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/`