Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6220–6227
Marseille, 11–16 May 2020

A French Corpus for Event Detection on Twitter

Béatrice Mazoyer[1,3] **, Julia Cagé**[2] **, Nicolas Hervé**[3] **, Céline Hudelot**[1]
[1]CentraleSupélec (Paris-Saclay University), MICS, Gif-sur-Yvette, France
[2]SciencesPo Paris, Department of Economics, Paris, France
[3]Institut National de l'Audiovisuel, Research Department, Bry-sur-Marne, France
{beatrice.mazoyer, celine.hudelot}@centralesupelec.fr, julia.cage@sciencespo.fr, nherve@ina.fr

Abstract

We present Event2018, a corpus annotated for event detection tasks, consisting of 38 million tweets in French (retweets excluded) including more than 130,000 tweets manually annotated by three annotators as related or unrelated to a given event. The 257 events were selected both from press articles and from subjects trending on Twitter during the annotation period (July to August 2018). In total, more than 95,000 tweets were annotated as related to one of the selected events. We also provide the titles and URLs of 15,500 news articles automatically detected as related to these events. In addition to this corpus, we detail the results of our event detection experiments on both this dataset and another publicly available dataset of tweets in English. We ran extensive tests with different types of text embeddings and a standard Topic Detection and Tracking algorithm, and detail our evaluation method. We show that tf-idf vectors allow the best performance for this task on both corpora. These results are intended to serve as a baseline for researchers wishing to test their own event detection systems on our corpus.

Keywords: Natural Language Processing, Twitter, Topic Detection and Tracking, Event detection, Dataset, French

1. Introduction

Twitter is known as a major source of information about public events: a recent study by McGregor and Molyneux (2018) shows that journalists using Twitter as part of their daily work consider tweets as newsworthy as headlines from the Associated Press. Twitter can thus help researchers understand public opinion, measure the interest raised by certain types of topics, or detect some unpredictable events in real time. Indeed, the social network has been used for flood prevention (de Bruijn et al., 2017), to detect local events (Wei et al., 2019), or to predict stock market movements (Pagolu et al., 2016), among many other applications.

However, the specificities of Twitter (short texts, use of slang, abbreviations, hashtags, images and videos, very high volume of data) make all automatic detection tasks very difficult on tweet datasets. Besides, research on general topic detection and tracking (without specification of the type of topic) lacks some publicly available tweet datasets to produce reproducible results. This is all the more true for non-English languages. Existing datasets may also have different definitions of event or topic, which leads to issues when comparing event-detection systems. Many works on event detection are actually focused on burst detection (detecting topics such as natural disasters, attacks, etc., that cause an unusual volume of tweets), and do not attempt to assess the relative size of events. We seek to detect all events, both those that generate a high volume of tweets and those that are little discussed, and to group together all tweets related to the same event. The very broad definition proposed by McMinn et al. (2013) therefore seems to us the most relevant:

"*Definition 1.* An *event* is a **significant** thing that happens at some specific time and place."

"*Definition 2.* Something is *significant* if it may be discussed in the media."

With this definition in mind, the topic detection and tracking task is conceptually similar to dynamic clustering. It requires a large scale evaluation corpus, in order to be representative of the task's complexity on real-world data. In this article, we present (1) Event2018[1] a corpus consisting of more than 38 million tweets in French, including more than 130,000 tweets manually annotated by three annotators as related or unrelated to a given event, (2) the titles and URLs of 15,500 news articles automatically detected as related to one of these events, and (3) our event detection results on Event2018 and on a similar corpus in English (McMinn et al., 2013). With these results, we aim at providing a common baseline for researchers of the field. The code of our experiments is publicly available online.[2]

2. Related Work

Twitter gives a limited access to its data, but still provides some API endpoints to retrieve tweets (which is not the case of other more popular social networks), hence the large number of works based on Twitter datasets. However, few of them provide access to their evaluation corpora. We detail available event detection collections in this section.

McMinn et al. (2013) created the largest available corpus on event detection. They used several methods to generate candidate events: two automatic event detection methods on their set of 120 million tweets in English and one method based on query expansion of Wikipedia events. The automatically generated events were then assessed using Amazon Mechanical Turk, firstly to evaluate if the automatically generated events corresponded to their definition of event, secondly to judge if the clustered tweets were all relevant to the event. They finally merged together the events from the

[1] The corpus is available at dataset.ina.fr. Please fill-in the agreement form and indicate the name of the corpus (Event2018) in your application. In accordance to Twitter's terms of use, we only provide the tweet ids. However, our github repository contains a python script to retrieve the full text of the tweets from the Twitter API.

[2] github.com/ina-foss/twembeddings

three candidate generation methods and removed duplicate events. The final corpus consists in more than 100,000 annotated tweets covering 506 events. However, this corpus dates from 2012. Because of tweets beeing removed and Twitter accounts being closed, a large part of this dataset can no longer be recovered. In August 2018, we could only retrieve 66.5 million tweets from the original collection (55%) and 72,790 tweets from the annotated corpus (72%).

The SNOW dataset (Papadopoulos et al., 2014) can also be used as test corpus for an event detection task. It is composed of 59 events from the headlines of the BBC RSS newsfeed and from NewsWhip UK published during one day (25 February 2014). However it does not provide comprehensive sets of tweets related to each event, only two to five "representative" tweets from a collection of more than 1 million tweets emitted on that date.

The Signal-1M corpus consists of two datasets: a dataset of 1 million news articles from September 2015 (Corney et al., 2016), and a dataset of tweets related to 100 randomly selected articles (Suarez et al., 2018). The tweets dataset contains approximatively 6000 tweets.

The News Event Dataset (Mele and Crestani, 2019) contains 147,000 documents from 9 news channels and published on Twitter, RSS portals and news websites from March to June 2016. 57 media events were annotated on a crowdsourcing platform, to label 4300 documents, including 744 tweets.

The last two datasets do not contain enough tweets to allow a large-scale evaluation of any event detection and tracking system. In addition, events in SNOW, Signal-1M and the News Events Dataset are selected from media sources only, which restricts the definition of an event. Only the corpus by McMinn et al. relies on a topic detection step among collected tweets before proceeding to the annotation step. This method allows for a greater variety of events, but it is likely to influence the evaluation: indeed, event detection systems similar to those used by McMinn et al. for corpus creation may get better results when tested on this corpus.

We tried to avoid these biases when creating our own corpus. The methodology used to build our dataset is detailed in the following section.

3. Methodology

3.1. Corpus Construction

3.1.1. Tweets

In order to obtain a large volume of tweets in French, we relied on a collection method (Mazoyer et al., 2018) based on the use of neutral words (the most common words used in French on Twitter) to query the Twitter ”streaming” API. Our experiments show that the distribution of tweets in our collection is similar to the distribution of tweets in samples collected randomly using the Twitter ”sample” API. Thus, we can consider that we do not introduce any bias in our dataset (some events being more represented than others, for example) linked to the collection method. Besides, we compared our collection with the corpus of tweets built by Cardon et al. (2019), that consists of tweets containing URLs from a curated list of 420 media sources. Our collection contains 61% of these tweets, which allows us to affirm that we collect nearly 60% of all tweets issued in French at a given moment. This method allowed us to retrieve more than 110 million tweets in French, including 38 million original tweets (tweets that are not retweets), in a period of 23 days, from July 16th to August 7th, 2018.

3.1.2. News articles

As part of OTMedia, a wider project on news propagation (Hervé, 2019), we tracked every piece of content published online by French news media during the same period. We did not perform any manual annotation on these documents but we used the First Story Detection algorithm (see section 5.1) to group together articles related to the same topic.

3.2. Corpus Annotation

In order to annotate a large variety of topics, we mixed events drawn randomly from press articles and from events trending on Twitter. 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. We did not either use Wikipedia to select important events, considering that an event detection system should be able to detect topics of different scales, from everyday Twitter gossips to worldwide important news.

3.2.1. Press Events Selection

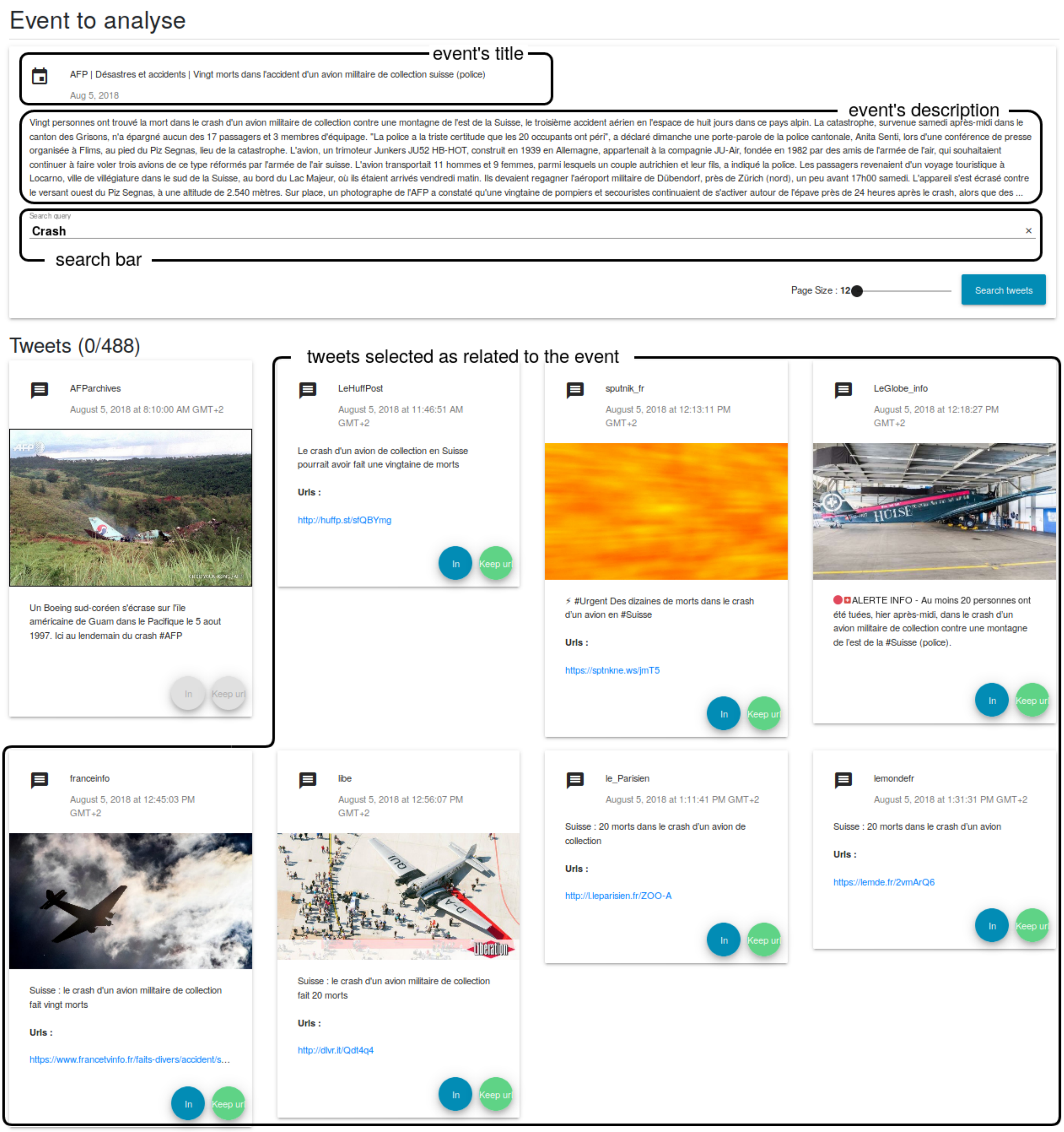
In practice, we drew 30 events a day, two thirds from the *Agence France Presse* (AFP), which is the world third largest news agency, and one third from a pool of major French newspapers (*Le Monde, Le Figaro, Les Échos, Libération, L'Humanité, Médiapart*). Duplicates, that is articles covering the same event, were manually removed. 30 events seemed to us to be the maximum number of events the annotators could process in one day. In reality they did not have time to annotate most of them, and only processed the beginning of the list each day (see section 3.2.3 for more details on annotation instructions).

3.2.2. Twitter Events Selection

We did not want to miss events in the Twitter sphere that would receive little coverage from 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 to identify "significant terms" in a subset of documents (foreground set) compared to the rest of the collection (background set). In our case, the score simply compares for each term t its frequency of appearance on a given day to its frequency on every other days. 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.

3.2.3. Manual Annotation Procedure

Three political science graduate students were hired for a month to annotate the corpus. All three of them were Twitter users and had a good knowledge of French news. Every day they were presented with the new list of events. We developed a user interface (see Figure 1) presenting each event in the form of a title and a description text.



Event to analyse

AFP | Désastres et accidents | Vingt morts dans l'accident d'un avion militaire de collection suisse (police)

Aug 5, 2018

Vingt personnes ont trouvé la mort dans le crash d'un avion militaire de collection contre une montagne de l'est de la Suisse, le troisième accident aérien en l'espace de huit jours dans ce pays alpin. La catastrophe, survenue samedi après-midi dans le canton des Grisons, n'a épargné aucun des 17 passagers et 3 membres d'équipage. "La police a la triste certitude que les 20 occupants ont péri", a déclaré dimanche une porte-parole de la police cantonale, Anita Senti, lors d'une conférence de presse organisée à Flims, au pied du Piz Segnas, lieu de la catastrophe. L'avion, un trimoteur Junkers JU52 HB-HOT, construit en 1939 en Allemagne, appartenait à la compagnie JU-Air, fondée en 1982 par des amis de l'armée de l'air, qui souhaitaient continuer à faire voler trois avions de ce type réformés par l'armée de l'air suisse. L'avion transportait 11 hommes et 9 femmes, parmi lesquels un couple autrichien et leur fils, a indiqué la police. Les passagers revenaient d'un voyage touristique à Locarno, ville de villégiature dans le sud de la Suisse, au bord du Lac Majeur, où ils étaient arrivés vendredi matin. Ils devaient regagner l'aéroport militaire de Dübendorf, près de Zürich (nord), un peu avant 17h00 samedi. L'appareil s'est écrasé contre le versant ouest du Piz Segnas, à une altitude de 2.540 mètres. Sur place, un photographe de l'AFP a constaté qu'une vingtaine de pompiers et secouristes continuaient de s'activer autour de l'épave près de 24 heures après le crash, alors que des ...

Search query: Crash ×

Page Size : 12

Tweets (0/488)

AFParchives
August 5, 2018 at 8:10:00 AM GMT+2
Un Boeing sud-coréen s'écrase sur l'île américaine de Guam dans le Pacifique le 5 aout 1997. Ici au lendemain du crash #AFP

LeHuffPost
August 5, 2018 at 11:46:51 AM GMT+2
Le crash d'un avion de collection en Suisse pourrait avoir fait une vingtaine de morts
Urls :
http://huffp.st/sfQBYmg

sputnik_fr
August 5, 2018 at 12:13:11 PM GMT+2
⚡ #Urgent Des dizaines de morts dans le crash d'un avion en #Suisse
Urls :
https://sptnkne.ws/jmT5

LeGlobe_info
August 5, 2018 at 12:18:27 PM GMT+2
● ALERTE INFO - Au moins 20 personnes ont été tuées, hier après-midi, dans le crash d'un avion militaire de collection contre une montagne de l'est de la #Suisse (police).

franceinfo
August 5, 2018 at 12:45:03 PM GMT+2
Suisse : le crash d'un avion militaire de collection fait vingt morts
Urls :
https://www.francetvinfo.fr/faits-divers/accident/s...

libe
August 5, 2018 at 12:56:07 PM GMT+2
Suisse : le crash d'un avion militaire de collection fait 20 morts
Urls :
http://dlvr.it/Qdt4q4

le_Parisien
August 5, 2018 at 1:11:41 PM GMT+2
Suisse : 20 morts dans le crash d'un avion de collection
Urls :
http://l.leparisien.fr/ZOO-A

lemondefr
August 5, 2018 at 1:31:31 PM GMT+2
Suisse : 20 morts dans le crash d'un avion
Urls :
https://lemde.fr/2vmArQ6

Figure 1: View of the annotation interface

For media events we used the title and the first paragraphs of the drawn 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 the search bar to enter keywords and find the collected tweets containing those exact keywords. For every event, they were asked to search for related tweets on the user interface, using a large variety of keywords.

We 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. (2013), we asked the annotators to mark tweets as related to the event if it referred to it, even in an implicit way.

Some major events would have required days of work to be fully treated. We therefore instructed the students not to spend more than an hour on an event. This has of course an impact on the maximum number of tweets per event that could be annotated. Even so, the students were only able to process an average of 16 events per day (21 working days over a month, 327 events in total). 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.

3.2.4. Maximizing Annotation Efficiency

The annotation interface has been designed to take advantage of the annotators' intelligence and avoid repetitive tasks. Thus, it seemed unnecessary to make them annotate tweets that contained exactly the same text. For tweets containing the same url, we added a "keep url" checkbox (see the green buttons on Figure 1), which was checked by default. If annotators felt that the url present in the tweet did not refer to content related to the event, they had to uncheck the box. For most other tweets (for which the url did refer to an event-related article), tweets containing the same url were no longer shown to the annotators in the interface. Here are all the rules we have used to avoid repetition. For a given event: (1) tweets longer than 4 words containing the same text as an already annotated tweet were no longer displayed; (2) tweets containing the same url as an already annotated tweet were no longer displayed, unless the "keep url" checkbox was unchecked for that tweet; (3) retweets and responses to a tweet, as well as tweets quoting a previously annotated tweet were not displayed. These rules played an important role to improve the quality of the corpus in terms of tweets diversity within a given event. The following section details some quality evaluation metrics.

4. Corpus Characteristics

4.1. Descriptive Statistics

In total, 137,757 tweets were annotated (found by annotators using search keywords), and 95,796 tweets were considered linked to one or several events. 327 events were selected, including 31 "Twitter events" (detected using the term frequency on Twitter) and 296 "media events" (drawn randomly in our collection of online news). For the "Twitter events", we could manually associate 27 of them to a news article from our collection. From these 323 articles (296 + 27), 167 were automatically clustered with other articles from our pool. In total, we provide the titles of 15,544 news articles from 61 media outlets and their URLs, when available.[3] Since the events were discovered day by day, we manually merged some of them to obtain 257 "macro-events". A "macro-event" contains 296 tweets on average. 0.5% of the tweets were linked to several events. Additional descriptive statistics are presented in Table 1.

To describe the distribution of events across categories we used the classification by the French news agency AFP.

	Annotated	**Linked to daily event**	**Linked to macro event**
tweet count	137757	95796	95796
event count	327	327	257
mean	428	296	376
std	434	449	1324
min	7	2	2
25%	150	30	22
50%	280	100	76
75%	494	344	241
max	2913	2906	18534

Table 1: Distribution of the number of tweets per event. **Annotated** tweets were found by annotators using search keywords but not necessarily considered as **linked to an event**.

IPTC category	**Event count**
arts, culture, entertainment and media	14
conflict, war and peace	25
crime, law and justice	76
disaster, accident and emergency incident	11
economy, business and finance	53
education	4
environment	5
health	3
human interest	6
labour	3
lifestyle and leisure	2
politics	46
religion and belief	1
science and technology	2
society	17
sport	56
weather	3
Total	**327**

Table 2: Distribution of daily events across the 17 top IPTC Information Interchange Model Media Topics.

AFP dispatches are labeled using the IPTC Information Interchange Model[4] categories. This taxonomy is used internationally by media companies to apply metadata to text, images and videos. The distribution of events across these categories is detailed in Table 2.

4.2. Intra-event diversity

The quality of the dataset should also be measured in terms of variety of the tweets within a given event: what proportion of the tweets are simply the headline of the same linked article, for instance? Indeed, the proportion of tweets containing an url is high among the annotated tweets (71%), compared to the entire corpus of 38 million tweets (36%). However, due to the design of the annotation interface (see Section 3.2.4), very few annotated tweets share the same url. Only 4.6% of tweets linked to an event share the same url as another tweet in the same event. However, we did not anticipate during annotation that different urls can be linked to the same page. After redirection, 9.3% of tweets share the same page as another tweet in the same event, but 95% of pages are shared less than 3 times in a given event.

[3] AFP dispatches are not publicly available online, therefore we cannot provide their URL.

[4] en.wikipedia.org/wiki/IPTC_Information_Interchange_Model

4.3. 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 (2005) recommend to use Fleiss' kappa (Fleiss, 1971) 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 (Randolph, 2005) 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. (2013). In our corpus, $\kappa_{free} = 0.79$ which indicates a strong level of agreement among annotators.

5. Event Detection Baseline

This dataset was designed to allow the scientific community to compare its event detection results. To enable researchers to evaluate their work, we share an evaluation baseline. To do so, we present in this section the standard event detection algorithm, an evaluation metric, and the results of our experiments with different text embedding methods. In order to assess the validity of our approach in several languages, we run the same experiments on the dataset by McMinn et al. (2013).

5.1. Event Detection Algorithm

We model the task of event detection as a dynamic clustering problem. For this type of clustering, algorithms need to take into account both the thematic similarity of the documents and their temporal proximity, so as not to group together tweets issued at very different times. Moreover, the number of topics (and therefore clusters) is usually not known in advance. To meet these constraints, techniques such as the "First Story Detection" algorithm, used in the UMass system (Allan et al., 2000), are often applied to tweets. This is the case in the works of Petrović et al. (2010), Hasan et al. (2016), and Repp and Ramampiaro (2018). Petrović et al. (2010) propose a method based on Locality Sensitive Hashing to accelerate search, while Repp and Ramampiaro (2018) introduce mini-batches. This "FSD" algorithm is detailed below (see Algorithm 5.1) because it is also the one we use to establish our baseline.

Similarly to Repp and Ramampiaro (2018), we use the FSD algorithm with mini-batches of 8 tweets to accelerate the computation time. Other parameters of this algorithm are w (number of past tweets among which we look for a nearest neighbor) and t, the distance threshold above which a tweet is considered far enough away from past tweets to form a new cluster. The value of w has been set differently for each corpus: it is set to match aproximately the size of one day of tweets, depending on the average number of tweets per day in each corpus. Different t values were then tested for each type of lexical embedding. In general, lower t values result in smaller clusters, and therefore greater intra-cluster homogeneity, but can lead to over-clustering.

Algorithm 1 "First Story Detection"

```
Input: threshold $t$, window size $w$, corpus $C$ of documents
    in chronological order
Output: thread ids for each document
$T \leftarrow [\,]\,; i \leftarrow 0$
while document $d$ in $C$ do
    if $T$ is empty then
        $thread\_id(d) \leftarrow i$
        $i \leftarrow i + 1$
    else
        $d_{nearest} \leftarrow$ nearest neighbor of $d$ in $T$
        if $cosine(d, d_{nearest}) < t$ then
            $thread\_id(d) \leftarrow thread\_id(d_{nearest})$
        else
            $thread\_id(d) \leftarrow i$
            $i \leftarrow i + 1$
        end if
    end if
    if $|T| \geq w$ then
        remove first document from $T$
    end if
    add $d$ to $T$
end while
```

5.2. Evaluation Metric

The clustering performance is evaluated by a measure that we call "best matching F1". It is defined by Yang et al. (1998): once the clustering is done, the F1 score of each (cluster, event)[5] pair is computed. Each event is then matched to the cluster for which the F1 score is the best. Each event can only be associated with one cluster. The "best matching F1" score corresponds to the average of the F1 values of matching pairs.

5.3. Experiments

We test the FSD algorithm with different text embeddings. This subsection details the models used.

Tf-idf (Sparck Jones, 1972). Due to the inherent brevity of tweets, we simplified the calculation of tf-idf to a simple calculation of idf, since it is rare for a term to be used several times in the same tweet. The form used to calculate the weight of a term t in a tweet is therefore $idf(t) = 1+log(n+1/df(t)+1)$, where n is the total number of documents in the corpus and $df(t)$ is the number of documents in the corpus that contain t. We have distinguished two calculation modes for n and $df(t)$: **tfidf-dataset** denotes the method that counts only annotated tweets, and **tfidf-all-tweets** denotes the calculation method that takes into account all tweets in the corpus (38 million tweets) to obtain n and $df(t)$. For each method, we restrict the vocabulary with a list of *stop-words* and a threshold df_{min}, the minimum number of tweets that must contain t for it to be included in the vocabulary. In all our experiments, $df_{min} = 10$. We

[5] We call "clusters" the groups of documents formed by the clustering algorithm, and "events" tweets related to the same topic as defined by manual annotation.

thus obtain a vocabulary of nearly 330,000 words in English and 92,000 words in French for **tfidf-all-tweets**, and 5,000 words in English and 9,000 words in French for **tfidf-dataset**.

Word2Vec (Mikolov et al., 2013). We used pre-trained models for English, and trained our own French models. For each corpus, we distinguish between **w2v-twitter**, models trained on tweets, and **w2v-news**, models trained on press articles. For English, w2v-twitter is a pre-trained model published by Godin et al. (2015)[6] (400 dimensions) and w2v-news is a model trained on Google News and published by Google[7] (300 dimensions). In French, w2v-twitter was trained with the CBOW algorithm on 370 million tweets collected between 2018 and 2019, and w2v-news was similarly trained on 1.9 million AFP dispatches collected between 2011 and 2019. Both models have 300 dimensions. As Word2Vec provides word embeddings and not sentence embeddings, the representation of tweets is obtained by averaging the word vectors of each word. Two methods were used for averaging: a simple average, and an idf-weighted average using the **tfidf-all-tweets** calculation method.

ELMo (Peters et al., 2018). For English, we used the model published on TensorFlow Hub[8]. For French, a model trained on French published by Che et al. (2018)[9]. In each case, we use the average of the three layers of the network as a representation of each word. The representation of a tweet is produced by averaging these vectors (of dimension 1024).

BERT (Devlin et al., 2018). Google provides an English model and a multilingual model[10]. In order to improve the performance of the multilingual model on French tweets, we continued training for 150,000 steps on tweets collected in June 2018. We refer to the simple multilingual model as **bert** and the model trained on tweets as **bert-tweets**. In each case, we used the penultimate layer of the network (of dimension 768) as embedding, by averaging the tokens to obtain a tweet representation.

Universal Sentence Encoder (Cer et al., 2018). The provided models[11,12] (english and multilingual) are designed to provide sentence embeddings, so we were able to use them as is, without averaging vectors like in the previous representations. The vectors are of dimension 512.

5.4. Results

The results of the FSD algorithm are summarized in Table 3. The tf-idf (tfidf-all-tweets) vectors give the best results for both datasets. This is probably due to the shape of the tf-idf vectors, which are particularly suited to cosine similarity calculations, as well as to event characteristics in the datasets: the same terms are obviously widely used among

Model	English		French	
	t	*F1*	*t*	*F1*
bert	0.04	39.22	0.04	44.79
bert-tweets	-	-	0.02	50.02
elmo	0.08	22.48	0.20	46.08
tfidf-all-tweets	0.75	**70.10**	0.70	**78.05**
tfidf-dataset	0.65	68.07	0.70	74.39
use	0.22	55.71	0.46	74.57
w2v-news	0.30	53.99	0.25	66.34
w2v-news tfidf-weights	0.31	61.81	0.30	75.55
w2v-twitter	0.16	43.20	0.15	57.53
w2v-twitter tfidf-weights	0.20	53.45	0.25	71.73

Table 3: "best matching F1" score (in percentage) with the FSD algorithm for each embedding model and with the best threshold t. The value of the best threshold parameter is indicated in column t.

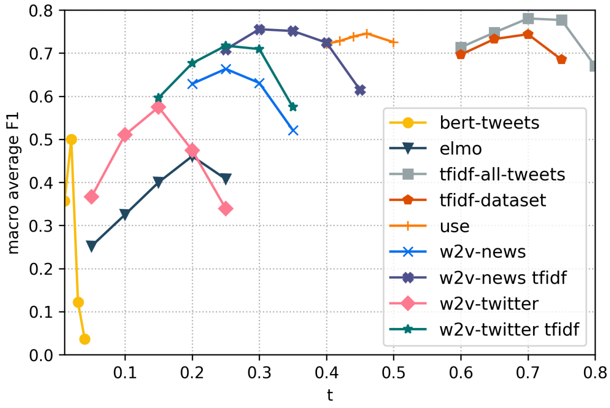


Figure 2: "best matching F1" score depending on the threshold parameter t on our corpus (tweets in French).

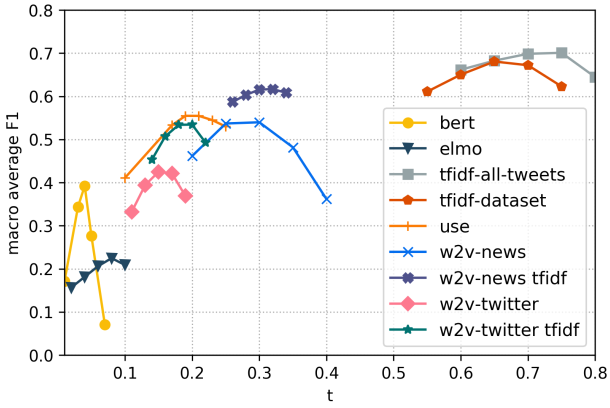


Figure 3: "best matching F1" score depending on the threshold parameter t on the McMinn et al. corpus (tweets in English).

[6] github.com/loretoparisi/word2vec-twitter

[7] code.google.com/archive/p/word2vec/

[8] tfhub.dev/google/elmo/2

[9] github.com/HIT-SCIR/ELMoForManyLangs

[10] github.com/google-research/bert models: bert-large, uncased and bert-base, multilingual cased

[11] tfhub.dev/google/universal-sentence-encoder-large/3

[12] tfhub.dev/google/universal-sentence-encoder-multilingual-large/1

the tweets of the same event. These results are consistent with those of Cagé et al. (2020), who came to similar conclusions regarding event detection in press articles.

The best deep network model (Universal Sentence Encoder), does not improve the results, and is still less effective than w2v-news model weighted by tf-idf. The best embeddings (tf-idf, Word2Vec with tf-idf weights, Universal Sentence Encoder) also have the property to be less sensitive to variations of the threshold t, as shown in Figures 2 and 3. Moreover, the optimal value of t for a given embedding seems to be approximately the same for each corpus (0.7 for tf-idf). This result may indicate that the First Story Detection algorithm could be applied to other (not annotated) tweet datasets without changing the threshold value.

6. Conclusion and Future Work

In this work, we present a French corpus of tweets annotated for event detection. We detail the corpus construction and annotation methodology, and provide information about the corpus characteristics. We then describe the results of our event detection experiments on this corpus and on another publicly available dataset. We show that with the First Story Detection algorithm, plain tf-idf vectors outperform more recent embedding techniques based on neural networks for this task.

This new dataset opens many possible research directions. It can be used as a baseline to test new event detection algorithms, including systems that would take into account the multimodal aspect of tweets (text, image, video, urls, mentions), or the graph of retweets and replies. This very large volume of tweets can also serve as a training corpus for language models.

7. Acknowledgements

We would like to thank Liza Héliès, Siegrid Henry and Antoine Moutiez for their great annotation work, and Denis Rakulan for developping the user interface's front- end. Many thanks also to Benjamin Ooghe-Tabanou, Dominique Cardon, Jean-Philippe Cointet and Guillaume Plique for sharing their dataset (Cardon et al., 2019).

This research was generously supported by the French National Research Agency (ANR-17-CE26-0004).

8. Bibliographical References

Allan, J., Lavrenko, V., Malin, D., and Swan, R. (2000). Detections, bounds, and timelines: Umass and tdt-3. In *Proceedings of topic detection and tracking workshop*, pages 167–174.

Cagé, J., Hervé, N., and Viaud, M.-L. (2020). The production of information in an online world. *The Review of Economic Studies*.

Cardon, D., Cointet, J.-P., Ooghe, B., and Plique, G. (2019). Unfolding the multi-layered structure of the french mediascape.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Che, W., Liu, Y., Wang, Y., Zheng, B., and Liu, T. (2018). Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 55–64, Brussels, Belgium, October. Association for Computational Linguistics.

Corney, D., Albakour, D., Martinez-Alvarez, M., and Moussa, S. (2016). What do a million news articles look like? In *NewsIR@ ECIR*, pages 42–47.

de Bruijn, J., de Moel, H., Jongman, B., and Aerts, J. (2017). Towards a global flood detection system using social media. In *EGU General Assembly Conference Abstracts*, volume 19, page 1102.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Godin, F., Vandersmissen, B., De Neve, W., and Van de Walle, R. (2015). Multimedia lab @ acl wnut ner shared task: Named entity recognition for twitter microposts using distributed word representations. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 146–153.

Hasan, M., Orgun, M. A., and Schwitter, R. (2016). TwitterNews: Real time event detection from the Twitter data stream. *PeerJ PrePrints*.

Hervé, N. (2019). OTMedia – the French transmedia news observatory. In *FIAT/IFTA Media Management Seminar.* www.herve.name/research/nherve_ina_fiat_stockholm_otmedia_preprint.pdf.

Mazoyer, B., Cagé, J., Hudelot, C., and Viaud, M. (2018). Real-time collection of reliable and representative tweets datasets related to news events. In *Proceedings of the First International Workshop on Analysis of Broad Dynamic Topics over Social Media (BroDyn 2018) co-located with the 40th European Conference on Information Retrieval (ECIR 2018), Grenoble, France, March 26, 2018*, volume 2078 of *CEUR Workshop Proceedings*, pages 23–34.

McGregor, S. C. and Molyneux, L. (2018). Twitter's influence on news judgment: An experiment among journalists. *Journalism*.

McMinn, A. J., Moshfeghi, Y., and Jose, J. M. (2013). 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.

Mele, I. and Crestani, F. (2019). A multi-source collection of event-labeled news documents. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 205–208. ACM.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*.

Pagolu, V. S., Reddy, K. N., Panda, G., and Majhi, B.

(2016). Sentiment analysis of twitter data for predicting stock market movements. In *2016 international conference on signal processing, communication, power and embedded system (SCOPES)*, pages 1345–1350. IEEE.

Papadopoulos, S., Corney, D., and Aiello, L. M. (2014). Snow 2014 data challenge: Assessing the performance of news topic detection methods in social media. In *Proceedings of the SNOW 2014 Data Challenge co-located with 23rd International World Wide Web Conference (WWW 2014)*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Petrović, S., Osborne, M., and Lavrenko, V. (2010). 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.

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

Repp, Ø. and Ramampiaro, H. (2018). Extracting news events from microblogs. *Journal of Statistics and Management Systems*, 21(4):695–723.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Suarez, A., Albakour, D., Corney, D., Martinez, M., and Esquivel, J. (2018). A data collection for evaluating the retrieval of related tweets to news articles. In *European Conference on Information Retrieval*, pages 780–786.

Wei, H., Zhou, H., Sankaranarayanan, J., Sengupta, S., and Samet, H. (2019). Delle: Detecting latest local events from geotagged tweets. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Analytics for Local Events and News*, pages 1–10.

Yang, Y., Pierce, T., and Carbonell, J. G. (1998). A study of retrospective and on-line event detection. In *Proc. of ACM-SIGIR*, pages 28–36.