
Data Science Bootcamp

# La classification de textes en entreprise

Sebastian Paulo et Thierry Linde

en coopération avec

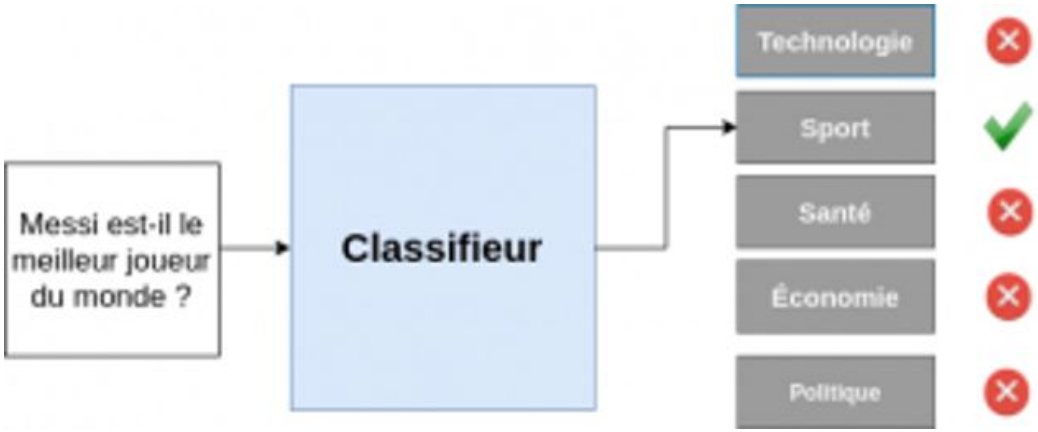
Holi

# La classification de textes peut-elle faciliter la gestion d’informations en entreprise ?

# La classification de textes - un domaine clé du NLP

## Contexte holi.io

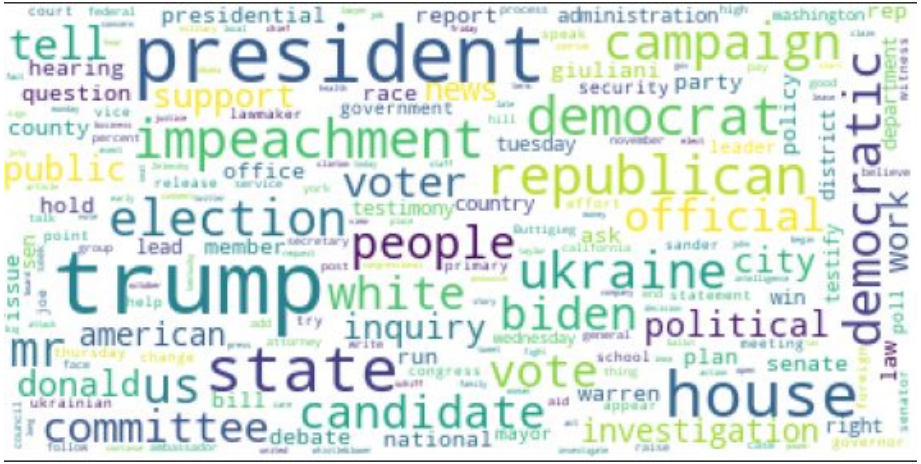
- Une méthode supervisée qui relie des textes à des catégories

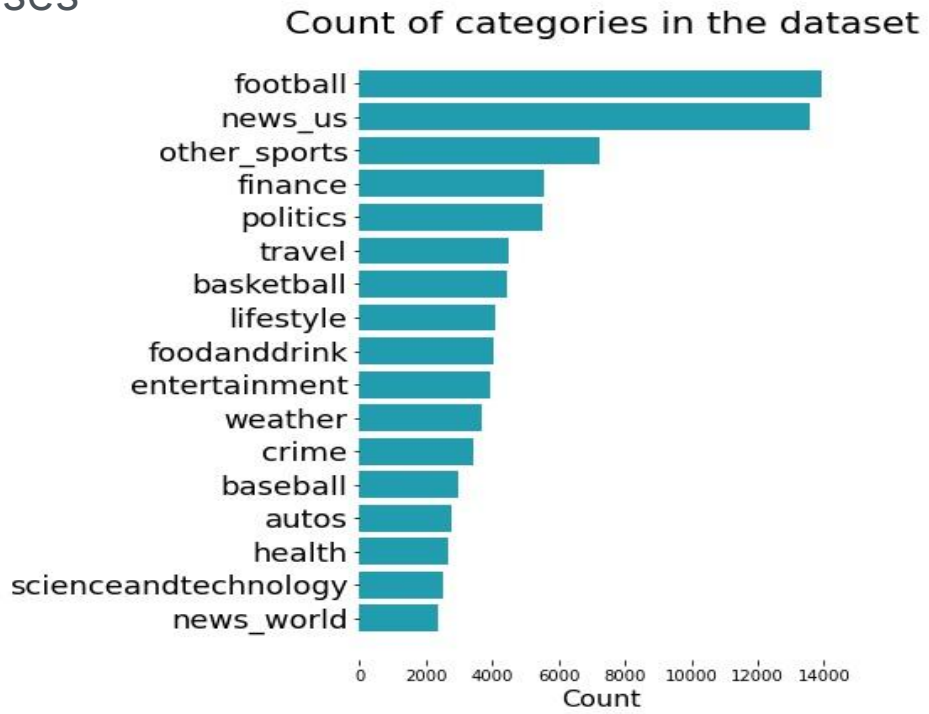



- Classification de textes en entreprise permettant un meilleur flux d'informations:
  - identification de contenu
  - recommandations ciblées

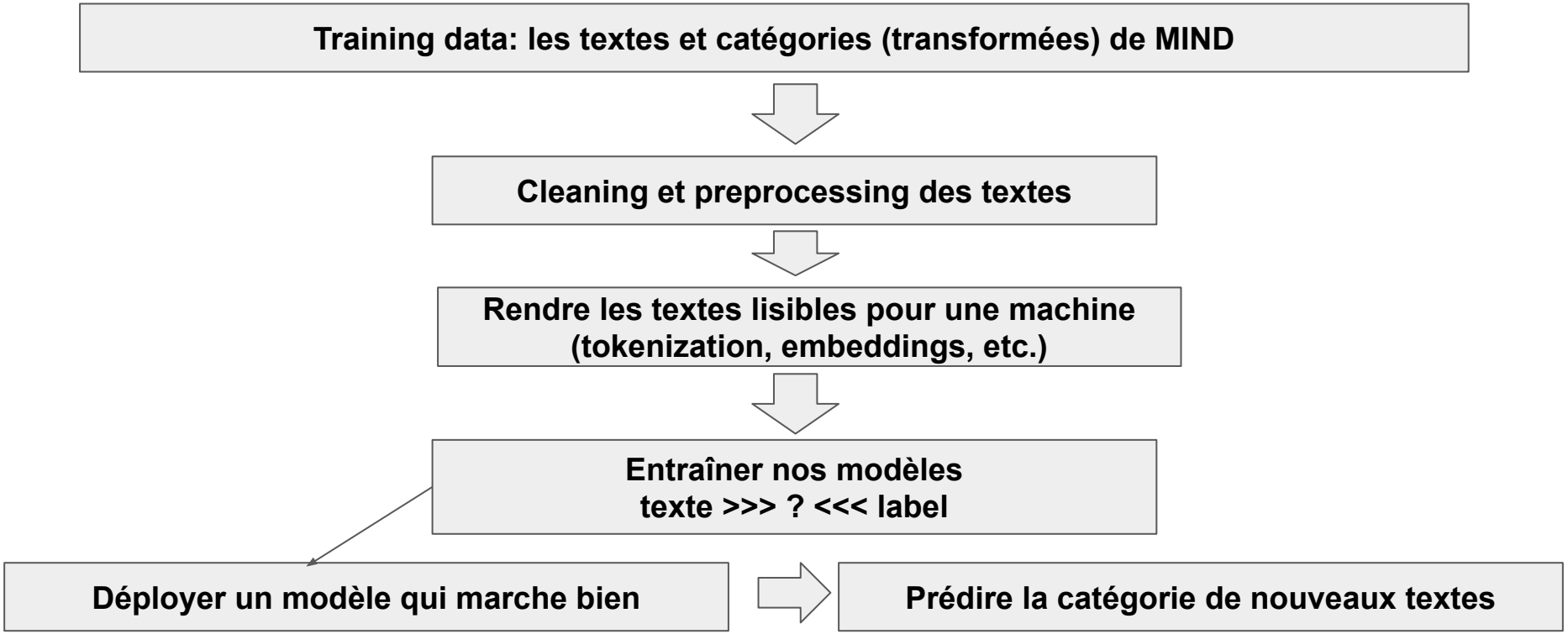
# Notre dataset: MIND (Microsoft News - msn)

- accès limité à des datasets d'entreprises
- exploration du potentiel de classification sur un dataset "news"



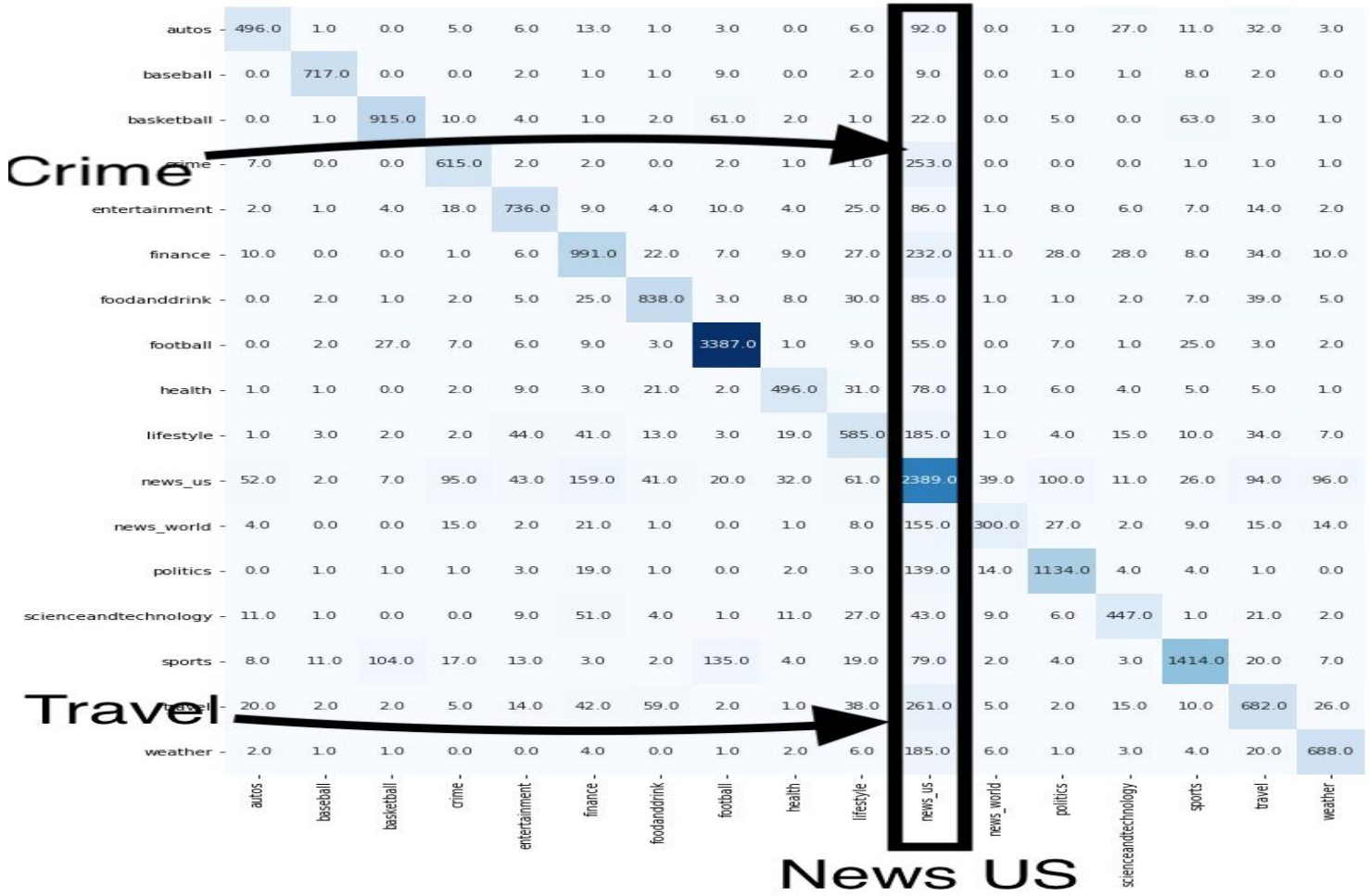



# Notre pipeline de classification



# Choix des modèles vs data quantité/qualité

| **Model** | Régression Log. | SVM | Réseau Conv1d | Réseau LSTM |
|---|---|---|---|---|
| **F1-score (avec 'news_us')** | 0.772 | 0.765 | 0.748 | 0.772 |
| **F1-score (sans 'news_us')** | 0.840 | 0.843 | 0.812 | 0.842 |


Crime
Travel
News US
autos
baseball
basketball
crime
entertainment
finance
foodanddrink
football
health
lifestyle
news_us
news_world
politics
scienceandtechnology
sports
travel
weather

## Pistes d’amélioration

- Qualité des labels
- Importance du preprocessing
- Optimisation des modèles/modèles plus avancés

## Recommandations - contexte entreprise

- Datasets spécifiques au domaine
- Quantité suffisante de données bien labellisées comme obstacle?
- Transfer Learning

Jedha

# Merci,

à bientôt !