Grégory LANG - 19/01/2020

Etudiant Ingénieur Machine Learning OpenClassrooms

DÉTECTION DE FAKE NEWS À L'AIDE DU MODÈLE CAMEMBERT

(BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS)

PROJET IML 7

SOMMAIRE

LE CONTEXTE

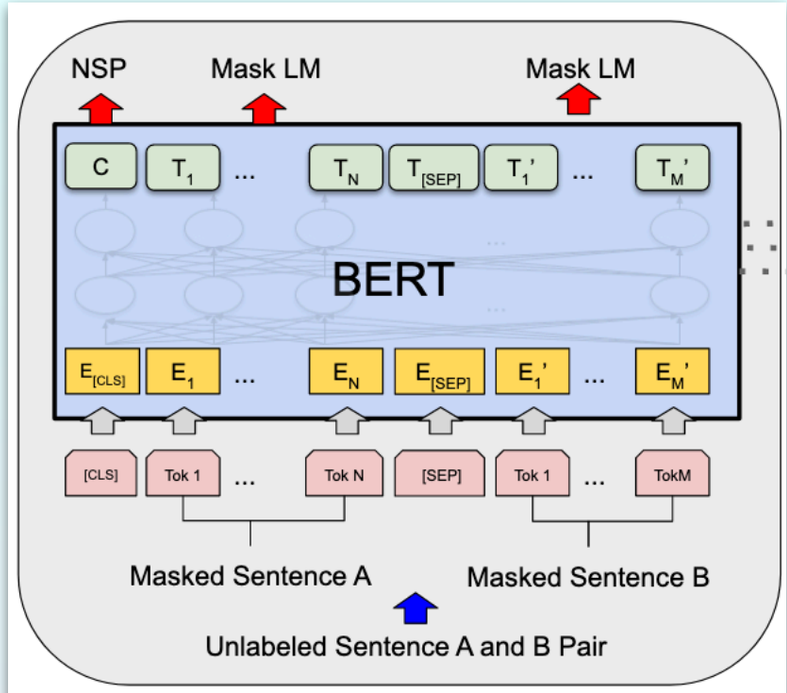
- Développer une preuve de concept
- Effectuer **une veille thématique**
  - choix : NLP : domaine des classifiers de texte
- Mettre en pratique un nouvel algorithme de façon autonome
  - Choix fait :
    - de réutiliser du code
    - avec un dataset original

LA PROBLÉMATIQUE ET SON INTERPRÉTATION

- Problématique proposée : détection de fake news dans les articles de journaux en ligne
- Détection à travers le contenu des articles
  - texte du titre + body
- Modèle choisi : camemBERT : nouveau modèle pré-entrainé sur un corpus français.
  - Transfert learning : classer les articles
    - « fake » ou « fiable »
- Le comparer à une baseline

LE MODÈLE CAMEMBERT

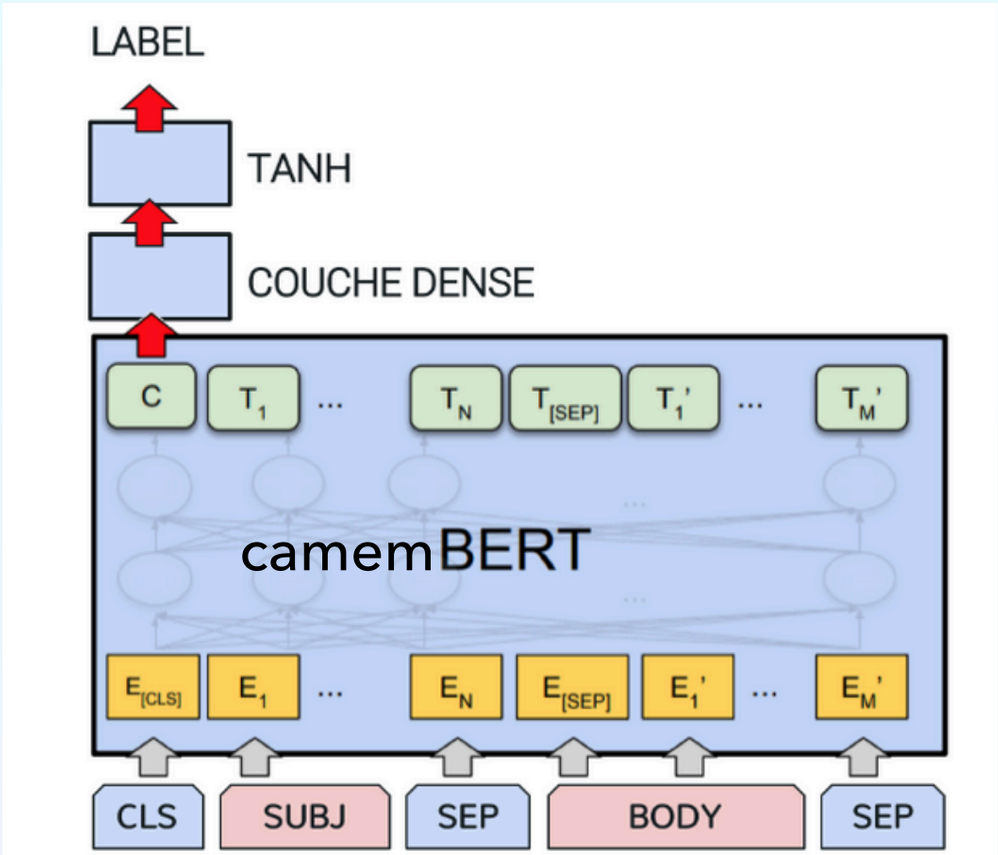
- Type RoBERTa issu de BERT
  - pré-entrainé sur le sous-corpus français d'OSCAR
    - 46,896,036,417 mots (282 Go)
- Architecture des modèles identiques
  - BERT == RoBERTa == camemBERT
    - Modèles de type Transformers : Encoder-Decoder with Attention
      - 110 Millions de paramètres (12 layers / 768 hidden dimensions /12 attention heads)
- Mais différents types d'entrainements / hyper-paramètres
  - RoBERTa = BERT entrainé plus longtemps avec plus de données (entre autres)
  - camemBERT = RoBERTa entrainé avec un tokenizer different…
- Traitement parallèle possible contrairement aux RNN « Encoder-Decoder with Attention »



Pre-training

TRANSFERT LEARNING AVEC CAMEMBERT

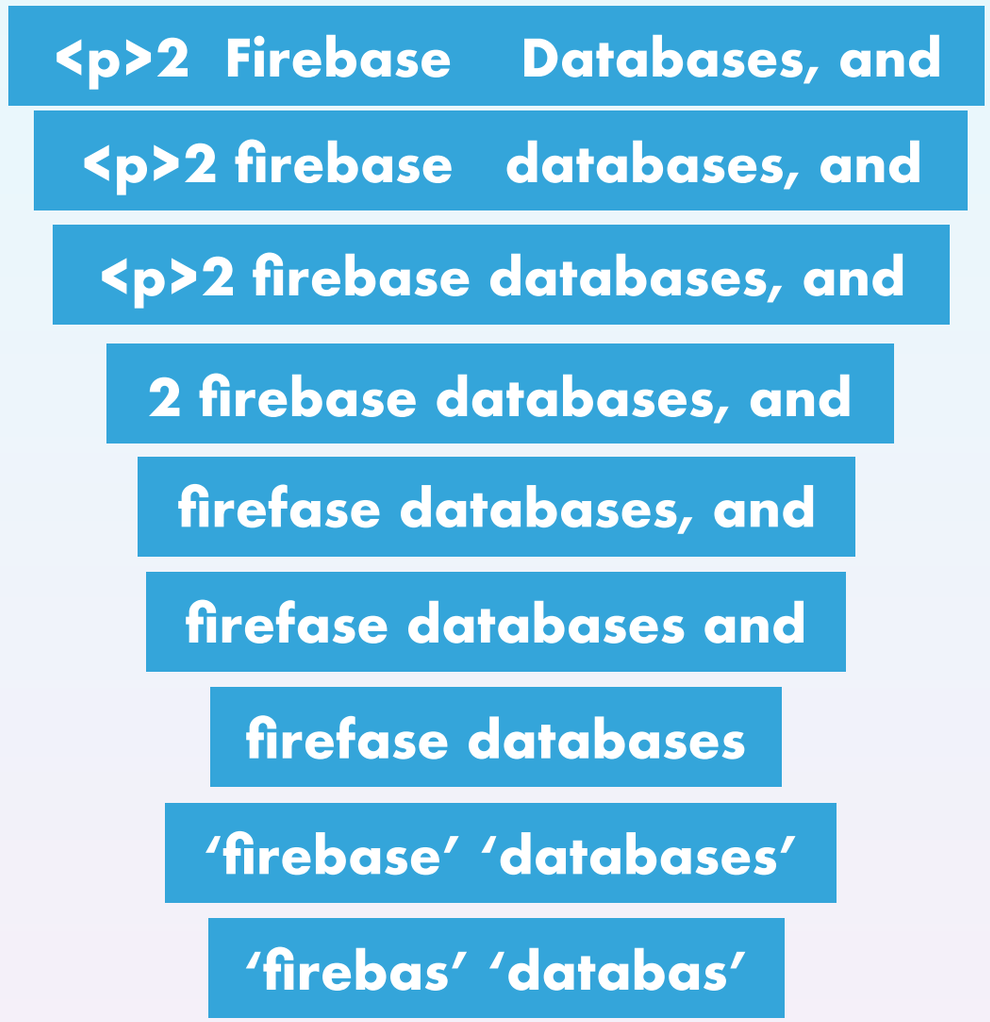
- Modèle utilisé en mode classification
  - Ajout en sortie de camemBERT :
    - en entrée : **n** états cachés du premier token [CLS] (sortie de camemBERT)
    - *Couche intermédiaire « dropout »*
    - Couche **dense** avec **n** entrées / **n** sorties
      - fonction d'activation : **tanh**
    - *Couche intermédiaire « dropout »*
    - 1 dernière couche linéaire : **n** entrées / **1** seule sortie



- Fonction de perte : erreur quadratique

MODÈLE BASELINE

- Modèle régression logistique
  - Features : TF-IDF
  - Avec préparation du texte pour les features :
    - Agrégation du « Title » et du « Body »
    - Mise en minuscule
    - Suppression caractères « espace» « tabulation » non désirés
    - Suppression tag HTML
    - Suppression des nombres
    - Suppression de la ponctuation
    - Suppression des mots les plus utilisés (stopwords)
    - Création de tokens
    - Extraction de la racine (Stemming)



CRÉATION DU DATASET

- Création de différents « spiders » Scrapy pour :
  - 20 minutes
  - Le Gorafi
  - buzzbeed.com
  - NordPresse.be

PARCOURANT LES PAGES

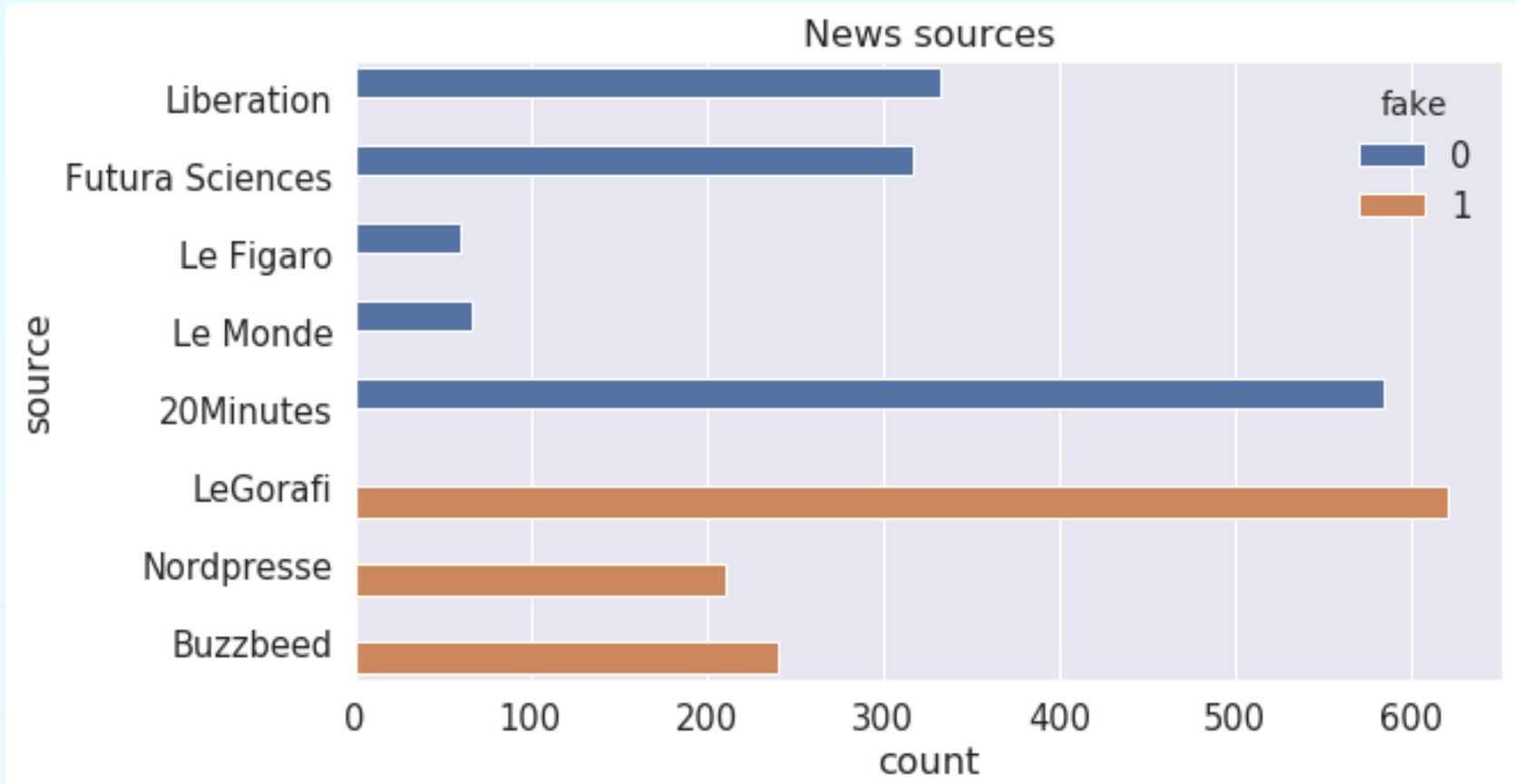
- Utilisation du GitHub **gbolmier/newspaper-crawler** pour
  - *An autonomous French newspaper crawler based on Scrapy framework*
  - Le Monde (avec correction)
  - Le Figaro
  - Libération
  - Futura Sciences

PARCOURANT LES FLUX RSS

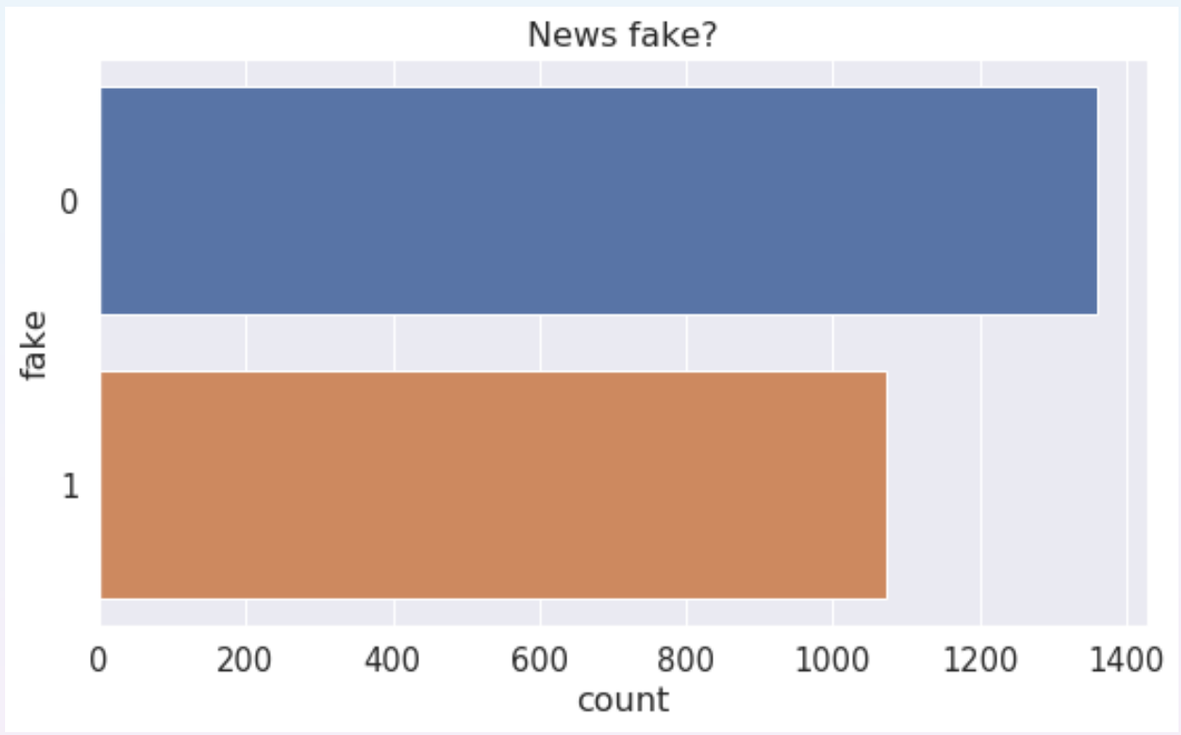
EXPLORATION

- 2432 articles

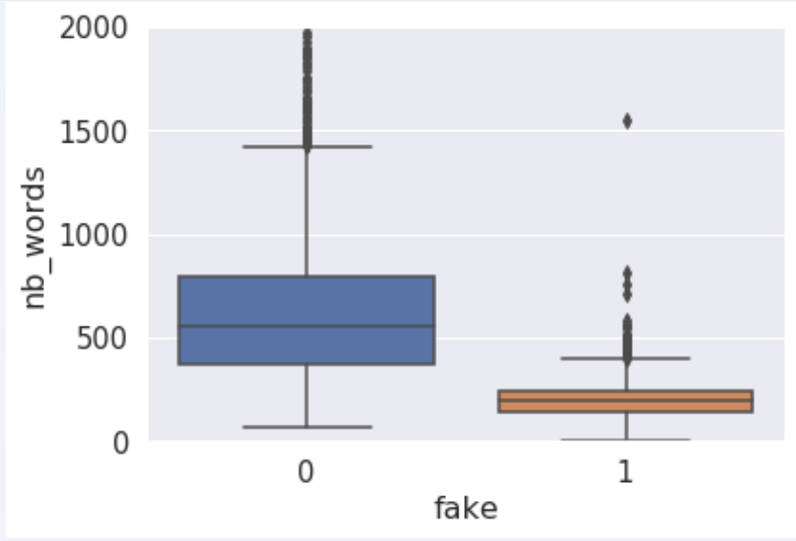
20 MINUTES & LEGORAFI PLUS REPRÉSENTÉS



- 45% fake



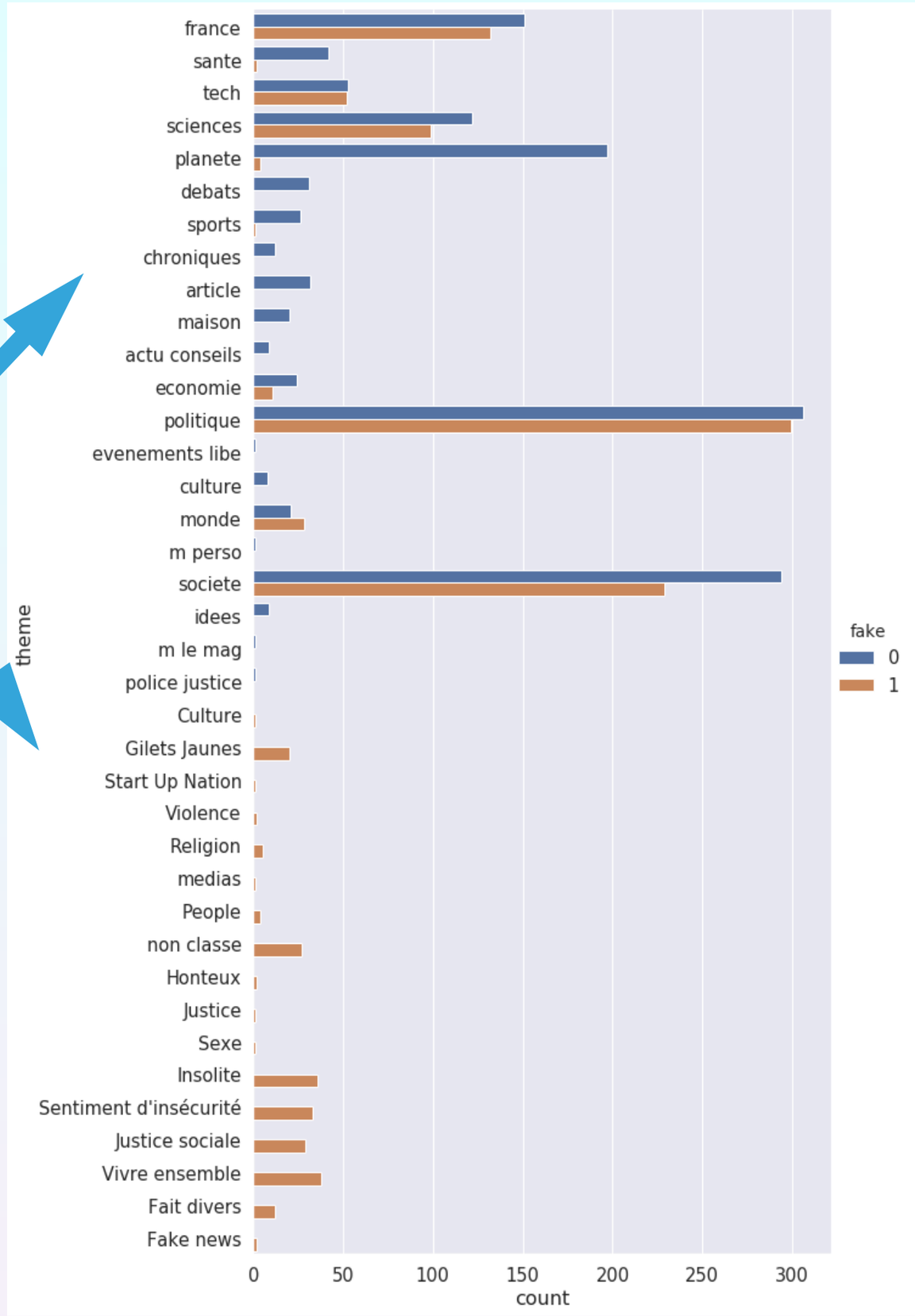
FAKE NEWS : 60% DE MOTS EN MOINS



EXPLORATION DU DATASET

- 3/4 des news dans un thème équilibré
  - France
  - technologies
  - sciences
  - politique
  - société
  - monde

NOMBREUX THEMES PEU REPRÉSENTÉS

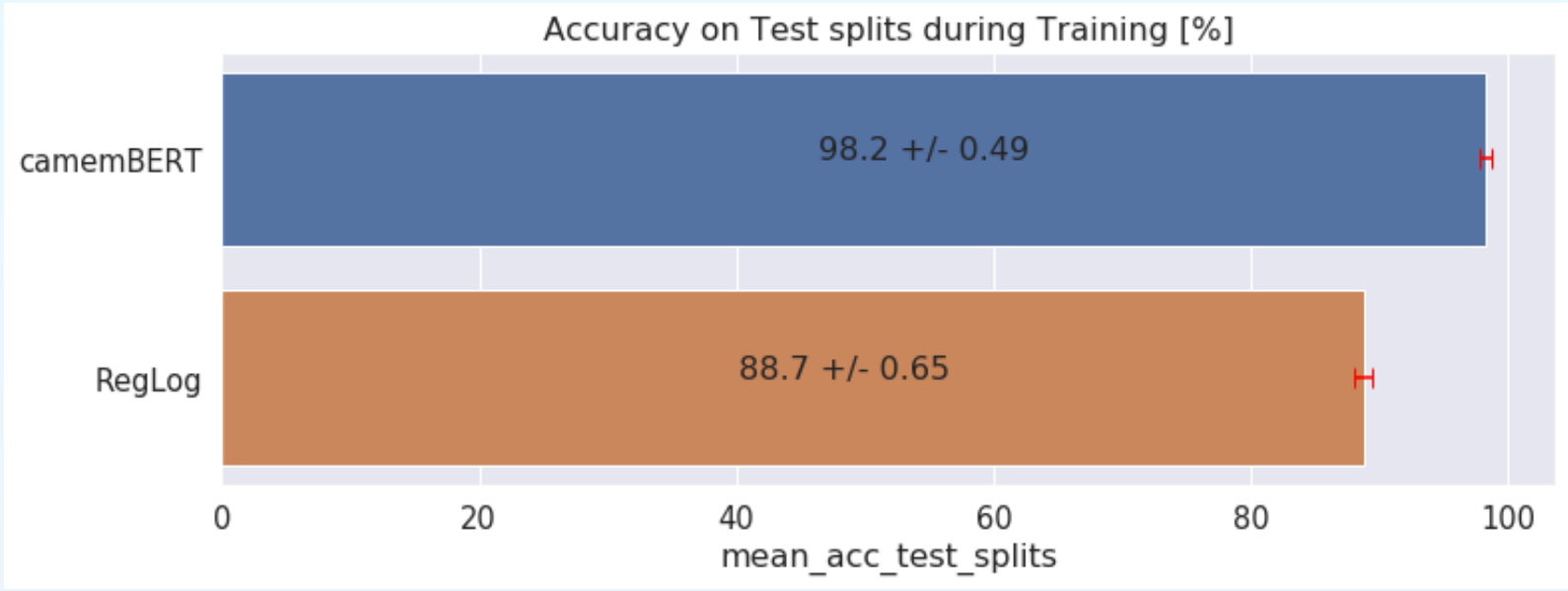


RÉSULTAT DES MODÈLES

70% TRAIN / 30% TEST

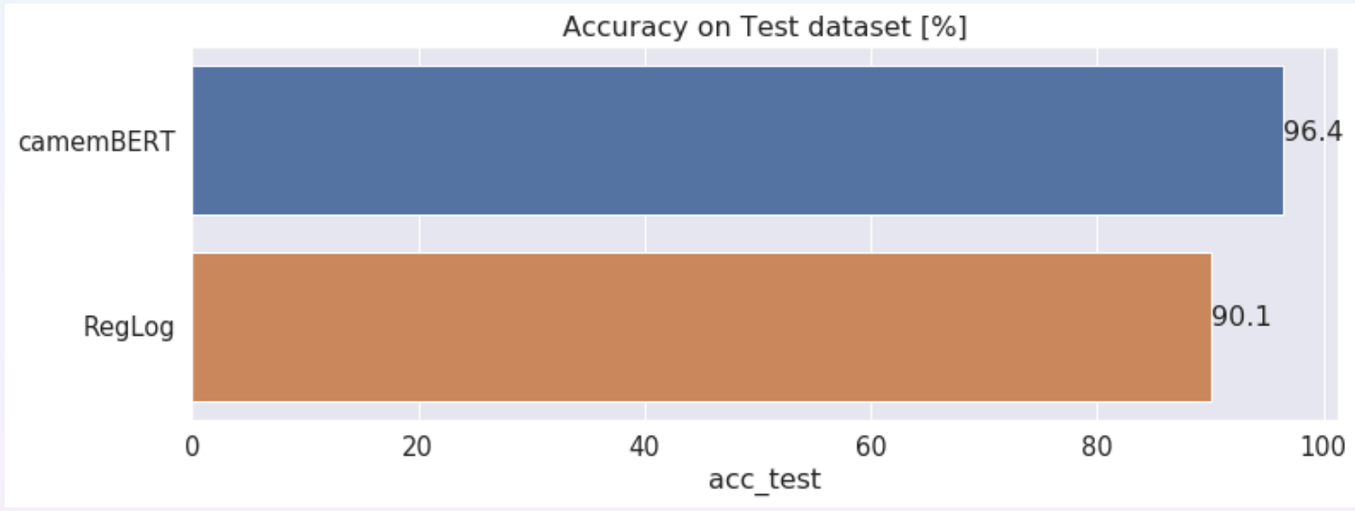
- Bonne accuracy de camemBERT sur les **test splits** > **98%**
  - Validation croisée sur 5 splits avec StratifiedShuffleSplit
  - Test splits sur 30% du jeu de « train »

FAIBLE MARGE D'ERREUR



camemBERT +10% D'ACCURACY SUR LES TEST SPLITS

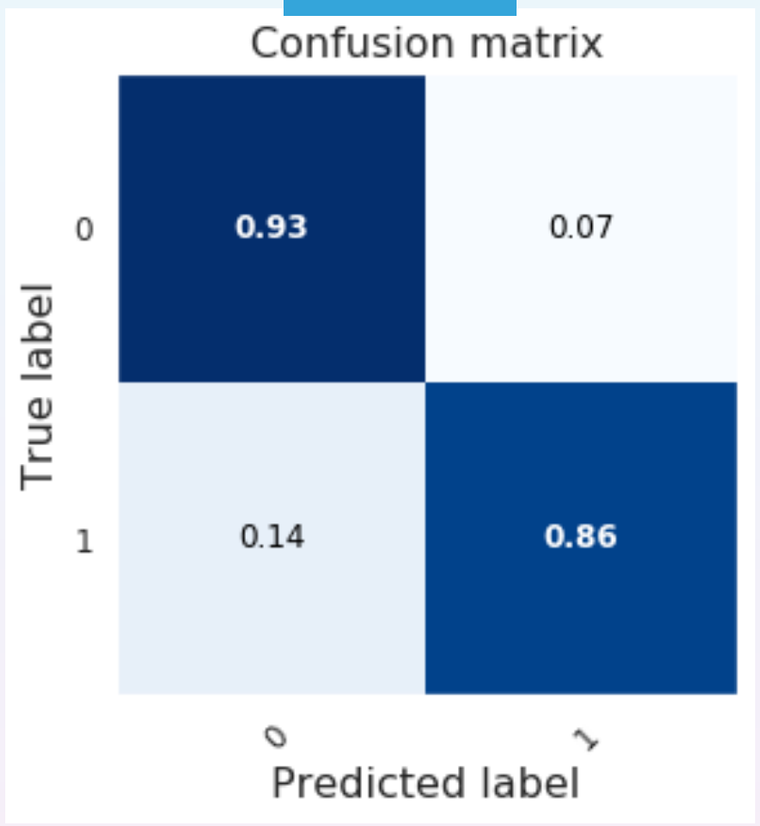
- Bonne accuracy de camemBERT sur jeu de « test » : > 96%
  - => bonne généralisation attendue.



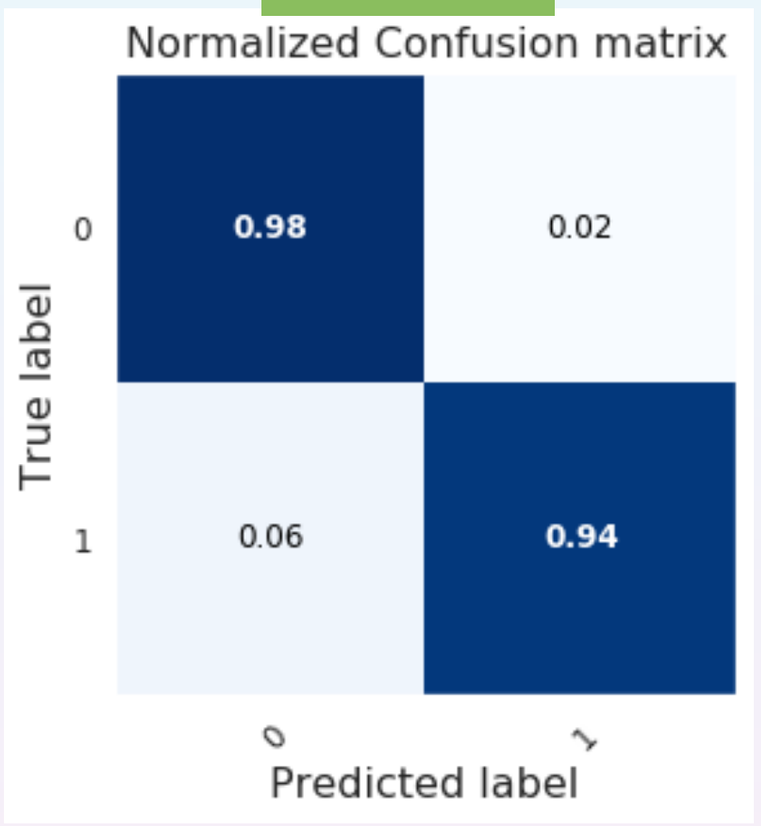
RÉSULTAT DES MODÈLES : MATRICES DE CONFUSION

- camemBERT assez équilibré
- Baseline moins efficace pour détecter les fake news
  - camemBERT 94% / Baseline 86% de vrais positifs (fake news bien prédites)
- - de « fake » prédite « fiable » avec camemBERT (6%) / Baseline (14%)

BASELINE



CAMEMBERT

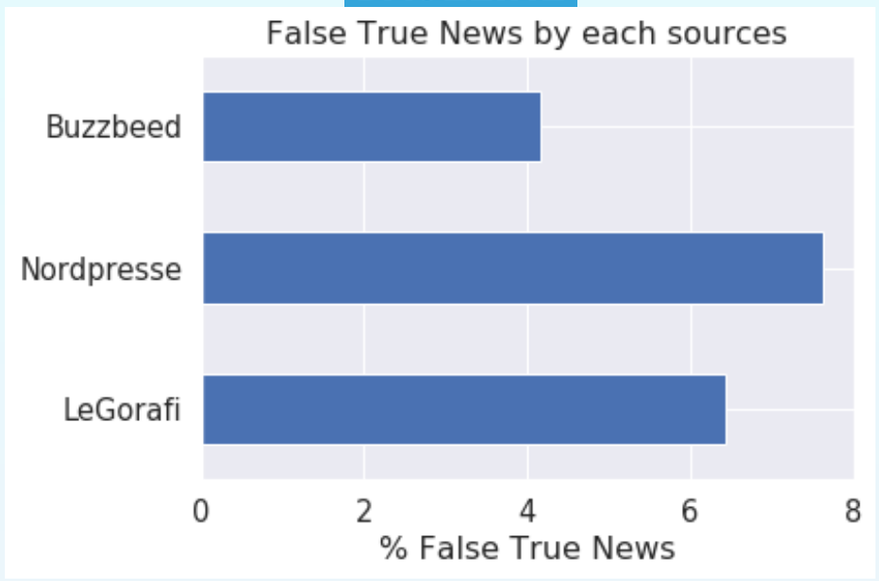


RÉSULTATS DES MODÈLES : FAUX NÉGATIFS

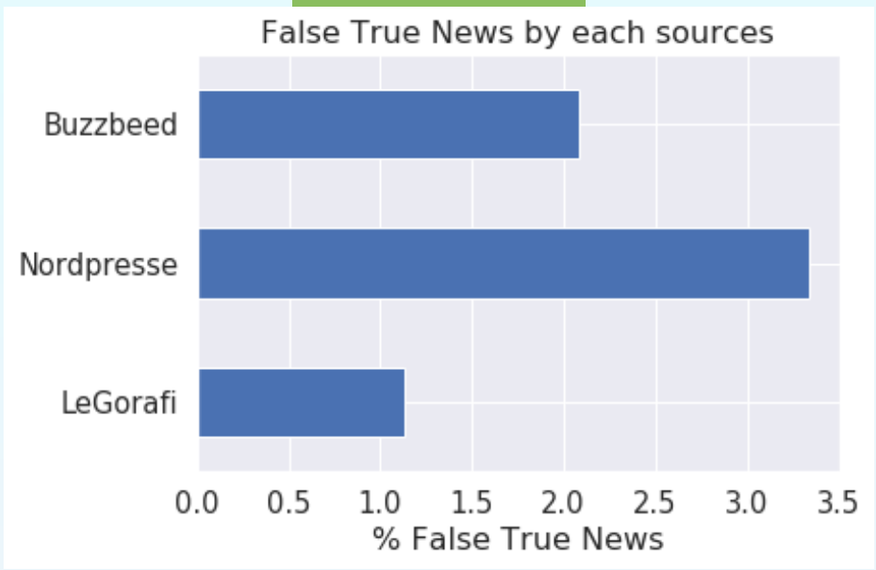
FAKE PRÉDITE FIABLE

- Plus de faux négatifs pour Nordpresse

BASELINE

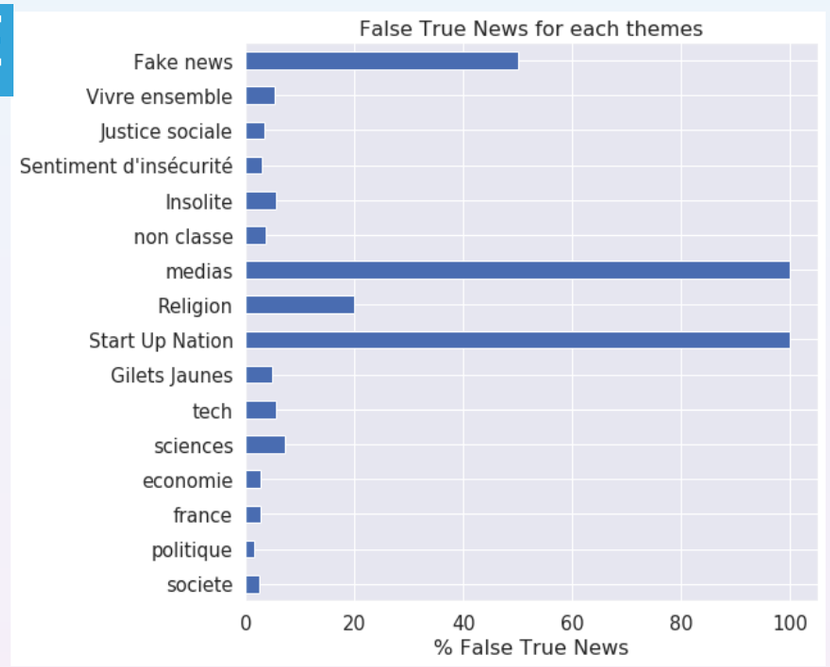


CAMEMBERT

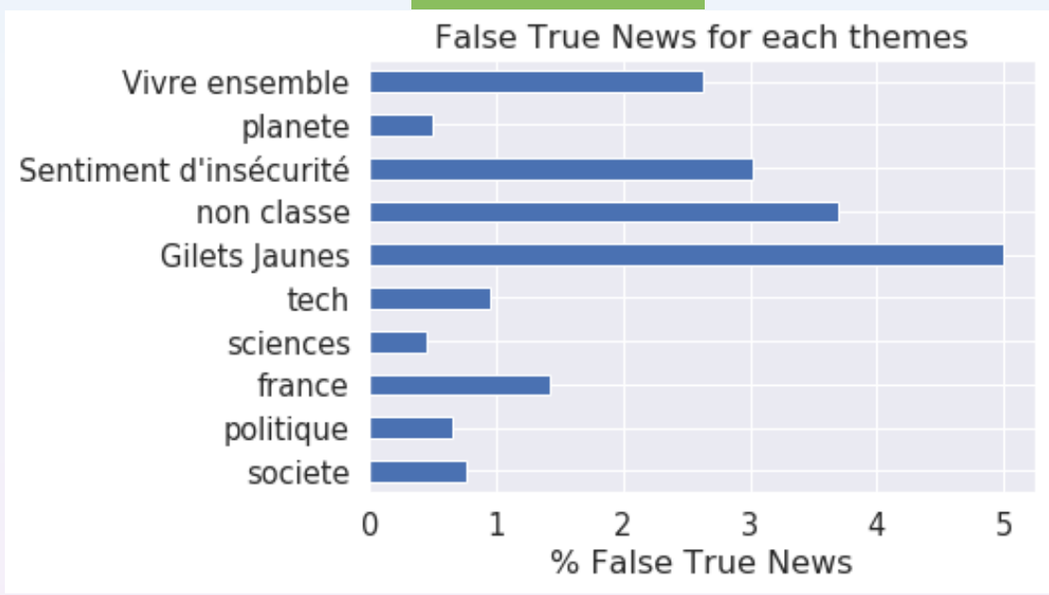


- Thèmes différents entre 2 modèles

BASELINE



CAMEMBERT

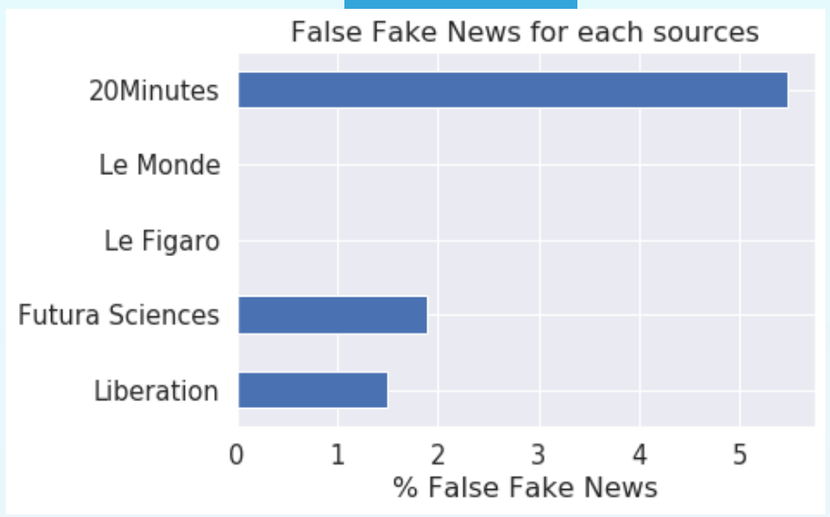


RÉSULTATS DES MODÈLES : FAUX POSITIFS

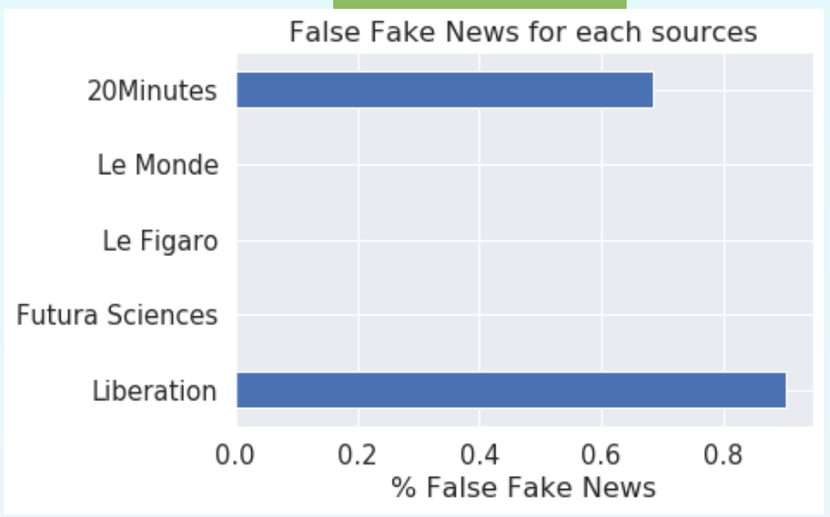
FIABLE PRÉDITE FAKE

- pas de faux positif pour Le Monde et le Figaro

BASELINE

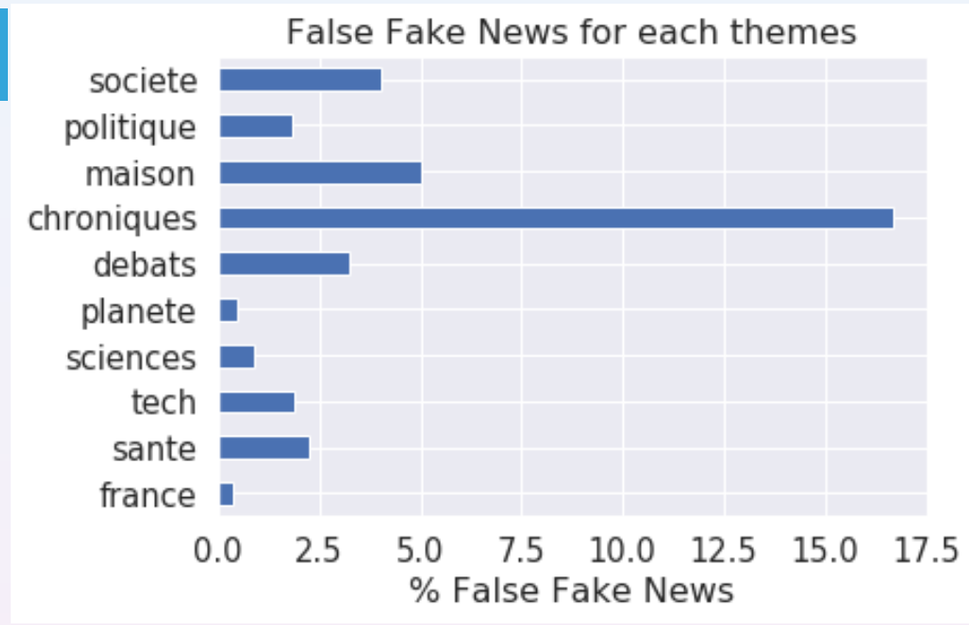


CAMEMBERT

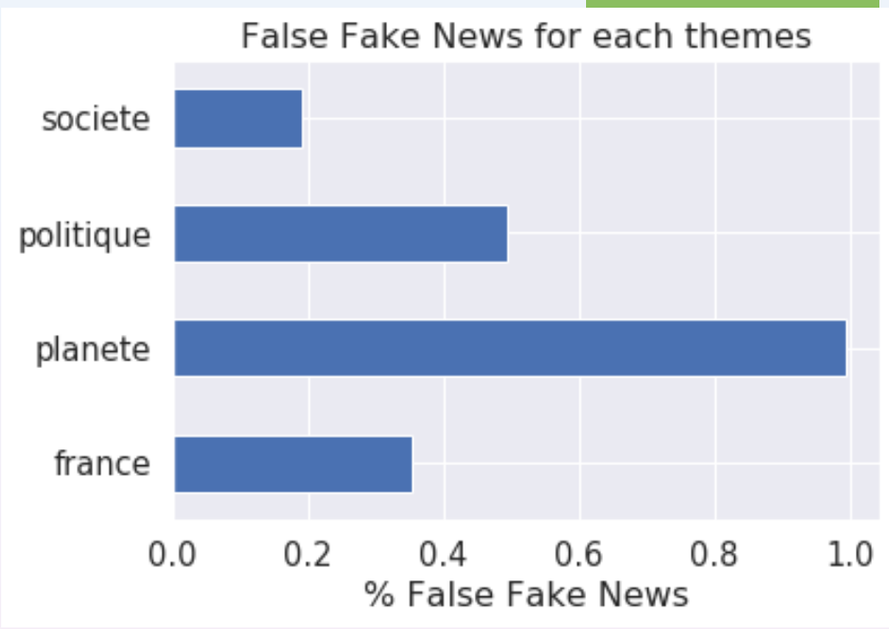


- thème chroniques le + difficile pour baseline

BASELINE



CAMEMBERT



CONCLUSIONS

- camemBERT : un très bon classifier
  - accuracy +10% / baseline
- Mais modèle lourd à entrainer
  - GPU et espace disque important (> 40 Go voir bcp plus)
  - timing train : 50 min environ sur Google Colab env. GPU

AXES D'AMELIORATION

- Tester le modèle avec d'autres sources inconnues du modèle
- Lancer l'entrainement sur plus d'« epocs »
  - en 1 fois plutôt que actuellement 10 fois
- Fake news assez évidentes en générale
  - plutôt parodiques
  - labels intermédiaires : fiable - plutôt fiable - plutôt fake - fake ?
- plus de news en améliorant le scraping des journaux par pages (plutôt que RSS)
  - attention : Le Monde interdit cela
- Créer un site web (API) de détection de la fiabilité d'un article