Détection de fausses informations dans les réseaux sociaux : l'utilité des fusions de connaissances

Cédric Maigrot [1,2] — Ewa Kijak [1,2] — Vincent Claveau [1,3]

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

RÉSUMÉ. *Les réseaux sociaux permettent une diffusion massive et rapide des informations. Un des problèmes principaux de ces canaux de communication est l'absence de vérification associée à la viralité de l'information partagée. C'est ce problème difficile que les participants de la tâche* Verifying Multimedia Use *du workshop* Mediaeval *ont abordé. Pour cela, ils ont proposé plusieurs stratégies et types d'indices relevant de différentes modalités (texte, image, informations sociales). Dans cet article, nous explorons l'intérêt de combiner et fusionner ces approches pour évaluer le pouvoir prédictif de chaque modalité et tirer parti de leur éventuelle complémentarité.*

ABSTRACT. *Social networks make it possible to share rapidly and massively information. Yet, one of their major drawback comes from the absence of verification of the piece of information, especially with viral messages. This is the issue addressed by the participants to the* Verification Multimedia Use *task of* Mediaeval 2016*. They used several approaches and clues from different modalities (text, image, social information). In this paper, we explore the interest of combining and merging these approaches in order to evaluate the predictive power of each modality and in order to make the most of their potential complementarity.*

MOTS-CLÉS : *Détection de hoax, Fusion de connaissances, Analyse du texte, Analyse de l'image, Crédibilité de la source*

KEYWORDS: *Hoax detection, Knowledge fusion, Text analysis, Image analysis, Source credibility*

1. Introduction

Les réseaux sociaux tiennent une part croissante dans nos vies professionnelles ou personnelles, notamment par leurs capacités à nous tenir informés d'événements transmis par nos connaissances ou contacts. Il est devenu commun que des nouvelles importantes soient d'abord diffusées dans les réseaux sociaux avant d'être traitées par les médias traditionnels. Cette vitesse de propagation de l'information alliée au nombre de personnes la recevant définissent la viralité de l'information. Mais cette viralité, caractéristique majeure des réseaux sociaux, a un revers : les utilisateurs ne vérifient que rarement la véracité des informations qu'ils partagent. Il est donc commun de voir circuler des informations fausses et/ou manipulées (on parle alors d'*hoax*, de rumeurs, de légendes urbaines, ou de *fake*). De plus, une information catégorisée comme étant un hoax mais déjà partagée un grand nombre de fois peut être difficile à arrêter.

Le projet dans lequel s'inscrit ce travail a pour but de détecter automatiquement la véracité d'une information virale. Le but final est de créer par exemple un système qui préviendra l'utilisateur avant qu'il ne partage une fausse information. Partant du constat que ces informations virales sont souvent composées d'éléments multimédias (texte accompagné d'images ou de vidéos), nous proposons un système multimodal. Dans les travaux que nous présentons, nous proposons d'une part des approches exploitant le contenu textuel, le contenu des images ou les sources citées dans les messages, ainsi que des stratégies de combinaison de ces approches mono-modales. Ces différentes approches sont évaluées et comparées expérimentalement sur les données du challenge *MediaEval2016 Verifying Multimedia Use* portant précisément sur cette problématique mais aussi comparées aux soumissions des autres équipes participantes à cette tâche. D'autre part, à partir des soumissions de toutes les équipes participant à cette tâche de *MediaEval2016*, nous explorons de nouvelles fusions pour analyser la capacité de prédiction d'un système collaboratif.

Après une revue de l'état de l'art en section suivante, nous présentons dans la section 3 la tâche *Verifying Multimedia Use* (VMU) du workshop *Mediaeval* dont sont extraites les données utilisées dans ces travaux. Nous présentons ensuite dans la section 4 les systèmes mis en place ainsi que les systèmes proposées par les autres équipes participantes à la tâche *VMU*. Par la suite, la section 5 présente successivement les expérimentations et les résultats obtenus. Enfin, la section 6 synthétise les observations faites et évoque les pistes possibles pour l'avenir.

2. État de l'art

L'analyse de la véracité des informations est un axe de recherche qui est étudié dans le cadre de plusieurs projets. Nous ne nous intéressons ici qu'aux informations virales circulant dans les réseaux sociaux. Il convient de préciser que d'autres travaux, portant notamment sur le *fact checking* ne sont pas abordés ici ; même s'ils partagent un but commun de vérification, les différences de nature des informations (source,

mode de diffusion), et de finalité (aide au journalisme) impliquent des méthodes différentes de celles employées pour les hoax.

Le projet européen *PHEME* (Derczynski *et al.*, 2015) s'intéresse à la détection des rumeurs sur les réseaux sociaux ou les médias en ligne. Plusieurs travaux de ce projet étudient les liens entre messages sur les réseaux sociaux. Ces travaux s'intéressent aux réponses et réactions aux tweets pour en décider la véracité. Ce projet n'a pas pour objectif, comme nous dans cet article de classer le tweet sur la base de son unique contenu, mais vise plutôt, selon les auteurs, le *crowdsourced verification*. Cependant, les travaux menés au sein de ce projet sur la normalisation des hashtags ou la détection d'entités nommées (Declerck et Lendvai, 2015) peuvent être utiles pour la tâche ciblée dans cet article.

Dans le même axe, le projet européen *Reveal Project* (Middleton, 2015b), a pour objectif de développer des outils et des services de vérification de l'information dans les réseaux sociaux, selon une perspective journalistique et professionnelle. Différents médias tels que l'image (Zampoglou *et al.*, 2015), la vidéo (Middleton, 2015a), et le texte sont analysés. Il s'agit cependant de développer des outils non pas automatiques mais d'aide aux journalistes. Les travaux font par ailleurs extensivement usage de ressources externes (Gottron *et al.*, 2014).

Enfin, *InVid* [1] est un projet européen qui s'intéresse à la détection automatique de fausses vidéos, en travaillant sur des images issues des vidéos analysées (Foteini *et al.*, 2016). L'analyse des vidéos est importante et utile dans notre cas, cependant ces travaux ne sont pas applicable en tant que tel car leur approche ne s'intéresse pas aux textes diffusant ces vidéos et aux aspects réseaux sociaux à proprement parler.

Un autre angle d'étude des réseaux sociaux dans la littérature porte sur l'analyse des sources des messages et des relations entre membres. (Golbeck et Hendler, 2006) proposent ainsi une mesure de confiance entre utilisateurs des réseaux sociaux, qui caractérise la confiance d'une relation entre deux utilisateurs. Cette relation de confiance peut ainsi servir d'indice pour juger de la fiabilité des informations transmises. Plusieurs approches ont d'ailleurs été proposées afin de déterminer la crédibilité d'une source. (Gupta *et al.*, 2012) proposent une application de l'algorithme *PageRank* (Page *et al.*, 1999) sur un graphe représentant les relations entre les tweets, les auteurs de ces tweets et les événements associés à ces tweets. Ces approches nécessitent cependant une connaissance extensive du réseau qui les rendent difficilement applicables en pratique pour des réseaux sociaux commerciaux et grand public.

En ce qui concerne l'analyse des images circulant sur les réseaux sociaux afin de déterminer leur véracité, le problème est multiple. Une image peut avoir été modifiée intentionnellement (falsification), ou être utilisée pour illustrer un propos avec lequel elle n'a aucun rapport (détournement). Deux catégories d'approches pour aborder ces problèmes existent. D'une part, celles qui se basent sur une analyse des statistiques de l'image permettant de détecter des modifications. Par exemple, dans le cas des

1. Voir `http://www.invid-project.eu/`

images au format JPEG, il est possible de repérer une double compression (Bianchi et Piva, 2012) et donc une modification partielle de l'image. (Goljan *et al.*, 2011) se basent sur la connaissance de l'empreinte de l'appareil de capture d'une image, qui sera modifiée dans le cas d'une modification de l'image. L'autre catégorie d'approche utilise des informations externes à l'image pour déterminer son intégrité. Il s'agit dans ce cas de rechercher dans une base de données (ou le web) les images similaires ou identiques afin de déterminer si l'image a été modifiée ou détournée. Le problème de la recherche d'images similaires est un domaine actif dont les derniers travaux se basent sur des réseaux de neurones convolutionnels profonds pour décrire et comparer les images (Wan *et al.*, 2014). C'est cette dernière approche que nous utilisons en section 4.3.

3. Présentation de la tâche *Verifying Multimedia Use*

La tâche *Verifying Multimedia Use* (VMU) de la campagne d'évaluation *Mediaeval* en 2016, proposait de classer des messages provenant de *Twitter*[2] selon leur véricité entre les classes *vrai* et *faux* avec la possibilité d'utiliser une classe *inconnu* si le système ne permet pas de prendre de décision. Cela a pour but de garder une haute précision des deux autres classes sur les prédictions réalisées (Boididou *et al.*, 2016b). Par constitution de la base de données d'évaluation, tous les messages sont labelisés soit *vrai*, soit *faux*, et sont accompagnés soit d'une ou plusieurs images, soit d'une vidéo (cf. figure 1). Ainsi, aucun message n'est dépourvu d'un contenu multimédia. À l'inverse, plusieurs messages peuvent partager la même image. Ainsi, si certaines images ne sont utilisées que par un unique message, d'autres sont partagées par plus de 200 messages. De plus, les messages sont regroupés par événement. La taille des événements n'est pas équilibrée comme le montre la figure 2. Ainsi, le plus grand événement est *Paris Attack* avec 580 messages pour 25 contenus multimédias alors que les plus petits sont les événements *Soldier Stealing* et *Ukrainian Nazi* avec un unique message et une seule image. Le tableau 1 présente la répartition des données entre les ensembles d'apprentissage et de test, ainsi que le nombre d'images et de vidéos par ensemble.

Plusieurs descripteurs étaient proposés lors de cette tâche. Ces descripteurs étaient de trois catégories : textuel, utilisateur ou image.

Les descripteurs textuels proposés, noté $\mathcal{T}$, sont des descripteurs de surface : nombre de mots, longueur du texte, occurence des symboles *?* et *!*, présence des symboles *?* et *!* ainsi que d'émoticônes heureux ou malheureux, de pronoms à la première, deuxième ou troisième personne, le nombre de majuscules, le nombre de mots à opinion positifs et de mots à opinion négatifs, le nombre de mentions *Twitter*, de hashtags, d'urls et de retweets.

2. `https://twitter.com/`

Figure 1. *Exemple d'un tweet de la campagne MediaEval*

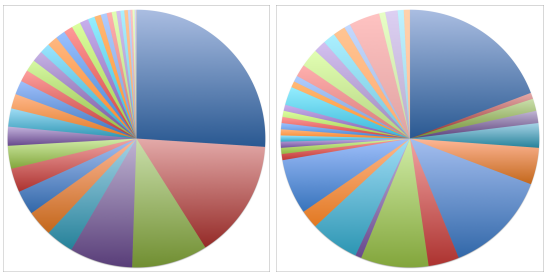

Figure 2. *Répartion des messages, à gauche, et des contenus multimédias (images et vidéos), à droite, par événement dans le jeu de test de la tâche* VMU *(35 événements)*

L'ensemble de descripteurs basé sur l'utilisateur, noté $\mathcal{U}$, est constitué des informations suivantes : nombre d'amis, nombre de *followers*, ratio du nombre d'amis sur le nombre de *followers*, si le compte contient une url, si le compte est vérifié et le nombre de messages postés.

L'ensemble des descripteurs images, noté $\mathcal{FOR}$, provient de méthodes issues du domaine des *forensics* : indices de double compression JPEG (Bianchi et Piva, 2012), Block Artifact Grid (Li *et al.*, 2009), Photo Response Non-Uniformity (Goljan *et al.*, 2011) et coefficients de Benford-Fourier (Pasquini *et al.*, 2014).

Notre équipe *Linkmedia* (noté par la suite *LK*) ainsi que trois autres équipes équipes (*MMLAB*, *MCG-ICT* et les organisateurs de la tâche : *VMU*) avons participé pour un total de 14 soumissions.

Ensemble d'apprentissage 15 821 messages				Ensemble de test 2 228 messages			
Événements : 17				Événements : 35			
Vrai 6 225 messages		Faux 9 596 messages		Vrai 998 messages		Faux 1 230 messages	
Images 193	Vidéos 0	Images 118	Vidéos 2	Images 54	Vidéos 10	Images 50	Vidéos 16

Tableau 1. *Description des ensembles d'apprentissage et de test pour la tâche VMU.*

4. Présentation des systèmes *Linkmedia*

Nous présentons dans cette section les approches que nous avons développées, puis brièvement les approches proposées par les autres équipes participantes à la tâche.

Dans notre approche, tous les messages partageant la même image ont la même classe *vrai*, *faux* ou *inconnu*. Il suffit donc de déterminer la classe de chaque image et de reporter la classe prédite 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. Nous proposons trois approches : la première est basée sur le contenu textuel de la publication ; la seconde sur les sources ; la troisième sur les images. Aucune n'utilise les descripteurs $\mathcal{T}$,$\mathcal{U}$ ou $\mathcal{FOR}$ présentés en section 3.

4.1. *Approche textuelle (`LK-T`)*

Cette approche exploite le contenu textuel des publications et ne fait pas appel à des connaissances externes supplémentaires. Comme expliqué précédemment, un tweet est classé à partir de l'image associée. Une image est elle-même décrite par l'union des contenus textuels des publications qui utilisent cette image. 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'apprentissage (*e.g* "it's photoshopped") ou des aspects plus stylistiques (*e.g* présence d'émoticones, expressions populaires…).

Soit I_q la description pour une image inconnue et $\{I_{d_i}\}$ l'ensemble des descriptions des images de l'ensemble d'apprentissage. La classe de I_q est décidée par vote des k images dont les descriptions sont les plus similaires dans $\{I_{d_i}\}$ (classification par les k-plus-proches voisins). Le calcul de similarité entre les descriptions textuelles est le cœur de cette approche. Pour cela, nous utilisons une technique considérée comme état-de-l'art en recherche d'information, à savoir la similarité Okapi-BM25 (Robertson *et al.*, 1998). Celle-ci calcule un score de *Retrieval Status Value* (RSV) en fonction des termes communs à une requête (dans notre cas le texte à classifier I_q) et à un document (ici un texte de $\{I_{d_i}\}$) ; voir équation 1.

$$\text{RSV}_{\text{Okapi}}(I_q, I_{d_i}) = \sum_{t \in I_q} qTF(t) * TF(t, I_{d_i}) * IDF(t) \quad [1]$$

$$\begin{aligned} qTF(t) &= \frac{(k_3+1) * qtf}{k_3 + qtf} \\ TF(t, I_{d_i}) &= \frac{tf * (k_1+1)}{tf + k_1 * (1 - b + b * \frac{dl(I_{d_i})}{dl_{avg}})} \\ IDF(t) &= \log \frac{n - df(t) + 0.5}{df(t) + 0.5} \end{aligned} \quad [2]$$

avec t un terme présent dans la requête, qtf le nombre d'occurrences du terme dans la requête, tf le nombre d'occurrences dans le document, dl_{avg} la taille moyenne des documents, n le nombre de documents dans la collection, et $df(t)$ le nombre de documents contenant le terme t. Les paramètres k_1, k_3 et b sont des constantes, avec des valeurs par défaut $k_1 = 2$, $k_3 = 1000$ et $b = 0,75$.

Un système de détection de la langue (basé sur le service *Google Translate*) est utilisé pour trouver et traduire les publications non écrites en anglais. Un autre pré-

traitement est la normalisation de l'orthographe et des smileys développé par l'équipe. Le paramètre du nombre de voisins k est déterminé à 1 par validation croisée sur l'ensemble d'apprentissage.

4.2. *Prédiction basée sur la confiance des sources (`LK-S`)*

Cette approche, similaire à (Middleton, 2015a), se base sur une connaissance (statique) externe. Comme pour l'approche précédente, la prédiction est réalisée au niveau de l'image et l'image est représentée par l'union des contenus textuels (traduits en anglais si nécessaire) des messages où elle apparaît. La prédiction 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. Pour le premier type de source, nous déterminons une liste d'agences de presse dans le monde, journaux (principalement francophones et anglophones), réseaux télévisuel d'information (francophones et anglophones). Pour le second type, nous définissons manuellement plusieurs patrons d'extraction, comme `photographed by + Name`, `captured by + Name`, ... Enfin, une image est classée comme *inconnue* par défaut sauf si une source de confiance est trouvée dans sa description.

4.3. *Recherche d'images similaires (`LK-I`)*

Dans cette approche, seul le contenu des images est utilisé pour réaliser une prédiction. Les tweets contenant des vidéos ne sont pas traités par cette approche et obtiennent la classe *inconnu*. Nous utilisons une approche de type recherche d'images similaires dans une base d'images de références, répertoriées comme *fausses* ou *vraies*. Une image requête donnée (dont on cherche la classe) reçoit la classe de l'image la plus similaire de la base (si elle existe). Sinon, l'image requête reçoit la classe *inconnu*.

La base de référence a été construite en collectant des images présentes sur cinq sites spécialisés dans le référencement de fausses informations : *www.hoaxbuster.com*, *hoax-busters.org*, *urbanlegends.about.com*, *snopes.com* et *www.hoax-slayer.com*. La base contient environ 500 images originales (*vraies*) et 7500 images *fausses*.

Les descripteurs de l'image sont calculés en utilisant un réseau de neurones convolutionnel profond (Simonyan et Zisserman, 2014). Les images sont d'abord redimensionnées à la taille standard de 544×544 et passées dans les couches convolutionnelles du réseau (Tolias *et al.*, 2016). Ensuite, les deux premières couches entièrement connectées sont mises sous forme de noyau et appliquées au tenseur de sortie, produisant un nouveau tenseur de dimension $11 \times 11 \times 4096$. Enfin, nous appliquons un filtre moyenneur et une normalisation $\mathcal{L}2$ qui nous permet d'obtenir un vecteur de description de dimension 4096. Une fois les descripteurs d'images obtenus, une similarité par cosinus est calculée entre les images requêtes et les images de la base.

Le système de recherche retourne donc une liste d'images ordonnées par similarité. Considérer que deux images sont suffisamment similaires nécessite de prendre une décision sur la similarité entre deux images. La décision est prise par rapport à un seuil de similarité de $0,9$ (déterminé de façon empirique sur l'ensemble d'apprentissage).

4.4. *Présentation des autres approches*

Trois autres équipes ont participé à la tâche *VMU*. Pour chaque prédiction, nous décrivons ci-dessous le type de données sur lequel est basée cette prédiction ainsi que l'approche utilisée.

4.4.1. *Équipe* VMU

Cinq méthodes de prédictions ont été proposées, qui reposent sur deux systèmes (Boididou *et al.*, 2016a), dont elles sont des variantes.

`VMU-F1` et `VMU-F2` sont basées sur un premier système qui est un méta-classifieur dans lequel deux ensembles de descripteurs sont utilisés séparément par deux classifieurs, entrainés sur l'ensemble d'apprentissage. Chaque classifieur prédit alors *vrai* ou *faux* pour chaque message, ce qui permet donc d'obtenir deux prédictions par message. Les messages prédits sur l'ensemble de test sont alors traités selon deux cas : accord entre les deux prédictions ou non. Les messages de l'ensemble de test ayant reçu des prédictions différentes sont alors analysés par un troisième classifieur entrainé sur l'union de l'ensemble d'entrainement et des messages de l'ensemble de test ayant reçu des prédictions en accord sur les deux premiers classifieurs. `VMU-F1` utilise les descripteurs $\mathcal{T}$ et $\mathcal{U}$ pour les deux premiers classifieurs, tandis que `VMU-F2` utilise l'union de $\mathcal{T}$ et $\mathcal{U}$ pour l'un des classifieurs, et $\mathcal{FOR}$ pour l'autre.

Les prédictions `VMU-S1` et `VMU-S2` utilisent un second système qui se base sur deux listes de sources connues : la première est une liste de sources de confiance alors que la seconde regroupe des sources de non-confiance. Toutes ces sources ont reçu une pondération de confiance ou non-confiance. Lorsqu'une prédiction basée sur les sources n'est pas possible, le premier système est utilisé pour fournir une prédiction.

Enfin, la dernière prédiction `VMU-B` est une baseline obtenue par l'application d'un classifieur sur la concaténation des descripteurs $\mathcal{T}$, $\mathcal{U}$ et $\mathcal{FOR}$.

4.4.2. *Équipe* MMLAB

La prédiction `MML-T` représente la sortie du module de traitement de texte proposé par (Phan *et al.*, 2016). Les auteurs ne précisent pas l'utilisation qu'ils font des ensembles de descripteurs $\mathcal{T}$ et $\mathcal{U}$.

Le second module est celui associé aux contenus multimédias (images et vidéos). Pour cela, les auteurs appliquent un traitement semblable aux images et aux vidéos. Pour les images, un moteur de recherche inversé est utilisé et une mesure de fréquence (*TF-IDF*) est appliquée sur les textes des sites les plus pertinents retournés par le

moteur pour obtenir un ensemble de mot-clés associés à l'image. Dans le cas d'une vidéo *Youtube*, la mesure de fréquence est appliquée aux commentaires de la vidéo. Les autres vidéos ne sont pas analysées. Un traitement supplémentaire est appliqué aux images, orienté sur la détection de modifications dans l'image en se basant sur les descripteurs $\mathcal{FOR}$. La prédiction `MML-I` est la sortie de ce deuxième module.

Enfin `MML-F` est la fusion sur les sorties de `MML-T` et `MML-I` avec des coefficients respectifs de $0,2$ et $0,8$ afin de favoriser le second module mais aussi assurer une prédiction dans le cas d'une incapacité du second module à prédire (e.g. vidéo ne provenant pas de *Youtube*).

4.4.3. *Équipe* MCG-ICT

La première approche proposée par (Cao *et al.*, 2016) se base sur le contenu textuel des messages. L'idée associé au système proposé est d'une part de réunir les messages partageant le même événement (indication donnée par la tâche) mais aussi de séparer ces événements en sous-événements par un classifieur non supervisé de type k-means et en posant l'hypothèse que les messages contenus dans un même sous-événement obtiennent la même classe. En plus des descripteurs $\mathcal{T}$ proposé par la tâche, les auteurs ajoutent les descripteurs suivants : nombre de tweets dans le sous-événement (cluster), nombre de tweets distincts (afin de discriminer les retweets), ratio de tweets distincts, ratio de tweets contenant une URL ou une mention et enfin le ratio de tweets contenant plusieurs URLs, mentions, hashtags ou points d'interrogation. Ce module correspond à la prédiction `MCG-T`.

Le second module traite les images et les vidéos différemment. Les auteurs utilisent les descripteurs $\mathcal{FOR}$, sans préciser le traitement qu'ils en font. Les vidéos sont traitées selon un arbre de décision construit manuellement selon (Silverman, 2014). Les descripteurs utilisés sont : présence de logos, la longueur de la vidéo, la résolution de la vidéo, le nombre de coupures brusques dans la vidéo, ratio de résolution et ratio de contraste. La prédiction `MCG-I` correspond à cette approche.

Enfin, `MCG-F` est une fusion basée sur ces deux prédictions précédentes.

5. Expérimentations et discussions

5.1. *Protocole expérimental*

Les données utilisées pour évaluer ces systèmes sont celles issues de l'ensemble de test de la tâche *VMU* présentée dans la section 3 (cf. tableau 1). Nous n'utilisons cependant pas la même mesure d'évaluation que la tâche (i.e. *F-Mesure* sur la classe *faux*) car cette dernière n'est pas discriminante entre les prédictions des messages *vrais* et *inconnus*. De plus, cette mesure se base sur la classe majoritaire *faux*, ce qui représente un biais (i.e. *F-Mesure* sur la classe *faux* de $71,14$ % sur l'ensemble de test en prédisant tout le temps *faux*). Nous utilisons à la place la micro-*F-Mesure* et le taux de bonnes classifications qui sont des mesures globales sur l'ensemble des

classes à prédire. Une image pouvant être utilisée par plusieurs messages, l'évaluation est faite par validation croisée sur les événements, de sorte à garantir que tous les messages utilisant une même image se retrouvent dans le même pli afin de ne pas biaiser l'évaluation.

Les résultats des méthodes décrites en section 4, ré-évalués selon le protocole décrit ci-dessus, sont présentés dans le tableau 2.

	$F\text{-}Mesure$	$Taux_{B.C}$		$F\text{-}Mesure$	$Taux_{B.C}$
LK-T	**68,95 %**	$67,19\,\%$	VMU-F1	$91,28\,\%$	$90,75\,\%$
LK-I	$11,54\,\%$	$6,46\,\%$	VMU-F2	$80,26\,\%$	$79,89\,\%$
LK-S	$61,94\,\%$	**75,56 %**	VMU-S1	**93,37 %**	**93,04 %**
			VMU-S2	$92,28\,\%$	$91,92\,\%$
			VMU-B	$67,59\,\%$	$61,27\,\%$
	$F\text{-}Mesure$	$Taux_{B.C}$		$F\text{-}Mesure$	$Taux_{B.C}$
MML-T	$63,32\,\%$	$63,43\,\%$	MCG-T	$67,59\,\%$	$61,27\,\%$
MML-I	$69,89\,\%$	$62,34\,\%$	MCG-I	$61,82\,\%$	$60,73\,\%$
MML-F	**79,60 %**	**78,55 %**	MCG-F	**69,83 %**	**68,09 %**

Tableau 2. *Performances des soumissions des équipes* Linkmedia *(LK),* VMU*,* MM-LAB *(MML) et* MCG-ICT *(MCG) à la tâche VMU selon le taux de bonne classification et la micro-F-Mesure*

5.2. *Fusion des soumissions*

Une fusion directe des 14 prédictions du tableau 2 est d'abord réalisée dans cette partie. Pour réaliser cette fusion, nous utilisons pour décrire chaque message les prédictions *vrai*, *faux* ou *inconnu* des différents systèmes, afin d'apprendre une combinaison des prédictions. Les fusions des prédictions sont réalisées par trois algorithmes de classification : 1) SVM linéaire ; 2) arbre de décision ; 3) Random Forest (avec 100 arbres). Ces trois classifieurs sont comparés à une baseline correspondant au vote majoritaire sur les 14 prédictions (i.e. parmis les 14 prédictions, la classe prédite la plus fréquemment est associée au message). Les résultats sont présentés dans le tableau 3.

On note que la baseline ne permet pas de surpasser les meilleures prédictions à la tâche, contrairement aux classifieurs utilisant les 14 prédictions. Cela montre que toutes les prédictions n'ont pas la même importance et que les classifieurs permettent d'apprendre des pondérations. Le meilleur classifieur (*SVM linéaire*) permet une augmentation de 2 points du taux de bonne classification, ce qui correspond à 45 messages supplémentaires bien classés sur les 2228 messages.

Il est important de noter que 140 messages ont été identifiés comme difficiles à classer (i.e. une grande majorité des classifieurs prédisent la mauvaise classe). Une fusion aura alors de grande chance de se baser sur cette gande majorité pour prendre sa décision ce qui engendrera une erreur de prédiction. On peut alors poser une limite

Fusion directe	Baseline	SVM	Arbre de déc.	Random Forest
F-Mesure	$79,12\ \%$	**95,50 %**	$94,49\ \%$	$95,03\ \%$
Taux de B.C.	$80,12\ \%$	**95,47 %**	$94,48\ \%$	$95,02\ \%$

Tableau 3. *Fusion basée sur les* 14 *prédictions soumises à la tâche VMU*

maximum de $2\ 088$ messages sur les $2\ 228$ pouvant être classés, ce qui représente un taux de bonne classification de $93,72\ \%$. Or nous avons vu que d'après la fusion directe des différentes prédictions, nous obtenons le score de $95,47\ \%$ ce qui tend à montrer la difficulté d'obtenir un score plus élevé. Ces messages difficiles à classe présentent deux caractéristiques communes pouvant expliquer cette difficulté : 1) des urls réduites qui cachent la source citée (e.g. utilisation du site *goo.gl*) ; 2) une grande partie de ces messages proviennent des événements *Paris attacks* et *Fugi Lenticular* qui sont des événements ayant des messages *vrais* et *faux* et sont donc ambigus.

Nous avons vu que les approches pouvaient se compléter afin d'améliorer les scores de prédiction. Cependant la fusion proposée utilise l'intégralité des prédictions alors que l'information véhiculée par chaque classifieur peut être redondante (e.g. les prédictions `MCG-T` et `MCG-I` influent sur la prédiction `MCG-F`). Par ailleurs, nous n'obtenons aucune information sur les apports de chaque approche lors de la fusion directe.

5.3. *Comparaison des différentes approches*

Afin d'analyser et tester la complémentarité des différentes approches et des différents type de données (texte, source ou image), nous considérons dans la suite uniquement les prédictions basées sur le même type de données, en excluant les prédictions faisant intervenir des fusions entre modalités. Ainsi seules les prédictions `LK-T`, `LK-I` et `LK-S` seront gardées parmis nos prédictions, `MML-T`, `MML-I`, `MCG-T` et `MCG-I` pour les prédictions des équipes *MMLAB* et *MCG-ICT*. Enfin, les prédictions de l'équipe *VMU* sont toutes basées sur de la fusion (cf. section 4.4). Nous retenons cependant `VMU-S1` qui basée sur les sources et qui est la prédiction obtenant les meilleurs performances. Ces huit prédictions, notées *élémentaires*, seront utilisées dans la suite.

5.3.1. *Approches textuelles*

Trois prédictions peuvent être associées à une approche textuelle : `LK-T`, `MML-T` et `MCG-T`. La prédiction `LK-T` tend à classer tous les messages comme *faux*, ce qui peut s'expliquer par le fort déséquilibre des classes dans l'ensemble d'apprentissage (trois fois plus de messages *faux* que *vrais*) sur lequel le classifieur est appris. Ainsi, *636* messages réels sont classés comme étant *faux*. À l'inverse, les prédictions `MML-T` et `MCG-T` ont tendance à plus se tromper sur la classification des messages *faux* classés comme *vrais* (i.e. respectivement 557 et 457 messages *faux* sur les 1230 sont classés *vrais*). On peut aussi noter une différence entre ces trois prédictions quant aux descrip-

teurs utilisés. Alors que les prédiction MML-T et MCG-T se portent sur des descripteurs de surface, ou descripteurs statistiques, (essentiellement l'ensemble de descripteurs $\mathcal{T}$), la prédiction LK-T utilise des descripteurs de contenu (i.e. des patterns précis dans le texte). Ces prédictions sont donc possiblement adaptées à une fusion afin recouper ses capacités de prédictions différentes.

5.3.2. *Approches basées sur les sources*

Deux prédictions sont identifiées comme utilisant des sources : LK-S et VMU-S1. Alors que les deux approches se basent sur une liste de sources de confiance, la prédiction VMU-S1 considère en plus une source de non-confiance. On peut noter que les deux listes de source de confiance n'étant pas identiques, ces dernières peuvent donc se compléter. Une seconde différence se fait quant au choix de la classe à attribuer en cas d'absence de source. Alors que VMU-S1 choisit la classe *faux*, qui est la classe majoritaire de l'ensemble d'apprentissage, la prédiction LK-S fait le choix de la classe *inconnu* qui donnera obligatoirement un message mal classé (puisqu'aucun message ne possède réellement cette classe) mais qui permet une haute précision des messages classés comme *vrai* ou *faux* (respectivement $100.00\ \%$ et $92,97\ \%$) aux dépens du rappel (respectivement $41,22\ \%$ et $87,47\ \%$).

5.3.3. *Approches basées sur les contenus multimédias*

Les approches multimédias sont les plus diversifiées. On compte trois prédictions dans lesquelles les images et/ou les vidéos sont utilisées : LK-I, MML-I et MCG-I.

Ainsi même si les approches multimédias présentent les résultats les plus faibles individuellement, elles ont une plus grande capacité de complémentarité par une fusion car elles sont très différentes sur leur approche. LK-I recherche les images répertoriées comme étant *fausses* ou *vraies* dans une base d'images de référence et ne se prononce que lorsque l'image associée à un message a été retrouvée. Cela ne permet de classer que peu de messages (170 messages sur les 2228) mais d'obtenir une précision élevée ($97,30\ \%$ sur la classe *faux*). Les messages pour lesquels aucune image similaire n'a été trouvée obtiennent la classe *inconnu*. De plus, tous les messages ayant pour illustration une vidéo reçoivent également la classe *inconnu*. MCG-I est la seule approche à proposer un traitement sur les vidéos alors que les messages accompagnés par une vidéo représentent $48,43\ \%$ du jeu de données. Tout comme LK-I, cette soumission contient des prédictions associées à la classe *inconnu*. Plusieurs phénomènes peuvent expliquer les faibles performances des systèmes. Premièrement, dans le cas d'une différence légère entre l'image originale (réelle) et l'image modifiée (fausse), les images peuvent être confondues par le système de recherche car elles seront très similaires. Cela impactera directement les soumissions LK-I et MML-I qui recherchent des images similaires dans des bases de connaissances. Deuxièmement, les images référencées sur les sites spécialisés sont parfois altérées : il peut s'agir par exemple de l'ajout d'un "tampon" 'faux', 'rumeur' ou 'vrai', ou de modifications afin d'améliorer la compréhension (*e.g* un cercle rouge sur la zone photoshoppée). De même, les images diffusées sur les réseaux sociaux subissent également souvent ce même

type d'édition. Ces modifications font décroître la similarité entre l'image requête et l'image de la base, et de ce fait dégradent les performances du système (cf. Fig 3).

Figure 3. *Exemple d'une image requête (à gauche) ayant un vrai positif dans la base (à droite) qui n'a pas été retrouvé par la recherche d'images similaires, les artefacts d'édition de l'image requête faisant chuter le score de similarité entre ces 2 images.*

Au vu des résultats des approches basées sur les images, il semble que l'utilisation d'une recherche d'images similaires (prédiction `MML-I`) apporte plus d'information que l'utilisation des descripteurs $\mathcal{FOR}$ (prédiction `MCG-I`). De plus, les prédictions `VMU-F1` et `VMU-F2` (voir section 4.4) diffèrent principalement par l'utilisation ou non de l'ensemble de descripteurs $\mathcal{FOR}$. L'utilisation de ce dernier amenant à une baisse des scores de prédiction (tableau 2). Cependant aucune approche ne propose de pré-traitements ou de post-traitements sur la comparaison des images similaires trouvées. Il serait intéressant de voir dans quelle mesure les descripteurs $\mathcal{FOR}$ pourraient aider de tels pré-traitements ou post-traitements (e.g capacité supplémentaire de vérification des contenus similaires retrouvés et détection des modifications).

5.4. *Fusion précoce et fusion tardive des modalités*

À partir de l'ensemble des 8 prédictions élémentaires définit précédemment (`LK-T`, `LK-I`, `LK-S`, `VMU-S1`, `MML-T`, `MML-I`, `MCG-T` et `MCG-I`), nous proposons deux fusions : 1) une fusion précoce consistant en l'application d'un classifieur sur l'ensemble des 8 prédictions ; 2) une fusion tardive représentée par un méta-classifieur à deux niveaux dans lequel les messages sont regroupés par type d'approche (texte, source ou image) puis un classifieur regroupe les trois prédictions de 1^{er} niveau.

5.4.1. *Fusion tardive*

Le tableau 4 présente les résultats des trois classifieurs de premier niveau (prédiction au niveau du texte, source ou image) ainsi que les trois fusions de second niveau (prédiction finale à partir des trois prédictions précédentes). Les trois fusions de second niveau sont modélisées par des règles : 1) La règle $\mathcal{RE}_1$ prédit *faux* si au moins une prédiction en entrée vaut *faux* ; 2) La règle $\mathcal{RE}_2$ prédit *vrai* ou *faux* si au moins

deux prédictions en entrée sont en accord et correpond donc à un vote majoritaire ; 3) La règle $\mathcal{RE}_3$ prédit *vrai* ou *faux* lorsque les trois prédictions en entrée sont en accord, si ce n'est pas le cas *inconnu* est prédit. Une première constatation est le résultat encourageant du classifieur réalisant la fusion des prédictions *image*. À l'inverse, la fusion entre les deux prédictions basées sur les sources (`LK-S` et `VMU-S1`) donne un résultat moins prometteur (i.e. score de micro-*F-Mesure* de $93,04$ % pour la prédiction `VMU-S1` seule contre $90,53$ % pour la fusion des prédictions basées sur les sources). Une explication possible est une éventuelle contradiction sur la confiance de certaines sources entre les différentes listes utilisées par les systèmes. Il serait alors intéressant de confronter directement les différentes listes de sources de confiance utilisées. De plus, ces approches possèdent une autre limite associée au fait que l'information de la source peut ne pas être disponible ce qui rend impossible une prédiction dans ce cas là. Enfin, on constate que l'approche visant à n'attribuer une classe qu'à condition d'obtenir cette prédiction par les trois classifieurs de premier niveau obtient un score relativement faible mais représente une classification sûre puisque $1\ 400$ messages sont ainsi classés sur les $2\ 228$ possibles et seulement 4 erreurs sont présentes dans ces $1\ 400$ prédictions.

1^{er} niveau de prédiction		SVM	Arbre de déc.	Random Forest
Texte	F-Mesure	$64,68$ %	$74,41$ %	**74,51 %**
	Taux de B.C.	$64,95$ %	$74,73$ %	**74,82 %**
Source	F-Mesure	$89,14$ %	**90,53 %**	$89,57$ %
	Taux de B.C.	$89,17$ %	**90,56 %**	$89,54$ %
Image	F-Mesure	$84,44$ %	$85,45$ %	**85,73 %**
	Taux de B.C.	$84,49$ %	**85,73 %**	**85,73 %**
2^{eme} niveau de prédiction		$\mathcal{RE}_1$	$\mathcal{RE}_2$	$\mathcal{RE}_3$
	F-Mesure	$78,74$ %	**88,43 %**	$76,25$ %
	Taux de B.C.	$80,07$ %	**88,60 %**	$62,66$ %

Tableau 4. *Fusion tardive basée sur les huit prédictions élémentaires*

5.4.2. *Fusion précoce*

Les résultats d'une fusion précoce des 8 prédictions élémentaires retenues sont présentés dans le tableau 5. La baseline est de nouveau le vote majoritaire sur les 8 prédictions en entrée. Dans le cas d'une égalité, la classe *inconnu* est utilisée. On note alors que malgré le retrait de la moitié des prédictions en entrée (8 prédictions sur les 14 initiales), il est possible de classer correctement $91,83$ % des tweets, soit $2\ 046$ tweets sur $2\ 228$.

Fusion précoce	Baseline	SVM	Arbre de déc.	Random Forest
F-Mesure	$55,26\ \%$	$86,66\ \%$	**91,85 %**	$91,00\ \%$
Taux de B.C.	$66,29\ \%$	$86,58\ \%$	**91,83 %**	$90,98\ \%$

Tableau 5. *Fusion précoce basée sur les huit prédictions élémentaires*

6. Conclusion

Dans cet article, un système de fusion est proposé en se basant sur les prédictions réalisées par les quatres équipes participantes à la tâche *Verifying Multimedia Use* de la campagne d'évaluation *Mediaeval 2016*. Ainsi, nous avons vu que les approches basées sur la crédibilité de la source obtiennent de bons scores de prédiction mais peuvent se contredire lors d'une fusion entre elles, ce qui ne permet pas une amélioration des scores de prédictions . Nous avons aussi vu que les prédictions réalisées sur les images obtiennent de meilleures performances grâce à la fusion des prédictions des différents systèmes basés image (approximativement $20\ \%$ sur les scores de micro-*F-Mesure* et du taux de bonne classification). Enfin, cet article montre que la fusion des différentes prédictions permet une augmentation des scores de micro-*F-Mesure* et du taux de bonne classification passant respectivement de $93,37\ \%$ et $93,04\ \%$ à $95,50\ \%$ et $95,47\ \%$.

Les travaux futurs viseront à corriger certains problèmes mis en évidence par nos expérimentations au niveau des systèmes par modalités (e.g images modifiées considérée comme similaire à l'image réelle initiale), ainsi qu'au niveau des fusions. Nous prévoyons ainsi d'étendre la couverture de la base d'images et d'améliorer le module de comparaison de contenu, notamment en effectuant des post-traitements pour éliminer les faux-positifs lors de la reconnaissance d'images similaires. D'autres pistes de recherche possibles sont les applications et l'évaluation de ces prédictions, élémentaires et fusions, à d'autres types de données ou de contexte (e.g analyse en temps réel).

7. Bibliographie

Bianchi T., Piva A., « Image forgery localization via block-grained analysis of JPEG artifacts », *IEEE Transactions on Information Forensics and Security*, 2012.

Boididou C., Middleton S., Papadopoulos S., Dang-Nguyen D.-T., Riegler M., Boato G., Petlund A., Kompatsiaris Y., « The VMU Participation @ Verifying Multimedia Use 2016 », *MediaEval 2016 Workshop*, 2016a.

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*, 2016b.

Cao J., Jin Z., Zhang Y., Zhang Y., « MCG-ICT at MediaEval 2016 : Verifying Tweets From Both Text and Visual Content », *MediaEval 2016 Workshop*, 2016.

Declerck T., Lendvai P., « Processing and Normalizing Hashtags », *Recent Advances in Natural Language Processing (RANLP)*, 2015.

Derczynski L., Maynard D., Rizzo G., van Erp M., Gorrell G., Troncy R., Petrak J., Bontcheva K., « Analysis of named entity recognition and linking for tweets », *Information Processing & Management*, 2015.

Foteini, Mezaris V., Patras Ioannis B.-M., « Online multi-task learning for semantic concept detection in video », *Image Processing (ICIP), 2016 IEEE International Conference on*, IEEE, p. 186-190, 2016.

Golbeck J., Hendler J., « Inferring binary trust relationships in web-based social networks », *ACM Transactions on Internet Technology (TOIT)*, 2006.

Goljan M., Fridrich J., Chen M., « Defending against fingerprint-copy attack in sensor-based camera identification », *IEEE Transactions on Information Forensics and Security*, vol. 6, n° 1, p. 227-236, 2011.

Gottron T., Schmitz J., Middleton S., « Focused exploration of geospatial context on linked open data », *3rd International Conference on Intelligent Exploration of Semantic Data (ICIESD)*, 2014.

Gupta M., Zhao P., Han J., « Evaluating Event Credibility on Twitter », *2012 SIAM International Conference on Data Mining*, 2012.

Li W., Yuan Y., Yu N., « Passive detection of doctored JPEG image via block artifact grid extraction », *89th Signal Processing*, 2009.

Middleton S., « Extracting attributed verification and debunking reports from social media : mediaeval-2015 trust and credibility analysis of image and video », *Mediaeval 2015 Workshop*, 2015a.

Middleton S. E., « REVEAL Project-trust and credibility analysis », *Mediaeval 2015 Workshop*, 2015b.

Page L., Brin S., Motwani R., Winograd T., « The PageRank citation ranking : bringing order to the web », *Stanford InfoLab*, 1999.

Pasquini C., Pérez-González F., Boato G., « A Benford-Fourier JPEG compression detector », *IEEE International Conference on Image Processing (ICIP)*, 2014.

Phan Q.-T., Budroni A., Pasquini C., De Natale F., « A hybrid approach for multimedia use verification », *MediaEval 2016 Workshop*, 2016.

Robertson S. E., Walker S., Hancock-Beaulieu M., « Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive », *7th Text Retrieval Conference (TREC)*, 1998.

Silverman C., *Verification Handbook : An Ultimate Guideline on Digital Age Sourcing for Emergency Coverage*, The European Journalism Centre (EJC), 2014.

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

Tolias G., Sicre R., Jégou H., « Particular object retrieval with integral max-pooling of CNN activations », *4th International Conference on Learning Representations (ICLR)*, 2016.

Wan J., Wang D., Hoi S. C. H., Wu P., Zhu J., Zhang Y., Li J., « Deep learning for content-based image retrieval : A comprehensive study », *22nd ACM international conference on Multimedia (ICM)*, 2014.

Zampoglou M., Papadopoulos S., Kompatsiaris Y., « Detecting image splicing in the wild (WEB) », *Multimedia & Expo Workshops (ICMEW)*, 2015.