PRÉDICTION DES MESSAGES D'URGENCE EN HAÏTI À L'AIDE DU NLP

Projet de Data Science – Capstone

Amee Hashley JEUDY – ameehashleyjeudy@gmail.com
Woodnalie Saviola JOSEPH – woodnaliesjoseph@gmail.com

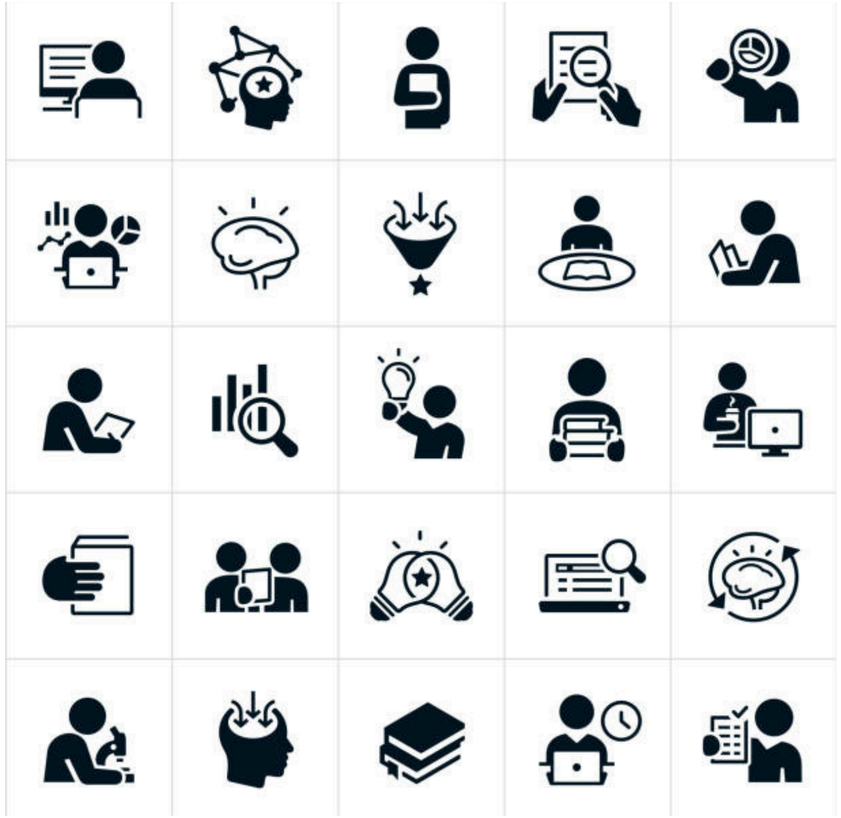
SOMMAIRE

CONTEXTE

Après le séisme de 2010, des milliers de messages ont été envoyés pour demander de l'aide.
L'objectif est d'automatiser le tri entre :

- les messages de type "Request" (aide demandée) et,
- les messages "Info" (information générale)

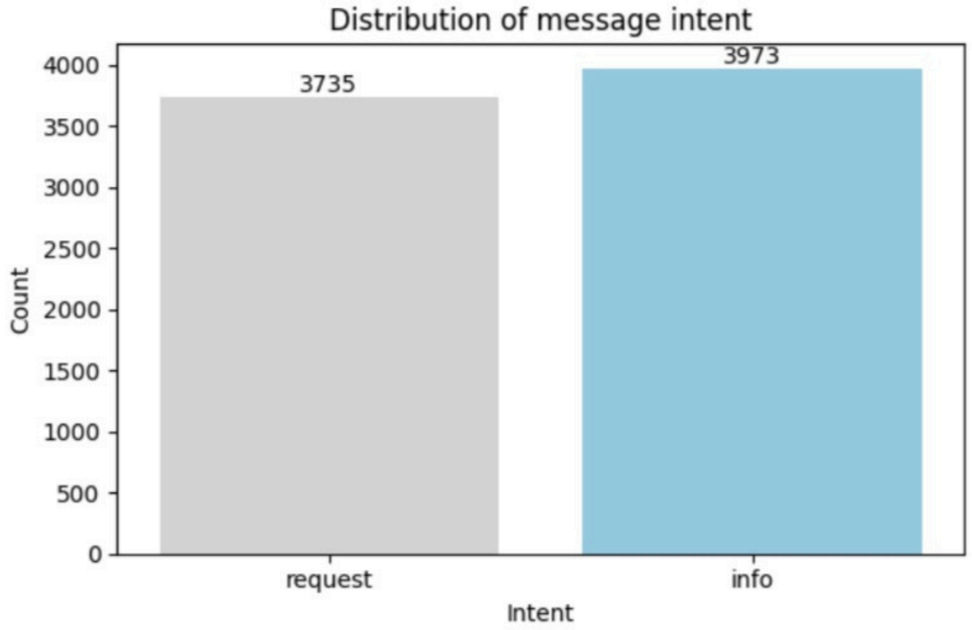
OBJECTIFS



- Construire un modèle de classification NLP binaire.
- Comparer plusieurs algorithmes (Logistic Regression, Naive Bayes, Random Forest).
- Identifier les mots les plus caractéristiques des requêtes.
- Soutenir la prise de décision humanitaire rapide.

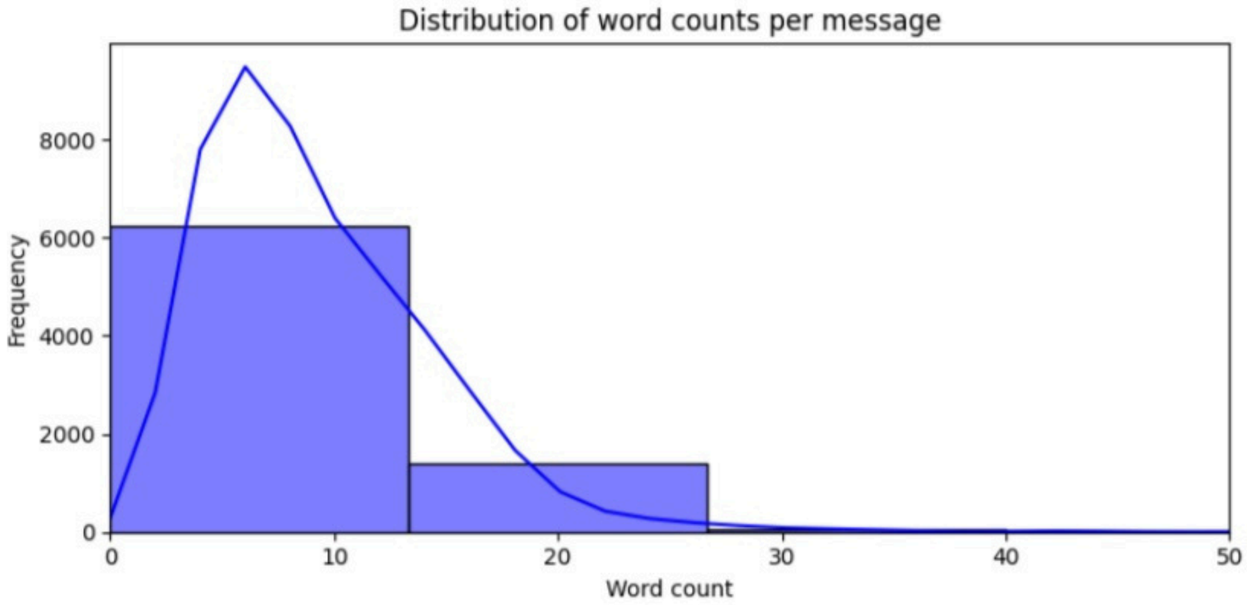
DONNÉES

- Dataset Kaggle “Disaster Response Messages”
- 26 382 messages, dont 15 420 liés à Haïti
- Cible : Request (1) vs Info (0)
- 36 catégories initiales → filtrées pour binaire



Préparation & Nettoyage

- Suppression des doublons et valeurs manquantes
- Nettoyage du texte (URLs, ponctuation, minuscules)
- Tokenisation, suppression des stopwords, lemmatisation
- TF-IDF vectorisation (1 à 3-grammes)



Méthodologie NLP

- Collecte et préparation des données
- Représentation du texte
- Modélisation
- Pipeline :

Message brut → TF-IDF → Modèle ML → Classe prédite

- Évaluation

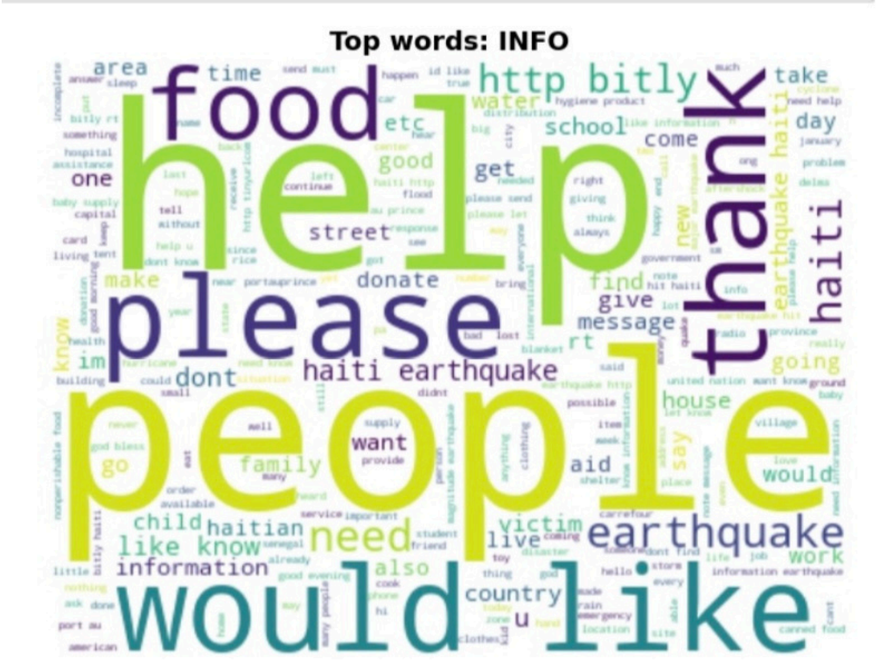
Résultats comparés

Modèle	Accuracy	Macro F1	ROC-AUC
Logistic Regression	0.771	0.771	0.863
Naive Bayes	0.764	0.764	0.852
Random Forest	0.774	0.773	0.859
Logistic Regression (Tuned)	**0.773**	**0.773**	**0.862**

- Modèle retenu: Logistic Regression (Tuned)
  - Stable, rapide, interprétable, transparent,...

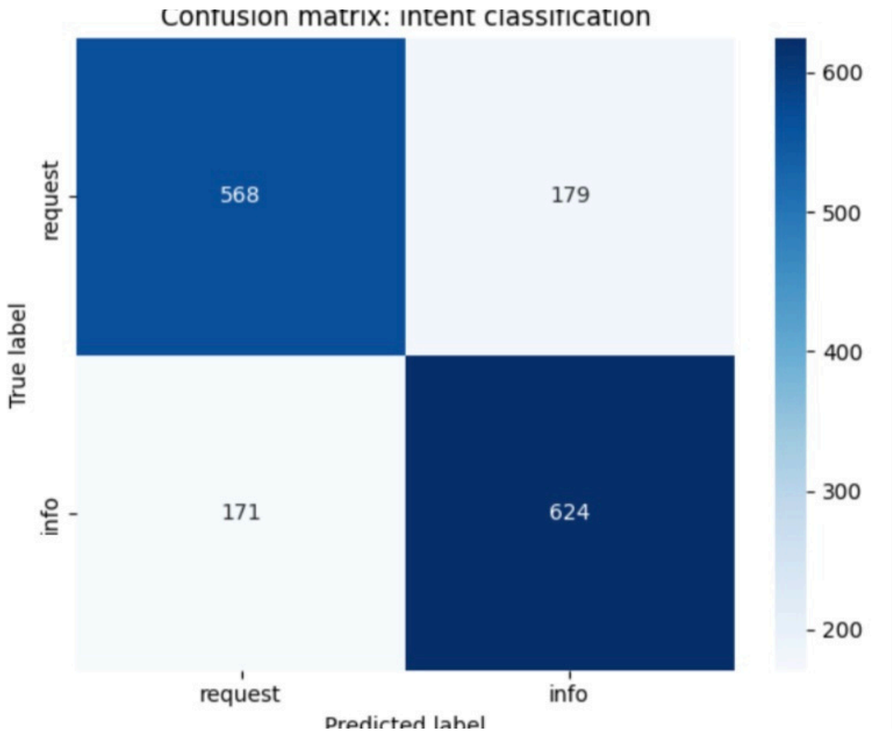
INTERPRÉTATION LINGUISTIQUE

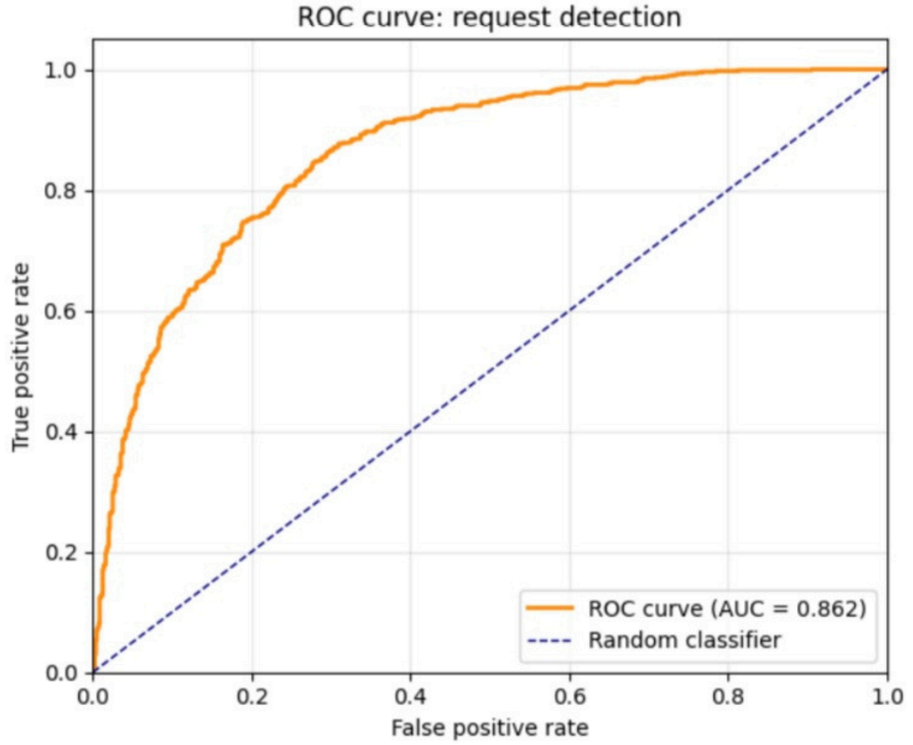
- Mots-clés des requêtes (“Request”) : help, need, hungry, food, water, tent, medicine
- Mots-clés des infos (“Info”) : haiti, service, donate, cyclone, storm, news



PERFORMANCE DU MODÈLE

- Matrice de confusion équilibrée
- ROC-AUC = 0.862 → excellente séparation entre classes
- F1-score = 0.773 → modèle robuste aux déséquilibres





CONCLUSION

Le modèle Logistic Regression classifie correctement 77 % des messages.
Il permet de filtrer automatiquement les requêtes d’aide lors d’une crise.

Perspectives :

- Étendre à plusieurs langues (français, créole).
- Tester BERT ou DistilBERT.
- Intégrer le modèle dans une API pour ONG humanitaires.

MERCI POUR VOTRE ATTENTION

Amee Hashley JEUDY – ameehashleyjeudy@gmail.com
Woodnalie Saviola JOSEPH – woodnaliesjoseph@gmail.com

SolidariTerre