Fusion par apprentissage pour la détection de fausses informations dans les réseaux sociaux

Cédric Maigrot, Ewa Kijak, Vincent Claveau

Fusion par apprentissage pour la détection de fausses informations dans les réseaux sociaux

Cédric Maigrot, Ewa Kijak, Vincent Claveau

IRISA, Univ. Rennes 1, CNRS, Campus de Beaulieu, 35042 Rennes, France
prenom.nom@irisa.fr

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 drawbacks comes from the absence of verification of the pieces 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 possibility of combining and merging these approaches in order to evaluate the predictive power of each modality and to take advantage 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.*

DOI:10.3166/DN.1.2-3.1-26

1. Introduction

Les réseaux sociaux prennent 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 sur 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, même une information identifiée comme étant un hoax peut être difficile à arrêter lorsqu'elle est déjà partagée un grand nombre de fois.

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 dans ces travaux un système multimodal. Nous présentons à la fois des approches exploitant uniquement le contenu textuel, le contenu des images ou les sources citées dans les messages, et des stratégies de combinaison de ces approches mono-modales. Ces différentes approches sont évaluées et discutées sur les données de la tâche *Verifying Multimedia Use* du challenge *MediaEval2016* portant précisément sur cette problématique. D'autre part, à partir des méthodes de toutes les équipes ayant participé à cette tâche de *MediaEval2016*, nous explorons différentes stratégies de fusion pour analyser l'apport des différentes approches et 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 challenge *MediaEval* dont sont extraites les données utilisées dans ces travaux. Nous présentons ensuite dans la section 4 les approches que nous avons mises en place, ainsi que les systèmes proposés par les autres équipes participantes à la tâche *VMU*. La section 5 présente le protocole expérimental et les résultats obtenus par les différentes approches. Différentes stratégies de fusion sont testées et discutées dans la section 6. Enfin, la section 7 résume les principales observations 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 sous plusieurs angles. 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. En particulier, le *fact-checking* a souvent pour objet la détection des énoncés de personnalités susceptibles d'être vérifiées dans des déclarations de personnalités, appelé *checkworthiness* (Hassan *et al.*, 2015), alors que nous classons des messages postées sur les médias sociaux, par des inconnus, de manière globale. Les approches utilisées par le *fact checking* pour ensuite juger de l'énoncé suspect fait souvent appel à des bases de connaissance (Ciampaglia *et al.*, 2015) alors que nous n'exploitons pour notre part que les données d'entraînement (les messages fournis dans le cadre du challenge MediaEval).

Plusieurs familles d'indices ont été exploitées pour détecter les *hoax* dans les réseaux sociaux. On peut notamment citer :

– les indices textuels : le message lui-même apporte évidemment des informations potentiellement exploitables.

– les indices multimédias, dans le cas de messages contenant des images ou des vidéos : ces contenus multimédias peuvent parfois être analysés en vue de détecter des modifications (*photoshoppage*).

– les indices de diffusion dans le réseau social, qui peuvent se décliner en deux questions : quelle est la source de l'information, quel est le parcours (mode de diffusion) du message.

L'analyse des sources des messages et des relations entre membres a ainsi fait l'objet de plusieurs travaux. (Golbeck, 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 étendue du réseau qui les rendent difficilement applicables en pratique pour des réseaux sociaux commerciaux et grand public.

L'analyse du mode de diffusion des messages a également fait l'objet de plusieurs travaux, dont le but est de distinguer des rumeurs de messages classiques en observant leur propagation dans le réseau social (Vosoughi, 2015 ; Wang, Terano, 2015 ; Jin *et al.*, 2013 ; Zubiaga *et al.*, 2015). Ces analyses reposent sur des modèles de diffusion mais nécessitent d'avoir accès à une grand part du réseau pour suivre ces messages, ce qui est rarement possible avec les réseaux sociaux grand public que nous visons.

L'analyse des indices multimédias est au cœur du projet européen *InVid* [1]. Ce projet s'intéresse à la détection automatique de vidéos truquées dans les réseaux sociaux. Pour ce faire, Foteini *et al.* (2016) travaillent sur des images issues des vidéos analysées avec des approches issues du domaine des *forensics* permettant de détecter des

1. Voir http://www.invid-project.eu/. The InVID project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687786.

modifications de l'image. L'analyse des vidéos est importante et utile dans notre cas, cependant ces travaux ne sont pas applicables en tant que tel car leur approche ne s'intéresse pas aux textes diffusant ces vidéos ni aux aspects réseaux sociaux à proprement parler.

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 images au format JPEG, il est possible de repérer une double compression (Bianchi, 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.

D'autres travaux exploitent plusieurs de ces indices conjointement. C'est le cas de (Middleton, 2015b), qui, dans le cadre du européen *Reveal Project*[2] , 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 largement usage de ressources externes (Gottron *et al.*, 2014).

Le projet européen *PHEME*[3] (Derczynski *et al.*, 2015) s'intéresse à la détection des rumeurs sur les réseaux sociaux ou les médias en ligne en exploitant des indices textuels et des indices de diffusion. Plusieurs travaux de ce projet étudient les liens entre les messages sur les réseaux sociaux, c'est-à-dire les 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*, c'est-à-dire s'appuyer sur des analyses (humaines) produites par les utilisateurs du réseau et postées en réaction au tweet étudié. Cependant, les travaux menés au sein de ce projet sur la normalisation des

2. Voir https://revealproject.eu/. Le projet *Reveal Project* est financé par un schéma de projet collaboratif sous l'accord de subvention (*grant agreement*) No 610928.

3. Voir https://www.pheme.eu/. Le projet *Pheme* est financé par la Commission Européenne sous l'accord de subvention (*grant agreement*) No 611233. Ce projet a commencé le 1er janvier 2014 pour 36 mois.

hashtags ou la détection d'entités nommées (Declerck, Lendvai, 2015) peuvent être utiles pour la tâche ciblée dans cet article.

3. Présentation de la tâche *Verifying Multimedia Use* du challenge *MediaEval2016*

La tâche *Verifying Multimedia Use* (VMU) de la campagne d'évaluation *MediaEval* en 2016, proposait de classer des messages provenant de *Twitter*[4] selon leur véracité entre les classes *vrai* et *faux*, avec la possibilité d'utiliser une classe *inconnu* si le système ne permet pas de prendre de décision. Autoriser le système à ne pas se prononcer peut permettre d'obtenir une forte précision pour les classes *vrai* et *faux* (Boididou, Papadopoulos *et al.*, 2016).

Concernant la classe attribuée à chaque message, la règle suivante est appliquée : Un message est considéré comme *faux* si il partage un contenu multimédia qui ne représente pas l'événement dont il fait référence.

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). Tous les messages ont au moins un contenu multimédia (image ou vidéo). Plusieurs messages peuvent cependant partager la même image, mais il est important de noter qu'une vidéo ou une image aura toujours la même classe (biais créé lors de la constitution du jeu de données par les organisateurs). 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 3. Ainsi, le plus grand événement dans cette collection est *Paris Attack* avec 580 messages partageant 25 contenus multimédias différents, 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. Il faut noter que dans la section 5, les résultats présentés sont ceux obtenus sur l'ensemble de test, et que les techniques de fusion présentées en section 6 sont utilisées sur les prédictions des systèmes des participants sur l'ensemble de test.

Plusieurs descripteurs ont été proposés par les organisateurs lors de cette tâche. Ces descripteurs relèvent 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 positive et de mots à opinion négative, le nombre de mentions *Twitter*, de hashtags, d'urls et de retweets.

4. https://twitter.com/

Figure 1. Exemples de deux tweets de la tâche VMU de la campagne MediaEval, partageant la même image

Figure 2. Exemples de deux tweets de la tâche VMU de la campagne MediaEval, sur le même événement (ouragan Sandy) que dans la figure 1

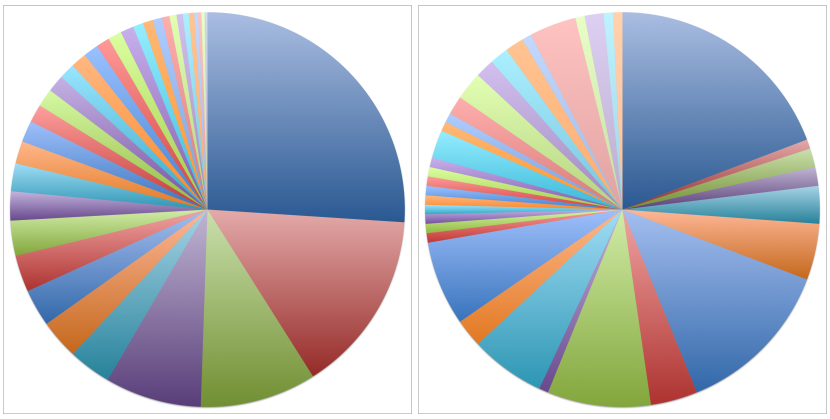

Figure 3. Répartition des messages, à gauche, et des contenus multimédias (images et vidéos), à droite, par événement dans le jeu de test de la tâche VMU *(35 événements)*

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

L'ensemble des descripteurs associés aux images, noté $\mathcal{FOR}$, provient de méthodes issues du domaine des *forensics* : indices de double compression JPEG (Bianchi, 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).

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

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

4. Présentation des systèmes ayant participé à la tâche *Verifying Multimedia Use* du challenge *MediaEval2016*

Quatre équipes ont participé à la tâche pour un total de 14 soumissions. Les équipes sont dénotées par la suite *LK* (notre équipe), *MMLAB*, *MCG-ICT* et *VMU* (organisateurs de la tâche). 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 nos approches, tous les messages partageant la même image sont associés à 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 du message ; 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. La fin de cette section sera consacrée à la présentation des approches des autres équipes participantes. Une étude comparative entre nos approches et celles à la tâche sera proposé plus tard dans l'article.

4.1. Approche textuelle (`LK-T`)

Cette approche exploite le contenu textuel des messages 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 messages qui utilisent cette image. L'idée à l'œuvre dans cette approche est de capturer les commentaires similaires entre un message du jeu de test et ceux 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 d'une image inconnue (i.e. l'union des contenus textuels des messages qui utilisent cette image) et $\{I_{d_i}\}$ l'ensemble des descriptions des images de l'ensemble d'apprentissage. La classe de I_q est déterminée 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 donc au cœur de cette approche. La similarité utilisée est 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 classer I_q) et à un document (ici un texte de $\{I_{d_i}\}$) ; voir équation 1.

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

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

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 prétraitement est la normalisation de l'orthographe et des smileys développé par l'équipe pour le challenge DeFT 2017 (Claveau, Raymond, 2017). 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 externe (statique). 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 dans lesquels 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) en s'appuyant sur des listes établies [5], réseaux télévisuel d'information (francophones et anglophones) [6]. 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` et `LK-I2`)

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 (c'est-à-dire *vraies*) et 7 500 images *modifiées*.

Les descripteurs que nous générons à partir des images sont calculés en utilisant un réseau de neurones convolutionnel profond (Simonyan, 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

5. https://en.wikipedia.org/wiki/List_of_news_agencies
6. https://en.wikipedia.org/wiki/Lists_of_television_channels

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 cosinus est calculée entre les images requêtes et les images de la base.

Le système de recherche retourne donc une liste d'images ordonnée 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).

Dans l'approche que nous notons `LK-I`, si aucune image de la base n'est jugée similaire, l'image requête reçoit la classe *inconnu*. Du fait de la faible taille de la base de référence, ce cas est courant. Une version alternative de cette approche, notée `LK-I2` par la suite, assigne à ces images incertaines, la classe de probabilité a priori maximale, à savoir la classe *faux*.

4.4. Présentation des autres approches

Pour chacune des autres équipes participantes, nous décrivons ci-dessous le type de données ainsi que l'approche utilisée pour prédire la classe des messages.

4.4.1. Équipe VMU

Cinq méthodes ont été testées par les organisateurs de la tâche. Ces méthodes reposent sur deux systèmes, dont elles sont des variantes (Boididou, Middleton *et al.*, 2016).

`VMU-F1` et `VMU-F2` s'appuient sur un premier système qui est un méta-classifieur dans lequel deux ensembles de descripteurs sont utilisés séparément par deux classifieurs, entraînés 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 entraîné sur l'union de l'ensemble d'entraînement et des messages de l'ensemble de test ayant reçu des prédictions en accord sur les deux premiers classifieurs. Les classifieurs utilisés sont des forêts aléatoires. `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 $\mathcal{FOR}$ pour l'un des classifieurs, et $\mathcal{U}$ pour l'autre.

`VMU-S1` et `VMU-S2` sont basés sur `VMU-F1`, auquel est ajouté un second système qui exploite deux listes de sources connues : la première est une liste de sources de confiance alors que la seconde regroupe des sources de non-confiance. Lorsqu'une prédiction basée sur les sources n'est pas possible, le premier système est utilisé pour fournir une prédiction.

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

4.4.2. Équipe MMLAB

L'approche proposée repose sur deux classifieurs de type forêt aléatoire (Phan *et al*., 2016). Le premier classifieur, appelé MML−T, prend en entrée la concaténation des descripteurs $\mathcal{T}$ et $\mathcal{U}$ proposés par les organisateurs de la tâche.

Le second classifieur, dénoté MML−I, utilise les contenus multimédias (images et vidéos) associés aux messages. Il prend en entrée la concaténation de descripteurs *forensics* (l'ensemble $\mathcal{FOR}$) et de descripteurs textuels obtenus en utilisant une base de connaissances externe. Pour chaque évènement, une liste des termes les plus pertinents en relation avec cet évènement est établie en utilisant la mesure *TF-IDF* sur les textes des sites les plus pertinents retournés par un moteur de recherche textuel en ligne. Pour chaque image, un moteur de recherche inversé (*Google image search*) est ensuite utilisé et des mesures de fréquence (i) des termes pertinents précédemment identifiés, (ii) des termes de polarité positive et négative (issue de lexique utilisée en analyse de sentiment) sont appliquées sur les textes des sites les plus pertinents retrouvés. Dans le cas d'une vidéo *Youtube*, ces mesures de fréquence sont appliquées aux commentaires de la vidéo. Les autres vidéos ne sont pas analysées.

Enfin MML−F est la fusion (combinaison linéaire) des scores de chaque classe fournis par 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. Les descripteurs $\mathcal{T}$ et $\mathcal{U}$ proposés par les organisateurs de la tâche sont utilisés, et un nouvel indice est ajouté à cet ensemble. Le calcul de ce nouvel indice repose sur la séparation d'un évènement en thèmes; un thème est défini comme l'ensemble des messages partageant la même image ou vidéo. Chaque thème est décrit par les moyennes des descripteurs $\mathcal{T}$ et $\mathcal{U}$ de ses messages, complétées de nouvelles statistiques comme le nombre de messages dans le thème, le nombre de messages (*hashtags*) distincts (afin de discriminer les retweets), les ratios de messages distincts, de messages contenant une URL ou une mention, et de messages contenant plusieurs URLs, mentions, hashtags ou points d'interrogation. À partir de ces caractéristiques, un classifieur au niveau des thèmes est construit, et indique la probabilité qu'un message soit *vrai* ou *faux*. Cette probabilité est le nouvel indice ajouté à chaque message. Le classifieur au niveau des messages, construit sur les descripteurs textuels enrichis, est dénommé MCG−T.

Un second module évalue la crédibilité du contenu visuel. Pour les images, les auteurs utilisent les descripteurs $\mathcal{FOR}$ (sans préciser le classifieur utilisé). Les vidéos sont traitées différemment. En se référant à Silverman (2014), les auteurs définissent quatre caractéristiques pour décrire les vidéos : une mesure de la netteté de l'image, le

rapport de contraste, défini comme le rapport de la taille d'une vidéo sur sa durée, la durée de la vidéo et la présence de logos. Ces quatre caractéristiques sont combinées par un arbre de décision binaire. On note `MCG-I` les prédictions correspondant à cette approche.

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

5. Résultats et discussions des différentes approches

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). La mesure d'évaluation utilisée dans la tâche est la *F-Mesure* sur la classe *faux*. Cependant, cette mesure n'est pas discriminante entre les prédictions *vrai* et *inconnu* des messages. De plus, elle se base sur la classe majoritaire *faux*, ce qui représente un biais (i.e. *F-Mesure* sur la classe *faux* est de $71,14\ \%$ sur l'ensemble de test en prédisant systématiquement *faux*). Nous utilisons à la place la micro-*F-Mesure* et le taux de bonnes classifications (*accuracy*) qui sont des mesures globales sur l'ensemble des classes à prédire.

D'autre part, 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 paquet afin de ne pas biaiser l'évaluation. Pour mettre en œuvre cette validation croisée, l'ensemble des événements est subdivisé aléatoirement en n paquets. L'évaluation rend donc compte de la performance que l'on peut espérer lors du traitement d'un nouvel évènement, engendrant son lot de messages pouvant être vrais ou faux. Les résultats des méthodes décrites en section 4, ré-évalués selon le protocole décrit ci-dessus, sont présentés dans le tableau 2 pour une évaluation par message et dans le tableau 3 pour l'évaluation par groupe de messages.

Entre les deux modes d'évaluation (par message, ou par groupe de messages partageant un même contenu multimédia), on observe de grandes différences pour certaines méthodes. En effet, les approches assignant des classes contradictoires à différents messages partageant un même contenu multimédia sont pénalisées dans notre deuxième cadre d'évaluation (chute de rappel). À l'inverse, notre approche `LK-I2` bénéficie de sa stratégie par défaut pour les contenus multimédia classés 'inconnu' par `LK-I`. Les résultats de chacune des approches sont discutés dans les sous-sections suivantes.

5.2. Comparaison des différentes approches selon les modalités exploitées

En complément des résultats chiffrés de l'évaluation fournis précédemment, nous examinons les approches selon le type d'indices qu'elles exploitent (modalité texte, source ou image) et leur éventuelle complémentarité pour les expériences de fusion présentées dans la section suivante. Nous excluons de cette étude les prédictions fai-

Tableau 2. Performances des soumissions des équipes Linkmedia *(LK),* VMU*,* MMLAB *(MML) et* MCG-ICT *(MCG) à la tâche VMU selon le taux de bonne classification (%) et la micro-F-mesure (%) (écart-type entre parenthèses) avec une évaluation par message*

	F-mesure	Taux B.C.
LK-T	77,5 (22,1)	72,5 (22,3)
LK-I	41,9 (32,6)	33,0 (30,3)
LK-I2	43.4 (28.4)	33,7 (28,5)
LK-S	**88,5 (16,1)**	**87,1 (15,8)**
VMU-F1	90,2 (5,9)	87,0 (10,22)
VMU-F2	82,9 (24,0)	85,6 (17,1)
VMU-S1	89,1 (9,2)	**89,5 (7,0)**
VMU-S2	**90,5 (6,3)**	87,0 (10,8)
VMU-B	82,6 (24,3)	78,8 (25,8)
MML-T	54,8 (19,1)	53,2 (14,8)
MML-I	77,1 (25,9)	71,4 (25,5)
MML-F	**83,3 (14,9)**	**78,3 (17,4)**
MCG-T	**72,4 (29,6)**	**69,0 (32,1)**
MCG-I	59,9 (35,1)	62,4 (33,5)
MCG-F	66,6 (35,9)	64,4 (36,6)

Tableau 3. Performances des soumissions des équipes Linkmedia *(LK),* VMU*,* MMLAB *(MML) et* MCG-ICT *(MCG) à la tâche VMU selon le taux de bonne classification (%) et la micro-F-mesure (%) (écart-type entre parenthèses) avec une évaluation par groupe de messages partageant un même contenu multimédia*

	F-mesure	Taux B.C.
LK-T	71,7 (36,9)	69,5 (36,9)
LK-I	47,5 (45,6)	45,8 (45,6)
LK-I2	80,7 (33,5)	78,8 (35,0)
LK-S	**81,9 (33,8)**	**84,3 (30,6)**
VMU-F1	28,9 (39,7)	40,8 (39,0)
VMU-F2	71,1 (40,4)	**74,4 (36,4)**
VMU-S1	40,0 (43,8)	50,9 (41,1)
VMU-S2	33,5 (41,2)	43,6 (40,3)
VMU-B	**77,2 (33,7)**	74,1 (35,2)
MML-T	9,25 (23,3)	13,8 (23,9)
MML-I	70,4 (36,4)	67,3 (36,4)
MML-F	**71,3 (36,7)**	**71,5 (34,3)**
MCG-T	**66,4 (42,9)**	**67,5 (41,3)**
MCG-I	55,9 (42,9)	59,9 (40,5)
MCG-F	62,6 (43,4)	66,6 (41,0)

sant déjà intervenir des fusions entre modalités. Ainsi seules les prédictions LK−T, LK−I et LK−S seront gardées parmi 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* reposent toutes sur de la fusion (cf. section 4.4). Nous retenons cependant VMU−S1 qui se fonde sur les sources et qui est la prédiction obtenant les meilleures performances. Ces huit prédictions, notées *élémentaires* (ou plus précisément 7 élémentaires plus VMU−S1), seront utilisées dans la suite.

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

5.2.2. 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 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 forte précision des messages classés comme *vrai* ou *faux* (respectivement 100.00 % et $92,97$ %) aux dépens du rappel (respectivement $41,22$ % et $87,47$ %).

5.2.3. 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 peuvent présenter une complémentarité pour une fusion car elles utilisent des indices très différents. 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 texte en surimpression (typiquement sous forme d'un tampon 'faux', 'rumeur' ou 'vrai') ou de modifications afin d'améliorer la compréhension (*e.g* un cercle rouge sur la zone photoshoppée pour aider le lecteur à la trouver). 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 4).

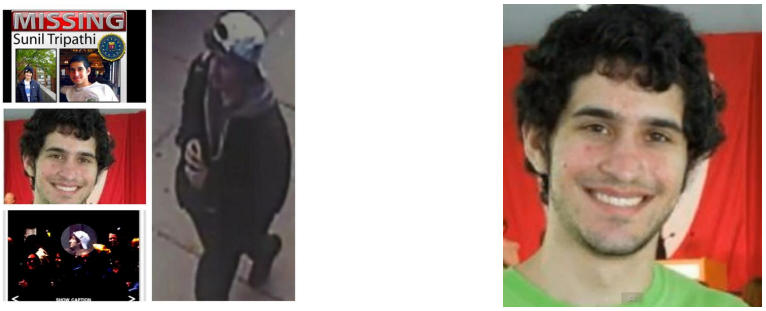

Figure 4. Exemple d'une image requête (à gauche) ayant un vrai positif dans la base (à droite) qui n'a pas été retrouvé par la recherche d'images similaires, les artefacts d'édition de l'image requête faisant chuter le score de similarité entre ces 2 images.

Au vu des résultats des approches basées sur les images, il semble que l'utilisation d'une recherche d'images similaires (prédictions `LK-I` et `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 amène 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 sup-

plémentaire de vérification des contenus similaires retrouvés et détection des modifications).

Les faibles résultats obtenus par l'approche `LK-I2` fondée sur l'image s'expliquent en partie par la faible taille de la base d'images. En effet, seulement environ 25 % des images à classer étaient représentées dans la base au moment de la soumission des résultats pour le challenge. Le grand nombre d'images pour lesquelles aucune décision n'a été prise (classe *inconnu*) impacte fortement les résultats en terme de rappel.

Pour analyser l'influence de la taille de cette base sur les résultats, nous reportons dans la figure 5 l'évolution des mesures de performance (précision, rappel et F-mesure ; évaluation par message) en fonction du nombre d'images dans la base. Pour chaque taille de base considérée, les expériences ont été répétées dix fois et les résultats moyennés. Pour chaque expérience, les images sélectionnées sont désignées aléatoirement. La base utilisée est légèrement plus grande que celle utilisée lors du Challenge MediaEval 2016 (2 000 images supplémentaires), ce qui explique les résultats légèrement supérieurs quand 100 % de la base est utilisée.

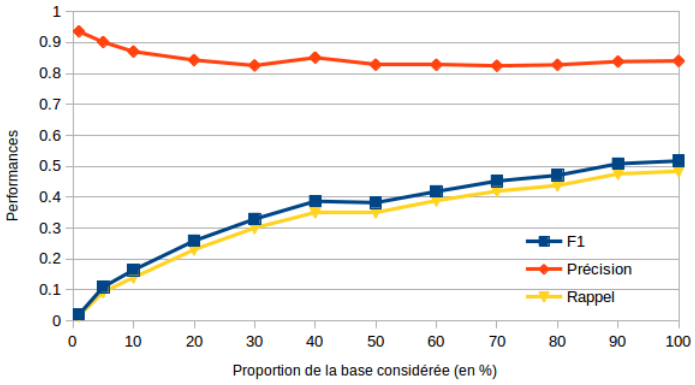

Figure 5. Évolution des performances de l'approche image en fonction de la taille de la base d'image (en pourcetage).

Sans surprise, il en ressort clairement la dépendance à la taille, et donc à la couverture, de la base. En effet, si la précision de l'approche reste relativement stable à un niveau élevé, une petite base implique un faible rappel. Il est cependant intéressant d'observer la pente de la courbe qui laisse espérer un gain de f-mesure conséquent avec des tailles de bases d'images raisonnables. Notons tout de même que ces bases sont constituées à partir de ressources développées manuellement (sites web) ; cet apport d'informations externes riches est discuté en sous-section 6.3.

6. Stratégies de fusion

6.1. Fusion simple des soumissions

Une fusion directe des prédictions du tableau 2 est d'abord étudiée 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 quatre algorithmes de classification :

1. SVM linéaire ;
2. arbre de décision ;
3. *Random Forest* (forêt aléatoire, avec 500 arbres de profondeur 2) ;
4. réseau de neurones (une couche *Dropout*, une couche cachée dense de taille 20 et une couche de sortie avec fonction d'activation sigmoïde).

En plus de ces classifieurs, nous indiquons les résultats d'un système référence correspondant au vote majoritaire sur les prédictions des participants (i.e. parmi les prédictions, la classe prédite la plus fréquemment est associée au message).

Le protocole d'évaluation est le suivant : pour chaque classifieur, nous évaluons ses performances lorsqu'il est entraîné sur tous les messages de tous les événements sauf un événement dont les messages servent de jeu de test. Nous faisons tourner cet événement test et moyennons les résultats. C'est donc l'équivalent du *leave-one-out*, sauf que nous raisonnons au niveau des événements et non pas des exemples pris individuellement. Nous adoptons par ailleurs les deux cadres d'évaluation vus précédemment : une prédiction par message, et une prédiction par groupe de message partageant le même contenu multimédia. Les résultats sont respectivement présentés dans les tableaux 4 et 5. Nous notons avec une astérisque les résultats statistiquement significatifs (test de Wilcoxon avec $p < 0,05$) par rapport au système de référence.

On note que le système de référence ne permet pas de surpasser les meilleures prédictions à la tâche, contrairement aux classifieurs utilisant toutes les méthodes des participants. Cela montre que toutes les prédictions n'ont pas la même importance et que les classifieurs permettent d'apprendre des pondérations adaptées à chacune des méthodes, voire des combinaisons non linéaires plus complexes. À ce titre, le meilleur classifieur (réseau de neurones) permet une augmentation significative du taux de bonne classification, tout en offrant plus de constance (écart-type des mesures de performances plus faible).

Certains messages sont plus difficiles à classer que d'autres et cela se retrouve bien sûr sur les résultats de la fusion. Dans la figure 6, nous indiquons sous forme d'histogramme la répartition des tweets selon le nombre de méthodes les classant correctement. Comme on peut le voir, tous les messages sont correctement classés par au moins une des méthodes des participants. Mais pour certains messages, une grande partie des méthodes les classent incorrectement : il y a notamment 263 messages pour lesquels la majorité des méthodes se trompe. Une stratégie de fusion simple aura alors

Tableau 4. F-Mesure moyenne et taux de bonne classification (%) sur les messages et écarts-types de la fusion basée sur les prédictions soumises à la tâche Verifying Multimedia Use *de* MediaEval 2016*; évaluation par message*

Fusion directe	Majorité	SVM	Arbre de décision	Random Forest	NN
F-Mesure	$87,5\ (26,3)$	$87,1\ (24,8)$	$86,0\ (25,3)$	$86,3(23,5)$	$89,5(23,9)^*$
Taux de B.C.	$87,9\ (22,4)$	$87,2\ (22,5)$	$86,6\ (21,4)$	$88,6(13,7)^*$	$90,2(19,1)^*$

Tableau 5. F-Mesure moyenne et taux de bonne classification (%) sur les images et écarts-types de la fusion basée sur les prédictions soumises à la tâche Verifying Multimedia Use *de* MediaEval 2016*; évaluation par groupe de messages partageant un même contenu multimédia*

Fusion directe	Majorité	SVM	Arbre de décision	Random Forest	NN
F-Mesure	$82,6\ (31,6)$	$90,9\ (23,9)^*$	$84.3\ (28,8)$	$90,5(24,6)^*$	$91,4(23,7)^*$
Taux de B.C.	$84,0\ (28,3)$	$95,1\ (11,7)^*$	$86,9\ (23,0)^*$	$95,1(12,9)^*$	$96,3(10,5)^*$

de grande chance de se baser sur cette majorité pour prendre sa décision, ce qui engendrera une erreur de prédiction. Un examen des cas d'échec montre que ce sont bien ces quelques messages qui trompent les classifieurs et les modules de fusion. Ces messages difficiles à classer présentent l'une des trois caractéristiques suivantes pouvant expliquer cette difficulté :

1. tweets écrits dans langues non prises en charge par les traitements (extraction d'information) et rendant les calculs de similarité inadaptés (trop peu de tweets dans

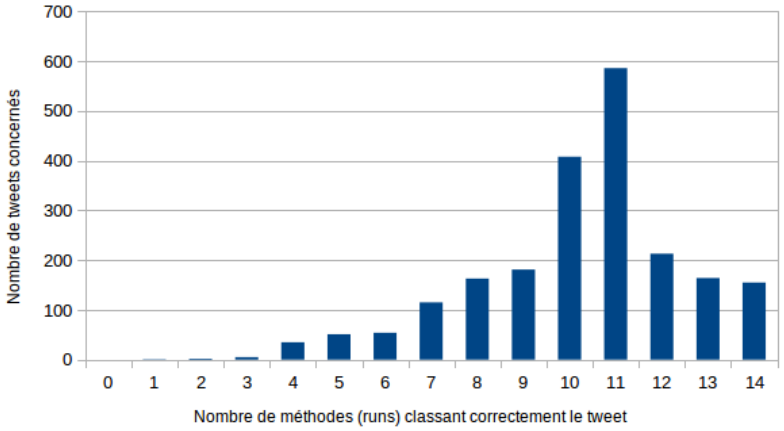

Figure 6. Histogramme des messages selon le nombre de méthodes les classant correctement ; évaluation par groupe de messages partageant un même contenu multimedia

cette langue) ;

2. des URL réduites qui cachent la source citée (e.g. utilisation des alias courts d'URL tels que goo.gl, t.co ou bit.ly) ;

3. une grande partie de ces messages proviennent d'événements ayant des messages *vrais* et *faux* et sont donc ambigus (*Paris attacks* et *Fugi Lenticular*).

Nous donnons en figure 7 deux exemples de tels tweets.

Figure 7. Exemple de messages difficiles à classer (mal classés par plus de 12 méthodes des participants et mal classés par les modules de fusion).

Pour étudier les contributions à la fusion de chacune des méthodes, nous pouvons observer les classifieurs produits. Dans la suite, nous nous focalisons sur les *Random Forest* qui obtiennent à la fois de bons scores et qui permettent d'étudier ces contributions facilement. La contribution d'un attribut (dans notre cas la prédiction d'une méthode) est définie comme l'importance selon l'indice de Gini, appelé aussi *mean decrease impurity* et tel que défini par (Breiman *et al.*, 1984), moyenné sur l'ensemble des arbres de la forêt aléatoire et normalisé entre 0 et 100 %. Nous présentons ces contributions dans la figure 8, et nous les mettons en regard des performances des soumissions prises indépendamment.

Il est surprenant d'observer que ce ne sont pas les meilleurs systèmes qui servent de base à la fusion. En effet, `VMU-F1`, `VMU-S1` et `VMU-S2` représentent plus de 60 % des contributions à la fusion, alors que leurs scores sont parmi les plus faibles. Ces trois systèmes ont en effet pour caractéristique d'être les plus précis, mais d'avoir un faible rappel (beaucoup de messages sont classés 'inconnu'), ce qui explique leur faibles résultats globaux. La fusion des prédictions permet d'exploiter leur très grande précision quand ils se prononcent (prédiction 'vrai' ou 'faux') et de se reporter sur d'autres systèmes sinon.

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

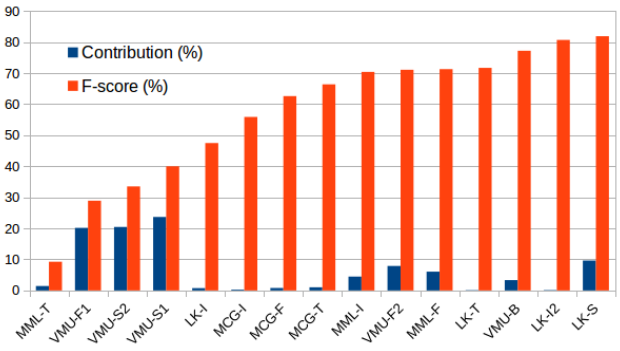

Figure 8. Contributions de chacun des systèmes dans la fusion mesurée par indice de Gini sur les forêt aléatoires, mis en regard de la F-mesure de ces systèmes ; évaluation par groupe de messages partageant un même contenu multimédia

n'obtenons aucune information sur les apports de chaque approche lors de la fusion directe. Nous examinons ces deux points dans les sous-sections suivantes.

6.2. Fusion des prédictions élémentaires

Les résultats d'une fusion directe des huit prédictions élémentaires définies précédemment (`LK-T`, `LK-I`, `LK-S`, `VMU-S1`, `MML-T`, `MML-I`, `MCG-T` et `MCG-I` ; voir section 5.2) sont présentés dans le tableau 6. Le système de référence est de nouveau le vote majoritaire sur les huit prédictions en entrée. Dans le cas d'une égalité entre *vrai* et *faux*, la classe *inconnu* est utilisée.

Tableau 6. Performances (%) de la fusion sur les huit prédictions élémentaires ; évaluation par groupe de messages partageant un même contenu multimédia

Fusion	Majorité	SVM	Arbre de déc.	Random Forest	NN
F-Mesure	$88,5$	$88,6$	$88,5$	$90,0^*$	$91,3^*$
Taux de B.C.	$93,3$	$92,8$	$92,3$	$95,0^*$	$95,9^*$

On note alors que, malgré le retrait de la moitié des prédictions en entrée, il reste possible de classer correctement $95,0\,\%$ des images et de leur tweets associés. La fusion apporte donc encore un gain absolu de $10\,\%$ par rapport au meilleur système (`LK-S` dans ce scénario d'évaluation). Il est également intéressant de comparer ces résultats à ceux du tableau 5. On obtient notamment de meilleurs résultats avec le système de référence en ne retenant que les prédictions élémentaires. Cela s'explique aisément, puisque par définition le vote par majorité est sensible aux doublons (et plus largement aux corrélations) induits par les runs incluant déjà de la fusion. Les méthodes de classification réputées peu sensibles à ces phénomènes de corrélations entre attributs, comme les *Random Forest*, obtiennent logiquement des résultats équivalents. La fusion dans ce cas repose en partie sur des systèmes différents de ceux vus

précédemment, comme on peut l'observer dans la figure 9, mais offre finalement des performances identiques.

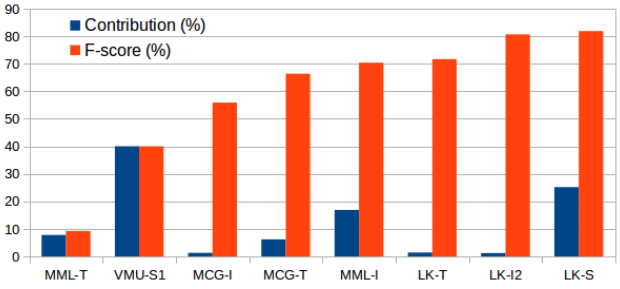

Figure 9. Contributions de chacun des systèmes dans la fusion des systèmes élémentaires, mesurée par indice de Gini sur les forêt aléatoires, et mis en regard de la F-mesure de ces systèmes

6.3. Influence des connaissances externes dans la fusion

Certains des huit systèmes élémentaires exploitent des connaissances externes aux données d'entraînement. Il s'agit d'une part des approches se fondant sur l'identification des sources, pour lesquels des listes blanches ou noires de sources ont été compilées manuellement pour construire ces systèmes (`LK-S` et `VMU-S1`). Et d'autre part, cela concerne les approches `MML-I` et `LK-I` dans lesquelles des bases d'images externes sont utilisées pour comparaison. Il est légitime de s'interroger sur l'influence de ces connaissances externes dans les résultats obtenus, notamment du fait de la grande contribution des approches fondées sur les sources dans l'expérience précédente. Nous proposons dans le tableau 7 les résultats obtenus par les mêmes expériences de fusions, restreintes aux approches élémentaires n'utilisant aucune ressource externe aux données d'entraînement.

Tableau 7. Performances (%) de la fusion sur les prédictions n'utilisant pas de connaissances externes ; évaluation par groupe de messages partageant un même contenu multimédia

Fusion	Majorité	SVM	Arbre de déc.	Random Forest	NN
F-Mesure	63,3	60,0	59,4	60,4	61,1
Taux de B.C.	61,8	59,8	60,3	60,7	62,1

Les performances sont cette fois-ci inférieures aux précédentes tentatives de fusion, et même inférieures à certaines des méthodes prises isolément (cf. tableau 3). Ce dernier point montre d'une part que les quatre méthodes restantes prédisent des classes différentes (cela transparaît aussi avec les scores du système de référence), et qu'il est difficile de trouver une régularité pour privilégier une méthode plutôt qu'une autre (scores des techniques d'apprentissage inférieurs à la référence). Enfin, il ressort clairement l'importance des ressources externes utilisées dans certains systèmes des

participants, puisque leur absence entraîne une chute de 25% des performances de la fusion.

6.3.1. Fusion par modalités

À partir de l'ensemble des huit prédictions élémentaires, nous proposons une fusion à deux niveaux dans laquelle les messages sont classés selon les trois modalités (texte, source ou image) puis un classifieur regroupe ces trois prédictions de 1^{er} niveau.

Le tableau 8 présente tout d'abord les résultats des trois classifieurs de premier niveau (prédiction au niveau du texte, source ou image). Une première constatation est le résultat encourageant du classifieur réalisant la fusion des prédictions *texte*. En effet, les résultats sont nettement supérieurs à ceux des systèmes pris individuellement. Pour les sources, le gain de la fusion est là aussi présent. En revanche, la fusion des approches image a plutôt tendance à produire des résultats moins bons que le meilleur système.

Tableau 8. Performances (%) de la fusion selon les différents niveaux et les différentes modalités ; évaluation par groupe de messages partageant un même contenu multimédia.

1^{er} niveau de prédiction		SVM	arbre de dec.	Random Forest	NN
Texte	F-Mesure	$89,7$	$76,8$	$89,7$	$88,9$
	Taux de B.C.	$94,3$	$82,0$	$94,3$	$93,8$
Source	F-Mesure	$89,6$	$83,2$	$89,6$	$89,2$
	Taux de B.C.	$93,9$	$87,9$	$93,9$	$93,9$
Image	F-Mesure	$68,8$	$56,6$	$68,2$	$68,4$
	Taux de B.C.	$68,7$	$63,0$	$67,1$	$68,9$

Pour implémenter la fusion de second niveau, nous nous appuyons sur les réseaux de neurones, qui sont simples à mettre en place et donnent de bons résultats dans toutes les expériences de fusion précédentes. L'architecture du réseau reflète notre approche à deux niveaux : les trois réseaux de neurones correspondant à la fusion de chacun des trois groupes de système (texte, source et image) sert à alimenter un réseau de neurones de second niveau (même architecture que dans les autres expériences). Pour entraîner ce réseau, nous testons deux approches (notée entraînement 1 et 2 par la suite) :

1. les réseaux texte, image et sources sont entraînés individuellement, et le réseau de second niveau est ensuite entraîné à partir de leurs prédictions ;

2. tout le réseau est entraîné d'un seul bloc.

Nous présentons les résultats de la fusion de second niveau avec les deux stratégies d'entraînement dans le tableau 9.

Comme on peut le constater, les résultats dans les deux cas sont très bons, mais on constate un intérêt à entraîner tout le réseau d'un bloc plutôt que par niveau. La différence est statistiquement significative (test de Wilcoxon avec $p = 0,05$). En effet,

Tableau 9. Performances (%) de la fusion à deux niveaux par réseau de neurones selon la stratégie entraînement ; évaluation par groupe de messages partageant un même contenu multimédia.

Prédiction en deux niveaux	entraînement 1	entraînement 2
F-Mesure	$91,2$	$94,2^*$
Taux de B.C.	$95,1$	$97,8^*$

avec la stratégie d'entraînement 1, les résultats sont du même niveau que ceux d'une fusion directe de toutes les méthodes.

7. Conclusion

Dans cet article, nous proposons et examinons plusieurs stratégies de fusion 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 reposent sur des ressources externes (listes blanches ou noires de sources) dont la construction et l'entretien peut ne pas sembler crédible dans une application à très large échelle (tweets venant de différents pays, en différentes langues, par exemple). Les approches fondées sur l'analyse des images obtiennent en général des résultats individuels décevants du fait de leur incapacité à se prononcer sur de nombreux cas. En revanche, fusionnées à d'autres approches, elles peuvent se révéler apporter une information complémentaire améliorant les performances globales d'un système. Plus largement, nous avons d'ailleurs constaté que ce ne sont pas forcément les approches réalisant les meilleurs scores individuels qui contribuent le plus au système de fusion. Les systèmes de fusion par apprentissage que nous avons proposés permettent en effet d'exploiter la grande précision de certains systèmes tout en compensant leur faible rappel avec d'autres méthodes. Enfin, le résultat principal de cet article est l'intérêt de proposer des systèmes fusionnant des approches différentes. La stratégie la plus performante semble de le faire par niveau en groupant les méthodes travaillant sur le même type d'information (texte, image, source). Une mise en oeuvre de cette approche à deux niveaux avec un réseau de neurones donne en effet de très bons résultats, significativement meilleurs que les autres approches explorées dans cet article.

Beaucoup de pistes restent ouvertes à l'issue de ce travail. Nous développons actuellement des jeux de données devant permettre de confronter les approches existantes à des cas plus nombreux et plus variés (tweets, mais aussi articles de blogs ou de sites d'opinion et de journaux). Ces jeux de données sont mis à disposition sur le site http://hoaxdetector.irisa.fr/.

D'un point de vue technique, les travaux futurs viseront à corriger certains problèmes, mis en évidence par nos expérimentations, des systèmes s'appuyant sur les images (e.g. images modifiées considérées comme similaires à l'image réelle initiale, images non retrouvées). Nous prévoyons ainsi d'étendre la couverture de la base

d'images. Nous explorons également des pistes d'amélioration du module de comparaison de contenu, notamment en effectuant des post-traitements pour éliminer les faux-positifs lors de la reconnaissance d'images similaires, et le repérage des zones modifiées dans ces images (Maigrot *et al.*, 2017). 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).

Enfin, d'un point de vue applicatif, la présentation des informations à l'utilisateur doit aussi être étudiée. Il semble peu opportun qu'un système implémente une censure stricte de messages jugés 'faux', mais la présentation d'éléments douteux soulève des défis d'ordre cognitif (acceptation du jugement de la machine), d'interface homme machine, mais aussi d'apprentissage, notamment lorsque la décision est, comme nous l'étudions dans cet article, issue de multiples systèmes fusionnés par des techniques permettant difficilement l'explicativité de la décision finale (notamment pour les réseaux de neurones).

Remerciements

Ce travail a bénéficié d'une aide de l'État attribuée au labex COMIN LABS et gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'avenir » portant la référence ANR-10-LABX-07-01.

Bibliographie

Bianchi T., Piva A. (2012). Image forgery localization via block-grained analysis of jpeg artifacts. *IEEE Transactions on Information Forensics and Security*, p. 1003–1017.

Boididou C., Middleton S., Papadopoulos S., Dang-Nguyen D.-T., Riegler M., Boato G. *et al.* (2016). The VMU participation @ verifying multimedia use 2016. In *Mediaeval 2016 workshop.* Amsterdam.

Boididou C., Papadopoulos S., Dang-Nguyen D.-T., Boato G., Riegler M., Middleton S. E. *et al.* (2016). Verifying multimedia use at mediaeval 2016. In *Mediaeval 2016 workshop.*

Breiman L., Friedman J. H., Olshen R. A., Stone C. J. (1984). *Classification and regression trees*. Belmont, California, U.S.A., Wadsworth Publishing Company.

Cao J., Jin Z., Zhang Y., Zhang Y. (2016). Mcg-ict at mediaeval 2016: Verifying tweets from both text and visual content. *MediaEval 2016 Workshop*.

Ciampaglia G. L., Shiralkar P., Rocha L. M., Bollen J., Menczer F., Flammini A. (2015). Computational fact checking from knowledge networks. *CoRR*, vol. abs/1501.03471. Consulté sur http://arxiv.org/abs/1501.03471

Claveau V., Raymond C. (2017, juin). IRISA at DeFT2017 : classification systems of increasing complexity . In *DeFT 2017 - Défi Fouille de texte*, p. 1-10. Orléans, France. Consulté sur https://hal.archives-ouvertes.fr/hal-01643993

Declerck T., Lendvai P. (2015). Processing and normalizing hashtags. *Recent Advances in Natural Language Processing (RANLP)*, p. 104–109.

Derczynski L., Maynard D., Rizzo G., Erp M. van, Gorrell G., Troncy R. *et al*. (2015). Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, vol. 51, n° 2, p. 32–49.

Foteini, Mezaris V., Patras B.-M., Ioannis. (2016). Online multi-task learning for semantic concept detection in video. In *Image processing (icip), 2016 IEEE international conference on*, p. 186–190.

Golbeck J., Hendler J. (2006). Inferring binary trust relationships in web-based social networks. *ACM Transactions on Internet Technology (TOIT)*, vol. 6, n° 4, p. 497–529.

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

Gottron T., Schmitz J., Middleton S. (2014). Focused exploration of geospatial context on linked open data. In *3rd international conference on intelligent exploration of semantic data (ICIESD)*.

Gupta M., Zhao P., Han J. (2012). Evaluating event credibility on twitter. In *2012 SIAM International Conference on data mining*, p. 153–164.

Hassan N., Li C., Tremayne M. (2015). Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th acm international on conference on information and knowledge management*, p. 1835–1838. New York, NY, USA, ACM. Consulté sur http://doi.acm.org/10.1145/2806416.2806652

Jin F., Dougherty E., Saraf P., Cao Y., Ramakrishnan N. (2013). Epidemiological modeling of news and rumors on twitter. In *Proceedings of the 7th workshop on social network mining and analysis*, p. 8.

Li W., Yuan Y., Yu N. (2009). Passive detection of doctored jpeg image via block artifact grid extraction. *89th Signal Processing*, p. 1821–1829.

Maigrot C., Kijak E., Sicre R., Claveau V. (2017). Tampering detection and localization in images from social networks: A CBIR approach. In *Image analysis and processing - ICIAP 2017 - 19th international conference, catania, italy, september 11-15, 2017, proceedings, part I*, p. 750–761. Consulté sur https://doi.org/10.1007/978-3-319-68560-1_67

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

Middleton S. (2015b). Reveal project-trust and credibility analysis. *Mediaeval 2015 Workshop*.

Page L., Brin S., Motwani R., Winograd T. (1999). The pagerank citation ranking: bringing order to the web. *Stanford InfoLab*.

Pasquini C., Pérez-González F., Boato G. (2014). A benford-fourier jpeg compression detector. In *Ieee international conference on image processing (icip)*, p. 5322–5326.

Phan Q.-T., Budroni A., Pasquini C., De Natale F. (2016). A hybrid approach for multimedia use verification. In *Mediaeval 2016 workshop*.

Robertson S. E., Walker S., Hancock-Beaulieu M. (1998). Okapi at TREC-7: Automatic Ad Hoc, Filtering, VLC and Interactive. In *7th text retrieval conference (trec)*, p. 253–264.

Silverman C. (2014). *Verification handbook: An ultimate guideline on digital age sourcing for emergency coverage* (C. Silverman, Ed.). The European Journalism Centre (EJC).

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

Tolias G., Sicre R., Jégou H. (2016). Particular object retrieval with integral max-pooling of cnn activations. In *4th international conference on learning representations (iclr)*.

Vosoughi S. (2015). *Automatic detection and verification of rumors on twitter*. Thèse de doctorat non publiée, Massachusetts Institute of Technology.

Wan J., Wang D., Hoi S. C. H., Wu P., Zhu J., Zhang Y. *et al*. (2014). Deep learning for content-based image retrieval: A comprehensive study. In *22nd ACM international conference on multimedia (icm)*, p. 157–166.

Wang S., Terano T. (2015). Detecting rumor patterns in streaming social media. In *Big data (big data), 2015 ieee international conference on*, p. 2709–2715.

Zampoglou M., Papadopoulos S., Kompatsiaris Y. (2015). Detecting image splicing in the wild (web). In *Multimedia & expo workshops (icmew)*, p. 1–6.

Zubiaga A., Liakata M., Procter R., Bontcheva K., Tolmie P. (2015). Towards detecting rumours in social media.