TELECOM Paris

IP PARIS

kaggle

Présentation P3 - Projet fil rouge

Toxic Comment Classification

28/04/2020

Hamza AMRI, Camille COCHENER, Sophie LEVEUGLE, Rodolphe SIMONEAU

Plan de la présentation

1. Contexte et présentation du nouveau sujet
2. Etat de l'art et méthodes
3. Premiers résultats obtenus
4. Difficultés et ajustements
5. Prochaines étapes

Contexte et présentation du nouveau sujet

Rappels des objectifs du sujet initial

Analyser et valoriser les données audio des centres d'appels

Reconnaissance vocale | NLP | Apprentissage

1 Transcription des données audio en texte

2 Extraire des insights pertinents pour améliorer la satisfaction client et la performance de vente des conseillers

Rappels des objectifs du sujet initial

1 – **Transcription des données audio en texte**

2 – **Choix de trois cas d'usage pertinents**

- Identification de l'argumentaire clé de vente avec assurance
- Segmentation des conversations afin d'identifier le sentiment du client
- Prédiction de la volumétrie d'appel

Featured Prediction Competition

Toxic Comment Classification Challenge

Identify and classify toxic online comments

$35,000

Prize Money

Jigsaw/Conversation AI · 4,550 teams · 2 years ago

Overview | Data | Notebooks | Discussion | Leaderboard | Rules | Team | My Submissions | Late Submission

Overview

Description

Evaluation

Discussing things you care about can be difficult. The threat of abuse and harassment online means that many people stop

Toxic Comment Classification Challenge

"FUCK YOUR FILTHY MOTHER IN THE ASS, DRY!"

"Stupid peace of shit stop deleting my stuff asshole go die and fall in a hole go to hell!

Toxic Comment Classification Challenge

Conversation AI

Développement d'outils pour améliorer les conversations en lignes
Etude de commentaires toxiques

Beaucoup de modèles disponibles sur l'**API Perspective**
MAIS

Les modèles font encore des erreurs

Pas de sélection possible du type de toxicité

Toxic Comment Classification Challenge

Construire un **modèle multi-labels** qui permet de **détecter plusieurs types de toxicité**...

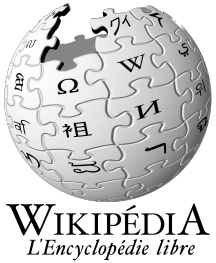
Menaces

Obscénité

Insultes

Haine

à partir de **commentaires des pages de discussion de Wikipedia**



Rappel de notre équipe

Hamza AMRI

Camille COCHENER

Sophie LEVEUGLE

Rodolphe SIMONEAU

Support Académique

Geoffroy PEETERS
(Spécialiste des données audio)

Béatrice BIANCARDI
(Spécialiste des sciences cognitives)

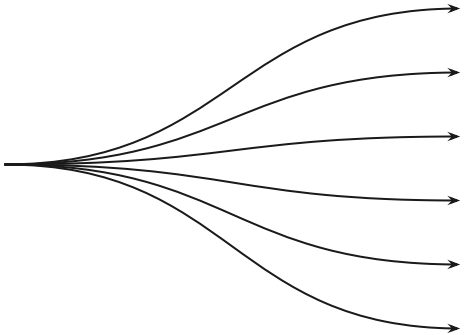
Etat de l’art et méthodes

Le problème de classification multi-label

Entrée

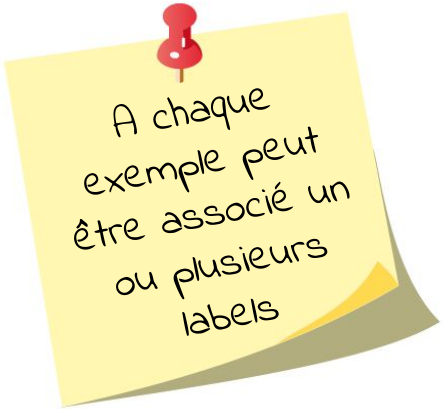
X

Commentaires des pages Wikipedia (texte)



Plusieurs réponses binaires

Toxic
Severe toxic
Obscene
Threat
Insult
Identity hate



Trouver une fonction qui relie les entrées X à plusieurs vecteurs binaires en sortie

Transformation du problème en...

Binary Relevance

- 1 classifieur par label
- Suppose l'**indépendance entre les labels**

Label Powerset

- Considération de **combinaison de labels**
- 1 classifieur pour une combinaison de labels
- Prise en compte des **corrélations** entre labels
- Coûteux en calcul

Classifier Chain

- Chaînes de classifieurs binaires
- Prise en compte de la **sortie du classifieur précédent**
- Prise en compte des **corrélations** entre labels

Customisation des algorithmes

- Changement de fonction de coût et/ou de fonction de décision

Préparation des données textuelles

"D'aww! He matches this background colour I'm seemingly stuck with. Thanks. (talk) 21:51, January 11, 2016 (UTC)"

"In nationality terms it refers to the Republic of Ireland""" no such country . so nationality terms it refers to Ireland . â€" Preceding unsigned comment added by 109.76.191.188"

Préparation des données textuelles

"the", "is", "in", "for"...

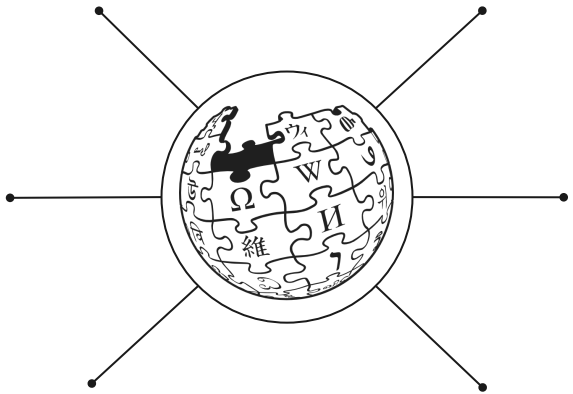
Stopwords

Chiffres

"109.76.191.188"
"21:51"

"!", "!!", "?!"
"@!$%§"

Ponctuation



Lemmatisation

be, is, are, were

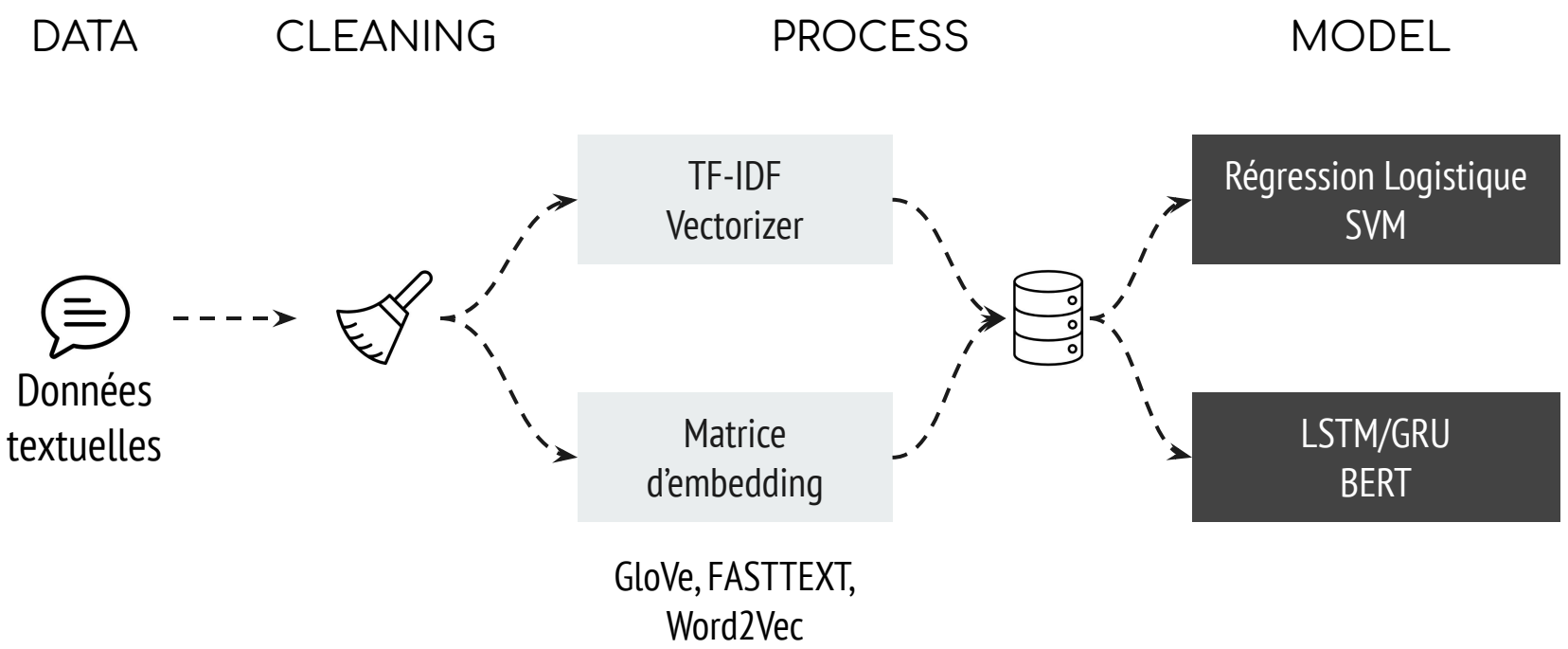
"GET FUCKED UP. GET FUCKEEED UP..."

Majuscules

Stemmatisation

cook, cook-ed, cook-ing

Techniques d'apprentissage identifiées

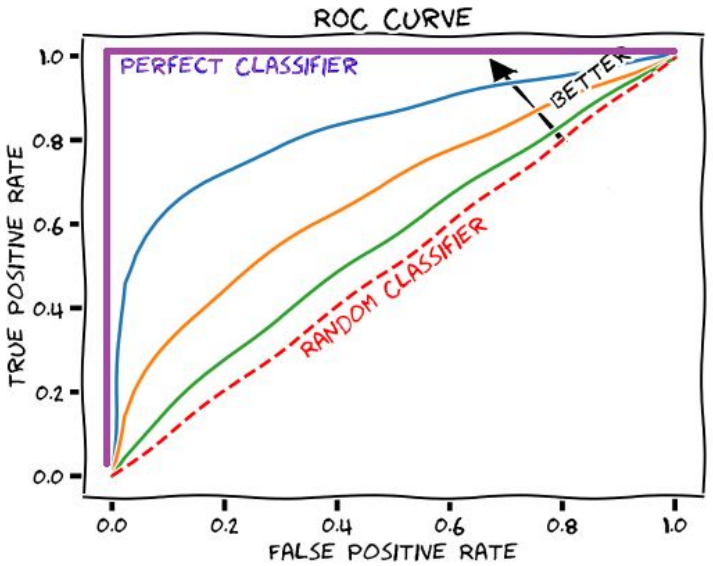


Métriques d'évaluation

Moyenne des AUC individuelles de chaque colonne prédite

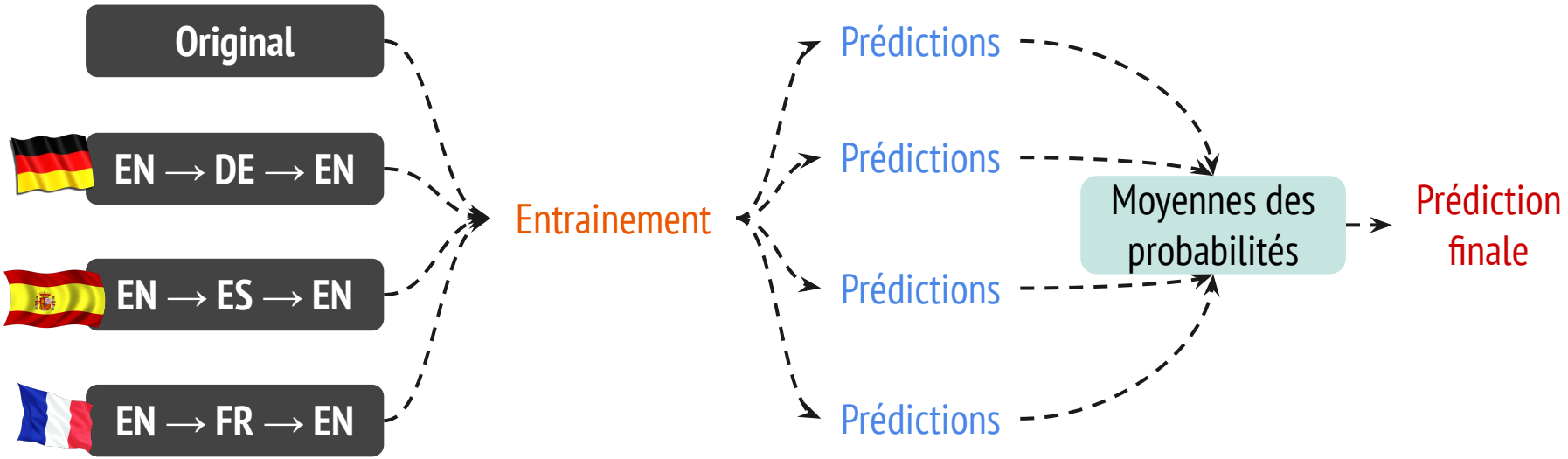
$$\frac{1}{m}\sum_{i=1}^{6} AUC_i$$

- Données déséquilibrées
- Classification multi-label



Méthodes supplémentaires envisagées

Train et Test-Time Augmentation (TTA)



Quels environnements de développement ?

Développement du code pour les analyses

Librairies majoritaires utilisées

NLTK

Seaborn

Plateforme de calcul ?

Premiers résultats obtenus

Exploration des données (1)

Analyse des labels à prédire

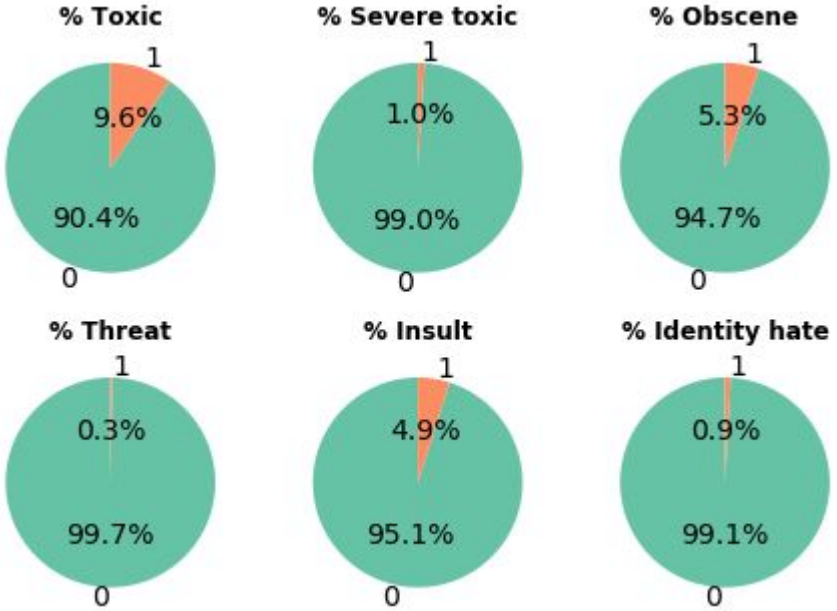
Classes déséquilibrées

89.8% Sans toxicité

10.2% Avec toxicité

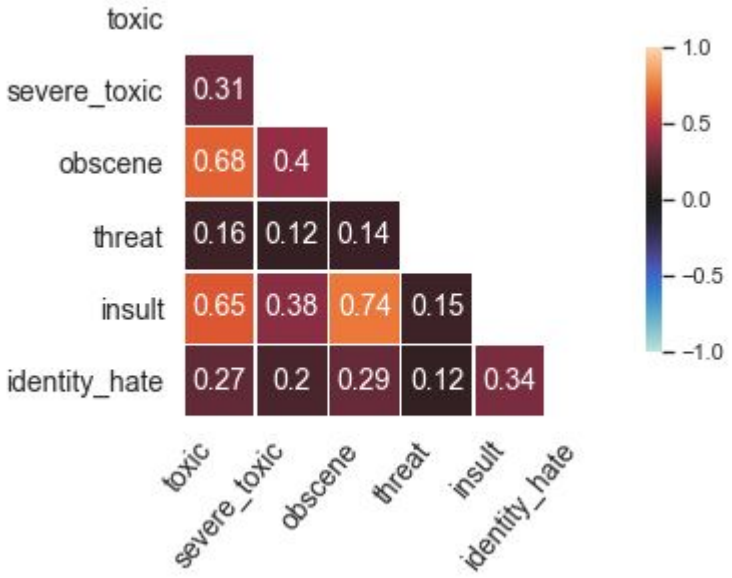
Majorité de labels **"toxic", "insult", "obscene"**

Pas de données manquantes



Exploration des données (2)

Analyse des labels à prédire



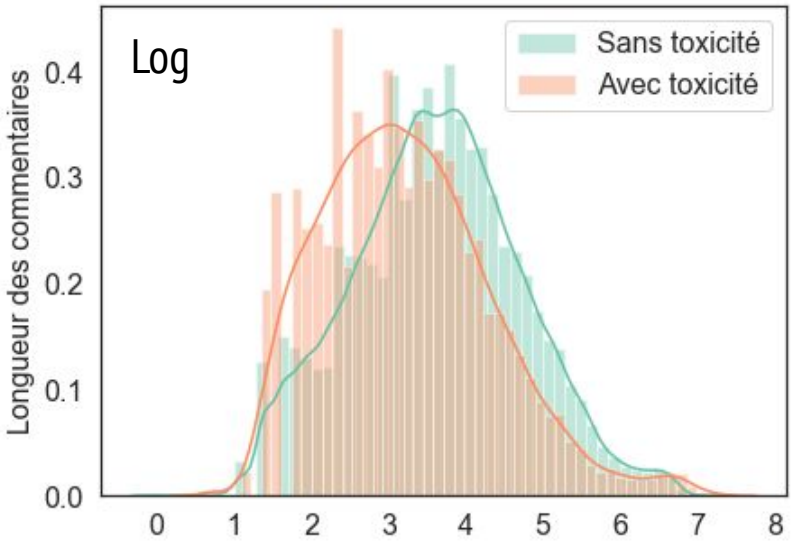
Corrélations entre labels

De **fortes corrélations** entre :

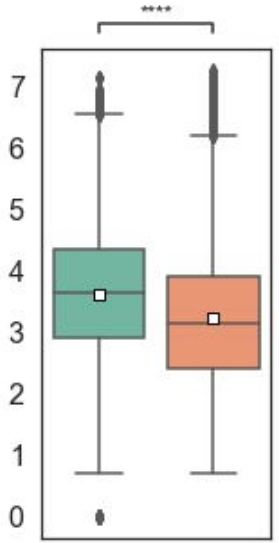
- "toxic" et "insult" : **0.65**
- "toxic" et "obscene" : **0.68**
- "insult" et "obscene" : **0.74**

Exploration des données (3)

Longueurs des commentaires



Légère ≠ entre les longueurs moyennes des commentaires sans toxicité et avec toxicité



*Cohen's d = 0.35
(Petit effect size)

Description des commentaires

- Longueur moyenne : *67.3 termes*
- Longueur min : *1 terme*
- Longueur max : *1411 termes*
- *75%* des commentaires ont une *longueur < 75 termes*

Exploration des données (4)

Suppression des stopwords

fuck,
shit,
fucking,
nigger...

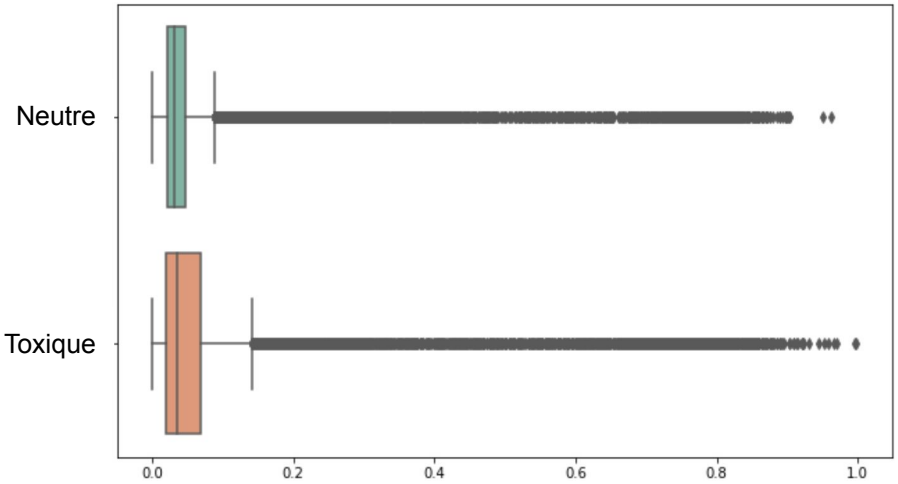
≠

article,
wikipedia,
page,
talk...

Exploration des données (5)

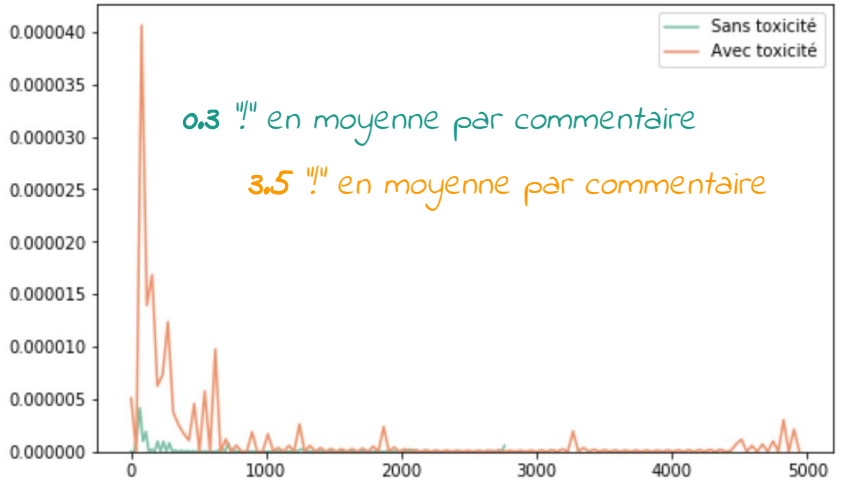
Analyse des commentaires

Une utilisation sensiblement plus marquée des majuscules dans les commentaires très toxiques



Taux de majuscules par commentaire

Une utilisation plus marquée du point d'exclamation dans commentaires toxiques

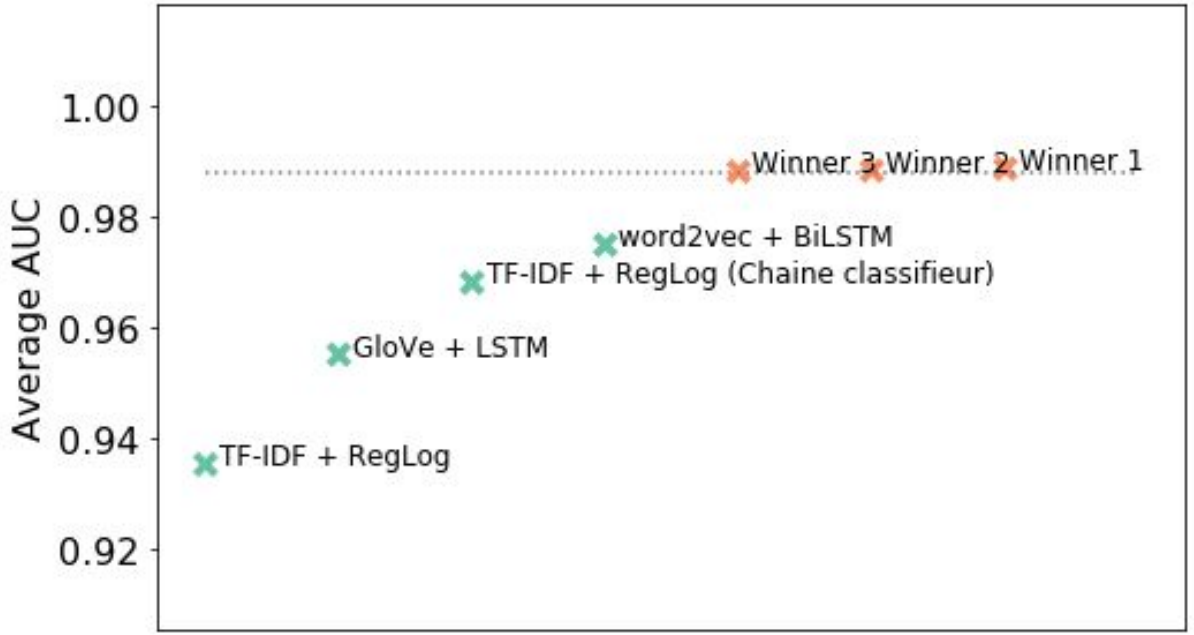


Nombre de "!" par commentaire

Résultats des premiers modèles

Essai n°	Préparation	Transformation	Algorithme	AUC
1	**Nettoyage classique** Lower Stopwords Lemmatisation ...	TF-IDF	Binary relevance Régression Logistique	0.93518
2		TF-IDF	Chaîne de classifieur Régression Logistique	0.968
3		Tokenizer Padding Embedding (GloVe)	LSTM Simple Dense	0.95503
4		Tokenizer Padding Embedding (GloVe, Word2vec, Fasttext)	Bidirectional LSTM MaxPool Dropout Dense	0.97483

Positionnement dans le challenge



🎯

Score de l'équipe gagnante

0.98856

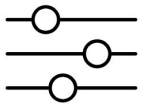
Difficultés et ajustements

Difficultés et ajustements réalisés

Changement de sujet au bout de 6 mois de projets

↓

Prise de décision rapide sur un nouveau sujet



Travail à distance dû au confinement

↓

Github, visioconférence, drive...

Limites dans la puissance de calcul

↓

Utilisation d'outils gratuits (Collab,...)

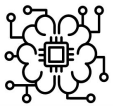
Limites d'un projet Kaggle

↓

Recherche de complexité (création d'une API...)

Prochaines étapes

MAI



P1 - Élargir notre benchmark de modèles
P1 - Choisir une méthode et l'optimiser
P1 - Développer un code propre prêt à être déployer

P1 - Choisir la forme finale de la solution

JUIN

Dashboard

P2 - Développer la forme finale de la solution
P2 - Travailler sur les supports de rendu final

Merci pour votre attention !