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

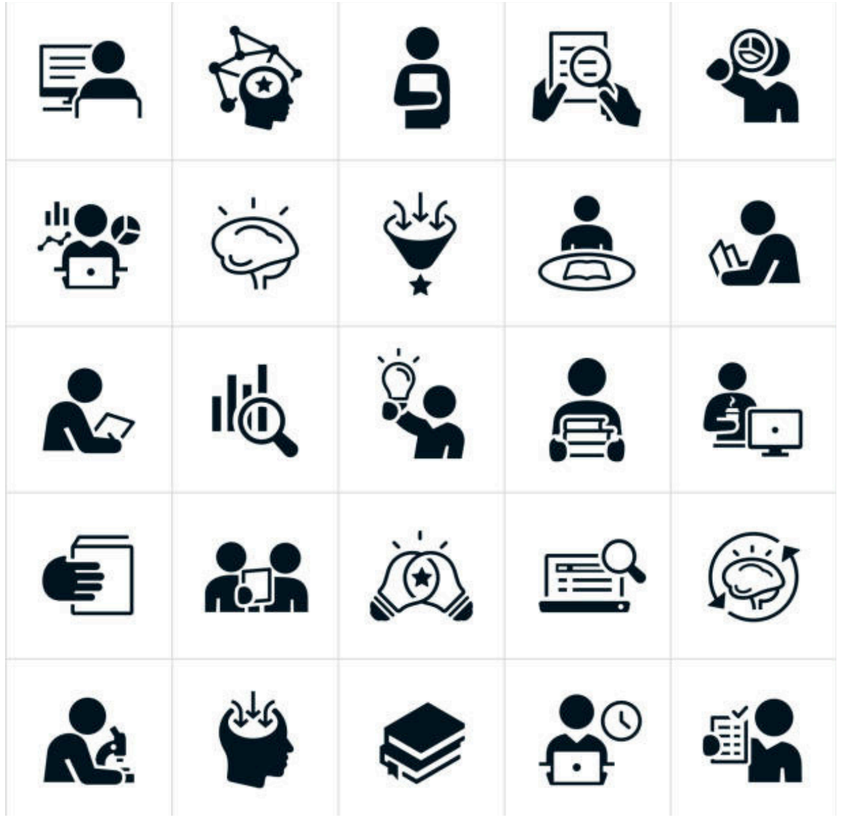
Contexte

Après le séisme dévastateur de 2010 en Haïti, des milliers de messages ont été envoyés par SMS, réseaux sociaux et canaux communautaires.

Le tri manuel de ces messages était :

- Lent : surcharge des équipes
- Inégal : dépendant de la disponibilité humaine
- Critique : chaque retard pouvait aggraver la situation des victimes

Objectifs



Le projet vise à développer un système NLP interprétable pour classer les messages en deux catégories :

- request
- info

Les objectifs spécifiques sont :

- Accélérer le tri des messages en temps réel
- Maintenir la transparence des décisions algorithmiques
- Réduire la charge des équipes de terrain

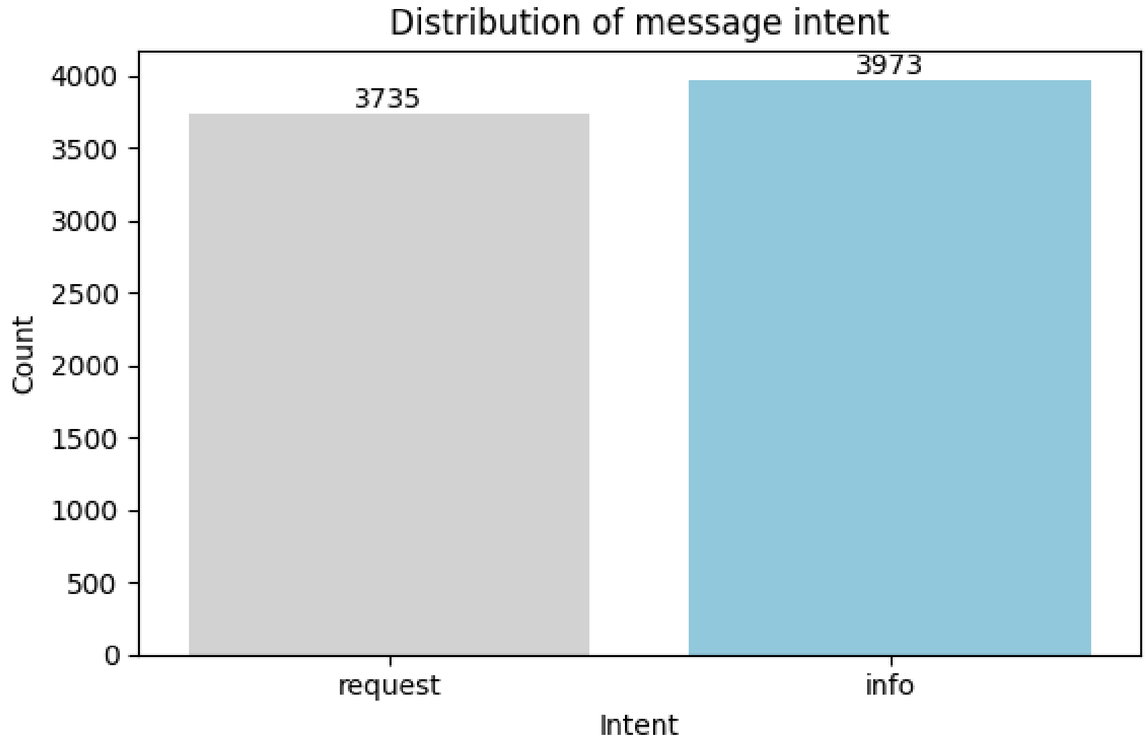
Données

- Données issues du jeu Figure Eight Disaster Response (Kaggle), annotées manuellement pour 36 catégories humanitaires.
- 26 382 entrées
- Filtrage contextuel :
  - Seuls les messages liés au séisme de 2010 en Haïti ont été conservés
  - Genre = "direct" (messages de première main)
  - Résultat : 15 420 messages pertinents
- Cible : Request (1) vs Info (0)

Préparation & nettoyage

- Conversion en minuscules
- Suppression des URLs, mentions, hashtags, ponctuation
- Tokenisation et suppression des stopwords
- Lemmatisation (ex. "needs" → "need")
- Suppression des doublons

- Création de la variable cible :
  - request si request = 1
  - info sinon
- Distribution +/- équilibrée :
  - 51.5% info,
  - 48.5% request



Modélisation NLP

- Modèles testés
  - Régression Logistique (linéaire, interprétable)
  - Naive Bayes (probabiliste, rapide)
  - Random Forest (non-linéaire, robuste)

- Pipeline NLP
  - TF-IDF vectorisation avec trigrams
  - Filtrage du vocabulaire (min_df=2, max_df=0.95)
  - Stopwords supprimés, pondération sublinéaire

- Tuning
  - Régression Logistique optimisée par validation croisée (GridSearchCV)
  - Critère : F1-macro pour équilibrer les classes

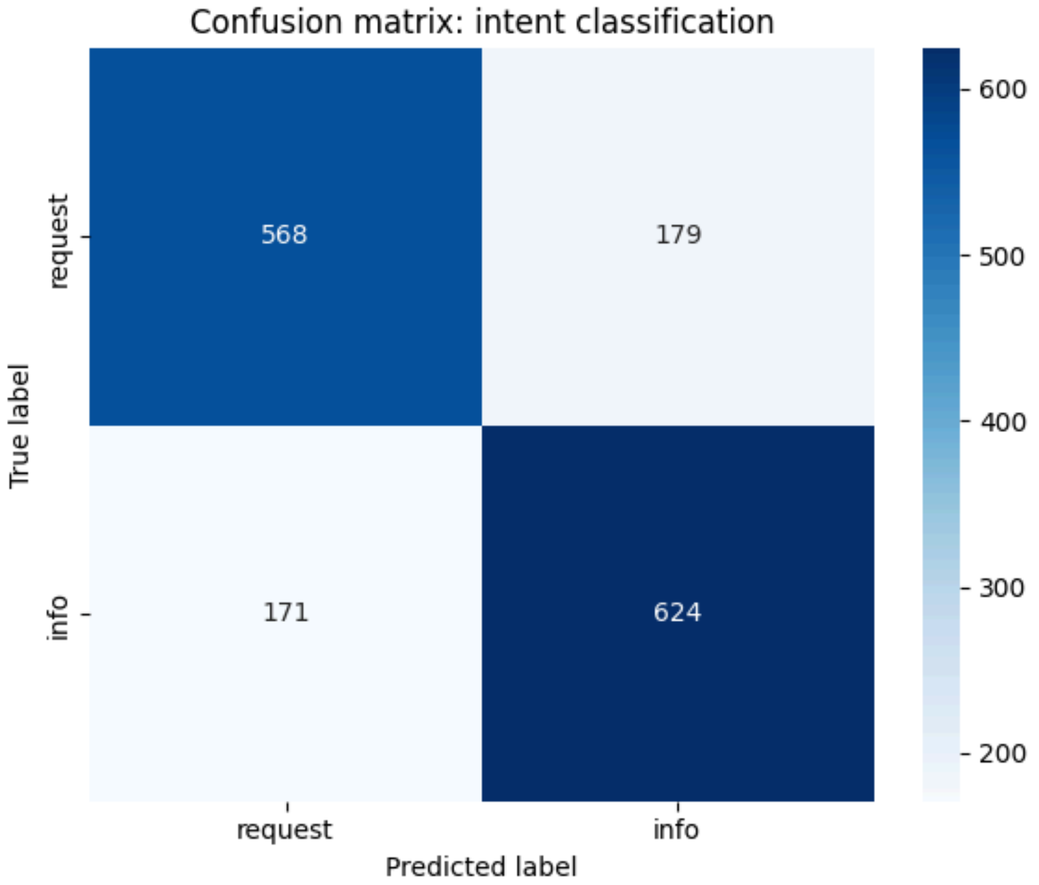
Résultats comparés

Model	Accuracy	Macro F1	ROC-AUC
Logistic Regression	0.77108	0.77085	0.86270
Naive Bayes	0.76394	0.76394	0.85151
Random Forest	0.77367	0.77327	0.85877
Logistic Regression (Tuned)	0.77302	0.77272	0.86245

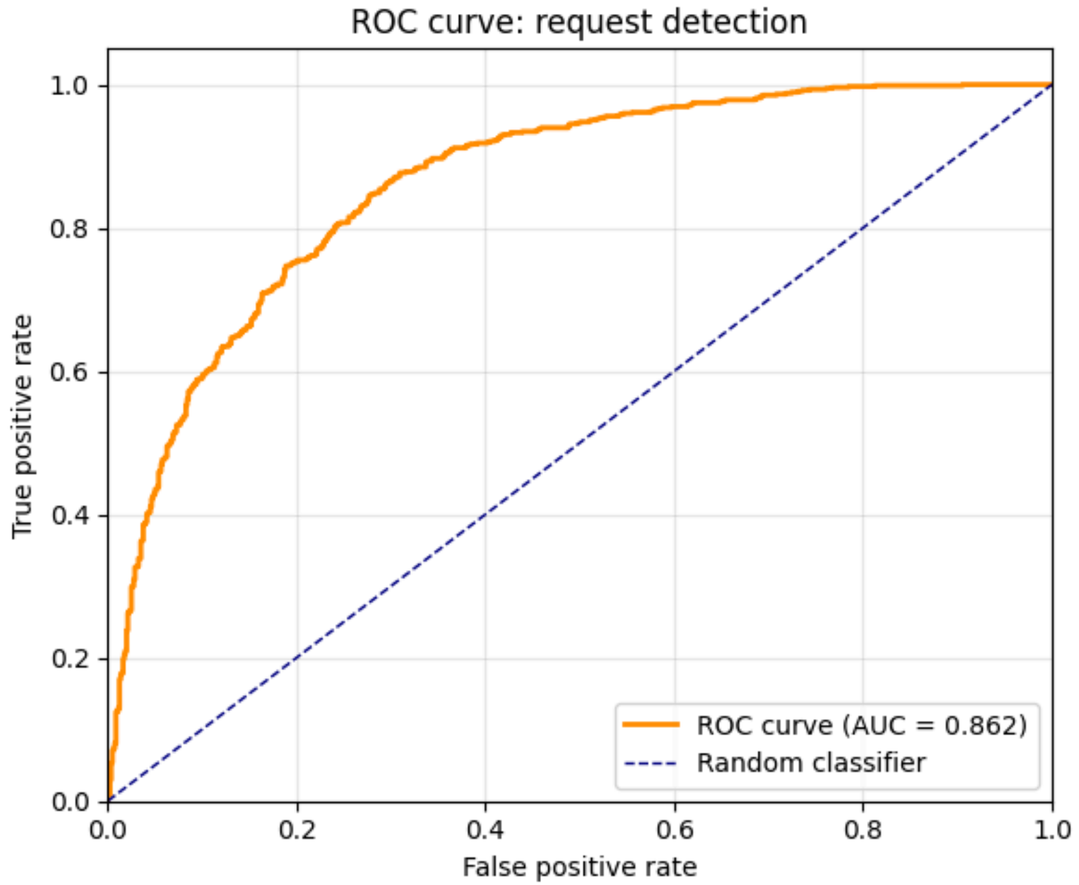
- Le modèle final est la Régression Logistique optimisée, car il est à la fois performant et interprétable

Performance du modèle

- Matrice de confusion équilibrée



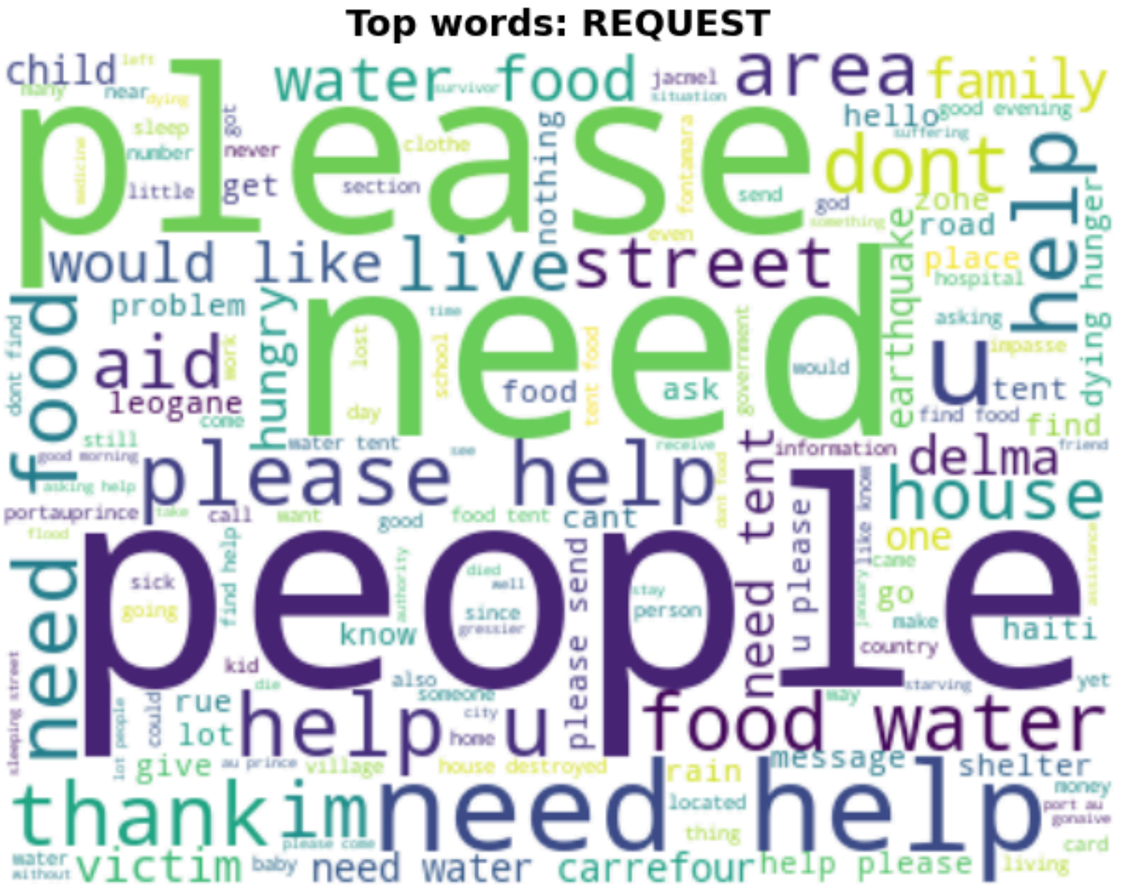
- AUC = 0.862 → excellente séparation entre les classes



Interprétation linguistique

- Modèle interprétable
  - La régression logistique permet d'identifier les mots et expressions qui influencent directement la prédiction

- Signaux forts pour les **demandes urgentes (request)**
  - Mots-clés fréquents :
    - "help", "need", "water", "food", "tent", "baby", "send", "house", "people"
  - Expressions typiques capturées par les trigrams
    - "need water now", "please help us", "no food here"



Interprétation linguistique

- Contraste lexical clair avec les messages **info**
  - Vocabulaire plus général
    - “haiti”, “earthquake”, “information”, “know”, “please”, “thank”
  - Moins d’actions concrètes, plus de contexte ou de commentaires

Conclusion

Le projet démontre qu'un modèle NLP simple mais bien calibré peut :

- Identifier efficacement les demandes urgentes dans un contexte humanitaire
- Offrir des performances solides (F1 = 0.77, AUC = 0.86)
- Rester interprétable et donc utilisable par les acteurs de terrain

Recommandations

Pour les équipes humanitaires

- Déployer le modèle dans les outils de tri terrain (WhatsApp bots): Le modèle est léger, rapide et interprétable; idéal pour les environnements à faible ressources
- Former les agents à l'interprétation des sorties du modèle: Les mots-clés comme "need water" ou "tent please" sont directement exploitables pour l'action
- Utiliser le modèle comme outil d'appui, non de remplacement: Maintenir le jugement humain pour les cas ambigus ou sensibles

Pour les partenaires technologiques

- Intégrer le modèle dans des flux multilingues: Étendre la couverture aux messages en créole et en français pour éviter les biais de traduction
- Ajouter une classification multi-label: Identifier les types d'aide demandés (eau, nourriture, abri...) pour affiner la réponse logistique
- Mettre en place un suivi en temps réel des performances: Ajuster le modèle selon les retours du terrain et les évolutions du langage

MERCI DE VOTRE ATTENTION

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

SolidariTerre