Qualification géographique d'un corpus de tweets pour la détection d'un événement climatique

Pierrick Bruneau*, Etienne Brangbour*,**, Stéphane Marchand-Maillet**

* LIST
Esch-sur-Alzette, Luxembourg
prenom.nom@list.lu,
** Université de Genève
Genève, Suisse
stephane.marchand-maillet@unige.ch

Résumé. Dans le cadre de l'analyse et de la prévention des catastrophes naturelles, le couplage entre modélisation physique de ces événements et leur représentation sur les réseaux sociaux est l'objet de nombreuses études actuellement. Une étape nécessaire est alors de collecter et de qualifier des corpora de contenus émis sur les réseaux sociaux. En particulier, dans cette communication nous nous intéressons à la collecte de contenus sur Twitter selon des critères spatio-temporels, et à la fiabilité des champs de nature spatiale qui caractérisent les tweets. Nous illustrons notre propos en étudiant le lien entre la fréquence des lieux renseignés dans le corpus et les surfaces associées, ou encore l'influence de l'application à l'origine des tweets. Nous utilisons nos observations pour estimer la pertinence d'une zone d'intérêt hydrologique.

1 Introduction

La modélisation et la prédiction de la survenue et de l'étendue des crues sont des activités essentielles en vue de limiter l'impact écologique et économique de ces événements. L'approche classique à ces prédictions repose sur des simulations de l'écoulement et de la saturation résultant des eaux pluviales (Bates et De Roo, 2000). Plus récemment, des modèles ont été proposés afin d'assimiler des sources d'information externes, telles qu'issues de l'imagerie satellite multispectrale, aux prédictions des modèles physiques classiques (Hostache et al., 2015). Le projet Publimape propose d'étudier l'extension de cette approche à des données issues de réseaux sociaux tels que Twitter ou Instagram. L'ouragan Harvey, survenu en 2017 et abondamment commenté alors [1], est utilisé pour tester et valider nos développements au cours du projet.

La mise en oeuvre de données de réseaux sociaux pour servir la détection d'événements a été considérée depuis deux perspectives dans la litérature. D'un côté, le *Volunteered Geographical Information* (VGI) demande explicitement à des utilisateurs d'aller sur le terrain et de capturer de l'information liée à l'événement (Griesbaum et al., 2017). En un sens, ce point

1. https://fr.wikipedia.org/wiki/Ouragan_Harvey

de vue est proche d'une campagne de *crowdsourcing*. De l'autre côté, les utilisateurs de Twitter sont considérés comme des unités d'un réseaux de capteurs, l'information pertinente étant alors collectée de manière passive (i.e. *Participatory Sensing* (Burke et al., 2006; Crooks et al., 2013)). Nous adoptons cette dernière perspective dans le cadre du projet Publimape.

Dans cette communication, nous nous concentrons plus particulièrement sur la dimension spatiale d'un corpus associé à l'ouragan Harvey, collecté par nos soins. En particulier, l'objectif ici est d'arriver à une première appréciation de la finesse et de la fiabilité de l'information géographique présente dans le corpus collecté, du point de vue du cas d'utilisation du projet. Après un état de l'art des méthodes et contraintes liées à la collecte de tweets pour la détection d'événements en section 2, nous exposons ensuite notre méthodologie de collecte, de stockage et d'analyse en section 3. Nous y procédons à une étude de cas formulée à partir de nos discussions avec les partenaires du projet spécialistes en hydrologie, nous permettant d'arriver à une première qualification de nos données pour l'application ciblée dans le projet.

2 Etat de l'art de la collecte de tweets

Twitter expose une API permettant de collecter les messages émis en temps réel [2]. Il est ainsi possible de restreindre la sélection selon des hashtags, des mot-clés, ou des coordonnées géographiques. Il est également possible de requérir un échantillon aléatoire du flux complet émis sur Twitter, sachant que cet échantillon est limité en volume à 1% du flux complet (Cheng et Wicks, 2014). Il est également possible de requérir une collection de tweets sur une période passée, mais cette fonctionnalité est payante (i.e. API *Enterprise* de Twitter [3]).

Les tweets collectés via les API de Twitter (gratuite ou *Enterprise*) sont obtenus sous la forme d'objets JSON, c'est-à-dire d'un dictionnaire, dont les champs peuvent être soit des types littéraux (nombres ou chaines de caractères), soit eux-mêmes des dictionnaires, implémentant une structure hiérarchique.

Parmi les champs à la racine d'un objet représentant un tweet (environ 30), vu le contexte de notre communication nous prêterons plus particulièrement attention aux champs à caractère géographique :

— *coordinates* : le géotag (i.e. coordonnées GPS) du tweet,
— *place.full_name* : le lieu attaché au tweet,
— *place.bounding_box* : l'enveloppe convexe du lieu attaché au tweet (*bbox* dans le reste du document),
— *user.location* : le lieu attaché au profil de l'utilisateur à l'origine du tweet. Plus spécifiquement, c'est un champ texte renseigné par l'utilisateur sur son profil.

Le partage et la mise à disposition de bases de données de tweets sont a priori interdits. En revanche, sous réserve de ne partager que les identifiants de tweets, il est acceptable de partager des ensembles d'identifiants de tweets, laissant au destinataire la charge de regénérer la collection de tweets de son côté. Dans le cas de projets de recherche à but non-lucratif, il est a priori possible de partager ainsi des collections contenant un nombre illimité d'identifiants [4].

2. https://developer.twitter.com/en/docs/tweets/filter-realtime/overview
3. https://developer.twitter.com/en/enterprise
4. https://developer.twitter.com/en/developer-terms/agreement-and-policy.html

Dans le contexte de la détection d'événements dans une collection de tweets, les auteurs opposent la collecte par mots-clés à celle par filtre géographique (Ozdikis et al., 2017). La collecte par mots-clés ou hashtags est communément employée (Starbird et al., 2010), par exemple dans le contexte d'un tremblement de terre survenu en 2011 (Crooks et al., 2013). Une distribution spatio-temporelle de cet ensemble initial peut ensuite être estimée (Sriram et al., 2010; Helwig et al., 2015).

Dans (Cheng et Wicks, 2014), les auteurs soutiennent que l'analyse de tweets collectés par mot-clés néglige l'impact de phénomènes de diffusion, où les utilisateurs copient et collent du texte sans passer par le mécanisme de *retweet*. Même si les expériences de (Sakaki et al., 2013) montrent que cet impact est faible, la collecte par filtre géographique reste moins biaisée (Ozdikis et al., 2017). Des filtres géographiques sont également utilisés par (Gao et al., 2018). Les retweets et les textes dans une langue autre que l'anglais sont alors exclus. Notons que la collecte de tweets utilisant l'API streaming gratuite souffre de limitations par rapport à l'API *Enterprise*, dont l'impossibilité de collecter les contenus dont le profil de l'utilisateur est dans la Région d'Intérêt (RI) : seuls les contenus explicitement localisés dans la RI sont obtenus. Nous utilisons des filtres géographiques en section 3.1.

Des initiatives ont récemment vu le jour pour favoriser le partage persistent de collections de tweets. Ainsi, le portail Harvard Dataverse [5] référence de tels ensembles afin de favoriser la recherche reproductible. En fait, il est possible d'y trouver un ensemble de 35M d'identifiants associés aux ouragans Harvey et Irma (Littman, 2017). Ceux-ci ont été collectés en filtrant par mots-clés et hashtags : ce corpus est donc en quelque sorte complémentaire à celui que nous présentons en section 3.1. Il a par ailleurs servi de base à une tâche MediaEval récemment (Bischke et al., 2018), où les administrateurs de la tâche en ont filtré un sous-ensemble d'environ 10K élements contenant nécessairement une image, et fait annoter ces images selon qu'elles représentent une voie praticable ou non grâce à la plateforme Figure Eight [6].

3 Méthodologie

3.1 Collecte et stockage du corpus Harvey

De concert avec les partenaires hydrologues du projet Publimape, nous avons défini la région d'intérêt de la collecte comme indiqué sur la figure 1a, pour la période du 19 Août 2017 au 21 Septembre 2017. En figure 1b, on voit que cette période est liée à un pic d'intérêt pour le terme *harvey*.

L'utilisation d'un tel ensemble d'enveloppes convexes rectangulaires est requis pour l'utilisation des API de Twitter. Nous requerrons le contenu pour lequel les champs *coordinates*, *places*, ou *user.location* ont un recouvrement avec la RI. Nous avons ainsi collecté environ 7.5M de tweets.

Nous stockons ces données dans une base de données NoSQL en tant que collection de documents. Afin de faciliter leur analyse du point de vue géographique, nous créons des annotations dans une collection distincte. Une annotation est constituée d'un identifiant de tweet et d'un type (*geotag* ou *bbox*). Pour les annotations de type *geotag*, les coordonnées GPS sont ajoutées directement en tant que champ de l'annotation. Nous avons aussi créé une collection

5. `https://dataverse.harvard.edu/`
6. `https://www.figure-eight.com`

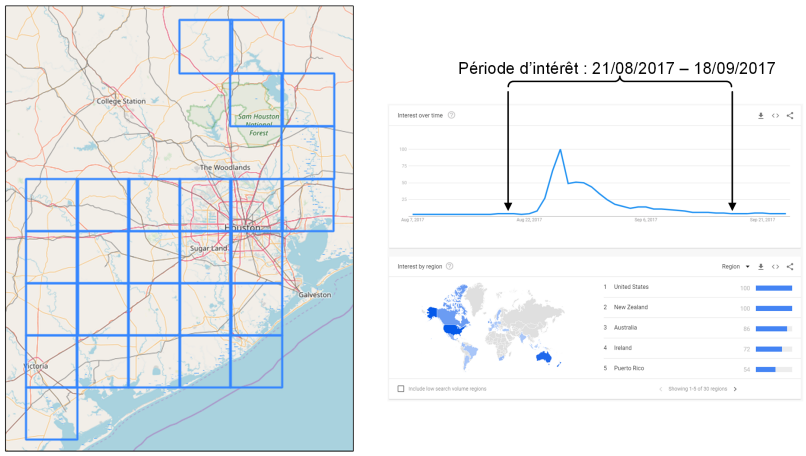

FIG. 1 – a) *Région d'intérêt (RI) ayant servi de filtre à la collecte du corpus.* b) *Intérêt pour le terme* harvey *pendant la période d'intérêt d'après* Google Trends.

où sont stockés les informations des lieux rencontrés dans les champs *bbox* (dont les enveloppes convexes). Les annotations *bbox* font référence aux identifiants de cette collection.

Les tweets obtenus par l'API *Enterprise* ont un champ étendu pour l'utilisateur (*user.derived*), qui contient notamment des noms de lieux normalisés, mais ne fournit pas les enveloppes convexes associées, seulement le *geotag* du barycentre géographique de l'entité nommée. Afin d'encoder des annotations d'un type similaire à *bbox*, nous avons utilisé Nominatim [7], un service de geocodage permettant de retrouver les enveloppes convexes. Le risque de biais est limité car les entités nommées fournies par Twitter ont a priori fait l'objet d'un filtrage par un moteur similaire. Nous sélectionnons ainsi le premier résultat retourné par Nominatim dont l'enveloppe convexe contient le *geotag* barycentrique. Nous stockons ainsi des annotations de type *pbbox* (i.e. *profile bounding box*) afin de différencier l'origine des enveloppes. Tous types confondus, nous avons ainsi stocké environ 8.3M d'annotations, faisant référence à 8434 lieux.

Les tweets ont été collectés selon que leur champ *coordinates*, *place* ou *user* a un recouvrement avec la RI définie sur la figure 1a. Notons cependant que du contenu associé à la fois à un profil dans la zone d'intérêt, et à un géotag ou une enveloppe convexe hors de la zone d'intérêt n'est vraisemblablement pas pertinent pour notre cas d'utilisation (e.g. habitant de Houston en vacances en Europe). Nous réalisons ainsi un post-traitement sur les annotations, excluant le cas de figure ci-dessus. Environ 170K annotations et 4700 lieux sont ainsi exclus. Remarquons qu'environ 2% des annotations et plus de 50% des lieux sont exclus de l'analyse. Cela est cohérent avec l'interprétation indiquée plus haut, et permet a priori d'enlever du bruit de l'analyse présentée dans la prochaine section.

7. https://nominatim.openstreetmap.org/

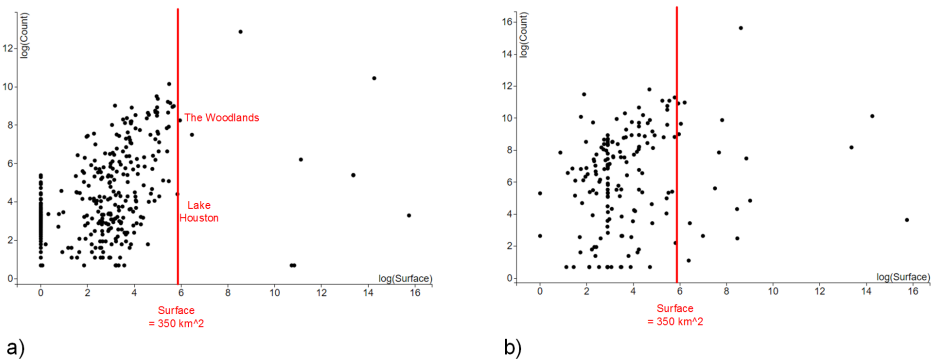

FIG. 2 – *Nuages de points représentant l'effectif en fonction de la surface des annotations* bbox *(*a*) et* pbbox *(*b*).*

3.2 Analyse de la dimension géographique du corpus

L'objectif du projet présenté en introduction est d'associer du contenu émis sur Twitter avec le plus de précision spatiale possible. Dans un premier temps, nous supposons que les annotations de type *geotag* sont exactes. Ces annotations représentent environ 1% du corpus, comme indiqué en figure 4 et conformément à des observations déjà formulées dans la littérature (Middleton et al., 2014).

Nous nous intéressons alors aux surfaces des lieux identifiés dans le corpus, et plus particulièrement au lien entre la surface d'une enveloppe convexe, et sa fréquence dans le corpus : un lieu est-il d'autant plus fréquent qu'il désigne une surface étendue ? Les nuages de points en figure 2 représentent la fréquence d'un lieu dans notre corpus en fonction de la surface de ce lieu. Tant les fréquences des lieux que leur surface respective sont assez bien modélisés par une distribution exponentielle peu lisible. Les nuages de points sont donc représentés dans un repère log-log pour rendre le graphique plus lisible. Les visuels en figure 2 sont issus d'une application interactive où le survol des points révèle le lieu et sa surface, facilitant ainsi l'exploration. Nous représentons séparément les annotations de type *bbox* et *pbbox*.

Dans le cas de *bbox*, il existe une correlation significative selon les tests de Pearson et Kendall ($p < 10^{-10}$), avec un coefficient de corrélation de Pearson estimé à 0.73. En inspectant la figure 2a, on voit qu'assez naturellement, des lieux très spécifiques comme *Cypress Park High School* ne sont mentionnés que 3 fois, alors que *Houston, TX* et *Texas, USA* apparaissent pour respectivement 3.9M et 34K tweets. Dans le contexte de notre projet, ces dernières annotations ne sont pas assez spécifiques. Nous utilisons ensuite la figure 2a pour établir un seuil de spécificité, au-delà duquel sous sommes certains qu'une mention géographique n'a aucun utilité pour notre cas d'étude. Nous établissons qualitativement ce seuil à 350 km^2, ce qui revient à exclure les valeurs isolées à droite du nuage de points en figure 2a. En pratique, cette valeur sépare *Lake Houston* de *The Woodlands* (voir figure 3) : la surface associée est considérable, et rend a priori négligeable le risque d'exclure de l'information utile vu notre cas d'étude.

Nous avons également reporté ce seuil sur la représentation des annotations de type *pbbox* en figure 2b. Dans ce dernier cas, les corrélations de Pearson et de Kendall sont faibles (0.19

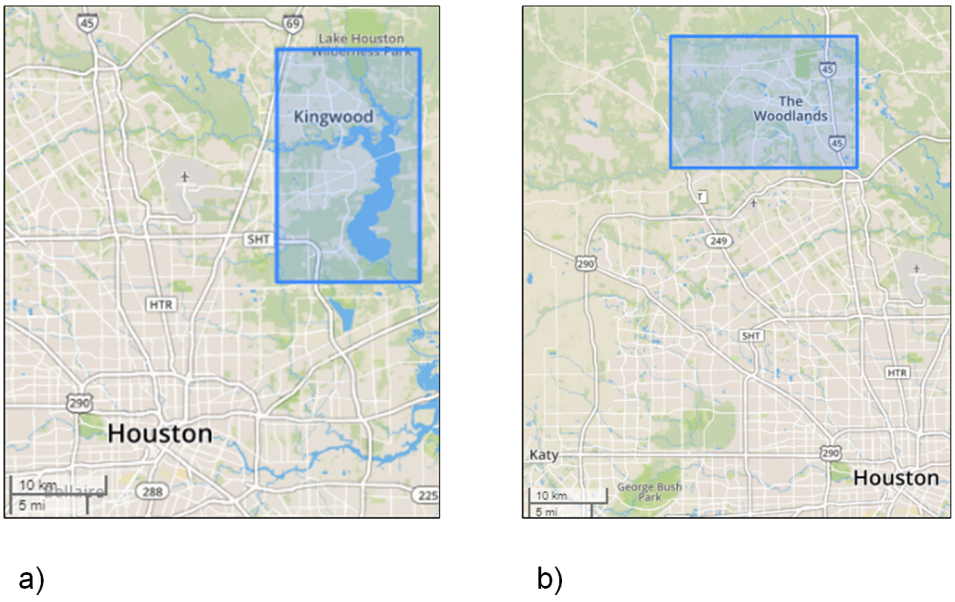

FIG. 3 – *Enveloppes convexes pour les lieux* Lake Houston *(*a*) et* The Woodlands *(*b*).*

dans les deux cas) et assez peu significatives ($p = 0.01$ pour le test de Pearson). Les utilisateurs indiquent souvent une ville comme localisation de profil, ce qui est en général moins spécifique que les lieux donnés précédemment en exemple. Le seuil choisi en conserve néanmoins une grande partie.

Nous utilisons ce seuil pour définir des sous-catégories aux annotations d'enveloppes convexes (*s* pour *small* et *l* pour *large*). La distribution croisée entre les types d'annotations est représentée en figure 4. On voit ainsi que seule 17.4% de l'information géographique renseignée en tant que champs de tweets est utilisable dans notre contexte applicatif.

Sur la figure 4, nous avons également représenté le champ *source*, qui indique l'application à l'origine du tweet. Globalement, l'immense majorité des tweets est émise depuis les clients Twitter iPhone, Android et web (76% pour ces 3 catégories cumulées). Ensuite viennent les contenus émis depuis d'autres plateformes sociales (3% pour Facebook et 2% pour Instagram) et les clients Twitter moins populaires (iPad et TweetDeck avec 1% chacun). Sur l'ensemble du corpus, les sources automatisées de contenus apparaissent comme minoritaires : les deux premières, SocialOomph et IFTTT représentent 1% du corpus chacune.

Grâce à la vue interactive de la figure 4, nous nous concentrons sur les tweets ayant une information géographique pertinente a priori. Parmi les tweets ayant un *geotag*, Instagram est majoritaire (63% de ces derniers). La restriction aux tweets ayant un *geotag* met par ailleurs en avant quelques sources minoritaires potentiellement intéressantes de notre point de vue, par exemple des rapports météorologiques (*CWIS Twitter Feed*) et de traffic automobile (*TTN HOU Traffic*). Nous devrons cependant évaluer dans quelle mesure les *geotags* émis de sources automatisées sont fiables quant à la localisation de l'information véhiculée. En effet, si on peut

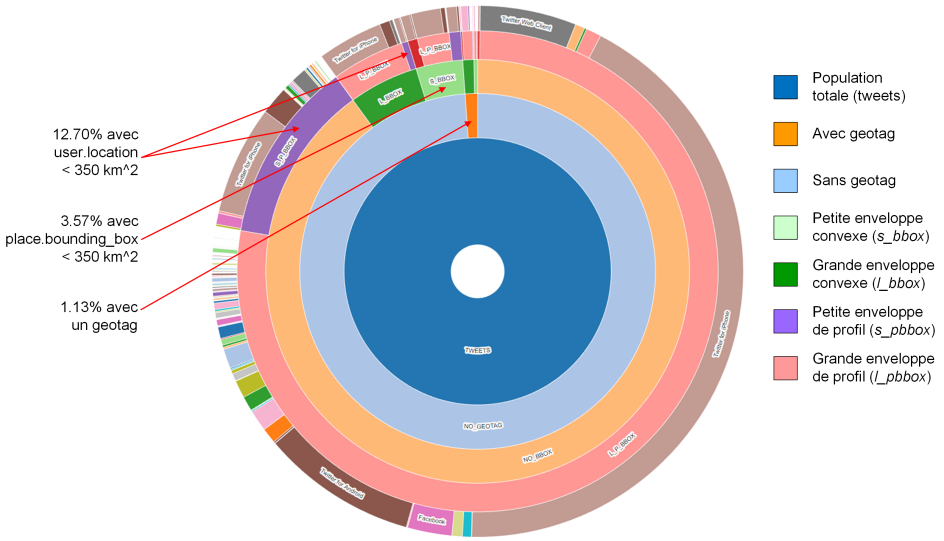

FIG. 4 – *Vue* sunburst *de la distribution croisée des types d'annotations géographiques du corpus. Le champ* source *est également indiqué.*

attacher une relative fiabilité aux applications officielles de Twitter et des autres principaux réseaux sociaux, avec les API de Twitter la localisation peut potentiellement être renseignée de manière arbitraire par un robot.

Par ailleurs, l'écrasante majorité des tweets dans la catégorie *s_bbox* est émise par les clients Twitter officiels. Ces clients officiels représentent les 3 quarts des tweets dans la catégorie *s_pbbox*, le dernier quart regroupant les clients moins populaires mentionnés plus hauts, ainsi que de nombreuses sources minoritaires (dont l'essentiel du contenu émis par IFTTT). Il semble donc raisonnable dans un premier temps de concentrer l'analyse sur les annotations *geotag* et *s_bbox*.

Afin de faciliter la discussion avec nos partenaires hydrologues, nous avons filtré les tweets contenant au moins un des mots-clés *flood* ou *harvey* dans les catégories *geotag* et *s_bbox*. Nous représentons les 19K tweets ainsi sélectionnés sur une carte de densité *heatmap* en figure 5a. En particulier, les hydrologues aimeraient savoir si il nous serait a priori possible d'extraire de l'information sociale pour la région de Wharthon. En zoomant sur cette région, on identifie 81 tweets potentiellement intéressants (voir figure 5b). En enlevant le filtre textuel, ce nombre passe à 3600 environ. Ainsi, cela isole assez logiquement une portion minime de notre corpus, la densité des contenus étant forcément liée à la densité de population. Cependant, ce sous-ensemble est suffisamment petit pour permettre de réaliser une inspection manuelle des contenus concernés, ce qui peut être utile pour réaliser des tests qualitatifs.

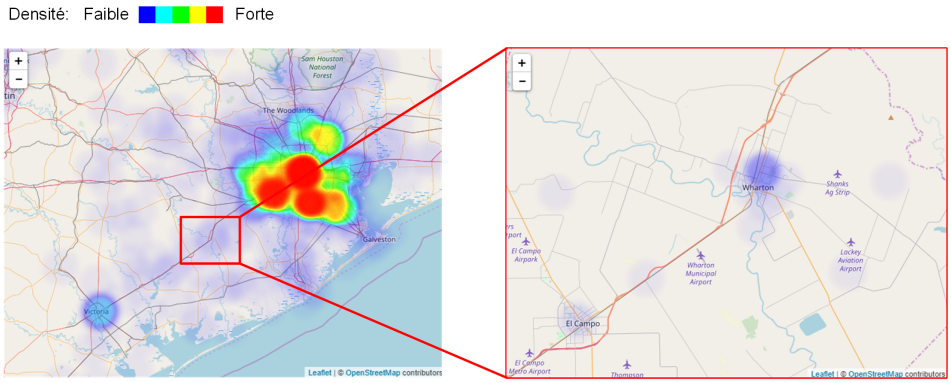

FIG. 5 – *Carte de densité des catégories* geotag *et* s_bbox *parmi les tweets dont le texte contient* flood *ou* harvey *(*a*), zoomée sur la région de Wharthon (*b*).*

4 Conclusion

Nous avons pu voir que seule une faible proportion de notre corpus peut être géolocalisée de manière satisfaisante compte tenu de notre cas d'utilisation. Nous pourrions toutefois tirer parti de l'information géographique présente dans le texte des tweets, e.g. *"The Intersection of Asford Pkwy and Dairy Ashford Rd is singificantly higher than yesterday"* dans notre corpus. Une telle extraction a déjà été considérée dans la litérature sur l'extraction d'événements dans un corpus de tweets (Middleton et al., 2014). Nous pourrions bénéficier des derniers progrès de l'état de l'art en reconnaissance d'entités nommées, e.g. (Chiu et Nichols, 2015). D'autres travaux ont exploité la géographicité des tags et du langage (Kordopatis-Zilos et al., 2016), mais il serait a priori difficile de transférer de telles approches à notre échelle.

L'objectif du projet est in fine d'agréger le contenu classifié sur une grille 2D. Des approches de classification spatio-temporelle ont été proposées dans la littérature (Helwig et al., 2015; Anantharam et al., 2015; Tamura et Ichimura, 2013), parfois couplées à des techniques permettant de prendre en compte la densité globale d'émission de tweets (Gao et al., 2018), et la détection de pics thématiques (Atefeh et Khreich, 2015; Cordeiro et Gama, 2016). En pratique, nous allons explorer l'apport possible de techniques d'active learning et de crowdsourcing afin de caractériser les contenus de manière plus fine que par mots-clés tel que vu dans la section 3.2, et tirer parti de la multi-modalité du contenu étudié (i.e. texte, coordonnées géographiques, temps, image), sous la contrainte de valeurs manquantes (Brangbour et al., 2018).

5 Remerciements

Ce travail a été réalisé dans le contexte du projet Publimape, financé par le programme CORE du Fonds National de la Recherche luxembourgeois (FNR).

Références

Anantharam, P., P. Barnaghi, K. Thirunarayan, et A. Sheth (2015). Extracting City Traffic Events from Social Streams. *ACM Trans. Intell. Syst. Technol. 6*(4), 43 :1–43 :27.

Atefeh, F. et W. Khreich (2015). A Survey of Techniques for Event Detection in Twitter. *Computational Intelligence 31*(1), 132–164.

Bates, P. et A. De Roo (2000). A simple raster-based model for flood inundation simulation. *Journal of hydrology 236*(1-2), 54–77.

Bischke, B., P. Helber, Z. Zhao, J. de Bruijn, et D. Borth (2018). The Multimedia Satellite Task at MediaEval 2018. In *MediaEval 2018*, pp. 3.

Brangbour, E., P. Bruneau, et S. Marchand-Maillet (2018). Extracting Flood Maps from Social Media for Assimilation. In *IEEE eScience*.

Burke, J., D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, et M. Srivastava (2006). Participatory sensing. In *ACM WSW*.

Cheng, T. et T. Wicks (2014). Event Detection using Twitter : A Spatio-Temporal Approach. *PLOS ONE 9*(6), e97807.

Chiu, J. P. C. et E. Nichols (2015). Named Entity Recognition with Bidirectional LSTM-CNNs. *arXiv :1511.08308 [cs]*.

Cordeiro, M. et J. Gama (2016). Online Social Networks Event Detection : A Survey. In S. Michaelis, N. Piatkowski, et M. Stolpe (Eds.), *Solving Large Scale Learning Tasks. Challenges and Algorithms : Essays Dedicated to Katharina Morik on the Occasion of Her 60th Birthday*, pp. 1–41. Springer International Publishing.

Crooks, A., A. Croitoru, A. Stefanidis, et J. Radzikowski (2013). #Earthquake : Twitter as a Distributed Sensor System. *Transactions in GIS 17*(1), 124–147.

Gao, Y., S. Wang, A. Padmanabhan, J. Yin, et G. Cao (2018). Mapping spatiotemporal patterns of events using social media : a case study of influenza trends. *International Journal of Geographical Information Science 32*(3), 425–449.

Griesbaum, L., S. Marx, et B. Höfle (2017). Direct local building inundation depth determination in 3-D point clouds generated from user-generated flood images. *Natural Hazards and Earth System Sciences 17*(7), 1191–1201.

Helwig, N. E., Y. Gao, S. Wang, et P. Ma (2015). Analyzing spatiotemporal trends in social media data via smoothing spline analysis of variance. *Spatial Statistics 14*, 491–504.

Hostache, R., G. Corato, M. Chini, M. Wood, L. Giustarini, et P. Matgen (2015). A new approach for improving flood model predictions based on the sequential assimilation of SAR-derived flood extent maps. In *EGU General Assembly Conference Abstracts*, Volume 17.

Kordopatis-Zilos, G., A. Popescu, S. Papadopoulos, et Y. Kompatsiaris (2016). Placing Images with Refined Language Models and Similarity Search with PCA-reduced VGG Features. In *MediaEval 2016*, pp. 3.

Littman, J. (2017). Hurricanes Harvey and Irma Tweet ids. `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/QRKIBW`.

Middleton, S. E., L. Middleton, et S. Modafferi (2014). Real-Time Crisis Mapping of Natural Disasters Using Social Media. *IEEE Intelligent Systems 29*(2), 9–17.

Ozdikis, O., H. Oğuztüzün, et P. Karagoz (2017). A survey on location estimation techniques for events detected in Twitter. *Knowledge and Information Systems 52*(2), 291–339.

Sakaki, T., M. Okazaki, et Y. Matsuo (2013). Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development. *IEEE Transactions on Knowledge and Data Engineering 25*(4), 919–931.

Sriram, B., D. Fuhry, E. Demir, H. Ferhatosmanoglu, et M. Demirbas (2010). Short Text Classification in Twitter to Improve Information Filtering. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 841–842. ACM.

Starbird, K., L. Palen, A. Hughes, et S. Vieweg (2010). Chatter on the Red : What Hazards Threat Reveals About the Social Life of Microblogged Information. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, CSCW '10, New York, NY, USA, pp. 241–250. ACM.

Tamura, K. et T. Ichimura (2013). Density-Based Spatiotemporal Clustering Algorithm for Extracting Bursty Areas from Georeferenced Documents. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 2079–2084.

Summary

In the context of the analysis and prevention of natural hazards, coupling physical modelling and social media data extraction has raised considerable interest recently. A necessary step is to collect and qualify corpora of content emitted on social media platforms. In this paper specifically, we focus on the collection of Twitter content using spatio-temporal filters, and the reliability of spatial data fields featured in tweets. We illustrate our discussion by studying the link between the frequency of places names in the corpus and their surface, as well as the influence of the application that originated the tweets. We build upon our observations to estimate the relevance of a hydrological region of interest.