Analyse des discours sur Twitter dans une situation de crise

Étude de l'incident à l'usine Lubrizol de Rouen

Hiba Abou Jamra, Annabelle Gillet, Marinette Savonnet, Éric Leclercq

Laboratoire d'Informatique de Bourgogne - EA 7534
Univ. Bourgogne Franche-Comté
9, Avenue Alain Savary, F-21078 Dijon - France
Hiba_Abou-Jamra@etu.u-bourgogne.fr

RÉSUMÉ. *Les données des réseaux sociaux ont un potentiel de valeur que les outils d'analyse doivent révéler. Twitter, en tant que plateforme de microblogging, facilite les interactions entre ses utilisateurs, permet la diffusion rapide d'informations. L'analyse des grandes tendances, des évènements, à partir des données permet de comprendre ou d'agir sur le monde réel. Dans cet article, nous présentons une méthodologie de travail et des résultats d'analyse pour l'étude de la communication dans les domaines de l'alimentation et de la santé qui ont pour cadre le projet inter-disciplinaire COCKTAIL incluant plusieurs partenaires industriels. Nous identifions un cas d'utilisation général pour différents types d'analyses des discours et nous montrons comment les algorithmes permettent de construire un observatoire afin de proposer des indicateurs macroscopiques et d'étudier les évènements, les communautés, les utilisateurs influents. Un exemple concret autour de l'évènement Lubrizol servira de fil conducteur pour illustrer les fonctionnalités et les problématiques traitées.*

ABSTRACT. *Social media data has value potential that analytics tools need to reveal. Twitter, as a microblogging platform, facilitates interactions between its users, allows rapid dissemination of information. The analysis of major trends, events, from the data makes it possible to understand or act on the real world. In this article, we present a working methodology and analysis results for the study of communication in the fields of food and health, which are part of the interdisciplinary COCKTAIL project involving several industrial partners. We identify use cases for different types of analysis and we show how the analysis algorithms make it possible to build an observatory in order to propose macroscopic indicators and to study events, communities, influential users. A concrete example around the Lubrizol event will serve as a common thread to illustrate the functionalities and the addressed problems.*

MOTS-CLÉS : *Analyses des réseaux sociaux, Twitter, détection des communautés, mesures de centralité, détection des évènements, séries temporelles*

KEYWORDS: *Social network analysis, Twitter, community detection, centrality measure, event detection, time series*

1. Introduction et problématique

Les données des réseaux sociaux sont créées à partir des interactions entre les individus, ces interactions étant amplifiées par les relations personnelles. Ces réseaux possèdent des caractéristiques intéressantes en terme de valorisation des données. Cependant, ils ont des propriétés spécifiques (distribution en loi de puissance, petit monde, assortativité, attachement préférentiel, etc.) qui nécessitent des outils d'analyse plus sophistiqués que les approches classiques des systèmes d'information ne proposent pas. Par exemple, la structure communautaire des réseaux sociaux est une des propriétés fondamentales. Cette structure peut être utilisée pour comprendre les interactions entre les utilisateurs mais aussi pour expliquer des évènements. En effet, des observations ont mis en évidence que certains évènements émergent plus vite à travers les réseaux sociaux qu'à travers d'autres médias plus traditionnels comme les sites Web, la radio et la télévision (Aiello *et al*., 2013). Dans ce cadre, Twitter est reconnu comme un révélateur d'évènements importants souvent quelques minutes ou heures après qu'ils se soient produits (Fedoryszak *et al*., 2019) et comme une chambre de résonance à forte influence qui propage rapidement l'information dans des communautés polarisées.

Notre objectif est de proposer une plateforme de collecte et d'analyse prenant en compte la richesse des données Twitter afin d'appréhender les usages, la circulation de l'information et la construction des discours. Si le fonctionnement de Twitter est simple – des tweets produits par des utilisateurs à l'aide de quelques opérateurs – les liens créés sont, selon le domaine ou les intentions de l'utilisateur, sémantiquement différents (Azaza *et al*., 2019). Les workflows doivent être reproductibles, pour cela nous utilisons Jupyter [1] pour construire des squelettes d'analyses dans lesquels chaque série de traitements est clairement identifiée : provenance des données, description des données de sortie, algorithmes, conditions d'application des algorithmes, contraintes pour l'interprétation des résultats produits.

Notre démarche de travail repose sur différents niveaux d'analyse (macroscopique, mésoscopique puis microscopique) et cherche à comprendre comment les espaces des utilisateurs et des hashtags sont constitués et comment ils interagissent, ceci afin d'éclairer la structure de la communication sur Twitter. Elle sera décrite à partir de l'étude d'un évènement qui est l'incendie de l'usine Lubrizol à Rouen. Pour l'analyse macroscopique, des chiffres significatifs et des fréquences brutes seront identifiés, puis l'étude mésoscopique se concentrera sur les communautés d'utilisateurs et la centralité de ces derniers dans chacune des communautés et enfin pour l'analyse microscopique nous analyserons le cas particulier de la construction du discours de responsabilité institutionnelle lors d'une crise.

1. https://jupyter.org

L'article est organisé comme suit, la section 2 présente des travaux similaires traitant de l'analyse d'évènements et de crises sur Twitter. La section 3 présente les principes du projet interdisciplinaire COCKTAIL, nous y discutons de la collecte et du nettoyage des données. La section 4 détaille notre approche appliquée aux tweets concernant l'incendie de l'usine Lubrizol à Rouen avec des analyses sur l'utilisation des hashtags, de la structuration des communautés d'utilisateurs et comment les institutions (État, métropole, agence sanitaire et de santé, rectorat, etc.) ont réagi et communiqué. La section 5 présente la mise en œuvre des workflows reproductibles à l'aide de l'outil Jupyter. Finalement, la section 6 conclut l'article et dresse les perspectives dégagées par ce travail.

2. Travaux connexes

Twitter est l'un des réseaux sociaux les plus utilisés, faisant de lui un modèle représentatif pour les Sciences Humaines et Sociales. Plusieurs études ont analysé les discours sur Twitter afin d'identifier les réactions des individus face à des crises naturelles ou des évènements remarquables dans la politique, le sport, l'économie ou la société. Nous pouvons citer les études de MacEachren *et al*. (2011) qui analysent dans le cadre d'une crise les activités dans l'espace et le temps des utilisateurs de Twitter, et Öztürk et Ayvaz (2018) qui ont exploré les opinions et les sentiments des utilisateurs de Twitter à l'égard de la crise des réfugiés syriens. Dans la suite de cette section, nous présentons plus en détail des travaux sur le Brexit qui se rapprochent de notre démarche.

Le Brexit a donné lieu à plusieurs études ayant pour but d'analyser ses causes et ses conséquences à partir des messages échangés sur Twitter. Mora-Cantallops *et al*. (2019) ont étudié l'influence du Brexit sur la façon dont les informations sont discutées sur le réseau, mais également la création de messages et la forme du réseau lui-même. 4 037 684 tweets, émis par des utilisateurs localisés au Royaume-Uni et catégorisés en positif (quitter), négatif (rester) ou neutre, ont été collectés entre le 12 Mai et le 23 Juin 2016. Ces tweets, regroupés par intervalles d'une heure, ont permis de construire des graphes dont les nœuds sont les utilisateurs et les liens sont les relations retweet, reply et quotes. Ensuite, pour comprendre la volatilité des interactions face à un tel évènement, les auteurs ont utilisé le modèle statistique GARCH (*Generalized Autoregressive conditional heteroscedasticity*) pour analyser les séries temporelles. Des mesures de centralité pour détecter les utilisateurs influents et des détections de communautés ont aussi été effectuées. La limite de cette étude est que les données sont incomplètes car les tweets étudiés ont été obtenus à partir d'un processus d'échantillonnage.

Vasiliu *et al*. (2016) ont surveillé et analysé les interactions sur Twitter autour du Brexit en utilisant la plateforme du projet SSIX (*Social Sentiment analysis financial IndeXes*), qui fournit aux entreprises européennes un ensemble de logiciels afin d'analyser et de comprendre les sentiments exprimés sur les médias sociaux. Les données ont été collectées à partir de 75 critères (mots-clés, hashtags et utilisateurs) entre le 4

et 30 Mai 2016, et archivées dans une base non-relationnelle. L'analyse s'est déroulée sur trois périodes: 1) avant le vote, 2) le jour du vote, et 3) après le vote. Pour les périodes avant et après le vote, les auteurs ont observé les tendances sur une période de 3 à 4 jours, alors que pour le jour du vote, ils observaient les tendances toutes les deux à trois heures. Les tweets ont été classés en deux catégories : quitter et rester. Ils ont découvert que leur résultat différait du vote de 9,4%, probablement en raison de la tranche d'âge des utilisateurs de Twitter, de la localisation des tweets collectés et du fossé éducatif entre les utilisateurs de Twitter et la population britannique.

Nous remarquons que les travaux décrits précédemment ont étudié les données de Twitter, pour un évènement donné en utilisant un ensemble d'outils spécifiques pour l'analyse. Il est important de noter que ces études se sont concentrées sur des critères particuliers du réseau Twitter comme par exemple la polarisation, ce qui les rend incomplètes et peu reproductibles en termes d'analyses des discours relatifs à des évènements et des incidents importants.

3. Le projet COCKTAIL : contexte, objectifs

Le projet COCKTAIL, lauréat d'un appel à projet ISITE-BFC (*Initiatives Science Innovation Territoires Économie en Bourgogne-Franche-Comté*), vise à créer un observatoire en temps réel des tendances, des singularités et des signaux faibles circulant dans les discours sur Twitter. Les analyses effectuées ont pour objectif de détecter les discours critiques pouvant devenir viraux, les communautés clés d'acteurs liées au domaine étudié (alimentation et santé) avec leurs liens/interactions, les évènements ou précurseurs d'évènements, les discours émergents caractérisés par des signaux faibles ainsi que les tendances culturelles liées aux pratiques alimentaires et à la santé. Le consortium comprend des chercheurs en Sciences de l'Information et de la Communication, en Informatique, en Sciences Cognitives, et en Sciences des Aliments, le pôle de compétitivité Vitagora spécialisé dans l'agro-alimentaire et des entreprises du domaine informatique. L'observatoire, logiciel libre, permettra aux chercheurs et aux stratèges des secteurs public et privé de bénéficier des avancées scientifiques du projet et aux partenaires de développer des services adaptés à leur modèle économique.

Le diagramme de cas d'utilisation (*use case*) de la figure 1 présente les acteurs et les grandes fonctionnalités de la plateforme. COCKTAIL fait intervenir plusieurs acteurs dont le *data engineer* qui conçoit et gère l'architecture de collecte et de stockage des données. Il classe les données recueillies en fonction des besoins exprimés par le commanditaire et aide le *data scientist* dans la phase de nettoyage des données. Le *data analyst* et le *data scientist* traduisent et formalisent le questionnement métier du commanditaire. Le *data scientist* croise les données, sélectionne, modifie les algorithmes et valide les analyses. Finalement le *data analyst* transforme les analyses en informations métier pour le commanditaire.

Dans la suite, nous expliquons brièvement l'architecture Hydre développée pour la collecte en continu et le stockage des données issues de Twitter (Gillet *et al.*, 2019). Un système de stockage de type polystore a été développé et intégré dans la plateforme.

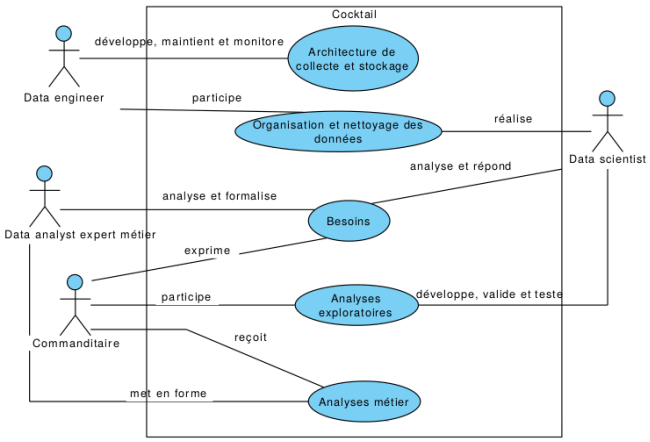

Figure 1. Diagramme de cas d'utilisation de niveau 1 décrivant les fonctionnalités de la plateforme COCKTAIL et les acteurs correspondants

L'architecture de la plateforme s'inspire du patron de Lambda Architecture (Marz, Warren, 2015) et s'appuie sur les composants suivants :

– des processus de collecte implémentés avec le modèle d'acteurs Akka [2] et déployés sur un cluster. Ces processus exploitent les deux APIs que Twitter met à disposition : l'API search et l'API stream. Ces APIs sont exploitées avec la librairie Twitter4j. Les critères de collecte sont des comptes, des hashtags et des mots-clés. Le système Kafka [3] est utilisé pour servir d'intermédiaire entre la partie collecte et les couches Speed et Batch et permet de conserver les données pendant 7 jours, ce qui contribue à la capacité de résistance aux pannes de l'architecture ;

– la couche Speed permet de calculer des indicateurs macroscopiques en temps réel en utilisant Kafka Streams [4], afin d'établir des séries temporelles en temps réel sur les éléments importants constitutifs des tweets comme les hashtags, les mentions et de servir de base pour détecter des évènements ;

– la couche Batch enregistre les données brutes dans Hadoop HDFS (*Hadoop Distributed File System*) [5] et parallèlement par micro-batch dans le polystore. Le stockage dans HDFS permet de lancer des traitements de reprises depuis les données brutes si le schéma de données du polystore est modifié suite à des évolutions fonctionnelles ou techniques ;

– la couche Serving est implantée par un polystore qui met à disposition les données dans des formats adaptés à leur analyse. Le polystore comprend les bases de don-

2. https://akka.io/
3. https://kafka.apache.org/
4. https://kafka.apache.org/documentation/streams/
5. https://hadoop.apache.org/

nées PostgreSQL [6] (pour le traitement des données attributaires), ArangoDB [7] (pour l'analyse des graphes) et TimescaleDB [8] (pour l'analyse des séries temporelles).

L'architecture Hydre est capable de supporter le flux moyen de Twitter (6 000 tweets par seconde) permettant des collectes importantes. À la fin février 2020, le nombre de tweets collectés dans le polystore atteint environ 75M, pour un volume de 1,9 To de données brutes JSON et 171 Go dans PostgreSQL.

Hydre est la suite de SNFreezer, une plateforme multi-paradigme de collecte, de stockage et d'analyse de tweets (Basaille *et al*., 2016), dans laquelle nous avons comparé les différentes fonctionnalités de SNFreezer avec certains projets concurrents.

4. Étude de cas : l'incendie de l'usine Lubrizol à Rouen

Dans cette section, nous étudions les prises de forme des discours sur Twitter dans le cadre de l'incendie de l'usine Lubrizol à Rouen, et nous analysons plus particulièrement la construction du discours de responsabilité institutionnelle suite à cet incendie. Pour cette étude nous développons une méthodologie d'analyse sur trois niveaux (macroscopique, mésoscopique et microscopique), qui est implantée sous la forme de *notebooks* dans Jupyter.

4.1. Chronologie de l'incident et première apparition sur Twitter

Le 26 septembre 2019, entre 2h40 et 2h50 du matin, une partie de l'usine Lubrizol à Rouen et trois bâtiments de Normandie Logistique ont été ravagés par un incendie qui a provoqué une énorme fumée noire de 22 km. Lubrizol [9], fabriquant des additifs pour lubrifiants industriels et pour les carburants, est classée Seveso « à haut risque ». Le sinistre a fait aucune victime, mais sa cause reste pour le moment inconnue, même la localisation précise du départ du feu n'a pas pu être déterminée [10]. Dès le début, Twitter a joué un rôle très important dans la diffusion de l'information sur cet incendie. Alors que l'incendie a débuté vers 2h40, le premier tweet relatif à l'incendie a été publié à 3h03 par Thomas Schonheere, journaliste à France Bleu Normandie, le mot lubrizol apparaît à 3h24, et à 4h15 le hashtag `#lubrizol` surgit pour la première fois, le premier tweet d'une institution (le préfet de la Seine Maritime) est émis à 4h50 demandant d'éviter le secteur.

6. https://www.postgresql.org/
7. https://www.arangodb.com/
8. https://www.timescale.com/
9. https://france.lubrizol.com/fr-FR/About
10. https://www.lexpress.fr/actualite/societe/incendie-de-l-usine-lubrizol-a-rouen-ce-que-l-on-sait-un-mois-apres_2104941.html

4.2. Nettoyage des données

L'intérêt de l'évènement nous a amené à lancer une collecte pour laquelle 47 hashtags et 111 utilisateurs ou comptes ont été fournis par les chercheurs en sciences humaines et sociales du projet COCKTAIL comme critères de collecte :

Hashtags : `#CELubrizol`, `#lubrizolrouen`, `#lubrizoltransparence`, `#seveso`, `#VeritePourRouen`, `#WarrenBuffettDoitPayer`, etc.

Utilisateurs : `@atmonormandie`, `@fbleunormandie`, `@Min_Ecologie`, `@prefet76`, `@regionNormandie`, `@damienadam76`, etc.

Ces critères ont permis de collecter environ 2 millions de tweets entre le 26 septembre et le 26 novembre 2019. Pour diminuer le bruit dans la collecte, nous avons nettoyé ces données en filtrant celles nécessaires pour notre étude. Dans cette étape, nous avons gardé les tweets correspondants aux critères fournis (hashtags et comptes) filtrés avec les mots `lubrizol` ou `rouen`. Par exemple nous retenons les tweets contenant `lubrizol` ou `rouen` avec `controlessanitaires`, ou encore `rouen` avec `desastreecologique`. En revanche avec cette méthode les tweets du compte `@fbleunormanandie` sans rapport avec l'incident sont éliminés. À l'issue de cette opération, le corpus est réduit à 558 895 tweets. Parmi ces tweets, 73 126 sont des tweets originaux et 485 769 sont des retweets émis par 141 177 comptes. Durant la seule journée du 26 septembre, 29 495 tweets dont 26 940 retweets ont été produits sur l'évènement, confirmant bien l'effet chambre d'écho de Twitter.

4.3. Exploration macroscopique du corpus

Tableau 1. TOP 5 des utilisateurs institutionnels les plus actifs

Utilisateurs	Nombre de tweets originaux
Sénat Direct	61
Sénat	58
Préfet de la Seine-Maritime	53
Atmo Normandie	20
Gouvernement	17

Utilisateurs	Nombre de retweets
Métropole Rouen Ndie	78
Ministère de la Solidarité et de la Santé	41
Ville de Rouen	35
Préfet de la Seine-Maritime	28
ARS Normandie	27

Utilisateurs	Nombre de tweets retweetés
Préfet de la Seine-Maritime	135
Gouvernement	43
Minitère de l'intérieur	25
ARS Normandie	15
Académie de Rouen	8

Nous avons réalisé des analyses exploratoires de niveau macroscopique sur le corpus en calculant des chiffres significatifs comme le nombre total des tweets comprenant des tweets originaux, des retweets, des réponses et des citations. Ainsi, nous avons calculé les fréquences d'apparition des producteurs de tweets, des retweeteurs, des utilisateurs qui sont les plus retweetés et des hashtags, à l'aide de requêtes analytiques SQL lancées sur la base de données relationnelle. Le tableau 1 montre les utilisateurs institutionnels les plus actifs. Afin d'étudier comment ces utilisateurs communiquent, nous avons construit le graphe d'interaction $utilisateur-utilisateur_{retweet}$, et nous avons ensuite généré une visualisation avec le logiciel Gephi [11]. Sur la figure

11. https://gephi.org/

2, la taille des nœuds est proportionnelle à leur centralité. Il est à noter que les utilisateurs `Senat` et `Senat_direct` sont isolés, cela peut s'expliquer par la constitution d'une commission d'enquête sénatoriale sur Lubrizol le 10 octobre.

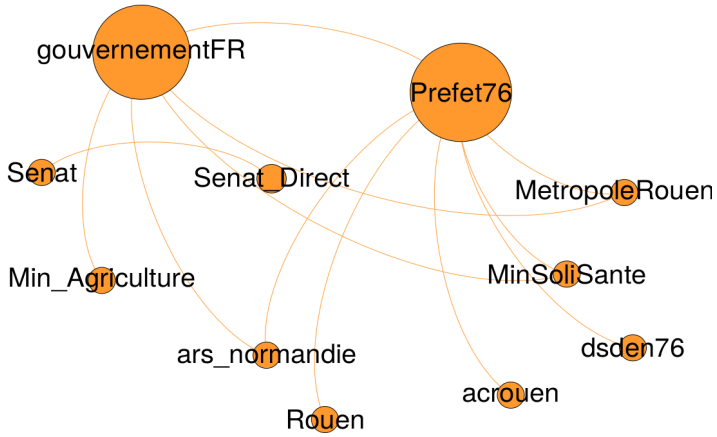

Figure 2. Graphe de la relation retweet pour les utilisateurs institutionnels les plus actifs

Nous avons réalisé une étude globale macroscopique des hashtags.La forte présence du hashtag `#lubrizol` nous a amené à travailler sur les co-occurences de celui-ci ainsi que ses composés. Le tableau 2 montre l'inquiétude concernant les risques encourus avec l'apparition des hashtags `ecologie`, `hydrocarbures` et `climat`. Cette inquiétude est associée à un besoin de transparence mais aussi à une rapide politisation d'après les hashtags `climatstrike`, `giletsjaunes`, et `act46` et à l'interpellation du pouvoir en place avec les hashtags `gouvernement`, `castaner`, `larem` et `enmarche`. Il est à remarquer un détournement de ce hashtag pour mettre en avant l'acte 46 des gilets jaunes (manifestation du 28 septembre). En effet, 14 834 utilisateurs différents ont relayé ou émis des tweets relatifs aux gilets jaunes dans notre corpus.

4.4. Analyses mésoscopiques

Dans ce qui suit, nous allons décrire la détection des communautés dans différents types de graphes, le calcul des scores de hubs et autorités, et les mesures de centralité pour étudier les phénomènes d'influence.

4.4.1. Détection des communautés

Un processus pertinent dans l'analyse des réseaux sociaux est de découvrir des nœuds qui partagent des caractéristiques communes dans un graphe. Ce processus, appelé détection de communautés, consiste à trouver des groupes ayant une forte densité de liens. Plusieurs travaux ont proposé différentes méthodes de détection de com-

Tableau 2. Trois co-occurents du hashtag `#lubrizol` *et ses composés*

#lubrizol et composés	Hashtag2	Hashtag3	Hashtag4	Fréquence
lubrizol	incendierouen	giletsjaunes	act46	2363
lubrizol	incendierouen	giletsjaunes	acheres	1303
lubrizolrouen	lubrizol	gouvernement	castaner	905
lubrizol	ecologie	climat	castaner	904
lubrizol	incendie	hydrocarbures	actionclimat	760
lubrizoltransparence	lubrizolrouen	lubrizol	gouvernement	757
lubrizol	ecologie	climatestrike	climat	750
lubrizol	lubrisol	giletsjaunes	act46	655
lubrizoltransparence	lubrizolrouen	lubrizol	larem	592
lubrizol	lrem	enmarche	agriculteurs	515

munautés. Nous avons utilisé l'algorithme Louvain basé sur la maximisation de la modularité (Blondel *et al.*, 2008). Nous avons choisi cet algorithme car il s'exécute en O(nlogn) et il exploite des informations globales sur la topologie du réseau. Nous avons aussi employé l'algorithme Walktrap créé par Pons et Latapy (2005) pour détecter des communautés. Il se base sur l'idée qu'une marche aléatoire à partir d'un nœud reste pendant un certain temps dans la communauté du nœud. Nous avons utilisé deux algorithmes comparables car ils utilisent la modularité mais s'appuient sur des principes différents. Le tableau 3 synthétise les résultats produits par les deux méthodes pour le graphe $utilisateur - utilisateur_{retweet}$. Nous remarquons que les méthodes Louvain et Walktrap trouvent deux communautés identiques qui correspondent pour l'une aux médias nationaux comme `BFMTV`, `FranceInfo` et `CNEWS` et pour l'autre aux médias locaux comme `76actu` et `paris_normandie`. Les tailles des communautés correspondantes appuient cette ressemblance. Les nœuds `AiphanMarcel`, `Loran076` , `CerveauxNon` et `Rouendanslarue` appartenant à une communauté détectée par Walktrap se retrouvent dans deux communautés pour Louvain. L'utilisateur `DuPouvoirDachat` est fortement considéré central dans l'une des communautés avec Louvain, tandis que nous ne le retrouvons pas comme un nœud central (importance négligeable) avec Walktrap (valeur `NA`).

Tableau 3. Détection de communautés avec Louvain et Walktrap dans le graphe $utilisateur - utilisateur_{retweet}$

Utilisateur	Communauté Louvain	Communauté Walktrap
AiphanMarcel	Taille communauté: 550	Taille communauté: 716
Loran076		
CerveauxNon	Taille communauté: 181	
Rouendanslarue		
BFMTV	Taille communauté: 222	Taille communauté: 134
FranceInfo		
CNEWS		
76actu	Taille communauté: 204	Taille communauté: 200
paris_normandie		
DuPouvoirDachat	Taille communauté: 155	NA

4.4.2. Étude des relations entre les utilisateurs

Pour l'étude des relations entre les utilisateurs, nous avons étudié les phénomènes d'influence. Sous une forme plus algorithmique, il s'agit de découvrir les hubs et autorités dans le graphe des retweets en utilisant l'algorithme HITS (*Hyperlink-Induced Topic Search*) qui calcule deux scores pour chaque nœud, appelés score de hub et score d'autorité, uniquement en fonction des liens présents entre les nœuds (Kleinberg, 1999). Des résultats montrent que HITS fournit un calcul précis des nœuds d'autorité et des hubs (Devi *et al*., 2014). Nous avons appliqué HITS sur chacune des communautés détectées par la méthode Louvain (voir le tableau 4). Les utilisateurs qui sont des autorités sont des médias nationaux et locaux et des institutionnels. Les utilisateurs qui ressortent comme hub ont la particularité de beaucoup retweeter.

Tableau 4. Scores hub et autorité et communautés détectées par l'algorithme Louvain

Graphe de la relation retweet							
Autorités				**Hubs**			
Utilisateur	**Score**	**Rang**	**Communauté Louvain**	**Utilisateur**	**Score**	**Rang**	**Communauté Louvain**
BFMTV	1	1		YohannCrn	1	1	
franceinfo	0.682	2		bah_9	0.133	2	
Prefet76	1	1		Quinsolo	1	1	
gouvernementFR	0.298	2		lau68951920	0.661	2	
AiphanMarcel	1	1		MetropoleRouen	1	1	
76actu	1	1		DuPouvoirDachat	1	1	
raphtual	0.354	2		pythoncxde	0.350	2	

L'algorithme PageRank est aussi largement utilisé pour mesurer la centralité des utilisateurs. Nous avons ainsi étudié l'influence des utilisateurs dans leur communauté. En reprenant les communautés détectées par Louvain, nous avons calculé, communauté par communauté, le score PageRank de chaque utilisateur. La figure 3 reprend les communautés détectées avec le même code couleur que dans le tableau 3, la taille des nœuds est proportionnelle au score de Page Rank obtenu. Les nœuds ayant un score Page Rank élevé sont considérés comme autoritaires, d'ailleurs les comptes médiatiques (`BFMTV` et `FranceInfo`) sont les plus retweetés. D'un autre côté, `DuPouvoirDachat` est l'un des comptes qui retweetent le plus et il est vu comme un hub.

L'application de deux algorithmes permet aux chercheurs en Sciences Humaines et Sociales d'affiner leur interprétation dans les relations entre utilisateurs. Par exemple, l'utilisateur `76actu` qui est un média local est trouvé comme autoritaire avec l'algorithme HITS mais ne paraît pas comme un nœud central avec Page Rank dans sa communauté.

4.5. Analyses microscopiques : construction d'un discours de responsabilité institutionnelle dans une situation de crise

L'objectif de cette analyse est de voir comment les institutions s'expriment sur Twitter de manière responsable, comment elles incarnent la responsabilité d'information et de protection vis-à-vis des citoyens. Pour cela, le corpus Lubrizol a été restreint

Figure 3. Centralités intra-communautaires calculées après l'application de l'algorithme Louvain

pour étudier le discours des institutions (État, ministère, Préfet, académie, organismes publics). Pour se faire, l'équipe en Sciences Humaines et Sociales a sélectionné 23 comptes catégorisés par « institutions » et « organismes », le hashtag `#lubrizol` et le mot lubrizol réduisant alors le nombre total de tweets à 665.

Après une analyse qualitative en double aveugle, ces tweets ont été classés, suivant leur contenu, par modalité énonciative : « information », « conseil » et « obligation ». La figure 4 montre deux exemples de tweets catégorisés, avec les extraits de texte qui ont contribué à cette classification.

La figure 5 présente une synthèse de quelques utilisateurs par modalité énonciative en identifiant aussi les hubs et les autorités. À partir de ces résultats, les chercheurs en Sciences Humaines et Sociales ont fait l'analyse suivante : l'État, *via* ses différents acteurs institutionnels, semble chercher à construire une image « responsable » grâce aux informations et aux conseils apportés, faisant appel ainsi à la responsabilité du citoyen pour adopter une conduite appropriée. Le discours institutionnel porte assez peu sur l'obligation puisque le citoyen responsable n'a pas besoin de subir des injonctions : il agit en bon citoyen en connaissance de cause, en fonction des informations et des conseils reçus. Certains ont accusé l'État d'irresponsabilité en raison d'une stratégie de gestion de la peur, consistant à communiquer pour éviter la panique, tout en minimisant certains risques pour éviter l'affolement de la population.

Figure 4. Exemples de tweets classifiés

	% (nb de tweets contenant ce type de discours /nb de discours par compte)		
	Information	Conseil	Obligation
Académie de Rouen	60,42	35,42	4,17
ARS Normandie	72,92	18,75	8,33
Assemblée nationale	80,00	20,00	0,00
Atmo France	69,23	23,08	7,69
Gouvernement	85,37	9,76	4,88
Métropole Rouen Ndie	68,91	23,53	7,56
Ministère des Solidarités et de la Santé	82,14	16,07	1,79
Ministère de l'Intérieur	54,55	36,36	9,09
Préfet de la Seine-Maritime	70,75	25,47	3,77
Préfet de la Somme	93,33	6,67	0,00
Région Normandie	80,00	10,00	0,00

% (nb de tweets contenant ce type de discours /nb de discours par classe)		
Information	Conseil	Obligation
73,16	21,50	4,99

Hubs	Authorities

Figure 5. Pourcentage de tweets par modalité énonciative

La figure 6 met en perspective le nombre de tweets émis et catégorisés par modalité énonciative par rapport à des évènements significatifs qui se sont déroulés durant la période de l'étude. Nous notons que chaque évènement réel donne lieu à une émission de tweets.

Nous avons effectué à la fois une analyse quantitative et qualitative portant sur trois niveaux de granularité. L'étude macroscopique a mis en évidence l'interaction entre les utilisateurs qui communiquent sur Lubrizol. Elle a aussi mis en évidence le détournement du hashtag `#lubrizol` pour mettre en avant d'autres évènements. Et finalement, grâce à l'étude microscopique, nous avons étudié la communication institutionnelle lors d'une crise.

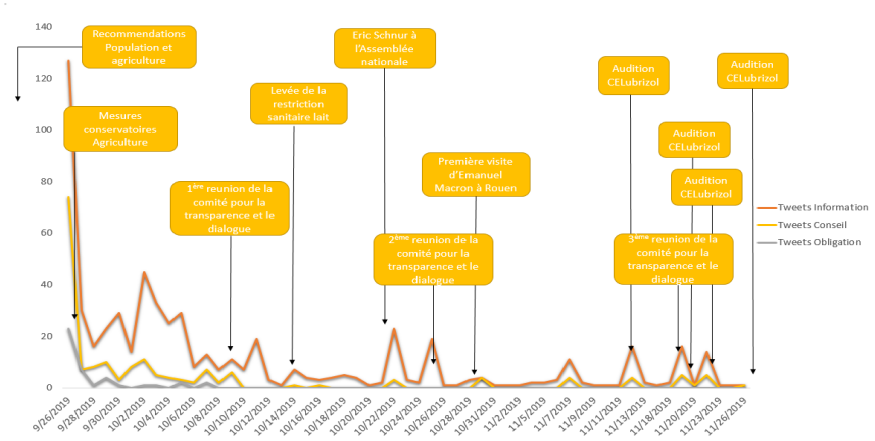

Figure 6. Série temporelle de l'évolution du nombre de tweets par modalité énonciative

5. Implémentation de la méthodologie sous une plateforme numérique

Le *notebook* Jupyter est un outil distribué sous licence BSD (*Berkeley Software Distribution*) utilisé pour l'analyse de données. Il permet aux *data scientists* de créer des scripts combinant du code, du texte et des interfaces graphiques (Perez, Granger, 2015). Des noyaux spécifiques à différents langages de programmation s'exécutent indépendamment et interagissent avec Jupyter, dont Python, R et Scala. Nous considérons le *notebook* Jupyter comme un environnement interactif qui nous permet de rassembler une description des données en entrée, de développer avec plusieurs langages de programmation dans un même noyau, et puis d'enregistrer et convertir les résultats vers d'autres formats que les fichiers structurés JSON tels que HTML et PDF.

Nous avons divisé l'implémentation en deux sections principales. La première consiste à la préparation pour effectuer les analyses, dans laquelle en utilisant le noyau Python, des tables sont créées sur une base PostgreSQL, après filtrage et regroupement des données brutes du corpus Lubrizol. La deuxième section comprend la partie analytique où les deux noyaux Python et R sont utilisés pour le calcul des indicateurs et des fréquences globaux, puis l'étude des graphes (communautés, centralité). À la fin nous utilisons une fonctionnalité Python pour visualiser les résultats sous forme de graphe à l'aide d'une connexion à Gephi. La figure 7 montre une capture du *notebook* Jupyter de l'analyse sur le corpus lubrizol, dans laquelle nous trouvons à gauche un panneau de navigation qui montre les sections qui composent ce *notebook*. À droite se trouvent les codes écrits en Python ou R correspondant aux sections du *notebook*.

D'autre part, une interface front-end est développée afin de fournir une visualisation des résultats aux *data analysts*/experts métiers, ce qui leur permet d'interpréter et de valider la méthodologie adaptée. Cette interface est développée en JavaEE, et

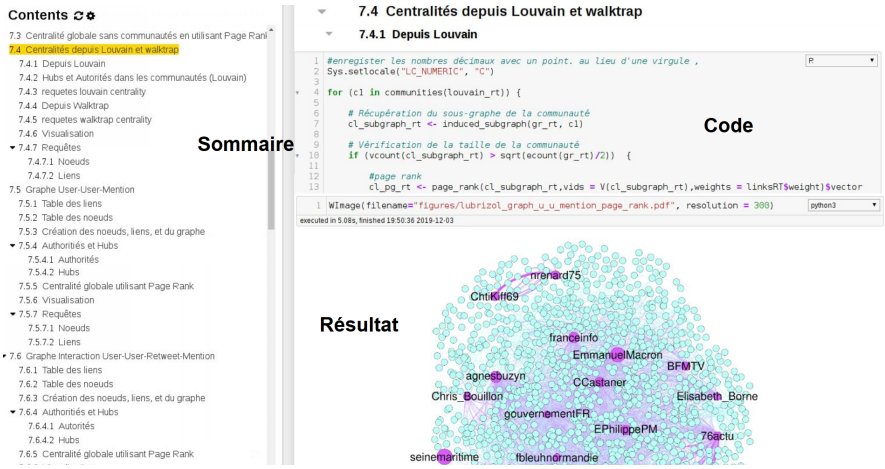

Figure 7. Le notebook Jupyter sur l'étude du corpus Lubrizol

les langages HTML, CSS et JQuery. Pour obtenir les résultats, deux méthodes sont possibles. La première consiste à extraire des séries temporelles d'un élément particulier de la base TimescaleDB (par exemple `#lubrizol`, `@Prefet76`, etc.). La deuxième méthode consiste à extraire le TOP k pour une catégorie d'éléments de la base (par exemple hashtag, mention, etc.). Ces deux méthodes utilisent les données enregistrées dans la base TimescaleDB qui est alimentée par la couche Speed décrite dans la section 3.

La figure 8 représente une série temporelle qui montre l'évolution du nombre des tweets par type (original, retweet, reply et quote), le *data scientist* peut aussi choisir de ne visualiser qu'un seul type, en fonction d'un intervalle de temps. Cet indicateur permet de suivre en temps réel l'émission de tweets sur une collecte et de voir si un évènement a lieu. Le même type d'indicateur est fourni pour suivre l'évolution en temps réel des hashtags.

6. Conclusion

Dans cet article, nous avons présenté notre démarche de travail et nos résultats d'analyse des tweets en l'appliquant à l'incendie de l'usine Lubrizol. Nos analyses ont été opérées à différents niveaux de granularité afin d'extraire des indicateurs significatifs. Nos analyses ont pour objectif de comprendre les interactions entre les utilisateurs, les communautés avec les utilisateurs influents, quels utilisateurs sont considérés comme des utilisateurs émettant une information fiable (autorité) et quels sont les utilisateurs qui relaient l'information (hub). Un focus sur le discours de responsabilité des institutions a été réalisé avec une approche qualitative, nous avons montré que la plupart des tweets étaient des tweets d'information.

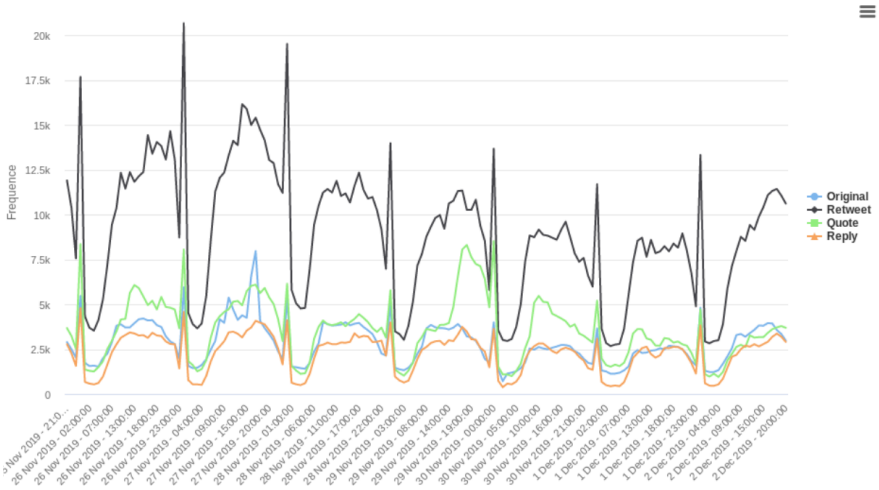

Figure 8. Interface front-end montrant l'évolution du nombre des tweets par intervalle d'une heure en fonction du temps

Dans la suite de notre travail, nous voulons détecter l'existence de signaux faibles. Un signal faible est une information d'alerte précoce, de faible intensité, pouvant être annonciatrice d'une tendance ou d'un évènement important (Ansoff, 1975). Le challenge porté par les signaux faibles est qu'ils sont difficiles à détecter et faciles à négliger, l'outillage algorithmique classique n'étant pas alors suffisant pour les détecter. C'est pourquoi nous voulons utiliser comme cadre théorique les « Graphlets » (Pržulj *et al.*, 2004). Les graphlets sont de petits (2 à 5 nœuds) sous-graphes induits connectés. Notre hypothèse est qu'ils sont une indication d'un début d'information précoce comme le début de la formation d'une communauté ou un mouvement agglomératif autour d'un hashtag.

Remerciement Ce travail est soutenu par le programme « Investissements d'Avenir », projet ISITE-BFC (contrat ANR-15-IDEX-0003). Le projet Cocktail est piloté scientifiquement par Gilles Brachotte, laboratoire CIMEOS EA-4177, Université de Bourgogne.

Bibliographie

Aiello L. M., Petkos G., Martin C., Corney D., Papadopoulos S., Skraba R. *et al.* (2013). Sensing trending topics in twitter. *IEEE Transactions on Multimedia*, vol. 15, nº 6, p. 1268–1282.

Ansoff H. I. (1975). Managing Strategic Surprise by Response to Weak Signals.

Azaza L., Leclercq É., Savonnet M. (2019). Modèle de réseaux multiplexes pour l'étude de l'influence sur twitter. In *Informatique des organisations et systèmes d'information et de décision (INFORSID)*, p. 255–270.

Basaille I., Kirgizov S., Leclercq É., Savonnet M., Cullot N. (2016). Towards a twitter observatory: A multi-paradigm framework for collecting, storing and analysing tweets. In *Conference on research challenges in information science (RCIS)*, p. 1–10.

Blondel V., Guillaume J.-L., Lambiotte R., Lefebvre E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, vol. 2008, n° 10, p. P10008.

Devi P., Gupta A., Dixit A. (2014). Comparative study of hits and pagerank link based ranking algorithms. *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 3, n° 2, p. 5749–5754.

Fedoryszak M., Frederick B., Rajaram V., Zhong C. (2019). Real-time event detection on social data streams. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*, p. 2774–2782.

Gillet A., Leclercq E., Cullot N. (2019). Lambda architecture pour une analyse à haute performance des données des réseaux sociaux. In *Informatique des organisations et systèmes d'information et de décision (INFORSID)*, p. 223–238.

Kleinberg J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, vol. 46, n° 5, p. 604–632.

MacEachren A. M., Robinson A. C., Jaiswal A., Pezanowski S., Savelyev A., Blanford J. *et al*. (2011). Geo-twitter analytics: Applications in crisis management. In *25th international cartographic conference*, p. 3–8.

Marz N., Warren J. (2015). *Big Data: Principles and best practices of scalable real-time data systems*. Manning.

Mora-Cantallops M., Sánchez-Alonso S., Visvizi A. (2019). The influence of external political events on social networks: The case of the brexit twitter network. *Journal of Ambient Intelligence and Humanized Computing*, p. 1–13.

Öztürk N., Ayvaz S. (2018). Sentiment analysis on twitter: A text mining approach to the syrian refugee crisis. *Telematics and Informatics*, vol. 35, n° 1, p. 136–147.

Perez F., Granger B. E. (2015). Project jupyter: Computational narratives as the engine of collaborative data science. *Retrieved September*, vol. 11, n° 207, p. 108.

Pons P., Latapy M. (2005). Computing communities in large networks using random walks. In *International symposium on computer and information sciences*, p. 284–293.

Pržulj N., Corneil D. G., Jurisica I. (2004). Modeling interactome: scale-free or geometric? *Bioinformatics*, vol. 20, n° 18, p. 3508–3515.

Vasiliu L., Freitas A., Caroli F., McDermott R., Zarrouk M., Hürlimann M. *et al*. (2016). In or out? Real-time monitoring of Brexit sentiment on Twitter. *SEMANTiCS*.