



HAL
open science

#Idéo2017: une plateforme citoyenne dédiée à l'analyse des tweets lors des événements politiques

Claudia Marinica, Julien Longhi, Nader Hassine, Abdulhafiz Alkhouli, Boris Borzic

► **To cite this version:**

Claudia Marinica, Julien Longhi, Nader Hassine, Abdulhafiz Alkhouli, Boris Borzic. #Idéo2017: une plateforme citoyenne dédiée à l'analyse des tweets lors des événements politiques. Extraction et Gestion des Connaissances (EGC), Jan 2018, Paris, France. hal-01699423

HAL Id: hal-01699423

<https://hal.science/hal-01699423v1>

Submitted on 2 Feb 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.

#Idéo2017 : une plateforme citoyenne dédiée à l'analyse des tweets lors des événements politiques

Claudia Marinica^{*,***}, Julien Longhi^{**}, Nader Hassine^{*,**}
Abdulhafiz Alkhouli^{*}, Boris Borzic^{*}

^{*}ETIS UMR 8051, Université Paris Seine, Université de Cergy-Pontoise, ENSEA, CNRS
F-95000, Cergy, France

{claudia.marinica, boris.borzic, abdulhafiz.alkhouli}@ensea.fr

^{**}AGORA, 33 Bd du Port, 95000 Cergy-Pontoise, France

nader2hassine@gmail.com, julien.longhi@u-cergy.fr

^{***}Equipe DUKe, Laboratoire des Sciences du Numérique de Nantes, UMR CNRS 6004
Nantes, France

Résumé. Cette plateforme a pour objectif de permettre aux citoyens d'analyser par eux-mêmes les tweets politiques lors d'événements spécifiques en France. Pour le cas de l'élection présidentielle de 2017, #Idéo2017 analysait en quasi temps réel les messages des candidats, et fournissait leurs principales caractéristiques, l'usage du lexique politique et des comparaisons entre les candidats.

1 Introduction

Les réseaux sociaux font partie du quotidien, notamment en ce qui concerne la « consommation » de l'information (Mercier (2014)). Le service de microblogging Twitter a permis aux réseaux sociaux de prendre une nouvelle dimension car il permet d'évoluer les réactions de ses utilisateurs sur des sujets sociaux (Longhi et al. (2016)), politiques (Conover et al. (2011)), etc.

L'analyse des tweets politiques lors des campagnes électorales ou d'événements spécifiques peut être considéré comme un genre de discours politique spécifique (Longhi (2013)). Dans ce contexte, Roginsky et Cock (2015) proposent une analyse qualitative d'interactions sur Twitter, mais se limitent à une analyse discursive et communicationnelle. De plus, des nombreux travaux ont étudié la prédiction du résultat des élections présidentielles (ou pourquoi cela n'est pas possible) (Tumasjan et al. (2010), Gayo-Avello et al. (2011)).

Ces travaux, relevant des sciences informatiques ou des sciences humaines et sociales, se font parfois difficilement écho. En outre, ces résultats sont difficilement accessibles pour les citoyens, et les analyses accessibles sont déjà agrégées par les médias, ou elles intègrent des traitements simples (comme le compte @TwitterFrance).

Cet article présente la plateforme #Idéo2017 qui, innovante par l'accès à l'information qu'elle permet aux citoyens, a pour objectif de proposer un outil d'analyse des tweets et des prises de parole (relayées sur Twitter) d'un ensemble de comptes Twitter emblématiques lors d'événements politiques spécifiques (par exemple, les comptes officiels des 11 candidats pour l'élection présidentielle en France en 2017). Dans le cadre des élections, #Idéo2017 traitait les

#Idéo2017 : une plateforme citoyenne dédiée à l'analyse des tweets politiques

tweets des candidats en constituant un corpus en quasi temps réel (mis à jour chaque 24 heures). Par l'utilisation d'outils et de fonctionnalités issues de la linguistique outillée, ce traitement peut souligner les principales caractéristiques de ce corpus, et permettre des comparaisons entre les différents candidats. De plus, #Idéo2017 permet également une navigation à facettes dans les tweets et une visualisation graphique intuitive, ainsi que l'extraction des corpus.

2 Description de la plateforme #Idéo2017

Dans cette section, nous décrivons les fonctionnalités de la plateforme, puis, nous montrons un cas d'utilisation et ensuite nous présentons les contraintes de développement.

2.1 Description des fonctionnalités

#Idéo2017 propose cinq fonctionnalités qui regroupent plusieurs traitements. La plateforme est adaptable à divers événements politiques et elle a été testée dans le cadre de l'élection présidentielle 2017 avec les comptes des 11 candidats officiels (<http://ideo2017.ensea.fr/plateforme/>), des législatives 2017 avec les comptes des 9 partis politiques, et du quinquennat actuel avec les comptes du Président, des ministres et des représentants de l'opposition (liens disponibles en ligne).

Analyse par mot employé. Cette fonctionnalité permet à l'utilisateur de choisir un mot parmi les mots souvent employés dans la politique (Alduy (2017)). Dans le cas de l'élection présidentielle, 13 mots sont étudiés : *France, état, république, peuple, loi, travail, liberté, démocratie, sécurité, immigration, terrorisme, islam et laïcité*. Cet ensemble de mots est mis à jour en fonction de l'événement politique couvert par la plateforme avec des mots spécifiques.

Le choix du mot donne accès à quatre analyses possibles sur les tweets (exemple pour l'élection présidentielle) : (1) l'usage fait de ce mot par les différents candidats sous forme de sur-/sous-emploi du mot, (2) les mots associés à ce mot extraits à base d'une analyse de cooccurrences, (3) l'emploi du mot et de ses dérivés par les différents candidats, et (4) le nuage de mots pour afficher le lexique complet des tweets.

Analyse par compte. Cette analyse permet à l'utilisateur de choisir un compte Twitter et d'analyser ses tweets. En fonction de l'événement, ce compte peut être celui d'un candidat à l'élection présidentielle, d'un parti pour les législatives ou d'un ministre pour le quinquennat.

Les analyses linguistiques disponibles à réaliser sur le compte Twitter choisi sont les suivantes : (1) la détection des mots les plus utilisés, (2) l'émergence de thématiques en regroupant les mots proches sémantiquement ; cette analyse produit un tableau contenant la liste des mots pour chaque thème et un graphe qui permet de faire émerger les grands domaines lexicaux et de dégager les principaux thèmes, (3) la détection des relations entre les mots qui permet d'analyser la similitude des mots sous forme graphique, (4) la mise en place d'un nuage de mots qui affiche sous cette forme le lexique des mots, et, si l'utilisateur a choisi d'analyser tous les comptes en même temps, (5) l'identification des spécificités des différents comptes : les mots et catégories spécifiques, et de les comparer.

Moteur de recherche à base de facettes. Cette troisième fonctionnalité de la plateforme est sous forme d'un moteur de recherche autonome et en temps réel. Il permet à l'utilisateur de chercher librement sur les tweets de l'événement à l'aide de plusieurs facettes : les comptes, les hashtags et les mentions, et de trier les résultats selon la date ou l'engagement, mais aussi

de connaître la répartition exacte du nombre de tweets par compte ou par grande thématique. Cette connaissance de la distribution de tweets nous offre une contextualisation globale pour chaque requête car notre objectif est autant de réaliser un moteur de recherche que d'offrir à nos utilisateurs une plateforme de veille concurrentielle entre les différentes stratégies de communication des comptes.

Visualisation graphique. En plus des analyses linguistiques, nous avons ajoutés des visualisations graphiques de données pour les comptes choisis à l'aide de l'outil de visualisation de données Kibana¹. Pour cela, nous avons proposé, pour le compte choisi, quatre graphes représentant : les 20 hashtags les plus utilisés, les 20 mentions les plus utilisées, et l'évolution du nombre de tweets et du nombre de retweets par jour.

Extraction d'un corpus. Cette dernière fonctionnalité permet à l'utilisateur de télécharger un sous-corpus de tweets de l'ensemble global de tweets. Cette fonctionnalité est très importante pour les chercheurs en sciences humaines et sociales car ils n'ont pas accès à des outils leur permettant d'extraire et de structurer leur corpus avant utilisation.

2.2 Cas d'utilisation

#Idéo2017 est une plateforme destinée au plus grand nombre et facile d'utilisation. Un cas d'utilisation est décrit dans la suite. Considérons que l'utilisateur choisit de s'intéresser à l'élection présidentielle et au mot *islam*. Dans la suite nous allons décrire les étapes que l'utilisateur suit afin de comprendre l'utilisation de ce mot par les candidats à l'élection.

Étape 1. Etant donné l'intérêt pour un mot précis, la fonctionnalité "**analyse par mot employé**" est utilisée. Parmi les analyses proposées, l'utilisateur choisit de comparer le **sur-emploi ou sous-emploi** de ce mot par les différents candidats. L'utilisateur remarque que le mot *islam* est sous-employé par *Marine Le Pen*, et modestement employé par *François Fillon*.

Étape 2. Supposons maintenant que l'utilisateur soit surpris du sous-emploi du mot *islam*. Ainsi, la plateforme lui permet, toujours dans la même fonctionnalité, de chercher **l'emploi de ce mot et de ses dérivés**, et observer leur fréquence. Il verra ainsi (fig. 1) que le nombre de tweets dans lesquels se trouve cette forme est très important chez ces deux candidats.

Étape 3. Ceci se confirme d'ailleurs en regardant **les mots associés** au mot *islam*, proposées sous forme graphique. Ce mot est en effet très lié à différents réseaux ou noeuds :

- islamisme / immigration / communautarisme,
- islamiste / fondamentaliste / idéologie / attentat / terrorisme,
- islamique / totalitarisme / Syrie.

Étape 4. Devant ces résultats, un retour au corpus s'impose pour l'utilisateur, et il est rendu possible grâce à la fonctionnalité **moteur de recherche** dans laquelle l'utilisateur peut chercher le mot qui l'intéresse, et les propositions de tweets s'affichent dynamiquement avec un accès direct aux tweets (fig. 2). Ceci confirme le résultat graphique obtenu dans les premières requêtes, puisqu'il est question, dans le tweet de Marine Le Pen, de *terrorisme islamiste*, et donc elle utilise souvent des dérivés du mot *islam*.

1. <https://www.elastic.co/fr/products/kibana>

#Idéo2017 : une plateforme citoyenne dédiée à l'analyse des tweets politiques

2.3 Développement

Pour la mise en place de la plateforme #Idéo2017, nous avons suivi plusieurs étapes : (1) l'extraction de l'ensemble de tweets, (2) la mise en place d'un sauvegarde des tweets, (3) l'indexation des tweets pour faciliter la recherche dans l'ensemble de tweets, (4) l'application d'un ensemble d'analyses linguistiques sur les tweets (avec l'utilisation et la modification des scripts d'IRaMuTeQ²), (5) la mise en place d'un moteur de recherche sur l'ensemble de tweets, et (6) l'affichage des résultats sur une page web.

Afin de traiter un grand nombre de tweets, nous avons utilisé Elasticsearch³ pour la mise en place d'un index sur les tweets. Celui-ci est exploité dans l'implémentation des analyses linguistiques, et la mise en place du moteur de recherche à facettes. Elasticsearch permet la sélection rapide et en temps réel des tweets contenant un mot spécifique, ou résultant du croisement entre les recherches simple et par facettes.

3 Conclusion

Cet article a présenté la plateforme #Idéo2017 qui, adaptable aux différents événements politiques, permet aux citoyens d'analyser par eux-mêmes les tweets politiques en lien avec l'événement. Elles permettent aux citoyens d'appréhender les données sociales relatives à des événements, et de leur donner du sens. Plusieurs versions de la plateforme ont été testées : pour l'élection présidentielle, les législatives, et le quinquennat.

Références

- Alduy, C. (2017). *Ce Qu'ils Disent Vraiment. Decoder Le Discours Des Presidentiables*. SEUIL.
- Conover, M., J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini, et F. Menczer (2011). Political polarization on twitter. In *Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*.
- Gayo-Avello, D., P. T. Metaxas, et E. Mustafaraj (2011). Limits of electoral predictions using twitter. In L. A. Adamic, R. A. Baeza-Yates, et S. Counts (Eds.), *ICWSM*. The AAAI Press.
- Longhi, J. (2013). Essai de caractérisation du tweet politique. *L'information grammaticale*.
- Longhi, J., C. Marinica, et N. Haddioui (2016). *Res per nomen 5, Négation et référence*, Chapter Extraction automatique de phénomènes linguistiques dans un corpus de tweets politiques : quelques éléments méthodologiques et applicatifs à propos de la négation. EPURE.
- Mercier, A. (2014). Twitter l'actualité : usages et réseautage chez les journalistes français. recherches en communication. *Recherches en communication* 40, 111–132.
- Roginsky, S. et B. D. Cock (2015). Faire campagne sur twitter. modalités d'énonciation et mises en récit des candidats à l'élection européenne. *Les Cahiers du numérique* 11(4), 119–144.

2. <http://www.iramuteq.org>

3. <https://www.elastic.co>

Tumasjan, A., T. O. Sprenger, P. G. Sandner, et I. M. Welpé (2010). Predicting elections with twitter : What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*.

Annexe

Annexe 1

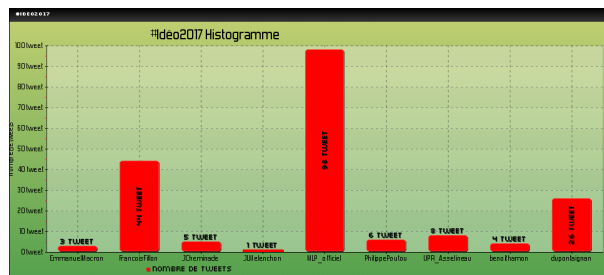


FIG. 1 – L'emploi du mot islam et de ses dérivés par les candidats à l'élection présidentielle.

Annexe 2



FIG. 2 – Moteur de recherche à facettes - recherche du mot islam.

Summary

This platform aims at allowing to the citizens to analyze by themselves the political tweets in specific political events in France. For the presidential election in 2017, #Idéo2017 analyzes

#Idéo2017 : une plateforme citoyenne dédiée à l'analyse des tweets politiques

in almost real time the candidates' messages, and provides their main characteristics, how the political lexicon is used and comparisons between the candidates.