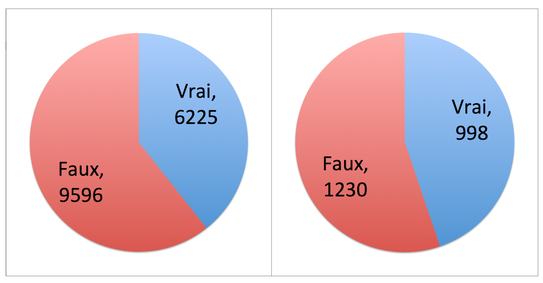
Détection de fausses informations dans les réseaux sociaux : vers des approches multi-modales

Cédric Maigrot, Vincent Claveau, Ewa Kijak, Ronan Sicre

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



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

- ➙ *Verifying Multimedia Use* (VMU) - Campagne d'évaluation *Mediaeval* 2016 *[1]*
- ➙ But : classer des messages provenant de *Twitter* selon leur véracité en trois classes : *vrai*, *faux* ou *inconnu*.
- ➙ Évaluation sur le score de *F-Mesure* de la classe *faux*
- ➙ Chaque message est accompagné d'un contenu multimédia (image ou vidéo)
- ➙ Utilisation d'une image ou vidéo par un maximun de 580 messages

Approche textuelle (run-T)

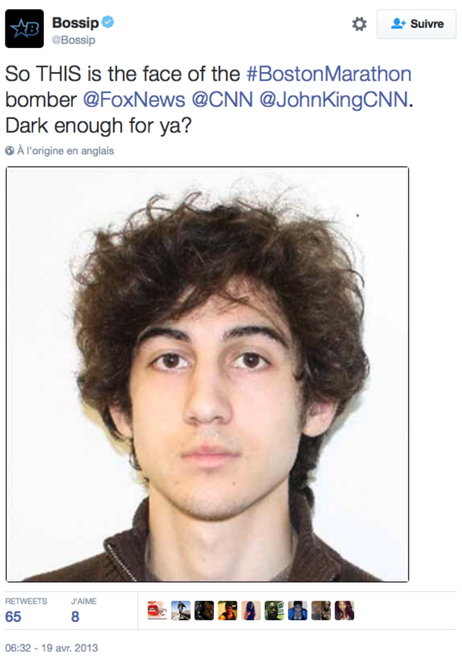
Détecte si le message possède un style d'écriture typique des hoax

- ➙ Repère les commentaires similaires entre l'image à classer et les images de l'ensemble d'entrainement (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 sources (run-S)

Détecte si le message est lié à une source de confiance

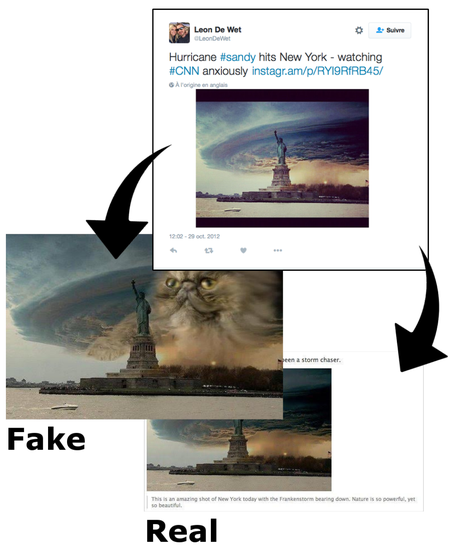
- ➙ Deux types de sources recherchées : les organismes liées aux actualités (e.g. agence de presse) et les sources explicites de l'image (e.g. le motif *photographed by + Nom*) *[2]*
- ➙ Prédit *vrai* si une source de confiance est détectée, *faux* sinon



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



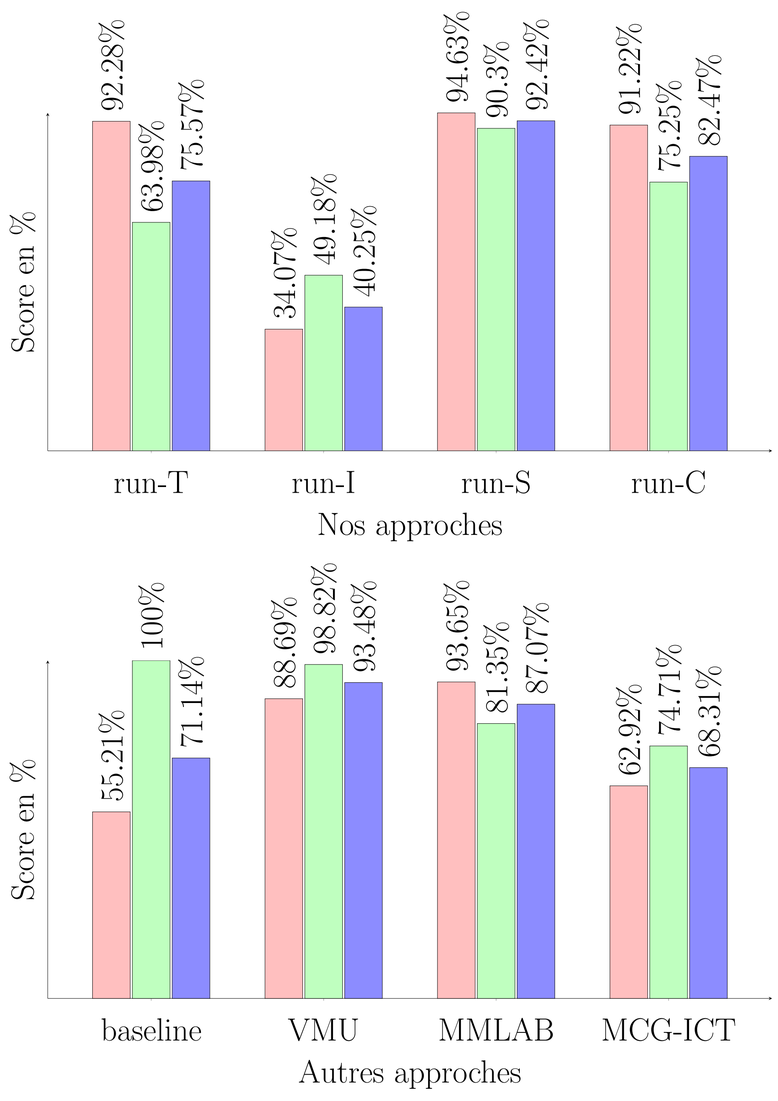
Combinaison des prédictions (run-C)

- ➙ 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'entrainement)

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'entrainement. 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

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 différencer 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