

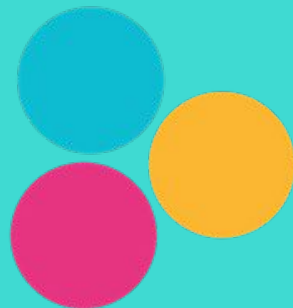


Data Science Bootcamp

La classification de textes en entreprise

Sebastian Paulo et Thierry Linde

en coopération avec



Holi

A man in a dark suit stands on a massive, towering stack of papers that fills the right side of the frame. He is reaching up with his right hand to touch the top layer of the stack. The stack of papers is composed of many thin, white sheets, creating a sense of depth and scale. The background is a cloudy sky.

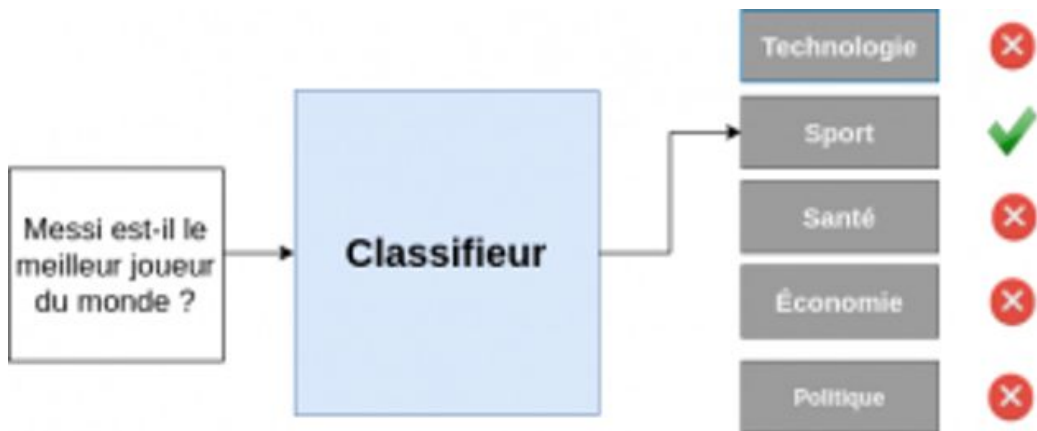
La classification de textes peut-elle faciliter la gestion d'informations en entreprise ?



La classification de textes - un domaine clé du NLP

Contexte holi.io

- Une méthode supervisée qui relie des textes à des catégories

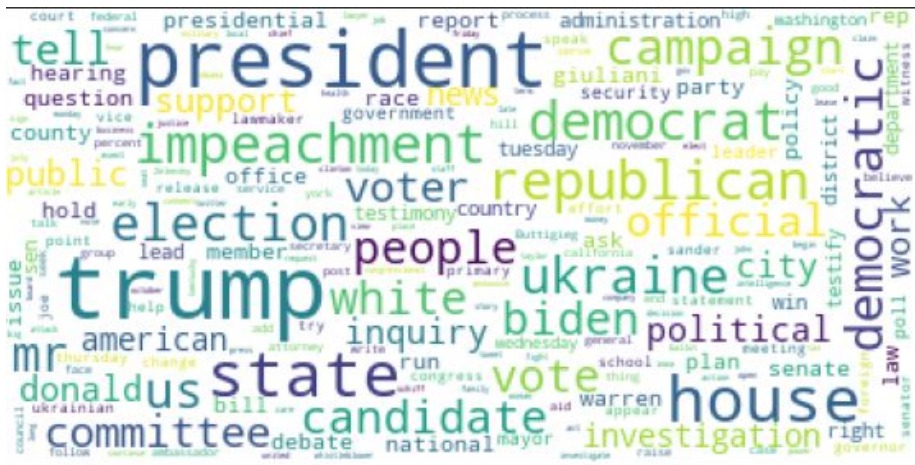


- Classification de textes en entreprise permettant un meilleur flux d'informations:
 - identification de contenu
 - recommandations ciblées

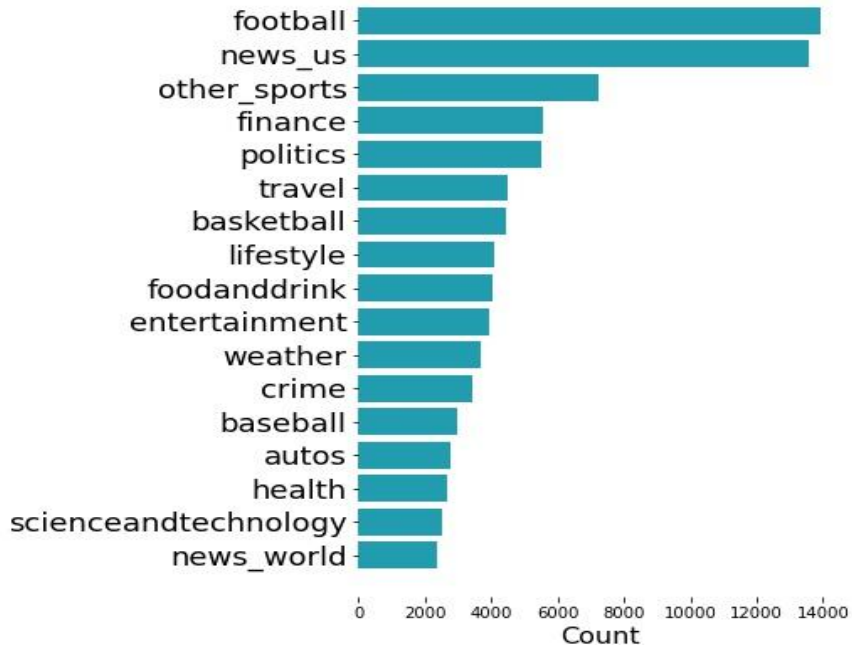


Notre dataset: MIND (Microsoft News - msn)

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



Count of categories in the dataset





Notre pipeline de classification

Training data: les textes et catégories (transformées) de MIND



Cleaning et preprocessing des textes



**Rendre les textes lisibles pour une machine
(tokenization, embeddings, etc.)**



**Entraîner nos modèles
texte >>> ? <<< label**

Déployer un modèle qui marche bien

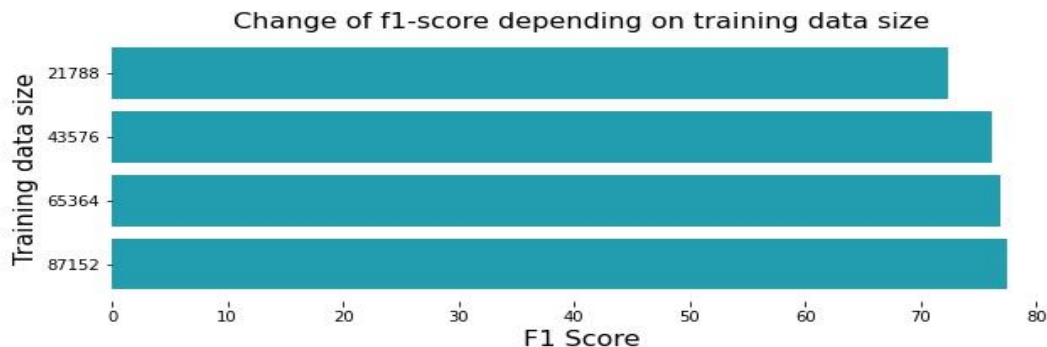


Prédire la catégorie de nouveaux textes



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

autos -	496.0	1.0	0.0	5.0	6.0	13.0	1.0	3.0	0.0	6.0	92.0	0.0	1.0	27.0	11.0	32.0	3.0
baseball -	0.0	717.0	0.0	0.0	2.0	1.0	1.0	9.0	0.0	2.0	9.0	0.0	1.0	1.0	8.0	2.0	0.0
basketball -	0.0	1.0	915.0	10.0	4.0	1.0	2.0	61.0	2.0	1.0	22.0	0.0	5.0	0.0	63.0	3.0	1.0
crime -	7.0	0.0	0.0	615.0	2.0	2.0	0.0	2.0	1.0	1.0	253.0	0.0	0.0	0.0	1.0	1.0	1.0
entertainment -	2.0	1.0	4.0	18.0	736.0	9.0	4.0	10.0	4.0	25.0	86.0	1.0	8.0	6.0	7.0	14.0	2.0
finance -	10.0	0.0	0.0	1.0	6.0	991.0	22.0	7.0	9.0	27.0	232.0	11.0	28.0	28.0	8.0	34.0	10.0
foodanddrink -	0.0	2.0	1.0	2.0	5.0	25.0	838.0	3.0	8.0	30.0	85.0	1.0	1.0	2.0	7.0	39.0	5.0
football -	0.0	2.0	27.0	7.0	6.0	9.0	3.0	3387.0	1.0	9.0	55.0	0.0	7.0	1.0	25.0	3.0	2.0
health -	1.0	1.0	0.0	2.0	9.0	3.0	21.0	2.0	496.0	31.0	78.0	1.0	6.0	4.0	5.0	5.0	1.0
lifestyle -	1.0	3.0	2.0	2.0	44.0	41.0	13.0	3.0	19.0	585.0	185.0	1.0	4.0	15.0	10.0	34.0	7.0
news_us -	52.0	2.0	7.0	95.0	43.0	159.0	41.0	20.0	32.0	61.0	2389.0	39.0	100.0	11.0	26.0	94.0	96.0
news_world -	4.0	0.0	0.0	15.0	2.0	21.0	1.0	0.0	1.0	8.0	155.0	300.0	27.0	2.0	9.0	15.0	14.0
politics -	0.0	1.0	1.0	1.0	3.0	19.0	1.0	0.0	2.0	3.0	139.0	14.0	1134.0	4.0	4.0	1.0	0.0
scienceandtechnology -	11.0	1.0	0.0	0.0	9.0	51.0	4.0	1.0	11.0	27.0	43.0	9.0	6.0	447.0	1.0	21.0	2.0
sports -	8.0	11.0	104.0	17.0	13.0	3.0	2.0	135.0	4.0	19.0	79.0	2.0	4.0	3.0	1414.0	20.0	7.0
travel -	20.0	2.0	2.0	5.0	14.0	42.0	59.0	2.0	1.0	38.0	261.0	5.0	2.0	15.0	10.0	682.0	26.0
weather -	2.0	1.0	1.0	0.0	0.0	4.0	0.0	1.0	2.0	6.0	185.0	6.0	1.0	3.0	4.0	20.0	688.0
autos -											news_us -						
baseball -												news_world -					
basketball -													politics -				
crime -														scienceandtechnology -			
entertainment -															sports -		
finance -																travel -	
foodanddrink -																	weather -
football -																	
health -																	
lifestyle -																	

News US



Pistes d'amélioration

Recommandations - contexte entreprise

- Qualité des labels
 - Importance du preprocessing
 - Optimisation des modèles/modèles plus avancés
-
- Datasets spécifiques au domaine
 - Quantité suffisante de données bien labellisées comme obstacle?
 - Transfer Learning



Jedha

Merci,
à bientôt !

