An Instrumented Methodology to Analyze and Categorize Information Flows on Twitter Using NLP and Deep Learning: A Use Case on Air Quality

Brigitte Juanals, Jean-Luc Minel

An Instrumented Methodology to Analyze and Categorize Information Flows on Twitter Using NLP and Deep Learning : A Use Case on Air Quality

B. Juanals[1,2], J.L. Minel[3]
[1]Centre Norbert Elias, Aix Marseille University - CNRS - EHESS, Marseille, France
[2]UMI IGlobes, CNRS - University of Arizona, Tucson, USA
[3]MoDyCo, University Paris Nanterre - CNRS, Nanterre, France

Abstract. This article focuses on the development of an instrumented methodology for modeling and analyzing the circulation message flows concerning air quality on the social network Twitter. This methodology aims at describing and representing, on the one hand, the modes of circulation and distribution of message flows on this social media and, on the other hand, the content exchanged between stakeholders. To achieve this, we developed Natural Language Processing (NLP) tools and a classifier based on Deep Learning approaches in order to categorize messages from scratch. The conceptual and instrumented methodology presented is part of a broader interdisciplinary methodology, based on quantitative and qualitative methods, for the study of communication in environmental health. A use case of air quality is presented.

Keywords: Air quality; Instrumented methodology; Circulation of information; Deep Learning ; Social network; Twitter

1 Introduction

This article proposes the development of an instrumented methodology for modeling and analyzing the flow of messages about air quality on the Twitter social network platform. In the areas of health and the environment, organizations' commitment to digital media and the social web is one of the new forms of mediation and media coverage developed for the public. Organizations have incorporated a social media activity into their communication policies; they follow the evolutions of the media practices of the public. However, the use of the socio-digital networks raises new methodological problems related to the description and analysis of new information-communication practices. They concern the specificities of the editorialization of information on these devices, the volume of messages and the interactions they allow.

The purpose of our work is to conceive, by relying jointly on methods anchored in social sciences and digital humanities, a representation of the modes of

circulation and distribution of message flows on Twitter in relation with stakeholders and the content exchanged. Twitter was chosen because, besides being a widely used social media, it offers the possibility, through an API, of collecting messages from a set of user accounts from associated hashtags, which cannot be done on other social media platforms (Facebook, Instagram). The methodology is based on the contribution of work on the communication of organizations on health and environmental issues on the social web [17] [19], social media platforms and media communication to the understanding of phenomena such as the mediation and circulation of information in these sociotechnical environments. The outline of this paper is the following. First, in Section 2, we present the literature review. In Section 3, we will present our methodology to collect and analyze flows of tweets. In Section 4, a use case on air quality is presented. Finally, we conclude in Section 5.

2 Literature Review

As mentioned in [5], tweet analysis has led to a large number of studies in many domains such as ideology prediction in Information Sciences [4], spam detection [20], dialog analysis in Linguistics [2], and natural disaster anticipation in Emergency [16], while work in Social Sciences and Digital Humanities has developed tweet classifications [18]. Recently, several studies on tweet classification have been carried out in NLP [10]. Basically, these analyses aim at categorizing open-domain tweets using a reasonable amount of manually classified data and either small sets of specific classes (e.g. positive versus negative classes in sentiment analysis) or larger sets of generic classes (e.g. News, Events and Memes classes in topic filtering). Associating NLP and machine learning techniques, [6] have classified institutional tweets in communication categories. Until recently, the most commonly used models were supervised learning, Support Vector Machine (SVM), Random Forest, Gradient Boosting Machine and Naive Bayes (NB) [13]. In supervised machine learning, features are extracted from tweets and metadata and then vectorized as training examples to build models. But lately, deep learning models applied to natural language processing tasks have achieved remarkable results [15]. Moreover, [8] reported on a series of experiments with convolutional neural networks and showed "that a simple CNN with little hyperparameter tuning and static vectors achieves excellent results on multiple benchmarks". Our model of deep learning classifier is largely inspired of this work.

3 An instrumented Methodology

The proposed methodology aims to describe and analyze the informational and communicational dynamics at work on the Twitter platform. It explores the circulation patterns of message flows and exchanges, apprehended as a dynamic process, as well as the relationships which are established between different stakeholders. Our aim is to question the forms of engagement, participation

and relationships between organizations and audiences by analyzing the flow of their messages. We consider that the field of environmental health and the socio-technical device Twitter contribute to configuring the relations and the interactions between the participants. In this perspective, hashtags on Twitter may be seen as meeting points of different categories of stakeholders interested in the same themes while being anchored in different spheres - the environment and health, politics, the media, industry, the economy, etc.

Our analysis focuses on Twitter messages (called tweets) sent by accounts of organizations and non-institutional stakeholders. Concerning messages, we will call a message sent by a twitter account an 'original tweet' and a message sent by an account different from the issuing account a 'retweet'. The current Twitter API gives access to the original tweet (and its sending account) of a retweet. The generic term tweet includes 'original tweet' and 'retweet'. Regarding stakeholder qualification, we distinguished accounts managed by institutions (called 'organizational account'), and accounts managed by individuals (called 'private account').

Our methodology is broken down into several steps. First, analyze the set of data collected by applying a t-SNE analysis to get a global picture of all the data. Second, train a classifier based on a convolutional neural networks (cf. below) to categorize the data and analyze the modes of involvement and interaction between organizational and private accounts. Third, carrying out a lexical analysis to identify the topics addressed. Finally, identifying networks and passing accounts. We implemented our whole methodology by building a workflow based on open access tools. We also developed some scripts python to manage the interoperability between all these different tools.

About the second step, the main drawback of shallow supervised machine learning approaches, is that they require a very time consuming step to identify linguistic or semiotic features and raise issues about the relevance of these features. Recently, new approaches based on Deep Learning techniques and especially on convolutional neural networks (convets), which no longer require researchers to identify features, have been proposed. A second advantage of convets is that they obtain better results in terms of accuracy than shallow machine learning systems [8,14]. It is for these reasons that we developed a classifier based on convets.

The architecture of our classifier is composed of several layers [7]. The first layer is a pre-trained word embedding as proposed by [15] with a kernel that matches the 5 words used as neighbors. A word embedding is a distributed representation where each word is mapped to a fixed-sized vector of continuous values. The benefit of this approach is that different words with a similar meaning will have a similar representation. A fixed-vector size of 100 was chosen. The following layers, a Conv1D with 200 filters and a MaxPooling1D are based on the works reported in [8,3]. The back-end of the model is a standard Multilayer Perceptron layer to interpret the convets features. The output layer uses a softmax activation function to output a probability for each of the three classes affected at the tweet processed. Finally, only the class with the highest probability is

kept. The evaluation using the standard cross-validation 10-fold test [9] gave an accuracy of 0.97 in line with the state of the arts [8].

4 A Use Case on Air Quality

Environmental health is an emerging and hotly debated topic that covers several fields of study such as pollution in urban or rural environments and the consequences of these changes on health populations. The environmental factors analyzed fall into four broad thematic dimensions relating to polluted sites and soils, water quality and air quality and habitat. Among these factors, we focused on air quality, which was the subject of many alerts in major cities at the end of 2016 and which is becoming an international concern with the regular peaks of fine particles matter in urban areas.

The data acquisition stage consisted in harvesting tweets with the following hashtags: the hashtag #Air and one other hashtag among the following list: #pollution, #santé (health), #qualité(quality) or #environnement(environment). In this paper, we limit the analysis to French tweets, by using the "lang" features in Tweeter API, sent between the first of November 2017 and the 30th of July 2018. This period of time is considered as a proof of concept and we intent to use the classifier to process all the tweets that will be sent during the year 2018. The main figures are the following: 4 832 tweets of which 39% of original tweets and 61% of retweets sent by 2517 participants (405 organizational accounts, 2112 private accounts). More specifically: 41% of organizational accounts and only 30% of private accounts produced original tweets. Participation for private accounts was largely limited to the action of retweeting (75% of tweets) the messages sent by the institutional partners. In order to obtain global picture of all the accounts, we applied the t-SNE algorithm [12] based on 5 dimensions proposed by [6]: type of account, number of original tweets sent, relayed score, relaying score, mentioned score. The main interest of this algorithm is to take a set of points in a high-dimensional space and find a faithful representation of those points in a 2D plane (cf. figure 1). We tuned the algorithm features as recommended in [12] and finally perplexity = 30 and iteration = 500 were chosen. In other words, it offers a global vision of practices of production, diffusion and interaction (mention, retweet, quote).

Categorical analysis relying on automatic classification is a relevant processing method to characterize the semantics of the messages as it makes possible to analyze the modes of engagement of the stakeholders on Twitter. Automatic classification implies a prior human classification (supervised machine learning). Taking into account the size of the corpus of tweets, a human analysis would still have been possible but first, as mentioned in [11], the inter-coder minimum reliability is usually around 0.74, and secondly, we intend to process in real time all the tweets that will be sent during the year 2018. These two reasons argue in favour of developing an automatic classification.

In order to build a classifier, a classification analysis of the contents of a sample of 350 randomly chosen original tweets was carried out in two stages

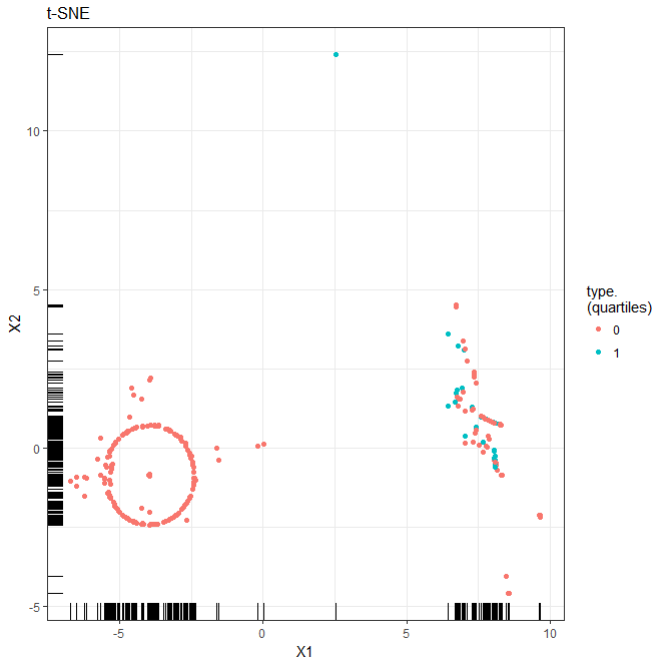

Fig. 1. t-SNE analysis

[7]. First, these 350 tweets were annotated by hand by two experts using three classes: “informative”, “promoting”, “humorous” (in the sense of emotional or expressive). In the second stage, a classifier, based on convets (see Section 3), was trained on the annotated samples. The training loss decreases with every epoch, and the training accuracy increases with every epoch; to prevent overfitting, the training was stopped after four epochs. The classifier was applied to the corpus of all tweets to categorize them. Table 1 show the findings. There is a main differ-

Table 1. Automatic classification of original tweets

Categorization	Institutional accounts	Private accounts
Informative	58%	53%
Promoting	32%	25%
Humorous	10%	22%

ence between institutional and individuals accounts concerning the “humorous” class. Private accounts sent twice as many “Humorous” tweets as institutional accounts. Similarly, institutional accounts sent much more “Promoting” tweets.

Observing the flow of information through the circulation of messages involves looking at the modes of stakeholder participation. They are materialized in information-communication practices. To answer this question, the analysis is based on the classification of the accounts.

In order to characterize Twitter accounts we used some of attributes proposed by [6] : "relayed", "relaying", "mentioned" and "passing". As pointed out by [6], the value of this index is not significant in itself; it simply provides a means of comparing accounts. We identified six passing accounts that had a significant score [7]. The analysis of these passing accounts makes it possible to identify some of their characteristics. These are all accounts of organizations with the exception of one influencer. It is remarkable these key influential accounts do not share their communities of accounts.

From the whole corpus, the data were partioned in restricted subcorpora built according to the criteria of the type of stakeholder (organizational or individual). These limited corpora enable future analyses (content analyses or discourse analyses) related to the status and role of the stakeholders to be carried out. In the space of this article, we focus on the analysis of several graphs. The aggregated communities are computed by applying the Louvain algorithm [1]; they highlights several points. From the whole graph (cf. on the left figure 2), if we select the organizational accounts (cf. on the middle figure 2), communities are linked by their retweet policy. One community is related with Anne Hidalgo (Mayor of Paris), a second one is related with Ambassad'AIr (a NGO), a third one is related with Air Paca (another NGO) and finally another community is related to the Prefect of the Occitanie Region (representing the French goverment). These communities, which share a few links between them, correspond to the territorial and administrative organization of the French regions whose communication activity is most apparent on twitter. If we select the private accounts from the whole graph (not shown), three accounts can be identified; they do not share any users. Private accounts do not communicate with each other, they do not mention or retweet each other (cf. on the far right figure 2); these three accounts are retweeted, quoted or mentioned. Their participation in the flow of information is limited to a subset of their followers.

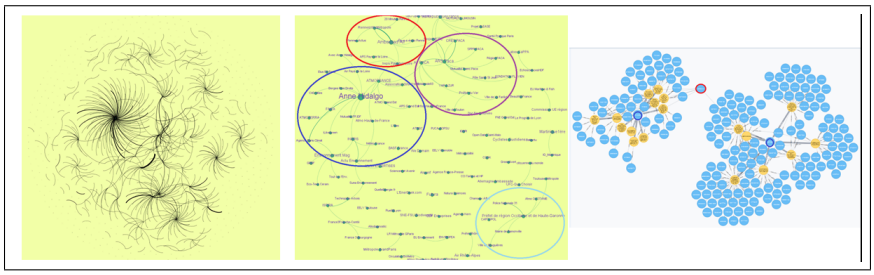

Fig. 2. Whole graph, organizational accounts graph and two communities

We computed the first ten significant terms and hashtags in the tweets to identify the main themes of the exchanges. The 10 most frequent terms in the text of the tweets (after deleting prepositions, conjunctions and hashtags) are the following. Two words "villes" (towns) and "intérieur" (inside) evoke places of observed pollution. Three words "défi" (challenge), "vigilance" (vigilance) "informer" (to inform), evoke the concerns about air quality and the concerns about air quality, monitoring pollution, and informing the public. Three words "préfectoral" (prefectural), "ballon" (balloon) and "dispositif" (device) evoke the administrative and technological tools used to monitor air quality. It is surprising that the widely discussed causes of air pollution related to "pesticides" (pesticides) and "particules" (particles matter) are hardly mentioned (they have a very low number of occurrences). The 10 most frequent hashtags in the text of the tweets (after deleting hashtags used to query the Twitter API)are the following. Five hashtags refer to geographical places. Three cities "Paris", "Rennes" and "Marseille" and two regions "Martinique" and "Haute Garonne" are the most frequently mentioned and show a higher communication activity related to political positions. Two hashtags refer to pollution events "PicPollution", (PeakPollution) and public or activist exhortations in favor of limiting activities that generate pollution "stopPollution". The hashtag "mobilité" (mobility) refers to the issue of the urban traffic.

5 Conclusions

We described an instrumented methodology for the analysis of the flow of messages on the Twitter platform. This methodology is based on a multi-dimensional approach in order to apprehend the complexity of the modes of circulation of messages. We developed a workflow to implement our methodology; an essential part of this workflow is a classifier based on convolutional neural networks which categorize the flow of tweets.

This research is part of a broader framework of ongoing work on the evolution of environmental communication in the public space. We are applying the same methodology on English tweets, during the same period of time, in order to carry out a comparative analysis between anglophone and francophone audience.

6 Acknowledgments

This study is partially funded by iGlobes (UMI 3157).

References

1. Blondel, V., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics 2008(10) (2008)
2. Boyd, D., Golder, S., Lotan, G.: Tweet, tweet, retweet conversational aspects of retweeting on twitter. In: 43rd Hawaii International Conference on System Sciences, HICSS. pp. 1–10 (2010)

3. Brownlee, J.: Deep Learning for Natural Language Processing. Machine Learning Mystery, Vermont, Australia (2017)
4. Djemili, S., Longhi, J., Marinica, C., Kotzinos, D., Sarfat, G.E.: What does twitter have to say about ideology? In: NLP 4 CMC: Natural Language Processing for Computer-Mediated Communication. pp. 16–25 (2014)
5. Foucault, N., Courtin, A.: Automatic classification of tweets for analyzing communication behavior of museums. In: LREC 2016. pp. 3006–3013 (2016)
6. Juanals, B., Minel, J.L.: Analysing cultural events on twitter. In: Lecture Notes in Artificial Intelligence, Computational Collective Intelligence. vol. 10449, pp. 376–385. Springer (2017)
7. Juanals, B., Minel, J.L.: Categorizing air quality information flow on twitter using deep learning tools. In: Lecture Notes in Artificial Intelligence, Computational Collective Intelligence. vol. 11055, pp. 1–10. Springer (2018)
8. Kim, Y.: Convolutional neural networks for sentence classification. CoRR abs/1408.5882 (2014), http://arxiv.org/abs/1408.5882
9. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. pp. 1137–1143. Morgan Kaufmann (1995)
10. Kothari, A., Magdy, W., Darwish, K., Mourad, A., Taei, A.: Detecting comments on news articles in microblogs. In: Kiciman, E., al. (eds.) 7th In-ternational Conference on Web and Social Media (ICWSM). The AAAI Press (2013)
11. Lachlan, K., Spence, P., Lin, X., M., D.G.: Screaming into the wind:examining the volume and content of tweets associated with hurricane sandy. Communication Studies 65(5), 500–518 (2014)
12. Van der Maaten, L., Hinton, G.: Visualizing high-dimensional data using t-sne. Journal of Machine Learning Research 9, 2579–2605 (2008)
13. Malandrakis, N., Falcone, M., Vaz, C., Bisogni, J.J., Potamianos, A., Narayanan, S.: Sail: Sentiment analysis using semantic similarity and contrast features. In: 8th International Workshop SemEval. pp. 512–516 (2014)
14. Manning, C.D.: Computational linguistics and deep learning. Computational Linguistics 41(4), 701–707 (2015), https://doi.org/10.1162/COLI_a_00239
15. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013), http://arxiv.org/abs/1301.3781
16. Sakaki, T., Okazaki, M., Matsuo, Y.: Tweet analysis for real-time event detection and earthquake reporting system development. IEEE Transactions on Knowledge and Data Engineering 25(4), 919–931 (2013)
17. Schmidt, C.: Trending now: Using social media to predict and track disease outbreaks. Environment Health Perspectives 120-1 (2012)
18. Shiri, A., Rathi, D.: Twitter content categorisation: A public library perspective. Journal of Information and Knowledge Management 12(4) (2013)
19. Thackeray, R., Neiger, B., Smith, A., Van Wagenen, S.: Adoption and use of social media among public health departments. BMC Public Health 12:242 (2012)
20. Yamasaki, S.: A trust rating method for information providers over the social web service: A pragmatic protocol for trust among information explorers, and provider in formation. In: 11th Annual International Symposium on Applications and the Internet (SAINT'11). pp. 578–582 (2011)