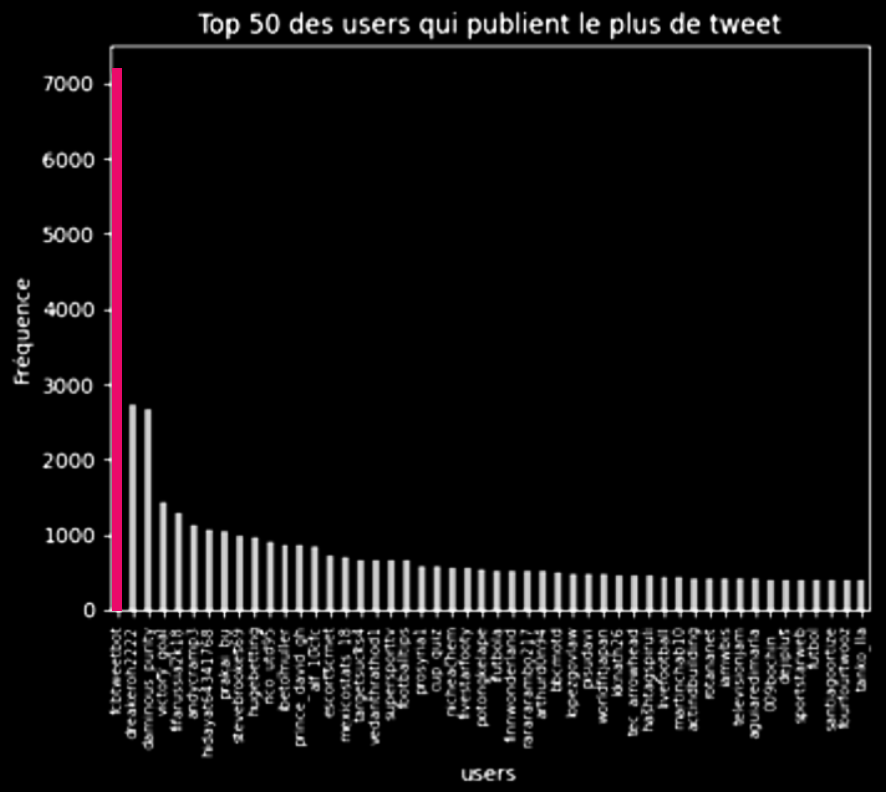
Deployment

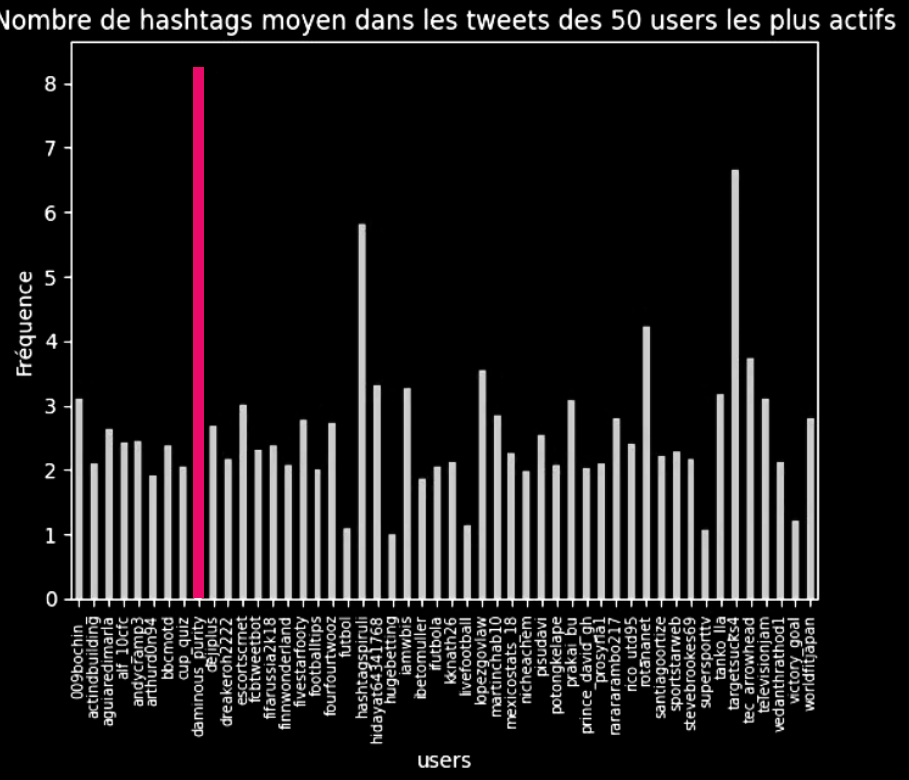
Analyse exploratoire sur la fréquence des hashtags



Insight ➞ Surreprésentation du hashtag WorldCup

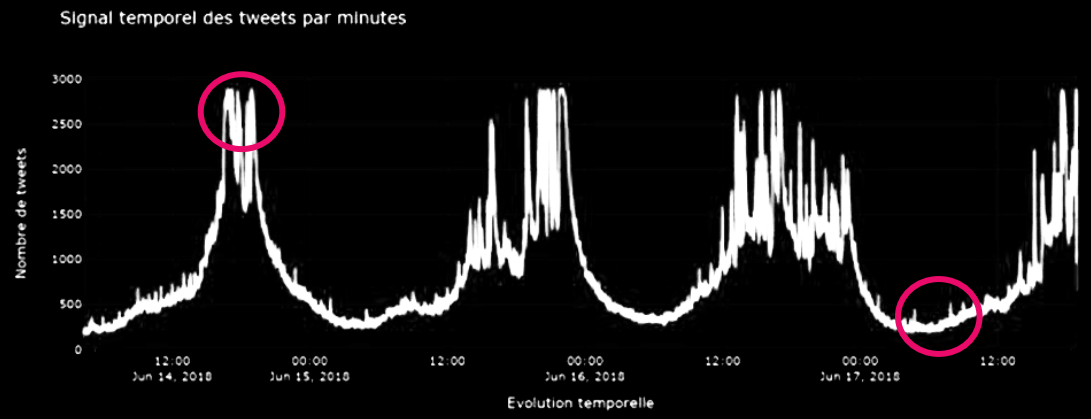
Analyse exploratoire sur l’activité utilisateur

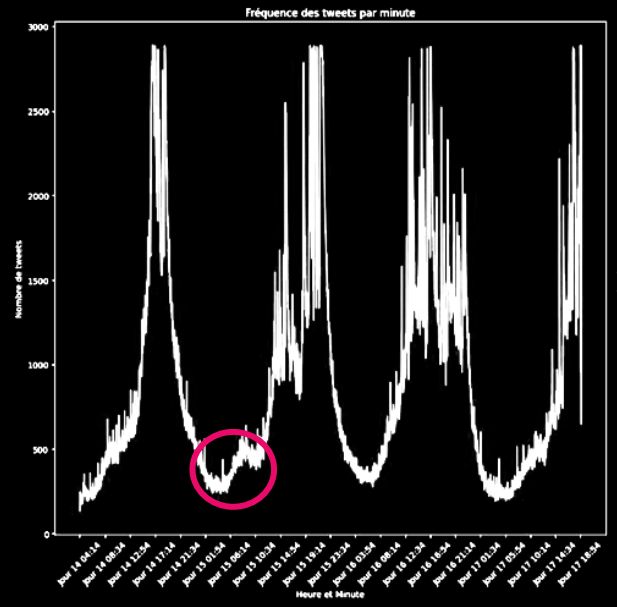


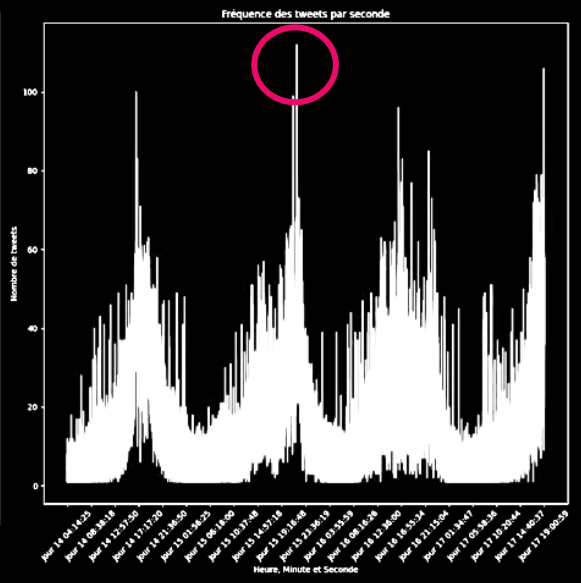


Insight → Fcbtweetbot est le compte / bot le plus actif et l’occurrence des hashtags dans les tweets est de maximum 8

Analyse exploratoire sur la fréquence du signal







Insight ➞ Des pics sont observables entre 12h et 15h avec 2500 tweets maximum et des baisses le sont aussi sur des plages horaires entre 4h et 11h

CRISP-DM Canva

Business Understanding

- Quels événements participent à une multiplication des anomalies sur Twitter ?
- Quels Tweets participent à cette multiplication ?
- Quelles zones géographiques sont touchées par ces anomalies ?

Data Understanding

- Signal de l'activité de l'opérateur téléphonique Orange pendant une plage de temps
- Requête à l'API de Twitter (JSON, plusieurs Go)
- Récupérer les identifiants de l'émetteur du tweet, les hashtags utilisés, les mentions associées et la date de création du Tweet

- Réalisation d'un dictionnaire de variables

Data Preparation

Création de 2 datasets supplémentaires :

- ✓ Un dataset concentré sur les hashtags
- ✓ Un dataset concentré sur les liaisons entre les tweets (les retweets et les quoted tweets)

- Listage des dimensions intéressantes
- Récupération des dimensions sous Python
- Nettoyage par suppression des manquants

- ✓ Suppression des tweets avec le hashtag WorldCup

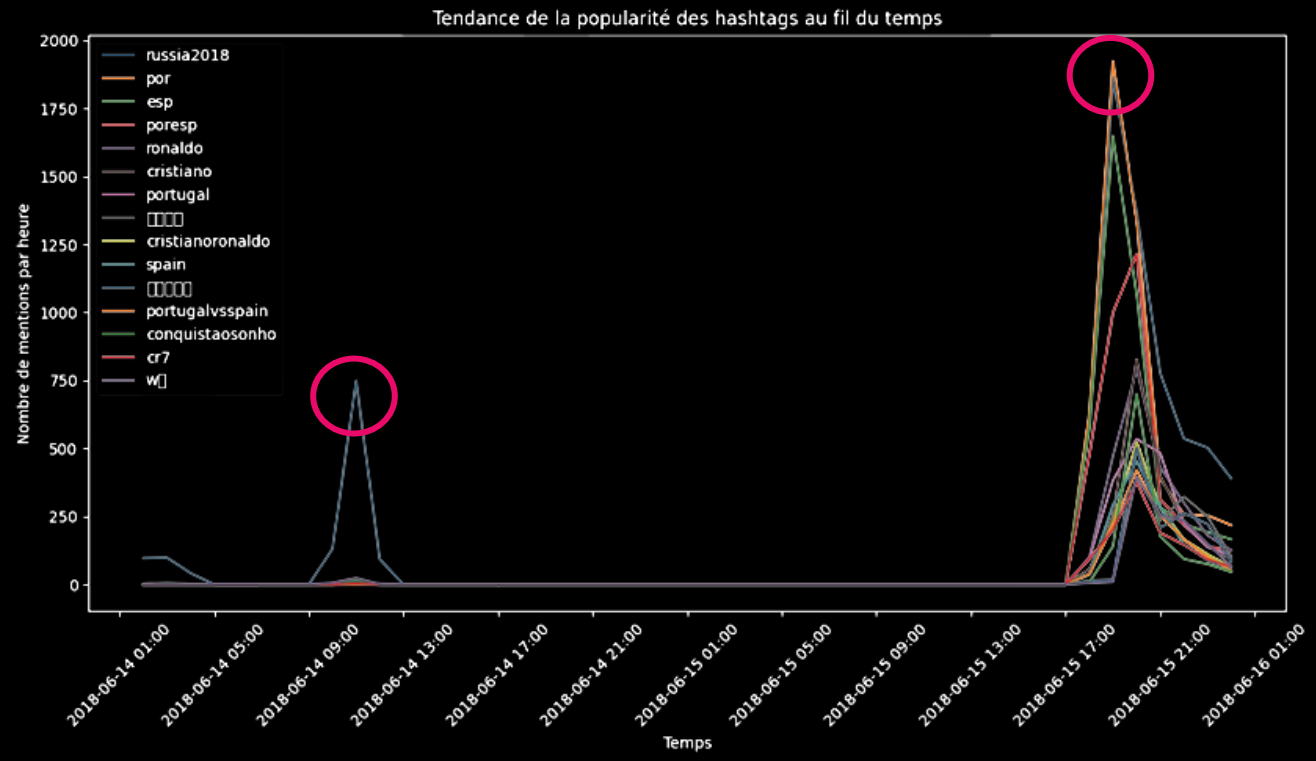
Modelling

- ✓ TF-IDF sur tweet_hashtag, user_mentions_screen_name et retweet_hashtags
- ✓ Encodage de la variable temporelle
- ✓ Modèle d'Isolation Forest avec ajustement des hyperparamètres n_estimators, max_samples, contamination
- ✓ Optimisation bayésienne en dix itérations sur le score moyen d'anomalie
- ✓ Algorithme de Louvain pour la détection de communautés de hashtags

Evaluation

Deployment

Analyse exploratoire sur la popularité des hashtags

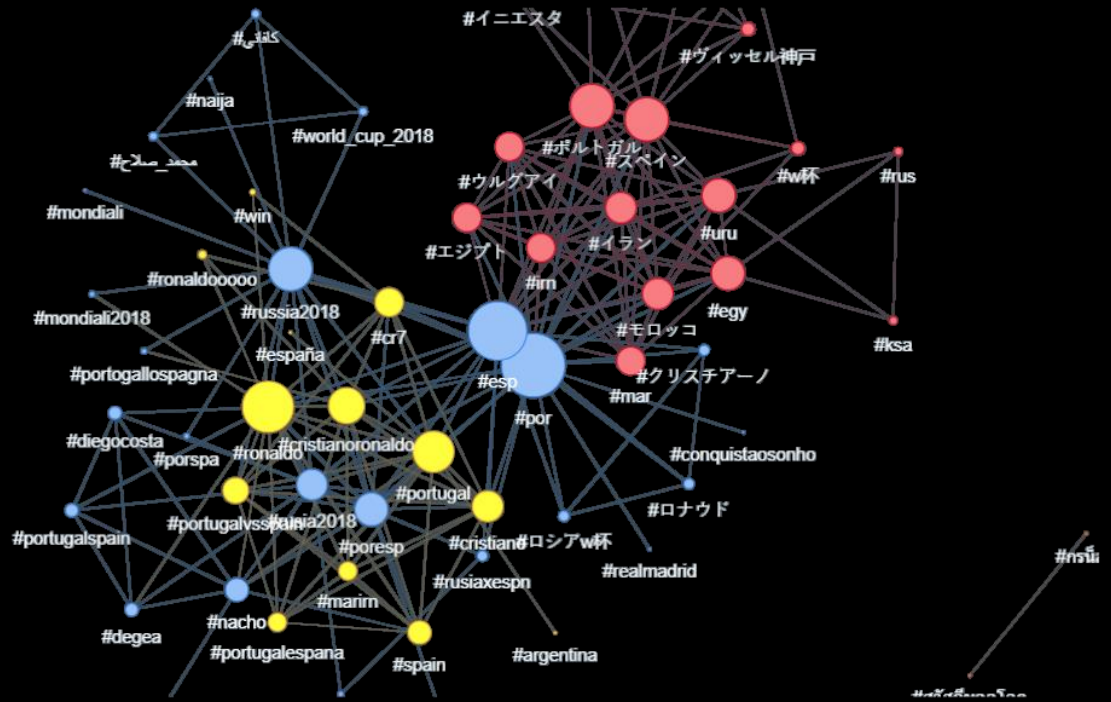


Insight ➞ Le hashtag #russia2018 a été mentionné en moyenne 154 fois par heure. Les hashtags #por, #esp et #poresp ont été mentionnés respectivement 106, 88 et 74 fois par heure.

Co-occurrences des hashtags

Insight → Les hashtags #por et #esp sont les plus cités ensemble, correspondants au match entre les deux équipes

Graphe de connaissance pour la détection de communauté



Insight ➞ Les nœuds rouges forment une communauté dense et semblent être centrés autour de hashtags en sinogrammes, ce qui pourrait indiquer un groupe d'utilisateurs principalement chinois

Insight ➞ Les nœuds jaunes et bleus pourraient représenter des communautés avec un intérêt particulier pour les matchs ou joueurs liés à l'Espagne et au Portugal, comme le suggèrent les hashtags #cristianoronaldo, #portugalspain et #diegocosta

CRISP-DM Canva

Business Understanding

- Quels événements participent à une multiplication des anomalies sur Twitter ?
- Quels Tweets participent à cette multiplication ?
- Quelles zones géographiques sont touchées par ces anomalies ?

Data Understanding

- Signal de l'activité de l'opérateur téléphonique Orange pendant une plage de temps
- Requête à l'API de Twitter (JSON, plusieurs Go)
- Récupérer les identifiants de l'émetteur du tweet, les hashtags utilisés, les mentions associées et la date de création du Tweet

- Réalisation d'un dictionnaire de variables

Data Preparation

Création de 2 datasets supplémentaires :

- ✓ Un dataset concentré sur les hashtags
- ✓ Un dataset concentré sur les liaisons entre les tweets (les retweets et les quoted tweets)

- Listage des dimensions intéressantes
- Récupération des dimensions sous Python
- Nettoyage par suppression des manquants

- ✓ Suppression des tweets avec le hashtag WorldCup

Modelling

- ✓ TF-IDF sur tweet_hashtag, user_mentions_screen_name et retweet_hashtags

- ✓ Encodage de la variable temporelle

- ✓ Modèle d'Isolation Forest avec ajustement des hyperparamètres n_estimators, max_samples, contamination

- ✓ Optimisation bayésienne en dix itérations sur le score moyen d'anomalie

- ✓ Algorithme de Louvain pour la détection de communautés de hashtags

Evaluation

Pour le ministère de la culture et le ministère de l'économie :

- ➢ Il existe un targeting possible sur une forte communauté chinoise (soft power culturel et déploiement commercial – French Tech)

Pour le ministère des armées :

- ➢ Il est nécessaire de prendre des précautions avec ces utilisateurs chinois au vu de la situation géopolitique

Deployment

Formalisation du cas d’usage

Contexte : Aider le secteur public à identifier des anomalies sur les réseaux sociaux afin de prévenir des menaces ou de trouver des opportunités selon les organismes. Le modèle mis en place permettra d’identifier des proximités entre utilisateurs et d’obtenir des insights stratégiques.

Utilisateurs cibles

- Marketing / Communication / Analystes
- OSINT
- Services de sécurité
- Managers

Bénéfices du cas d’usage

- Ciblage commercial
- Campagnes culturelles
- Sécurité proactive
- Promotion de l’écosystème économique français

Données nécessaires

- API Twitter
- Données mobiles

Limites/challenges

- Volume de la donnée
- Qualité de la donnée
- Analyses sectorielles

Valeur

★ ★ ★

Effort

★ ★ ☆

Valeur métier

- Augmentation globale de la proactivité pour les fonctions métiers
- Développement économique et culturel

Complexité technique

- Maintien de la plateforme
- Déploiement sectoriel
- Enrichissement de la donnée

Statut

- ☑ à lancer
- ☐ à expérimenter
- ☐ à arrêter
- ☐ à industrialiser

Quick win ?

CRISP-DM Canva

Business Understanding

- Quels événements participent à une multiplication des anomalies sur Twitter ?
- Quels Tweets participent à cette multiplication ?
- Quelles zones géographiques sont touchées par ces anomalies ?

Data Understanding

- Signal de l'activité de l'opérateur téléphonique Orange pendant une plage de temps
- Requête à l'API de Twitter (JSON, plusieurs Go)
- Récupérer les identifiants de l'émetteur du tweet, les hashtags utilisés, les mentions associées et la date de création du Tweet

- Réalisation d'un dictionnaire de variables

Data Preparation

Création de 2 datasets supplémentaires :

- ✓ Un dataset concentré sur les hashtags
- ✓ Un dataset concentré sur les liaisons entre les tweets (les retweets et les quoted tweets)

- Listage des dimensions intéressantes
- Récupération des dimensions sous Python
- Nettoyage par suppression des manquants

- ✓ Suppression des tweets avec le hashtag WorldCup

Modelling

- ✓ TF-IDF sur tweet_hashtag, user_mentions_screen_name et retweet_hashtags

- ✓ Encodage de la variable temporelle

- ✓ Modèle d'Isolation Forest avec ajustement des hyperparamètres n_estimators, max_samples, contamination

- ✓ Optimisation bayésienne en dix itérations sur le score moyen d'anomalie

- ✓ Algorithme de Louvain pour la détection de communautés de hashtags

Evaluation

Pour le ministère de la culture et le ministère de l'économie :

- ➢ Il existe un tageting possible sur une forte communauté chinoise (soft power culturel et déploiement commercial – French Tech)

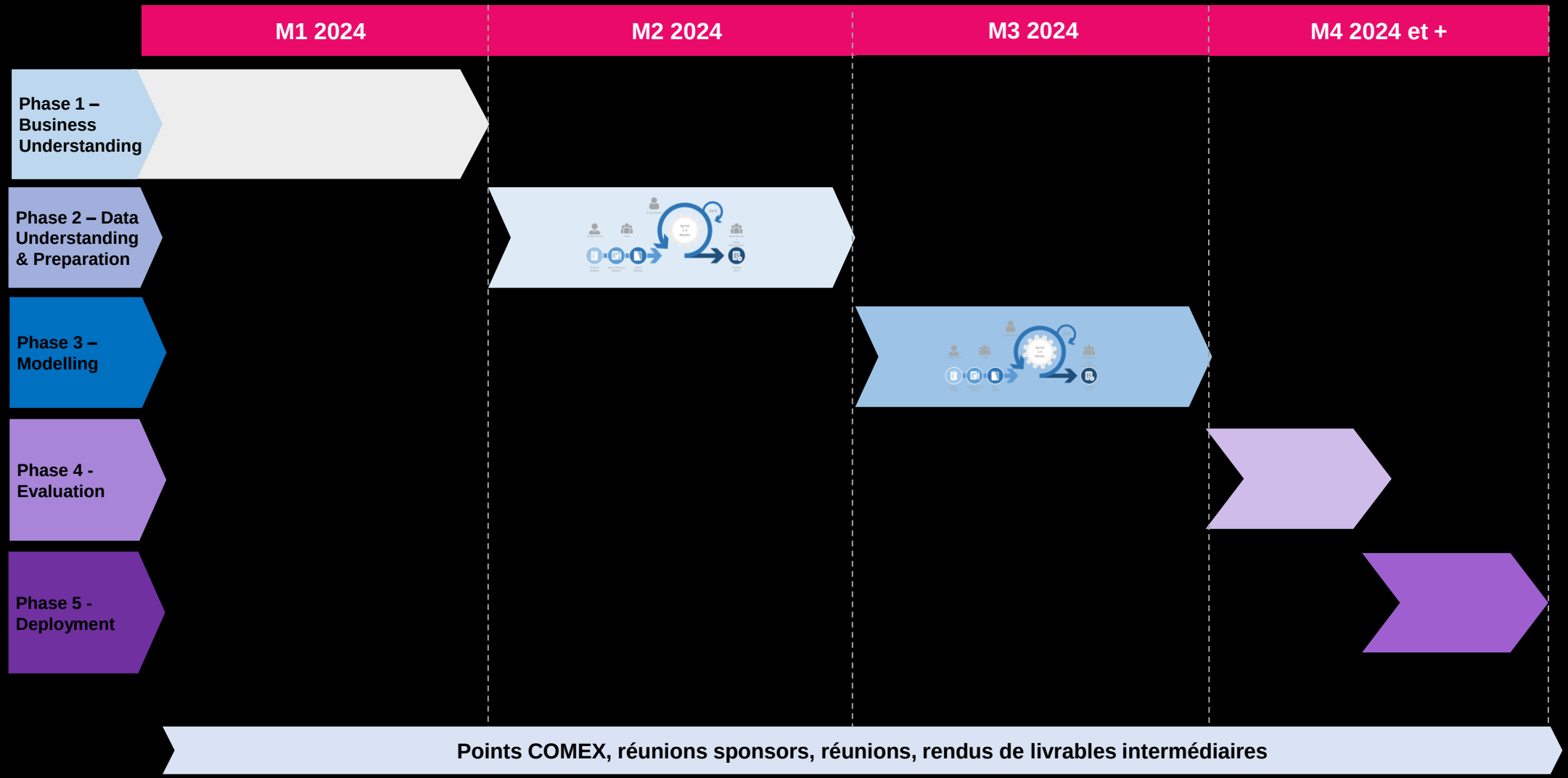
Pour le ministère des armées :

- ➢ Il est nécessaire de prendre des précautions avec ces utilisateurs chinois au vu de la situation géopolitique

Deployment

- ➢ Echéancier sur S2 2024 – Budget proposé : 400 000€ initiaux estimés (contrats en régie)
- ➢ Entretien et enrichissement à prévoir

Planning prévisionnel



Equipe mobilisée

Luka LEINENWEBER – *AMOA Data*

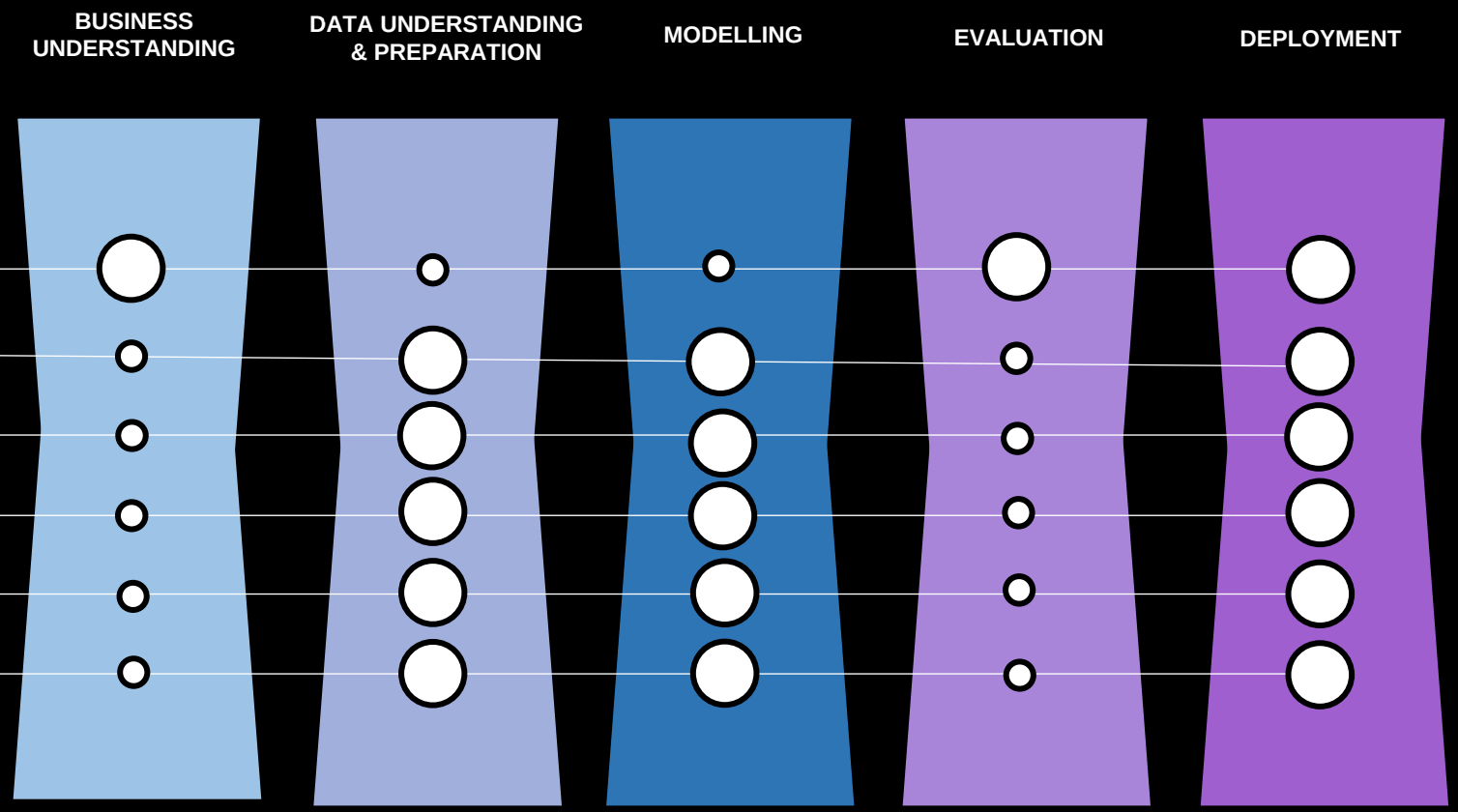
Youssef DOUKKALI – *BI Expert*

Yannick DONGUE – *Data Scientist / PM*

Adam's OUATTARA – *Data Scientist / Tech Lead*

Nathanael KALA – *Data Engineer*

Issa AISSATOU – *Data Scientist*



Merci