Réduire les biais dans la collecte de tweets

B. Mazoyer, N. Hervé, C. Hudelot, J. Cagé

Journée DAHLIA, Nantes, 28 juin 2019

Obtenir des données de Twitter

Dans quel but, contexte de l'étude :

- Étude du graphe social ou des interactions ?
- Domaine d'étude bien circonscrit et limité ? (comptes, mots-clés, période temporelle, …)
- Captation *a posteriori* ou en temps réel ?
- Sampling ou exhaustivité ?

Dans notre cas :

- Analyse de la propagation d'information en français
- Comparaison avec les médias traditionnels

Approches possibles

Accès aux tweets :

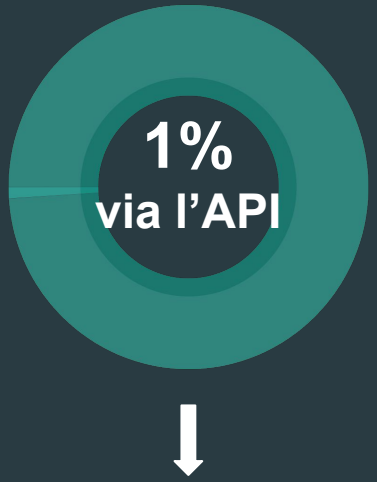
- Payant : firehose, broker et API premium / enterprise
- Gratuit : API restreinte et scraping web

Les API :

- search : 7 jours d’historique
- streaming sample : 1% de tous les tweets
- streaming filter : critères de filtre sur requête

Volumétrie avec l’API sample

500 millions de tweets par jour



5 millions de tweets collectés

5 millions de tweets collectés

90 000 tweets en français

90 000 tweets en français

180 tweets par jour

Approche proposée : optimisation de l'API filter

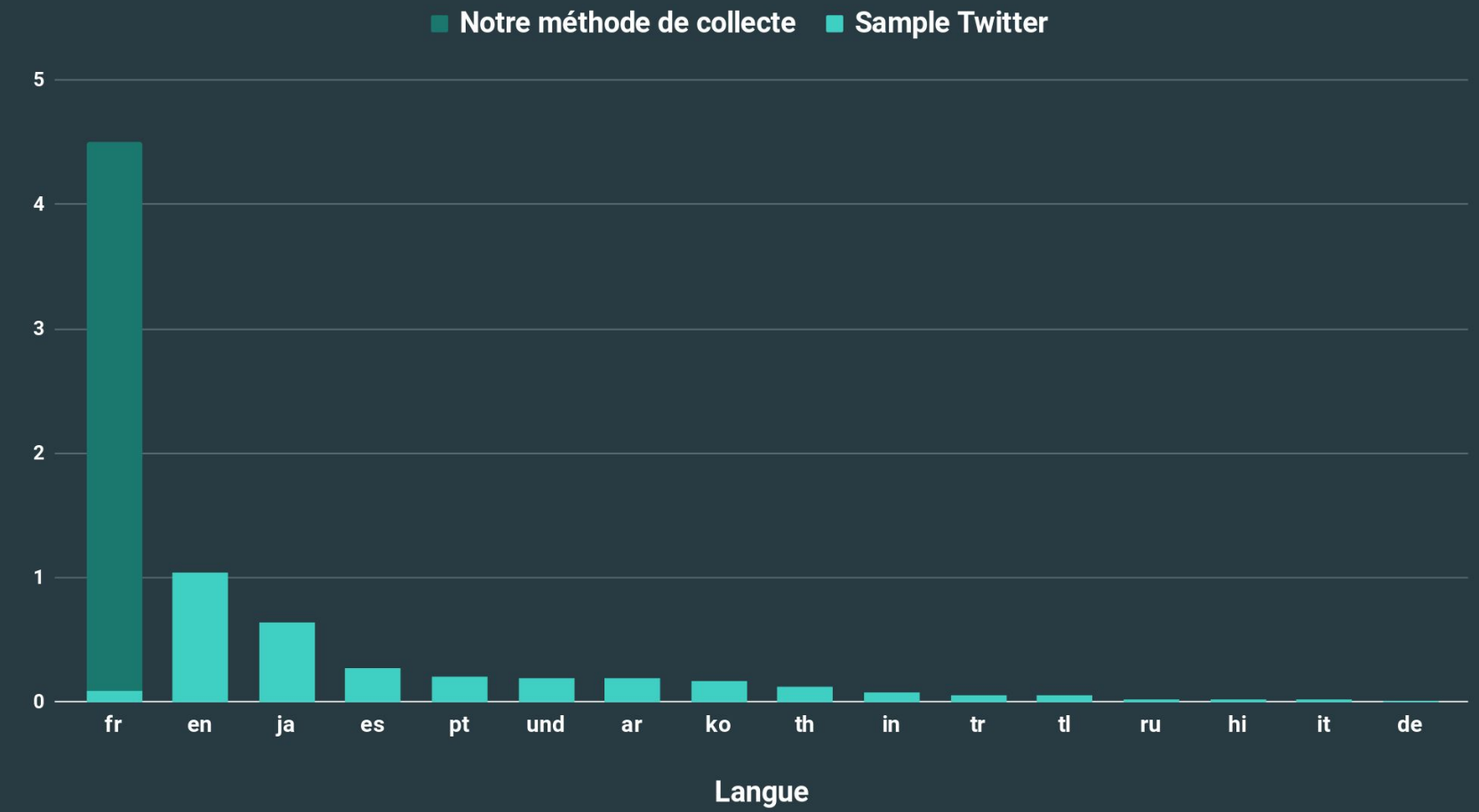
Objectif : couverture maximale des tweets en français sans biais de captation

- Utilisation de plusieurs clés d'accès
  - Dans la limite de ce qui est autorisé par Twitter (1 clé par chercheur sur le projet)
  - Dans la limite de ce qui est gérable pour les resources machine
- Requête de mots clés neutres (stop words)
  - Détermination de ces mots clés par analyse statistique du flux sample
- Répartition des mots clés de filtre sur les clés d'API disponibles
  - Aléatoirement
  - En tenant compte de leur cooccurrence
- Nombre maximal de tweets atteignable pour une clé d'API si requête idéale
  - 1%/1.8% = 55% des tweets français (~ 5 millions sur les 9 millions émis par jour)

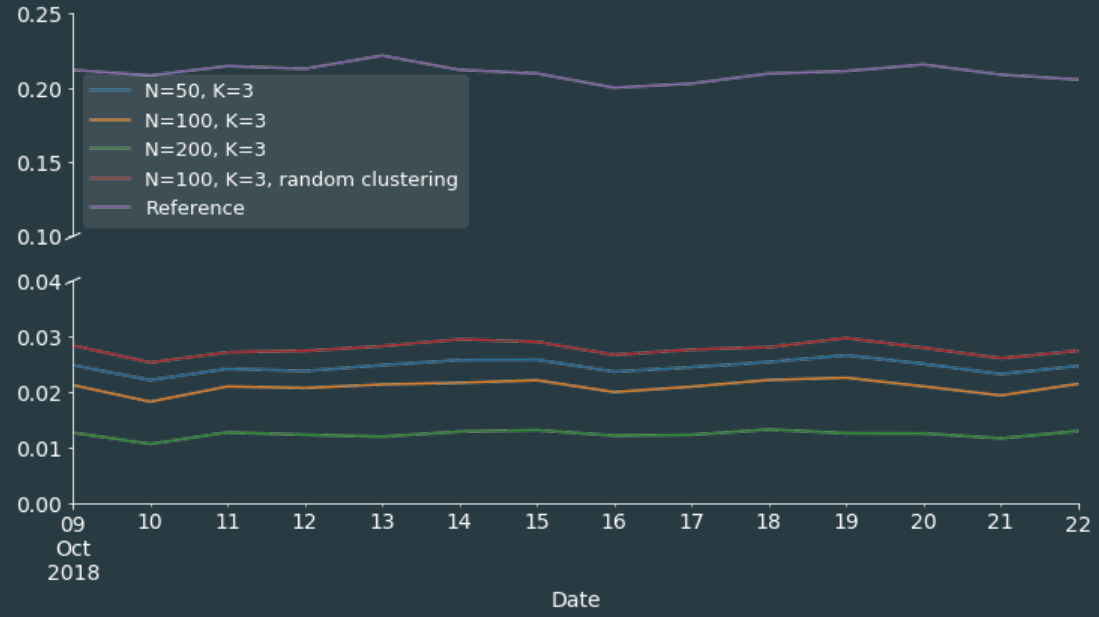
Répartition aléatoire

Répartition selon la cooccurrence

Nombre de tweets moyen par jour (en millions)



Approche proposée : évaluation



Évolution jour par jour de la divergence de Kullback-Leibler entre la distribution des mots dans l'échantillon C_1 et la distribution des mots dans chaque échantillon collecté

$$D_{K,L}(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)}$$

Approche proposée : évaluation

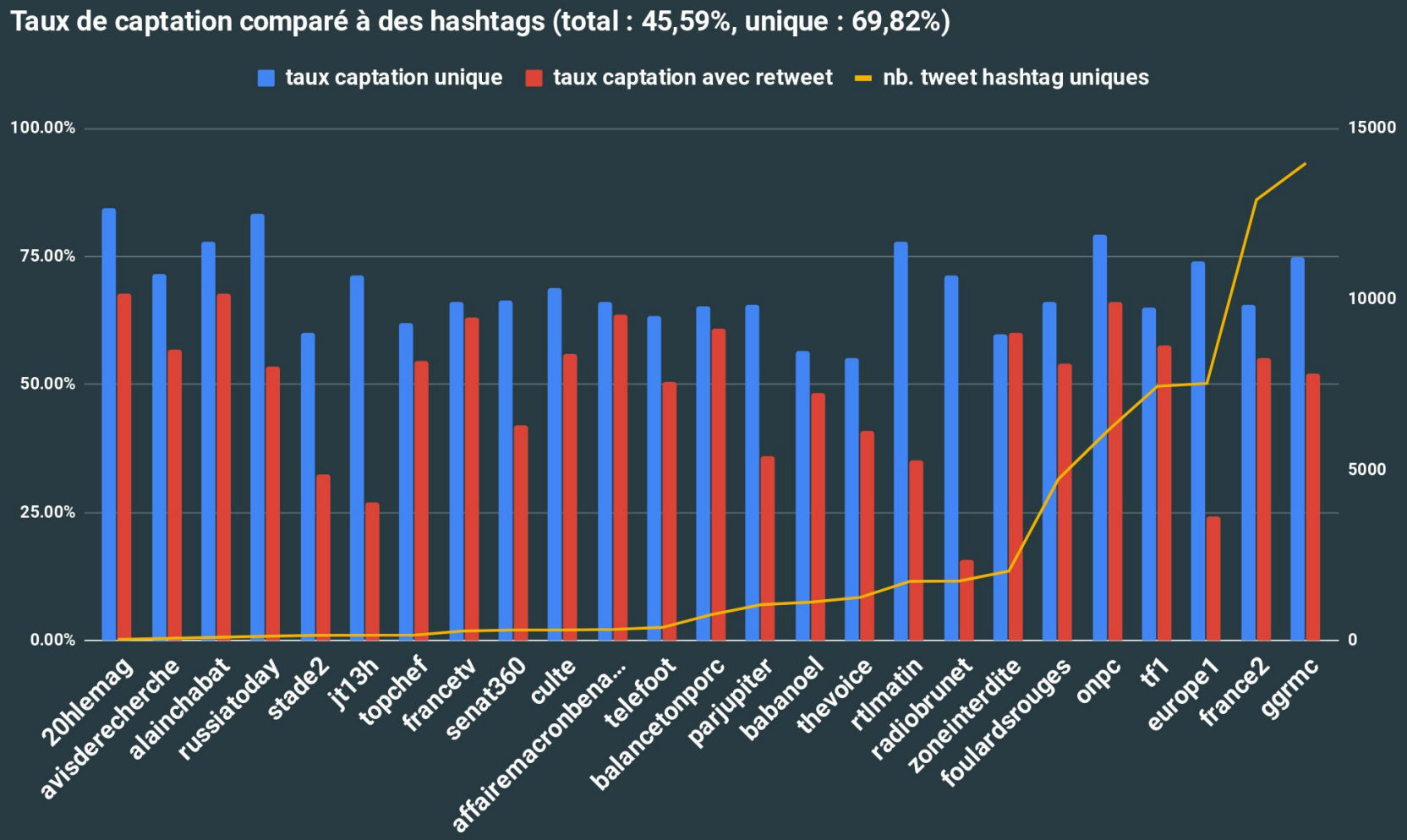
	$C_{1French}$	$C_{3,200} \cup C_{1French}$	$C_{3,100} \cup C_{1French}$	$C_{3,50} \cup C_{1French}$	$R_{3,100} \cup C_{1French}$
Number of characters	**116**	7***	8***	10***	9***
		(0)	(0)	(0)	(0)
Share of retweets	**0.61**	0.02***	0.02***	0.02***	0.02***
		(0.00)	(0.00)	(0.00)	(0.00)
Share of quotes	**0.18**	0.01***	0.01***	0.02***	0.01***
		(0.00)	(0.00)	(0.00)	(0.00)
Share of replies	**0.19**	-0.02***	-0.01***	-0.01***	-0.02***
		(0.00)	(0.00)	(0.00)	(0.00)
Number of URLs	**0.24**	0.01***	0.01***	0.01***	0.01***
		(0.00)	(0.00)	(0.00)	(0.00)
Number of hashtags	**0.24**	-0.01***	-0.05***	-0.06***	-0.05***
		(0.00)	(0.00)	(0.00)	(0.00)
Share of verified users	**0.01**	-0.00*	-0.00	-0.00**	0.00***
		(0.00)	(0.00)	(0.00)	(0.00)
Number of followers	**3,073**	-68	-72	-165*	59
		(74)	(74)	(72)	(76)
Number of friends	**692**	-9*	-34***	-44***	-30***
		(4)	(4)	(3)	(4)
Number of lists	**38**	0	0	-0	2***
		(0)	(0)	(0)	(0)
Observations	**875,630**	63,296,610	59,118,730	64,272,167	55,405,602

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$. Standard errors in parentheses.

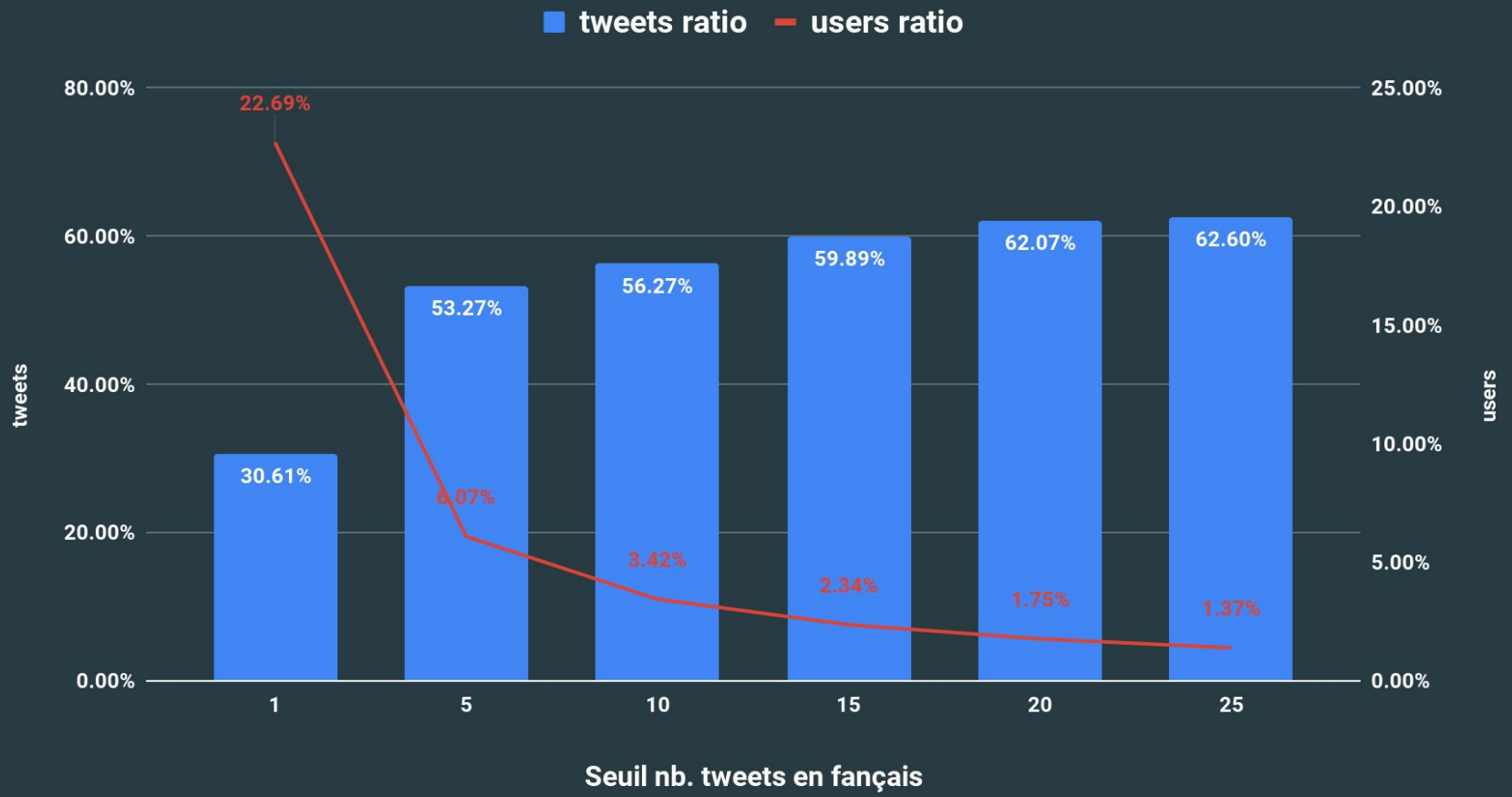
Approche proposée : évaluation

Comment évaluer la couverture de la captation ?

- On ne connaît pas le volume exact de tweets en français
- Tenir compte ou pas des retweets ?
- Étude comparée avec d’autres approches de captation
- Estimation par utilisateur


Taux de captation comparé à des hashtags (total : 45,59%, unique : 69,82%)
taux captation unique
taux captation avec retweet
nb. tweet hashtag uniques
100.00%
75.00%
50.00%
25.00%
0.00%
15000
10000
5000
0
20hlemag
avisderecherche
alainchabat
russiatoday
stade2
jt13h
topchef
francetv
senat360
culte
affairemacronbena...
telefoot
balancetonporc
parjupiter
babanoel
thevoice
rtlmatin
radiobrunet
zoneinterdite
foulardsrouges
onpc
tf1
europe1
france2
ggrmc

Taux de captation par utilisateur (~ 25 millions de tweets)



Approche proposée : limitations

- Tweets sans texte (y compris via sample)
- Identification de la langue parfois incertaine
- Capacité de gestion de gros volumes de tweets :-)

Architecture de la solution de captation