# Online Appendix to the Paper: Social Media and Newsroom Production Decisions

Julia Cagé*[1], Nicolas Hervé[2], and Béatrice Mazoyer[2,3]

[1]Sciences Po Paris and CEPR
[2]Institut National de l'Audiovisuel
[3]CentraleSupélec

June 2020

## Contents



*Corresponding author. julia [dot] cage [at] sciencespo [dot] fr.

# A Data sources

## A.1 Tweet collection: Additional details

**Share of collected tweets** As described in the core of the article, we use three different methods to evaluate the share of tweets we have collected. These evaluation methods are quickly presented in Section 2.1.1. Here, we provide more details on the methods based on the number of tweets per user, and report the associated Figures A.2 and A.3

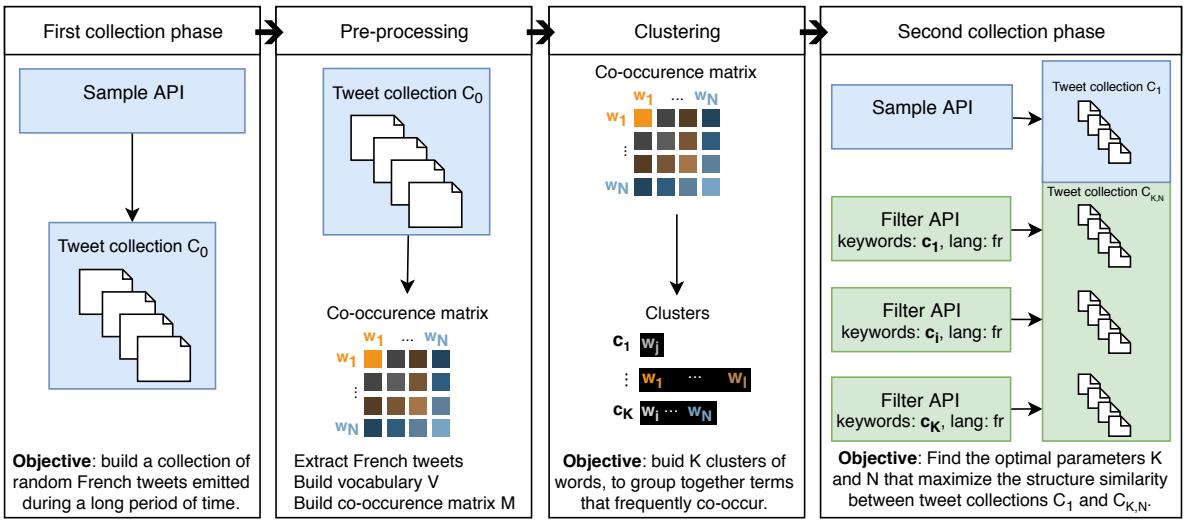


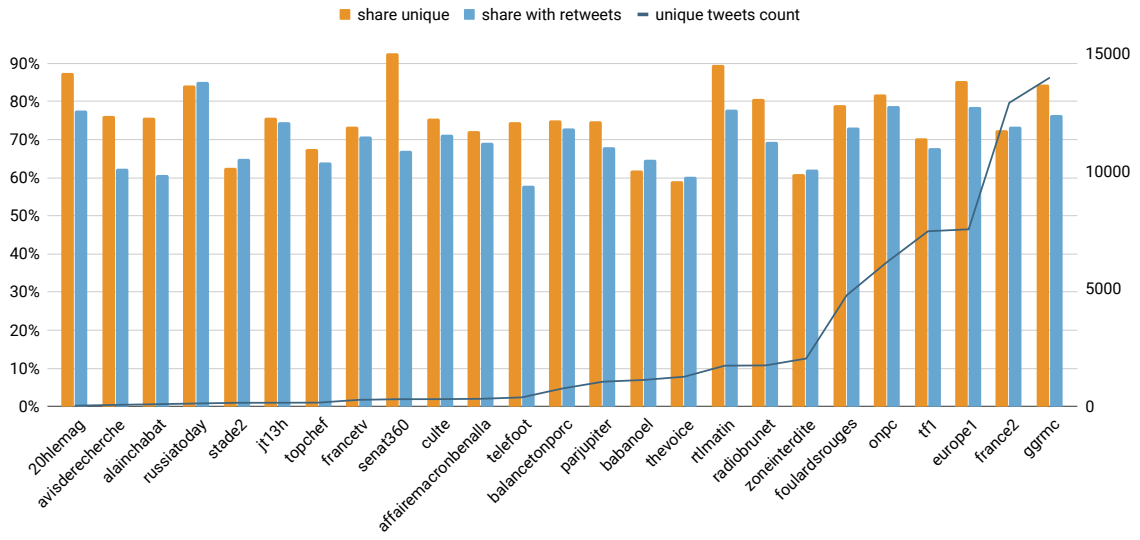

Figure A.1: Diagram of our experimental setup to select the best tweet collection method

In order to select users that write mostly in French (tweets written in other languages are not collected with our method), we used the OpenStreetMap API to locate users depending on what they indicate in the "location" field. We obtained 920,000 users localized in France that emitted 241 million tweets in three months, according to the "number of tweet" field. With our collection method, we captured 147 million tweets from these users, *i.e.* 61% of the real number of emitted tweets. We found the same percentage with the sample of users who geolocate their tweets in France (27,000 users). This method gives us a high estimate of the real number of emitted tweets in French, since some of these users probably write in other languages than French, even if they are located in France.

**List of stop words** To compute the average number of words included in the tweets, we have first removed the stop-words listed in Figure **??**.

## A.2 News media content data

The content data is from the OTMedia research projet. This projet was subsidized by the *Agence Nationale de la Recherche* (ANR – National Agency for Research), a French institution tasked with funding scientific research. The INA (*Institut National de l'Audiovisuel* – National




**Notes:** This figure plots the share of tweets from the DLWeb that we were also able to capture using our collection method. Blue columns represent the ratio for all tweets, yellow columns represent the ratio for original tweets (*i.e.* retweets excluded). The grey line shows the number of original tweets (*i.e.* retweets excluded) captured by the DLWeb for each hashtag. Tweets were collected from December 1st to December 31st, 2018.

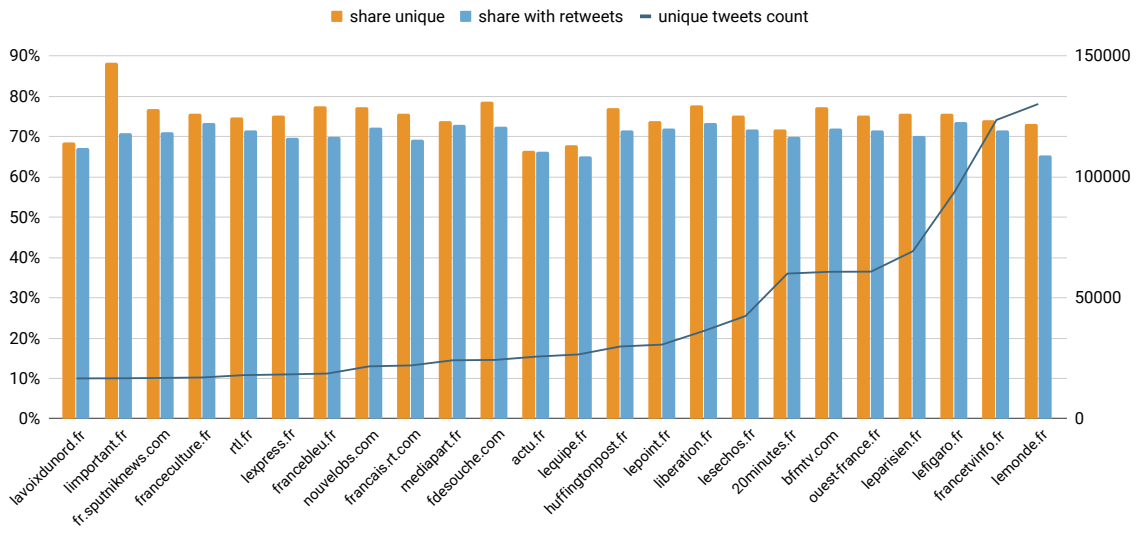
Figure A.2: Share of DLWeb tweets captured using our collection method for a set of 25 hashtags

Audiovisual Institute, a repository of all French radio and television audiovisual archives) was the project leader. The OTMedia research projet used the RSS feeds of the media outlets to track every piece of content they produced online. For the media outlets whose RSS feeds were not tracked by INA, we complete the OTMedia data by scrapping the Sitemaps of their website. Finally, we get all the AFP dispatches (respectively all the Reuters dispatches in French) directly from the AFP (from Reuters).

Our dataset includes the following media outlets:

**Local daily newspapers:**

1. *L'Ardennais*;
2. *Aisne Nouvelle*;
3. *Le Berry Republicain*;
4. *La Charente Libre*;
5. *Le Courrier Picard*;
6. *La Depeche Du Midi*;
7. *Est Eclair*;
8. *L'Eveil De La Haute Loire*;
9. *L'Independant Pyrenees Orientales*;
10. *Le Journal De La Haute Marne*;
11. *Le Midi Libre*;
12. *Monaco Matin*;
13. *La Montagne*;
14. *Nice Matin*;
15. *La Nouvelle Republique Des Pyrenees*;




**Notes:** This figure plots the share of tweets from the Mdialab that we were also able to capture using our collection method for the first 25 domain names in terms of original tweets in their dataset. Blue columns represent the ratio for all tweets, yellow columns represent the ratio for original tweets (*i.e.* retweets excluded). The grey line shows the number of original tweets (*i.e.* retweets excluded) captured by the Mdialab for each domain. Tweets were collected from December 1st to December 31st, 2018.

Figure A.3: Share of tweets from the Mdialab also captured using our collection method for 25 domain names

16. *La Nouvelle Republique Du Centre Ouest*;
17. *Ouest France*;
18. *Le Parisien*;
19. *Le Petit Bleu D'Agen*;
20. *La Provence*;
21. *La Republique Des Pyrenees*;
22. *Sud Ouest*;
23. *Le Telegramme*;
24. *L' Union*;
25. *Var Matin*;
26. *La Voix Du Nord*;
27. *Yonne Republicaine.*

**National daily newspapers:**

1. *La Croix*;
2. *Les Echos*;
3. *L'Equipe*;
4. *Le Figaro*;
5. *France Soir*;
6. *La Gazette Des Communes Des Departements Et Des Regions*;
7. *L'Humanite*;
8. *Liberation*;
9. *Le Monde*;
10. *Le Quotidien De L Art*;

```
STOP_WORDS_FR = ['0', '1', '2', '3', 'a', 'ah', 'ai', 'aime', 'aller', 'alors', 'ans', 'apres', 'après', 'as', 'au',
    'aussi', 'autre', 'autres', 'aux', 'avais', 'avait', 'avant', 'avec', 'avez', 'avoir', 'b', 'bah', 'bcp',
    'beaucoup', 'bien', 'bon', 'bonjour', 'bonne', 'bref', 'c', "c'est", "c'était", 'ca', 'ce', 'cela',
    'celle', 'celui', 'ces', 'cest', 'cet', 'cetait', 'cette', 'ceux', 'chaque', 'chez', 'co', 'comme',
    'comment', 'compte', 'contre', 'coup', 'cours', 'crois', 'cétait', 'c’est', 'd', 'dans', 'de', 'deja',
    'depuis', 'des', 'detre', 'deux', 'dire', 'dis', 'dit', 'dm', 'dois', 'doit', 'donc', 'du', 'déjà',
    'dêtre', 'e', 'eh', 'elle', 'elles', 'en', 'encore', 'entre', 'envie', 'es', 'est', 'estce', 'et', 'etais', 'etait',
    'etc', 'ete', 'etes', 'etre', 'eu', 'f', 'faire', 'fais', 'fait', 'faites', 'faut', 'fois', 'font', 'g',
    'genre', 'gens', 'grave', 'gros', 'gt', 'h', 'hein', 'https', 'i', 'il', 'ils', 'j', "j'ai", "j'aime",
    "j'avais", "j'me", "j'suis", "j'vais", 'jai', 'jaime', 'jamais', 'javais', 'je', 'jen', 'jme', 'jour',
    'journee', 'journée', 'jsp', 'jsuis', 'jte', 'juste', 'jvais', 'jveux', 'jetais', 'jétais', 'j’ai', 'k', 'l', 'la',
    'le', 'les', 'leur', 'leurs', 'lol', 'lui', 'là', 'm', 'ma', 'maintenant', 'mais', 'mal', 'mdr', 'mdrr',
    'mdrrr', 'mdrrrr', 'me', 'mec', 'meme', 'merci', 'merde', 'mes', 'met', 'mettre', 'mieux', 'mis', 'mm',
    'moi', 'moins', 'moment', 'mon', 'monde', 'mtn', 'même', 'n', 'na', 'nan', 'ne', 'nest', 'ni', 'nn',
    'non', 'nos', 'notre', 'nous', 'o', 'of', 'oh', 'ok', 'on', 'ont', 'ou', 'ouais', 'oui', 'où', 'p', 'par',
    'parce', 'parle', 'pas', 'passe', 'pcq', 'pense', 'personne', 'peu', 'peut', 'peutetre', 'peutêtre', 'peux',
    'plus', 'pour', 'pourquoi', 'pq', 'pr', 'prend', 'prendre', 'prends', 'pris', 'ptdr', 'ptdrrr', 'ptn',
    'pu', 'putain', 'q', 'qd', 'qu', "qu'il", "qu'on", 'quand', 'que', 'quel', 'quelle', 'quelque', 'quelques',
    'quelquun', 'qui', 'quil', 'quils', 'quoi', 'quon', 'r', 'rien', 'rt', 's', 'sa', 'sais', 'sait', 'sans',
    'se', 'sera', 'ses', 'sest', 'si', 'sil', 'soir', 'soit', 'son', 'sont', 'suis', 'super', 'sur', 't',
    'ta', 'tas', 'te', 'tellement', 'temps', 'tes', 'tete', 'the', 'tjrs', 'tjs', 'toi', 'ton', 'toujours',
    'tous', 'tout', 'toute', 'toutes', 'tres', 'trop', 'trouve', 'trouvé', 'très', 'tt', 'tu', 'tête', 'u',
    'un', 'une', 'v', 'va', 'vais', 'vas', 'veut', 'veux', 'via', 'vie', 'viens', 'voila', 'voilà', 'voir',
    'vois', 'voit', 'vont', 'vos', 'votre', 'vous', 'vrai', 'vraiment', 'vs', 'vu', 'w', 'wsh', 'x', 'xd',
    'y', 'ya', 'z', 'à', 'ça', 'ça', 'étais', 'était', 'été', 'êtes', 'être', '——', '-', '']
```

**Notes:** The figure plots the number of users entering our sample depending on the date of their Twitter account creation.

Figure A.4: Twitter users: Number of followers depending on the date of their account creation

11. *La Tribune*.

**Free (national daily) newspapers:**

1. *20 Minutes*.

**Weekly (national & local) newspapers:**

1. *10 Sport*;
2. *Agefi*;
3. *Argus*;
4. *Auto Hebdo*;
5. *L'Avenir De Artois*;
6. *Capital*;
7. *Challenges*;
8. *Closer*;
9. *Courrier International*;
10. *Creuse Agricole Et Rurale*;
11. *L'Echo De La Lys*;
12. *Echo Le Valentinois Drome Ardeche*;
13. *Elle*;

14. *Est Agricole Et Viticole*;
15. *L'Express*;
16. *Grazia*;
17. *Les Inrockuptibles*;
18. *Investir*;
19. *Jeune Afrique*;
20. *Le Journal De Millau*;
21. *Le Journal Du Dimanche*;
22. *L'Hebdo Du Vendredi*;
23. *La Manche Libre*;
24. *Marianne*;
25. *Le Monde Diplomatique*;
26. *Le Moniteur Des Travaux Publics Et Du Batiment*;
27. *L'Obs*;
28. *L'Observateur De Beauvais*;
29. *Paris Match*;
30. *Le Paysan Du Haut Rhin*;
31. *Le Point*;
32. *Point De Vue*;
33. *Le Republicain De L'Essonne*;
34. *La Semaine Dans Le Boulonnais*;
35. *Strategies*;
36. *Tele Z*;
37. *L'Usine Nouvelle*;
38. *Version Femina*;
39. *La Volonte Paysanne De L'Aveyron.*

**Monthly (national) newspapers:**

1. *Auto Infos*;
2. *Beaux Arts*;
3. *Causeur*;
4. *Connaissance Des Arts*;
5. *Le Courrier De Floride Etats Unis*;
6. *France Amerique Etats Unis*;
7. *Geo*;
8. *GQ Magazine*;
9. *Japon Infos*;
10. *Marie Claire*;
11. *Marie France*;
12. *Mon Viti*;
13. *Premiere*;
14. *Rav*;
15. *La Revue Des Deux Mondes*;
16. *Science Et Vie*;
17. *Sciences Et Avenir*;
18. *Sciences Humaines*;
19. *Tarx*;
20. *Tetu*;
21. *Vogue*;
22. *Zibeline.*

**TV:**

1. BFM TV;

2. Eurosport;
3. France 24;
4. LCI;
5. Public Senat;
6. TF1;
7. TV5 Monde.

**Radio:**

1. Europe 1;
2. France Bleu (Radio France);
3. France Culture (Radio France);
4. France Inter (Radio France);
5. France Musique (Radio France);
6. France Info (also TV);
7. Radio Classique;
8. RCF;
9. RFI;
10. RTL;
11. Tendance Ouest.

**News agencies:**

1. Agence France Presse.

**Pure online media:**

1. 01 Net;
2. Actu;
3. Aleteia;
4. AOC;
5. Arboriculture Fruitiere;
6. Basta;
7. Boursier Com;
8. Boursorama;
9. Bref Eco;
10. Buzzfeed;
11. C Net;
12. Cfnews;
13. Clubic;
14. Contrepoints;
15. Les Echos Du Touquet;
16. Echos Start;
17. Echosdunet;
18. Foot Mercato;
19. Football;
20. Gamekult;
21. Gamergen;
22. Generation Nouvelles Technologies;
23. Ginjfo;
24. Goodplanet Info;
25. Herault Tribune;
26. Huffington Post;
27. Influenth;
28. Informatique News;
29. L'ADN;
30. Le Libre Penseur;

31. Le Media;
32. Le Tribunal Du Net;
33. L'Explicite;
34. L'Incorrect;
35. L'Internaute;
36. LVSL;
37. Maddyness;
38. Made In Foot;
39. Made In Perpignan;
40. Marsactu;
41. Mashable;
42. Medialot;
43. Mediapart;
44. Meta Media;
45. Minutenews;
46. Mon Cultivar Elevage;
47. Mondafrique;
48. Monde Informatique;
49. Les Moutons Enrages;
50. Myeurop Info;
51. Newsly;
52. Numerama;
53. Numeriques;
54. Ohmymag;
55. Olivieranger;
56. Paris Depeches;
57. Le Petit Journal;
58. Pourquoi Docteur;
59. Pure Medias;
60. Purepeople;
61. Resistance Republicaine;
62. Rue 89 Lyon;
63. Rue89 Bordeaux;
64. Rue89 Strasbourg;
65. Slate;
66. Sputniknews;
67. Toulouse 7;
68. Toute La Culture;
69. La Tribune De L Art;
70. Up Magazine;
71. L'Usine Digitale.

**French-speaking foreign media** Further, we also gather the content produced online by the following French-speaking foreign media:

1. *20 Minutes Suisse* (Switzerland);
2. Quotidien Canada (Canada);
3. *Temps Suisse* (Switzerland);

4. Lequotidien (pure online media from Quebec);
5. Africa Intelligence;
6. Express Mu Ile Maurice;
7. Nouvelles Caledoniennes;
8. Nouvelle Tribune Benin;
9. Wort Luxembourg;
10. Infohaiti Net Haiti.

### A.3 IPTC topics

To define the subject of its dispatches, AFP uses URI, available as QCodes, designing 17 IPTC media topics. The IPTC is the International Press Telecommunications Council.

The 17 topics are defined as follows:

- **Arts, culture and entertainment**: matters pertaining to the advancement and refinement of the human mind, of interests, skills, tastes and emotions.
- **Crime, law and justice**: establishment and/or statement of the rules of behaviour in society, the enforcement of these rules, breaches of the rules and the punishment of offenders. Organisations and bodies involved in these activities.
- **Disaster and accident**: man made and natural events resulting in loss of life or injury to living creatures and/or damage to inanimate objects or property.
- **Economy, business and finance**: all matters concerning the planning, production and exchange of wealth.
- **Education**: all aspects of furthering knowledge of human individuals from birth to death.
- **Environment**: all aspects of protection, damage, and condition of the ecosystem of the planet earth and its surroundings.
- **Health**: all aspects pertaining to the physical and mental welfare of human beings.
- **Human interest**: items about individuals, groups, animals, plants or other objects with a focus on emotional facets.

- **Labour**: social aspects, organisations, rules and conditions affecting the employment of human effort for the generation of wealth or provision of services and the economic support of the unemployed.
- **Lifestyle and leisure**: activities undertaken for pleasure, relaxation or recreation outside paid employment, including eating and travel.
- **Politics**: local, regional, national and international exercise of power, or struggle for power, and the relationships between governing bodies and states.
- **Religion and belief**: all aspects of human existence involving theology, philosophy, ethics and spirituality.
- **Science and technology**: all aspects pertaining to human understanding of nature and the physical world and the development and application of this knowledge.
- **Society**: aspects of the life of humans affecting its relationships.
- **Sport**: competitive exercise involving physical effort. Organizations and bodies involved in these activities.
- **Conflicts, war and peace**: acts of socially or politically motivated protest and/or violence and actions to end them.
- **Weather**: the study, reporting and prediction of meteorological phenomena.

# B Algorithms

## B.1 Social media event detection

**Preprocessing** Each text embedding model takes different text formats as inputs: for example, models able to deal with sentences, such as BERT, Sentence-BERT, ELMo or Universal Sentence Encoder, take the full text with punctuation as input. For Word2Vec and tf-idf models, we lowercase characters and remove punctuation. Table B.1 summarizes all preprocessing steps depending on the type of model. Each column corresponds to a preprocessing step:

- Remove mentions: mentions are a Twitter-specific way of referring to another Twitter user in a tweet, so that she is notified that the tweet is talking about her or is addressed to her. Entries take the following form: @name_of_the_user. For tf-idf models, removing mentions is a way to reduce the size of the vocabulary. For most Word2Vec models, mentions are not part of the vocabulary, except for w2v_twitter_en.
- Unidecode: we use the Python module unidecode to convert Unicode characters to ASCII characters. In French, for example, all accents are removed: "Wikipédia" becomes "Wikipedia".
- Lower: we set the text in lowercase letters.
- Hashtag split: we split hashtags on capital letters. "#HappyEaster" becomes "Happy Easter". This step is of course applied before lowercasing.
- Remove long numbers: we remove numbers longer than 4 digits.
- Remove repeated characters: we limit the number of repeated characters inside a word to three. "looooool" becomes "loool".

## B.2 Mainstream media event detection algorithm

**Description of the algorithm** The goal of online topic detection is to organize a constantly arriving stream of news articles by the events they discuss. The algorithms place all the documents into appropriate and coherent clusters. Consistency is ensured both at the temporal and the semantic levels. As a result, each cluster provided by the algorithm covers the same topic (event) and only that topic. Following Allan et al. (2005) who have experienced their TDT system in a real world situation, we adopt the following implementation:

1. As in most natural language processing methods, we first pre-process our documents by removing very common words (called stop words) and applying a stemming algorithm so as to keep only the stem of the words.

| model | rm mentions | unidecode | lower | rm punctuation | hashtag split | rm long numbers | rm repeated chars | rm urls |
|---|---|---|---|---|---|---|---|---|
| tfidf_all_tweets | X | X | X | X | X | X | X | X |
| tfidf_dataset | X | X | X | X | X | X | X | X |
| w2v_afp_fr | X | X | X | X | X | X | X | X |
| w2v_twitter_fr | X | X | X | X | X | X | X | X |
| w2v_gnews_en | X | | | X | X | X | X | X |
| w2v_twitter_en | | | | X | X | X | X | X |
| elmo | | | | | X | X | X | X |
| bert | | | | | X | X | X | X |
| bert_tweets | | | | | X | X | X | X |
| sbert_sts | | | | | X | X | X | X |
| sbert_nli_sts | | | | | X | X | X | X |
| sbert_tweets_sts | | | | | X | X | X | X |
| sbert_tweets_sts_long | | | | | X | X | X | X |
| use | | | | | X | X | X | X |

Table B.1: Preprocessing applied for each model

2. Each document is then described by a semantic vector which takes into account both the headline and the text. We apply a multiplicative factor of five to the words of the title as they are supposed to describe the event well, resulting in an overweight in the global vector describing the document. A semantic vector represents the relative importance of each word of the document compared to the full dataset. A standard scheme is TF-IDF: term frequency-inverse document frequency, a numerical statistic intended to reflect how important a word is to a document in a corpus. The TF-IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus. More precisely, the weight of a word $w$ in a document $d$ is: $TFIDF(w,d) = wf\,(w,d) * log\,(N/df\,(w))$, where $wf$ is the frequency of word $w$ in the considered document, $dw$ is the number of documents in which it appears, and $N$ is the total number of documents. The total vector is $TFIDF(d) = [TFIDF(w_1,d), TFIDF(w_2,d), ..., TFIDF(w_n,d)]$, where $w_1, ..., w_n$ are the words occurring in the whole text stream to segment.

3. The documents are then clustered in a bottom-up fashion to form the events based on their semantic similarity. The similarity between two documents is given by the distance between their two semantic vectors. We use the cosine similarity measure (Salton et al., 1975).

4. This iterative agglomerative clustering algorithm is stopped when the distance between documents reaches a given threshold. We have determined this threshold empirically based on manually created media events.

5. A cluster is finalized if it does not receive any new document for a given period of time. We use a one-day window.[1]

**Performance of the algorithm** This event detection algorithm can be compared to other detection systems by its ability to put all the stories in a single event together. We test the quality of the algorithm by running it on a standard benchmark dataset: the Topic Detection and Tracking (TDT) Pilot Study Corpus. The TDT dataset contains events that have been created "manually": the goal is to compare the performance of the algorithm with that of humans. The goal of the TDT initiative is to investigate the state of the art in finding and following events in a stream of news stories (see e.g. Allan et al., 1998). To test the performance of our algorithm on the English corpus, we slightly adapt it. There is indeed no similar test corpus in French. First, we use an English stop-word list and an English stemming algorithm. Second, the time frame of the test corpus being wider than ours, the

[1] Events can last more than one day. But if during a 24-hour period of time no document is placed within the cluster, then the cluster is closed. Any new document published after this time interval becomes the seed of a new event cluster.

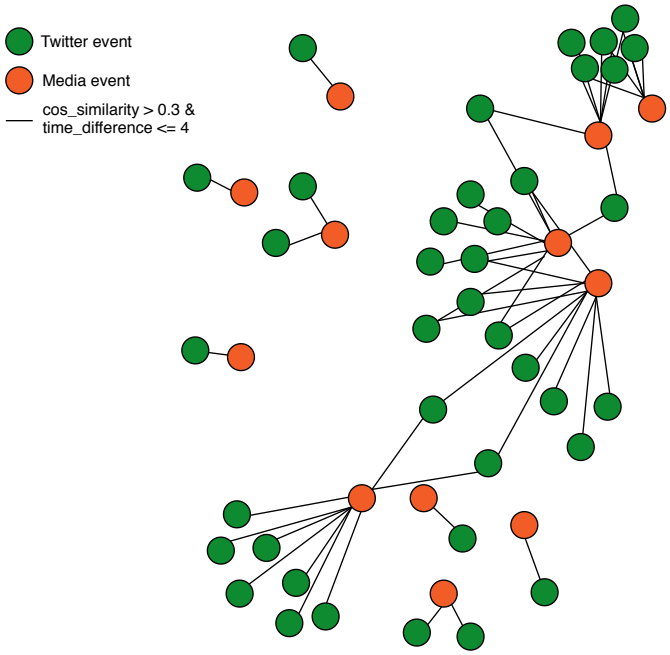
one-day window used to close clusters is not adapted. When testing our algorithm we thus follow the literature (Allan et al., 2005) and close a cluster when $2,500$ documents have been treated by the algorithm and none of them has been added to the cluster.

In the TDT Pilot Study Allan et al. (1998), two types of algorithms are evaluated: a "retrospective algorithm" and an "online algorithm". A retrospective algorithm needs to know all the articles in order to detect media events whereas an online algorithm is fed by the stream of articles, one by one. Given that the OTMedia platform must be able to manage articles in real time, we implemented an online algorithm.

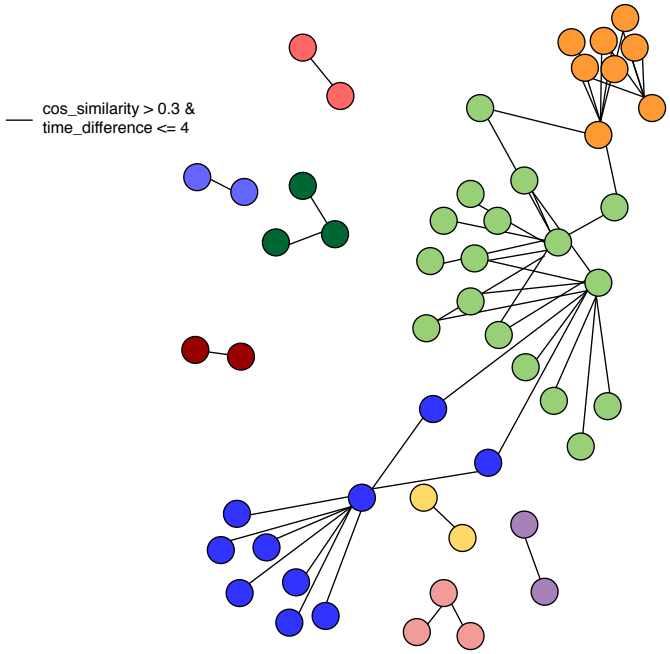
Note also that we find that the main parameter of our implementation, the distance threshold on semantic similarity, is the same for this English test corpus and our corpus of French news articles. While we were expecting these thresholds to be of similar order of magnitude (the TF-IDF representation of text is only based on word appearance frequencies; given both corpuses include articles that are of the same nature, it is not surprising to obtain relatively close thresholds for French and English), finding a similar threshold is nonetheless very reassuring as to the quality of our algorithm. In particular, it ensures that the news events that are detected by our algorithm are as close as possible to what a human would be able to do.

### B.3 Joint events

**BEATRICE** – decrire en details la methode utilisee pour les evenements joints




(a) **Building the similarity network**




(b) **Applying Louvain algorithm**

**Notes: A COMPLETER**.

Figure B.1: Graphical representation: Building the joint events

# C Additional tables

Table C.1: Summary statistics: Tweets – split sample (July 2018 - September 2018), before filters

| | Mean | St.Dev | P25 | Median | P75 | Max | Obs |
|---|---|---|---|---|---|---|---|
| **Characteristics of the tweet** | | | | | | | |
| Length of the tweet (nb of characters) | 101 | 52 | 60 | 97 | 140 | 1,121 | 428,338,133 |
| Number of words | 6.2 | 4.0 | 3.0 | 6.0 | 8.0 | 269 | 428,338,133 |
| =1 if tweet contains an URL | 0.13 | 0.33 | 0.000 | 0.000 | 0.000 | 1 | 428,338,133 |
| =1 if the tweet is a retweet | 0.63 | 0.48 | 0.000 | 1.000 | 1.000 | 1 | 428,338,133 |
| =1 if the tweet is a reply | 0.17 | 0.38 | 0.000 | 0.000 | 0.000 | 1 | 428,338,133 |
| =1 if the tweet is a quote | 0.19 | 0.39 | 0.000 | 0.000 | 0.000 | 1 | 428,338,133 |
| **Popularity of the tweet** | | | | | | | |
| Number of retweets | 2.2 | 110.4 | 0.000 | 0.000 | 0.000 | 117,389 | 159,932,748 |
| Number of replies | 0.2 | 6.5 | 0.000 | 0.000 | 0.000 | 47,892 | 159,932,748 |
| Number of likes | 3.7 | 177.6 | 0.000 | 0.000 | 0.000 | 449,881 | 159,932,749 |

**Notes:** The table gives summary statistics. Time period is July 2018 - September 2018. Variables are values for all the tweets included in our dataset before we applied the filters to remove the bots. Variables are described in more details in the text.

Table C.2: Summary statistics: Twitter users (full sample; last time the user is observed)

| | Mean | St.Dev | P25 | Median | P75 | Max |
|---|---|---|---|---|---|---|
| **User activity** | | | | | | |
| Total number of tweets | 15,174 | 40,642 | 286 | 2,265 | 12,762 | 6,183,567 |
| Nb of tweets btw first & last time | 99 | 445 | 4 | 9 | 38 | 61,203 |
| Nb of tweets user has liked | 8,520 | 23,184 | 158 | 1,220 | 6,655 | 2,831,010 |
| Nb of users the account is following | 688 | 4549 | 88 | 211 | 519 | 1672425 |
| **User identity** | | | | | | |
| Date of creation of the account | 2,014.469 | 2.742 | 2,012 | 2,015 | 2,017 | 2,018 |
| =1 if verified account | 0.005 | 0.074 | 0 | 0 | 0 | 1 |
| =1 if user is a journalist | 0.0012 | 0.034 | 0 | 0 | 0 | 1 |
| =1 if user is a media | 0 | 0 | 0 | 0 | 0 | 1 |
| **User popularity** | | | | | | |
| Nb of followers | 2,200 | 86,685 | 32 | 147 | 515 | 58,775,462 |
| Nb of public lists | 20 | 578 | 0 | 1 | 6 | 1,028,438 |
| Observations | 4,222,734 | | | | | |

**Notes:** The table gives summary statistics. Time period is July 2018 - July 2019. Variables are values for all the Twitter users included in our dataset the last time we observe them. Variables are described in more details in the text.

Table C.3: Summary statistics: Joint events – Depending on news breaker

| | Media first | Twitter first | Diff/se |
|---|---|---|---|
| Length of the event (in hours) | 408 | 529 | -120*** |
| | | | (15) |
| Number of documents in event | 5,678 | 4,719 | 959 |
| | | | (2,171) |
| **Twitter coverage** | | | |
| Nb of tweets in event | 5,623 | 4,676 | 947 |
| | | | (2,170) |
| Number of different Twitter users | 2,125 | 2,957 | -832 |
| | | | (467) |
| Average number of retweets of tweets in events | 2.6 | 2.5 | 0.0 |
| | | | (0.1) |
| Average number of replys of tweets in events | 0.3 | 0.3 | -0.0 |
| | | | (0.0) |
| Average number of favorites of tweets in events | 3.1 | 3.7 | -0.6** |
| | | | (0.2) |
| **Media coverage** | | | |
| Number of news articles in the event | 55 | 43 | 12*** |
| | | | (3) |
| Number of different media outlets | 18 | 16 | 1** |
| | | | (0) |
| Observations | 5,766 | | |

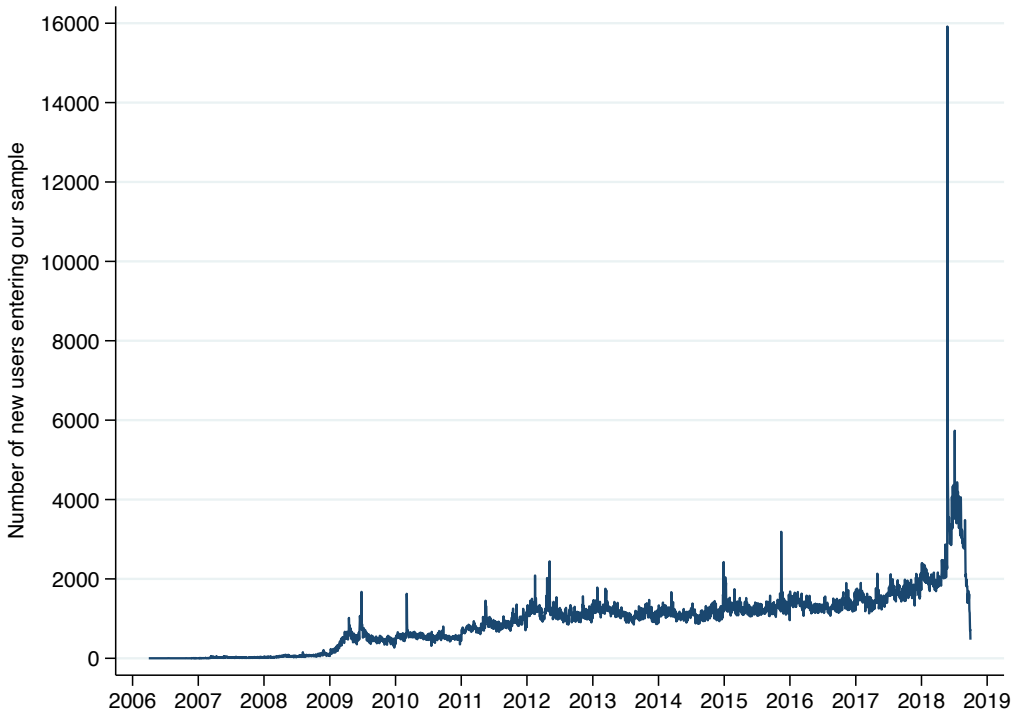
**Notes:** The table gives summary statistics. Time period is July 2018 - September 2018. The observations are at the event level. Column 1 presents the events that appear first on media. Column 2 presents the results for the events that appear first on Twitter. In column 3, we perform a $t$-test on the equality of means.

Table C.4: Summary statistics: Twitter users – Gatekeepers

| | Mean | St.Dev | P25 | Median | P75 | Max |
|---|---|---|---|---|---|---|
| **User activity** | | | | | | |
| Total number of tweets | 65,663 | 131,782 | 4,700 | 21,555 | 74,287 | 6,183,567 |
| Nb of tweets July-September 2018 | 112 | 580 | 4 | 9 | 43 | 46,013 |
| Nb tweets user has liked | 20,707 | 53,746 | 415 | 2,913 | 15,874 | 2,831,010 |
| Nb of users the account is following | 11,475 | 35,916 | 353 | 1,053 | 7,578 | 1,672,425 |
| **User identity** | | | | | | |
| Date of creation of the account | 2012 | 3 | 2010 | 2011 | 2014 | 2018 |
| =1 if verified account | 39.5 | 48.9 | 0.0 | 0.0 | 100.0 | 100 |
| =1 if user is a journalist | 8.45 | 27.82 | 0.00 | 0.00 | 0.00 | 100 |
| =1 if user is a media | 0.757 | 8.668 | 0.000 | 0.000 | 0.000 | 100 |
| **User popularity** | | | | | | |
| Nb of followers | 115,010 | 727,425 | 12,835 | 26,462 | 57,461 | 58,775,462 |
| Nb of public lists | 592 | 4,854 | 64 | 175 | 461 | 1,028,438 |
| Observations | 58,521 | | | | | |

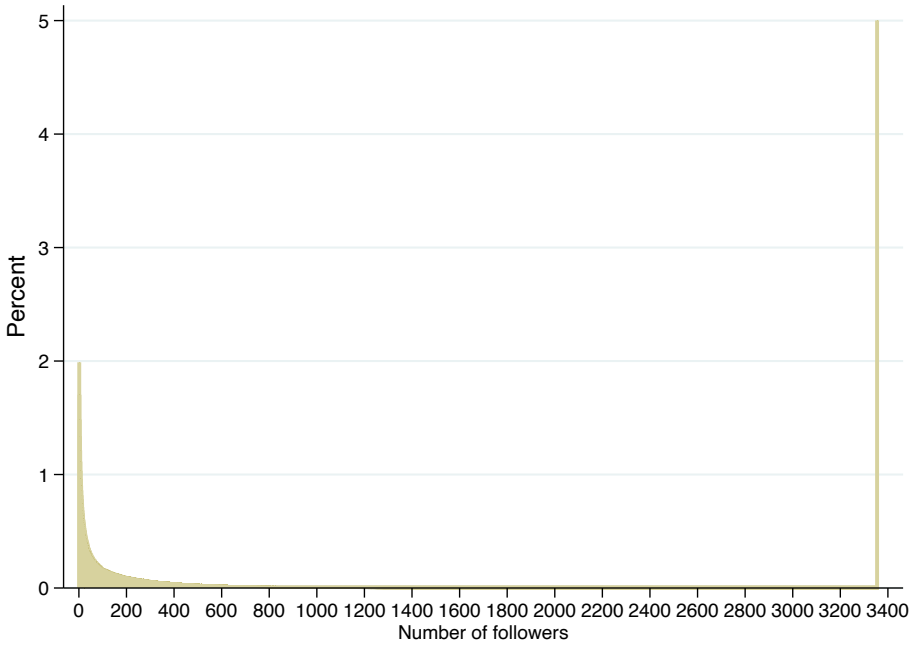
**Notes:** The table gives summary statistics. Time period is July 2018 - September 2018. Variables are values for all the "gatekeepers" included in our dataset the last time we observe them. Gatekeepers are defined as: **TO BE COMPLETED**

# D Additional figures

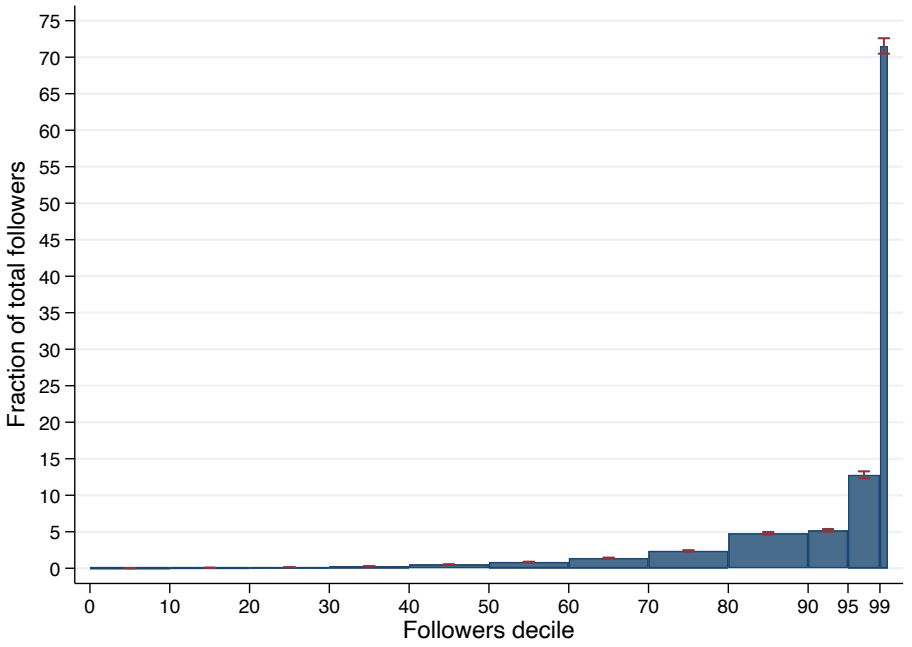



**Notes:** The figure plots the number of users entering our sample depending on the date of their Twitter account creation.

Figure D.1: Twitter users: Number of followers depending on the date of their account creation



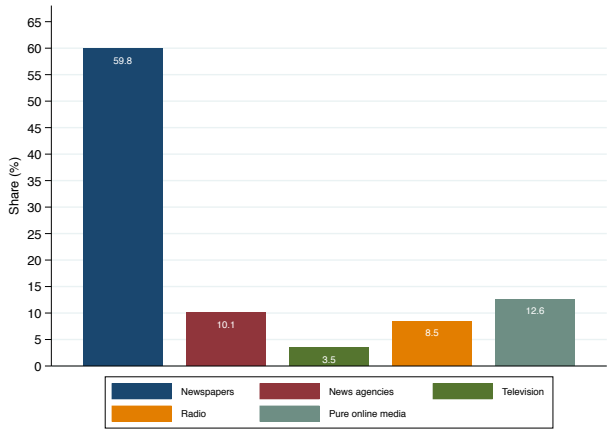

(a) **Distribution of the number of followers**



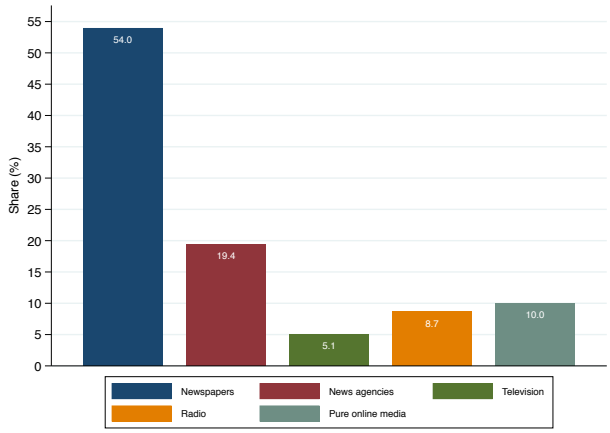

(b) **Cumulative distribution of the number of followers**

**Notes:** The upper Figure D.2a plots the distribution of the number of followers (winsorized at the 95th percentile, i.e. 3,355 followers) of the Twitter users in our dataset. The bottom Figure D.2b plots the cumulative distribution of the number of followers.

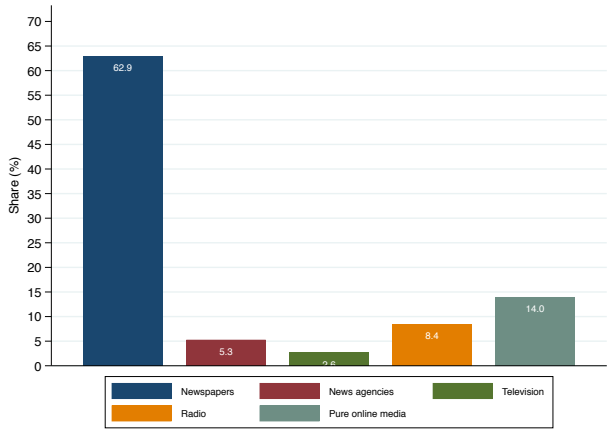
Figure D.2: Twitter users: Distribution of the number of followers




(a) All documents



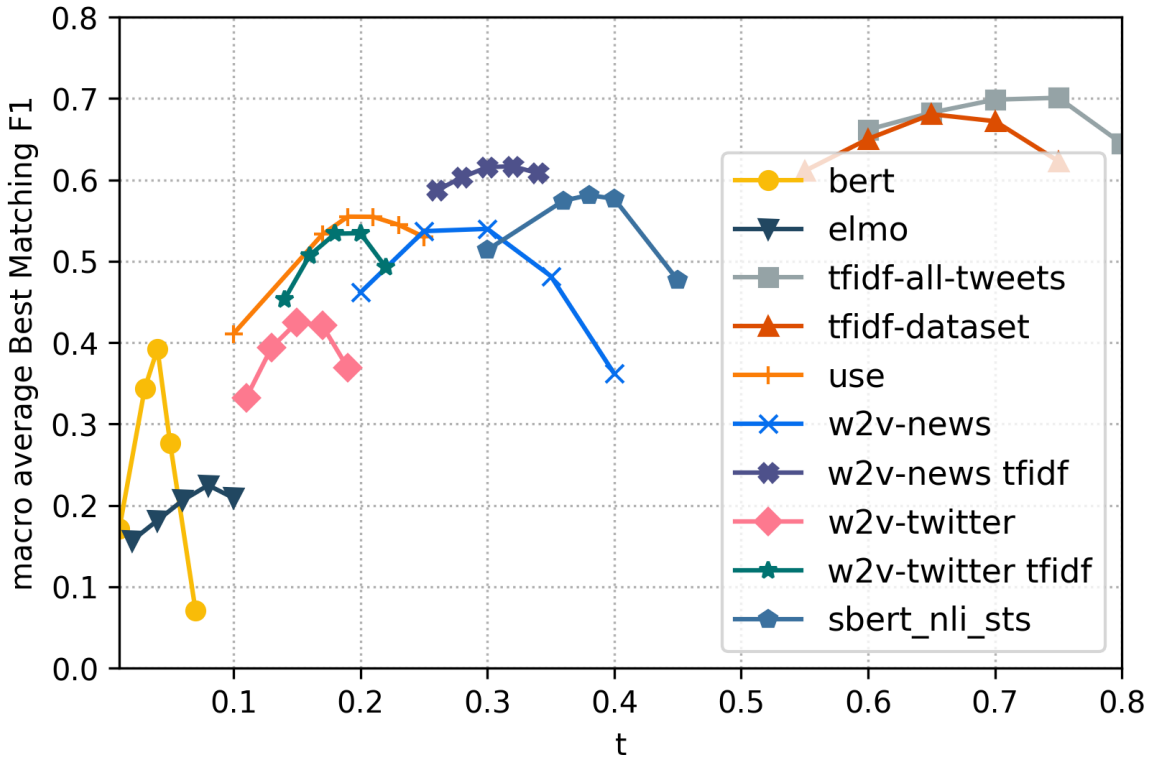

(b) Documents classified in events



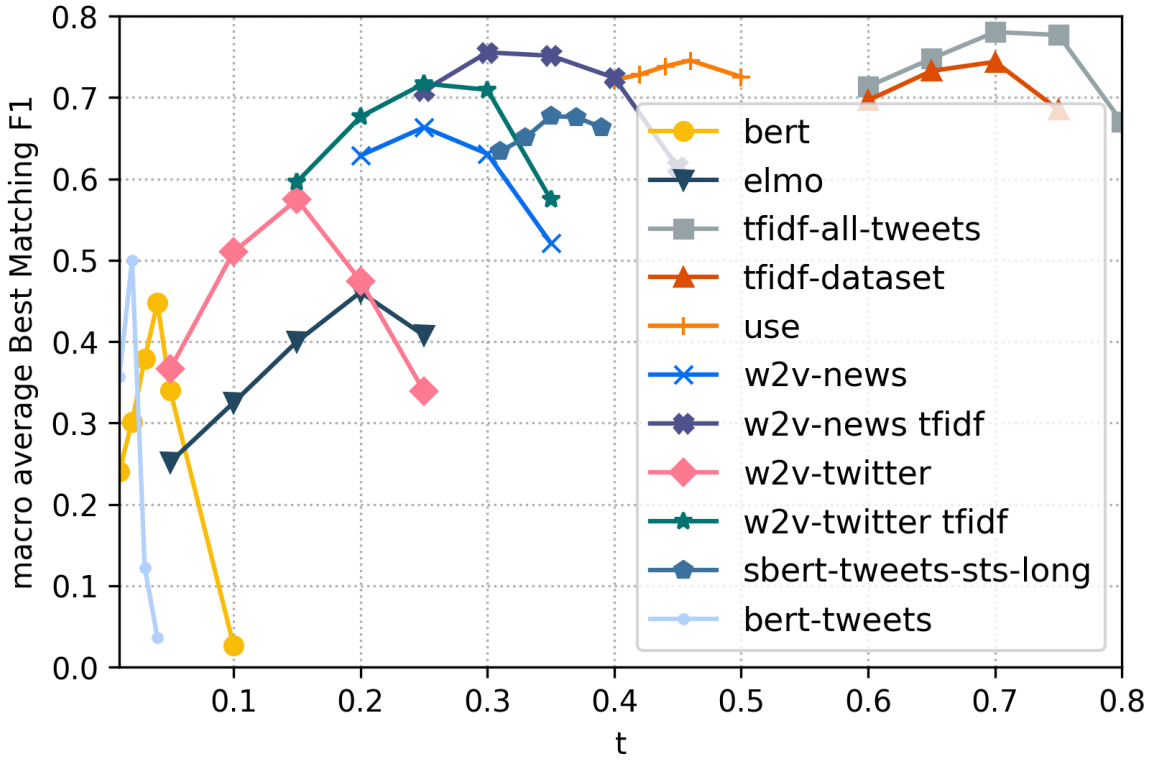

(c) Documents not classified in events

**Notes:** The figures plot the share of the documents depending on the offline format of the media outlet. The upper figure D.3a plots this number for all the documents; the middle figure D.3b for the documents classified in events; and the bottom figure D.3c for the documents not classified in events. News events are defined in detail in the text, and the list of the media outlets included in each category is given in Section A.

Figure D.3: Share of the documents by offline format

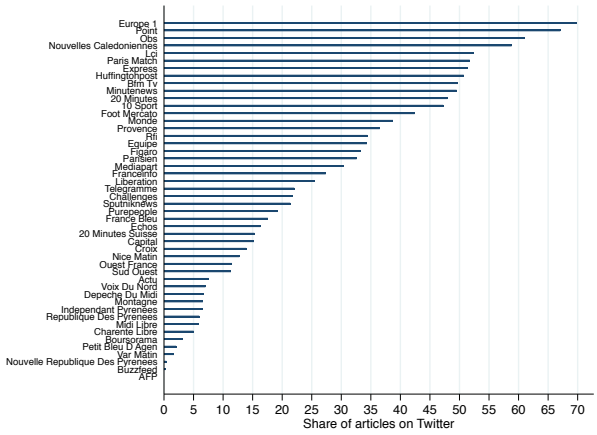



(a) **English**

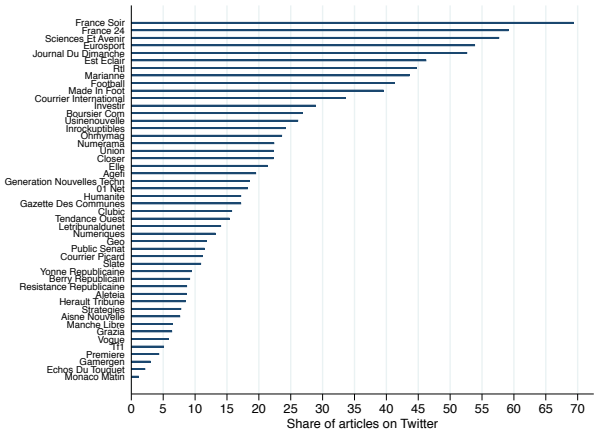



(b) **French**

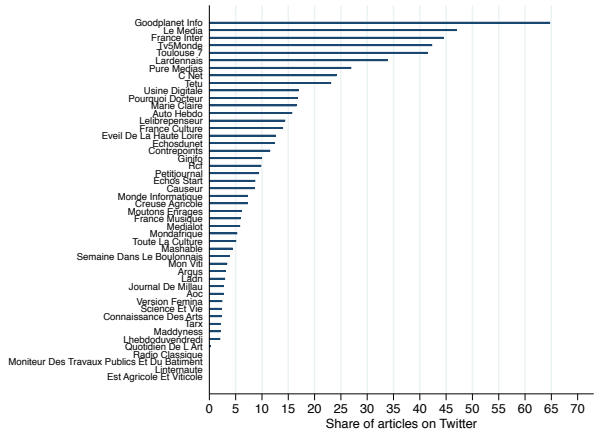
Figure D.4: Best Matching F1 score for FSD clustering depending on the threshold parameter $t$ for each corpus




(a) Fourth quartile of the number of articles distribution



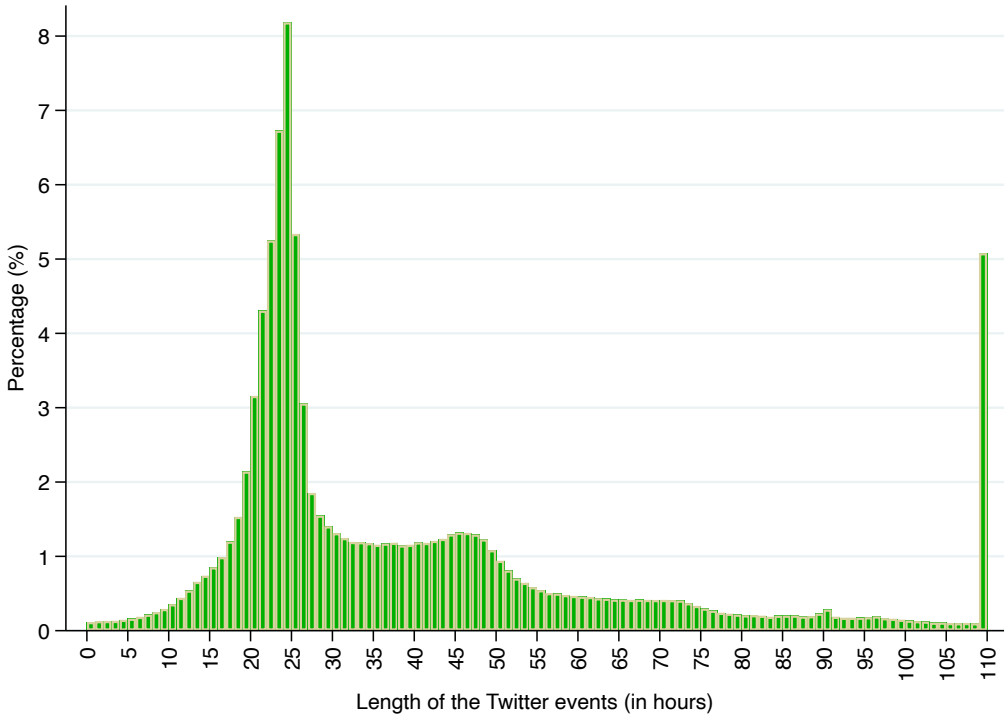

(b) Third quartile of the number of articles distribution




(c) Second quartile of the number of articles distribution

**Notes:** The figures plot the share of the articles published online that are on Twitter, depending on the media outlet. Media outlets are ranked depending on the number of articles they publish online between July 2018 and September 2018. The upper figure D.5a plots the share for the media outlets that are in the fourth quartile of the number of articles distribution; the middle figure D.5b for the media outlets that are in the third quartile; and the bottom figure D.5c for the media outlets that are in the second quartile.

Figure D.5: Share of the articles published online that are on Twitter, depending on the media outlet




**Notes:** The figure plots the distribution of the length of the Twitter events (in hours), Winsorized at the 95th percentile (=105 hours). **REPRENDRE PROPEMENT**

Figure D.6: Twitter events: Distribution of the length of the events (in hours), Winsorized at the 95th percentile (=109.8 hours)

## References

**Allan, James, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang**, "Topic Detection and Tracking Pilot Study Final Report," in "In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop" 1998, pp. 194–218.

**_ , Stephen Harding, David Fisher, Alvaro Bolivar, Sergio Guzman-Lara, and Peter Amstutz**, "Taking Topic Detection From Evaluation to Practice," in "Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 4 - Volume 04" HICSS '05 IEEE Computer Society Washington, DC, USA 2005.

**Salton, Gerard M., Andrew K. C. Wong, and Chungshu Yang**, "A Vector Space Model for Automatic Indexing," *Commun. ACM*, 1975, *18* (11), 613–620.