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

Cédric Maigrot[*,**] Vincent Claveau[*,***] Ewa Kijak[*,**]

[*]IRISA, {prenom}.{nom}@irisa.fr
[**]Université de Rennes 1 [***]CNRS

1 Introduction

Le projet dans lequel s'inscrit ce travail a pour but d'analyser automatiquement les informations partagées sur les réseaux sociaux, dans l'objectif de détecter les fausses informations. Partant du constat que ces dernières sont souvent composées d'éléments multimédias (texte accompagné d'images ou de vidéos), nous proposons un système multimodal. Nous présentons dans ce travail des approches exploitant le contenu textuel du message, les images associées et les sources citées dans les messages, ainsi qu'une combinaison de ces trois types d'indices. Les différentes approches proposées sont évaluées expérimentalement sur les données du challenge *MediaEval2016 Verifying Multimedia Use* [1], dont l'objectif est la classification en *vrai* ou *faux* de messages provenant du réseau *Twitter*.

2 Méthodologie

Le corpus de messages de la tâche *MediaEval2016 Verifying Multimedia Use* est divisé en un ensemble d'entrainement (15 821 messages) et un ensemble de test (2 228 messages). Par construction du corpus, les données présentent la propriété suivante : tous les messages partageant la même image ont la même classe. Il suffit donc de déterminer la classe de chaque image et de reporter sa prédiction sur les messages associés à cette image, selon la règle suivante : un message est prédit comme *vrai* si toutes les images associées sont classées *vraies*, *faux* sinon. Il est important de noter la distribution inégale de messages utilisant une image. La mauvaise classification d'une image n'aura pas le même impact sur les scores de classification des messages selon qu'elle soit partagée par beaucoup ou peu de messages.

Approche textuelle. Comme expliqué précédemment, la classe d'un message est déterminée à partir de la classe de l'image associée. Dans cette approche textuelle, une image est décrite par l'union des contenus textuels des messages qui utilisent cette image, puis classée par un classifieur entrainé sur l'ensemble d'apprentissage. L'idée à l'œuvre dans cette approche est de capturer les commentaires similaires entre une publication du jeu de test et celles du jeu d'entraînement (*e.g* "it's photoshopped") ou des aspects plus stylistiques (*e.g* présence d'émoticones, expressions populaires...).

1. Voir `http://multimediaeval.org/mediaeval2016/verifyingmultimediause/`

Approche basée sur la confiance des sources. La seconde approche , similaire à Middleton (2015), se base sur une connaissance (statique) externe. Comme dans l'approche précédente, une image est représentée par l'union des contenus textuels des messages où elle apparaît. La prédiction pour chaque image est faite par détection d'une source de confiance dans la description de l'image. Deux types de sources sont recherchés : 1) un organisme d'information connu ; 2) une citation explicite de la source de l'image. Si une source de confiance est trouvée dans sa description, l'image est classée *vraie*.

Approche basée sur la recherche d'images similaires. Dans cette approche, chaque image est utilisée comme requête pour interroger une base d'images de références, connues comme *fausses* ou *vraies*. Si il existe au moins une image similaire dans la base, l'image requête reçoit la classe de l'image la plus similaire. Sinon, l'image requête reçoit la classe *inconnu*. Les images sont décrites par un réseau de neurones convolutionnel pré-entrainé (Simonyan et Zisserman, 2014). Deux images sont considérées similaires si la similarité cosinus entre leurs descripteurs est supérieure à un certain seuil.

Combinaison des prédictions Cette dernière approche combine les prédictions des trois précédentes faites au niveau de l'image. Pour cette fusion, nous utilisons une approche par apprentissage artificiel, à savoir le *boosting* sur des arbres de décision (Laurent et al., 2014).

3 Résultats

Les modèles appris grâce aux données d'entrainement sont appliqués sur l'ensemble de test pour l'évaluation. Il est important de noter que les données d'entrainement et de test sont totalement distincts (i.e proviennent de rumeurs différentes). Les résultats présentés dans la table 1 correspondent aux mesures d'évaluation proposées dans la tâche de *MediaEval* à savoir la précision, le rappel et la F-mesure de la classe *faux*.

Méthode	baseline	texte	source	image	fusion	VMU	MMLAB	MCG
Précision	1	0,64	0,90	**0,97**	0,75	**0,99**	0,74	0,82
Rappel	0,55	0,92	**0,94**	0,12	0,91	0,88	0,94	0,94
F-mesure	0,71	0,76	**0,92**	0,21	0,83	**0,93**	0,83	0,87

TAB. 1 – *Performances des quatre qpproches proposées sur l'ensemble de test et comparaison à la meilleure soumission des autres participants à la tâche (VMU, MMLAB, MCG)*

Références

Laurent, A., N. Camelin, et C. Raymond (2014). Boosting bonsai trees for efficient features combination : application to speaker role identification. In *Proceedings of Interspeech 2014*.

Middleton, S. (2015). Extracting attributed verification and debunking reports from social media : mediaeval-2015 trust and credibility analysis of image and video. *Proceedings of the Mediaeval 2015 Workshop*.

Simonyan, K. et A. Zisserman (2014). Very deep convolutional networks for large-scale image recognition. *Processing of Computing Research Repository*.