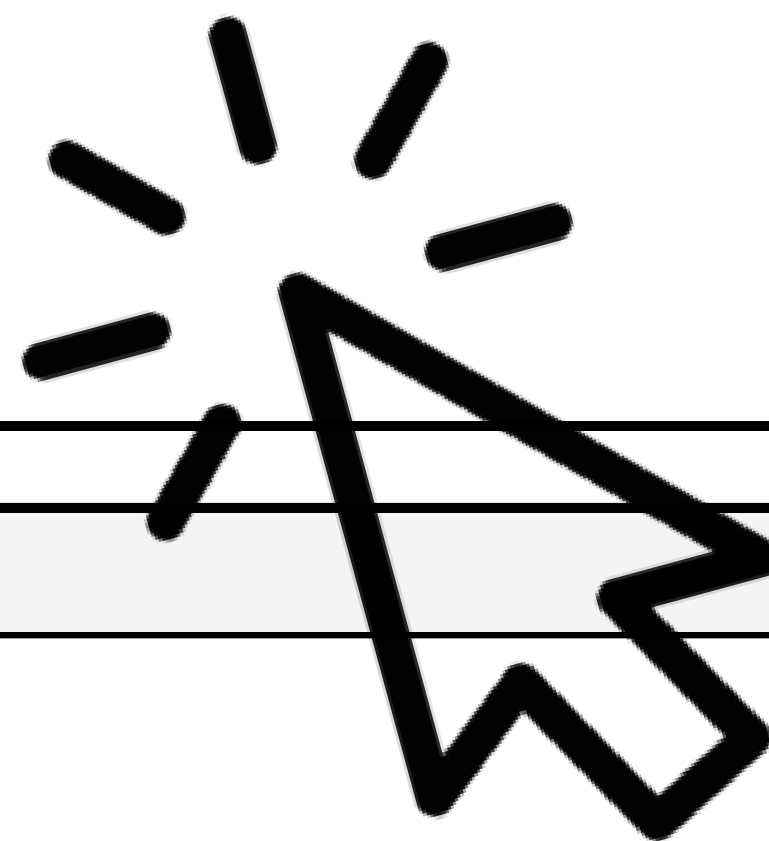


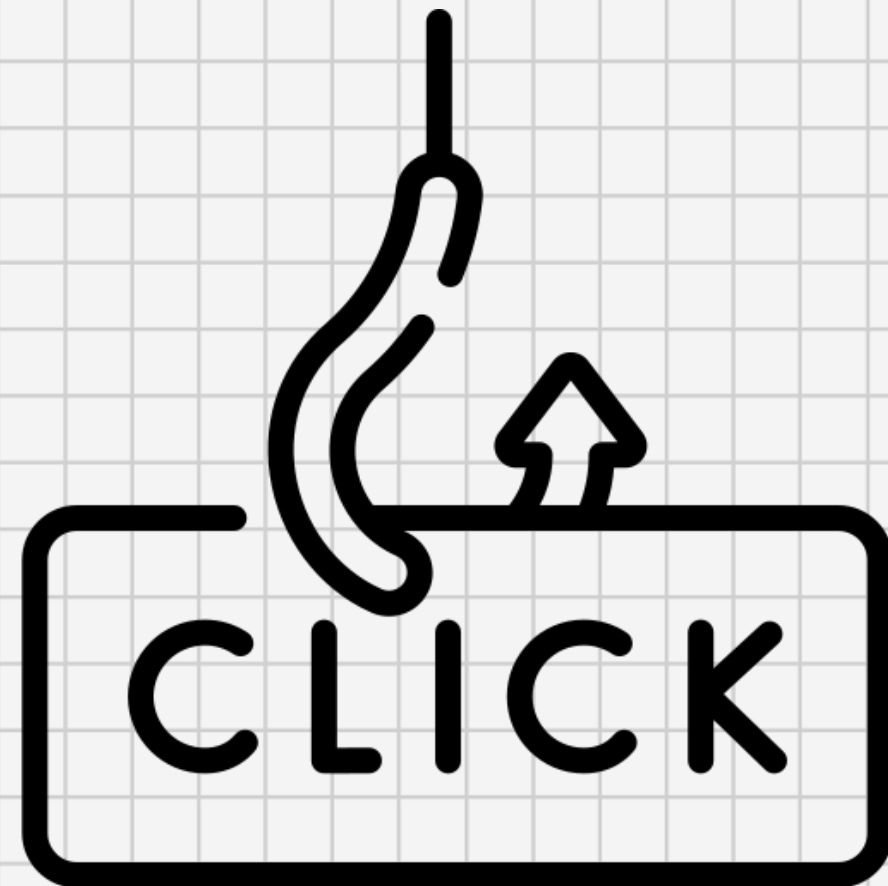


# Comprendre et détecter le clickbait



Heddier Soler - Krystel Laudon - Colin Llacer

# Sujets de la présentation



Introduction



Méthodologie



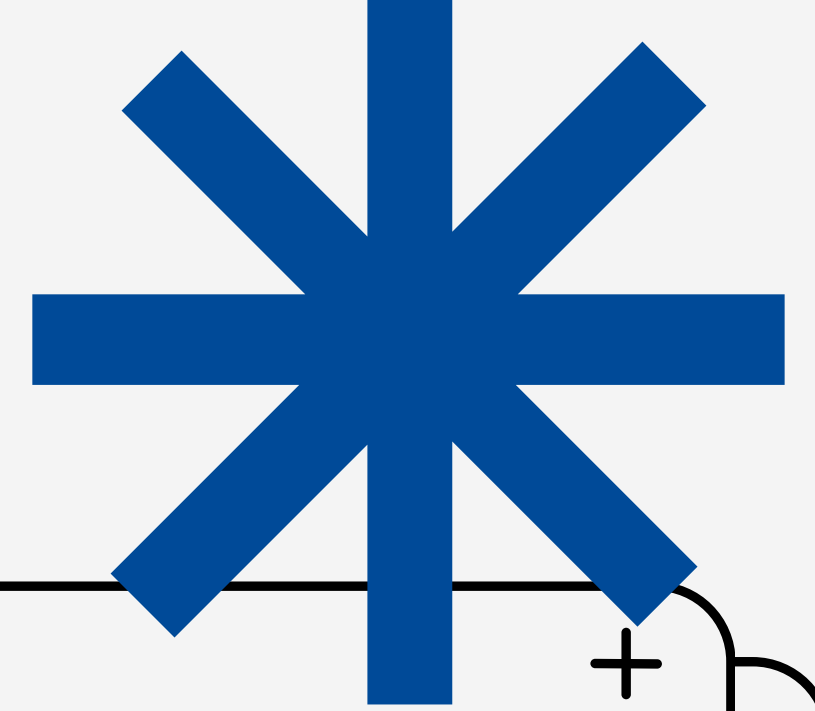
Analyses



Création d'un modèle de prédiction



Difficultés rencontrées et conclusion




# Introduction

# Qu'est-ce que le clickbait ?

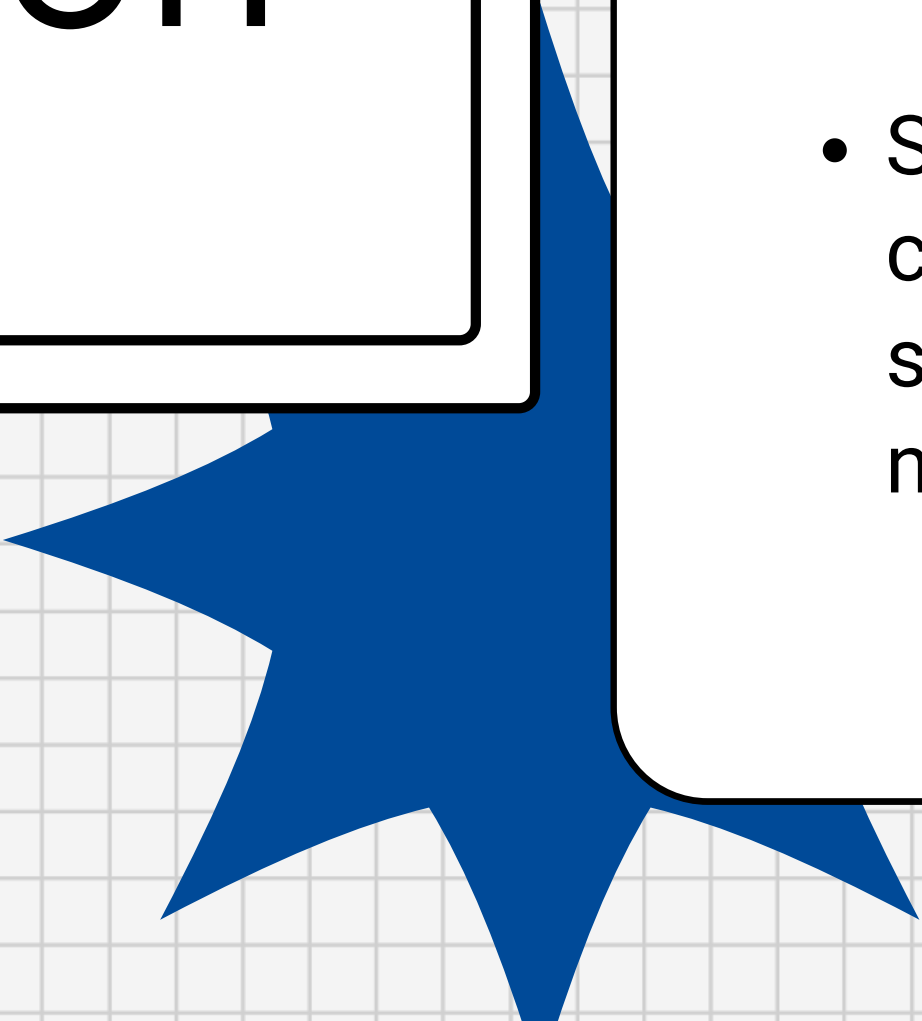


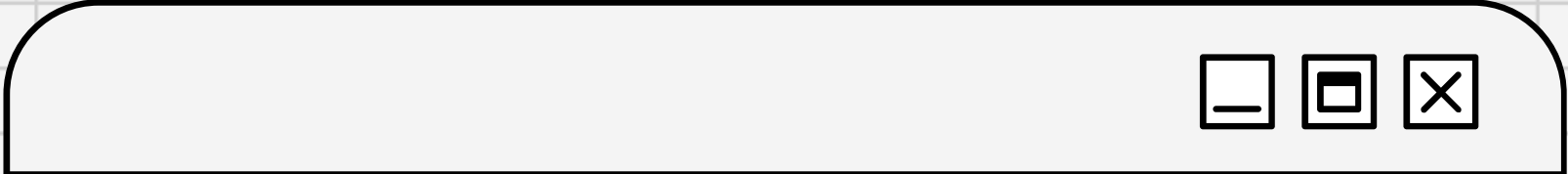
“ Hyperlien au **titre provocant ou intrigant** affiché sur une page Web, qui pique la curiosité des internautes et les incite à cliquer pour en connaître davantage, mais qui ne mène qu'à un **contenu peu informatif** et décevant ”

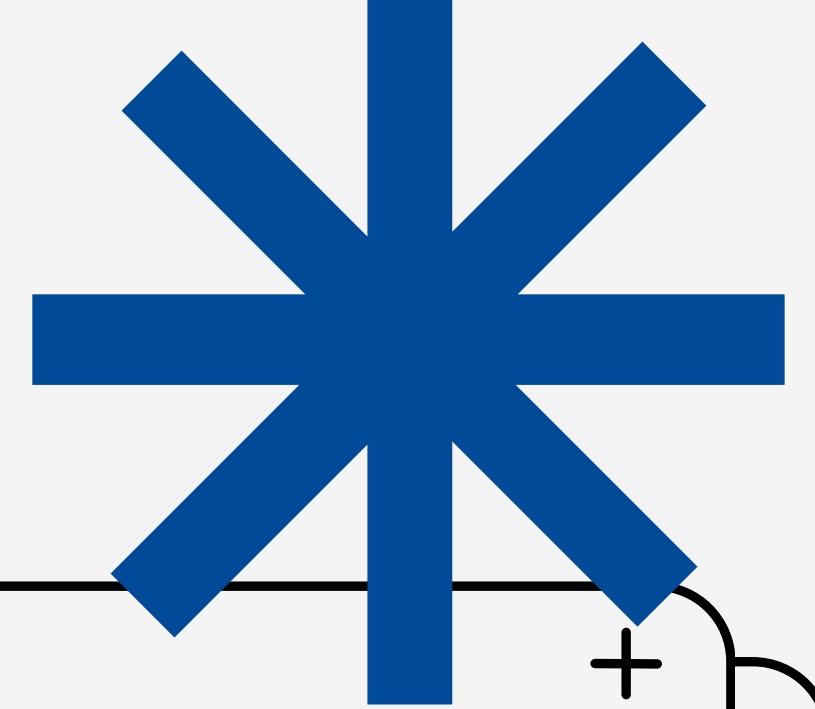
Office québécois de la langue française



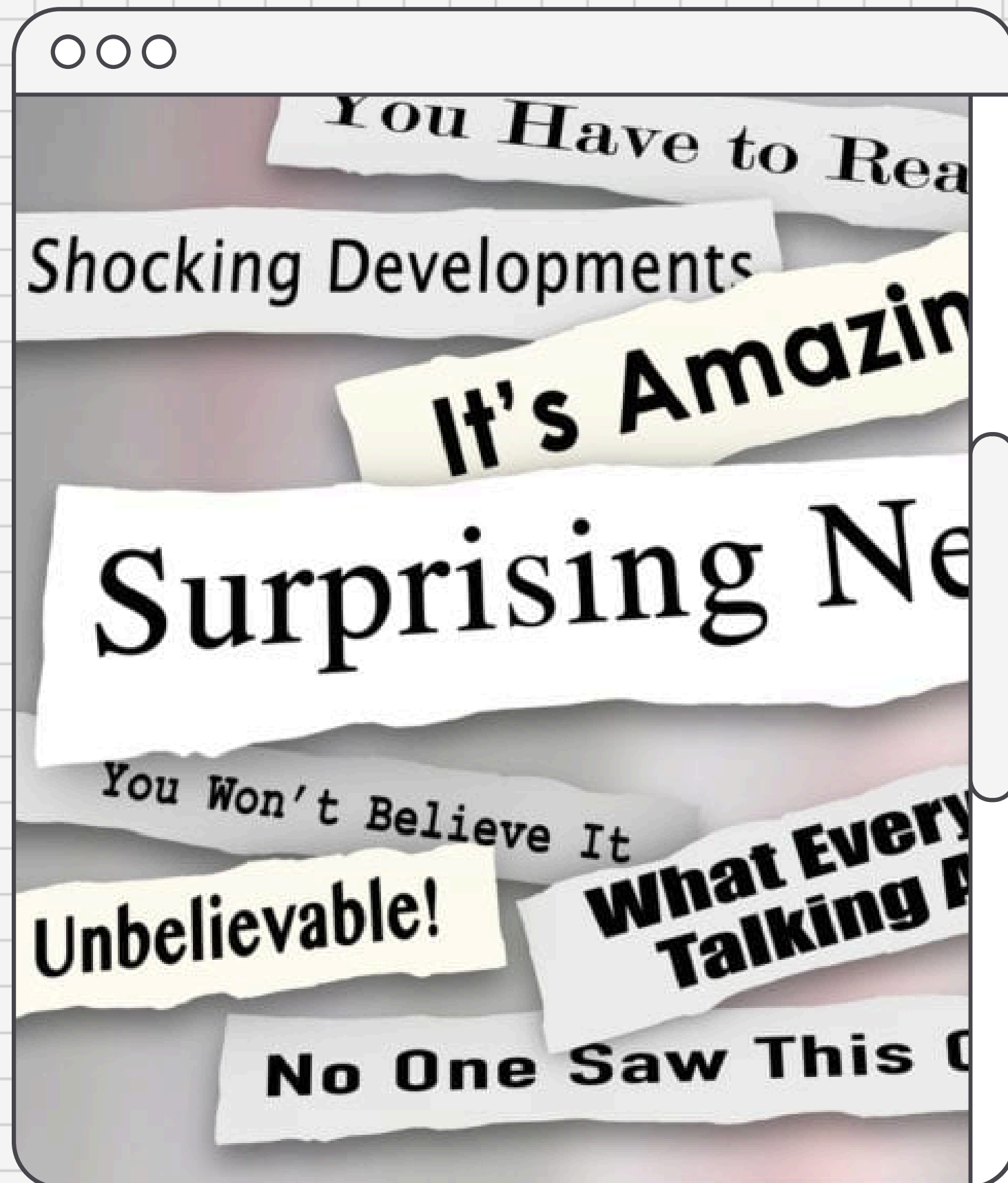
# Intuition



- 
- Les titres “clickbait” semblent suivre un certain nombre de règles et disposer d’une structure particulière
  - Si nous parvenons à identifier ces règles et structures avec suffisamment de précision, nous pourrions créer des filtres



# Méthodologie



## Étapes :



- Mettre les données en forme et les séparer en différents jeux de données selon le niveau de prétraitement



- Effectuer une multitude d'analyses sémantiques, grammaticales et syntaxiques à partir des différents jeux de données



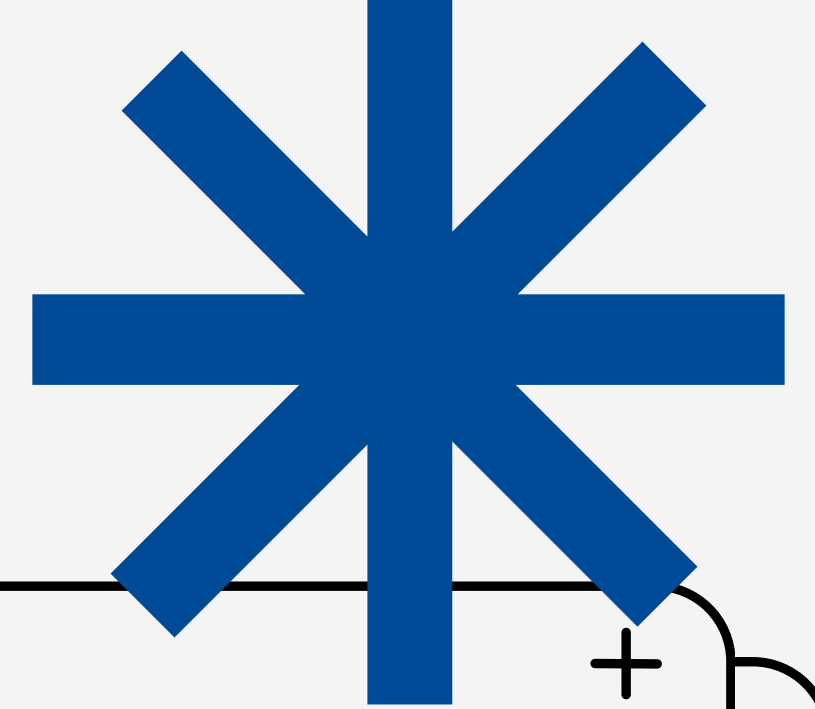
- En parallèle, à partir d'un modèle BoW, identifier les mots les plus pertinents pour notre classification



- Construire deux modèles pour chaque jeu de variables (indicateurs seulement, mots seulement et mélange)



- Utiliser le modèle le plus performant comme filtre anti-clickbait



# Analyses



# Indicateurs

## Longueur du titre

Les titres clickbait sont légèrement plus longs en moyenne (53 contre 59 caractères)

## Nombre de majuscules

On retrouve presque deux fois plus de majuscules dans les titres clickbait

## Diversité du vocabulaire

On retrouve 17 500 mots uniques dans les titres traditionnels, contre seulement 12 000 pour les titres clickbaits

## Longueur moyenne des mots utilisés

Les mots utilisés dans les titres traditionnels sont légèrement plus longs en moyenne (5,6 contre 5 caractères)

## Analyse de la ponctuation

Beaucoup plus de virgules (5x) et de points (x7) dans les titres traditionnels. Beaucoup plus de guillemets (x3) dans les titres clickbait.

## Analyse de sentiment

Les titres traditionnels sont en moyenne légèrement négatifs, les titres clickbait légèrement positifs

## Analyse grammaticale

On retrouve plus d'adjectifs (x3) dans les titres traditionnels, mais plus d'adverbes (x3) et de pronoms (x13) dans les titres clickbait. Peu de différence sur les superlatifs

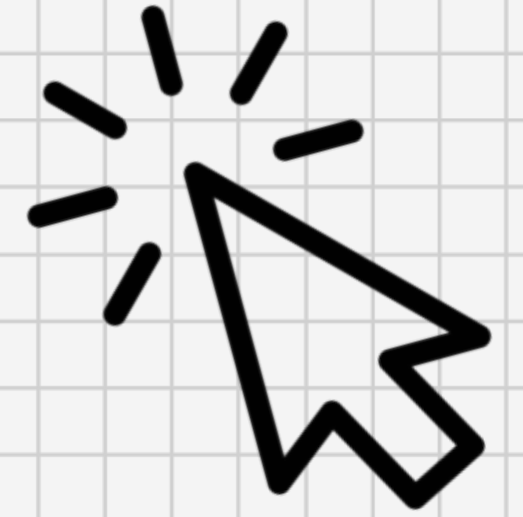
## Pertinence des mots





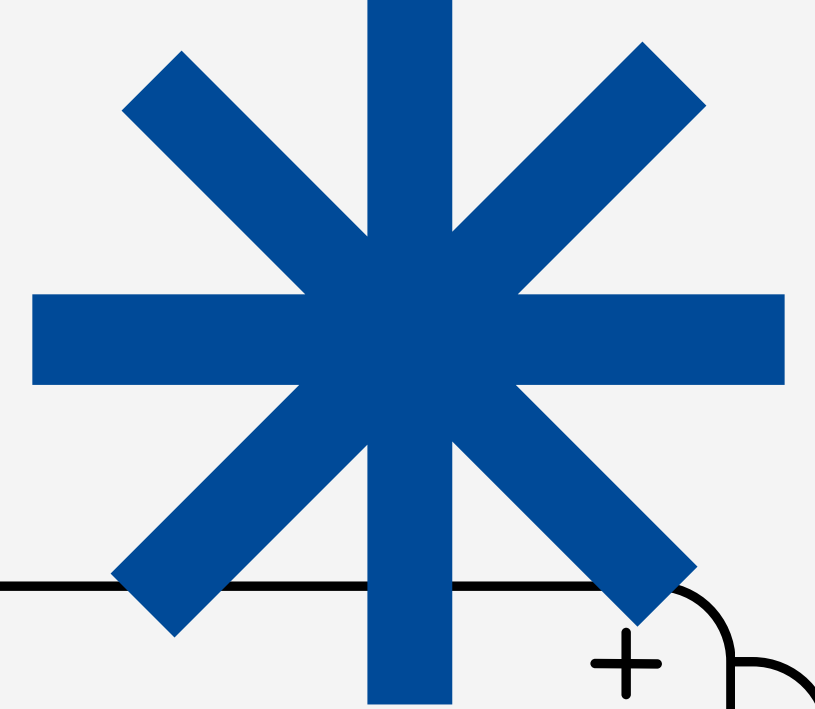
### Nuage de mot / N-grams :

- L'analyse du nuage de mot ainsi que de bi et trigrammes fait ressortir des sujets abordés très différents
- Les sites clickbait utilisent souvent des listes ("X choses", "X fois", "X personnes"), parlent beaucoup d'esotérisme et utilisent des références à la pop culture
- Les titres traditionnels contiennent souvent des informations géographiques, des avis de décès, ou des informations (géo)politiques



### Mots les plus discriminants (Chi-carré) :

- L'utilisation de chiffres dans le titre est l'indicateur le plus important pour un titre clickbait, même si beaucoup de nombre apparaissent aussi dans les titres traditionnels
- Les mots que l'on retrouve souvent dans les titres de clickbait sont généralement les plus discriminants ("thing", "know", "people", "based", "zodiac")
- Le vocabulaire lié aux drames semble également relativement important ("killed", "die", "crash", "dead")



# Modèles

----->

Quel modèle choisir ?



## Régression logistique

Explicabilité

	<u>Exactitude</u>	<u>F1</u>	<u>TFP</u>
<u>Indicateurs</u>	0,847	0,879	7,4%
<u>Mots</u>	0,818	0,885	<b>2,1%</b>
<u>Cumulé</u>	0,912	0,941	3,0%

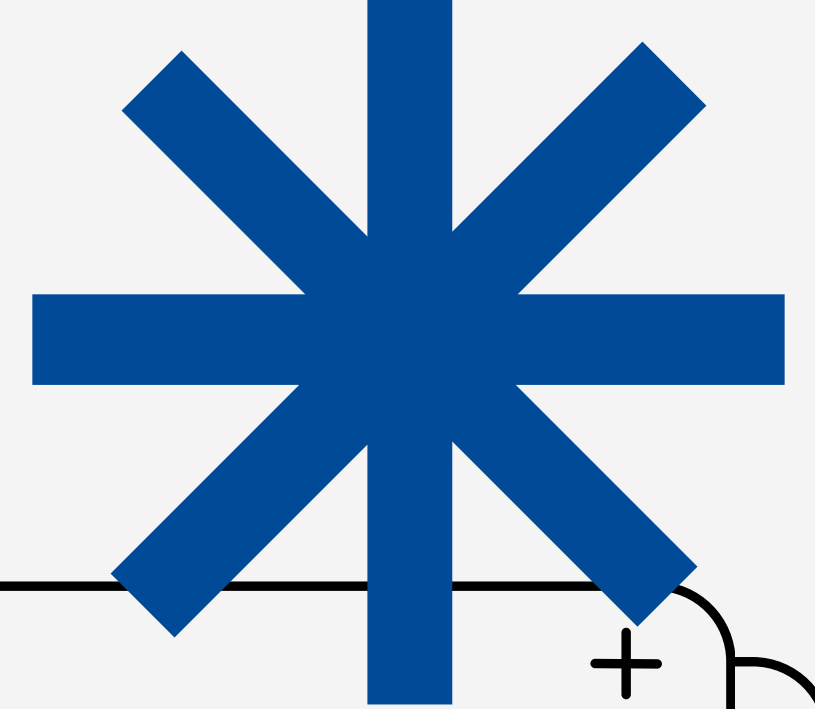
VS

## Foret aléatoire

Performance

	<u>Exactitude</u>	<u>F1</u>	<u>TFP</u>
<u>Indicateurs</u>	0,926	0,935	5,5%
<u>Mots</u>	0,824	0,886	2,8%
<u>Cumulé</u>	<b>0,948</b>	<b>0,960</b>	3,0%

Nous avons décidé de retenir la forêt aléatoire avec le jeu de variables le plus complet



# Difficultés rencontrées

# Difficultés et solutions

Recherche



- Le jeu de données utilisé ne comprend que 6 sites internet comme source de titres “clickbait”, et 4 pour les titres “normaux”
  - Notre approche basée à la fois sur un sac de mots et des indicateurs nous permet d’améliorer notre capacité de généralisation
- Certains indicateurs sont très différents selon le type de titre. Risque de classifier comme “clickbait” un journal réputé qui utiliserait une caractéristique surreprésentée dans les titres clickbait
  - Notre approche hybride permet de mitiger ces effets
  - Les modèles ne sont pas uniquement comparés sur leur exactitude, et le seuil pour la classification maximise le score F-1 (prend en compte les faux positifs)
- Nous avons également du revenir à de multiples reprises sur notre prétraitement après observation des n-grams, du nuage de mot et des mots les plus significatifs
  - Transformer les nombres en un tag, traiter les contractions...

Enregistrer

Annuler



# Merci !

Des questions ?

