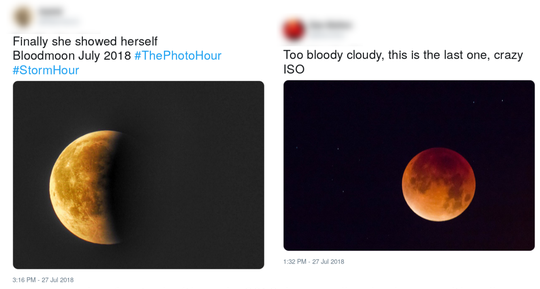
USING TEXT AND IMAGE FOR TOPIC DETECTION ON TWITTER

Béatrice Mazoyer[1-2], Nicolas Hervé[2], Céline Hudelot[1] and Julia Cagé[3]
[1] CentraleSupélec, MICS Laboratory
[2] Institut National de l'Audiovisuel
[3] Sciences Po, Department of Economics

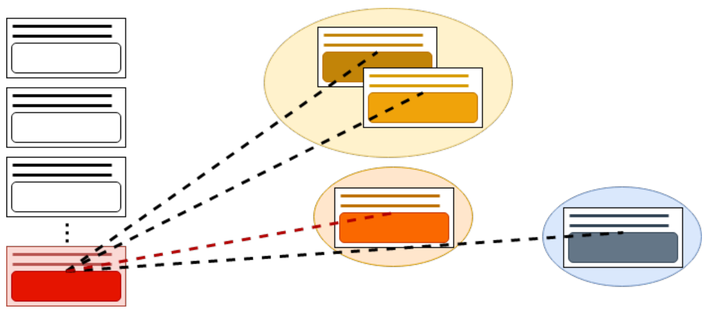
ABSTRACT

Social networks such as Twitter, Snapchat, Facebook and Instagram are popular platforms to gather and share information in the forms of short texts often associated with multimedia content (e.g. images or videos). However, in several social media analysis tasks, such as topic detection, this multimedia content tends to be overlooked. We explore various ways of using tweets pictures (in addition to text) for topic detection, either by creating **common representations of image and text,** or by **adapting the state of the art First Story Detection algorithm** to include several types of contents.

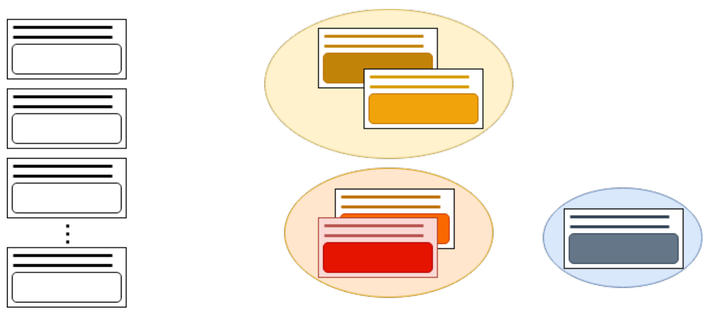


Example of a topic (the July 2018 lunar eclipse) where multimedia contents provide a critical information for topic detection

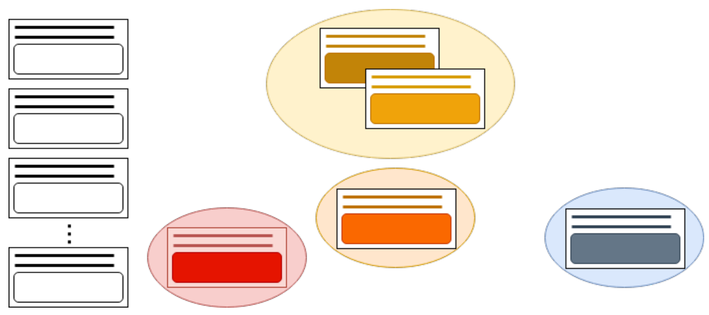
FIRST STORY DETECTION ALGORITHM



Documents are represented in a common vector space, and for each new document d, the algorithm looks for its nearest neighbor d′ among a window of the w previous documents.



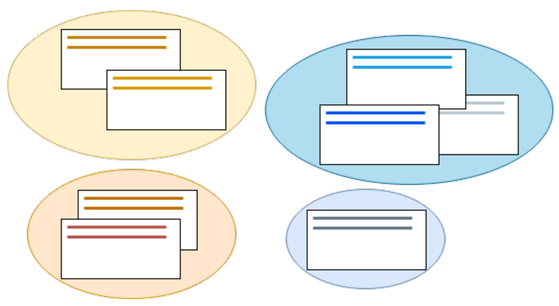
If the distance between d and d′ is lower than a pre-defined threshold t, d is attributed the same cluster as d′.



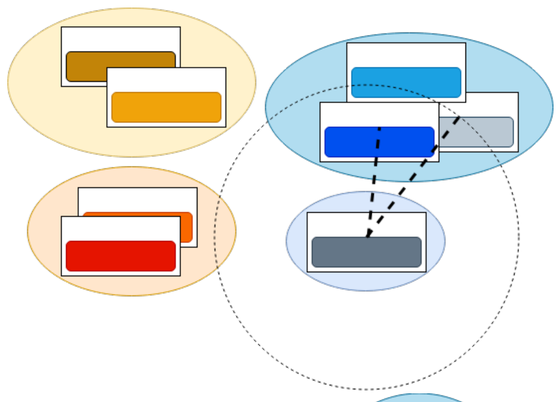
Else, a new cluster is created, that contains only d.

RECLUSTERING ALGORITHM

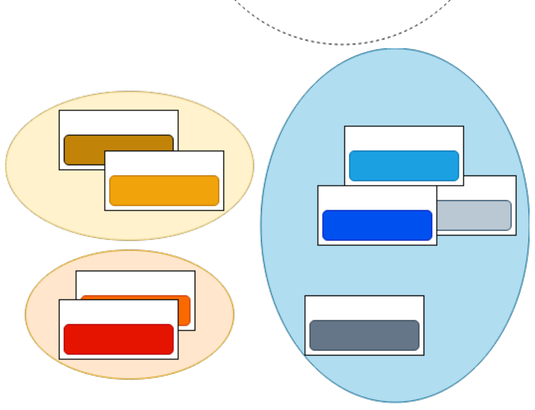
First clustering with the standard First Story Detection algorithm, using **only the textual content** to represent tweets.



For each cluster c, take all the documents in c and find the neighbors at a distance lower than a pre-defined threshold t_1. This time, we use **only the visual content** to represent tweets.



If a proportion p_1 of these neighbors is part of the same cluster c′, c and c′ are merged. This is repeated until the number of clusters stops decreasing.



TESTED EMBEDDINGS

IMAGE

-**ResNet Layer**: we use the penultimate layer of the ResNet50 network as the vector representation of images. Dimension of the image embedding: 2048.

-**SIFT features**: all images are described with SIFT features, which are then compressed to 128-bit binary hash codes. The distance between any two features can then be efficiently approximated by the Hamming distance between the hash codes. For each query image, we return the approximate KNN of each query feature. We use the number of matching features to compute a distance metric between images.

TEXT

-**TF-IDF**: a TF-IDF model is computed over the entire set of tweets (not only annotated tweets).

-**Word2Vec**: several types of Word2Vec models trained on documents in French and English are tested. Dimension of the text embedding: 300.

-**ELMo**: we use pre-trained ELMo models in French and in English. Dimension of the text embedding: 1024.

COMMON EMBEDDING

We build a common vector of text and image by concatenating the ResNet vectors with each type of textual representation.

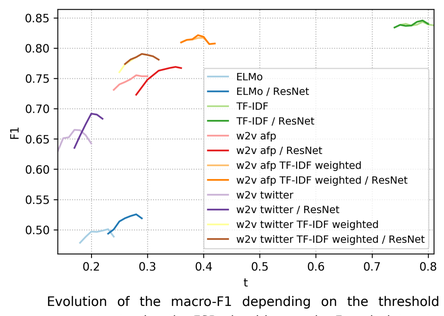
DATASETS

	Event2012	Our Corpus
Number of annotated tweets	150,000	88,000
Number of tweets containing images	570	17,000
Number of tweets in total corpus	120 million	40 million
Number of events	500	300
Collected in	Oct-Nov 2012	Jul-Aug 2018
Language of tweets	English	French

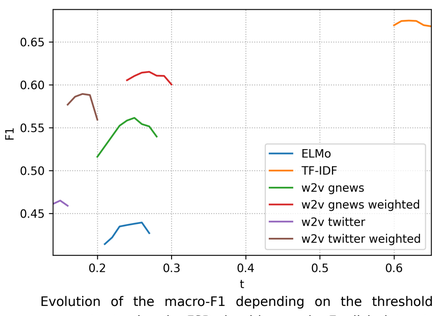
RESULTS

	Precision	Recall	F1
TF-IDF	**0.93**	**0.84**	0.84
TF-IDF - reclustering ResNet	0.92	0.77	0.80
TF-IDF - reclustering SIFT	0.97	0.74	0.81
ELMo - concatenated ResNet	0.80	0.50	0.53
W2V AFP - concatenated ResNet	0.87	0.78	0.77
W2V Twitter - concatenated ResNet	0.88	0.67	0.69
TF-IDF - concatenated ResNet	0.92	**0.84**	**0.85**

Best clustering results on the French dataset



Evolution of the macro-F1 depending on the threshold parameter t using the FSD algorithm on the French dataset



Evolution of the macro-F1 depending on the threshold parameter t using the FSD algorithm on the English dataset