université PARIS-SACLAY

CentraleSupélec

NNT : 20XXSACLXXXX

Thèse de doctorat

Thesis title. It can extend over several lines (even 4 or 5)

Thèse de doctorat de l'Université Paris-Saclay
préparée à Nom de l'établissement

École doctorale n°000 Dénomination (Sigle)
Spécialité de doctorat: voir spécialités par l'ED

Thèse présentée et soutenue à Ville de soutenance, le Date, par

FIRSTNAME LASTNAME

Composition du Jury :

Prénom Nom Statut, Établissement (Unité de recherche)	Président
Prénom Nom Statut, Établissement (Unité de recherche)	Rapporteur
Prénom Nom Statut, Établissement (Unité de recherche)	Rapporteur
Prénom Nom Statut, Établissement (Unité de recherche)	Examinateur
Prénom Nom Statut, Établissement (Unité de recherche)	Directeur de thèse
Prénom Nom Statut, Établissement (Unité de recherche)	Co-directeur de thèse
Prénom Nom Statut, Établissement (Unité de recherche)	Invité
Prénom Nom Statut, Établissement (Unité de recherche)	Invité

Dedication

Acknowledgements

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed sollicitudin massa vel venenatis dictum. Aliquam erat volutpat. Phasellus accumsan eu felis at luctus. Integer neque elit, venenatis sed iaculis in, tincidunt nec augue. Aliquam erat volutpat. Nulla sodales tortor non justo tincidunt, non varius risus mollis. Aliquam est purus, cursus at nulla ac, sollicitudin placerat diam. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Ut at leo eget metus scelerisque venenatis. Sed quis dui nisi. Morbi sodales, leo ac scelerisque malesuada, libero sem placerat ante, sit amet ullamcorper ligula nulla vestibulum tellus. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed sollicitudin massa vel venenatis dictum. Aliquam erat volutpat. Phasellus accumsan eu felis at luctus. Integer neque elit, venenatis sed iaculis in, tincidunt nec augue. Aliquam erat volutpat. Nulla sodales tortor non justo tincidunt, non varius risus mollis. Aliquam est purus, cursus at nulla ac, sollicitudin placerat diam. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Ut at leo eget metus scelerisque venenatis. Sed quis dui nisi. Morbi sodales, leo ac scelerisque malesuada, libero sem placerat ante, sit amet ullamcorper ligula nulla vestibulum tellus.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed sollicitudin massa vel venenatis dictum. Aliquam erat volutpat. Phasellus accumsan eu felis at luctus. Integer neque elit, venenatis sed iaculis in, tincidunt nec augue. Aliquam erat volutpat. Nulla sodales tortor non justo tincidunt, non varius risus mollis. Aliquam est purus, cursus at nulla ac, sollicitudin placerat diam. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Ut at leo eget metus scelerisque venenatis. Sed quis dui nisi. Morbi sodales, leo ac scelerisque malesuada, libero sem placerat ante, sit amet ullamcorper ligula nulla vestibulum tellus.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed sollicitudin massa vel venenatis dictum. Aliquam erat volutpat. Phasellus accumsan eu felis at luctus. Integer neque elit, venenatis sed iaculis in, tincidunt nec augue. Aliquam erat volutpat. Nulla sodales tortor non justo tincidunt, non varius risus mollis. Aliquam est purus, cursus at nulla ac, sollicitudin placerat diam. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Ut at leo eget metus scelerisque venenatis. Sed quis dui nisi. Morbi sodales, leo ac scelerisque malesuada, libero sem placerat ante, sit amet ullamcorper ligula nulla vestibulum tellus. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed sollicitudin massa vel venenatis dictum. Aliquam erat volutpat. Phasellus accumsan eu felis at luctus. Integer neque elit, venenatis sed iaculis in, tincidunt nec augue. Aliquam erat volutpat.

Nulla sodales tortor non justo tincidunt, non varius risus mollis. Aliquam est purus, cursus at nulla ac, sollicitudin placerat diam. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Ut at leo eget metus scelerisque venenatis. Sed quis dui nisi. Morbi sodales, leo ac scelerisque malesuada, libero sem placerat ante, sit amet ullamcorper ligula nulla vestibulum tellus.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed sollicitudin massa vel venenatis dictum. Aliquam erat volutpat. Phasellus accumsan eu felis at luctus. Integer neque elit, venenatis sed iaculis in, tincidunt nec augue. Aliquam erat volutpat. Nulla sodales tortor non justo tincidunt, non varius risus mollis. Aliquam est purus, cursus at nulla ac, sollicitudin placerat diam. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Ut at leo eget metus scelerisque venenatis. Sed quis dui nisi. Morbi sodales, leo ac scelerisque malesuada, libero sem placerat ante, sit amet ullamcorper ligula nulla vestibulum tellus.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed sollicitudin massa vel venenatis dictum. Aliquam erat volutpat. Phasellus accumsan eu felis at luctus. Integer neque elit, venenatis sed iaculis in, tincidunt nec augue. Aliquam erat volutpat. Nulla sodales tortor non justo tincidunt, non varius risus mollis. Aliquam est purus, cursus at nulla ac, sollicitudin placerat diam. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Ut at leo eget metus scelerisque venenatis. Sed quis dui nisi. Morbi sodales, leo ac scelerisque malesuada, libero sem placerat ante, sit amet ullamcorper ligula nulla vestibulum tellus.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed sollicitudin massa vel venenatis dictum. Aliquam erat volutpat. Phasellus accumsan eu felis at luctus. Integer neque elit, venenatis sed iaculis in, tincidunt nec augue. Aliquam erat volutpat. Nulla sodales tortor non justo tincidunt, non varius risus mollis. Aliquam est purus, cursus at nulla ac, sollicitudin placerat diam. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Ut at leo eget metus scelerisque venenatis. Sed quis dui nisi. Morbi sodales, leo ac scelerisque malesuada, libero sem placerat ante, sit amet ullamcorper ligula nulla vestibulum tellus.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

According to the Reuters Institute, 36% of French adults use social media as a source of news.[1] This share has declined since 2017, but this is mostly due to a decrease in the use of Facebook, while other networks are stable (like Twitter) or growing rapidly. This evolution of news consumption may reflect a growing interest for stories that are usually not covered by traditional news media, or that are covered in a different way. This raises the question of the type of news that is mostly shared on social networks: are the citizens differently informed when they use social network as a gateway to news?

In response to the transformation of news consumption, one should expect a change in the production of news by traditional media outlets. McGregor and Molyneux [14] find that journalists using Twitter as part of their daily work consider tweets as newsworthy as headlines from the Associated Press. Does this change in the perception of journalists has an effect on the type of stories they choose to report? Does the success of a story on social media impacts the news production by traditional news media outlets?

The objective of this thesis is to investigate the role of Twitter in the evolution of both news consumption and news production in recent years. We aim at understanding what kind of news are amplified by the sphere of social networks, and, conversely, to show in which cases events born on the social networks become a subject relayed by traditional media outlets. The challenge is to quantify and analyze precisely the relationships between the two spheres, in a context of very strong influence of each sphere on the other.

Indeed, stories do not spread only in one direction, from news media to social networks or from social networks to news media. The recent Benalla case in France can be used as an illustration of the different rebounds that an event can have in both spheres: videos of Alexandre Benalla wearing a police helmet and hitting a protester where published on social media on May 1, 2018. However, he was only identified as an aide from President Macron's office on July 18, by the newspaper *Le Monde*. After this first article, traditional media outlets started to investigate on the missions of Mr. Benalla. Both journalists and Twitter users published numerous pictures

[1] http://media.digitalnewsreport.org/wp-content/uploads/2018/06/digital-news-report-2018.pdf?x89475

Notes: These preliminary results were nuanced in the full study published on August 8, 2018

Figure 1.1: Tweet from DisinfoLab, an organization working on disinformation, about the "overactivity" of some Twitter accounts during the Benalla scandal

of Alexandre Benalla in official appearances of the President, including during the period when he was allegedly suspended. The newspaper "Sud Ouest" used the term "photo hunt" to qualify the attitude of social media users and photo reporters[2]. On July 30, a Belgian organization called DisinfoLab evoked an artificial swelling of the number of messages on Twitter related to the Benalla case.[3] In the partial results published on that day, the "overactivity" of some Twitter accounts and "pro-Russian" accounts were mentioned. On August 8, DisinfoLab published the entire study,[4] showing no evidence that an organized Russian intervention has sought to amplify the Benalla case on Twitter. However, several media outlets had already relayed the (wrong) information that "Russian bots" had influenced the reaction of the public on Twitter.[5]

This is only one example of the plurality of "interaction patterns" [17] between social networks and traditional media that can occur in the news agenda. Here, social networks are first used as a source by news outlets (most videos of Alexandre Benalla used by journalists where initially published on Twitter by witnesses of the May 1 demonstration). Then, they participate to the controversy raised by traditional media outlets and amplify it. Finally, the amplitude of the echo on Twitter is discussed by media outlets. Other types of interaction patterns exist,

[2] https://www.sudouest.fr/2018/07/23/affaire-benalla-quand-les-reseaux-sociaux-s-amusent-5256018-10458.php
[3] https://twitter.com/DisinfoEU/status/1023903729668575242
[4] https://spark.adobe.com/page/Sa85zpU5Chi1a/
[5] https://abonnes.lemonde.fr/les-decodeurs/article/2018/08/08/l-impossible-quete-des-bots-russes-de-l-affaire-benalla_5340540_4355770.html?

for example “break on Twitter first” stories (like the hashtag #MakeOurPlanetGreatAgain posted on June 2017 by Emmanuel Macron that was widely commented by traditional news media) or Twitter movements that criticize traditional media (for example, Twitter users reacted to the shocking images of victims of the Nice attack with the hashtag #CSACoupezTout, which led the *France Télévisions* group to apologize[6]). We aim at developing measurement tools and analysis criteria to characterize and quantify these interaction patterns.

The originality of this project is its bi-disciplinary nature. On the one hand, it will consist of an analysis in media economics, in order to determine the factors that influence the relative impact of a story on Twitter and in traditional news media. This thesis will have to take into account several types of measures of the media impact, both absolutely (number of tweets, number of articles), but also related to the membership networks of the users emitting the information: a story spreads more widely if it is issued within a majority group [5], and information is more easily relayed if it emanates from the account of a journalist or a politician [6]. On the other hand, it will require advanced computer science research to design novel approaches to Twitter event detection and clustering, using both Natural Language Processing and Image Processing.

1.1 Choice of Twitter

Why choose Twitter over another social network? Twitter is a micro-blogging website where users can post short messages called “tweets”. Tweets are limited to 280 characters, but can also contain pictures or videos. It is difficult to find reliable statistics on the number of tweets emitted every day worldwide. In 2014, Twitter announced the figure of 500 million tweets on average per day[7], but there has been no other official statement since. Several types of interactions are possible on this social network: users can “follow” other users (that is to say, subscribe to their account in order to see all the tweets they publish), they can “retweet” a tweet (re-publish it on their own account), “reply” to it, “quote” it (re-publish the tweet with a comment of their own) or “like” it. Users can refer to other users in their tweets using “mentions” (the user’s name preceded by ”@”), and they can tag specific words as important using “hashtags” (words preceded by “#”). Tweets are publicly visible by default, which is why Twitter is used by many public personalities like politicians or journalists. This is one of the main differences between Twitter and Facebook, where posts can by default only be read by one’s “friends” (usually people that one know in real life).

Twitter is not the most used social network in France. According to the Reuters Intitute, in 2018 it is even the 4th French social network (used by 16% of respondents), behind Facebook (63%), YouTube (51%) and Facebook Messenger (31%). Nevertheless, we chose this network for our analysis of the relationships between social networks and traditional news media for several reasons.

First of all, it is predominantly used for news access. The “News Use Across Social Media Platforms 2018” study

[6] https://www.francetvinfo.fr/economie/medias/france-televisions/edition-speciale-sur-l-attentat-de-nice-france-televisions-pres
1548057.html

[7] https://blog.twitter.com/official/en_us/a/2014/the-2014-yearontwitter.html

by the Pew Research Center[8] finds that 71% of American Twitter users get news on the platform, compared to 68% for Facebook and 38% for YouTube. There is no similar study on French social networks, but the structure of Twitter makes it a privileged tool for sharing news content, independently of the country. Indeed, if favors public statements rather than private messages to family and friends, and encourages the sharing of external content (reference to other pages through URLs) because of the brevity of tweets. Kwak et al. [12] even argue that the structure of Twitter makes it similar to a news media.

Secondly, Twitter has quickly become the preferred network of journalists, who use it both to easily contact sources and to build a connection with their audience [23]. In the sample of journalists studied by McGregor and Molyneux [14], 93% had a Twitter account. A report conducted at the request of the European Commission[9] shows a similar trend in Europe: the interviewed journalists make the distinction between Twitter, mostly used for work, and Facebook, more used in private life.

Finally, Twitter provides a larger access to its data than other social media platforms. Even if the volume of tweets that one can access through Twitter's APIs is limited, the company still provides a free access to a rather large volume of data. Conversely, despite the research effort recently launched by Facebook to protect elections,[10] it is still nearly impossible for researchers outside Facebook to get access to information on users' activity on the platform.

1.2 Choice of France

Working on French tweets and news contents is difficult due to the still little amount of available corpora for Natural Language Processing and tweet analysis compared to English. However, we can rely on the corpus provided by the French Institute for Audiovisual (INA) that contains all content published by French news media online (newspapers, radio, televisions and online pureplayers) with their precise publication date [2]. Besides, we have access to the universe of the AFP (Agence France Presse) news agency's dispatches, which gives us a proxy for the news stories that make it to traditional media outlets – with similarly the exact time of each dispatch.

Besides, the main empirical challenge for researchers using Twitter data comes from the fact that, because of the limits of the Twitter streaming API, it is impossible for researchers to capture the universe of tweets that are posted on the platform during a given period of time. Perhaps paradoxically, the advantage of France comes here from the fact that there is less data than for example for the United States: according to our estimates, around 9.2 million tweets are published every day in French. Using the collection methods detailed in chapter **??** section 2.3.2, we are able to capture a little bit more than 4.5 million tweets a day, which means that our dataset covers nearly half of all the tweets published in French.

[8] http://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/
[9] http://ec.europa.eu/commfrontoffice/publicopinion/archives/quali/journsm_en.pdf
[10] https://www.facebook.com/notes/mark-zuckerberg/preparing-for-elections/10156300047606634/

1.3 Characterization of events

What is a media event? One possible definition is: a fact that is important enough to be reported in the media. In contemporary societies, news media have therefore a role to play in the definition of events. According to the historian Pierre Nora, the emergence of the mass media has transformed the nature of events: *"Press, radio, images are not only means from which events would be relatively independent, they are the very condition of their existence."*[11][18]. The sociologist Patrick Champagne [3] shares the same view (*"The media build the events they report"*[12]) but highlights the fact that creating an event is a collective process : one media outlet alone cannot make the news if it is not picked up by others. A media event has thus to be reported by several sources to be defined as such.

With the appearance of social media, a new dimension has emerged: traditional news media cannot ignore a topic that is really bursting on Facebook or Twitter. In practice, many news events start nowadays on social media, like the #MeToo movement. Social events and media events tend to be the same in many cases, that is what we call *joint events*. However, any conversation or trend on social media cannot be considered as a *media event*. We therefore propose a distinction between *media events, social events* and *joint events* In this part, we provide definitions that formalize these intuitions.

1.3.1 Definitions

Let f be a **triggering fact**, i.e. a fact that causes an important **activity** in either the traditional media sphere or the social media sphere or the both. This fact can happen in real life or on the internet. Panagiotou et al. [19] discuss the relevance of a distinction between *real world events* and *virtual events* such as "memes, trends or popular discussions". These distinctions are unnecessary for our applications: for example, many politicians use social media to make public statements, that are not less "real" than the ones made during interviews or press conferences. A triggering fact can thus be a tweet, a Youtube video, a speech, a sport performance, a trial, a Facebook post, etc.

In our work, we consider that a **source** s can be a media outlet or a Twitter account. A **source** can publish one or several **content objects**.

A **content object** o is a tuple (x_o, s_o, d_o, f) with x_o the object in itself (i.e. a tweet, a news article,...), s_o its source, d_o its publication date and f the related triggering fact.

Two kinds of content object will be considered in this thesis :

- **Twitter object** $t = (x_t, s_t, d_t, f)$ in which x_t is the tweet (text and /or visual contents) and s_t is a Twitter account and d_t the publication date of the tweet.

[11] "Presse, radio, images, n'agissent pas seulement comme des moyens dont les événements seraient relativement indépendants, mais comme la condition même de leur existence."
[12] "les médias construisent les événements dont ils rendent compte"

- **Media object** $m = (x_m, s_m, d_m, f)$ in which x_m is the media content (press article, video, television or radio report...), s_m is a media outlet, and d_m the publication date of the object.

Using these different concept, we can now defined an **event** as a set of content objects from several sources that are related to the same fact f.

More precisely, a **Twitter event** E_T is a set of Twitter objects $\{t_1, ..., t_n\}$ discussing the same fact f and generated by a least k different sources (i.e. Twitter accounts) in a restricted time interval $[d_{start}, d_{end}]$. For the sake of clarity, we introduce the set $X_T = \{x_t^1, ..., x_t^n\}$, $S_T = \{s_t^1, ..., s_t^n\}$ and $D_T = [d_{t,start}, d_{t,end}]$ that are respectively the set of tweets, the set of sources and the time interval of E_T. E_T is a Twitter event if $|S_T| \geq k$ and if $D_T \subseteq [d_{start}, d_{end}]$.

Similarly, a **Media event** E_M is a set of media objects $\{m_1, ..., m_n\}$ discussing the same fact f and generated by a least l different sources (i.e. media outlets) in a restricted time interval $[d_{start}, d_{end}]$. We introduce the set $X_M = \{x_m^1, ..., x_m^n\}$, $S_M = \{s_m^1, ..., s_m^n\}$ and $D_M = [d_{m,start}, d_{m,end}]$ that are respectively the set of tweets, the set of sources and the time interval of E_M. E_M is a media event if $|S_M| \geq l$ and if $D_M \subseteq [d_{start}, d_{end}]$.

1.4 Detailed Outline

Chapter 2

Building a corpus for event detection on Twitter

2.1 Introduction: properties of a Twitter event detection corpus

2.1.1 Representativity of the collected data

In an ideal world, to compare news production on social media and on mainstream media, one would need the universe – during a given period of time (e.g. the year 2017) and a geographical location (e.g. France, the UK or the US) – of all documents published on the one hand on social media and on the other hand on mainstream media. Unfortunately, given the limitation of the Twitter API, it is not possible for the researcher to capture the universe of the documents (or tweets) published on Twitter. However, the researcher can work on a sample of the documents, as long as this sample is "representative". Why do we need representativity?

Assume that we get access to a subsample of the documents published on Twitter, but that this sample is not representative of the overall traffic. For example, assume that this sample of tweets is such that the tweets' characteristics (perhaps because the API provides the researcher with documents tweeted by users with more followers) are such that these documents have a higher probability to make it to the mainstream media. Then using this biased subsample will lead the researcher to overestimate the probability for a news story broken on social media to appear on mainstream media.

The same issue will arise if the researcher wants to tackle the follow-up question: what are the determinants of the success of a news story initially broken on social media? Imagine that the researcher is using a selected sample of tweets that is not representative. Imagine for example that this sample of tweets comes mainly from journalists working for a given media, e.g. *Le Monde*, and that, at the same time, within the set of tweets posted by

Le Monde's journalists, only the successful ones are part of the sample, then the results of the empirical analysis will be biased in favor of *Le Monde*. In other words, when the researcher will study the impact of the company for which the journalist works (independent variable) on the probability for the news story broken on Twitter to make it to mainstream media (dependent variable), the coefficient obtained for *Le Monde* will ovestimate the real causal impact of the company.

It seems very difficult to correct for this kind of biases. Hence the necessity to have a representative sample of tweets, i.e. a sample of tweets such that the tweets included in our sample do not differ from the tweets that are not included along all the dimensions that may have a direct impact on the dependent variable of interest.

2.1.2 Quality of the annotated events

The properties of a given corpus have an impact on the implementation of the detection algorithm: for example, if all events in our corpus tend to grow at a high rate (*i.e.* people react very quickly to that event on social media), a simple way to increase the performance of our program would be to select group of tweets that have a high growth rate and discard others as "non events". However, this would result in a program unable to detect other types of events and introduce bias in our results. Therefore, the choices made during the creation of the annotated corpus are critical to ensure that our program can detect a large variety of events. We propose here a number of desired features that would help reduce the bias of an event detection corpus.

Event's size: Some events generate a high number of documents, but most of them do not. McMinn et al. [15] reduce the number of events to annotate by cutting out clusters with less than 30 tweets. In our definition, there is no limit in the number of tweets to characterize an event. Therefore, it is important to keep "small" events in the corpus.

Entropy and user diversity: Petrović et al. [20] define entropy as:

$$Entropy = \sum_i \frac{n_i}{N} log \frac{n_i}{N}$$

where n_i is the number of times word i appears in a cluster of tweets and N is the total number of words in that cluster.

Kumar et al. [11] define user diversity as:

$$UserDiversity = \sum_i \frac{u_i}{T} log \frac{u_i}{T}$$

where u_i is the number of tweets published by account i in a cluster and T is the total number of tweets in that cluster. Both high entropy and high user diversity denote that a given topic has reached a large public. The higher the entropy, the more difficult it is to automatically detect a topic with natural language processing methods, since no unified vocabulary is used to debate the topic. Therefore, a good event detection corpus should contain at least some events with high entropy. In this respect, one of the events selection method by McMinn et al. [15], that consists in using the descriptions from Wikipedia events as queries to retrieve related tweets, is likely to produce clusters with low entropy.

Categories McMinn et al. [15] test the diversity of their corpus by examining the events' distribution across 8 categories. These categories where created from a mapping between TDT categories [1] and Wikipedia categories. They identify differences in the distribution depending on the event selection method : for example, approaches based on pre-detection produce more clusters about "Sport" and less about "Armed Conflicts and Attacks" than the Wikipedia-based approach. It is likely that sports events are more discussed on Twitter than by traditional news media which would explain why they were not selected as "events" by the Wikipedia community.

2.2 State of the art: building event detection corpora

2.3 Tweet collection

Our objective is to collect automatically and continuously a set of tweets representative to the real Twitter activity. In addition to being **representative**, this corpus must contain a **volume of tweets sufficient for media events to be represented** in it, in particular medium and small events. Besides, these tweets must be **in French**. The following sections present the methods used to achieve these objectives.

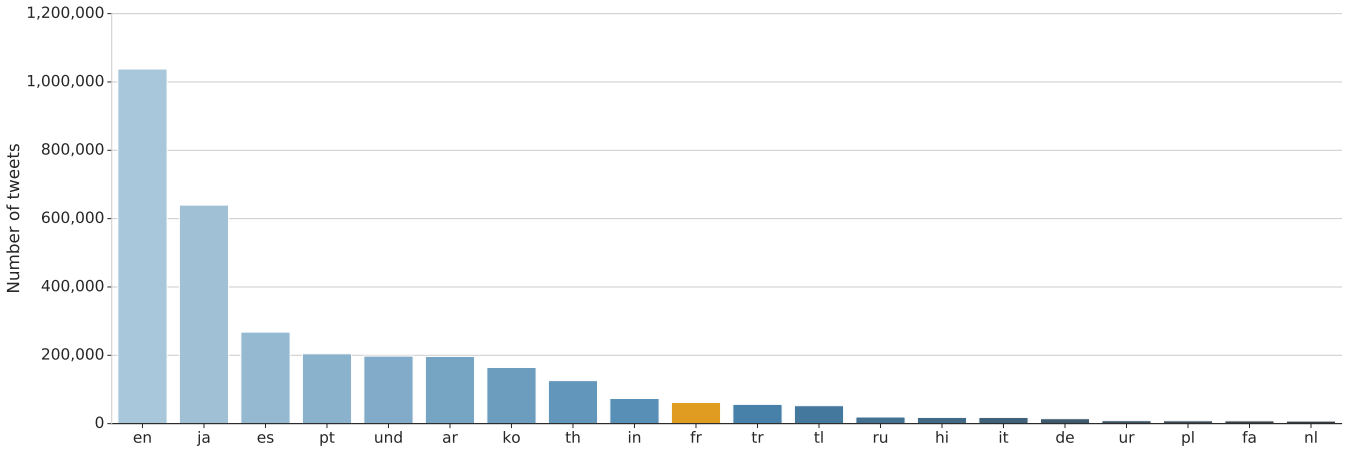
2.3.1 Constraints

Their are different ways of collecting large volumes of tweets, although collecting the full volume of tweets emitted during a given period is not possible. Indeed, even if Twitter is known for providing a larger access to its data than other social media platforms,[1] the Twitter streaming APIs are strictly limited in term of volume of returned tweets. These limitations are all constraints that we must integrate into our collection procedure.

Sample API: the Sample API[2] continuously provides 1% of the tweets posted around the world at a given moment. Once connected to the API at time t_0, the user gets 1% of all tweets emitted after t_0, and receives regular batches of

[1] In particular, despite the research effort recently launched by Facebook, it is still nearly impossible for researchers outside Facebook to get access to information on users' activity on the platform.

[2] `https://developer.twitter.com/en/docs/tweets/sample-realtime/overview/GET_statuse_sample.html`



Notes: This figure plots the average number of tweets collected per day using the Sample API in the 20 most frequent languages on Twitter. The language metadata is provided by Twitter. "und" stands for "undefined" language.

Figure 2.1: Average number of tweets in each language collected using the Sample API during one day

tweets as long as she stays connected. Twitter provides little information on how the sample is generated. However, Kergl et al. [9] have studied it by analyzing the ids of tweets (based on a timestamp in milliseconds and on a number of series to identify tweets issued during the same millisecond). They show that all tweets provided by the API were issued between the 657th and 666th milliseconds of each second, which should assure the user to receive a constant stream representing around 1% of the total. Another study done on the distribution of tweets in the Sample API in comparison with another paid API, which provides full access to all emitted tweets, shows no statistically significant difference between the two samples [16].

This API does not meet our needs, since the proportion of tweets in French is only 1.8% of the total sample on average (Figure 2.1 illustrate the distribution of tweets in different languages). Moreover, according to Liu et al. [13], the proportion of tweets concerning news is less than 0.2%. If we combine all these restrictions, we could only have access to 92,000 tweets in French a day, and less than 200 tweets a day concerning news if we were to simply use the Sample API provided by Twitter.

Filter API: the Filter API[3] continuously provides tweets corresponding to the input parameters (keywords, account identifiers, geographical area). The language of the returned tweets can be selected. Again, the API provides only about 1% of the total flow. However, this is sometimes enough to collect all the tweets containing a relatively little used keyword, and this is naturally enough to collect all the tweets from a given account. For a quite little represented language such as French, this API could theoretically provide us up to 55% ($\frac{1\%}{1.8\%} = 55\%$) of the tweets emitted in French. Given this observation, we worked at identifying the keywords that maximize the number of returned French tweets.

[3] https://developer.twitter.com/en/docs/tweets/filter-realtime/api-reference/post-statuses-filter.html

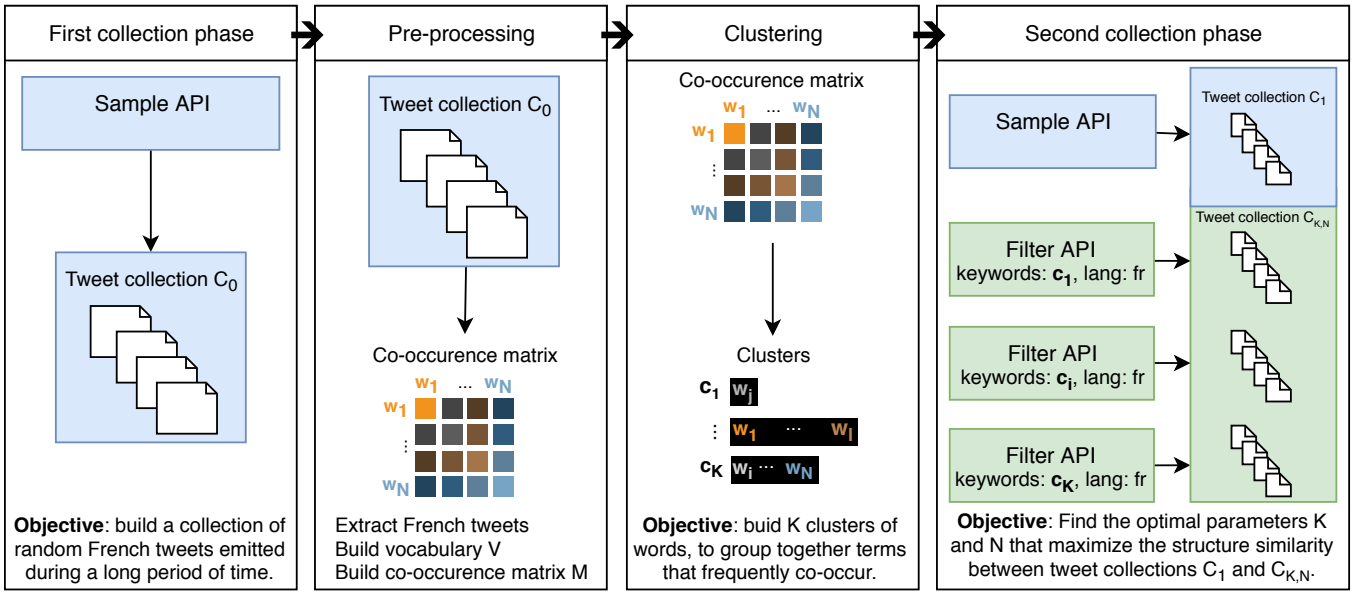


Figure 2.2: Diagram of our experimental setup to select the best tweet collection method

Joseph et al. [8] compare five samples collected through the Filter API with the same input keywords at the same time, using five different connection tokens[4]: they find that two connections to the Filter API at the same time with the same keywords as inputs are "nearly identical". It is hence not useful to try to get more tweets using a second access token with the same keywords. However, spreading different keywords over several API connections should return a higher number of tweets.

2.3.2 Proposed collection strategy

Given the constraints of the APIs, we decided to collect tweets by using the random stream proposed by the Sample API, but to increase the volume of collected data by using the Filter API as well, with "neutral" terms as keywords parameters, and "French" as language parameter. In order to further increase the volume, we decided to use multiple tokens to connect to the Filter API.

The choice of the keywords parameters was done to optimize two metrics: the number of collected tweets, and their representativity of the real Twitter activity. The selected terms had thus to be the most frequently written on Twitter, and we had to use different terms (and terms that do not co-occur in the same tweets) as parameters for each connection. This way, the multiple connections would return sets of tweets with little intersection, and thus a greater total volume.

The precise strategy was the following (it is schematized in figure 2.2): given a set of tweets $C_0 = \{t_1, \ldots, t_k\}$ collected using the Sample API during a time-interval $I = [d_{t,start}, d_{t,end}]$ we select tweets in French, creating a subset $C_{0French}$ and extract from them a vocabulary $V = \{w_j, \forall j \in [1, \ldots, M]\}$ of all unique words appearing

[4]To use the Twitter API, a connection token is required. Twitter limits the access to its data by generating only one connection token per Twitter account.

in $C_{0French}$. We extract a subset of the N words of V having the highest document-frequency. We build a co-occurrence matrix $\mathcal{M} = (m_{i,j}) \in \mathbb{N}^{N \times N}$ where $(m_{i,j})$ is the number of times w_i and w_j co-occur in the same tweet of $C_{0French}$. Using a clustering algorithm with $\mathcal{M}$ as adjacency matrix, we extract K clusters of terms. The K obtained clusters of words are then used as parameters of K different connections to the Filter API. By doing so, we aim at separating terms that are not frequently used together and thus to collect sets of tweets with the smallest possible intersection.

Section 2.3.4 presents the methods used to evaluate the similarity between $C_{K,N}$ the set of tweets collected using N keywords spread on K Filter API connections, and $C_1French$ the set of French tweets collected with the Sample API during the same period as $C_{K,N}$.

2.3.3 Experimental setup

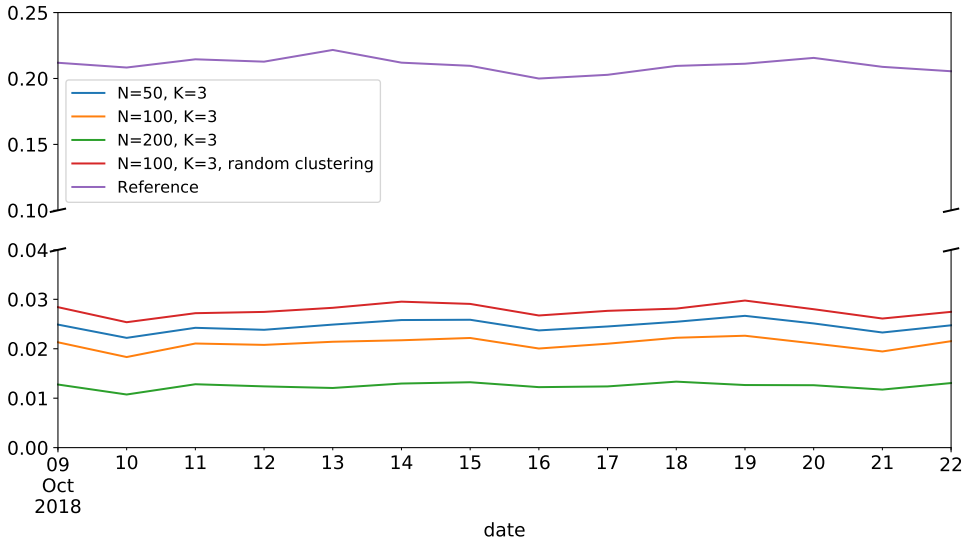
In practice, we collected our corpus C_0 of sample tweets between $d_{t,start}$ = 2018/01/15 and $d_{t,end}$ = 2018/02/15. The text of the collected tweets was tokenized on white spaces and on punctuation characters ("qu'il" was considered as two words, "qu" and "il"). The resulting vocabulary V was lowercased and accents were removed, since we noticed that accents and capital letters are not taken into account by the Twitter API. For example, it returns tweets containing both "à" and "a" if the parameter "a" is given as input. No other pre-processing such as stemming was applied (the vocabulary contains both "mdr" and "mdrrr"[5], for example), since the objective here is not to query the API with semantically different terms, but with the most frequently used terms.

We ran tests with $N \in \{50, 100, 200\}$ and $K \in \{2, 3\}$. This choice is motivated by our storage and CPU capacity. The clustering algorithm chosen is spectral clustering.[6] The clusters of terms for the different N and K values are presented in appendix **??**. The resulting clusters are imbalanced in size: some contain only a few words, others contain all remaining terms. In order to control the effect of imbalanced clusters, we also tested to distribute terms randomly in clusters of size $\frac{N}{K}$. The samples collected using random clustering are denoted $R_{K,N}$. We ran each test for a period of two weeks, during which we also collected tweets from the Sample API and extracted tweets in French. This last set of tweets is denoted $C_{1French}$.

2.3.4 Evaluation of the collection strategy

We compared the sets $C_{K,N}$ and $R_{K,N}$ collected with each collection method with the set $C_{1French}$ collected with the Sample API during the same time interval to assess the representativity of each test dataset. This approach is comforted by the study of Morstatter et al. [16], who have had access to the entire stream of tweets and compared it with the Sample API. They find that the tweets from the Twitter Sample API are "a representative sample of the true activity on Twitter".

[5] French abbreviations similar to "lol" and "loool". "mdr" stands for "mort de rire".
[6] We used the implementation from python module scikit-learn



Notes: This figure plots the daily evolution of the KL-divergence between the word distribution in collection $C_1 French$ and the distribution obtained using 4 collection methods. The "Reference" is obtained by splitting the collection $C_1 French$ in two sets and computing the KL-divergence between them. A KL-divergence of 0 indicates a perfect similarity between 2 distributions.

Figure 2.3: Daily evolution of the divergence between collection $C_1 French$ and the collections $C_{K,N}$ with $K = 3$

Several comparison methods can be used in order to assess the similarity between two collections of texts. A first approach consists in considering the number of times each word is used in each collection as a probability distribution, and to measure the difference between those distributions. We used Kullback-Leibler divergence [10] as comparison metric. For each $(N, K) \in \{50, 100, 200\} \times \{2, 3\}$, we computed the KL-divergence between the word distribution in $C_{K,N}$ and in $C_{1French}$ for the same collection period. In order to have a reference of what level of divergence can be accepted, we also split the corpus $C_{1French}$ in two sets (depending on whether the tweets had an even or odd id) and computed the KL-divergence between those sets. Figure 2.3 presents the results for $K = 3$. Overall, we found that the collection $C_{3,200}$ was the most similar to $C_{1French}$ using KL-divergence as comparison metric.

This first way of evaluating the collection methods considers only the text of the tweets, without taking their structure into account: a tweet has an author, it is retweeted or not, it contains hashtags, urls, etc. In order to take the tweet samples' structure into account, we use Student's t-tests to determine whether there is a significant mean difference between the two collections along several variables :

- number of characters per tweet
- share of retweets (i.e. proportion of collected tweets that are actually retweets)
- share of quotes (i.e. proportion of collected tweets that quote another tweet)
- share of replies (i.e. proportion of collected tweets that reply to another tweet),

	$C_{1French}$	$C_{3,200} \cup C_{1French}$	$C_{3,100} \cup C_{1French}$	$C_{3,50} \cup C_{1French}$	$R_{3,100} \cup C_{1French}$
Number of characters	**116**	7***	8***	10***	9***
		(0)	(0)	(0)	(0)
Share of retweets	**0.61**	0.02***	0.02***	0.02***	0.02***
		(0.00)	(0.00)	(0.00)	(0.00)
Share of quotes	**0.18**	0.01***	0.01***	0.02***	0.01***
		(0.00)	(0.00)	(0.00)	(0.00)
Share of replies	**0.19**	-0.02***	-0.01***	-0.01***	-0.02***
		(0.00)	(0.00)	(0.00)	(0.00)
Number of URLs	**0.24**	0.01***	0.01***	0.01***	0.01***
		(0.00)	(0.00)	(0.00)	(0.00)
Number of hashtags	**0.24**	-0.01***	-0.05***	-0.06***	-0.05***
		(0.00)	(0.00)	(0.00)	(0.00)
Share of verified users	**0.01**	-0.00*	-0.00	-0.00**	0.00***
		(0.00)	(0.00)	(0.00)	(0.00)
Number of followers	**3,073**	-68	-72	-165*	59
		(74)	(74)	(72)	(76)
Number of friends	**692**	-9*	-34***	-44***	-30***
		(4)	(4)	(3)	(4)
Number of lists	**38**	0	0	-0	2***
		(0)	(0)	(0)	(0)
Observations	**875,630**	63,296,610	59,118,730	64,272,167	55,405,602

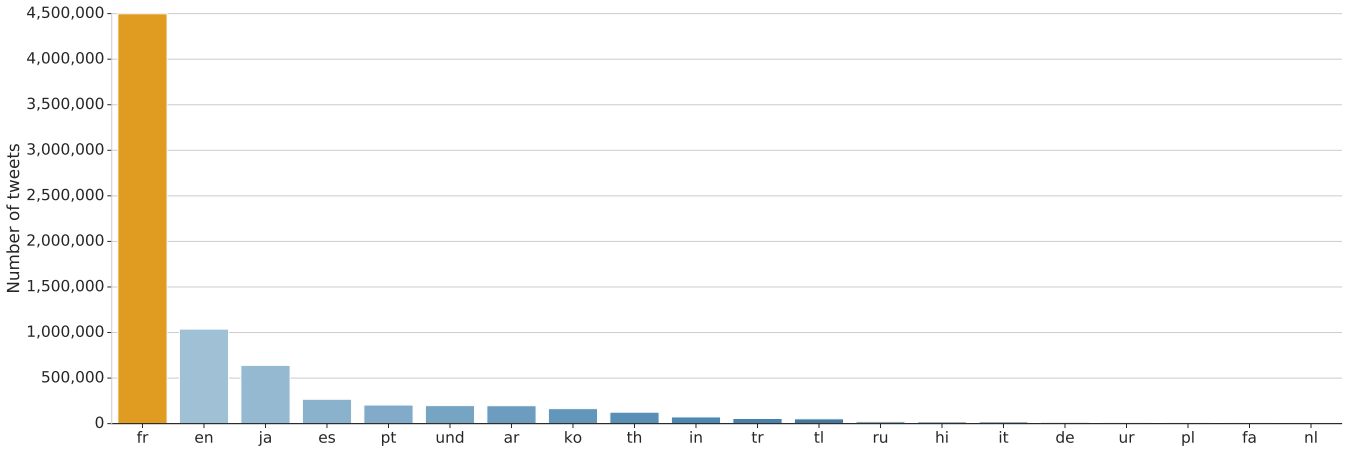
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$. Standard errors in parentheses.
Notes: The table presents summary statistics for the collected samples using Student's t-tests for the equality of means. The first column (in bold) presents the mean of the tested variables in the reference sample $C_{1French}$. The next columns show the mean differences between that reference sample and the samples collected with each method.

Table 2.1: Mean difference between each collection $C_{K,N}$, $K = 3$ and collection $C_{1French}$

.

- number of URLs per tweet
- number of hashtags per tweet
- share of "verified users" (i.e. proportion of collected tweets from a user whose identity has been verified by Twitter)
- number of "followers" of the tweet's author (i.e. accounts that subscribed to receive the tweets of this author)
- number of "friends" of the tweet's author (i.e. the accounts to which the author subscribed)
- number of public "lists" that the tweet's author is member of (any twitter account can create a public list of accounts, for example "list of famous pianists")

Table 2.1 summarize Student's t-tests for the samples collected with 3 clusters (K=3) compared to the random sample $C_{1French}$. All collection methods have a statistically significant mean difference to the random sample on all tested variables (except the number of followers, where there is no significant mean difference between the collection methods). However, the difference is small, particularly for the best collection method ($C_{3,200} \cup C_{1French}$): +0.01 for the number of URLs and -0.01 for the number of hashtags) which indicates a very similar structure of



Notes: This figure plots the average number of tweets collected per day using our best collection method in the 20 most frequent languages on Twitter. The language metadata is provided by Twitter. ”und” stands for ”undefined” language.

Figure 2.4: Average number of tweets in each language collected using the Sample API combined to our best collection method during one day

the conversations. The number of characters tends to be significantly higher in our collection methods than in the random sample (+7 characters for $C_{3,200} \cup C_{1French}$), which can be explained by the fact that our collection methods return tweets containing the keywords given as input parameters, excluding *de facto* all tweets that contain no word.

We decided to keep $C_{3,200} \cup C_{1French}$ as main collection method, since its similarity to the random sample was the highest with the two comparisons methods. Figure 2.4 illustrates the new distribution of language using that method combined to the Sample API.

2.4 Tweet annotation

We built our event detection corpus based on tweets collected with the best collection method (200 words spread on 3 clusters used as keywords for 3 connections to the Filter API, combined to one connection to the Sample API). We annotated these tweets depending on their relation to Twitter events and media events that took place in France at the time of the annotation.

2.4.1 Media events selection

To select media events, we decided to draw events randomly from the hundreds of events described in French press every day. We drew press articles for every day from July 15 to August 6, 2018, for a total of 23 days. We did not want to use any automatic detection method to generate events from the collected tweets, since it may bias the results of our evaluation tests (detection methods similar to the one used to generate events in the test set may be advantaged). We did not either use Wikipedia to select important events (like it is the case in McMinn et al. [15] and Petrović et al. [21]), considering that “an event detection system should also be able to detect newsworthy events at

a smaller scale" [7].

In practice, we drew 30 events a day, two thirds from the Agence France Presse (AFP), which is the third largest news agency in the world, and one third from a pool of major French newspapers (*Le Monde*, *Le Figaro*, *Les Échos*, *Libération*, *L'Humanité*, *Médiapart*). This selection method has the advantage of giving "big" events a higher chance of being selected, since they are covered by all news outlets, while also letting relatively "small" events emerge. Duplicates, that is to say articles covering the same event, were manually removed. We processed sub-events as separate events. For example on the 16th of July, after France's FIFA World Cup win, we drew articles about the press conference of the coach, the celebrations in the Paris metro, and violent riots during the celebrations. All three articles were considered as describing separate events, even though they were all linked to a larger "World Cup" event.

2.4.2 Twitter events selection

Since our final objective is to measure differences in the coverage of events by news media and by Twitter users, we did not want to miss important events in the Twitter sphere that would be little covered by traditional news media. We therefore monitored the trending terms on Twitter by detecting unusually frequent terms every day.

We chose a metric called JLH, that is used by Elasticsearch[7] to identify "significant terms" in a subset of documents (*foreground set*) compared to the rest of the collection (*background set*). It simply compares for each term t the frequency of appearance in the foreground set (p_{fore}) and in the background set (p_{back}). This metric is computed as:

$$f(t,d) = \begin{cases} (p_{fore}(t,d) - p_{back}(t))\frac{p_{fore}(t,d)}{p_{back}(t)} & if\ p_{fore}(t,d) - p_{back}(t) > 0 \\ 0 & elsewhere \end{cases}$$

Where $p_{fore}(t,d) = \frac{tf(t,d)}{u_d}$ (the number of different users mentioning term t on day d divided by u_d the total number of users tweeting on day d) and $p_{back}(t) = \frac{tf(t)}{U}$ (the number of different users mentioning term t in the total collection divided by U the total number of users, measured by the number of different authors of tweets in our collection). We did not use a standard tf-idf metric because it resulted poorly at identifying bursting terms for a given a day.

We computed the 20 terms having the best JLH scoring every day and went on Twitter to discover the underlying events causing a burst of these terms. We were then able to group together terms related to the same event. For example the terms "afcbom", "bournemouth", "bouom", "afcbournemouth", all related to a soccer match between the Association Football Club Bournemouth (AFCB) and the Olympique de Marseille (OM), were grouped together. We then excluded events:

- that were artificially amplified using automatic tools. In particular, the Q&A website Curious Cat was used to

[7] https://www.elastic.co/guide/en/elasticsearch/reference/current/search-aggregations-bucket-significantterms-aggregation.html#_jlh_score

post the same questions (“Where do you see yourself in tens years from now?”, “What is your favorite movie?”) to all Twitter users registered on Curious Cat. Many of them responded using the terms of the question (“My favorite movie is...”) causing a burst in the frequency of those terms.

- that had been already drawn from the media events selection process.

Once the media events and the Twitter events selected, the annotators' work could begin. In the following section, we explain the annotation procedure.

2.4.3 Annotation procedure

User interface

We developed a user interface presenting each event in the form of a title and a description text. For media events we used the title and the first paragraphs of the drawn corresponding press article. For Twitter events the title was constituted of the bursting terms detected with the JLH scoring and the description was a tweet manually selected because it described the event clearly. Under the title and the description, a search bar was presented. The user could use that bar to enter keywords and find the collected tweets containing those exact keywords. Twelve tweets per page were displayed, starting with the most retweeted. The user could select or unselect the tweets he considered related to the event. If the tweet contained an URL, the user could click or unclick a button under the tweet to indicate if the linked page was related to the event as well. Once the user had read all twelve tweets and selected the ones related to the event, the user could submit its answers and access to the next twelve tweets. Displayed tweets were not pre-selected by our program depending on their content. We only excluded retweets since it would only have displayed the same tweet several times, and displayed tweets emitted on the day of the event.

The interface also allowed to review tweets annotated by other users: once an annotator was finished with an event, she could access another page displaying the same event (same title, same description) and the tweets seen by other annotators in relation to the event. The user had to go through all these tweets and annotate them, without knowing if those tweets were marked as relevant or irrelevant to the event by the other users.

Annotation task

Three political science students were hired for a month to annotate the corpus. All three of them were Twitter users and had a good knowledge of media news. Every day they were presented the new list of events. They started on the 16th of July, 2018, to annotate events from the 15th of July. This first day of annotation was not included in the final dataset and served as a day of adaptation. Since the annotators did not work on Saturday or Sunday, every

day between July 15th and August 15th could not be annotated. We made the choice to annotate on a continuous period of time, from the 16th of July to the 6th of August.

For every event, they were asked to search for related tweets on the user interface, using a large variety of keywords. It was insisted on the importance of named entities (persons, locations, organizations) and on the specificity of Twitter (one person can be referred to using her real name or her Twitter user name, for example). Like McMinn et al. [15], we asked the annotators to mark tweets as related to the event if it referred to it, even in an implicit way. It appeared that annotators could not treat more than 20 events a day, and often no more than 10 events, depending on the volume of tweets generated by each event. Some major events would even have required days of work to be fully treated. We therefore instructed not to spend more than an hour on a subject. This has of course an impact on the maximum number of tweets per event that could be annotated.

In order to make the annotators work on the same tweets, we stopped the first annotation task after four hours of work every day, and asked them to go to the second part of the user interface, were they could find tweets already seen by at least one of the other annotators. They then had to annotate those tweets without knowing the judgment made by the others. This way, we could make sure that all tweets would be reviewed by all three students.

Annotation propagation

Even if the total number of tweets the annotators could deal with during one day was rather small, we designed some heuristics to increase the total number of annotated tweets: given one tweet t annotated as related or not related to one event, we extended the annotation to all tweets published on the same date that belonged to one of these categories:

- retweets of t
- quotes of t
- replies to t
- tweets containing the same URL as t, if the URL was marked as relevant to the event by the annotator
- tweets with the exact same text as t, if the text was longer than five words.

2.5 Evaluation of the created corpus

In this section, We first present the annotator agreement and discuss possible reasons for differences in agreement. We then describe the characteristics of the corpus, including the number of events, their distribution across different categories, and the number of tweets per event.

2.5.1 Annotator agreement

Annotator agreement is usually measured using Cohen's kappa for two annotators. Here we chose to hire three annotators in order to have an odd number of relevance judgments for each tweet. In the case of several annotators, Randolph [22] recommend to use Fleiss' kappa [4] in case of "*fixed marginal distributions*" (annotators know in advance the proportion of cases that should be distributed in each category) and free-marginal multirater kappa [22] if there is no prior knowledge of the marginal distribution. Indeed, we experienced some odd results using Fleiss' kappa on our corpus, in particular for events with a strong asymmetry between categories (when a large majority of tweets were annotated as "unrelated" to the event of interest, or the opposite). We hence decided to use free-marginal multirater kappa, which is also the measure used by McMinn et al. [15].

If one denote P_o the proportion of overall agreement among annotators, P_e the proportion of agreement expected by chance, free-marginal multirater kappa is expressed as

$$\kappa_{free} = \frac{P_o - P_e}{1 - P_e}$$

with

$$P_o = \frac{1}{Nn(n-1)}((\sum_{i=1}^{N}\sum_{i=j}^{k} n_{ij}^2) - Nn)$$

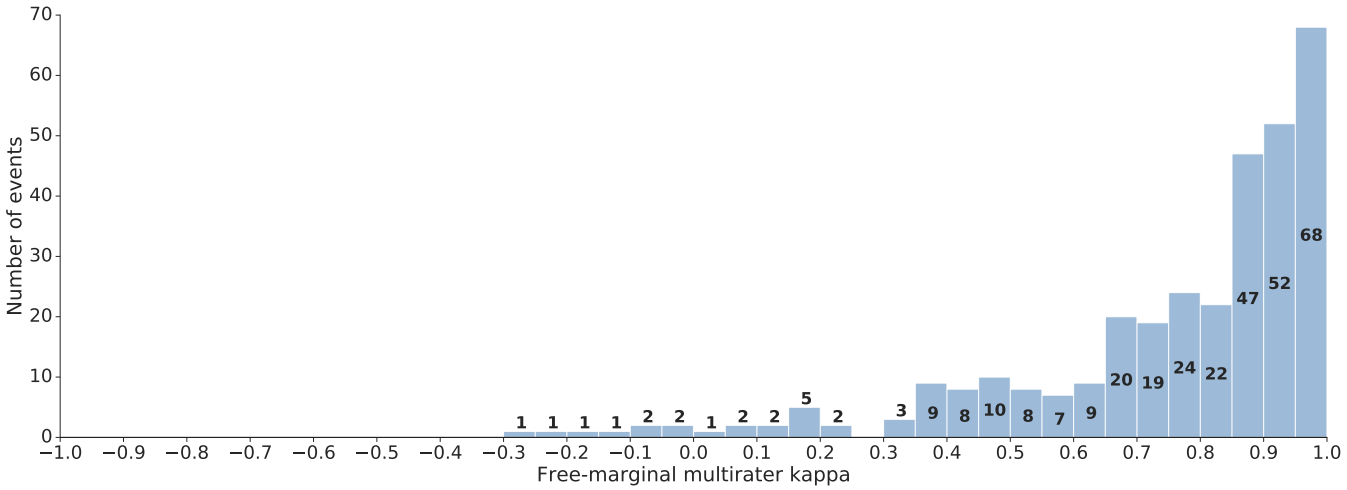
and

$$P_e = \frac{1}{k}$$

where N is the total number of cases (here the number of tweets to annotate), n is the number of annotators and k the number of categories (two in our case: *relevant* or *not relevant* to a given event). Like all kappa scores, its values vary between 0 and 1. A value of 0 indicates a level of agreement that could have been expected by chance, and a positive value indicates a level of agreement that is better than chance.

In our corpus, $\kappa_{free} = 0.79$ which indicates a strong level of agreement among annotators. We also computed the κ_{free} value for each individual event (taking into account all tweets that have been read by annotators while working on this event). Figure 2.5 describes the distribution of events depending on the κ_{free}. We observe that for some events, the agreement is very low: 8 events have a negative κ_{free} value, and 12 events have a κ_{free} value between 0 and 0.3.

We asked the annotators to re-read together the events were their agreement was particularly low, in order to understand why they did not annotate tweets the same way. The students admitted some errors in the annotation for 4 of the 17 examined events. For the other events, they explained that they had different views of the events' scope: for example, one article reported the fact that President of Nicaragua Daniel Ortega refused to resign in a context of severe crisis in his country. Two of the annotators included in the event the tweets related to the crisis in Nicaragua. One annotator restricted the event to the statement of Daniel Ortega. Faced with these differences in



Lecture Note: In our corpus, 68 events have a free-marginal multirater kappa higher than 0.95, 52 events have a free-marginal multirater kappa higher than 0.9, etc.

Figure 2.5: Distribution of the events depending on annotators' agreement, measured by free-marginal multirater kappa

opinion we could decide to remove from the corpus the tweets were the annotators disagree, or to remove events with a very low kappa. However it seems interesting to see how an algorithm behaves in such borderline cases.

2.5.2 Corpus characteristics

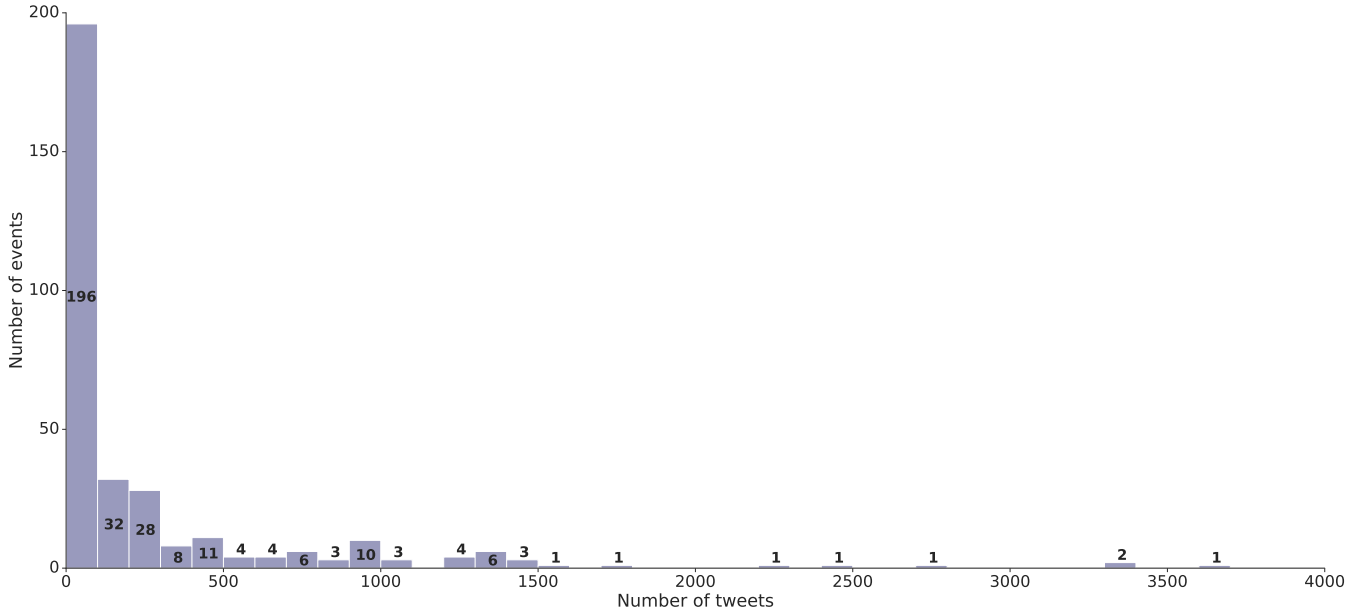
326 events were annotated, including 30 "Twitter events" (detected using the term frequency on Twitter). A total of 152 978 tweets were directly annotated, and 2 904 634 tweets were annotated using the annotation propagation described in 2.4.3.

Of course, even if annotators labeled a large number of tweets during the annotation procedure, many of them were annotated as "not related" to the event of interest. Nevertheless, we consider these annotations to be useful information for the purpose of evaluating an event detection algorithm: since these tweets were retrieved by the annotators because they contain terms associated to the event, they allow us to test our algorithm on "difficult" cases of tweets containing terms related to the event but not directly linked to it.

Figures 2.6 and 2.7 show the distribution of events depending on the number of associated tweets for both direct annotation and propagated annotation. For direct annotation (figure 2.6), despite a quite high number of annotated tweets in average (469 annotated tweets per event in average), if we consider only tweets annotated as "relevant" to the event, we get a vast majority of small events with less than 100 tweets. For one event, concerning a reform of the Polish electoral system, the annotators could not find any related tweet.

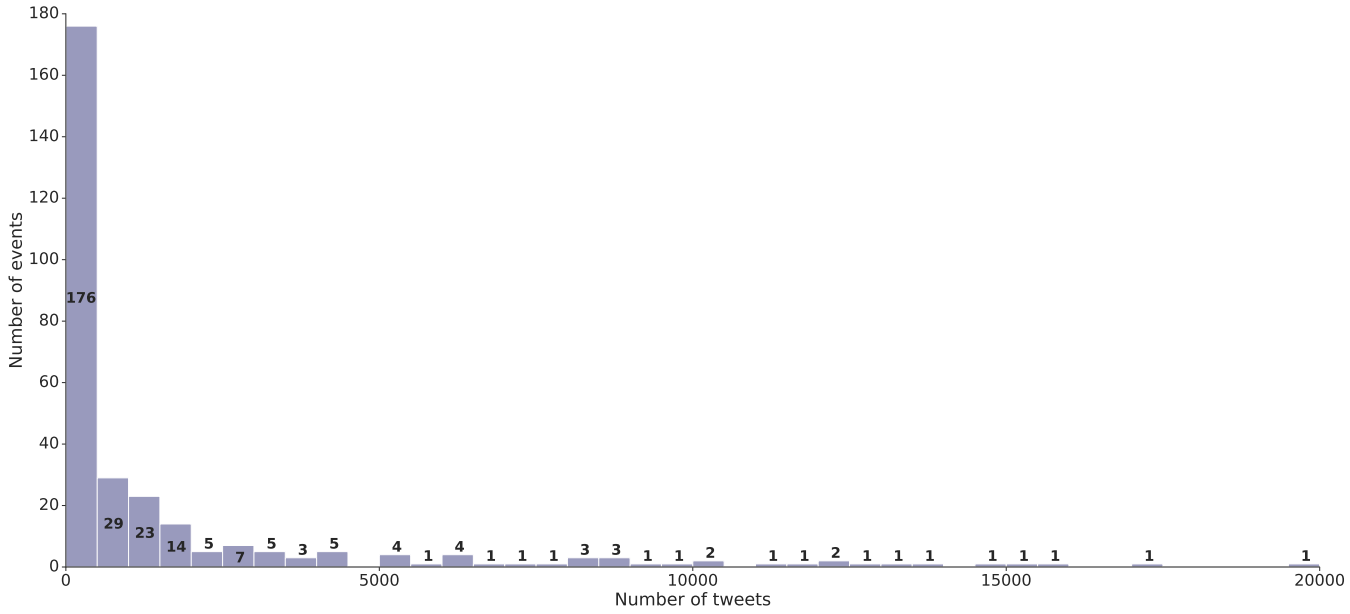
To describe the distribution of events across categories we used the classification by the French news agency AFP. AFP dispatches are labeled using the IPTC Information Interchange Model[8] Media Topics. This taxonomy is

[8] https://en.wikipedia.org/wiki/IPTC_Information_Interchange_Model



Lecture Note: 196 events of our corpus contain less than 100 directly annotated tweets. 32 events of our corpus contain less than 200 directly annotated tweets, etc.

Figure 2.6: Distribution of the events depending on the number of directly annotated tweets



Lecture Note: 176 events of our corpus contain less than 500 tweets annotated with propagation. 29 events of our corpus contain less than 1000 tweets annotated with propagation, etc.

Figure 2.7: Distribution of the events depending on the number of tweets annotated with propagation

English categories	French categories	Number of events
arts, culture, entertainment and media	Arts, culture, divertissement et médias	12
disaster, accident and emergency incident	Désastres et accidents	9
economy, business and finance	Economie et finances	47
education	Education	5
environment	Environnement	0
human interest	Gens animaux insolite	8
conflict, war and peace	Guerres et conflits	24
weather	Météo	0
crime, law and justice	Police et justice	71
politics	Politique	53
religion and belief	Religion et croyance	1
health	Santé	6
science and technology	Science et technologie	4
labour	Social	3
society	Société	21
sport	Sport	54
lifestyle and leisure	Vie quotidienne et loisirs	8

Table 2.2: Distribution of events across the 17 top IPTC Information Interchange Model Media Topics.

used internationally by numerous media companies to apply metadata to text, images and videos. The distribution of events across the 17 top Media Topics is detailed in figure 2.2. Among the 326 selected events, only 209 were drawn from the AFP and had a label. For the remaining 117 events (from other French press outlets or from Twitter events) we attributed a label manually.

Chapter 3

Detecting Twitter events

3.1 Introduction

3.2 State of the art: Twitter event detection

3.3 Algorithms

3.4 Text-only approaches

3.5 Multimodal approaches

Chapter 4

Linking Media events and Twitter events

Chapter 5

Analysis of the spread of news on Twitter and traditional media

Chapter 6

Conclusion

Appendix A

First Appendix

Appendix B

Second Appendix

Bibliography

[1] J. Allan. Introduction to Topic Detection and Tracking. In J. Allan, editor, *Topic Detection and Tracking: Event-based Information Organization*, pages 1–16. Springer US, Boston, MA, 2002.

[2] J. Cagé, N. Hervé, and M.-L. Viaud. The Production of Information in an Online World. Working Papers 15-05, NET Institute, 2015. URL `https://ideas.repec.org/p/net/wpaper/1505.html`.

[3] P. Champagne. L'événement comme enjeu. *Réseaux. Communication-Technologie-Société*, 18(100):403–426, 2000.

[4] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[5] Y. Halberstam and B. Knight. Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. *Journal of Public Economics*, 143:73–88, 2016. ISSN 0047-2727. doi: http://dx.doi.org/10.1016/j.jpubeco.2016.08.011. URL `http://www.sciencedirect.com/science/article/pii/S0047272716301001`.

[6] R. A. Harder, S. Paulussen, and P. Van Aelst. Making sense of twitter buzz: The cross-media construction of news stories in election time. *Digital Journalism*, 4(7):933–943, 2016.

[7] M. Hasan, M. A. Orgun, and R. Schwitter. A survey on real-time event detection from the twitter data stream. *J. Information Science*, 44(4):443–463, 2018.

[8] K. Joseph, P. M. Landwehr, and K. M. Carley. Two 1% s Don't Make a Whole: Comparing Simultaneous Samples from Twitter's Streaming API. In *SBP*, pages 75–83. Springer, 2014.

[9] D. Kergl, R. Roedler, and S. Seeber. On the endogenesis of twitter's spritzer and gardenhose sample streams. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2014, Beijing, China, August 17-20, 2014*, pages 357–364, 2014. doi: 10.1109/ASONAM.2014.6921610. URL `https://doi.org/10.1109/ASONAM.2014.6921610`.

[10] S. Kullback. *Information theory and statistics*. Courier Corporation, 1997.

[11] S. Kumar, H. Liu, S. Mehta, and L. V. Subramaniam. From tweets to events: Exploring a scalable solution for twitter streams. *arXiv preprint arXiv:1405.1392*, 2014.

[12] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. AcM, 2010.

[13] X. Liu, Q. Li, A. Nourbakhsh, R. Fang, M. Thomas, K. Anderson, R. Kociuba, M. Vedder, S. Pomerville, R. Wudali, R. Martin, J. Duprey, A. Vachher, W. Keenan, and S. Shah. Reuters Tracer: A Large Scale System of

Detecting & Verifying Real-Time News Events from Twitter. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, pages 207–216, 2016.

[14] S. C. McGregor and L. Molyneux. Twitter's influence on news judgment: An experiment among journalists. *Journalism*, page 1464884918802975, 2018.

[15] A. J. McMinn, Y. Moshfeghi, and J. M. Jose. Building a large-scale corpus for evaluating event detection on twitter. In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, pages 409–418, 2013.

[16] F. Morstatter, J. Pfeffer, and H. Liu. When is it biased?: assessing the representativeness of twitter's streaming API. In *23rd International World Wide Web Conference, \WWW '14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume*, pages 555–556. ACM Press, 2014. ISBN 978-1-4503-2745-9. doi: $10.1145/2567948.2576952$. URL `http://dl.acm.org/citation.cfm?doid=2567948.2576952`.

[17] Y. Ning, S. Muthiah, R. Tandon, and N. Ramakrishnan. Uncovering news-twitter reciprocity via interaction patterns. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015, Paris, France, August 25 - 28, 2015*, pages 1–8, 2015.

[18] P. Nora. L'événement monstre. *Communications*, 18(1):162–172, 1972.

[19] N. Panagiotou, I. Katakis, and D. Gunopulos. Detecting Events in Online Social Networks: Definitions, Trends and Challenges. In S. Michaelis, N. Piatkowski, and M. Stolpe, editors, *Solving Large Scale Learning Tasks. Challenges and Algorithms*, volume 9580. Springer International Publishing, Cham, 2016. URL `http://link.springer.com/10.1007/978-3-319-41706-6_2`.

[20] S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to Twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–189. Association for Computational Linguistics, 2010.

[21] S. Petrović, M. Osborne, and V. Lavrenko. Using paraphrases for improving first story detection in news and twitter. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 3-8, 2012, Montréal, Canada*, pages 338–346, 2012.

[22] J. J. Randolph. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*, 2005.

[23] A. Swasy. A little birdie told me: Factors that influence the diffusion of twitter in newsrooms. *Journal of Broadcasting & Electronic Media*, 60(4):643–656, 2016.

Titre: titre (en français)

Mots clés: 3 à 6 mots clés

Résumé: Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.
Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.
Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Title: titre (en anglais)

Keywords: 3 à 6 mots clés

Abstract: Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.
Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.
Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France