Détection des fausses informations dans les réseaux sociaux

Cédric Maigrot, Ewa Kijak, Vincent Claveau

Détection des fausses informations dans les réseaux sociaux

Cédric Maigrot, Ewa Kijak et Vincent Claveau
Cedric.Maigrot@irisa.fr
IRISA - DGA - Université Rennes 1 - CNRS

Problématique

- Utilisation croissante des réseaux sociaux (en 2015) :
  - Total sur les réseaux sociaux : 2 milliards d'utilisateurs
  - Facebook : 1 milliard et demi d'utilisateurs
  - Twitter : 300 millions d'utilisateurs
- Volume important de messages écrits :
  - Facebook : 4100 statuts partagés chaque seconde
  - Twitter : 500 millions de tweets envoyés chaque jour
  - Incapacité de vérifier manuellement la véracité des messages

Objectif : Déterminer automatiquement la véracité d'un message et justifier la décision

Exemples de détournements

#1

En réalité, il s'agit d'une photo du vol 1549 US Airways

Recherche d'informations contextuelles

#2

En réalité, il ne s'agit pas du compte officiel de Saud Al-FAisal

Analyse de la source: Compte non vérifié par Twitter

#3

En réalité, il s'agit de pétards pendant un mariage

Recherche d'informations contextuelles

#4

En réalité, le texte est ajouté sur la photo

Recherche d'informations: Images similaires mais non identiques

#5

En réalité, cette lettre est fausse

Analyse de l'image: Reconnaissance automatique des caractères

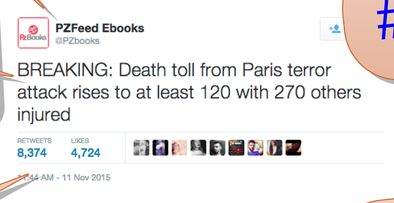
Analyse du texte: Reconnaissance d'entités nommées

#6

En réalité, il s'agit de personnes différentes

Vérification de la crédibilité: Comparaison avec une base de hoax connus

#7

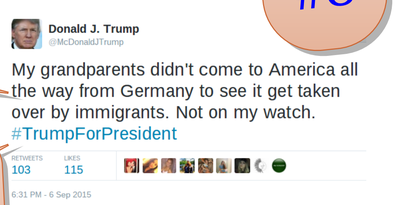


En réalité, ce compte publie des informations aléatoirement

Analyse de la source: Date antérieure à l'événement

Vérification de la crédibilité: Comparaison avec une base de hoax connus

#8



En réalité, il ne s'agit pas du compte officiel de Donald Trump

Analyse de la source: Compte non vérifié par Twitter

#9

En réalité, la mention "via iphone" a été ajoutée

Analyse de la source: Utilisation des méta-données envoyées par le réseau social

#10

En réalité, il s'agit d'une plaisanterie

Analyse de la source: Aveux du hoax dans un message plus récent de l'auteur

#11

En réalité, la personne présentée en photo est le boxeur Floyd Mayweather

Recherche d'informations contextuelles

Analyse de l'image: Reconnaissance faciale

#12

En réalité, cette photo a été prise à Londres devant un bureau de poste

Analyse de l'image: Détection de modifications dans l'image

Approches multimodales

- Analyse du texte, des images et de la source
- Recherche d'informations contextuelles basée texte et/ou image
- Apprentissage / classification
- Exemples d'informations extraites :
  - Détection des incohérences dans le texte [#1, #3, #5, #11]
  - Détection des modifications dans une image [#4, #5, #6, #11, #12]
  - Vérification de la crédibilité de la source [#2, #7, #8, #10]

[1] B. Tiziano, P. Alessandro, "Image Forgery Localization via Block-Grained Analysis of JPEG Artifacts", IEEE Transactions on Information Forensics and Security, Vol.7 No.3, p.1003-1017, June 2012

[2] C. Boididou, et al., "Verifying multimedia use at mediaeval 2015." Proceedings of the MediaEval 2015 Multimedia Benchmark Workshop. 2015.

Remerciements

Ces travaux sont soutenus par la Direction Générale de l'Armement (DGA), l'Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA) et l'Université de Rennes 1.