



HAL
open science

Détection de fausses informations dans les réseaux sociaux : vers des approches multi-modales

Cédric Maigrot, Vincent Claveau, Ewa Kijak, Ronan Sicre

► To cite this version:

Cédric Maigrot, Vincent Claveau, Ewa Kijak, Ronan Sicre. Détection de fausses informations dans les réseaux sociaux : vers des approches multi-modales. EGC 2017 - Extraction et Gestion des Connaissances, Jan 2017, Grenoble, France. pp.1. hal-01844067

HAL Id: hal-01844067

<https://inria.hal.science/hal-01844067>

Submitted on 19 Jul 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

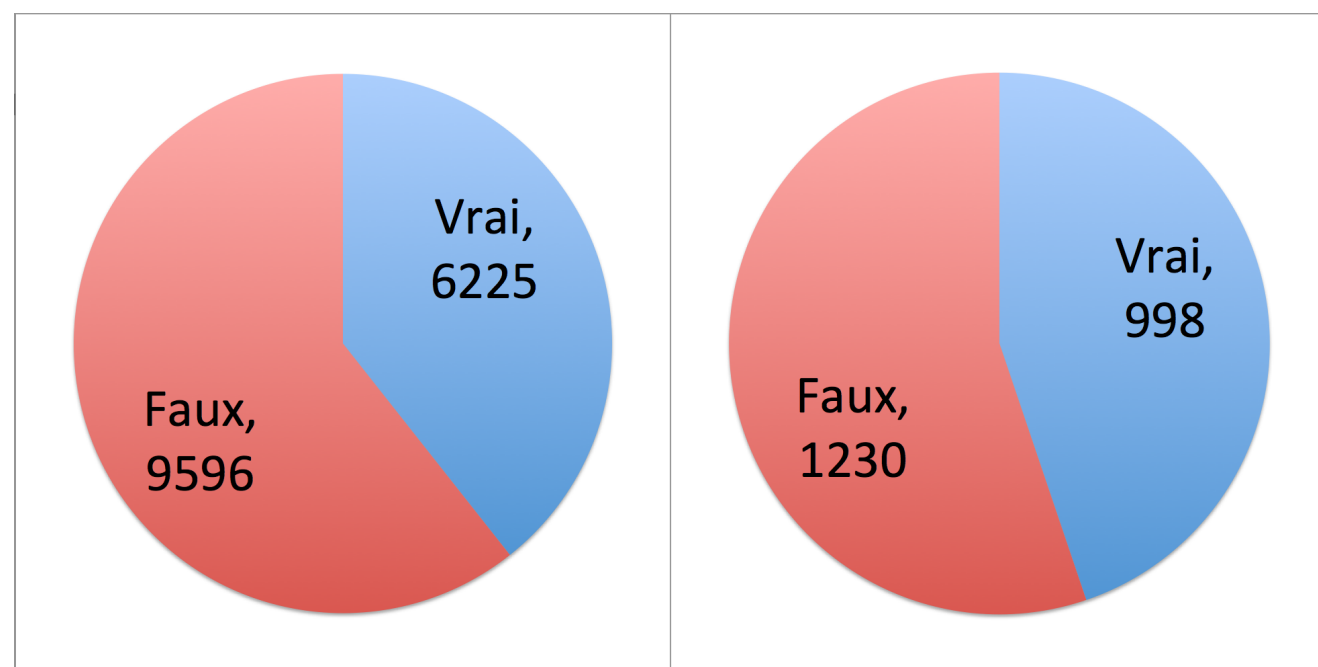
L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



DÉTECTION DE FAUSSES INFORMATIONS DANS LES RÉSEAUX SOCIAUX : VERS DES APPROCHES MULTI-MODALES

Cédric Maigrot Vincent Claveau Ewa Kijak Ronan Sicre
{Prénom} . {Nom}@irisa.fr

Tâche



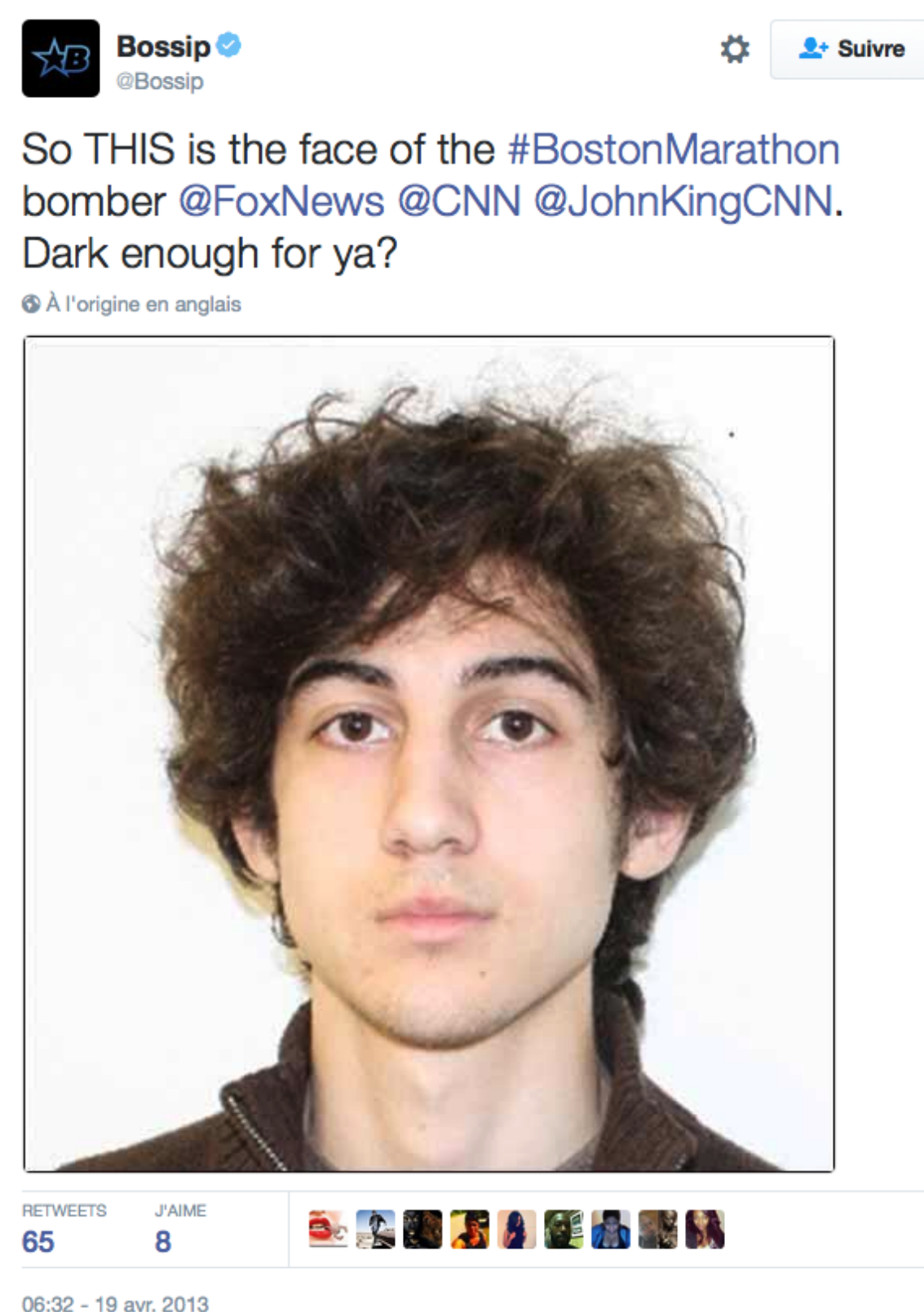
Répartition des messages selon leur véracité dans l'ensemble d'entraînement (gauche) et de test (droite)

- Chaque message est accompagné d'un contenu multimédia (image ou vidéo)
- Utilisation d'une image ou vidéo par un maximum de 580 messages

Approche basée sur les sources (run-S)

Détecte si le message est lié à une source de confiance

- Deux types de sources recherchées : les organismes liés aux actualités (e.g. agence de presse) et les sources explicites de l'image (e.g. le motif *photographed by + Nom*) [2]
- Prédit *vrai* si une source de confiance est détectée, *faux* sinon



Combinaison des prédictions (run-C)

- Fusion tardive : apprentissage de la meilleure combinaison
- Algorithme de Boosting (adaboost.MH, les paramètres de l'algorithme sont appris par validation croisée sur les données de l'ensemble d'entraînement)

Analyse

- Approche textuelle : résultats proches de ceux de l'approche basée sur les sources au niveau du score de rappel mais tend à classer chaque tweet comme *faux*
- Approche basée sur les images : faible précision comparée aux estimations réalisées sur l'ensemble d'entraînement. Causes : (1) la faible taille de la base d'images de référence; (2) la ressemblance entre les images originales et les versions modifiées; (3) la présence de tampons sur certaines images
- Combinaison des prédictions : ne permet pas d'augmenter les résultats du fait d'un surapprentissage

Approche textuelle (run-T)

Détecte si le message possède un style d'écriture typique des hoax

- Repère les commentaires similaires entre l'image à classer et les images de l'ensemble d'entraînement (e.g. *It's photoshopped*) et des caractéristiques de commentaires similaires (e.g. présence d'émoticônes)
- Prédiction réalisée par une approche *k-Plus-Proche-Voisin* (ici $k = 1$)



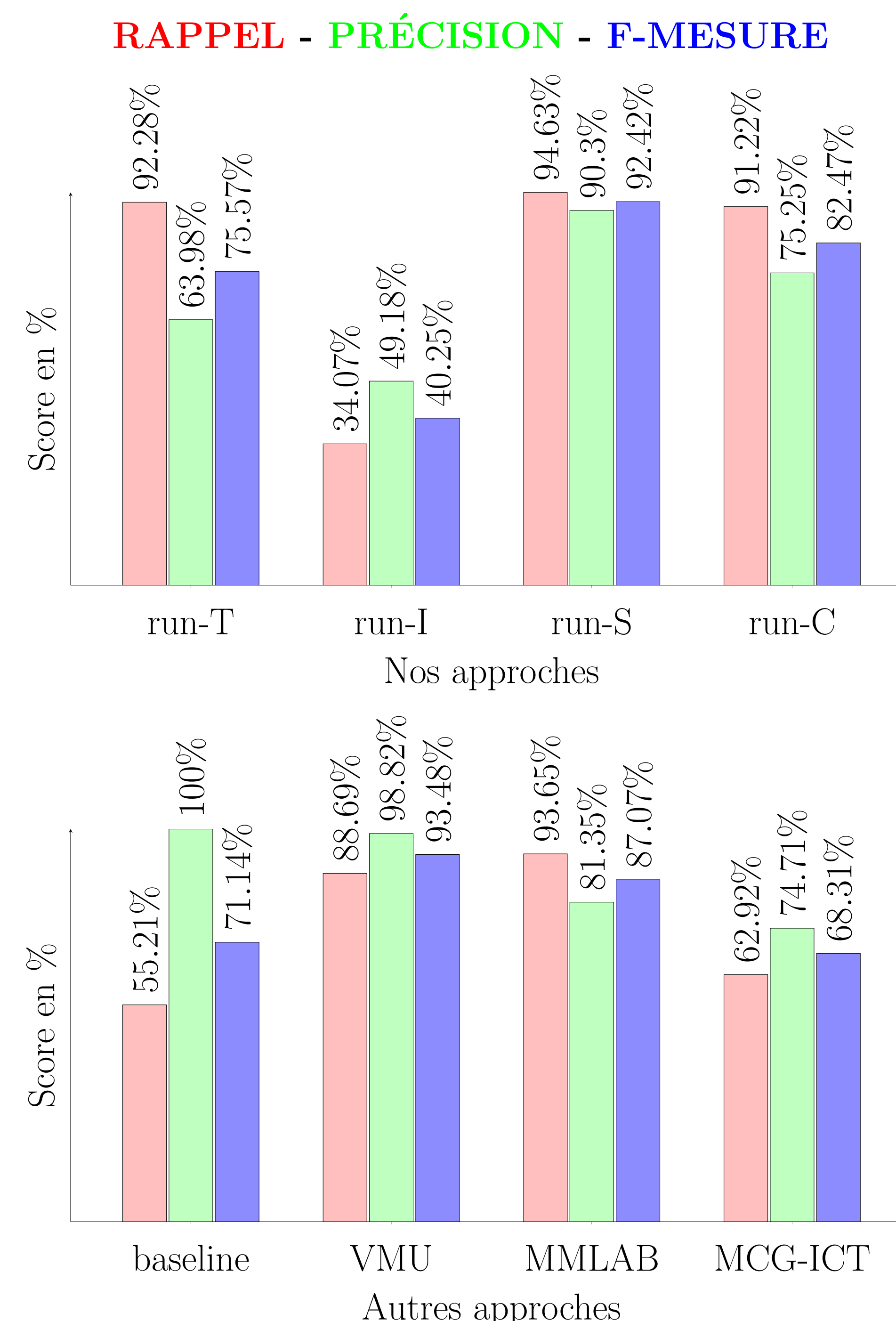
Approche basée sur les images (run-I)

Détecte les images connues

- Compare l'image à classer à une base d'image de 8 000 images connues (7 500 fausses and 500 images vraies)
- Base d'image est construite à partir de 5 sites spécialisés
- Description par une sortie d'une couche d'un CNN (vecteur de description de dimension 4096) [3]
- Prédit *vrai* (resp. *faux*) si une image similaire *vraie* (resp. *fausse*) est trouvée dans la base, *inconnu* sinon



Résultats



Perspectives

- confrontation de nos approches à celles employées par les autres équipes en étudiant le gain de prédiction lors de l'utilisation de toutes ces approches en même temps
- amélioration de la base d'images connues en collectant les sujets tendances sur Twitter ainsi que les articles d'actualité publiés par plusieurs sources d'informations
- mise en place de techniques de post-traitement pour détecter les zones de modification dans le but de différencier les images originales de leurs versions modifiées

Références

[1] Boididou C., Papadopoulos S., Dang-Nguyen D.-T., Boato G., Riegler M., Middleton S. E., Andreadou K., Kompatsiaris Y., "Verifying multimedia use at MediaEval 2016", *MediaEval 2016 Workshop*

[2] Middleton S., "Extracting attributed verification and debunking reports from social media : mediaeval-2015 trust and credibility analysis of image and video", *Mediaeval 2015 Workshop*

[3] Simonyan K., Zisserman A., "Very deep convolutional networks for large-scale image recognition", *Computing Research Repository (CRR)*, 2014