HAL
open science

# Détection de détournements coordonnés de hashtags et messages-images pour l'analyse des fils d'actualités sur les réseaux sociaux

Jean-Marc Francony, Philippe Mulhem, Florence Andreacola, Lorraine Goeuriot, Georges Quénot

## To cite this version:

Jean-Marc Francony, Philippe Mulhem, Florence Andreacola, Lorraine Goeuriot, Georges Quénot. Détection de détournements coordonnés de hashtags et messages-images pour l'analyse des fils d'actualités sur les réseaux sociaux. Modèles et analyse des réseaux : approches mathématiques et informatiques, Oct 2018, Avignon, France. hal-01915150

**HAL Id: hal-01915150**

**https://hal.science/hal-01915150v1**

Submitted on 7 Nov 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Détection de détournements coordonnés de hashtags et messages-images pour l'analyse des fils d'actualités sur les réseaux sociaux

Jean-Marc Francony, Philippe Mulhem, Florence Andreacola, Lorraine Goeuriot, Georges Quénot

November 7, 2018

## 1 Introduction

L'étude de phénomènes et d'évènements sur les réseaux sociaux passe par l'analyse de flux informationnels, dont les unités sont liées généralement par des marqueurs appelés "hashtags". La simple présence de ces marqueurs dans un message suffit à l'inclure dans le flux informationnel. Ainsi, certains contenus contiennent souvent du unités non-liées à la thématique principale du flux dans lequel il s'insère. Cette stratégie opportuniste viserait à profiter d'un flux informationnel dans une logique de visibilité au service d'objectifs commerciaux, militant ou encore malveillant.

Ce travail étudie l'apport de la détection d'une classe spécifique de messages dans les fils d'actualité sur des réseaux sociaux. Cette classe, que nous qualifions de "détournement de hashtag", ou DH, est associée à des publications sans lien sémantique avec le flux originel. Cette méthode de diffusion issue du marketing viral affecte la qualité des analyses automatiques et l'interprétation des flux d'information thématiques. Reposant parfois sur des techniques difficilement repérables, le DH peut également traduire des formes coordonnées de publication mobilisant un réseau d'acteurs selon une logique d'action. Dans ce travail, nous nous intéressons au cas du détournement "par l'image" ou DHI du flux d'information. C'est-à-dire que le sens des images utilisées est dénuée de lien avec les hashtags qui lui sont associés. Il s'agit d'une technique selon laquelle le détournement du contenu du message est supporté principalement par la composante image. Dans un tel cas, les approches fondées sur le traitement des textes ou la détection de communautés ne s'appliquent pas [3]. Plus spécifiquement, nous nous intéressons au cas particulier d'un DHI coordonné, c'est-à-dire relevant d'une action concertée au sein d'un réseau d'acteurs. Au delà de la détection des DH, nos travaux suggèrent d'accorder une place plus importante à l'étude des flux d'images dans l'analyse des fils d'actualités de Twitter, au bénéfice notamment de la compréhension des interactions sociales.

Dans cet article, nous présentons les corpus étudiés et nous montrons en particulier que la non-détection des DH pose le problème de la qualité d'extraction de statistiques sur des corpus réels de Tweets, comme les rafales de messages ou les thématiques. Nous présentons ensuite l'utilisation de deux méthodes simples de détections de duplicats images comme moyen d'identifier des réseaux de diffusion coordonnée, et discutons les détections proposées, avant de conclure.

# 2 Données sociales et militantisme scruté

## 2.1 Définition du cas

À l'occasion de l'ANR *Responsabilité Sociale des Journalistes - Médias, Diversité et Sport*, nous avons constitué deux corpus liés à des événements sportifs majeurs : les jeux olympiques de RIO en 2016 et les championnats du Monde d'Athlétisme de Londres en 2017. Mobilisant les principaux médias nationaux durant deux semaines sur de grandes plages de directs, au bénéfice d'une large audience internationale, ces événements médiatiques sont à l'origine d'une activité spécifique clairement identifiée sur les réseaux sociaux par des hashtags mis en avant pour l'occasion. Les hashtags #RIO2016 ou #LONDON2017 (voire également #LONDRES2017) ont constitué la trame officielle canalisant ces flux événementiels ; d'autres hashtags thématiques, tels que #JO2016 ou #EuroAthletics, etc. apparaissant comme des déclinaisons. Les flux importants d'actualité que ces événements agrègent en particulier sur Twitter en font les cibles privilégiées d'un détournement d'audience. Assimilable aux enjeux du *growth hacking* dans le marketing digital, l'objectif de ces techniques est d'utiliser le canal de diffusion associé à l'événement médiatique en cours pour maximiser la portée du message. Dans le cas présent, les deux corpus réalisés à un an d'intervalle ont fait l'objet d'une campagne de détournement de hashtag produit par un même collectif associé à la cause environnementale : *TiredEarth*[1]. Le compte TiredEarth se définit sur Twitter comme : *"an international group trying to improve awareness and increase the hope for a shiny future"*. L'identité collective défendue par TiredEarth ainsi que les modalités d'action en réseau que nous avons identifiées ressortent du militantisme. Le dispositif de publication mis en oeuvre en soutien de l'action militante sur Twitter se base sur un ensemble de comptes qui émettent des Tweets selon des flux sans pics prononcés, en respectant les cycles circadiens. Les textes des messages publiés sont exclusivement composés de hashtags d'actualité parmi les plus populaires (possiblement tronqués pour proposer de multiples variations), et d'une image porteuse d'un message lié à la pollution et la défense de l'environnement. Comme le message est contenu dans l'image et que les messages textes sont anodins, ces Tweets ne sont pas identifiés *a priori* comme des *spams* et se diffusent sur les fils d'actualité. Bien qu'aucune régularité n'apparaisse dans la logique de fonctionnement du dispositif de publication, la continuité de l'activité alternée entre les différents compte nous laisse penser qu'il s'agit d'un programme. Compte tenu des règles de bonnes pratiques que promeut Twitter, il est surprenant que de tels agissements contrevenant à ces principes se soient maintenus à un an d'intervalle[2]. Deux explications ressortent de nos études (ci-après) : le fait que le message effectif des publications soit inscrit dans une infographie ; l'existence d'un réseau de publication coordonnée qui dilue la responsabilité individuelle. En effet, porté par une banque d'images, les caractéristiques précédentes sont à l'origine d'un dispositif de publication dont la logique de fonctionnement respecte une lecture individuelle des conditions générales d'usage alors que le fonctionnement global les détourne. Partant de ce constat, l'objectif de notre travail est d'être en mesure de repérer au plus tôt le phénomène de DHI et d'identifier le réseau d'acteur contribuant à la diffusion.

[1] https://twitter.com/tiredearth

[2] Tous les comptes concernés dans le réseau de diffusion identifié dans notre étude sont désormais suspendus depuis août 2018

## 2.2 Jeu de données

Le corpus de référence associé à l'événement #London2017 est composé de 252 842 Tweets collectés du 3 au 15 août 2017. Le nombre d'images uniques associées au corpus est de 4 243 pour 7 389 fichiers diffusés ou rediffusés dans 80 961 tweets distincts, parmi lesquels 2 473 fichiers images (33.5%) correspondent à 76 images uniques produites par *Tiredearth*. Nous postulons que les caractéristiques des flux d'images Tweetées peuvent soutenir la détection du phénomène de DH.

A partir ce des données initiales, une étape initiale d'érosion est réalisée de la manière itérative suivante, jusqu'à convergence :

1. Dans un premier temps on commence par éliminer les utilisateurs qui ont émis un seul tweet. Ces utilisateurs ne sont pas des activistes d'après le processus utilisé;

2. Dans un second temps nous ne considérons que le sous-ensemble des tweets-images émis par les utilisateurs potentiellement activistes (cf. étape précédente). Nous considérons exclusivement les images publiées plus d'une fois. Une fois ces tweets-images retirés, il est alors possible que des utilisateurs restants n'aient envoyé qu'un tweet, on revient alors à l'étape 1 ci-dessus.

3. Si à la fin de l'étape 2 il n'y a aucun tweet-image de retiré, alors nous avons convergé et c'est sur les ensembles utilisateurs/tweets-images restants que l'on va se concentrer.

A titre d'exemple, sur notre corpus initial de 7031 tweets contenant une image, envoyés par 2391 utilisateurs pour un total de 4201 images uniques, le processus d'érosion permet de ne considérer finalement que 2722 tweets (ensemble noté $E_{tw}$), 122 utilisateurs (ensemble noté $U_{tw}$), c'est-à-dire 5,10% d'entre eux, et 297 images (ensemble noté $I_{tw}$), c'est-à-dire 7,07% d'entre elles.

Notre proposition est ainsi de produire des clusters d'images associés à des clusters de comptes. Pour évaluer les résultats, nous utilisons les deux mesures classiques de recherche d'information, le rappel et la précision par rapport à cet ensemble de 2473 images. ** PM : à modifier car pas cohérent avec au dessus **

**LG : trouver un article décrivant l'architecture classique d'un réseau et la proportion classique d'utilisateurs inactifs ou très peu actifs (followers et relayeurs uniquement) Permet de justifier l'érosion : ne nous intéressant qu'aux flux informationnels, nous excluons les utilisateurs non acteurs de notre corpus

# 3 Détection d'activisme par des données images

Le modèle d'activisme traité ici correspond à un ensemble de comptes Twitter qui émettent les mêmes images, mais sans retweeter les messages. Le principe que nous utilisons pour la détection de ces groupes se base donc sur les images associées au messages, en négligeant les textes de ces messages car ils agissent comme simple véhicule et ne sont pas significants. Comme nous avons constaté que les images envoyées sont exactement les mêmes, il est dès lors possible de se reposer sur des signatures de ces images. La signature considérée dans ce travail repose sur le MD5 (Message Digest 5, RFC1321[3]). Cette signature a le mérite d'être très connue, d'être calculée facilement par de nombreux logiciels, et

[3] http://www.ietf.org/rfc/rfc1321.txt

d'être même proposée par défaut lors de l'acquisition de flux d'informations des tweets. Nous nous focalisons sur deux catégories de données : les utilisateurs et les images. Plus précisément, nous cherchons donc ici à détecter les groupes d'utilisateurs activistes en nous basant sur les images qu'ils ont envoyés.

Pour cela, nous étudions l'utilisation de deux catégories d'approches : l'une, basée sur du clustering, utilise une représentation d'une catégories par l'autre (i.e., représenter les utilisateurs par les images de leurs tweets émis, ou bien représenter les images par les utilisateurs qui les ont émis dans au moins un tweet), l'autre, le co-clustering, se base sur l'utilisation conjointe de ces catégories, en représentant d'une manière unique les relations entre utilisateurs et images. Nous décrivons dans la suite ces deux approches.

## 3.1 Mesure d'évaluation

Nous décrivons ici comment nous avons choisi d'évaluer la qualité de la détection des utilisateurs activistes. Après une éude manulle du corpus d'utilisateurs $U_{tw}$, nous avons déterminer le sous ensemble $A_{tw}$ ($\subset U_{tw}$) des activistes de TiredEarth. La cardinalité de $A_{tw}$ est égale à 19. Nous allons donc utiliser des approches utilisant les données $E_{tw}$ émises par $U_{tw}$ pour les images de $I_{tw}$, en supposant que le résultat idéal est d'obtenir une partition de $U_{tw}$ composée de 2 clusters : l'un correpond aux activistes $A_{tw}$, et le second, $NA_{tw}$ est composé des autres utilisateurs de $U_{tw}$, ce qui donne $NA_{tw} = U_{tw} \setminus A_{tw}$.

Ce ensemble de clusters étant établi, il faut ensuite utiliser une mesure d'accord entre deux clusterings, pour savoir si le résultat d'un clutering automatique correspond ou non à l'idéal. Pour cela, nous utilisons mesure classique, l'indice de Rand [2] : considérons un ensemble de N éléments sur lequel deux partitions Y et Y' sont exécutées, l'indice de Rand utilise tous les couples d'éléments possibles ($=C_2^n$), et évalue combien de fois les couples sont dans le même cluster dans chaque partition Y et Y', et combien de fois ils sont dans des clusters différents dans Y et Y'. Cette valeur est normalisée dans l'intervalle $[0, 1]$ : la valeur de 1 dénote un accord parfait entre les partitions Y et Y', et une valeur de 0 un désaccord total entre les partitions. Dans notre cas, nous allons comparer la partition idéale avec celle obtenue par clustering.

## 3.2 Détection par clustering hiérarchique** A GARDER ? **

Dans cette étape, nous utilisons un simple clustering hiérarchique utilisant, pour chaque classe MD5 regroupant les images de même signature, l'ensemble des utilisateurs ayant envoyés des images de cette classe. Le résultat obtenu (en utilisant les paramètres suivants : seuil : 10, distance(c1,c2) : -jaccard(c1.users, c2.users)) est capable de regrouper les images envoyés par les mêmes personnes, avec un taux de précision de 99,80% et un taux de rappel de 100%, reflétant le fait que 2468 des 2473 images du cluster cibles sont retrouvés. Il est dès lors possible de retrouver par exploration les images des activistes, et donc le groupe.

Cette approche possède un désavantage qui est que les paramètres sont difficiles à être optimisés et/ou adaptés. Par exemple, le nombre de clusters obtenu est le nombre fixé au départ, et le fait d'agglomérer les clusters dans un second temps permet de modifier a posteriori le nombre obtenu.

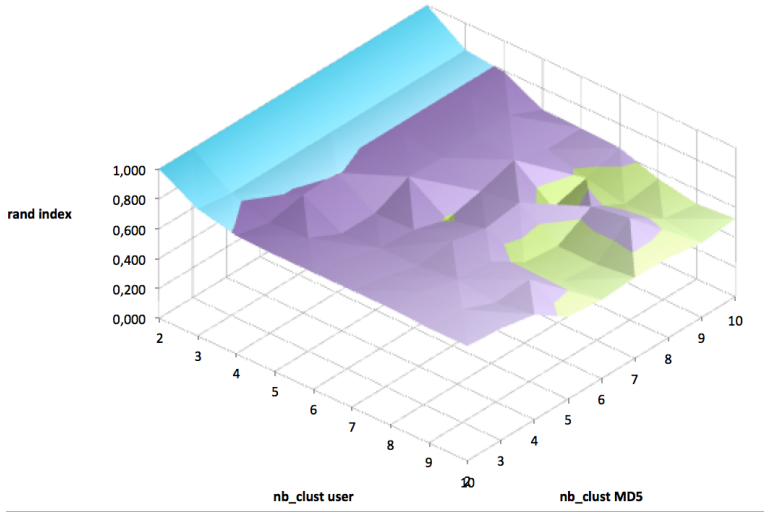


Figure 1: Indice Rand suivant les paramètres de nombre de clusters utilisateurs et nombre de clusters images.

### 3.3 Détection a posteriori par co-clustering

La détection par co-clustering repose sur une représentation matricielle binaire classe MD5 × utilisateur. Elle a pour principe de regrouper conjointement les images envoyées et les utilisateurs qui envoient les images. Pour utiliser ce co-clustering, nous utilisons l'approche à base de mixtures de blocs proposée par [1], en utilisant l'implantation de R Blockcuster 4.0[4] avec les paramètres par défaut. L'idée de base est de proposer des permutations de lignes et de colonnes amenant à des clusters qui tiennent compte à la fois des lignes et de colonnes. Elle repose sur des distributions de probabilités de Bernoulli, et utilise des maximisation de l'espérance alternativement sur les lignes et les colonnes. Ce co-clutering a en particulier des paramètres pré-estimant le nombre de clusters pour les utilisateurs et le images.

Afin d'étudier le comportement de cette détection nous avons choisi de réaliser des clusterings an faisant varier les valeur des clusters utilisteurs et images de 2 à 10, et de comparer par rapport à l'idéal décrit plus haut par l'indice de Rand. On suppose que cette éxpérimentation soit donner de meilleurs résultats quand le nombre de clusters utilisateurs attendu est de deux, mais ele doitêtre vérifié expérimentalement car le co-clustering est capable de s'adapter aux données et ne génère pas forcément le nombre attendu de clusters. Les résultats obtenus sont présentés en figure 1.

On constate sur la figure 1 que l'indice Rand des clusters obtenus pour les utilisateurs est de 1, quelque soit le nombre de clusters images recherché entre 2 et 10. Ce résultat est donc le clustering d'utilisateur idéal : tous les activistes sont regroupés, et tous les autres sont dans le second cluster. Par contre, dès que l'on demande plus de deux clusters utilisateurs, les scores d'indice Rand ne sont plus égaux à 1 : par exemple pour 3 clusters utilisteurs et 2 clusters images, l'indice de Rand obtenu est égal à 0,852. On constate sur cette figure que l'impact du nombre de clusters images et réduit : pour une valeur de nombre de clusters utilisteurs les valeurs sont assez stables, en particulier pour un nombre de clusters utilisateurs entre 2 et 5. Pour des nombres de clusters utilisateurs plus élevées, l'augmentation du nombre de cluster images est négatif : par exemple, pour

[4]https://CRAN.R-project.org/package=blockcluster

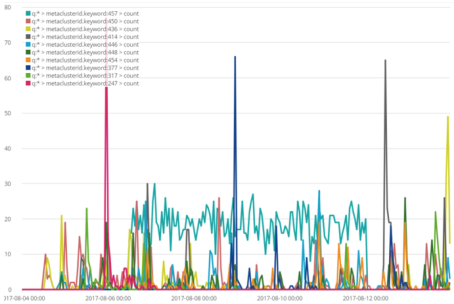

Figure 2: La prodution de tweets avec images par heure.

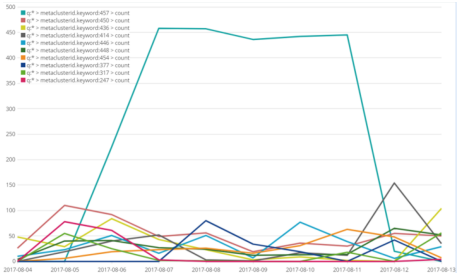

Figure 3: La prodution de tweets avec images par heure.

un nombre de clusters utilisateurs de 10, l'indice de Rand est égal à 0,769 pour 2 clusters images te 0,524 pour 10 clusters images.

* Dans discussion * On conclut donc de cette étude que l'utilisation de co-clutering avec une cible de 2 clusters pour les utilisateurs permet de détecter avec une très bonne qualité cette forme d'activisme, si on se base sur l'ensemble de données a posteriori.

### 3.4 Discussion

Les deux approches listées ci-dessus permettent d'obtenir le même résultat, c'est-à-dire un cluster regroupant toutes images du groupe d'activistes. Dans tous les cas, l'automatisation complète de cette détection n'est pas encore permise, car pour le moment nous ne pouvons retrouver les activistes si nous possédons déjà au moins une image (ou un compte utilisateur) d'activiste. Ce problème de "démarrage à froid" n'est pas traité ici.

## 4 Importance de la détection d'activisme sur l'étude d'un corpus de microblogs

Classiquement, des statistiques sont extraites des corpus afin de définir des axes d'analyse pour étudier un réseau social d'un point de vue sociologique. Si le corpus est bruité par des messages non pertinents, ici, ceux d'activistes, les statistiques peuvent être biaisées, amenant à des conclusions erronées sur le corpus considéré.

Il est courant, pour étudier un réseau, de se baser sur les pics de production (les rafales), afin de définir des seuils d'étude. Dans le cas qui nous intéresse, la prise en compte de la détection de activistes fournit des éléments très différents par rapport au données brutes (i.e., sans détection), comme le montrent les flux par heure dans la figure 2 et les flux par jour en figure 3.

Il résulte de ces figures qu'il est nécessaire de permettre une exploration la plus simple possible afin de faciliter le retrait de ces données d'activisme.

## 5 Conclusion

Le travail présenté visait à étudier des méthodes pour "nettoyer" un corpus de microblogs en se basant sur des éléments non-textuels. Cette détection a pour objectif de retirer les

microblogs et/ou utilisateurs opportunistes qui utilisent un événement populaire comme support afin occuper l'espace médiatique. Nous nous basons sur les signatures MD5 des images pour retrouver ces microblogs, et nous utilisons des techniques de clustering et de co-clutering pour la détection. Ces deux méthodes permettent de retrouver le groupe d'utilisateurs activistes. L'intérêt de cette détection est la suivante : le retrait des activismes détectés a un impact important sur les statistiques extraites pour les analyser : ne pas les intégrer biaise alors les analyses des réseaux sociaux d'un point de vue sociologique.

Ne peut-on pas monter en généralité ici par rapport aux activistes et parler de "spammeur", de "newsjaskeur"... FA

# References


[1] Gérard Govaert and Mohamed Nadif. Clustering with block mixture models. *Pattern Recognition*, 36(2):463–473, 2003.

[2] William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[3] Tingmin Wu, Sheng Wen, Yang Xiang, and Wanlei Zhou. Twitter spam detection: Survey of new approaches and comparative study. *Computers & Security*, 76:265–284, 2018.