

# **Rapport de Compétition Kaggle**

## **Natural Language Processing with Disaster Tweets**

**Dec 2022**

**Présenté par SEKPONA Kokou Sitsopé,  
Etudiant en Ingénierie Machine Learning  
à Openclassrooms/ Central Supélec**

# Introduction

Kaggle est un site de data science organisant des compétitions intéressantes à des buts d'apprentissage ou de résolution des problèmes réels de la vie.

En tant qu'étudiant Ingénieur Machine Learning,

Il nous est demandé de participer à une compétition Kaggle de notre choix.

Nous avons donc choisi de participer à ma compétition « **Natural Language Processing with Disaster Tweets** » qui est une compétition en traitement naturel de Langage.

Ce projet social rend service énormément à la communauté car permettra aux secouristes d'être alertés en cas de catastrophe à un endroit et donc vite intervenir, ce qui a suscité notre grand intérêt porté à cette compétition.

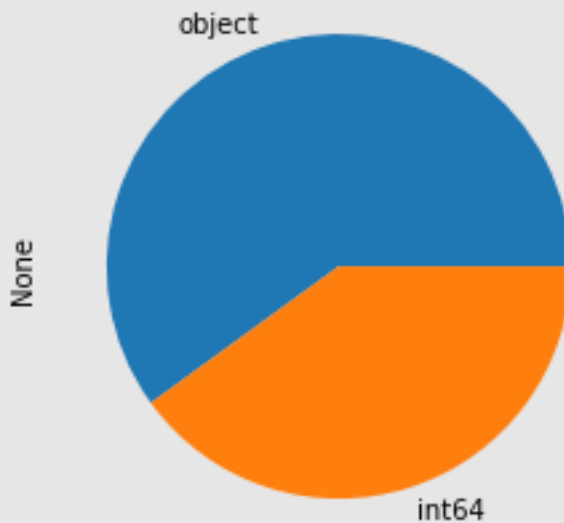
Nous allons donc présenter les détails concernant cette compétition à laquelle nous avons participé, les démarches et les résultats obtenus.

## Plan de notre travail

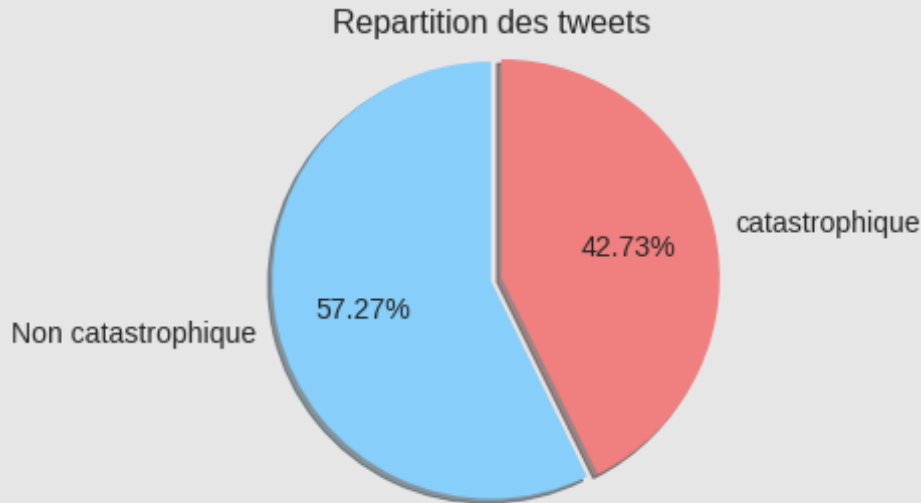
- Nettoyage et analyse exploratoire
- Feature extractions
- Test de différents modèles avec les différentes méthodes d'extraction de features utilisé
- Utilisation d'une métrique commune pour mesurer les résultats et comparaison
- Soumission du résultat de chaque model sur Kaggle

## A. Nettoyage et analyse exploratoire

- Notre dataset contient de train contient 5 colonnes et 7613 lignes. La colonne id ne nous sera pas utile. Nous allons la supprimer
- Distribution des colonnes selon les types



- Nous avons au total 70 Doublons dans notre dataset. Nous allons les supprimer
- Vérification de la distribution des données



Les Tweet sont bien réparties pour l'entrainement de notre model : Environ 50% dans chaque classe

## ➤ Prétraitement du Test

Nous nettoyons le texte et nous le transformons :

Suppression:

- URL
- Balises HTML
- Références de personnages
- Caractères non imprimables
- Valeurs numériques

Traitement

- Lemmatisons le texte
- Conversion en minuscules.
- Suppression des caractères répétés dans les mots allongés,
- Suppression des mots vides
- Conservation des hashtags car ils peuvent fournir des informations précieuses sur ce projet particulier.

## ➤ Feature Engineering

Nous créons 10 colonnes qui sont :

- Nombre de phrases
- Nombre de mots
- Nombre de caractères
- Nombre de hashtags
- Nombre de mentions
- Nombre de mots tout en majuscules
- Longueur moyenne des mots
- Nombre de noms propres (PROPN)
- Nombre de noms non propres (NOM)
- Pourcentage de caractères qui sont de la ponctuation

## B. Features extractions

Cette étape nous permet de convertir les documents en vecteur pour être utilisés par les modèles pour la prédiction

Nous utilisons 2 méthodes principales :

- **TfidfVectorizer**
- **Doc2vec**

## C. Modélisation : Résultats

### 1. Logistic Régression :

#### a. With TFIDFVectorizer

- Nous recherchons les meilleurs hyper paramètres avec GridSearchCV
- Accuracy : 0.7809139784946236

## b. With Doc2vec

- Nous recherchons les meilleurs hyper paramètres avec GridSearchCV
- Accuracy : 0.6503311258278146

## 2. Transformers

### a. With TfidfVectorizer

Accuracy: 0.569

### b. With Doc2vec

Accuracy: 0.558





## 3. Pre entrained model nnlm-en-dim50

Accuracy: 0.7775

### ➤ Soumission des résultats:

Search		
Overview	Data	Code
Discussion	Leaderboard	Rules
Team	Submissions	Submit Predictions
Complete · 11h ago · Doc2vec with Transformers		
sample_submission.csv	0.42966	
Complete · 12h ago · Tfidf with Transformers	0.42966	
sample_submission.csv	0.65337	
Complete · 13h ago · Doc2vec logistic Regressor with features engineering columns	0.64296	
sample_submission.csv	0.79098	
Complete · 2d ago · Logistic regression Tfidf Vectorizer commit		

## Rang à la compétition:

472	neer kumar		0.78639	2	1ms
473	Pandey Harsh		0.79098	2	4d
474	kokou sitsope		0.79098	8	10h
 Your Best Entry! Your submission scored 0.78639, which is not an improvement of your previous score. Keep trying!					

- Nous retenons le model de Régression Logistique utilisé avec TfidfVectorizer comme meilleur model car il a obtenu le meilleur score sur ce ensemble de données.

## Conclusion:

Cette compétition nous a permis d'améliorer notre expérience en Nlp et de découvrir d'autres modèles pré entraînés et aussi les Transformers qui nous ont permis au final d'obtenir un bon score par rapport aux résultats du projet 5 basé également sur le NLP. Nous avons appris également à participer à un concours actif, de Data science et à découvrir Kaggle, une plateforme idéale pour s'améliorer dans la Data science à cause des fonctionnalités comme le jupyter notebook avec un GPU et un TPU gratuit, des compétitions pour gagner des prix énormes. Ces découvertes nous permettent d'acquérir de nouvelles connaissances, ce qui est un atout pour notre carrière en tant qu'Ingénieur Machine Learning.