Analyse des comportements des utilisateurs de Twitter durant la campagne des élections européennes de 2014

Vincent Brault, Jean-Marc Francony, Adeline Leclercq-Samson, Matthieu Meynet

Analyse des comportements des utilisateurs de Twitter durant la campagne des élections européennes de 2014

Vincent Brault [1] & Jean-Marc Francony [2] & Adeline Leclercq-Samson [1] & Matthieu Meynet [3]

[1] *Univ. Grenoble Alpes, LJK, F-38000 Grenoble, France*
Univ. Grenoble Alpes, CNRS, F-38000 Grenoble, France
vincent.brault@univ-grenoble-alpes.fr et adeline.leclercq-samson@univ-grenoble-alpes.fr
[2] *Univ. Grenoble Alpes, UMR PACTE, F-38000 Grenoble, France*
jean-marc.francony@univ-grenoble-alpes.fr
[3] *Univ. Grenoble Alpes, GRESEC, F-38000 Grenoble, France*
matthieu.meynet@univ-grenoble-alpes.fr

Résumé. Dans cette présentation, nous proposons une analyse des comportements des utilisateurs de Twitter durant la campagne des élections européennes de 2014 dans le but de former des groupes d'utilisateurs qui avaient tendance à faire des actions aux mêmes moments durant cette campagne. Pour ce faire, nous effectuons une classification des jours de la campagne électorale puis, à l'aide de la classification obtenue, nous proposons une classification des utilisateurs dans le but d'étudier leurs comportements. Nous concluons cette présentation par une analyse des résultats obtenus.

Mots-clés. classification non supervisée, algorithme des K *means*, classification imbriquée, Twitter

Abstract. In this talk, we provide an analysis of the behavior of Twitter users during the 2014 European elections campaign. Our goal is forming user groups that tended to do the same action at the same time during this campaign. We propose a classification of the days of the electoral campaign and, using the classification obtained, we propose a classification of the users in order to study their behaviors. We conclude this talk with an analysis of the results obtained.

Keywords. unsupervised classification, K *means* algorithm, nested classification, Twitter

1 Introduction

Twitter sert de plateforme de réseau social (ou bien encore site de microblogage) qui a progressivement été investie par les agents du champs politique (hommes politiques, journalistes...) depuis 2008 et l'élection de Barack Obama (voir par exemple [4]). Dans ce

travail, nous étudions les *retweets* (c'est-à-dire les *tweets* qui citent d'autres *tweets*) et les *mentions* dans un tweet (les *tweets* dans lesquels sont mentionnés d'autres utilisateurs de Twitter) durant la campagne des élections européennes de 2014 dans le but de voir s'il y a des consignes au sein des partis et/ou des stratégies de communication militantes (voir par exemple [2]).

Pour ce faire, nous avons proposé une stratégie de classification des utilisateurs faite en deux temps : nous avons commencé par faire une classification des jours afin que les événements exceptionnels (comme le jour de l'élection) ne perturbent pas l'analyse des comportements puis une classification des utilisateurs afin d'étudier les comportements communs dans chaque groupe.

2 Données, modélisation et notations

Dans cette étude, nous disposons des *retweets* et *mentions* de $n = 31\,653$ utilisateurs comportant certains hashtags particuliers (comme **#EE2014** ou **#Europeennes2014**) durant la campagne de l'élection européenne 2014. Plus précisément, la période étudiée s'étale du 27 avril 2014 au 2 juin 2014 ; soit un peu avant le début officiel de la campagne et jusqu'à une semaine après le vote du dimanche 25 mai 2014 (soit un total de $J = 37$ jours). Notons qu'un même tweet peut contenir plusieurs *mentions* et donc faire référence à plusieurs utilisateurs, dans ce cas, le tweet en lui même sera compté pour chaque *mention*. Au total, nous avons 196 247 *retweets* et 297 617 *mentions*.

Afin de contourner le caractère instantané d'un tweet, nous avons choisi de regrouper les informations par plages horaires d'une heure et nous noterons $N^R_{i,j,h}$ (resp. $N^M_{i,j,h}$) le nombre de *retweets* (resp. *mentions*) faits par l'utilisateur $i \in \{1, \dots, n\}$ durant l'heure $h \in \{1, \dots, 24\}$ du jour $j \in \{1, \dots, J\}$.

Ainsi, le vecteur $N^R_{i,\cdot,h}$ représente tous les *retweets* faits par l'utilisateur i à l'heure h de chaque jour de la période étudiée. S'il a un comportement journalier régulier sur Twitter, toutes les valeurs de ce vecteur devraient être proches.

3 Classification

La classification se fait en deux temps : d'abord la classification des jours puis, étant donnée une classification des jours, la classification des utilisateurs.

Dans la suite, nous présenterons les procédures sur les *retweets* sachant que les *mentions* ont été traitées de la même façon.

3.1 Classification des jours

Nous avons d'abord cherché à classifier les jours en fonction du comportement global des utilisateurs. Pour cela, nous avons fait la somme sur tous les *retweets* de tous les utilisa-

teurs pour chaque heure de chaque jour, noté $N^R_{+,j,h}$. Nous obtenons ainsi 37 individus dans $\mathbb{N}^{24}$ et nous avons utilisé un algorithme des *K-means* (voir par exemple [3]) afin de regrouper les jours.

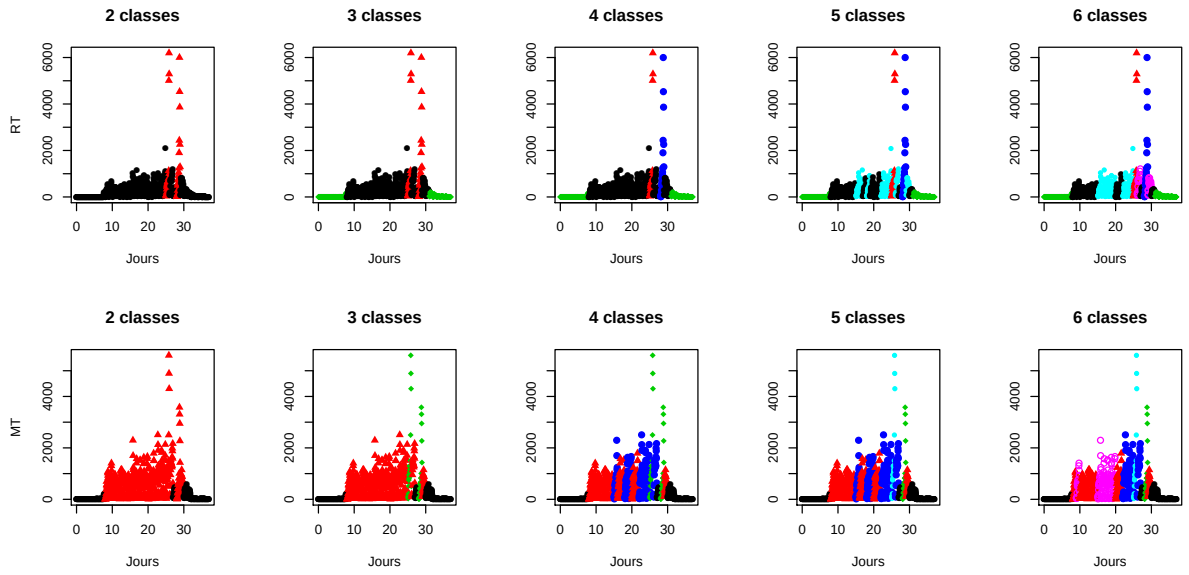

Figure 1: Représentation des classes de jours obtenues pour les *retweets* (ligne du haut) et les *mentions* (ligne du bas) pour un nombre de classes allant de 2 à 6 (colonnes) : pour chaque graphique, l'abscisse est mis à partir du premier jour d'étude (jour 0), chaque point représente une heure et les couleurs symbolisent les classes.

Nous pouvons voir sur la figure **??** les répartitions des classes. Les deux pics des *retweets* correspondent respectivement au jour du débat télévisé *Des paroles et des actes* sur France 2 (le 22 mai) et le jour de l'élection (le 25 mai). Nous voyons que l'algorithme a très vite séparé ces jours puis il a regroupé les jours d'avant et d'après la période électorale et a enfin partitionné progressivement les jours de la campagne (par exemple en séparant les jours suivants les deux jours importants de la campagne). À l'opposé, il semblerait que le jour de la semaine n'intervienne pas trop dans le comportement des utilisateurs de Twitter.

Du point de vue de l'analyse, il a été choisi de se concentrer sur 5 classes de jour. Toutefois, nous avons étudié également les cas avec un nombre de classes différent.

3.2 Classification des utilisateurs

Étant donnée une classification des jours, nous avons cherché à classifier les utilisateurs en groupes de telle sorte que :

- les utilisateurs d'une même classe aient des comportements similaires et différents de ceux des utilisateurs d'autres classes,
- pour une classe de jours donnée, chaque utilisateur ait le même comportement chaque jour de cette classe (par exemple, il va avoir l'habitude de *retweeter* vers 10h tous les jours d'une même classe).

Une fois ces choix faits, nous avons pris plusieurs points de vue pour diversifier les analyses :

- Entre 1h et 6h du matin, il y a peu d'activités. Nous avons regardé lorsque les actions durant ces heures sont concaténées ou non : cela signifie que la nuit peut soit compter pour une seule plage horaire, soit pour plusieurs.
- De même, certaines classes de jours comportent une seule journée et d'autres jusqu'à 14 jours. Dans le calcul des distances, nous avons le choix de considérer que chaque jour est unique (et donner plus de poids aux classes avec beaucoup de jours) ou de pondérer en considérant chaque classe avec le même poids (et redonner de l'importance aux jours dans les petites classes).
- Enfin, certains utilisateurs vont avoir tendance à beaucoup plus interagir que d'autres ce qui peut entrainer des classes avec un ou deux utilisateurs. Pour contrer ce problème, nous pouvons choisir de renormaliser par le nombre total de *retweets* faits par chaque utilisateur : dans ce cas, nous regarderons le comportement de chaque utilisateur *si nous avions fixé son nombre maximal d'utilisations* ; une autre façon de le voir est d'imaginer qu'il n'a le droit qu'à 100 *retweets* et de se demander quand va-t-il en faire.

Au final, nous avons donc 8 plans pour chaque type de données pour chaque partition de jours.

3.3 Évaluation des résultats

À la vue du nombre de plans, il est difficile de présenter les résultats en quelques graphiques ; nous avons donc créé une application shiny pour une meilleure visibilité. Nous avons commencé par regarder les profils des utilisateurs et nous pouvons faire quelques constatations générales :

- Lorsque les données ne sont pas normalisées, les utilisateurs *intensifs* sont mis dans des petits groupes.
- Lorsque les données sont normalisées, nous retrouvons des groupes d'utilisateurs qui ont eu une activité plutôt le soir du débat et/ou le soir de l'élection.

- Dans les deux cas, il reste presque tout le temps un groupe d'utilisateurs peu actifs et représentant plus de la moitié des personnes étudiées.

Nous avons ensuite étudié les répartitions des professions liées à la politique (journalistes, candidat·e·s...) ainsi que les étiquettes politiques (issues d'un travail d'indexation d'éléments contenus dans l'espace biographie de Twitter) dans chaque groupe et nous avons remarqué que tout semblait réparti plus ou moins équitablement ; ce résultat nous amène à penser que l'hypothèse consistant à dire : "ces individus qui *retweetent* au même moment affichent une proximité en terme d'engagement militant" ne semble pas vérifiée.

Enfin, pour apprécier la modélisation, nous avons regardé le graphe des échanges de tweets pour les classes ayant des pics lors du débat ou du jour de l'élection à l'aide de la plate-forme *Linkage*[1] appliquant les méthodes proposées par Bouveyron et al. [1]. Nous avons pu voir, par exemple, des groupes se former car les utilisateurs *retweetaient* les estimations proposées par *Radio Londres* qui a le droit de donner les résultats avant 20h puisque non soumis aux lois françaises.

4 Présentations

Dans cet exposé, nous présenterons la méthode utilisée puis les résultats obtenus à l'aide de notre application shiny. Nous conclurons par les nombreuses perspectives de ce travail.

Bibliographie

[1] Bouveyron, C., Latouche, P., & Zreik, R. (2018). *The stochastic topic block model for the clustering of vertices in networks with textual edges.* Statistics and Computing, 28(1), 11-31.
[2] Bruns, A. (2012). *Journalists and Twitter: How Australian news organisations adapt to a new medium.* Media International Australia, 144(1), 97-107.
[3] Hartigan, J. A. (1975). *Clustering algorithms.* John Wiley & Sons, Inc., New York, NY, USA, 99th edition, 1975. ISBN 047135645X.
[4] Heinderyckx, F. (2011). *Obama 2008: l'inflexion numérique.* Hermès, La Revue, (1), 135-136.

[1]Disponible sur `www.linkage.fr`