Détection de fausses informations dans les réseaux sociaux

Cédric Maigrot

UNIVERSITE BRETAGNE LOIRE MATHSTIC

UNIVERSITÉ DE RENNES 1

THÈSE DE DOCTORAT DE

L'UNIVERSITE DE RENNES 1
COMUE **U**NIVERSITE **B**RETAGNE **L**OIRE

École Doctorale N°601
Mathématique et Sciences et Technologies de l'Information et de la Communication
Spécialité : Informatique

Par

« Cédric MAIGROT »

« Détection de fausses informations dans les réseaux sociaux »

Thèse à soutenir le 1*er* avril 2019
Unité de recherche :
Thèse N° :

Rapporteurs avant soutenance :

Xavier TANNIER, Professeur à Sorbonne Université
Benoit HUET, Maître de conférences à Eurecom

Composition du jury :

Examinateurs : Xavier TANNIER, Professeur à Sorbonne Université
Benoit HUET, Maître de conférences à Eurecom
Julien VELCIN, Professeur à l'Université de Lyon 2
François TAÏANI, Professeur à l'Université de Rennes 1
Dir. de thèse : Laurent AMSALEG, Chercheur au CNRS

Co-encadrante : Ewa KIJAK, Maître de conférence à l'Université de Rennes 1

REMERCIEMENTS

Je tiens à remercier Madame Ewa Kijak, Maître de Conférences à l'Université de Rennes 1, et Monsieur Vincent Claveau, chercheur au CNRS, qui m'ont encadré tout au long de cette thèse et qui m'ont fait partager leurs remarques et conseils nombreux. Je souhaite, de plus, les remercier pour leur gentillesse et leur disponibilité permanente.

Je tiens aussi à remercier Monsieur Laurent Amsaleg, chercheur au CNRS, pour avoir dirigé ma thèse et pour ses conseils précieux à chaque fois que nous avons discuté.

J'adresse tous mes remerciements à Monsieur Xavier Tannier, Professeur à Sorbonne Université, ainsi qu'à Monsieur Benoit Huet, Maître de conférences à Eurecom, de l'honneur qu'ils m'ont fait en acceptant d'être rapporteurs de cette thèse.

Madame Sandra Bringay, Professeur des Universités à Montpellier, et Monsieur Jérôme Azé, Professeur des Universités à Montpellier, pour m'avoir initié à la recherche alors qu'ils encadraient mon stage de recherche dans le cadre de la validation de mon diplôme de Master et pour m'avoir convaincu de poursuivre mes études dans la réalisation d'une thèse.

J'exprime ma gratitude à Monsieur Julien Velcin, professeur à l'Université de Lyon 2, et Monsieur François Taïani, Professeur à l'Université de Rennes 1, qui a bien voulu être examinateurs.

Je tiens aussi à remercier Monsieur Jean-Marc Jézéquel, Directeur de l'IRISA et Professeur de l'Université de Rennes 1, qui m'a accueilli pendant deux ans au sein de son laboratoire.

Cette thèse n'aurait pu avoir lieu sans le financement de l'Université de Rennes 1 et de la Direction Générale de l'Armement (DGA) que je remercie chaleureusement.

Je tiens à remercier tous les membres de l'équipe Linkmedia, pour les discussions que nous avons eu durant toute la durée de la thèse sur ce sujet passionnant. Enfin, mes derniers remerciements vont à ma famille et mes amis proches sans qui, de part leur soutien moral sans faille, cette thèse n'aurait pu avoir lieu.

SOMMAIRE

LISTE DES TABLEAUX

TABLE DES FIGURES

CHAPITRE 1

CONTEXTE

CHAPITRE 2

ÉTAT DE L'ART

Contents

2.1 Introduction

La détection automatique de fausses informations dans les réseaux sociaux est devenue ces dernières années un sujet suscitant l'intérêt de nombreuses équipes de recherche. Cet intérêt croissant s'explique premièrement par l'actualité (voir chapitre 1), mais aussi par les défis que cela représente au niveau scientifique.

Cette problématique s'applique à de nombreux réseaux sociaux ayant des spécificités très diverses. Cependant nous pouvons représenter l'ensemble de ces messages étudiés d'une façon schématisée et générique comme présentée dans la figure 2.1 en nous basant sur les points communs de tous ces messages issus des différents réseaux sociaux.

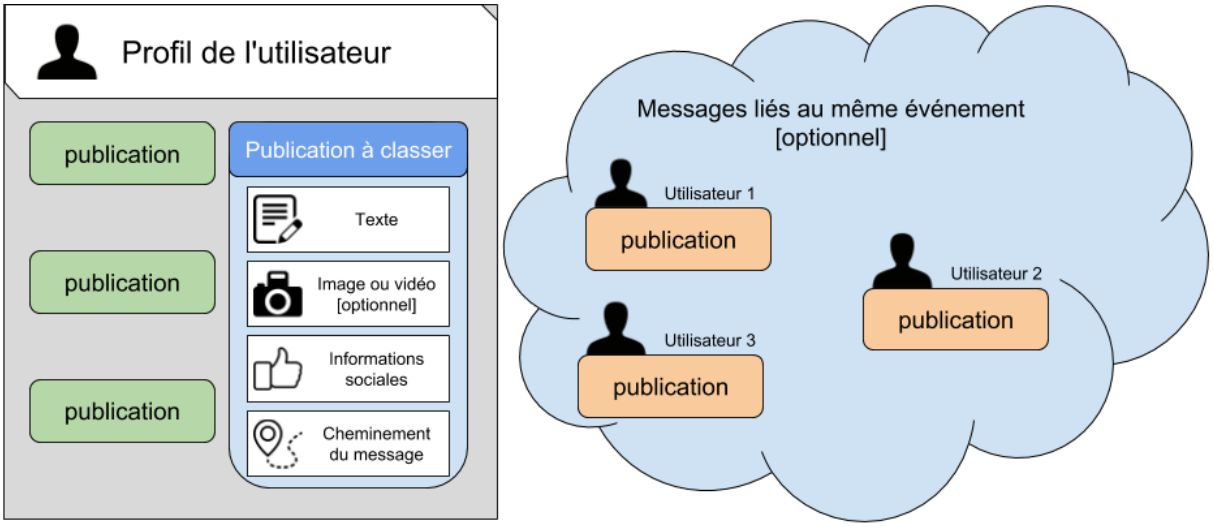


FIGURE 2.1 – Représentation générique d'une publication issue des réseaux sociaux

Chaque publication peut être vue comme la combinaison de cinq éléments :

— **Le texte** contenu dans la publication. Il peut contenir des liens vers des pages internets sous forme d'URLs, des émoticônes ou encore des éléments spécifiques au réseau social (par exemple des hashtags sur *Twitter*) ;

— **Le contenu multimedia** pouvant, si il existe, compléter le contenu textuel. Cela peut correspondre à une image ou une vidéo ;

— **Les informations sociales**, qui sont une spécificité des réseaux sociaux, représentent les interactions entre les utilisateurs. Ces actions se font pour la plupart sur les publications des autres utilisateurs. Parmi les actions courantes des réseaux sociaux, nous pouvons noter le partage de la publication (par exemple

le *retweet* sur *Twitter*) et l'approbation de la publication (par exemple la mention *J'aime* sur *Facebook*). Certaines informations peuvent aller plus loin que le simple approuvement ou désapprouvement. *Facebook* par exemple propose six réactions aux publications : *Like*, *Love*, *Haha*, *Wow*, *Sad* et *Angry* ajoutant alors du sentiment dans la réaction avec respectivement un approuvement neutre, de l'amour, de la joie, de la surprise, de la tristesse et de la colère. Le profil de l'utilisateur regroupant les informations connues sur l'auteur de la publication est associé à cette partie.

— **Le cheminement du message** fait intervenir l'historique du message, soit les personnes l'ayant publié précédemment. Ce cas peut intervenir par le partage de la publication ;

— **L'événement lié à la publication** peut être vu comme l'ensemble des messages évoquant le même sujet que la publication. Il n'y a pas obligatoirement d'autres messages si la publication aborde une sujet nouveau.

La détection automatique de fausses informations est présentée dans ce manuscrit comme un problème de classification illustré dans la figure 2.2. Ce problème se décompose en trois parties :

— Les **données** en entrée du système. Nous avons la possibilité d'utiliser chaque modalité des informations disponibles présenté précédemment ou n'importe quelle combinaison de modalités ;

— Le **classifieur** correspond à la partie qui apprend à reconnaître les fausses informations ;

— Les **classes** sont les différentes réponses possibles du classifieur : `vrai` ou `faux`. Il est important de noter que nous nous intéressons aussi, dans la mesure du possible, à la capacité d'expliquer le choix du classifieur.

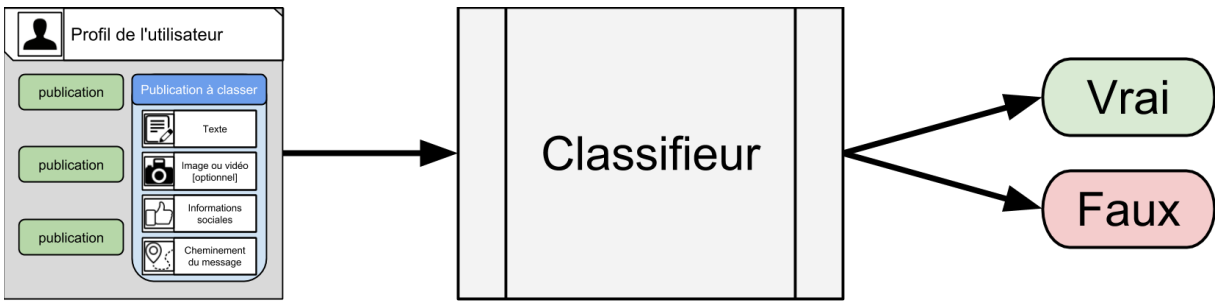


FIGURE 2.2 – Représentation du problème de classification des publications

La suite de cet état de l'art fournit un aperçu des différentes approches possibles de la détection de fausses informations. Deux sections sont consacrées aux modalités TEXTE et *image* qui sont les deux modalités sur lesquels ces travaux sont axés respectivement dans les sections 2.2 et 2.3. Une section est ensuite consacrée aux travaux sur les autres modalités 2.4. Enfin, ce chapitre se termine sur une discussion des choix de méthodes pour les travaux présentés dans ce manuscrit.

2.2 Approches basées sur le texte

Le texte est une des composantes obligatoires dans toutes les publications. Le traitement de cette modalité peut prendre plusieurs formes allant de la représentation utilisant des descripteurs lexicaux, des descripteurs syntaxiques et des descripteurs de sujet.

2.2.1 Descripteurs lexicaux

Les descripteurs lexicaux sont calculés au niveau des mots. On y retrouve les comptages (*e.g.* nombre de points d'interrogation), les motifs associés aux fausses informations et les lexiques de sentiments.

Une des premières études utilisant ce type de descripteurs est proposée par [CASTILLO, MENDOZA et POBLETE 2011] qui s'intéressent à définir des descripteurs de comptage au niveau des messages (*e.g.* nombre de mots dans le texte) pour classer des tweets selon leur crédibilité (utilisation de deux classes *crédible* et *non crédible*). Ces travaux ont montré des différences entre les descripteurs de surface, selon leur propagation, pouvant être utilisé pour la classification. Deux exemples des descripteurs les plus discriminants sont *"le tweet possède une URL"* et *"le tweet possède des points d'interrogation"*.

Basé sur ces travaux, [KWON, CHA, JUNG, W. CHEN et al. 2013] proposent une approche plus orientée sur texte et appliquée cette fois à la problématique de classer un message comme *vrai* ou *faux*. Les auteurs recherchent notamment la présence de pronoms de la première personne, pronoms de la deuxième personne et de pronoms de la troisième personne avec par exemple respectivement les pronoms *je*, *tu* et *il*. Certains éléments du texte permettent ainsi d'améliorer les prédictions par rapport aux descripteurs proposés par [CASTILLO, MENDOZA et POBLETE 2011]. Parmi ceux

là, on retrouve la présence de *"le tweet possède des mots avec une orientation de sentiment positive"*.

Ce type de descripteur est intéressant pour la détection de fausses informations, car il n'est pas possible de se baser exclusivement sur le contenu du message. L'entrainement d'un classifieur en utilisant exclusivement le contenu des messages pourrait engendrer un apprentissage basé sur les termes spécifiques aux fausses informations (*e.g.* une entité nommée particulière liée à une fausse information présente dans le jeu de données d'apprentissage). L'utilisation de descripteurs statistiques permet l'apprentissage de règle du type :

> "Les publications fausses sont généralement plus courtes que les informations vraies."

2.2.2 Représentation basée sur le contenu

[Z. Zhao, Resnick et Mei 2015] recherchent une forme d'interrogation par l'auteur et de correction d'une autre publication des messages de fausses informations. Les auteurs extraient les éléments les plus discriminants pour chaque classe (*vrai* ou *faux*) en utilisant une approche basée les caractéristiques du *Term Frequency* ou fréquence du terme (TF) des messages. Cette liste d'indices est ensuite étudiée par des experts qui sélectionnent des phrases indépendantes des événements comme modèles lexicaux finaux pour les fausses informations. Ces travaux présentent l'avantage de passer par une phase d'analyse par des experts ce qui renforce la cohérence des modèles lexicaux trouvés.

Les mots lexicaux exprimant des sentiments spécifiques sont également des indices très importants pour caractériser le texte. Dans [Castillo, Mendoza et Poblete 2011], les marques émotionnelles (point d'interrogation et point d'exclamation) et les émoticônes sont considérés comme des caractéristiques textuelles.

L'émotion véhiculée dans le texte est aussi étudiée par [Kwon, Cha, Jung, W. Chen et al. 2013] où de nombreuses caractéristiques lexicales associées aux sentiments sont proposées sur la base de dictionnaires. Après une étude comparative de ces caractéristiques, les auteurs constatent que certaines catégories de sentiments sont des caractéristiques distinctives de la détection des fausses informations, notamment les mots à effet positif, les mots d'action cognitive et les mots d'action provisoire.

Ce type d'approches peut nous permettre d'apprendre des motifs propres aux

textes associés à un fausse information comme par exemple : "C'est photoshopé !"

2.2.3 Utilisation de réseaux de neurones

L'utilisation de représentations par plongement de mots est de plus en plus courante. Ces représentations sont obtenues avec des réseaux de neurones entraînés pour reconstruire le contexte linguistique des mots [MIKOLOV et al. 2013]. Plusieurs travaux se sont inspirés de ces représentations pour la détection de fausses informations.

Certains mots malveillants dans le contenu peuvent être fortement liés à la catégorie des fausses informations. Pour mieux comprendre les mots auxquels le modèle prête plus d'attention, [T. CHEN et al. 2017] proposent une utilisation d'un mécanisme d'attention. L'une des hypothèses de leur travail est que les caractéristiques textuelles des fausses informations peuvent changer d'importance avec le temps et qu'il est crucial de déterminer lesquelles sont les plus importantes pour la tâche de détection. Semblable à [MA et al. 2016], ils regroupent d'abord les publications par intervalle de temps. À chaque pas de temps, l'état caché d'un *Recurrent Neural Network* ou réseau de neurones récurrents (RNN) se verra attribuer un paramètre de pondération pour mesurer son importance et sa contribution aux résultats. La performance des expériences démontre l'efficacité du mécanisme d'attention et montre que la plupart des mots liés à l'événement lui-même sont moins utilisés que les mots exprimant le doute, l'esquive et la colère des utilisateurs causés par la fausse information.

[NGUYEN, C. LI et NIEDERÉE 2017] se concentrent sur la détection au début de la propagation de la fausses informations et proposent un modèle basé sur un *Convolutional Neural Network* ou réseau neuronal convolutif (CNN) et un RNN comme montré dans la figure 2.3. Le CNN est appliquée sur les tweets pour créer une séquence de représentations de phrases de haut niveau afin d'apprendre les représentations cachées de tweets liés à des fausses informations et ainsi prédire la véricité au niveau de chaque tweet. Ensuite, la partie de RNN est utiliser pour analyser les séries temporelles (séries de prédictions au niveau des publications) obtenues par CNN pour obtenir une prédiction finale.

Une des limites à l'utilisation de ces approches par plongement de mots pour nos travaux est un problème courant des réseaux sociaux, c'est à dire la capacité pauvre des réseaux de neurones à expliquer le choix de la classification. Or, il s'agit d'un

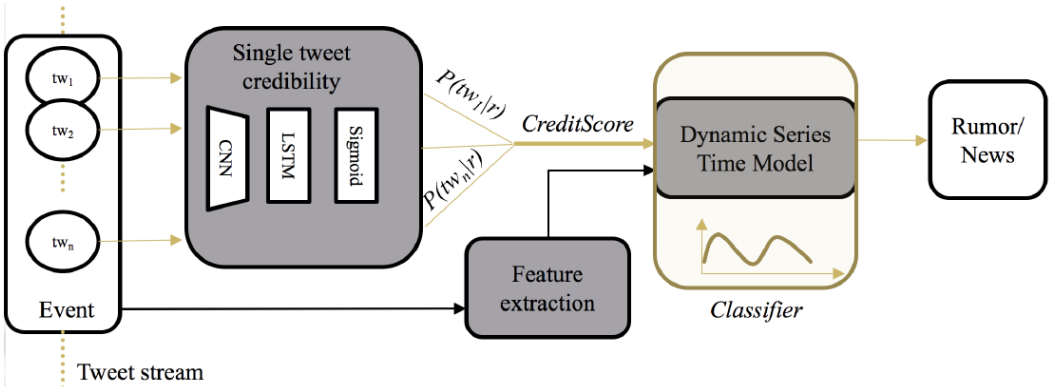


FIGURE 2.3 – Illustration de l'approche proposée par [NGUYEN, C. LI et NIEDERÉE 2017]

point important quant à la confiance que doit avoir l'utilisateur auprès du système lui signifiant qu'une information est fausse.

2.3 Approches basées sur le contenu multimédia

La détection de modifications dans une image est un sujet qui se révèle vaste du fait du large panel de possibilités quant aux types de modifications possibles.

La présentation des techniques de détection de modifications dans une image est faite dans cette section en suivant la classification la plus courante dans l'état de l'art. La hiérarchie utilisée dans ce chapitre est présentée dans la figure 2.4. Les images numériques peuvent être manipulées en utilisant deux familles d'attaques : des méthodes passives et actives, présentées successivement dans la suite de cette section.

2.3.1 Méthodes actives

La famille des attaques actives se décompose en deux sous-familles qui sont celles basées sur la détection d'une empreinte (*watermarking* en anglais) et celles utilisant la stéganographie. Elles ont aussi la particularité d'être des modifications invisibles à l'oeil humain.

La *stéganographie* est la technique permettant de cacher un message dans une image numérique. Ce type de techniques modifie les valeurs de quelques pixels permettant de cacher un message dans l'image sans modifier l'aspect visuel de l'image.

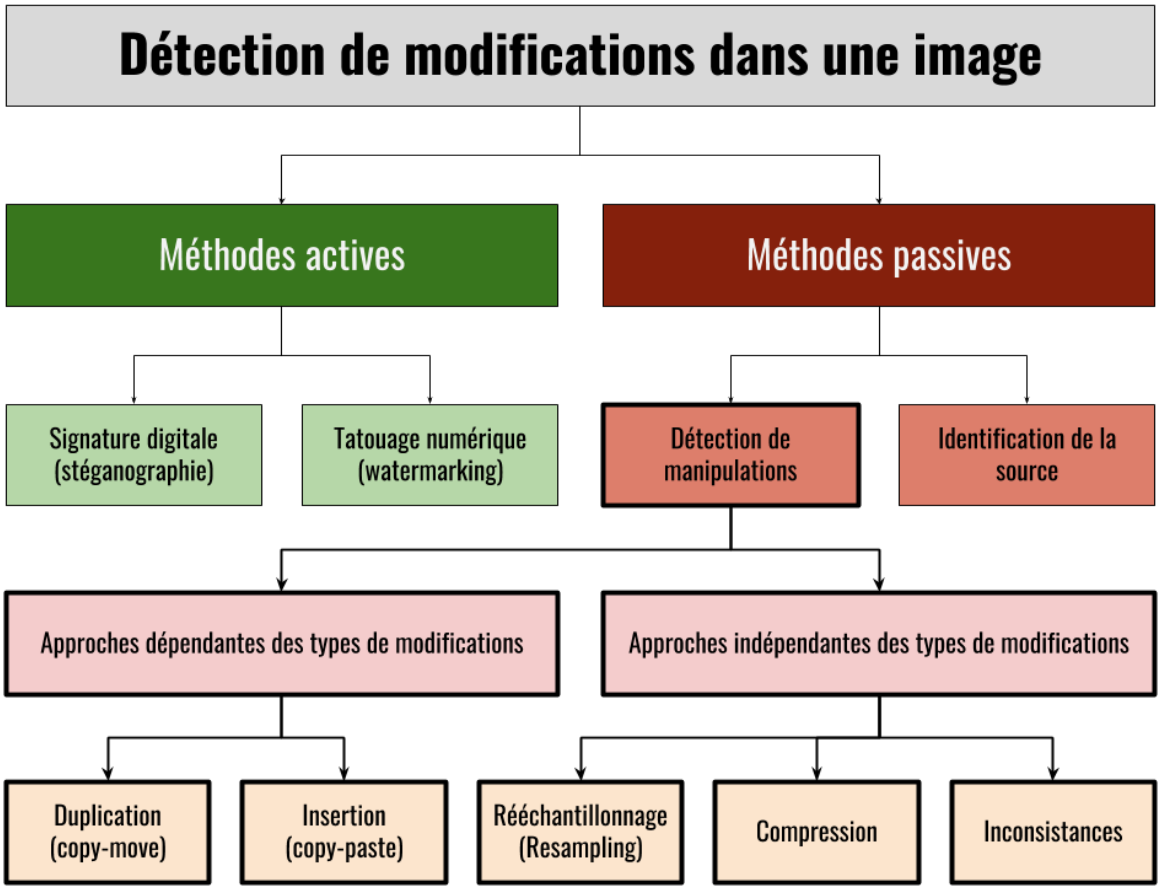


FIGURE 2.4 – Hiérarchie des types de méthode de détection de modifications dans une image inspirée par [BIRAJDAR et MANKAR 2013] et [MUSHTAQ et MIR 2014a]

Ces approches ne sont pas utiles dans notre problématique car nous nous intéressons aux images issues des réseaux sociaux ciblant le grand public et ces approches permettent de dissimuler un message dans l'image invisible à l'oeil nu. Si une telle attaque est présente sur une image se propageant sur les réseaux sociaux, personne ne le remarquera et cela n'aura aucun effet. Il n'est donc pas utile d'analyser ce type d'attaques dans notre cadre de travail.

La seconde catégorie utilise les *tatouages numériques*. Ces derniers correspondent à une technique permettant d'ajouter des informations de copyright ou d'identification à une image. En connaissant les informations servant à l'identification de l'image, il est possible de vérifier que ces mêmes informations n'ont pas été perturbées ce qui signifierait que l'image a été modifiée. De la même manière que la stéganographie, le

tatouage numérique ne vise pas à modifier l'aspect visuel de l'image. C'est pour cette raison que nous ne nous intéressons pas à ce type de modifications et par conséquence aux approches permettant de les détecter.

2.3.2 Méthodes passives

Ces méthodes visent à détecter et potentiellement localiser et compter le nombre de modifications visibles dans l'image. On distingue deux types de méthodes passives selon si la méthode mise en place est dépendante sur un type de modification ou non. Les différents types de modifications sont détaillés dans la sous-section décrivant les approches dépendantes sur un type de modification.

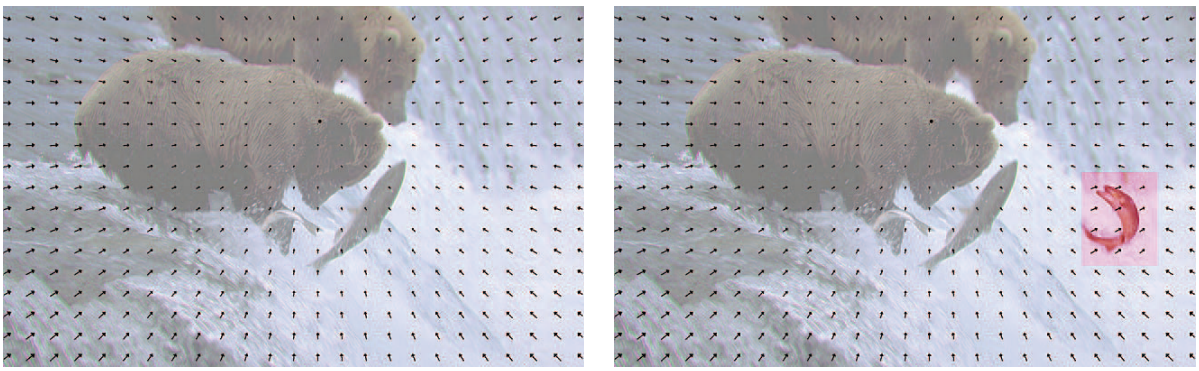
Approches indépendantes du type de modifications

Les approches indépendantes du type de modifications se décomposent en trois catégories : les approches se basant sur la compression de l'image, la détection d'inconsistances et la détection de rééchantillonnage. Cette dernière notion correspond à plusieurs définitions dans l'état de l'art, nous la définissions ici comme étant l'ensemble des modifications appliquées à l'image entière. Quelques exemples de modifications par rééchantillonnage est la rotation, la translation, le changement de couleur, etc.

Une première possibilité d'approches indépendantes du type de modifications est basée sur la compression de l'image. Ainsi, certaines approches utilisent les particularités des différents formats d'image pour détecter des incohérences dans les valeurs des pixels. Le format le plus étudié est *Joint Photographic Experts Group* (JPEG) de part sa capacité à compresser l'image et à ne pas mémoriser l'intégralité des données (perte de données engendrée par la compression). Plusieurs travaux, tels que [Bianchi et Piva 2012 ; Q. Wang et R. Zhang 2016 ; J. Li et al. 2018], s'intéressent à déterminer si une image est doublement compressée ou non. L'idée étant qu'une image est une première fois compressée lors de la prise de la photo, puis le sera une seconde fois au moment de l'enregistrement si la photo vient d'être modifiée. L'intérêt est que la zone modifiée sera compressée d'une seule fois (*i.e.* la seconde fois). Il est donc utile de savoir si une image est simplement compressée ou doublement compressée. Ce type d'approches permet aussi de localiser les modifications, le processus permettant de trouver les zones perturbées, mais nécessite de travailler avec un format d'images particulier ce qui n'est pas le cas des images circulant sur les réseaux

sociaux. L'utilisation des réseaux de neurones a aussi été abordée dans le cadre des approches basées sur le format de l'image par [B. Li et al. 2017]. Ici, les auteurs souhaitent utiliser les valeurs des coefficients *Discrete Cosine Transform* ou transformée en cosinus discrète en français (DCT) pour apprendre à prédire si une image au format JPEG est simplement ou doublement compressée.

Une seconde catégorie d'approches ne nécessitant pas de se focaliser sur un type d'attaque est basée sur les inconstances dans l'image. Ces approches se focalisent sur la détection d'aberrations chromatiques et incohérences de la lumière. Un exemple d'approche utilisant ce principe est présenté par [Johnson et Farid 2006] et est illustré dans la figure 2.5. Dans cette approche, les auteurs s'intéressent aux décalages entre les canaux rouges et verts dûs aux capteurs de l'appareil photo lors de la prise. On remarque que la zone modifiée dans l'image 2.5(b) possède un décalage entre les deux canaux imcompatibles avec le reste de l'image.



(a) (b)

Figure 2.5 – Exemple d'application de l'approche proposée par [Johnson et Farid 2006] pour la détection d'aberrations chromatiques sur une image non modifiée (à gauche) et une image modifiée (à droite).

D'autres approches utilisent le signal *Photo Response Non-Uniformity* (PRNU) qui est un signal apposé sur une photo par l'appareil au moment de la prise de la photo. En cas de modification de l'image, ce signal global à l'image est perturbé dans la zone attaquée [Chierchia et al. 2014]. Cependant, ces approches se basent sur une connaissance à priori qui est la connaissance de l'appareil photo utilisé pour prendre la photo et donc du signal apposé qui est propre à chaque modèle. L'utilisation de telles techniques nécessite donc deux hypothèses : 1) la connaissance de l'appareil utilisé ; 2) la connaissance du signal correspondant à l'appareil utilisé. Des listes réfé-

rençant le signal correspondant à chaque modèle d'appareil photo existent mais cela suppose de mettre régulièrement à jour de telles listes. Ces conditions ne sont pas envisageables dans une étude d'image provenant de réseaux sociaux du fait de la propagation de l'image. Des travaux visent pourtant à limiter ces conditions afin de se rapprocher des situations telles que l'analyse des réseaux sociaux : [COZZOLINO, MARRA et al. 2017] utilisent le signal PRNU sans ces hypothèses. Pour cela, ils proposent de réaliser un regroupement (*clustering* en anglais) d'images provenant d'une base pour laquelle nous savons que les images d'un même regroupement proviennent d'un même appareil photo et donc ayant le même signal PRNU. Il s'agit alors d'approximer le signal PRNU de l'image à analyser et de trouver le regroupement ayant le signal PRNU le plus similaire pour une image donnée. Le signal PRNU estimé de ce cluster sera utilisé sur l'image requête (*i.e.* l'image à analyser).

Approches dépendantes du type de modifications

On distingue deux types d'attaques dans une image : la duplication et l'insertion. En spécialisant une approche à un des deux types d'attaques, il est possible de profiter de spécificités propres à ce type d'attaque pour améliorer la détection et la localisation des modifications. Les méthodes spécialisées dans ces deux types d'attaque sont présentées ci-dessous.

Méthodes adaptées à la détection d'une insertion

Insertion

L'insertion, nommée *copy-paste* ou *splicing* dans la littérature anglophone, consiste au remplacement d'une partie de l'image par une région extraite d'une autre image. Un exemple est donné dans la figure 2.6 où une partie de l'image donneuse 2.6(a) est placée dans l'image hôte 2.6(b) afin d'obtenir l'image modifiée 2.6(c).

[W. WANG, DONG et TAN 2009] et [MUSHTAQ et MIR 2014b] s'intéressent à décrire les images à partir de matrices de co-occurence sur les niveaux de gris qui servent de descripteurs pour être classé par un modèle *Support Vector Machine* dans le but

(a) (b) (c)

FIGURE 2.6 – Exemple d'image modifiée par insertion (à droite) avec les deux images originales utilisées (à gauche et au centre).

d'obtenir une classification binaire entre les classes *réel* et *modifié*. La limite de ces approches est qu'elles ne permettent pas la localisation des modifications du fait de l'utilisation d'un classifieur.

[HSU et CHANG 2006] s'intéressent à la détection de rééchantillonnage comme étant une preuve de modification de l'image. Les auteurs partent du principe que la zone insérée aura subit des modifications telles qu'un redimensionnement ou une rotation afin de correspondre avec le contenu de l'image hôte. L'utilisation d'un algorithme de maximisation de l'espérance (en anglais *expectation-maximization algorithm*) sur des zones de l'image est utilisé pour détecter si le signal a subit un rééchantillonnage ou non. Cette approche est efficace, mais doit être appliquée à des images non compressées ce qui exclu les images JPEG de l'analyse.

Une méthode automatique a été introduite dans [QU, QIU et J. HUANG 2009] pour identifier les modifications d'insertion basée sur le système visuel humain, dans laquelle la saillance visuelle et la fixation sont utilisées pour extraire des caractéristiques. Un méta-classificateur est utilisé pour classer si une image est authentique ou falsifiée. Cependant, cette technique n'est pas robuste à l'application d'un flou sur l'image.

Méthodes adaptées à la détection d'une duplication

Partant du principe que les pixels de la zone dupliquée sont en double dans l'image, les méthodes présentées ci-dessous ont pour la plupart la même stratégie qui est la recherche de pixels ou de zones identiques. Une limite est à prendre en considération : en cas de chevauchement de la région copiée et de la région collée, certains pixels de

la zone copiée sont supprimés et ne seront donc pas en double dans l'image.

Duplication

La duplication, nommée *copy-move* dans la littérature anglophone, est similaire à l'insertion à la différence que la zone insérée provient de la même image. Un exemple est donné dans la figure 2.7 où une partie de l'image originale 2.7(a) est placée dans cette même image afin d'obtenir l'image modifiée 2.7(b).

(a) (b)

FIGURE 2.7 – Exemple d'image modifiée par duplication (à droite) avec l'image originale utilisée (à gauche).

[LUO, J. HUANG et QIU 2006] présentent une méthode dans laquelle ils utilisent des blocs produits grâce à une fenêtre glissante appliquée sur l'image. Un tableau est crée contenant les descriptions de chaque zone sous la forme de vecteurs. Ce tableau est trié et la similarité est calculée entre chaque vecteurs et ceux lui étant adjacents dans le tableau. Les vecteurs avec une haute similarité avec les vecteurs adjacents dans le tableau sont considérés comme issus d'une zone dupliquée. Les zones dupliquées obtenant des vecteurs identiques ou très similaires.

[AMERINI, BALLAN, CALDELLI, DEL BIMBO et al. 2011 ; AMERINI, BALLAN, CALDELLI, DEL BIMBO et al. 2010] ont proposé une méthode en trois étapes. La première est l'extraction de descripteurs *Scale-Invariant Feature Transform* (transformation de caractéristiques visuelles invariante à l'échelle) (SIFT) et l'appariement de descripteurs similaires. Pour cela, une recherche basée sur les deux plus proches voisins parmi les

points d'intérêt est effectuée. La deuxième étape utilise une structure hiérarchique pour former des regroupements de descripteurs basés sur leur similarité. La transformation géométrique est estimée dans la troisième étape. Cette méthode permet la détection de plusieurs régions dupliquées, mais l'algorithme ne parvient pas à localiser la modification avec précision et il ne peut pas détecter laquelle des deux régions est l'originale. De plus, cette méthode utilisant la détection de points d'intérêt est sensible aux régions lisses dans l'image.

[KAUR et SHARMA 2013] a proposé une méthode spécifique aux images JPEG. La méthode utilise la combinaison des coefficients DCT et des descripteurs SIFT. Les auteurs partent du principe que les coefficients DCT sont robustes à la compression JPEG et le bruit gaussien alors que les descripteurs SIFT sont robustes à la rotation et la mise à l'échelle. Par conséquent, la méthode proposée est capable de détecter les modifications dans les images, même si elles ont été soumises à différentes opérations de post-traitement. Le taux de détection d'images modifiées est augmenté de part l'union des capacités de détection de ces deux opérations.

[HASHMI, ANAND et KESKAR 2014] propose une approche qui utilise les descripteurs *Speeded Up Robust Features* (caractéristiques robustes accélérées en français) (SURF) et *Dyadic Wavelet Transform* (DyWT) (similaire à *Discrete Wavelet Transform* (DWT), mais étant invariant au décalage). Dans cette approche, DyWT est exécuté dans une image qui divise l'image en sous-images, les points d'intérêts et les descripteurs SURF sont ensuite calculés. Enfin, le système recherche les descripteurs identiques, ou très similaires, pour détecter les duplications dans une image. Les zones ainsi détectées sont marquées comme étant les régions modifiées.

2.3.3 Méthode générique

[S. D. LIN et T. WU 2011] ont développé une approche de détection de modifications utilisant une analyse de l'effet d'une double compression dans le domaine spatial et des coefficients DCT (figure 2.8). Dans cette méthode, les auteurs mettent en place deux traitements en parallèle pour détecter d'un part les modifications par duplication et d'autre part les modifications par insertion. Cependant, ils ne peuvent pas détecter des modifications dans une image lorsque des opérations de post-traitement avancées telles que des transformations géométriques sont appliquées.

L'utilisation d'auto-encodeur a aussi été étudié dans le cadre de la détection de mo-

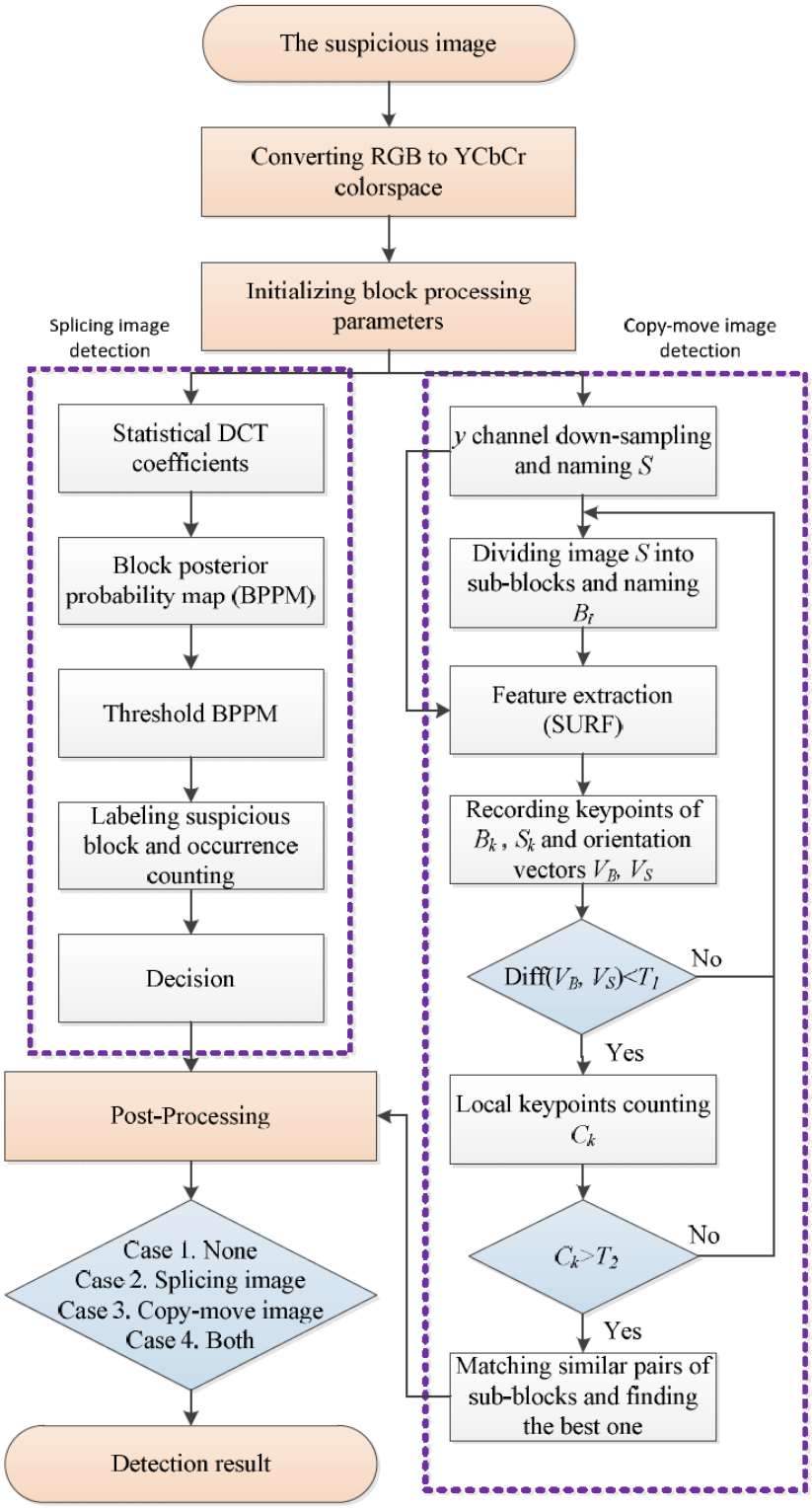


FIGURE 2.8 – Illustration de l'approche proposée par [S. D. LIN et T. WU 2011].

difications [COZZOLINO et VERDOLIVA 2016]. La figure 2.9 montre le modèle mis en place dans lequel les modifications dans l'image sont considérées comme des anomalies. Des descripteurs issus du bruit résiduel dans l'image sont extraits et envoyés à l'auto-encodeur qui va créer une représentation implicite de l'image et permettre la discrimination entre les zones originales et modifiées grâce à un processus de mise à jour des poids de manière itérative.

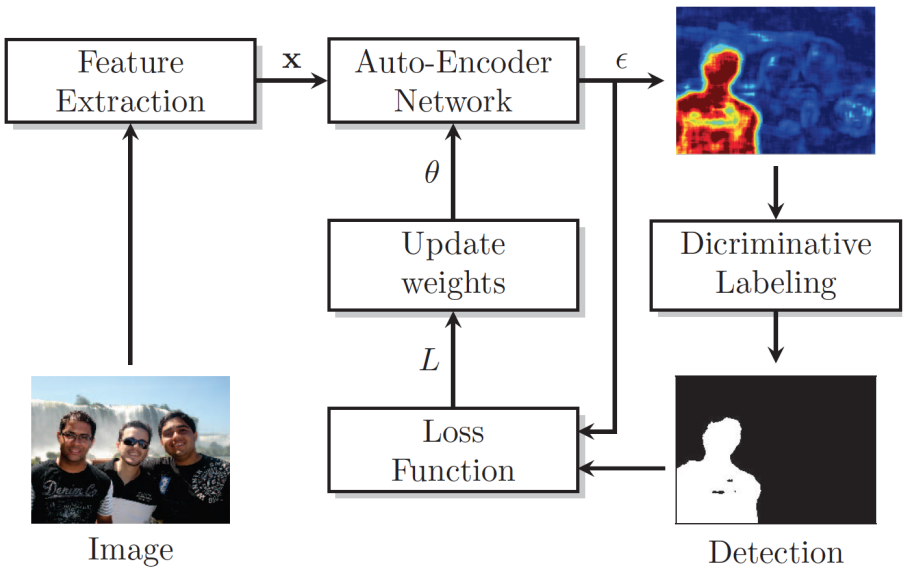


FIGURE 2.9 – Illustration de l'approche proposée par [COZZOLINO et VERDOLIVA 2016].

[SALLOUM, REN et KUO 2018] propose un réseau de neurones à convolution pour détecter et localiser les modifications dans une image (figure 2.10). Plusieurs réseaux sont testés, mais les meilleurs résultats sont obtenus grâce à un réseau multi-tâches prédisant d'une part la surface modifiée et d'autre part la limite entre la zone modifiée et non-modifiée.

2.4 Approches basées sur les autres modalités

Comme présenté en début de chapitre, les publications issues des réseaux sociaux fournissent d'autres informations en plus du texte et de l'image. Cette section vise à présenter succinctement les types d'approches existantes sur ces autres modalités. Bien que non étudiées (ou très brièvement) durant cette thèse, ces approches peuvent venir compléter celles proposées dans ce manuscrit.

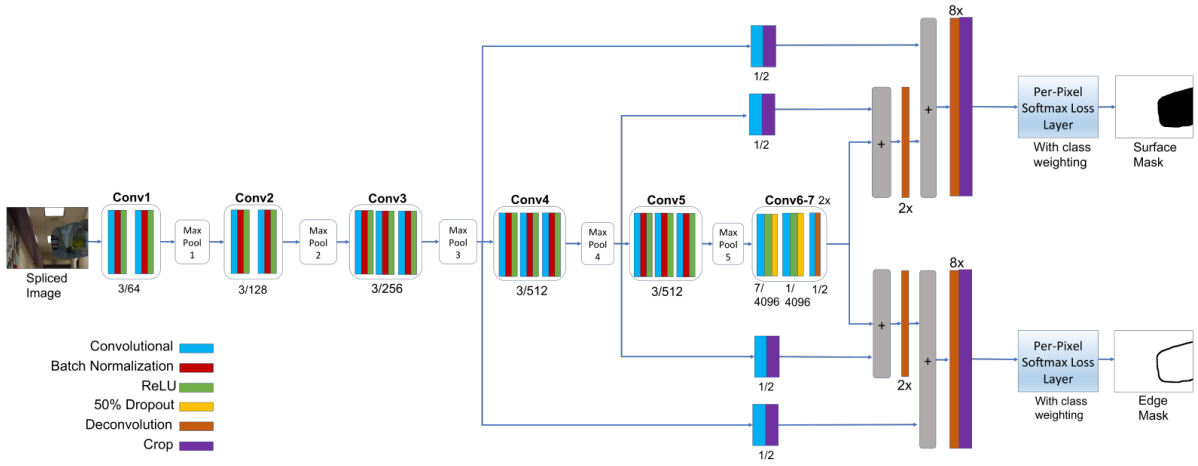


FIGURE 2.10 – Illustration de l'approche proposée par [SALLOUM, REN et KUO 2018].

2.4.1 Approches basées sur les informations sociales

L'une des principales caractéristiques des réseaux sociaux est la possibilité de réaliser toutes sortes d'interactions spécifiques aux réseaux sociaux. On peut lister trois types d'interactions sociales courantes sur les médias sociaux :

— les **interactions entre utilisateurs**, telles que *"ajouter un ami"* et *"suivre"*.
— les **liens entre les contenus** sont formés par des balises, des hashtag ou des liens URL ;
— les **interactions entre les utilisateurs et le contenu**, telles qu'une publication, l'ajout d'un commentaire ou le partage d'une publication.

De nombreux descripteurs sont dérivés de la caractéristique de connexion sociale des réseaux sociaux sur la tâche de détection des fausses informations. Les trois principaux types de descripteurs sociaux sont les descripteurs basés sur l'utilisateur, les descripteurs de propagation et les descripteurs temporels.

Les descripteurs *utilisateur* sont issus directement du réseau social d'où est extrait la publication à prédire. Les fausses informations sont créées par quelques utilisateurs et diffusées par de nombreux utilisateurs. Les descripteurs *utilisateur* visent à décrire les caractéristiques d'un utilisateur unique ou d'un groupe d'utilisateurs composé de plusieurs utilisateurs associés.

Les descripteurs individuels sont calculés sur un seul utilisateur, soit l'auteur de la publication. Ces descripteurs sont déterminés à partir du profil d'un utilisateur, telles

que la *date d'inscription*, l'*âge*, le *sexe* [MORRIS et al. 2012], ou des mesures de comptage, telles que le *nombre de followers* et le *nombre de messages postés* [CASTILLO, MENDOZA et POBLETE 2011].

[MORRIS et al. 2012] ont proposé deux descripteurs pour marquer le comportement de publication de l'utilisateur : le descripteur "client" correspond au logiciel utilisé par l'utilisateur et le descripteur "emplacement" indique si le message est envoyé à partir de l'endroit où l'événement s'est produit ou non.

Les descripteurs de groupe sont des descripteurs globaux d'un groupe dont les membres ont certains comportements similaires dans le processus de diffusion de la fausse information [F. YANG et al. 2012]. Ces descripteurs peuvent être obtenus en agrégeant les descripteurs d'un seul utilisateur, tels que *le ratio d'utilisateurs vérifiés* et *le nombre moyen de followers*.

Kwon et al. ont étudié la stabilité des caractéristiques dans le temps [KWON, CHA et JUNG 2017]. Ils ont constaté que, pour la détection des fausses informations, les descripteurs linguistiques et utilisateur sont adaptés aux stades initiaux, tandis que les caractéristiques structurelles et temporelles ont tendance à être performantes à long terme.

2.4.2 Approches basées sur le cheminement du message

Le cheminement du message correspond à l'historique du message, soit toutes les personnes ayant partagé ce message depuis son auteur original jusqu'à l'utilisateur associé au message à traiter.

Les approches présentées précédemment évaluent chaque message et chaque événement individuellement. Une observation simple est que des messages similaires ont tendance à avoir la même véracité dans un événement. Les approches basées sur la propagation sont proposées par des relations d'extraction des entités nommées et évaluent la crédibilité des messages et des événements dans leur ensemble. Le paradigme de la détection des fausses informations basée sur la propagation de la crédibilité comporte généralement deux étapes principales [GUPTA, P. ZHAO et HAN 2012 ; JIN, CAO, JIANG et al. 2014 ; JIN, CAO, Yongdong ZHANG et al. 2016] :

1. Construire un réseau de crédibilité. Les entités impliquées dans la détection de fausses informations, telles que les messages, les utilisateurs ou les événements, sont définies comme des nœuds du réseau. Chaque nœud a une valeur

de crédibilité initiale. Les liens entre ces entités sont définis et calculés en fonction de leur relation sémantique ou de leur relation d'interaction sur les médias sociaux (*e.g.* nombre de citations) ;

2. Propagation de la crédibilité. Sous certaines hypothèses de cohérence des nœuds et de régularité du réseau, les valeurs de crédibilité sont propagées sur le réseau construit selon des liens pondérés jusqu'à la convergence ce qui donne l'évaluation finale de la crédibilité pour chaque entité. Le problème de propagation est formé par une tâche où certains sommets sont connus comme crédibile ou non et le but est d'estimer la crédibilité des autres sommets [X. ZHU et GHAHRAMANI 2002 ; X. ZHU, GHAHRAMANI et LAFFERTY 2003 ; ZHOU et al. 2004]. Par rapport à la classification directe sur une entité individuelle, les approches basées sur la propagation peuvent tirer parti des relations entre entités et obtenir des résultats robustes.

[GUPTA, P. ZHAO et HAN 2012] ont introduit une méthode de propagation dans ce cadre là. Les auteurs construisent un réseau composé d'utilisateurs, de messages et d'événements sous deux intuitions : 1) Les utilisateurs crédibles n'offrent pas de crédibilité aux événements associés à des fausses informations en général ; 2) Les liens entre les messages crédibles ont des poids plus importants que ceux des messages de rumeurs, car les messages dans un événement associés à des fausses informations ne font pas de déclarations cohérentes. Les valeurs de crédibilité initiales de chaque message sont obtenues à partir des résultats d'un classificateur basé sur des caractéristiques similaire à ceux introduit par [CASTILLO, MENDOZA et POBLETE 2011] (voir la présentation des approches basées sur le texte en début de chapitre). Ils sont ensuite propagés sur ce réseau à l'aide d'itérations d'un algorithme de type *PageRank*.

Inspirés par l'idée de relier toutes les entités et de tirer parti des implications inter-entités pour la propagation de la crédibilité, [JIN, CAO, Yongdong ZHANG et al. 2016] ont proposé un réseau de crédibilité à trois couches construit à partir de différents niveaux sémantiques de contenu d'un événement : couche de message, couche de sous-événement et couche d'événement. Elles sont toutes basées sur le contenu et ont des relations directes avec la crédibilité des informations. Les sous-événements sont différents points de vue d'un même événement, qui sont des groupes de messages représentant des parties ou des sujets principaux d'un événement. Pour être précis, le réseau est formé comme suit : un message peut être lié à un sous-événement ; un sous-événement est lié à l'événement ; tous les messages sont liés

entre eux, de même que les sous-événements. En supposant que les entités avec un grand poids de lien entre elles aient des valeurs de crédibilité similaires, le problème de propagation de la crédibilité est formulé comme un problème d'optimisation de graphe selon les auteurs.

Les fausses informations se propagent rapidement sur les réseaux sociaux (voir chapitre 1) et le plus souvent, les personnes qui les partagent ne sont pas conscientes de leur erreur. La recherche et l'analyse de l'auteur peut être une solution pour déterminer plus efficacement la véracité de l'information. Cette approche utilise, en plus, l'hypothèse que les fausses informations sont souvent créés par les mêmes personnes.

2.4.3 Approches basées sur l'événement

Les descripteurs basés sur l'événement du message sont extraits au niveau de l'ensemble des messages et visent à comprendre les relations entre les messages au sein d'un corpus. [K. Wu, S. Yang et K. Q. Zhu 2015] définissent un ensemble de descripteurs de l'événement pour détecter les rumeurs sur *Weibo*[1], un réseau social célèbre en Chine. Les auteurs proposent un modèle *Latent Dirichlet Allocation* ou fréquence inverse du document (LDA) avec une distribution de 18 sujets sur tous les messages. Il est à noter que les auteurs permettent le fait que chaque message peut appartenir à un ou plusieurs sujets. Ils transforment le vecteur de distribution à 18 dimensions en vecteur binaire en définissant les k-sujets les plus probables à 1 et le reste des sujets à 0. La valeur de k est fixée par les auteurs.

[Jin, Cao, Yazi Zhang et al. 2015] regroupent les sujets en se basant sur l'événement auquel un message fait référence et extraient les descripteurs au niveau du message et de l'événement. Ils supposent que les messages sous un même sujet ont probablement des valeurs de vérité similaires. Sous cette hypothèse, ils regroupent les messages par événement et obtiennent les descripteurs au niveau du sujet en agrégeant les descripteurs obtenus au niveau du message (*e.g.* moyenne des valeurs d'un même descripteur sur tous les messages de l'événement). Ils avancent l'idée que ce type de descripteurs au niveau du sujet peut réduire l'impact des données bruitées tout en conservant la plupart des détails au niveau du message.

[F. Yu et al. 2017] constate que le RNN, utilisé dans certains travaux, n'est pas qualifié pour les tâches de détection précoce (*e.g.* détecter la fausse information avant

1. https://www.weibo.com

qu'elle ne se propage totalement sur le réseau social) avec peu de données et qu'il a tendance à privilégier les dernières publications traitées. Pour résoudre ces problèmes, ils proposent une approche basée sur un CNN pour la détection de fausses informations. Plus précisément, les auteurs proposent une méthode pour diviser chaque événement de fausses informations en plusieurs phases temporelles. La représentation de chaque phase est apprise par l'intermédiaire de `doc2vec`. Enfin, les vecteurs alimentent un CNN à deux couches, obtenant les résultats finaux de la classification à deux classes.

Les approches basées sur une analyse de l'événement associé à une publication présentent un intérêt dans le cadre de cette thèse car les publications peuvent posséder un texte d'une longueur limitée (*e.g.* publications issues de *Twitter*). Prendre en compte les messages évoquant le même événement que la publication initiale permet de développer le contexte et par exemple de vérifier les autres sources, crédibles ou non, évoquant cet événement.

2.4.4 Approches basées sur plusieurs modalités

[Ruchansky, Seo et Liu 2017] se concentrent sur trois caractéristiques des données de fausses informations : 1) le texte d'une publication ; 2) la réponse qu'il reçoit des utilisateurs ; 3) les utilisateurs qui font la promotion de la source, notamment en la partageant. Les auteurs proposent un modèle hybride, nommé `CSI`, qui combine les trois caractéristiques pour une prédiction plus précise. Le modèle est composé de trois modules comme présenté dans la figure 2.11 :

1. Le module *Capture* extrait une représentation temporelle de la publication en utilisant un RNN ;
2. Le module *Score* permet d'obtenir une représentation de chaque utilisateur et fusionne ces représentations pour obtenir une représentation de la publication ;
3. Le module *Integrate* concatène la sortie des deux autres modules et utilise le vecteur obtenu pour une classification.

[Jin, Cao, Guo et al. 2017] utilisent non seulement les informations textuelles mais aussi des informations visuelles et sociales et proposent un modèle basé sur une fusion multimodale. La système utilisé est présenté dans la figure 2.12. Un nombre croissant d'utilisateurs utilisent des images et des vidéos pour publier des informations en

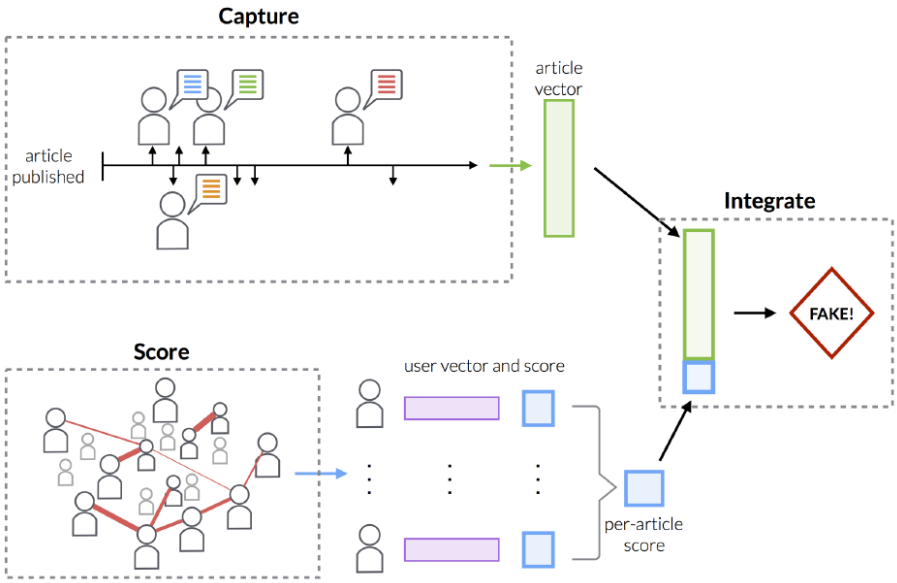


FIGURE 2.11 – Illustration de l'approche multimodale proposée par [RUCHANSKY, SEO et LIU 2017].

plus des textes. Par conséquent, pour un article donné, son texte et son contexte social sont d'abord fusionnés avec un réseau *Long Short-Term Memory* (réseau de neurones récurrents à mémoire court-terme et long terme en français) (LSTM). La représentation commune est ensuite fusionnée avec des caractéristiques visuelles extraites de `VGG-19` [SIMONYAN et ZISSERMAN 2014], un CNN. La sortie du LSTM à chaque pas de temps est utilisée comme couche d'attention au niveau des neurones pour aligner les caractéristiques visuelles avant la fusion avec la représentation du texte produite par un RNN. Les représentations basées sur le texte et l'image sont ensuite concaténées et utilisée comme entrées pour un classifier.

2.5 Discussions et choix d'approches

Comme nous l'avons vu dans ce chapitre, les approches présentes dans l'état de l'art sont très diverses et se basent sur toutes les modalités possibles des publications.

Nous nous focalisons sur les modalités **texte** et **image**.

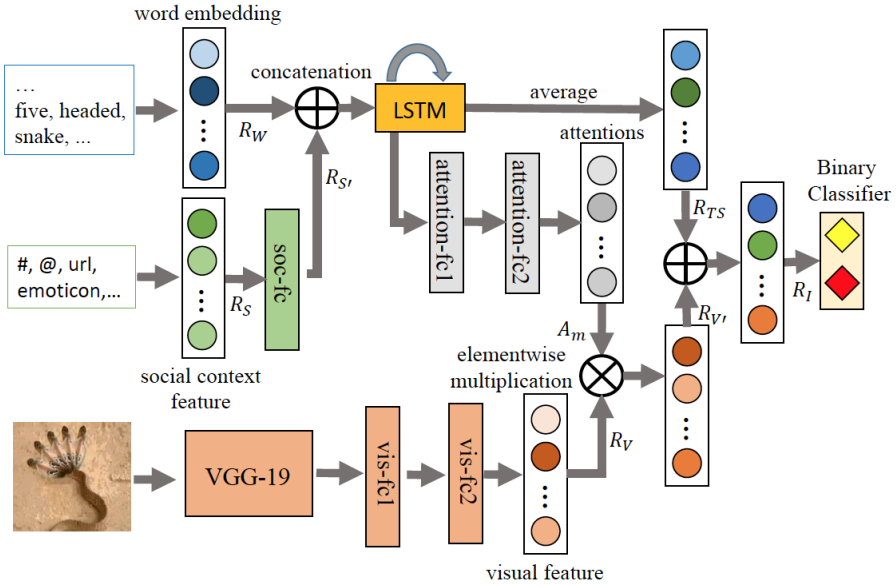


FIGURE 2.12 – Illustration de l'approche multimodale proposée par [JIN, CAO, GUO et al. 2017].

2.5.1 Analyse du texte

Nous souhaitons proposer une approche préliminaire orientée sur les articles publiés par les médias sur les réseaux sociaux. Ces médias peuvent être regroupés selon quatre types (définition de ces quatre types de médias détaillée dans le chapitre 3). Nous nous intéressons à mettre en place une méthode pour caractériser ces différents types de médias. Pour cela, nous nous focalisons sur le contenu textuel.

Les descripteurs textuels présents dans la littérature sont généralement de deux types : les comptages d'un élément, aussi appelé descripteurs *de surfaces*, et les descripteurs basés sur le contenu. Ces deux types de descripteurs permettent une représentation aussi bien de la structure du message par des descripteurs de surface que du contenu par des descripteurs basés sur les mots utilisés. Pour cela, nous nous intéressons aux différents descripteurs proposés dans la littérature, particulièrement pour la description en surface comme proposé par [CASTILLO, MENDOZA et POBLETE 2011 ; KWON, CHA, JUNG, W. CHEN et al. 2013 ; F. YANG et al. 2012].

Nous souhaitons aussi tester une discrimination par une description du contenu. Pour cela, nous orientons notre choix vers une utilisation des mesures *Term Frequency* ou fréquence du terme (TF) et *Inverse Document Frequency* ou fréquence inverse du document (IDF) dans le but de calculer le score TF-IDF afin de trouver des *n*-grammes

caractéristiques de chaque type de média.

Cette analyse fait indirectement intervenir une autre modalité, analysée dans la littérature, qui est l'estimation de crédibilité de la source de part le cheminement du message. Cependant, cela se fait sans la représentation sous forme de graphe commune aux travaux analysant le cheminement du message.

2.5.2 Analyse de l'image

Un constat sur les approches basées sur les images est que de nombreuses approches se basent soit sur une connaissance a priori (*e.g.* modèle de l'appareil photo ayant servi à prendre la photo) ou se concentrent sur un type de modification ou un format particulier (*e.g.* le format JPEG).

Or toutes ces approches ne sont pas applicables dans le cas d'images issues des réseaux sociaux. En effet, les formats d'images peuvent être variés ou changés par le réseau social lui-même lors de la soumission de la publication. Cela est aussi valable pour les approches utilisant une connaissance a priori puisque nous ne pouvons pas toujours connaitre l'origine de la photo. Il est alors impossible de récupérer des informations supplémentaires comme par exemple l'appareil photo ayant été utilisé. Cette remarque s'applique aussi à toutes les méta-données souvent supprimées par le réseau social, comme le lieu de la prise ou la date.

Il est donc quasiment indispensable de penser à une approche ne se basant que sur l'image en elle-même et n'étant pas basée sur un format en particulier.

Le système mis en place est basé sur la recherche d'une image requête dans une base d'images connues. Si la base comprend une image assez similaire pour être considéré comme une quasi-copie de la requête, les deux images sont comparées. Ce type d'approche a notamment été reprise depuis par d'autres travaux tels que [Brogan et al. 2017]. Cette approche est motivée par un constat qui est que des images modifiées apparaissant sur les réseaux sociaux sont souvent des versions modifiées d'anciennes images déjà connues sur les réseaux sociaux. Un exemple de ce phénomène est donné dans la figure 2.13.

Ce choix correspond à plusieurs avantages et inconvénients par rapport aux autres approches de l'état de l'art. Le principal avantage et qui aussi la principale motivation de notre choix, c'est à dire la généricité de l'approche. L'avantage de ne nécessiter d'aucune informations à priori sur l'image a cependant une contre-partie : la nécessité

(a)

(b)

(c)

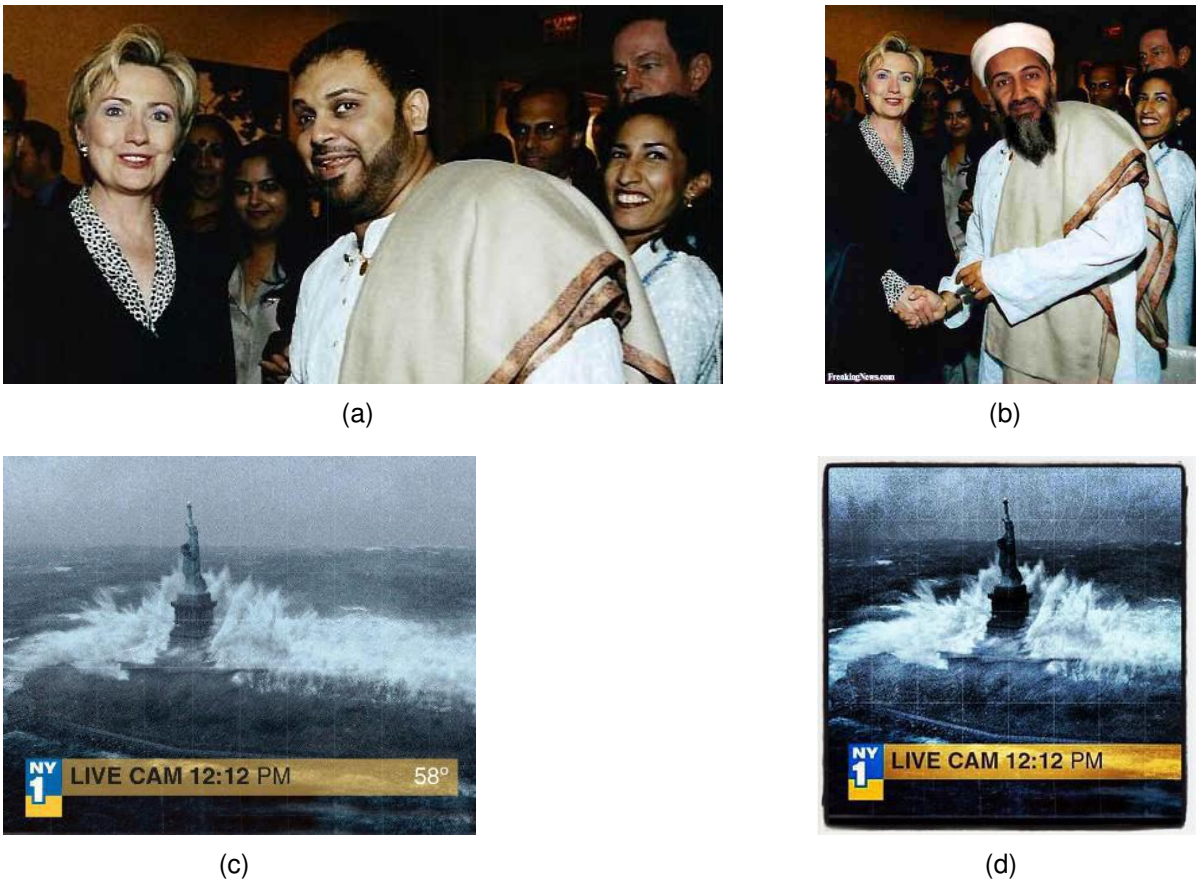
(d)

(e)

FIGURE 2.13 – Exemple d'une image réapparaissant régulièrement sur les réseaux sociaux avec le temps.

d'obtenir une image de référence pour réaliser la comparaison. Dans le cas où une telle image ne serait pas disponible, cette approche ne peut pas s'appliquer.

Une autre problématique de cette approche, et qui n'est pas non plus traitée par les autres travaux de la littérature, est l'estimation de la capacité d'une modification à rendre l'image fausse. Certaines modifications n'auront pas pour but de tromper l'utilisateur comme par exemple le changement de couleur de l'image 2.14(c) vers l'image 2.14(d). À l'inverse, le changement de visage de l'image 2.14(a) vers l'image 2.14(b) change la compréhension de l'image par l'utilisateur et donc son sens.



(a) (b)

(c) (d)

FIGURE 2.14 – Exemple de deux modifications. L'une changeant le sens de l'image (en haut) et une ne changeant pas son sens (en bas). Pour chaque modification, l'image originale est à gauche et la version modifiée à droite.

2.5.3 Recherches non abordées dans cet état de l'art

Bien que cet état de l'art présente un grand éventail de méthodes utilisées pour la détection de fausses informations et pouvant être comparées relativement directement aux travaux présentés dans ce manuscrit, certains champs de recherche n'ont pas été évoqués car annexe aux travaux présentés.

Premièrement, des travaux basés sur l'analyse de vidéos modifiées sont fréquemment proposés. L'analyse de la vidéo apporte une dimension supplémentaire à traiter, soit la dimension temporelle. Alors que les images ne peuvent être modifiée que spatialement, les vidéos peut aussi l'être au niveau temporel. Ces travaux s'intéressent à la fois à la modification d'une zone dans la vidéo (dimension spatiale) ou à l'insertion de courtes vidéos à un moment de la vidéo (dimension temporelle). En plus de l'aspect de modification des vidéos, on trouve le détournement d'information aussi possible avec des images. Une vidéo (resp. image) peut ne pas être modifiée en elle-même, mais être utilisée avec un texte donnant un contexte à la vidéo (resp. image) différent du contexte original à cette vidéo (resp. image).

Ensuite, la vérification de faits est un domaine de recherche à part entière. Cette tâche est encore trop difficile pour être pleinement automatisée comme le montre [GRAVES 2018]. Les auteurs de ces travaux évoquent le conclusion que les systèmes de vérification automatique de faits sont pour le moment particulièrement performants lorsqu'il est question d'assister un professionnel, mais beaucoup moins lorsque le système doit prendre une décision par lui-même. Les outils disponibles actuellement s'orientent en majorité sur une assistance aux journalistes afin d'accélérer la vérification de faits d'un article. Cela est principalement dû à la nécessité d'un jugement et d'une sensibilité au contexte qu'il n'est actuellement pas possible d'intégrer à un système de vérification de faits entièrement automatisé.

CHAPITRE 3

MÉDIAS TRADITIONNELS ET MÉDIAS DE RÉINFORMATION

Contents

3.1 Introduction

Beaucoup d'utilisateurs voient les réseaux sociaux comme une source d'informations. Un étude de 2013 [1] montre que Twitter est la source d'information la plus réactive. Le terme de *réactif* est utilisé ici pour signifier que l'information peut être trouvée d'abord sur les réseaux sociaux, puis dans les articles de médias plus ordinaires. Ainsi, il est courant d'être informé d'une actualité importante (*e.g.* attaque terroriste dans un lieu) ou non (*e.g.* score en temps réel d'un match sportif) en premier par des utilisateurs de *Twitter*, puis par les autres médias. C'est pourquoi de plus en plus de personnes considèrent ce réseau social comme leur source principale d'information. Cependant, ces utilisateurs ne vérifient que rarement les informations qu'ils partagent ce qui engendre beaucoup de mésinformation [2].

Une autre raison de ce changement d'habitude chez les utilisateurs est la présence des médias d'informations sur les réseaux sociaux. Parmi ces médias, on désigne par la suite ceux dits *traditionnels* qui correspondent aux médias connus du grand public. Le plus souvent, ils sont historiquement liés à d'autres canaux de diffusion (presse, papier, télévision, radio, etc.). Trois exemples de médias traditionnels sont donnés sur la ligne du haut de la figure 3.1.

Les réseaux ayant favorisé la rencontre et la communication entre des personnes souhaitant défendre une cause commune, des groupes thématiques de personnes se sont créés sur les réseaux sociaux. Ils se présentent eux-même comme des alternatives aux médias traditionnels et montrent une tendance à diffuser des informations en les interprétant à leur manière, voire en les modifiant de telle sorte qu'elles défendent les opinions du média (politiques, religieuses, etc.). Ils sont nommés par la suite comme étant des médias de *réinformation*. Trois exemples de médias traditionnels sont données sur la ligne du bas de la figure 3.1.

Cette communication massive passe par des articles diffusés sur les réseaux sociaux. De plus, les médias de réinformation s'opposent volontairement aux médias traditionnels qui, selon eux, déforment les faits et souhaite cacher la vérité au grand public. Ils se donnent donc pour mission de *rétablir la vérité et informer le grand public de la vérité*. Dans l'étude présentée dans ce chapitre, nous nous intéressons à déterminer dans quelle mesure les messages écrits sur les réseaux sociaux par les

1. https://www.20minutes.fr/high-tech/2234527-20180309-ligne-fake-news-repandent-six-fois-plus-vite-vraies
2. terme défini dans le chapitre 1

(a) Le Monde (b) BFM TV (c) Marianne

(d) Fdesouche (e) Les moutons enragés (f) Panamza

FIGURE 3.1 – Exemple de six médias d'informations présents sur les réseaux sociaux : trois médias traditionnels (en haut) et trois médias de réinformation (en bas)

médias traditionnels diffèrent de ceux des médias de réinformation. En effet, il existe des différences évidentes entre les deux types de structures, telles que le statut professionnel des médias traditionnels par exemple. Cependant, des médias (traditionnel comme réinformation) sont créés régulièrement et un utilisateur peut se retrouver face à un article écrit par un média qui lui est inconnu. Étant peu probable qu'il effectue une recherche sur le statut du média (ce réflexe pourtant utile et efficace n'est pas encore courant chez les utilisateurs des réseaux sociaux), cet utilisateur ne se basera que sur l'article qu'il est en train de lire.

Notre objectif est de savoir si des différences existent dans la manière de publier l'information par ces deux types de médias et d'estimer les capacités d'une détection automatique du type de média étant donné une publication. Ce travail de caractérisation de la source s'inscrit comme une étape intermédiaire dans un projet plus vaste de détection automatique des fausses informations sur les réseaux sociaux, en temps réel. Il est important de noter que l'objectif est donc de caractériser des messages des réseaux sociaux afin de déterminer automatiquement leur provenance (média traditionnel ou de réinformation) et non pas, à ce stade, leur véracité ou objectivité. Ce faisant, nous étudions l'influence de différents descripteurs en distinguant ceux de surface (*e.g.* longueur du message) de ceux portant sur le contenu du message [3].

Le chapitre est organisé comme suit. La section 2 présente la collecte des données

3. la différence entre ses deux types de descripteurs est définie plus loin dans ce chapitre au moment de la présentation de ces derniers.

et leur préparation ; La section 3 le protocole expérimental ; la section 4 les résultats obtenus. Enfin, la section 5 résume les conclusions de cette étude.

3.2 Constitution du jeu de données

Aucun jeu de données n'étant disponible pour cette tâche, nous avons été contraint d'en constituer un nouveau. Pour cela, plusieurs étapes ont été nécessaires en commençant par la sélection de pages *Facebook* et l'attribution d'un label à ces dernières correspondant au type de média dont fait partie cette page (traditionnel ou réinformation). Suite à cela, le contenu de ces pages est extrait pour être utilisé par notre système comme données d'entrée.

Une remarque importante quant à la constitution de ce jeu de données est le fait que ce dernier est constitué dans le but de répondre à une tâche plus large que celle traitée dans ce chapitre. Nous souhaitons évidemment collecter des données liées à des médias traditionnels et de réinformation, mais nous profitons du temps alloué à la constitution d'un corpus pour collecter des données de deux autres types de médias : les médias d'informations humoristiques et les médias de confiance.

3.2.1 Sélection et annotation des pages *Facebook* étudiées

La première phase de la constitution de ce jeu de données a été de lister le plus de médias possible selon nos connaissances personnelles. Cette liste a ensuite été complété par des listes trouvées sur internet [4]. Les travaux présentés ici étant aussi bien motivés par une étude de messages francophones que anglophones, des pages contenant des messages dans ces deux langues ont été sélectionnés. Suite à cela, une liste de plus de 100 médias est trouvée.

L'annotation de ces groupes est réalisée selon cinq labels :

1. médias de confiance ;
2. médias traditionnels ;
3. médias de réinformation ;
4. médias satiriques ;

4. Un exemple d'une telle liste est donnée ici : https://sk.ambafrance.org/Liste-des-principaux-medias

5. autres médias.

Les `médias de confiance` sont les sites listant les fausses informations déjà connues. Nous y ajoutons le média `AFP` qui est un média particulier et qui est jugé comme étant une source de confiance.

Les `médias traditionnel` correspondent à des organismes de presse réels et qui sont pour la grande majorité soit un journal papier, soit une chaine de télévision. Ces médias appartiennent à une société de presse identifiable.

Les `médias de réinformation` sont associés à des groupes ayant un point de vue sur l'actualité se voulant différent de celui des médias traditionnels et qui veulent promouvoir leur façon de penser par l'affichage explicite d'une volonté de réinformation ou toutes variantes autour du thème de la révélation de la vérité cachée par les médias de masse. Enfin, nous associons à ce type de média les pages partageant des informations classées comme fausses dans des sites spécialisés dans l'analyse des informations fausses ou trompeuses (par exemple `hoaxbuster.com`, `hoax-slayer.com`) ;

Les `médias humoristiques` sont des médias publiant des fausses informations basées sur l'humour. Ces sites déclarent le plus souvent ouvertement le coté humoristique des articles publiés.

Enfin, le label `autre` permet de mettre de coté les médias ne vérifiant aucune des quatre classes précédentes, ces médias sont par conséquent jugés non pertinents pour notre jeu de données. Un exemple est la radio *NJR* initialement présente dans la liste des pages sélectionnées, mais qui ne contient quasiment que des informations sur la radio et non des d'actualités.

La tâche d'annotation manuelle consiste alors en l'annotation des groupes selon ces cinq labels. Concernant les labels **humour** et **confiance**, l'annotation n'est pas nécessaire puisque nous appliquons des règles précises et automatiques sur ces deux labels (caractéristiques présentées précédemment dans la description de chaque type de média). Les autres groupes sont à annoter selon les classes `traditionnel`, `réinformation` et `autre`.

Cette tâche d'annotation est effectuée par trois annoteurs. Les accords inter-annoteurs obtenus sont élevés (κ de Fleiss [Joseph L Fleiss et Cohen 1973] = 0.874 ; α de Krippendorff [Krippendorf 1980] = 0.875). Ces deux accords ont pour but de représenter l'accord entre les différentes personnes ayant annoté les données et détermine indirectement la difficulté de la tâche en question pour un être humain (*i.e.* une tâche très facile recevra exactement les mêmes annotations quelque soit l'annoteurs).

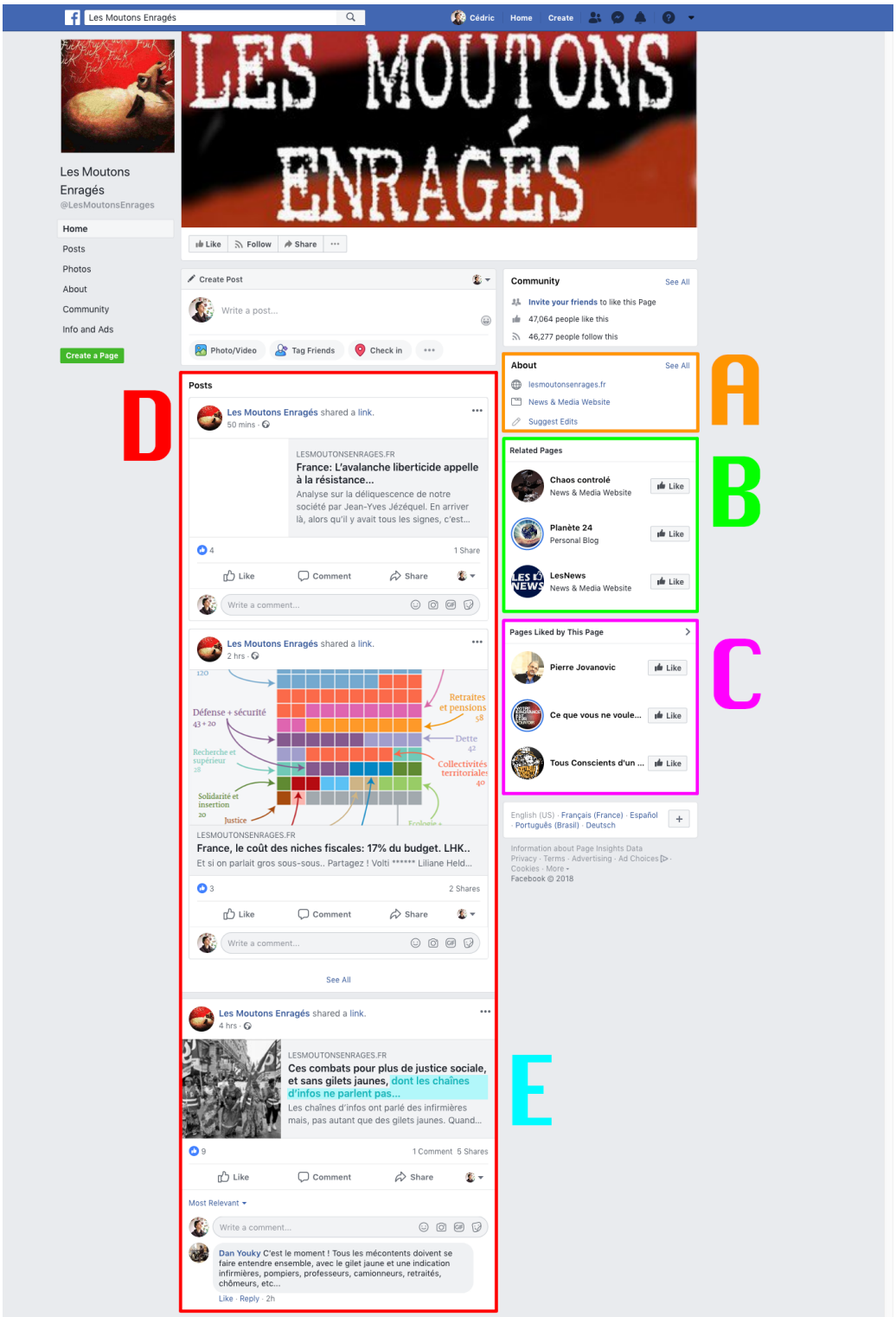


FIGURE 3.2 – Exemple de page de réinformation

La valeur très haute du κ de Fleiss correspond à un très bon score comme le montre la table 3.1 issue de [LANDIS et KOCH 1977]. De la même manière, l'α de Krippendorff est généralement considéré comme acceptable à partir de $0,6$ et bon à partir de $0,8$.

TABLE 3.1 – Interprétation du κ de Fleiss [DAVIES et Joseph L. FLEISS 1982] en fonction de sa valeur

Valeur de κ	Interprétation
< 0	Pauvre concordance
$0,01 - 0,20$	Faible concordance
$0,21 - 0,40$	Légère concordance
$0,41 - 0,60$	Concordance moyenne
$0,61 - 0,80$	Concordance importante
$0,81 - 1,00$	Concordance presque parfaite

Ces deux mesures évaluent la concordance des annotations en attribuant des scores entre 0 (désaccord complet) et 1 (annotations identiques). Ces scores élevés montrent que la tâche de discrimination entre les médias traditionnels et les médias de réinformation est possible pour un humain, sous réserve de faire quelques recherches sur ce média (action réalisée par les annoteurs pour la tâche d'annotation).

Les divergences ont ensuite été discutées pour décider par consensus de la classe à attribuer. La répartition des 79 groupes restants en fonction de ces quatre classes suite à l'harmonisation des annotations est donnée dans la figure 3.3. La figure 3.4 représente quant à elle la même répartition, mais au niveau des messages (l'annotation du groupe est reportée à tous les messages issus de ce groupe).

Pour la suite de cette étude, nous utilisons exclusivement les médias labellisés comme *traditionnel* ou *réinformation*, la problématique que nous souhaitons étudier étant la capacité d'un système automatique à distinguer des publications provenant de ces deux types de médias. Les autres groupes (`humour` et `confiance`) sont gardés pour des études futures.

3.3 Approche par apprentissage supervisé

Le but de cette étude est de tester les capacités d'un système de classification automatique à distinguer le type de média (*i.e. traditionnel* ou *réinformation*) dont provient un article journalistique sur les réseaux sociaux.

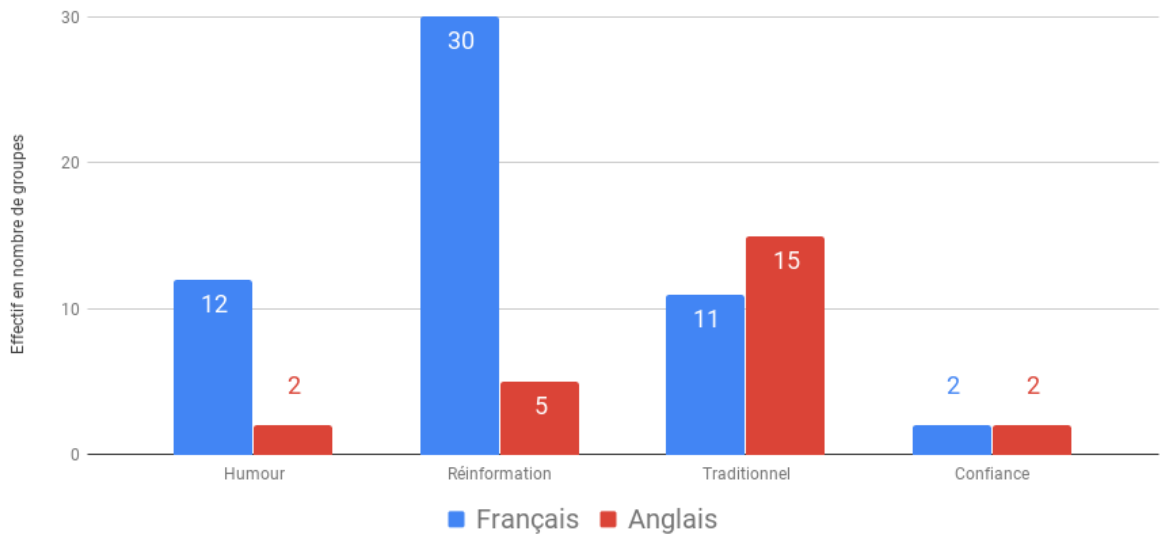


FIGURE 3.3 – Répartition des groupes récoltés et annotés dans les quatre types de groupes de leur langue

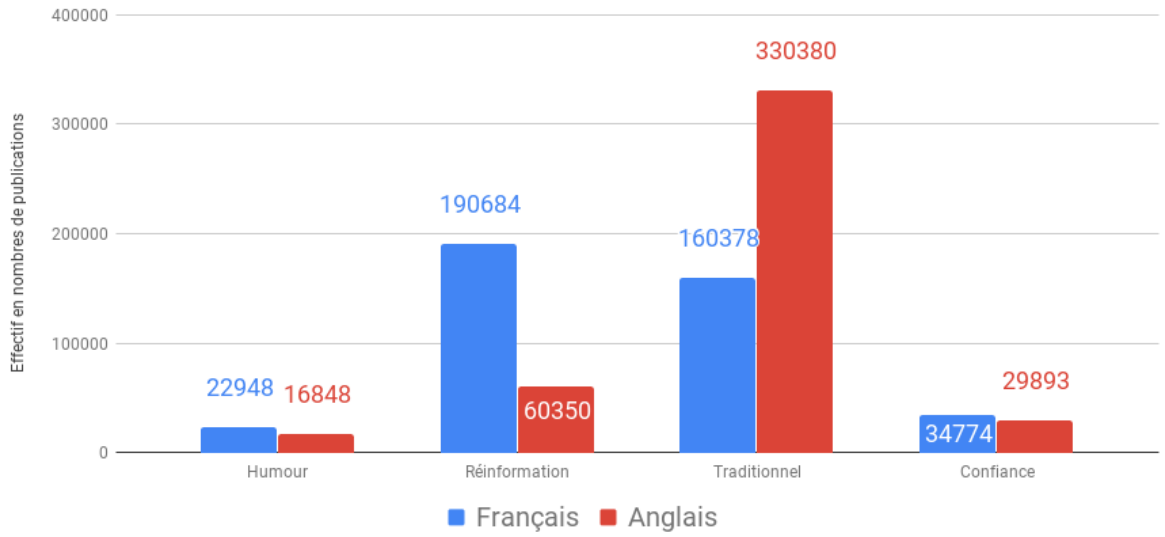


FIGURE 3.4 – Répartition des messages issus des groupes récoltés et annotés dans les quatre types de groupes en fonction de leur langue

La suite de la section est organisée comme suit : la sous-section 3.3.1 détaille les descripteurs utilisés. Ensuite, la sous-section 3.3.2 présente le protocole expérimentale basé sur ces descripteurs.

3.3.1 Descripteurs utilisés

Afin de représenter les messages, deux familles de descripteurs sont étudiées. Premièrement des descripteurs, dit *de surface*, qui correspondent à des mesures de comptage sur le texte (*e.g.* le nombre majuscules). Ensuite des descripteurs, dit *de contenu*, correspondant à une représentation directement basée sur les mots utilisés pour faire intervenir les liens sémantiques entre ces derniers. Un exemple de description d'une publication par ces deux types de descripteurs est donné dans la figure 3.5.

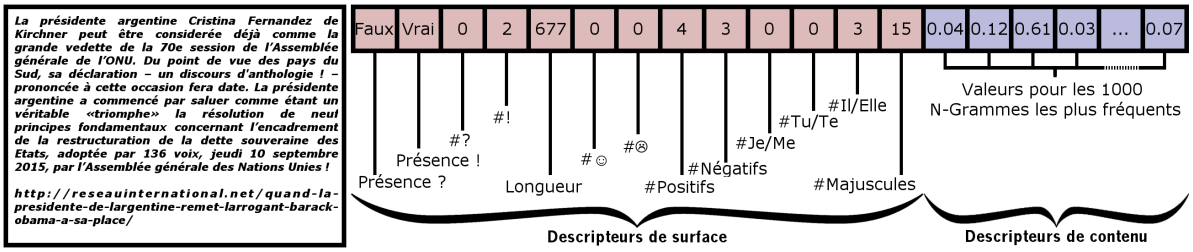


FIGURE 3.5 – Exemple de description d'une publication du jeu de données par les descripteurs de surface (en rouge) et de contenu (en bleu)

Descripteurs de surface

Ces descripteurs permettent de caractériser la structure des messages. Pour nos travaux, nous nous inspirons de ceux proposés dans l'état de l'art pour la détection de fausses informations sur les réseaux sociaux [BOIDIDOU, ANDREADOU et al. 2015]. Treize descripteurs sont calculés pour chaque message, caractérisant :

1. la longueur du texte en nombre de caractères ;
2. la présence du symbole " ?" (valeur booléenne) ;
3. la présence du symbole " !" (valeur booléenne) ;
4. l'occurence du symbole " ?" ;
5. l'occurence du symbole " !" ;
6. le nombre de pronoms de la première personne (exemple : je) ;
7. le nombre de pronoms de la deuxième personne (exemple : tu) ;
8. le nombre de pronoms de la troisième personne (exemple : il) ;
9. l'occurence d'émoticônes heureux ;

	$\mathcal{C}s_f$		$\mathcal{C}s_a$	
	réinformation	*traditionnel*	*réinformation*	*traditionnel*
Mots par message	1173, 64	4118, 89	621, 98	3067, 14
Hashtags par message	10, 27	26, 89	1, 48	10, 48
Majuscules par message (*)	1, 88%	2, 90%	3, 80%	2, 95%
Occurrence du symbole *?* (*)	0, 08%	0, 03%	0, 13%	0, 05%
Occurrence du symbole *!* (*)	0, 10%	0, 09%	0, 14%	0, 07%

TABLE 3.2 – Quelques caractéristiques de surface sur les corpus francophone et anglophone (en moyenne par message)

10. l'occurence d'émoticônes tristes ;
11. le nombre de majuscules ;
12. le nombre de mots à polarité positive ;
13. le nombre de mots à polarité négative.

Les descripteurs 6, 7, 8, 9, 10, 13 et 14 sont calculés à l'aide de dictionnaires. L'analyse des descripteurs 6, 7 et 8 permet de distinguer les textes qui évoquent un fait propre à l'auteur (utilisation des pronoms de la première personne), les textes interpellant le lecteur (pronoms de la deuxième personne) et les textes impersonnels (troisième personne)

Quelques-unes de ces caractéristiques sont données dans la table 3.2, mettant en évidence les différences de répartition des descripteurs en fonction des corpus. Les attributs notés avec un astérisque sont normalisés par la longueur des messages.

Descripteurs du contenu

Les descripteurs basés sur le contenu ont vocation à caractériser les médias par la présence de mots ou de séquences de mots spécifiques dans leurs messages. Pour ce faire, les messages sont lemmatisés avec TreeTaggger [H. SCHMID 1994] et les urls, hashtags et sources sont remplacés respectivement par les balises *[URL]*, *[HASHTAG]* et *[SOURCE]*, ensuite traités comme des mots.

Pour chaque message, les valeurs TF-IDF [SPARCK JONES 1972] des *n*-grammes de taille un à trois sont calculées. Les 1 000 *n*-grammes les plus discriminants, selon leur gain d'information [MITCHELL 1997], sont alors retenus pour constituer le vocabulaire de description, et chaque message est donc représenté par un vecteur de dimension 1 000.

3.3.2 Classification des publications

Cette sous-section décrit le protocole expérimental utilisé pour la tâche de classification visée, à savoir prédire l'origine (média traditionnel ou de réinformation) d'un message *Facebook*, les résultats obtenus, et une analyse de l'emploi des descripteurs dans les modèles de classification obtenus.

Les tests sont réalisés sur des sous-ensembles de la base de données, chacun constitué aléatoirement de 40 000 messages issus pour moitié de médias traditionnels et pour moitié de médias de réinformation. Le but est ici d'obtenir des classes équilibrées et de supprimer le fait que certains groupes publient plus souvent que d'autres et sont ainsi plus souvent représentés dans le corpus. Si un groupe très représenté possède une structure très particulière et qui lui est propre, cela engendrerait un biais dans l'analyse. Les descripteurs de surface et les descripteurs de contenu sont d'abord utilisés indépendamment, puis combinés.

La librairie Weka [HALL et al. 2009] est utilisée pour la classification. Plusieurs classifieurs de différentes familles sont testés : classifieur bayésien naïf (*Naive Bayes*), règles propositionnelles (*JRip*), Arbre de décision (*J48*), forêts aléatoires (*Random Forest*) avec $N = 100$, SVM (*SMO*) avec noyau RBF et k-plus proches voisins (IB_k) avec $k = 1$. Les descripteurs ont été normalisés selon la norme ℓ_2 pour ces deux dernières méthodes. Deux mesures complémentaires sont utilisées pour évaluer les résultats : la F-mesure (F_1) et le taux de bonnes classifications (*Taux BC*). Afin d'évaluer les classifieurs, une méthode de validation croisée à dix plis est utilisée.

3.4 Analyse des résultats

Les résultats obtenus sur les corpus anglais et français de 40 000 messages sont présentés respectivement dans les tableaux 3.3 et 3.4. La première constatation est que l'analyse du contenu retourne toujours des meilleurs résultats que l'analyse de surface seule. La combinaison des deux descripteurs améliore marginalement les résultats obtenus en utilisant les descripteurs sur le contenu seuls. On notera également que les résultats sur le corpus francophone sont légèrement meilleurs que sur le corpus anglophone. La supériorité des descripteurs basés contenu par rapport aux descripteurs de surface y est également plus marquée.

L'analyse des modèles de classification obtenus (lorsqu'elle est possible) permet

corpus $\mathcal{C}s_a$	*surface*		*contenu*		*surface + contenu*	
	F_1	*Taux BC*	F_1	*Taux BC*	F_1	*Taux BC*
Naive Bayes	52,80%	59,37%	80,36%	80,58%	80,42%	80,65%
J48	**65,04%**	**66,06%**	**85,14%**	**85,18%**	**84,94%**	**84,99%**
JRip	64,25%	65,34%	68,80%	70,31%	75,78%	76,15%
IB_1	54,17%	54,44%	61,28%	62,90%	77,09%	77,16%
Random Forest	62,92%	63,31%	75,90%	76,51%	79,53%	79,92%
SMO	51,61%	52,65%	76,91%	77,35%	76,93%	77,38%

TABLE 3.3 – F-mesure (F_1) et taux de bonne classification (*Taux BC*) des messages du corpus anglais (moyenne sur les 10 plis)

corpus $\mathcal{C}s_f$	*surface*		*contenu*		*surface + contenu*	
	F_1	*Taux BC*	F_1	*Taux BC*	F_1	*Taux BC*
Naive Bayes	38,58%	51,58%	72,38%	73,88%	73,12%	74,50%
J48	61,62%	61,76%	82,67%	82,81%	83,41%	83,54%
JRip	59,28%	59,41%	81,86%	82,02%	81,08%	81,33%
IB_1	61,57%	61,73%	83,75%	83,89%	83,74%	83,77%
Random Forest	**62,92%**	**63,31%**	75,90%	76,51%	79,53%	79,92%
SMO	13,82%	14,61%	**88,60%**	**88,62%**	**88,84%**	**88,86%**

TABLE 3.4 – F-mesure (F_1) et taux de bonnes classifications (*Taux BC*) des messages du corpus français (moyenne sur les 10 plis)

de comprendre les cas d'erreurs et de caractériser la pertinence des différents descripteurs.

Analyse des descripteurs de surface

La sélection des descripteurs les plus discriminants, effectuée par calcul de l'information mutuelle entre l'attribut et la classe, met en avant par ordre d'importance décroissante : 1) la longueur du texte ; 2) la présence de symboles de ponctuation *!* et *?* ; 3) l'orientation des pronoms personnels (*i.e.* première, deuxième ou troisième personne). Ces résultats sont corroborés par l'étude des descripteurs effectivement utilisés dans les arbres de décision et des règles prépositionnelles.

Analyse des descripteurs de contenu

L'étude des classifieurs interprétables (comme JRip, Random Forest, J48) générés à partir de ces descripteurs montre que les modèles de décision cherchent à caractéri-

ser principalement les médias traditionnels, le message étant classé en média de réinformation par défaut, c'est-à-dire lorsque qu'aucun de ses descripteurs ne l'a amené à être classé comme traditionnel. Cela semble indiquer qu'il est plus facile de déterminer des caractéristiques communes à tous les messages traditionnels qu'aux messages des médias de réinformation.

De ce fait, les erreurs sont majoritairement commises par manque de règles décrivant les messages de médias de réinformation. Cependant, cela permet d'obtenir une précision élevée sur la classe *traditionnel* (*i.e.* un message classé comme tel à une forte probabilité d'être bien classé) : par exemple, pour le corpus francophone, la précision de la classe *traditionnel* est de 94, 75% avec le classifieur *Naive Bayes* contre une précision de 66, 29% pour la classe *réinformation*.

Certains descripteurs discriminants de la classe *traditionnel* relèvent indirectement de l'aspect professionnel du site diffusant l'information ; il s'agit par exemple de la présence des termes *RSS* ou *votre abonnement* pour le français, et *accessiblity* ou *privacy* pour l'anglais. D'autres descripteurs notent le niveau de langue de certains sites de réinformation, faisant plus largement usage d'abréviation comme *WTF*, *DIY*, *pic*... Les médias traditionnels sont quant à eux caractérisés par une présence accrue de marques de citations, ou de mots comme *opinion*.

Analyse de la combinaison des descripteurs

Comme pour l'ensemble de descripteurs de contenu ci-dessus, les classifieurs cherchent à détecter les messages de médias traditionnels. Cette fois aussi la précision de la classe *traditionnel* est haute puisqu'elle vaut 94, 75% avec *Naive Bayes* sur le corpus francophone tout comme la précision de la classe *réinformation* qui obtient des résultats légèrement augmentés avec *Naive Bayes*, soit 66, 8%. Le score de la classe *traditionnel* étant beaucoup plus haut que celui de la classe *réinformation*, cela tend à montrer que les médias de réinformation veulent se faire passer pour des médias traditionnels en imitant leur structure. .

3.5 Conclusion et perspectives

Dans ce chapitre, nous avons analysé dans quelle mesure un système de classification automatique peut distinguer les différences entre deux types de médias en

se basant exclusivement sur le contenu textuel des publications de ces deux types de médias [MAIGROT, KIJAK et CLAVEAU 2016]. Cette étude intervient en début de thèse comme une étude préliminaire avant les travaux directement portés sur la détection de fausses informations (présentés dans les chapitres suivants). La distinction réalisée est faite entre deux types de médias : médias traditionnels et médias de réinformation. Pour cela, nous étudions deux aspects du texte avec deux ensembles de descripteurs : des descripteurs de surface représentant la forme du message par des mesures statistiques et des descripteurs de contenu basés directement sur les mots utilisés.

Nous avons ainsi vu que chacun des deux types de descripteurs permettaient de faire apparaître des différences entre les deux types de médias, cependant les descripteurs de contenu permettent une meilleure discrimination que les descripteurs de surface. Enfin, l'association des deux types de descripteurs permet une très légère augmentation des résultats.

Une amélioration possible est de prendre en compte l'URL possiblement présente dans un message et d'ajouter le contenu de cette URL au contenu textuel initial de la publication. Cela aura pour avantage d'augmenter la taille du texte et ainsi d'améliorer inévitablement la quantité d'informations associées à ce message. Cette perspective possède une contre-partie qui est de s'éloigner de la situation réelle dans laquelle un utilisateur voit une publication et doit décider de la provenance du message en ne se basant que sur la publication en elle même.

CHAPITRE 4

DÉTECTION DE FAUSSES INFORMATIONS PAR APPROCHES MULTIMODALES

Contents

4.1 Introduction

Le projet dans lequel s'inscrit ce travail a pour but de détecter automatiquement la véracité d'une information virale et dans la mesure du possible de justifier la classification. Le but final est de créer par exemple un système qui préviendra l'utilisateur avant qu'il ne partage une fausse information en lui indiquant le plus précisément possible ce qui est faux. Partant du constat que ces informations virales sont souvent composées d'éléments multimédias (texte accompagné d'images ou de vidéos), un système multimodal est proposé dans ce chapitre. Ce chapitre présente successivement les approches exploitant uniquement le contenu textuel, le contenu des images ou les sources citées dans les messages, puis 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 MediaEval 2016, des stratégies de fusion sont étudiées pour analyser l'apport des différentes approches et la capacité de prédiction d'un système collaboratif.

La section 4.2 présente la tâche *Verifying Multimedia Use* (VMU) du challenge *MediaEval* dont sont extraites les données utilisées dans ces travaux. Ensuite, la section 4.3 les approches mises en place, ainsi que les systèmes proposés par les autres équipes participantes à la tâche *VMU*. La section 4.4 présente le protocole expérimental et les résultats obtenus par les différentes approches. Différentes stratégies de fusion sont testées et discutées dans la section 4.5. Enfin, la section 4.6 résume les principales observations et évoque les pistes possibles pour l'avenir.

4.2 Présentation de la tâche *Verifying Multimedia Use* du challenge *MediaEval2016*

La tâche VMU de la campagne d'évaluation *MediaEval* en 2016, proposait de classer des messages provenant de *Twitter*[1] 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'ob-

1. https://twitter.com/

tenir une forte précision pour les classes *vrai* et *faux* [BOIDIDOU, PAPADOPOULOS, DANG-NGUYEN, BOATO, RIEGLER 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*. De plus, les messages sont soit accompagnés d'une ou plusieurs images, soit d'une vidéo (cf. figure 4.1). 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.

FIGURE 4.1 – Exemples de deux tweets de la tâche VMU de la campagne MediaEval, partageant la même image

De plus, les messages sont regroupés par événement. La taille des événements n'est pas équilibrée comme le montre la figure 4.2. 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 4.1 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 4.4, les résultats présentés sont ceux obtenus sur l'ensemble de test, et que les techniques de fusion présentées en section 4.5 sont utilisées sur les prédictions des systèmes des participants sur l'ensemble de test.

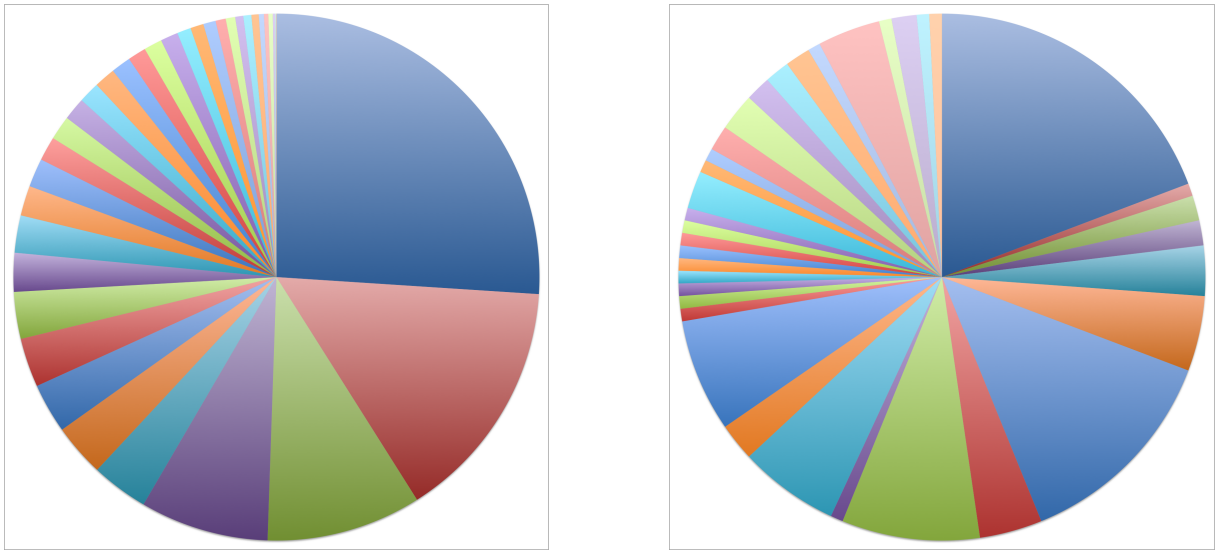


FIGURE 4.2 – Répartition des messages, à gauche, et des contenus multimédias (images et vidéos), à droite, par événement dans le jeu de test de la tâche *VMU* (35 événements)

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.

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 par l'auteur.

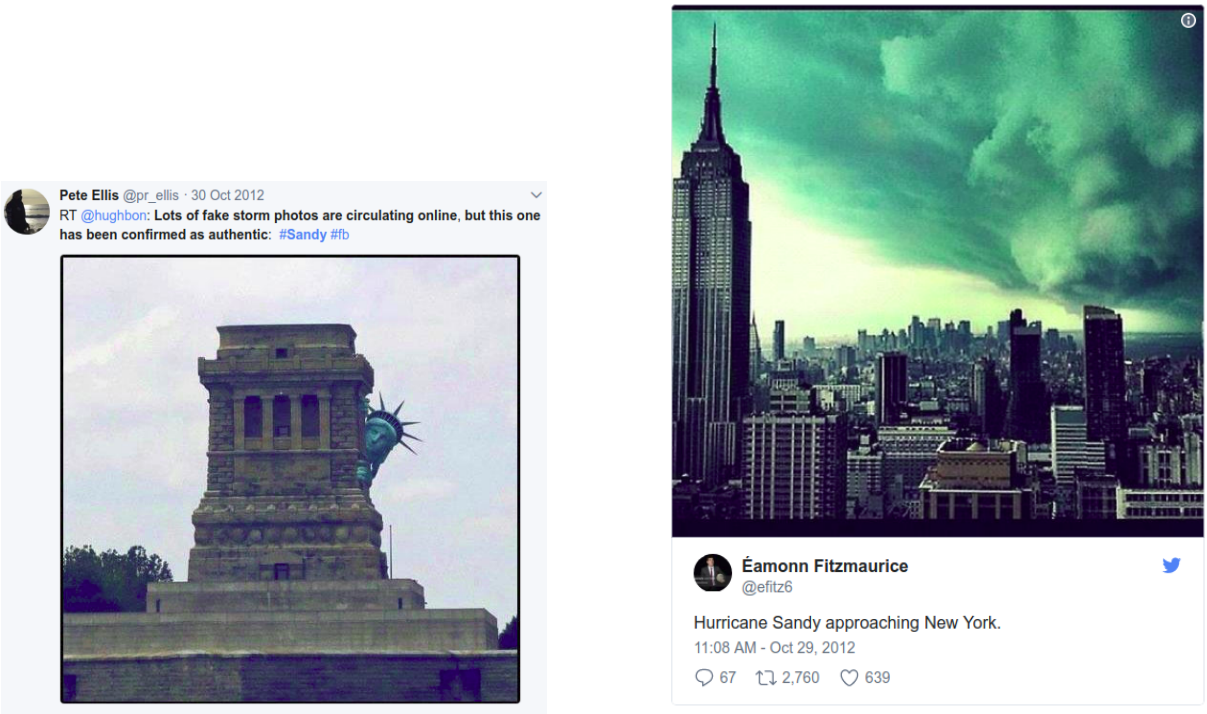


FIGURE 4.3 – Exemples de deux tweets de la tâche VMU de la campagne MediaEval, sur le même événement (ouragan Sandy) que dans la figure 4.1

TABLE 4.1 – 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

L'ensemble des descripteurs associés aux images, noté $\mathcal{FOR}$, provient de méthodes issues du domaine des *forensics* : indices de double compression JPEG [BIANCHI et PIVA 2012], Block Artifact Grid [W. LI, YUAN et N. YU 2009], Photo Response Non-Uniformity [GOLJAN, FRIDRICH et M. CHEN 2011] et coefficients de Benford-Fourier [PASQUINI, PÉREZ-GONZÁLEZ et BOATO 2014].

4.3 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* (les organisateurs de la tâche). Cette section présente les approches que nous avons développées, puis brièvement les approches proposées par les autres équipes participantes à la tâche.

Dans nos approches [MAIGROT, CLAVEAU, KIJAK et SICRE 2016], tous les messages partageant la même image sont associés à la même classe : *vrai*, *faux* ou *inconnu*. Partant du constat que les tweets partagent une même image ou vidéo, il est possible de déterminer la classe de chaque image et de reporter la classe prédite sur les messages associés à cette image ou vidéo, selon la règle suivante : un message est prédit comme *vrai* si toutes les images associées sont classées *vraies*, sinon *faux*.

Nous proposons trois approches : la première est basée sur le contenu textuel du message ; la seconde les sources ; la troisième les images. Aucune n'utilise les descripteurs $\mathcal{T}$, $\mathcal{U}$ ou $\mathcal{FOR}$ présentés en section 4.2. 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 le chapitre.

4.3.1 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'émoticônes, expressions populaires, ...).

Soit $\mathcal{I}$ la description d'une image en requête (*i.e.* l'union des contenus textuels des messages qui utilisent cette image) et $\mathcal{I}_{app}$ l'ensemble des descriptions des images de l'ensemble d'apprentissage. La classe de $\mathcal{I}$ est déterminée par vote des k images

dont les descriptions sont les plus similaires dans $\mathcal{I}_{app}$ (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, WALKER et HANCOCK-BEAULIEU 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 $\mathcal{I}$) et à un document (ici un texte de $\mathcal{I}_{app}$) ; voir équation 4.1.

$$\text{RSV}_{\text{Okapi}}(\mathcal{I}, \mathcal{I}_{app}) = \sum_{t \in \mathcal{I}} qTF(t) * TF(t, \mathcal{I}_{app}) * IDF(t) \tag{4.1}$$

$$qTF(t) = \frac{(k_3 + 1) * qtf}{k_3 + qtf} \tag{4.2}$$

$$TF(t, \mathcal{I}_{app}) = \frac{tf * (k_1 + 1)}{tf + k_1 * (1 - b + b * \frac{dl(\mathcal{I}_{app})}{dl_{avg}})} \tag{4.3}$$

$$IDF(t) = \log \frac{n - df(t) + 0.5}{df(t) + 0.5} \tag{4.4}$$

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 $k_1 = 2$, $k_3 = 1\,000$ 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 pour le challenge DeFT 2017 [CLAVEAU et RAYMOND 2017]. Le paramètre du nombre de voisins k est déterminé à 1 par validation croisée sur l'ensemble d'apprentissage.

4.3.2 Prédiction basée sur la confiance des sources (LK-S)

Cette approche, similaire à [S. E. MIDDLETON 2015], 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 des

messages dans lesquels elle apparaît (traduits en anglais si nécessaire). 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, une liste d'agences de presse dans le monde est déterminée, journaux (principalement francophones et anglophones) en s'appuyant sur des listes établies [2] ou réseaux télévisuel d'information (francophones et anglophones) [3]. Pour le second type, plusieurs patrons d'extraction sont déterminés manuellement (*e.g. 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.

4.3.3 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*. Une approche de type recherche d'images similaires est utilisée sur 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 (*i.e. vraies*) et 7 500 images *modifiées*.

Les descripteurs générés à partir des images sont calculés en utilisant un réseau de neurones convolutionnel profond `VGG-19` [SIMONYAN et ZISSERMAN 2014]. Les images sont d'abord redimensionnées à la taille standard de 544 $\times$ 544 et passées dans les couches convolutionnelles du réseau [TOLIAS, SICRE et JÉGOU 2016] qui permet à `VGG-19` de traiter des images de taille supérieur (plus de détails sont donnés dans le chapitre suivant consacré au traitement des images). Ensuite, les deux premières couches entièrement connectées sont mises sous forme de noyau et appliquées au

2. https://en.wikipedia.org/wiki/List_of_news_agencies
3. https://en.wikipedia.org/wiki/Lists_of_television_channels

tenseur de sortie, produisant un nouveau tenseur de dimension $7 \times 7 \times 4\,096$. Enfin, l'application d'un filtre moyenneur et d'une normalisation ℓ_2 permet d'obtenir un vecteur de description de dimension 4 096. 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 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 notée `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*.

4.3.4 Présentation des autres approches

Pour chacune des autres équipes participantes, les approches et le type de données utilisées pour prédire la classe des messages sont détaillés.

Équipe *VMU* [Boididou, S. Middleton et al. 2016]

Cinq méthodes ont été testées par les organisateurs de la tâche. Ces méthodes reposent sur deux systèmes. Le premier est adapté du système *Agreement-Based Retraining* (ARM) proposé lors de l'édition précédente de la tâche [Boididou, Papadopoulos, Dang-Nguyen, Boato et Kompatsiaris 2015]. Le second système, noté *Attribution based claim extraction* (ATT), est basé sur le système proposé par [S. E. Middleton 2015].

`VMU-F1` et `VMU-F2` s'appuient sur le système ARM qui est un méta-classifieur dans lequel deux ensembles de descripteurs sont utilisés séparément par deux classifieurs, entraînés sur l'ensemble d'apprentissage. Ce système est présenté dans la figure 4.4. 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.

Le second système, noté ATT, 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. `VMU-S1` est une combinaison des méthodes ARM et ATT. Dans cette approche, les publications qui ne peuvent pas être classées par la méthode ATT sont classés par la méthode ARM. `VMU-S2` utilise aussi les systèmes ARM et ATT en utilisant la sortie du second système en tant que caractéristique d'entrée supplémentaire pour le premier.

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

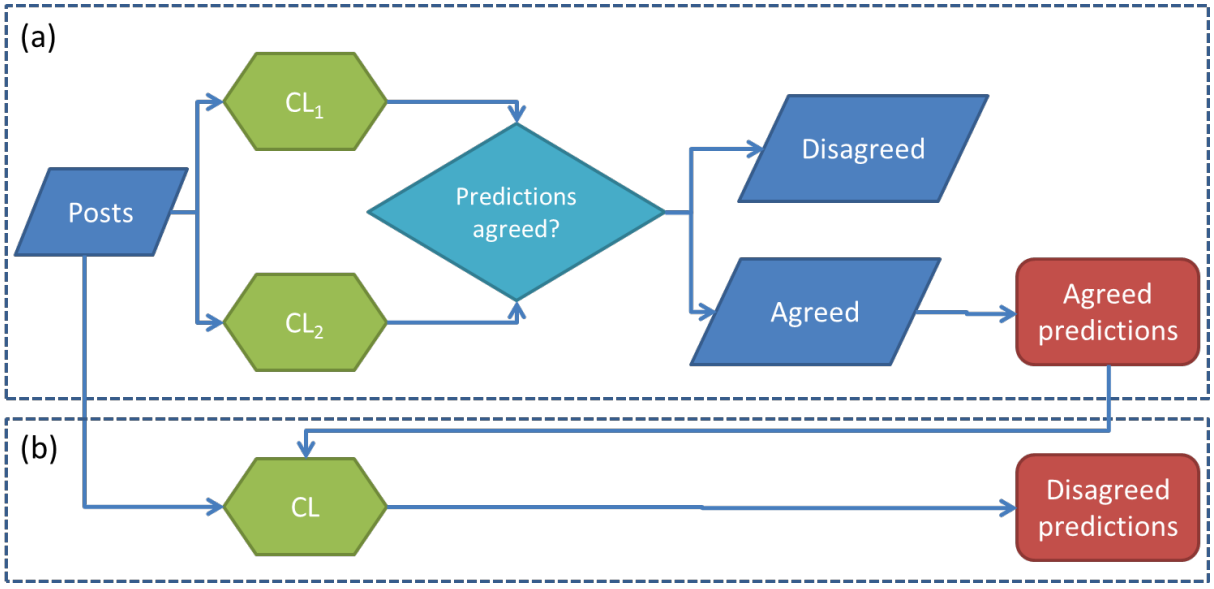


FIGURE 4.4 – Illustration de l'approche ARM [BOIDIDOU, S. MIDDLETON et al. 2016]

Équipe *MMLAB* [PHAN et al. 2016]

L'approche proposée repose sur deux classifieurs de type forêt aléatoire (figure 4.5). 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*).

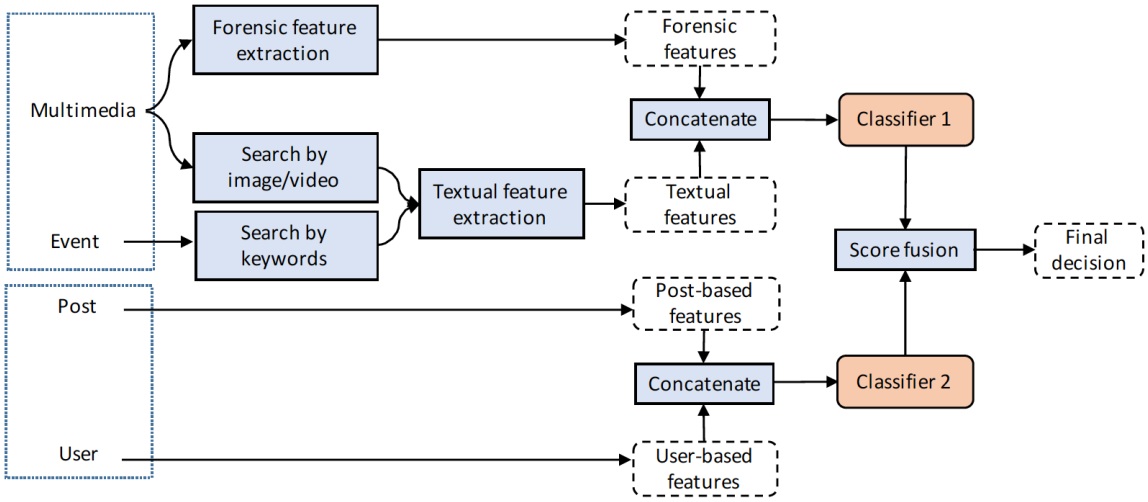


FIGURE 4.5 – Illustration de l'approche proposée par [PHAN et al. 2016]

Équipe *MCG-ICT* [CAO et al. 2016]

La première approche proposée se base sur le contenu textuel des messages et est illustrée dans la figure 4.6. 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`.

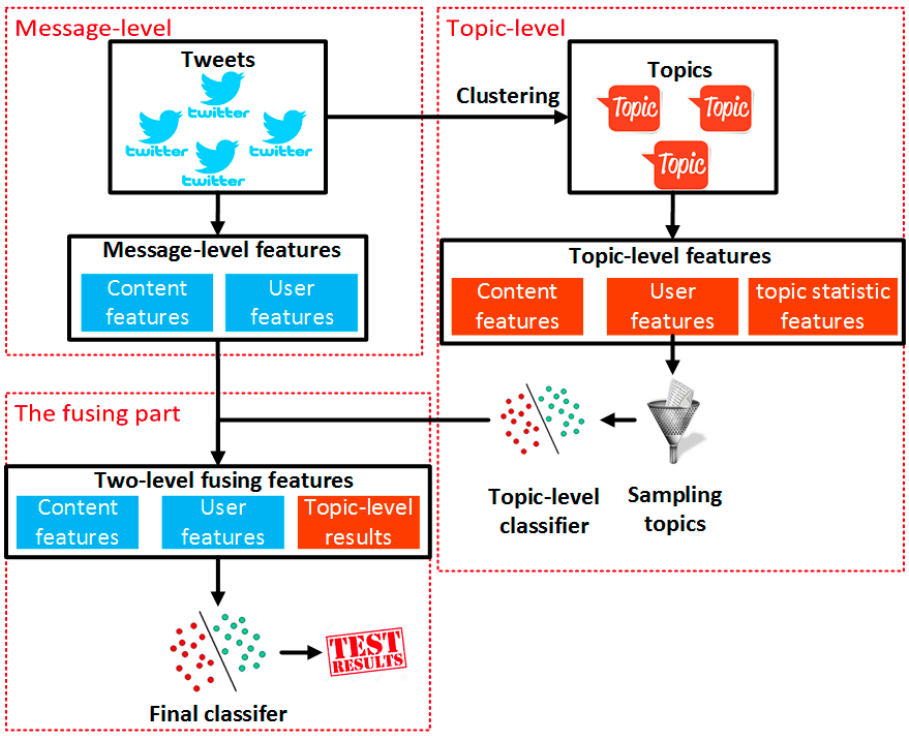


FIGURE 4.6 – Illustration de l'approche basée sur le texte et proposée par [CAO et al. 2016]

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 comme le montre la figure 4.7. 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.

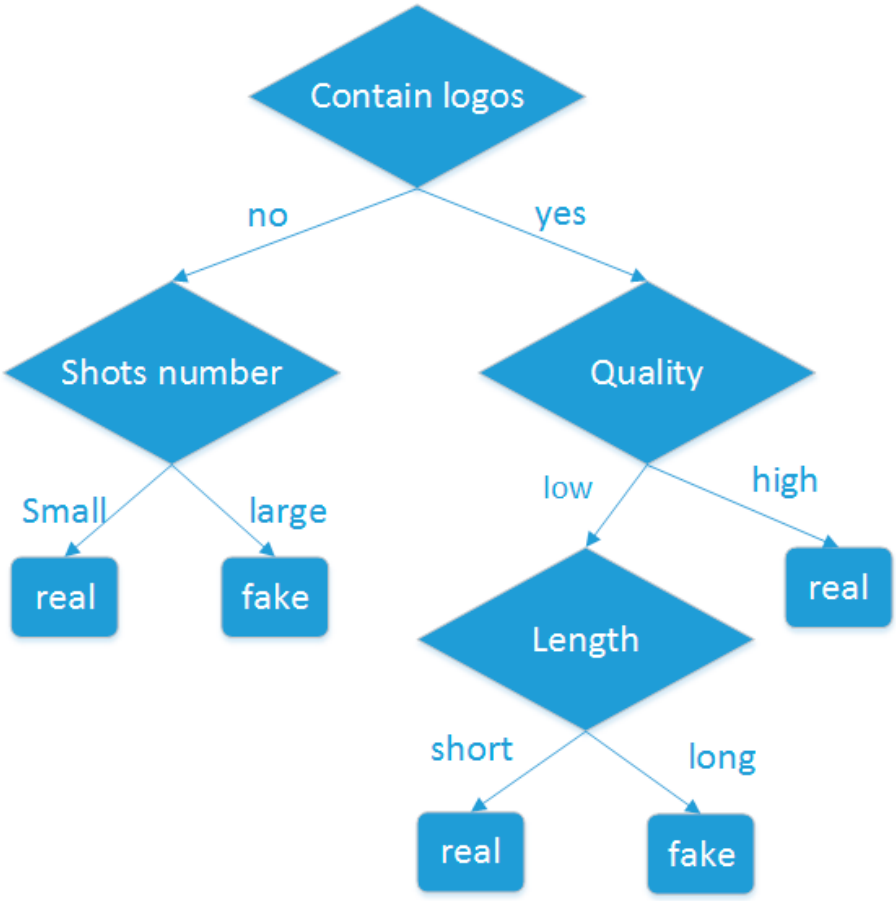


FIGURE 4.7 – Illustration de l'approche basée sur la vidéo et proposée par [CAO et al. 2016]

4.4 Résultats et discussions des différentes approches

4.4.1 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 4.2 (cf. tableau 4.1). 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*).

À la place,la micro-*F-Mesure* et le taux de bonnes classifications (*accuracy*) sont utilisées car ces dernières 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 4.3, ré-évalués selon le protocole décrit ci-dessus, sont présentés dans le tableau 4.2 pour une évaluation par message et dans le tableau 4.3 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.

TABLE 4.2 – 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)

TABLE 4.3 – 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 (%) (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)

4.4.2 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 faisant 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 4.3.4). 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 sept élémentaires plus `VMU-S1`), seront utilisées dans la suite.

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 1 230 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.

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 %).

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 2 228) 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

FIGURE 4.8 – 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

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 4.8).

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 4.3.4) 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 4.2). 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).

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, la figure 4.9 reporte 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 moyen-

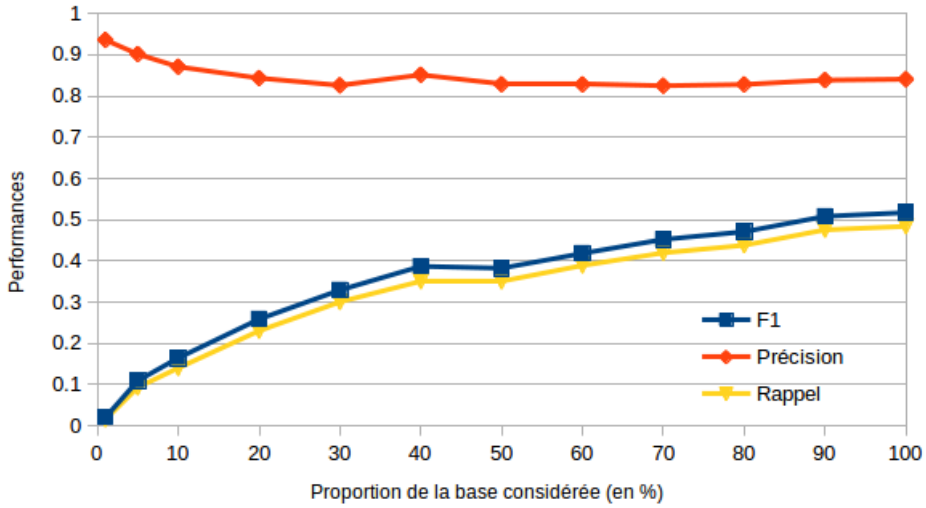


FIGURE 4.9 – Évolution des performances de l'approche image en fonction de la taille de la base d'image (en pourcentage).

nés. 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.

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 4.5.3.

4.5 Stratégies de fusion

4.5.1 Fusion simple des soumissions

Une fusion directe des prédictions du tableau 4.2 est d'abord étudiée dans cette partie. Pour réaliser cette fusion, chaque message est décrit par 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, un système référence est présenté 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, ses performances sont évaluées lorsqu'il est entraîné sur tous les messages de tous les événements sauf un événement dont les messages servent de jeu de test. Chaque événement est passé dans ce rôle de test, puis les résultats sont moyennés. C'est donc l'équivalent du *leave-one-out*, sauf que nous raisonnons au niveau des événements et non pas des exemples pris individuellement. Les deux cadres d'évaluation vus précédemment sont adaptés : une prédiction par message, et une prédiction par groupe de message partageant le même contenu multimédia. Les résultats sont respectivement présentés dans les tableaux 4.4 et 4.5. Une astérisque précise 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 4.10, la répartition des tweets est indiquée sous forme d'histogramme selon le nombre de méthodes les classant correctement. Comme on peut le voir, tous les messages sont correctement classés par au

TABLE 4.4 – 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)*

TABLE 4.5 – 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)*

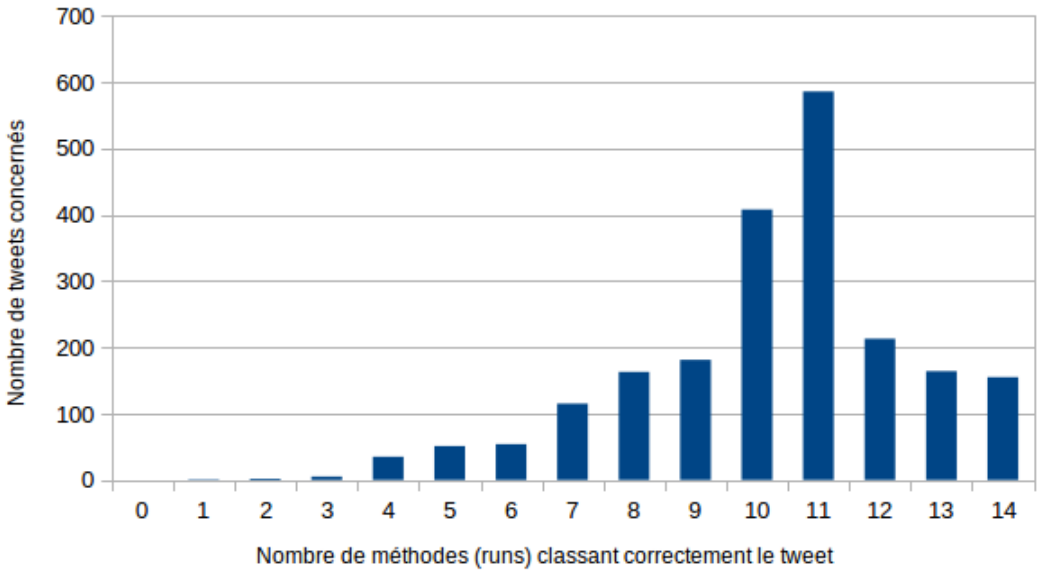


FIGURE 4.10 – Histogramme des messages selon le nombre de méthodes les classant correctement ; évaluation par groupe de messages partageant un même contenu multimedia

FIGURE 4.11 – Exemple de messages difficiles à classer (mal classés par plus de 12 méthodes des participants et mal classés par les modules de fusion).

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 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 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*).

Deux exemples de tels tweets sont donnés en figure 4.11.

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

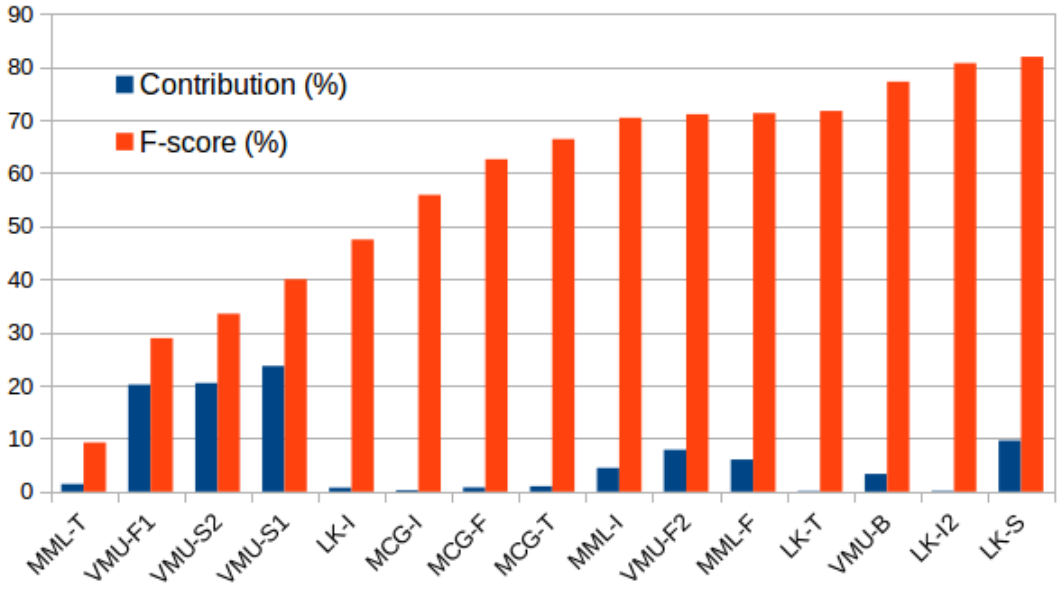


FIGURE 4.12 – Contributions de chacun des systèmes dans la fusion mesurée par indice de Gini sur les forêt aléatoires, mis en regard de la F-mesure de ces systèmes ; évaluation par groupe de messages partageant un même contenu multimédia

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 4.12, 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 n'obte-

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

4.5.2 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 4.4.2) sont présentés dans le tableau 4.6. 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.

TABLE 4.6 – 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 4.5. 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 4.13, mais offre finalement des performances identiques.

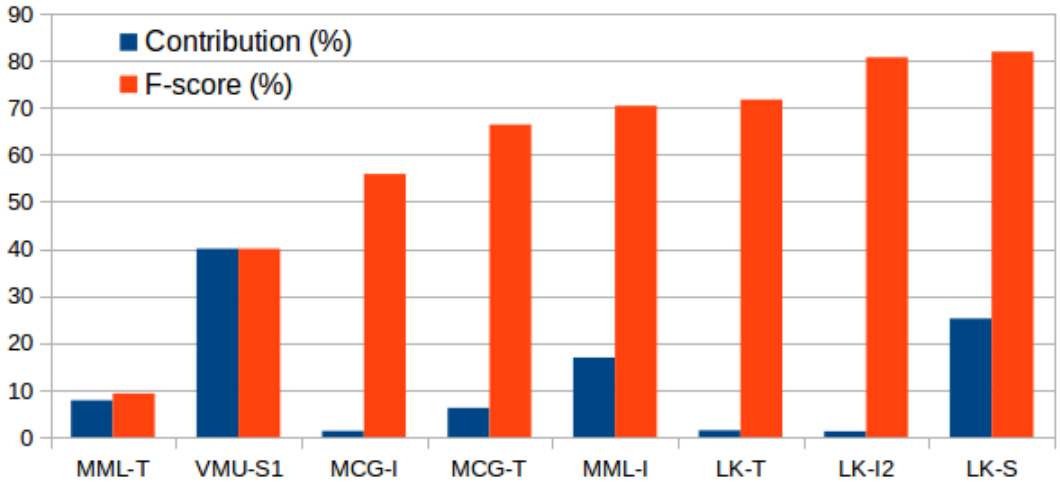


FIGURE 4.13 – Contributions de chacun des systèmes dans la fusion des systèmes élémentaires, mesurée par indice de Gini sur les forêt aléatoires, et mis en regard de la F-mesure de ces systèmes

TABLE 4.7 – 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

4.5.3 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 4.7 les résultats obtenus par les mêmes expériences de fusions, restreintes aux approches élémentaires n'utilisant aucune ressource externe aux données d'entraînement.

TABLE 4.8 – 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

Les performances sont cette fois-ci inférieures aux précédentes tentatives de fusion, et même inférieures à certaines des méthodes prises isolément (cf. tableau 4.3). 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.

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 4.8 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.

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 4.9.

TABLE 4.9 – 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*

Comme on peut le constater les résultats dans les deux cas sont très bons, mais il est plus intéressant d'entrainer tout le réseau d'un bloc que par niveau La différence est statistiquement significative (test de Wilcoxon avec $p = 0,05$). En effet, avec la stratégie d'entraînement 1, les résultats sont du même niveau que ceux d'une fusion directe de toutes les méthodes.

4.6 Conclusion

Dans ce chapitre, 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* sont proposées et étudiées [MAIGROT, KIJAK et CLAVEAU 2017 ; MAIGROT, CLAVEAU et KIJAK 2018]. 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. De plus, l'approche basée sur les images que nous avons proposé possède plusieurs biais importants. C'est pourquoi nous l'avons repris afin d'éviter ces biais et proposer une approche permettant la détection et la localisation de modifications dans une image dans le chapitre suivant (chapitre 5). 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 ce chapitre est l'intérêt de proposer des systèmes fusionnant des approches différentes. La stratégie la plus performante semble être 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 cette étude.

Beaucoup de pistes restent ouvertes à l'issue de ce travail. 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) sont mis à disposition sur le site http://hoaxdetector.irisa.fr/.

D'un point de vue technique, plusieurs problèmes peuvent être corrigés, mis en évidence par nos expérimentations, tout particulièrement les systèmes s'appuyant sur les images (*e.g.* images modifiées considérées comme similaires à l'image réelle initiale, images non retrouvées). Cette problématique fait l'objet de l'étude proposé dans le chapitre suivant (chapitre 5). En effet, le traitement de l'image doit être plus poussé afin notamment d'effectuer 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, CLAVEAU et KIJAK 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 ici, issue de multiples systèmes fusionnés par des techniques permettant difficilement l'explicativité de la décision finale (notamment pour les réseaux de neurones).

CHAPITRE 5

DÉTECTION DE MODIFICATIONS DANS UNE IMAGE

Contents

5.1 Introduction

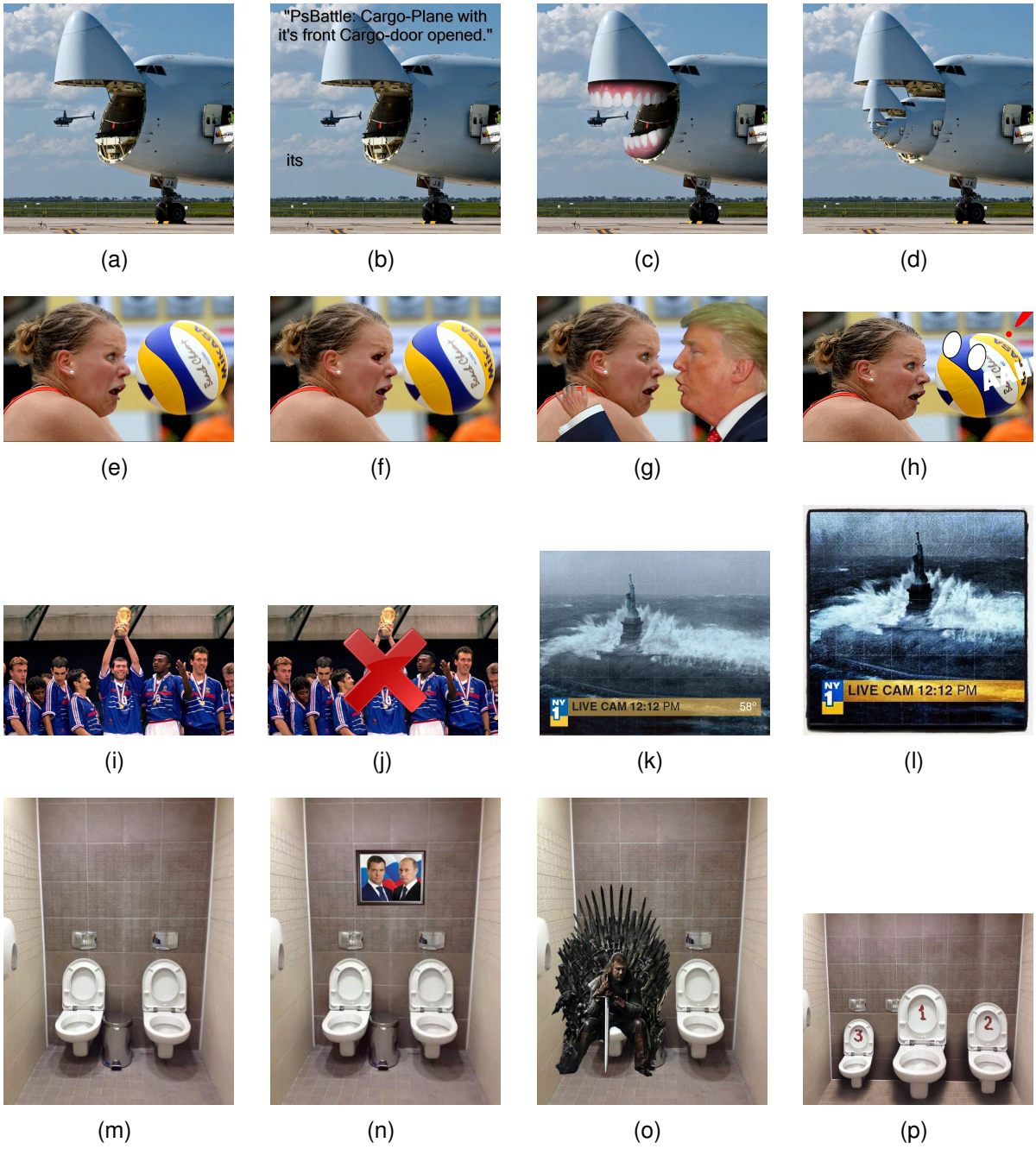
Les images présentes sur les réseaux sociaux sont nombreuses. Elles aident à illustrer un propos et à le préciser en apportant le plus souvent des informations supplémentaires au texte. Cependant, elles sont présentes dans de nombreux cas de fausses informations du fait de leur modification. Plusieurs exemples de modification dans une image sont donnés dans la figure 5.1. C'est pour cette raison qu'il est important pour un système de détection de fausses informations de pouvoir traiter une image individuellement et de se prononcer quant à son authenticité.

Comme défini dans l'état de l'art (chapitre 2), il existe trois types d'attaques sur une image que nous souhaitons étudier : la duplication d'une zone dans l'image, l'insertion d'un élément provenant d'une autre image et le rééchantillonnage. Une brève description de ces attaques est donné ci-dessous, des descriptions plus complète sont données dans le chapitre 2.

La **duplication** est la copie d'une zone de l'image pour être ajoutée ailleurs dans la même image. Un exemple d'image ayant reçue plusieurs duplications est l'image 5.1(d). L'**insertion** est très similaire à la duplication à la différence près que la zone ajoutée à l'image provient d'une autre image. Deux exemples d'images ayant reçues une insertion sont les images 5.1(g) et 5.1(n). Le **rééchantillonnage** correspond à toutes les transformations qui peuvent être appliquées sur l'ensemble de l'image (*e.g.* redimensionnement, rotation, etc.). Un exemple de rééchantillonnage est la modification appliquée à l'image 5.1(l) par rapport à sa version originale qui est l'image 5.1(k).

Cependant, ces trois catégories ne suffisent pas à définir pleinement une modification qui peuvent être décrites à un autre niveau en caractérisant le contenu sémantique de l'élément modifié. Cela est valable aussi bien pour une modification provenant d'une duplication que d'une insertion. Une différence peut alors être faite en distinguant les modifications représentant une partie de corps humain (image 5.1(f)), un objet (image 5.1(m)) ou encore un texte (image 5.1(b)). De même, chaque modification peut être caractérisée comme falsifiant l'image ou non. Par exemple, l'insertion d'une flèche dans l'image pour attirer l'attention de l'utilisateur ne semble pas modifier le sens de l'image (image 5.1(j)). À l'inverse, l'ajout d'une personne sur l'image change le sens de cette dernière et vise à tromper l'utilisateur (image 5.1(g)).

De telles caractérisations, plus complètes, peuvent être réalisées sur les images de la figure 5.1. Une caractérisation de la modification dans la figure 5.1(b) est *in-*



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

(m) (n) (o) (p)

FIGURE 5.1 – Cinq images non modifiées (images (a), (e), (i), (k) et (m)) et leurs versions modifiées (au centre et à droite).

sertion d'un texte ne modifiant pas le sens de l'image, alors que la modification de la figure 5.1(d) est *duplication d'un objet modifiant le sens de l'image*.

La majeure partie des approches dans la littérature concernant la détection de modifications dans une image se concentrent sur un seul type d'attaque (duplication ou insertion) ou nécessite une connaissance à priori sur l'image (par exemple *spécialisation aux images au format Joint Photographic Experts Group (JPEG)*). Nous proposons ici une approche visant à détecter le plus grand nombre de modifications sans connaissance sur l'image.

Le traitement proposé dans ce chapitre se fait à grâce à trois modules pouvant être utilisés successivement ou individuellement selon les connaissances sur l'image à analyser. Ce traitement est présenté dans la figure 5.2.

L'idée globale de l'approche mise en place est, étant donné une image requête $\mathcal{R}$ (repère **A** dans la figure 5.2) , de rechercher dans une base d'images connues (repère **B**) une ancienne version de $\mathcal{R}$, dite *candidate* et notée $\mathcal{C}$, grâce à un premier module de recherche d'images similaires (repère **I**). Si il existe une image $\mathcal{C}$ (repère **C**), une comparaison entre $\mathcal{R}$ et $\mathcal{C}$ est réalisée (repère **II**). Ce deuxième module produit une prédiction binaire $\mathcal{P}$ de chaque pixel comme étant modifié ou non (repère **D**). Les pixels prédits comme modifiés forment alors une ou plusieurs zones correspondant aux différences entre les deux images. Par application de $\mathcal{P}$, les zones détectées comme modifiées sont extraites (repère **E**). La carte binaire $\mathcal{P}$ permet aussi une visualisation des prédictions pour être, par exemple, montrée à l'utilisateur (repère **F**). Si il existe au moins une différence entre $\mathcal{R}$ et $\mathcal{C}$, alors $\mathcal{C}$ est considérée comme modifiée. Les différences trouvées sont traitées par un troisième module (repère **III**) qui vise à caractériser chaque modification en réalisant une classification de chaque imagette (partie de l'image $\mathcal{R}$) entre quatre classes qui sont *Visage*, *Texte*, *Forme* et *Autre* (repère **G**). Le repère **H** montre la vérité terrain $\mathcal{V}$ associé à $\mathcal{R}$ qui correspond à la prédiction parfaite et, de ce fait, à l'objectif à atteindre par le système.

La suite du chapitre est organisé comme suit : une présentation des jeux de données présents dans la littérature et ceux constitués durant cette thèse est réalisée dans la section 5.2. Les trois modules évoqués précédemment sont ensuite détaillés successivement dans les sections 5.3, 5.4 et 5.5. Enfin, la section 5.6 présente les conclusions à ces travaux.

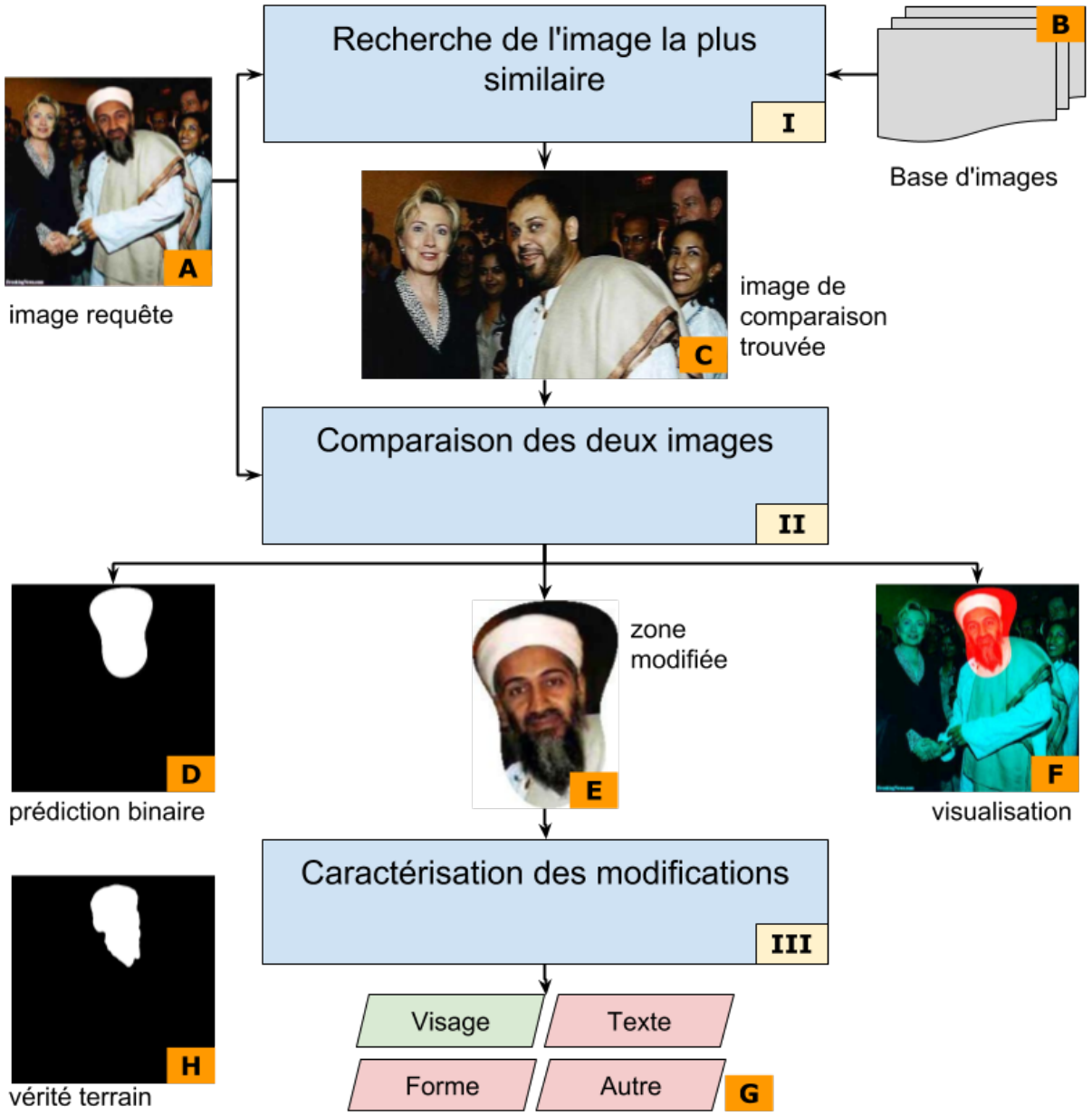


FIGURE 5.2 – Illustration de l'approche image mise en place

5.2 Données utilisées

Afin de tester le système illustré dans la figure 5.2, réunir des images modifiées ou non ne suffit pas du fait de la stratégie de comparaison avec une seconde image. Il est ainsi nécessaire d'utiliser des triplets d'images correspondant à une image requête, une image de comparaison et la vérité terrain de la modification. Ces trois images sont les images $\mathcal{R}$ (repère **A**), $\mathcal{C}$ (repère **C**) et $\mathcal{V}$ (repère **H**) dans la figure 5.2 .

Pour remplir ce rôle, plusieurs jeux de données ayant des tailles et difficultés différentes ont été proposés dans la littérature afin d'évaluer les méthodes de détection de modifications dans une image. Ces jeux de données diffèrent par le réalisme des modifications appliquées aux images (d'une simple modification directe à une modification associée à plusieurs post-traitements pour masquer les transformations), par le type de modifications (duplication, insertion et/ou rééchantillonage) et par la présence ou non des *vérités terrain*. Un exemple d'une vérité terrain est donnée dans la figure 5.3.

(a) (b) (c)

FIGURE 5.3 – Masque binaire de vérité terrain (à droite) pour la modification sur l'image modifiée (au centre) par rapport à l'image originale (à gauche).

5.2.1 Jeux de données issus de la littérature

MICC-F600 ($\mathbf{MICC_{F600}}$) Le jeu de données $\mathbf{MICC_{F600}}$ est composé de 440 images originales (non modifiées), de 160 images modifiées ainsi que des 160 masques de vérité terrain associés aux images modifiées [AMERINI, BALLAN, CALDELLI, BIMBO et al. 2013]. Dans la suite de ce manuscrit, les sous-ensembles des images originales et modifiées de $\mathbf{MICC_{F600}}$ sont respectivement notés $\mathbf{MICC^{ori}_{F600}}$ et $\mathbf{MICC^{mod}_{F600}}$

Les images originales sont issues des 1 300 images du jeu de données **MICC$_{F2000}$** [AMERINI, BALLAN, CALDELLI, DEL BIMBO et al. 2011] et les 160 images modifiées du jeu de données **SATS** [CHRISTLEIN, RIESS et ANGELOPOULOU 2010]. Ce jeu de données présente des images contenant des modifications réalistes et représentant un réel défi étant donné la présence éventuelle de plusieurs modifications dans une même image. Cependant, toutes les modifications consistent en des duplications. Des exemples d'images modifiées issues de ce jeu de données sont donnés dans la figure 5.4 accompagnées des vérités terrains associées.

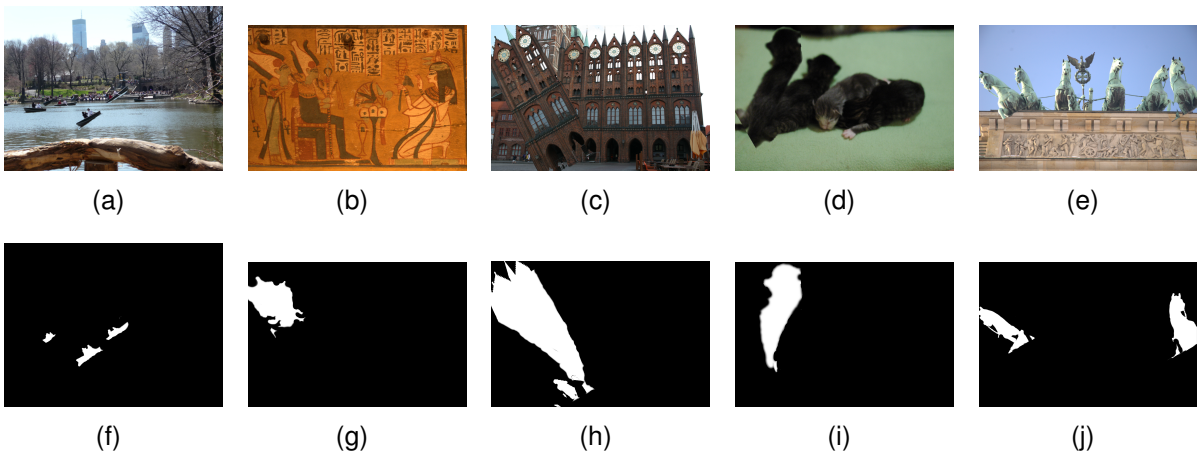


FIGURE 5.4 – Cinq exemples d'images modifiées du jeu de données **MICC$_{F600}$** (ligne du haut) et les masques de vérité terrain associés à ces images (ligne du bas).

Wild (WW) Le jeu de donnée initial est proposé par [ZAMPOGLOU, PAPADOPOULOS et KOMPATSIARIS 2015] et correspond à 80 images modifiées pour lesquelles les auteurs ont effectué des requêtes sur *Google Search* et *TinyEye* pour obtenir des images similaires présentes sur le web (suppression des copies parfaites) pour finalement obtenir 13 477 images. Le jeu de donnée ne propose pas une image originale pour les 80 images modifiées initiales. De plus, la recherche inversée sur *Google Search* et *TinyEye* retourne de nombreuses images identiques. Nous sélectionnons alors les images modifiées pour lesquelles nous avons au moins une image de comparaison et nous n'en gardons qu'une lorsque plusieurs images sont disponibles. De même, les auteurs proposent parfois plusieurs masques de vérité terrain lorsque plusieurs modifications sont présentes (un masque par modification). Lorsque c'est le cas, nous fusionnons les masques pour n'en obtenir qu'un. Nous nous intéressons ici à seule-

ment 77 images modifiées auxquelles nous associons une seule des images originales trouvées ainsi que le masque de vérité terrain associé à cette image modifiée. Le but est d'obtenir des triplets {*image modifiée*, *image originale* et *vérité terrain*} pour chaque image modifiée. Ce jeu de données présente l'avantage d'être varié et de correspondre à des cas réels d'images circulant sur les réseaux sociaux, mais est assez petit en nombre d'images modifiées. La notation **WW** dans la suite du manuscrit correspond au sous-ensemble de 77 triplets et non au jeu de données initial. Des exemples d'images de ce jeu de données sont donnés dans la figure 5.5.

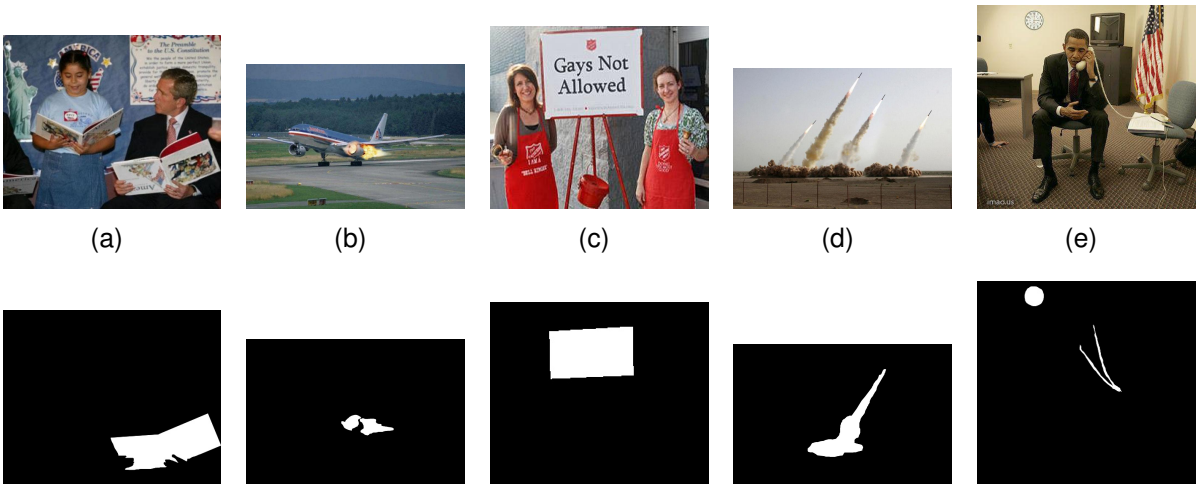


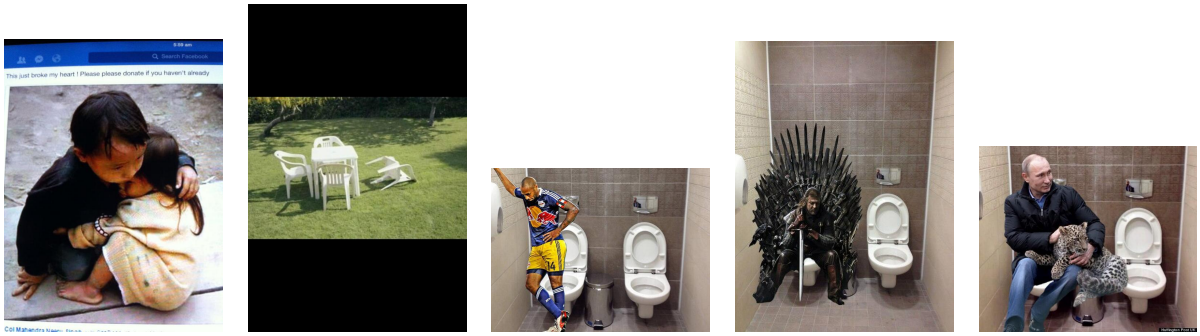
FIGURE 5.5 – Cinq exemples d'images modifiées du jeu de données **WW** (ligne du haut) et les masques de vérité terrain associés à ces images (ligne du bas)

MediaEval (Me) **Me** est composé de 316 images associées à des tweets utilisés dans le cadre de la tâche *Verifying Multimedia Use* de l'atelier *Mediaeval 2016*[1]. Des exemples d'images de ce jeu de données sont donnés dans la figure 5.6.

Parmi ces 316 images, nous séparons 18 images pour former un ensemble d'images, noté $\mathbf{Me_{req}}$ dans la suite du manuscrit, qui seront utilisées pour tester les systèmes. Ces images ont été sélectionnées manuellement et possèdent leur version originale dans le jeu de données **HB** (présenté par la suite). Cet autre jeu de données $\mathbf{Me_{req}}$ présente un réel défi en terme de détection au niveau des modifications des images qu'il comporte puisque les images qui le compose sont des vraies images

1. voir la présentation de la tâche dans le chapitre 4

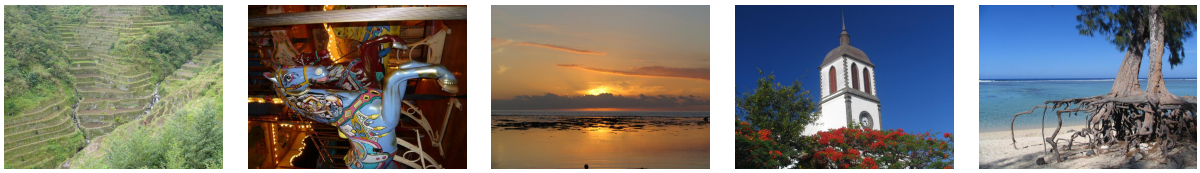
modifiées trouvées sur les réseaux sociaux. Les masques de vérité sont construits manuellement pour ces requêtes. Nous obtenons donc un autre jeu de données composé de duo {*image modifiée* et *vérité terrain*} qui permet d'augmenter légèrement le nombre d'images modifiées disponibles pour tester le système (les images originales étant dans le jeu de données **HB** présenté plus tard dans cette section).



(a) (b) (c) (d) (e)

FIGURE 5.6 – Présentation du jeu de données **Me**

Holidays (Ho) Ce jeu de données proposé par [JEGOU, DOUZE et C. SCHMID 2008] correspond à des photos réelles prises dans divers contextes et comprend 1 492 images non modifiées. Il est utilisé dans notre cas pour tester le système de recherche d'image similaire avec des requêtes négatives (*i.e.* ne possédant pas d'image similaire dans la base). Dans le cadre de notre étude, cela correspond à la situation d'une nouvelle fausse information sur les réseaux sociaux et n'étant pas encore connu de la base d'images. Des exemples d'images de ce jeu de données sont donnés dans la figure 5.7.



(a) (b) (c) (d) (e)

FIGURE 5.7 – Présentation du jeu de données **Ho**

5.2.2 Jeux de données constitués dans le cadre de la thèse

La plupart des jeux de données existants dans le domaine de la détection d'images modifiées se concentrent sur des attaques par duplication ou sont limités en taille (par exemple les jeux de données **WW** et **Me** présentés précédemment). Cependant, nous souhaitons orienter notre travail sur un plus grand nombre de modifications et avoir un nombre plus important de requêtes. Cela nécessite donc la constitution de jeux de données supplémentaires à ceux déjà présents dans la littérature. Nous avons ainsi constitué plusieurs jeux de données pour répondre à ces différentes attentes impossibles à vérifier autrement.

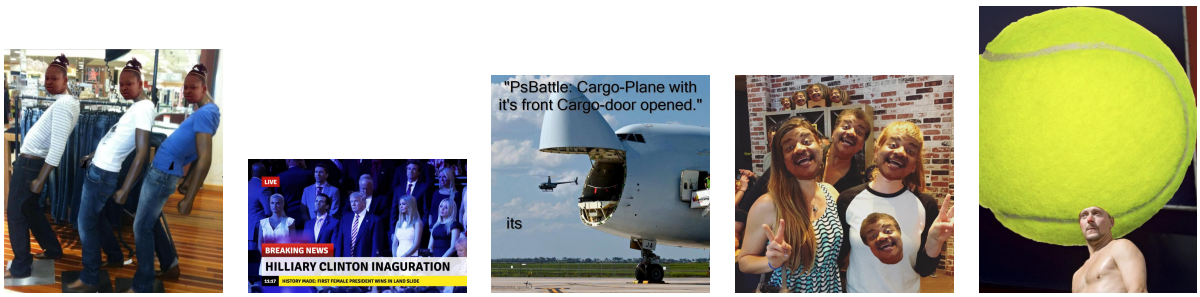
Reddit (Re) Reddit [2] est un site web communautaire de partage. Nous nous sommes intéressés ici aux *photoshop battles* durant lesquels un utilisateur poste une image initiale, puis d'autres utilisateurs soumettent des versions modifiées de cette image. Le but est de réaliser la meilleure modification possible de l'image initiale et cela peut prendre plusieurs formes (la plus drôle, la plus réaliste, etc.).

Cela a pour effet d'engendrer des versions de la première image avec des modifications très diverses (*e.g.* duplication, insertion, ajout de texte, etc.). La collecte des images s'est basée sur le principe que l'image initiale est dite *originale* et toutes les images en réponse sont des versions *modifiées* de cette image originale. Il est important de noter que l'image postée en première est possiblement une version déjà modifiée par rapport à une autre image non présente dans les images récoltées. Nous considérons ici que ce n'est pas le cas et nous prenons en compte exclusivement les modifications réalisées à partir de cette version *originale*.

Re est une collection de 129 images originales et leurs versions modifiées pour un total de 383 images. Parmi ces images, 106 images modifiées sont annotées manuellement par trois annoteurs avec un accord inter-annoteurs de $75,12\%$ en terme de score Jaccard ce qui correspond à un très bon accord. L'annotation de ce jeu de données porte sur la réalisation du masque de vérité terrain binaire des modifications dans l'image. Des exemples d'images de ce jeu de données sont donnés dans la figure 5.8.

Twitter (Tw) Ce jeu de données correspond à des publications issues des sujets tendance (*i.e. top tweet*) sur Twitter [3] entre le 1^{er} Janvier 2017 et le 31 Mars 2017.

2. http ://reddit.com/
3. https://twitter.com/



(a) (b) (c) (d) (e)

FIGURE 5.8 – Présentation du jeu de données **Re**

Pour cela, huit zones géographiques sont ciblées afin d'obtenir à la fois des messages francophones et anglophones : la France, la Belgique, la Suisse et le Québec pour le français et le Royaume-Unis, les États-Unis d'Amérique, le Canada et l'Australie pour l'anglais.

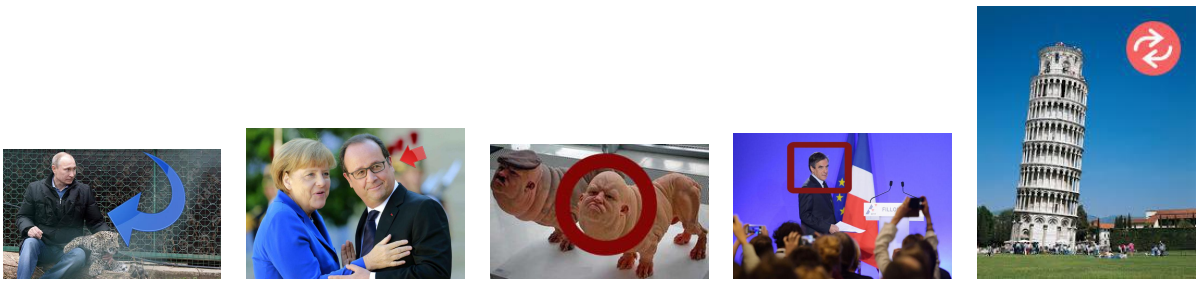
Nous supprimons ensuite les images totalement identiques afin de ne pas avoir de doublon parfait. Finalement, nous obtenons un jeu de données de 82 543 images représentant l'actualité de *Twitter* durant cette période sur 170 sujets (*i.e.* *top-tweets*) différents.

Toutes ces images sont utilisées pour représenter les sujets populaire sur Twitter entre le 1er Janvier 2017 et le 31 Mars 2017, mais aussi pour remplir la base d'images en tant que distracteurs.

Parmi ces images, nous sélectionnons 23 images sur lesquelles nous appliquons des modifications typiques des réseaux sociaux (assemblages de photos, recadrement ou insertion de texte ou icônes). Ces images modifiées forment un autre ensemble d'images requêtes pour tester notre système. Ce nouveau jeu de données est noté $\mathbf{Tw_{req}}$ dans la suite de ce manuscrit. Des exemples d'images de ce jeu de données sont donnés dans la figure 5.9.

Hoaxbuster (HB) **HB** est un jeu de données issu de cinq sites [4, 5, 6, 7, 8] de référencement des fausses informations connues. Nous avons collecté 8 035 images qui

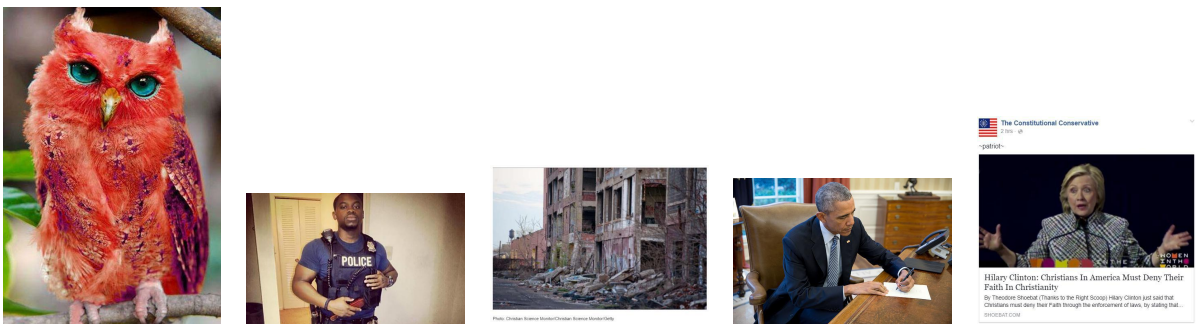
4. http://hoaxbuster.com
5. http://hoax-busters.org
6. http://urbanlegends.about.com
7. http://snopes.com
8. http://hoax-slayer.com



(a) (b) (c) (d) (e)

FIGURE 5.9 – Présentation du jeu de données **Tw**

sont toutes liées plus ou mois directement à une fausse information. Aucune information supplémentaire n'est connue sur ces images (images modifiées ou non), sauf qu'elles sont toutes liées à une fausse information. Cela veut dire qu'il peut aussi s'agir d'images originales ayant servies à réaliser des images modifiées. Des exemples d'images de ce jeu de données sont donnés dans la figure 5.10.



(a) (b) (c) (d) (e)

FIGURE 5.10 – Présentation du jeu de données **HB**

5.3 Recherche d'images similaires par le contenu

La première étape de l'analyse détermine de façon binaire si, étant donné une image requête $\mathcal{R}$, il existe une image $\mathcal{C}$ dans la base permettant une comparaison avec $\mathcal{R}$. Le but est ici de trouver l'image originale ayant servie à produire $\mathcal{R}$.

La mise en place de cette stratégie de comparaison avec une image de référence nécessite de déterminer quelle image appartenant à notre base d'images connues doit être utilisée pour remplir ce rôle d'image de référence. Le second rôle de cette première

étape est d'estimer si il est nécessaire de passer à l'étape 2, soit la comparaison des deux images.

Pour cela, un système de recherche d'images basée sur le contenu (*Content-Based Image Retrieval* (CBIR)) est mis en place.

La base d'images possédant une trop grande taille pour estimer en détail la similarité entre l'image $\mathcal{R}$ et chaque image de la base, une première étape consiste en la sélection d'une courte liste d'images similaires, puis à une analyse plus détaillée sur les images ainsi pré-sélectionnées. Cette approche est illustrée dans la figure 5.11.

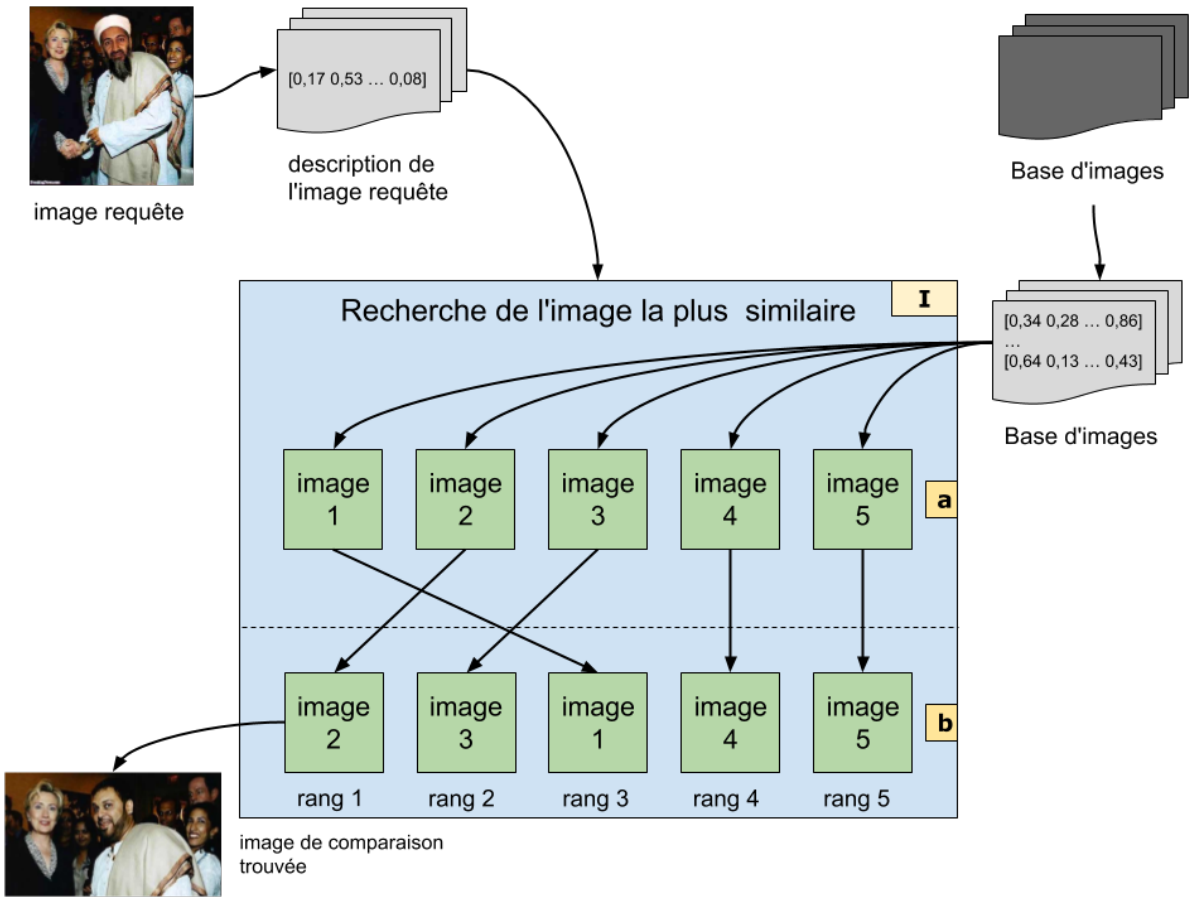


FIGURE 5.11 – Représentation de la première étape : Recherche d'une image de comparaison

5.3.1 Description des images

La recherche d'images similaires nécessite la création de représentations comparables pour toutes les images de la base et les requêtes. Nous utilisons une représentation issue d'un réseau de neurones et choisissons le réseau `VGG-19` [SIMONYAN et ZISSERMAN 2014] qui montre des performances hautes sur plusieurs problèmes de reconnaissance visuelle [TOLIAS, SICRE et JÉGOU 2016]. Le réseau `VGG-19`, représenté dans la figure 5.12, est entraîné sur le jeu de données **ILSVRC** [KRIZHEVSKY, SUTSKEVER et HINTON 2012] pour une tâche de reconnaissance de contenu.

Le jeu de données **ILSVRC** est un sous-ensemble de *ImageNet* contenant 1 000 classes avec 1 000 images par classe. La diversité des classes apprises permet une représentation globale des images intéressante. Quelques exemples de classes apprises sont *télévision*, *trimaran*, *parapluie*, *mésange*, *poule* et *requin marteau*[9].

Ces images ont reçu une annotation manuelle grâce à l'outil *Amazon's Mechanical Turk*. Les annotations réalisées au niveau de l'image indiquent la présence ou l'absence d'une classe représentant un objet dans cette image, par exemple "Présence d'un chien" ou "Absence d'un chien".

Le but de l'utilisation de ce réseau est d'obtenir un vecteur de description du contenu de l'image en passant les images dans ce réseau pré-entraîné et de récupérer le vecteur de sortie d'une des couches intermédiaires du réseau afin de l'utiliser comme une description d'une image en se basant sur le principe de *transfert learning* [BABENKO et al. 2014], notre utilisation du réseau et notre problématique de recherche n'étant pas exactement les mêmes que celles ayant servies leur de l'entrainement du réseau.

Les différentes couches d'un réseau de neurones permettent de décrire de différentes manières une image. Les vecteurs de données transmis entre les différentes couches des réseaux de neurones tendent à décrire des éléments plus ou moins précis en fonction de la profondeur des couches du réseaux. Ainsi, les dernières couches décrivent des formes alors que les premières couches décrivent les détails des images. Souhaitant obtenir une description en fonction du contenu, nous nous intéressons exclusivement aux vecteurs de sortie des dernières couches du réseaux.

Elles étudions la dernière couche convolutionnelle, soit la couche $conv_{5\_4}$, et les deux couches entièrement connectées (*fully connected* en anglais) fc_1 et fc_2 dans la

9. Liste des 1 000 classes : https://gist.github.com/yrevar/942d3a0ac09ec9e5eb3a#file-imagenet1000_clsid_to_human-txt

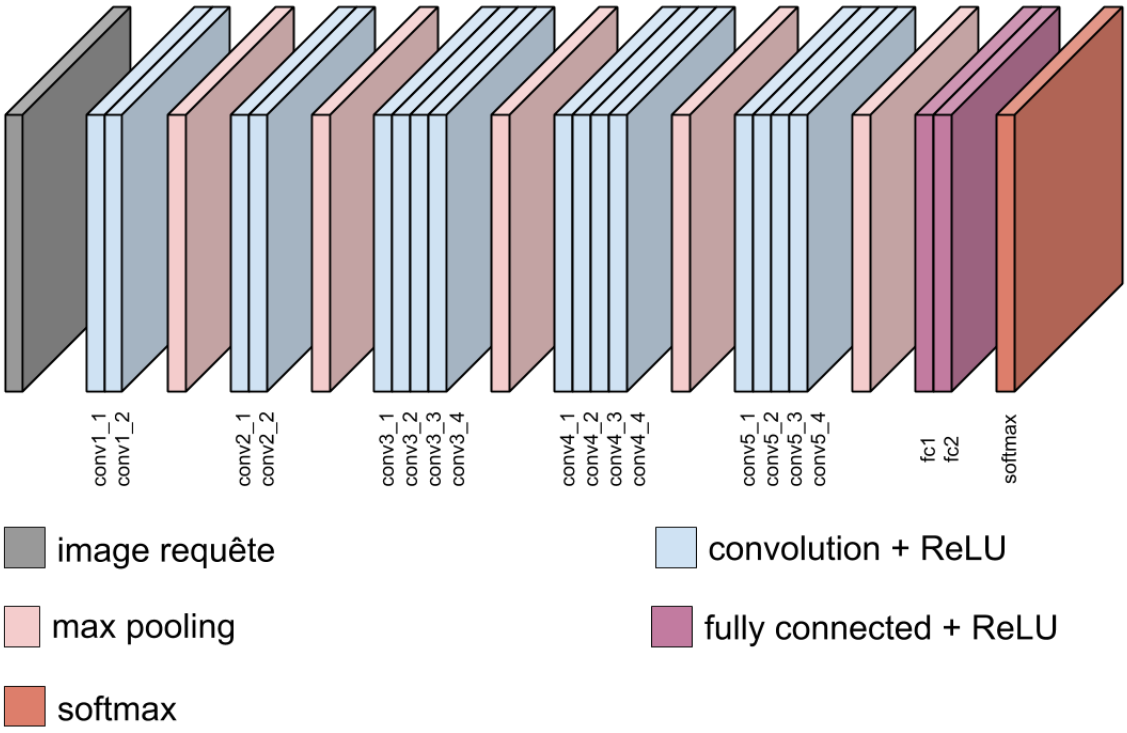


FIGURE 5.12 – Décomposition du réseau VGG-19 [SIMONYAN et ZISSERMAN 2014]

figure 5.12.

Il est à noter que cette approche nous fournit un descripteur global sur l'image à décrire. Le but ici est d'obtenir un seul descripteur permettant de représenter l'image entière.

Afin de pouvoir comparer correctement les vecteurs de description entre les images, il est nécessaire d'appliquer une normalisation commune à tous les vecteurs. Pour cela, deux types de normalisation sont pensées : une normalisation ℓ_2 et une normalisation *Power*. Cette dernière consiste en le remplacement de toutes les valeurs par la valeur de leur racine carré avant la normalisation ℓ_2.

Adaptation du réseau pour des images de plus grandes tailles

Une des limites du réseau VGG-19 est sa dépendance à analyser des images de petites tailles (224×224) ce qui nécessite un fort redimensionnement des images avant leur passage dans le réseau. Cela peut engendrer une trop grande perte d'information

	taille standard : 224×224	mise sous forme de noyau : 576×576
$conv_{5\_4}$	$7 \times 7 \times 512$	$18 \times 18 \times 512$
fc_6	$1 \times 1 \times 4096$	$7 \times 7 \times 4096$
fc_7	$1 \times 1 \times 4096$	$7 \times 7 \times 4096$

TABLE 5.1 – Taille du vecteur de description en fonction de l'approche utilisée et de la couche sélectionnée

du contenu de l'image. Pour éviter cela, [TOLIAS, SICRE et JÉGOU 2016] propose une mise sous forme de noyau des couches entièrement connectées.

Cela permet d'utiliser la première partie du réseau `VGG-19` composée de couches convolutionnelles avec des images de tailles différentes à 224×224. Les couches entièrement connectées, transformées en couches convolutionnelles sous forme de noyau, sont appliquées sur le vecteur en sortie de $conv_5$.

Le tableau 5.1 présente la taille du vecteur obtenu en fonction de la couche utilisée et de l'approche mise en place (*i.e.* standard ou sous la forme de noyau). Certaines couches fournissant un tensor, nous appliquons un *pooling* sur le vecteur obtenu dans le but de décrire les images par des vecteurs de forme $1 \times 1 \times x$, où x vaut soit 512 soit 4096.

5.3.2 Recherche des images candidates

Une fois toutes les images contenues dans la base et les requêtes décrites comme expliqué précédemment, chaque image requête est comparée aux images de la base par le biais de leur descripteur respectif. Le but est de passer de l'ensemble des images de la base à une liste d'images candidates à être l'image la plus comparable. Le descripteur calculé a pour but de décrire le contenu de l'image dans son ensemble et ainsi de trouver les images de la base représentant un contexte proche (*e.g.* une image de forêt permet de trouver les images représentant une forêt dans la base).

Pour cela, le choix du calcul de similarité utilisé est le produit scalaire entre les descripteurs, ces derniers ayant été normalisés. Les images de la base sont alors retournées sous la forme d'une liste, ordonnées par score décroissant de similarité avec l'image requête. Dans le but d'accélérer cette recherche, nous mettons en place une recherche par *KD-tree* [BENTLEY 1975]. La recherche par *KD-tree* retourne les images de la base triées.

Nous souhaitons maintenant retenir exclusivement les images les plus similaires. Pour cela, il est possible d'effectuer un tri grâce à un score minimum de similarité δ requis. Une valeur haute pour δ permet d'assurer une précision haute quant aux candidats retournés. Cependant plus la zone modifiée dans notre image requête est grande, plus le score de similarité avec l'image de la base considérée comme une bonne réponse attendue sera bas. Il est donc nécessaire de déterminer une valeur pour δ assez haute pour ne pas retourner trop de résultats incorrects mais assez tolérant quant au score de similarité obtenu entre les deux images lorsque la modification est grande. Le constat étant que lorsqu'une image considérée comme vrai positif est présente dans la base, elle sera le plus souvent en première position des images retournées mais il est nécessaire de savoir reconnaître lorsque la première image retournée est un faux positif. C'est pourquoi nous optons pour garder les 10 images les plus similaires de la base et effectuons un filtrage des candidats en testant individuellement chaque image candidate retenue par cette première étape.

5.3.3 Filtrage des candidats

Il est nécessaire de tester la cohérence d'une comparaison entre l'image requête et chacune des images candidates afin de trouver l'image la plus comparable parmi la liste d'images similaires trouvées (*reranking*). Nous mettons en place une méthodologie calculant le score *s* de comparabilité d'une image $\mathcal{R}$ par rapport à une image $\mathcal{C}$. Ce filtrage est basé sur le calcul d'une homographie $\mathcal{H}$ de $\mathcal{R}$ sur $\mathcal{C}$ et permet la réalisation d'un filtrage spacial pour éliminer les faux positifs.

Afin de se placer dans le contexte des réseaux sociaux dans lequel les images sont redimensionnées pour ne pas surcharger les serveurs du réseau social (problématique directement liée à la forte fréquentation de ces sites), nous traitons les images de sorte que la hauteur ou la largeur des images n'excèdent pas 900px. Ce traitement est appliqué indépendamment à $\mathcal{R}$ et $\mathcal{C}$.

De nombreux descripteurs sont proposés dans la littérature pour décrire localement une image. Le plus connu est *Scale-Invariant Feature Transform* (transformation de caractéristiques visuelles invariante à l'échelle) (SIFT) [Lowe 2004]. Cependant, nous utilisons à la place des descripteurs *Speeded Up Robust Features* (caractéristiques robustes accélérées en français) (SURF) [Bay, Tuytelaars et Van Gool 2006] qui ont l'avantage d'être plus rapide et plus robustes aux différentes transformations d'images

par rapport aux descripteurs SIFT.

En plus du choix des descripteurs, il est nécessaire de déterminer les points d'intérêt servant à décrire l'image. Pour cela deux approches existent : l'extraction détectée et dense. Une extraction détectée se base sur la détection des points d'intérêts. À l'inverse, une extraction dense échantillonne les points d'intérêt selon une grille régulière.

Une fois les ensembles $\mathcal{D}_1$ et $\mathcal{D}_2$ calculés, respectivement les points d'intérêt des images $\mathcal{R}$ et $\mathcal{C}$, nous souhaitons réaliser un appariement entre chaque descripteurs de $\mathcal{D}_1$ et le descripteur le plus similaire de $\mathcal{D}_2$. L'ensemble des appariements est noté $\mathcal{M}$.

Il faut maintenant déterminer dans quelle mesure $\mathcal{R}$ peut se projeter dans $\mathcal{C}$, pour cela nous devons calculer une homographie grâce à $\mathcal{M}$ en utilisant l'algorithme RANSAC [FISCHLER et BOLLES 1981], mais tous les appariements ne doivent pas être utilisés. C'est pourquoi, nous sélectionnons uniquement les appariements vérifiant le critère de Lowe [LOWE 2004] détaillé ci-dessous. Le critère de Lowe permet de ne pas prendre en compte les appariements basés sur des points d'intérêts issus de zones lisses qui ont tendance à rendre difficile l'estimation de l'homographie.

Definition 1 *Critère de Lowe [LOWE 2004] : Pour chaque descripteur a^i, on associe les deux descripteurs plus proches de la seconde image, notés b^{j1} et b^{j2} respectivement pour le descripteur le plus proche et le deuxième plus proche. Un appariement est valide si le ratio entre les scores de proximités mesuré par $D(a^i, b^{j1})/(a^i, b^{j2}) < r$ où $D(x, y)$ est le score de proximité entre les descripteurs x et y et r un seuil entre 0 et 1 donné. L'ensemble des appariements valides $\mathcal{D}$ est alors :*

$$\mathcal{D} = (a^i, b^{J1(i)}, i \in \{1, \dots, N_Q\} : \frac{D(a^i, b^{j1(i)})}{D(a^i, b^{j2(i)})} \leq r$$

avec :

$$J1(i) = \underset{j \in \{1,\dots,N_C\}}{\operatorname{argmin}} D(a^i, b^j)$$

et

$$J2(i) = \underset{j \in \{1,\dots,N_C\} \setminus J1(i)}{\operatorname{argmin}} D(a^i, b^j)$$

Nous obtenons $\mathcal{M}_{strict}$ ne possédant que les appariements entre les couples de descripteurs se ressemblant le plus. Nous appliquons l'algorithme RANSAC sur ce sous-ensemble $\mathcal{M}_{strict}$ afin de calculer une homographie $\mathcal{H}$ de $\mathcal{R}$ vers $\mathcal{C}$. Un exemple d'homographie calculée est donné dans la figure 5.13. Enfin, nous souhaitons évaluer $\mathcal{H}$

afin d'obtenir un score correspondant à la cohérence d'une comparaison entre l'image requête $\mathcal{R}$ et cette image candidate $\mathcal{C}$.

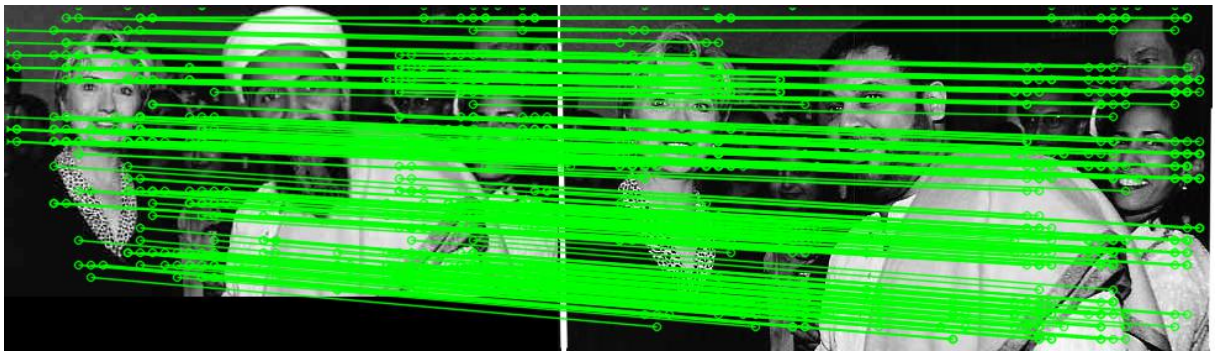


FIGURE 5.13 – Exemple de projection produit par application de l'homographie $\mathcal{H}$

Afin d'attribuer un score à cette homographie, nous appliquons $\mathcal{H}$ sur l'ensemble $\mathcal{M}_{strict}$. Un match est dit *vérifiant l'homographie* si la distance entre le point appareillé issu de $\mathcal{C}$ et le point théorique calculé par l'homographie $\mathcal{H}$ est inférieur à 10% de la taille de la diagonale de $\mathcal{C}$.

Nous calculons le score de comparabilité entre une image requête et chacun de ses candidats et gardons le candidat obtenant le meilleur score.

5.3.4 Expérimentations

La plupart des systèmes de recherche d'images similaires sont évalués sur des jeux de données représentant plusieurs points de vue d'un même object. Cependant, nous souhaitons tester dans quelle mesure notre système est capable de retourner une quasi-copie de notre image requête ou non si aucune copie existe. Nous évaluons en outre le comportement de notre système avec des images modifiées et dégradées.

L'ensemble d'images requêtes est composé d'images modifiées diverses et d'images non modifiées. Nous utilisons un ensemble de 2 151 images. Cet ensemble est à la fois composé de requêtes positives (possédant une réponse dans la base) et négatives (ne possédant pas de réponse dans la base) : 160 images de $\textbf{MICC}^{\textbf{mod}}_{\textbf{F600}}$, 106 images modifiées de **Re** et 40 images modifiées de $\textbf{Me}_{\textbf{req}}$ composent les requêtes positives et modifiées. S'ajoutent à ces dernières les 440 images de $\textbf{MICC}^{\textbf{ori}}_{\textbf{F600}}$ qui sont aussi des images requêtes positives mais non modifiées. Le jeu de données **Ho** forment les requêtes négatives.

La base d'images utilisée contient les images correspondantes aux versions originales des images modifiées. De plus, des images sont utilisées pour jouer le rôle de distracteurs (bruit). Plus spécifiquement, l'ensemble d'images associées à la base est composé de 93 121 images :

— 82 543 images provenant du jeu de données **Tw** ;
— 8 035 images provenant du jeu de données **HB** ;
— 316 images provenant du jeu de données **Me** ;
— 129 images originales provenant du jeu de données **Re** ;
— 98 images provenant du jeu de données **SATS** ;
— 2 000 images provenant du jeu de données $\mathbf{MICC_{F2000}}$ qui contiennent les images originales associées au jeu de données $\mathbf{MICC^{ori}_{F600}}$.

Contrairement à la majorité des systèmes de recherche d'images basés sur le contenu qui sont évalués en terme de précision (P@k, mAP, ...), nous évaluons notre systèmes en terme de taux de bonne classification, calculé sur toutes les requêtes à partir de la formule suivante :

$$taux_{bonne\ classification} = \frac{nombre\ d'images\ bien\ classées}{nombre\ d'images}$$

où *"nombre d'images bien classées"* correspond au nombre de cas où le système a le bon comportement. Cela ce caractérise par deux réponses possibles du système :

1. Aucune image retournée, si il s'agit d'une requête négative ;
2. L'image attendue, dans le cas d'une requête positive.

En effet, nous souhaitons que le système de recherche d'images similaires trouve l'image la plus similaire à l'image requête mais ne se prononce pas si aucune quasi-copie n'est trouvée dans la base d'images.

La figure 5.14 montre le taux de bonne classification du système pour différentes valeurs de seuil δ avec ou sans le système de filtrage. Nous observons que le meilleur seuil est $\delta = 0,9$ avec un taux de bonne classification de $91,91\%$ avec l'étape de filtrage géométrique et $81,08\%$ sans. Ces tests ont pour but de montrer d'estimer une valeur de δ pour minimiser le nombre d'images retournées par la première partie tout en gardant de bonnes performances (au lieu de systématiquement garder les 10 images les plus similaires).

La Table 5.2 montre les performances du système sur chaque jeu de données pour un seuil δ donné (au lieu de garder les 10 images les plus similaires). Nous observons

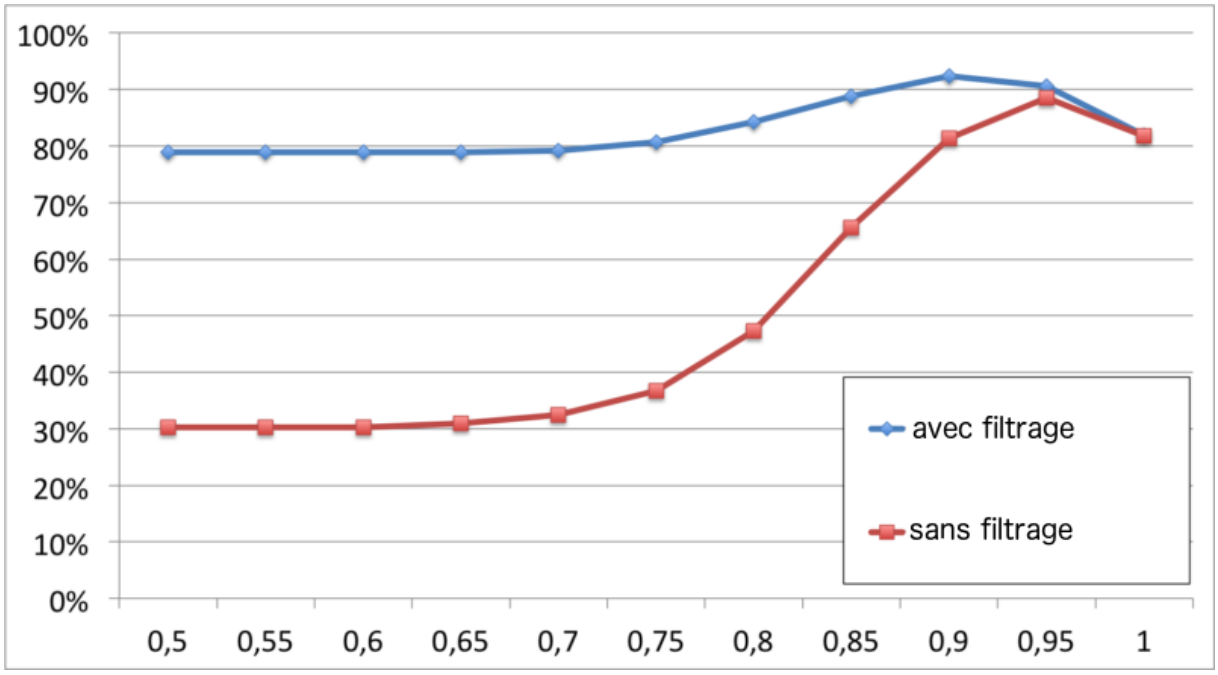


FIGURE 5.14 – Taux de bonne classification du système en fonction du seuil de similarité δ

un gain de performance lorsque δ est bas sur les jeux de données **Re**, **MICC$^{\text{mod}}_{\text{F600}}$** et **Me** (requêtes positives). À l'inverse, **Ho** offre de meilleures performances lorsque le seuil δ est haut. Nous vérifions l'hypothèse qu'un seuil bas favorise les requêtes positives mais génère de nombreux faux positifs.

Nous observons que le système de recherche d'images échoue principalement lorsque la zone modifiée est très grande par rapport à l'image. Ceci est particulièrement illustré par de mauvaises performances sur **Me**. Ce petit ensemble de requêtes a été spécialement choisi pour défier le système, qui est perturbé par des modifications trop grandes (plus de 50% de la taille de l'image) ou des insertions de bordures / bannières. Un exemple d'association réussie malgré une falsification assez importante et un faux positif sont donnés dans figure 5.15. Ici, le faux positif est supprimé par le filtrage géométrique.

TABLE 5.2 – Taux de bonne classification du système de recherche d'images similaires par jeu de données pour différentes valeurs du seuil $\mathcal{T}$

$\mathcal{T}$	**Re**	**MICC$_{F600}$**	**Me**	**Ho**
0.75	**73.62%**	**99.83%**	**32.50%**	74.68%
0.80	73.23%	**99.83%**	**32.50%**	80.58%
0.85	71.65%	99.50%	**32.50%**	88.41%
0.90	64.57%	98.50%	20.00%	96.09%
0.95	37.80%	94.00%	15.00%	**100.00%**

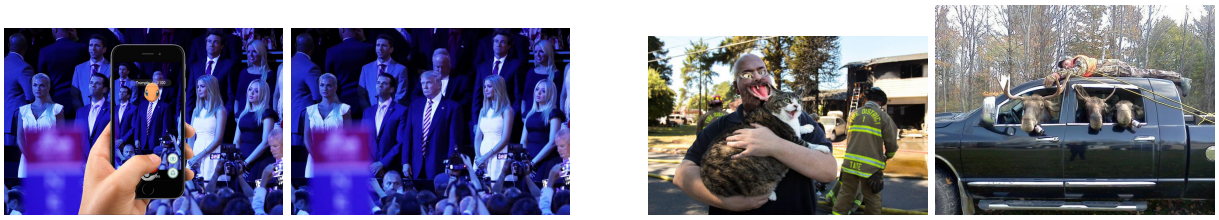


FIGURE 5.15 – Exemples de vrais positifs (à gauche) et faux positifs (à droite)

5.4 Détection et localisation des modifications

Ce deuxième module réalise la comparaison entre deux images. Plus précisément, le module détecte les modifications dans une image $\mathcal{R}$ en considérant l'image $\mathcal{C}$ comme référence. Nous posons l'hypothèse dans cette approche que l'image $\mathcal{R}$ est une version modifiée de $\mathcal{C}$. Les cas ne vérifiant pas cette hypothèse sont discutés plus tard dans le chapitre.

Les zones modifiées dans les images sur internet et les réseaux sociaux sont souvent la cible de transformations telles que des rotations, des redimensionnements, des recadrements ou des transformations affines. Le but de ce module est de proposer une approche permettant de détecter le plus grand nombre de modifications et étant limité par le moins de pré-traitements possibles. Pour cela, le système de filtrage de l'étape précédente est repris en grande partie.

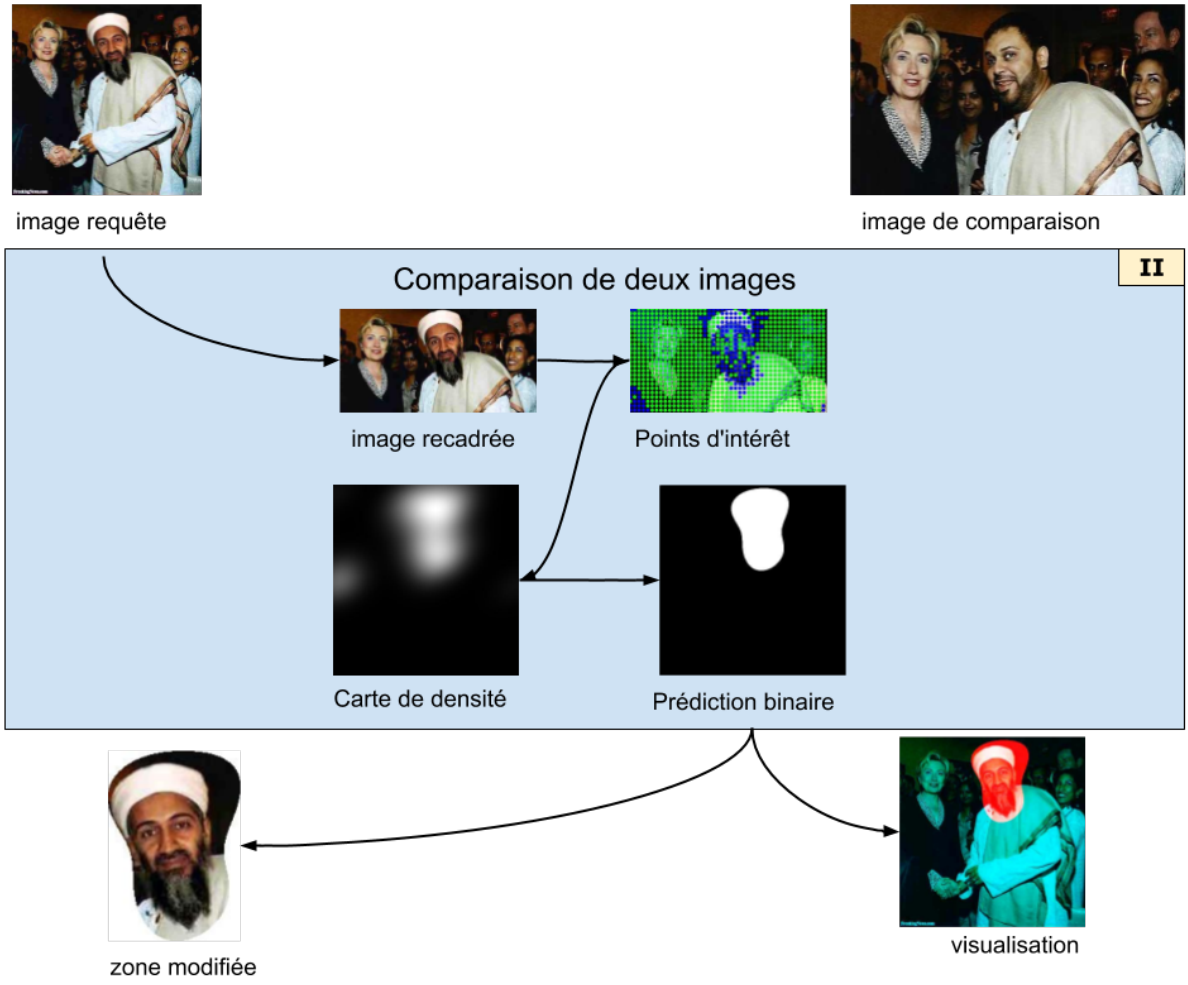


FIGURE 5.16 – Représentation de la deuxième étape : Comparaison des deux images

5.4.1 Approche basée sur un appariement des descripteurs locaux

Dans notre implémentation, les descripteurs SURF sont utilisés comme descripteurs locaux. Les descripteurs SURF sont calculés et appariés de $\mathcal{R}$ vers $\mathcal{C}$ selon le critère de Lowe comme introduit par [LOWE 2004] [10] qui permet de neutraliser la plus grande partie des mauvais appariements. Le critère de Lowe permet de ne pas prendre en compte les appariements basés sur des points d'intérêts issus de zones lisses qui ont tendance à rendre difficile l'estimation de l'homographie.

10. formule du critère de Lowe donné dans la présentation du premier module.

En effet, les zones lisses génèrent beaucoup de point d'intérêt ayant des valeurs très proches se qui engendre des erreurs d'appariement lors de la recherche du plus proche voisin.

Basé sur ces appariements, l'homographie $\mathcal{H}$ entre les deux images est estimée en utilisant l'algorithme *RAndom SAmple Consensus* (RANSAC) qui est robuste aux *outliers*.

Nous évaluons une extraction des points d'intérêt (*keypoints*) dense et détectés. La détection des points d'intérêt engendre souvent un grand taux d'appariements qui vérifient l'homographie, mais donne moins de points et par conséquence moins d'appariements ce qui cause des problèmes lorsque le nombre de points détectés est trop faible. L'extraction dense des points d'intérêt est donc utilisée dans cette étude (avec une grille régulière de 10 pixels) et les descripteurs sont calculés sur quatre échelles différentes, comme dans d'autres problèmes de reconnaissance [CHATFIELD et al. 2011].

Une fois l'homographie $\mathcal{H}$ estimée, chaque point d'intérêt de $\mathcal{R}$ est associé à deux points :

1. son plus proche voisin parmi les points d'intérêt de $\mathcal{C}$;
2. sa projection dans $\mathcal{C}$ selon $\mathcal{H}$.

Un point d'intérêt de $\mathcal{R}$ est considéré comme un *outlier* si la distance entre sa projection dans $\mathcal{C}$ selon $\mathcal{H}$ et les points d'intérêt appariés pour chaque échelle de descripteur est supérieur à $0.1 \times \text{diag}$, où diag est la taille de diagonale de l'image. Un point d'intérêt est un *inlier* si au moins l'appariement d'une des échelles de descripteur vérifie $\mathcal{H}$ (Fig. 5.17(d)).

La prédiction est finalement obtenue par estimation de la densité des *outliers* en utilisant un noyau d'estimation de densité (*Kernel Density Estimation*) avec un noyau gaussien (*gKDE*), la règle de Scott est utilisé pour estimer la largeur de la bande passante (*bandwidth*). La prédiction obtenue attribue une probabilité à chaque pixel d'être modifié (valeur de chaque pixel comprise entre 0 et 1). Cette prédiction est appelée *Tampering Heat Maps* (THM) dans la suite de ce manuscrit. Il est maintenant nécessaire de transformer cette prédiction THM en une prédiction binaire afin de prédire une modification au niveau du pixel.

Le but de cette étape est double :

1. unifier la prédiction : les *outliers* déterminés suite à l'estimation de l'homographie

$\mathcal{H}$ sont espacés de 10px ce qui donne une prédiction similaire à celle présentée dans la figure 5.17(d) ;

2. éliminer le bruit : une erreur d'appariement peut avoir lieu et produire un *outlier* isolé.

On compare cette approche à une approche plus simple, appelée *morpho*. L'approche consiste en l'application d'opérateurs morphologiques (dilatation, ouverture, fermeture et remplissage des cavités) sur la carte binaire donnée par les *outliers*. Ceci peut être vu comme une approximation d'une estimation de densité de noyau avec un noyau uniforme (rectangulaire). L'application d'ouverture puis d'une fermeture supprime les points isolés et améliorent la connectivité spatiale (Fig. 5.17(g)).

Dans certains cas, $\mathcal{H}$ est mal estimée : lorsque $\mathcal{R}$ a été retourné (*flip*) ou lorsque l'image originale retrouvée $\mathcal{C}$ est une version recadrée de $\mathcal{R}$: la zone de $\mathcal{R}$ qui n'existe pas dans $\mathcal{C}$ sera détectée comme étant modifiée alors que ce n'est pas le cas. Ces cas sont facilement détectable après l'estimation de l'homographie $\mathcal{H}$ et sont prévenus par des rotations et recradrements de $\mathcal{R}$ en fonction de $\mathcal{H}$.

Recadrage

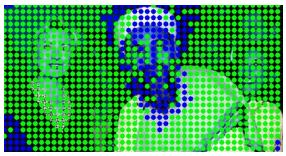
Dans le cas d'une image $\mathcal{C}$ recadrée, la zone retirée dans $\mathcal{C}$ est considérée comme modifiée dans $\mathcal{R}$ car elle n'est pas retrouvée dans $\mathcal{C}$. Ce cas est visible sur les figures 5.17(a) et 5.17(b) où la partie basse de la figure 5.17(a) n'est pas présente dans la figure 5.17(b).

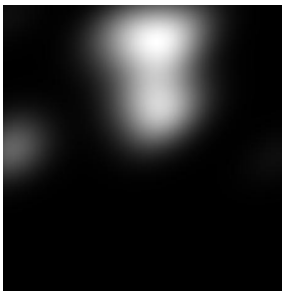
Cependant, il n'est pas nécessaire de considérer la partie retirée comme une duplication ou une insertion et de la traiter lors de la troisième partie de notre approche. Nous souhaitons donc traiter les parties retirées différemment en les détectant en tant que recadrage. Il est donc nécessaire de savoir détecter ces parties retirées pour ne pas les traiter par la suite. Pour cela, nous utilisons un appariement des points d'intérêt comme précédemment, mais de $\mathcal{C}$ vers $\mathcal{R}$ cette fois-ci. En estimant l'homographie $\mathcal{H}'$ de $\mathcal{C}$ vers $\mathcal{R}$ grâce à cette appariement il est possible de retrouver l'opération de recadrage et ainsi traiter les deux images recadrées de la même façon (par exemple les figures 5.17(c) et 5.17(b).


(a) image modifiée $\mathcal{R}$


(b) image originale $\mathcal{C}$


(c) image recadrée de $\mathcal{R}$


(d) outliers


(e) carte THM par *gKDE*


(f) binarisation de *gKDE*


(g) carte binaire *morpho*

FIGURE 5.17 – Exemple de résultat de la méthode de détection et localisation de modifications. Dans (d) : les *outliers* sont en bleus et les *inliers* en vert.

5.4.2 Expérimentations

Protocole expérimental

L'évaluation des performances consiste en la comparaison de la vérité terrain avec la prédiction binaire THM. Cette évaluation est donc sensible à deux paramètres : le seuil de binarisation et la qualité du masque de vérité terrain. Différentes métriques sont utilisées, chacune permettant l'évaluation d'une propriété. Pour chaque métrique, les prédictions sont binarisées en utilisant plusieurs valeurs de seuil couvrant les différentes valeurs possibles.

Scores au niveau des pixels : Une méthode d'évaluation habituelle pour la localisation lors d'une comparaison au niveau des pixels entre une sortie binaire et un masque de vérité terrain. La localisation de modifications est évaluée par le taux de faux positifs (*False Positive Rate (FPR)*) et le taux de faux négatifs (*False Negative Rate (FNR)*) qui doivent être minimisées.

Les pixels faux positifs peuvent soit venir d'un seuil de binarisation défavorable donné par des régions trop grande comparées à la vérité terrain, soit d'une région faus-

sement détectée. Ainsi, certains taux d'erreurs égaux peuvent ne pas avoir la même signification dans les deux cas. La même observations est valable pour les faux négatifs. Une autre difficulté est que l'évaluation ne prend pas en compte les composantes connexes et l'habilité à séparer deux zones modifiées proches.

Score basé sur les composantes connexes : Pour ces raisons, nous avons également utilisé une évaluation basée sur des composantes connexes telle que définie dans d'autres tâches de reconnaissance visuelles telles que les défis de localisation d'objets [EVERINGHAM et al. 2010]. Un "critère de chevauchement" est défini comme une intersection sur une union supérieure à $0,5$. Les composantes connexes de la prédiction sont affectées à des composantes connexes de la vérité terrain et jugées comme étant des vrais / faux positifs en mesurant leur chevauchement, étant donné la formule suivante :

$$\mathcal{A}_o = \frac{|CC_p \cap CC_{gt}|}{|CC_p \cup CC_{gt}|} \tag{5.1}$$

où CC_p correspond à la composante connexe prédite, CC_{gt} la composante connexe de la vérité terrain, $|.|$ correspond à l'aire et $CC_p \cap CC_{gt}$ (resp. $CC_p \cup CC_{gt}$) est l'intersection (resp. union) des composantes connexes prédite et de la vérité terrain. Pour être considérée comme une détection correcte, l'aire du chevauchement $\mathcal{A}_o$ doit être supérieure à un seuil $\mathcal{A}$. Ce seuil est habituellement définie à 0,5 (50%) dans les tâche de localisation d'objets.

Par conséquent, chaque composante connexe prédit est soit un vrai positif (TP) soit un faux positif (FP), et chaque composante connexe à la vérité terrain est soit un vrai positif (TP) soit un faux négatif (FN).

Score issu de WW : Pour une question de capacité de comparaison avec les approches similaires de l'état de l'art, nous utilisons aussi la méthodologie d'évaluation présentée dans [ZAMPOGLOU, PAPADOPOULOS et KOMPATSIARIS 2015]. La similarité entre une prédiction THM et le masque de vérité terrain est aussi une évaluation réalisée au niveau des pixels selon la formule donnée par les auteurs :

$$E(P, V) = \frac{\Sigma(P \cap V)^2}{\Sigma(P) \times \Sigma(V)}$$

où P correspond à la prédiction binaire réalisée par le système, V le masque binaire correspondant à la vérité terrain et $\Sigma(x)$ l'aire de la zone active du masque binaire x. Tout prédiction THM binarisée, atteignant un score $E(P, V)$ supérieur à un seuil

donné, est considéré comme une détection réussie. La performance de l'algorithme est donnée par le nombre de cas considérés comme correctement détectés.

Résultats

Les résultats associés à cette partie du système sont présentés dans la table 5.3 pour les jeux de données **Re**, **WW** et $\mathbf{Me_{req}}$. Nous comparons les résultats pour différentes valeurs du seuil de binarisation σ utilisé pour l'approche utilisant *gKDE*. Ce seuil correspond à la valeur minimum que doit avoir un pixel après application de *gKDE*. Si un pixel à une valeur supérieure ou égale à σ, alors le pixel prend la valeur 255 (valeur maximale), sinon le pixel prend la valeur 0 (valeur minimale). Plus la valeur de σ est grande, plus la prédiction sera petite et probable. Concernant la comparaison au niveau des pixels, plus la valeur de σ est grande, plus nombre de faux positifs diminue alors que le nombre de faux négatifs augmente.

L'approche basée sur *gKDE* est sensible à la valeur de σ et tend à produire des prédictions plus larges que `morpho`. Si au moins deux *outliers* sont détectés, l'approche basée sur *gKDE* génère toujours une prédiction. À l'inverse, le bruit est relativement bien filtré par l'approche morphologique, mais cette approche est sensible aux éléments structurants utilisés par les opérateurs morphologiques. La binarisation morphologique tend à sur-segmenter, ce qui peut être très pénalisé par les mesures d'évaluation. La sévérité de ces mesures dépend aussi de la granularité des masques de la vérité terrain . Nous évaluons donc à la fois $\mathcal{A} = 0.5$ et $\mathcal{A} = 0.1$, où $\mathcal{A}$ est le ratio de chevauchement minimum pour valider un appariement de composante connexe.

Enfin, nous comparons l'extraction des points d'intérêt détectée et dense et l'appariement pour l'évaluation de l'homographie. L'utilisation des points d'intérêt détectés améliore légèrement les résultats finaux, au prix de plus d'échecs dans l'estimation de l'homographie quand il n'y a pas assez de points d'intérêt.

L'approche proposée fonctionne relativement bien, mais nous observons plusieurs limitations correspondant à des méthodes basées sur l'appariement de caractéristique. Par exemple, une modification consistant à changer la couleur d'un objet ne peut pas être caractérisée par des descripteurs SURF et ne sera donc pas détectée. Les grandes zones uniformes favorisent l'inadéquation des descripteurs, ce qui entraîne des taux plus élevés de valeurs aberrantes. Les images inversées (effet miroir) ne sont pas traitées par l'estimation de l'homographie. Tous ces cas sont présents dans le jeu de données **WW** et sont les principales causes d'échec.

En général, les très petites régions modifiées sont difficiles à détecter. Elles peuvent être filtrés lors de la création de la carte binaire ou simplement ne pas être détectés comme valeurs aberrantes. Cependant, la méthode est robuste pour les grandes zones altérées.

TABLE 5.3 – Localisation des modifications en fonction de l'extraction des descripteurs détecté/dense et du seuil utilisé pour la binarisation pour différentes valeurs de $\mathcal{A}$.

Méthode		Comparaison au niveau des composantes						
			Re		**WW**		**Me$_{req}$**	
		$\mathcal{A}$	Préc.	Rappel	Préc.	Rappel	Préc.	Rappel
Dense	*gKDE*	0.5	38..5	27.87	31.48	29.82	35.56	33.33
Dense	σ = 0.5	0.1	69.92	**50.82**	62.04	58.77	**57.78**	54.17
Dense	*gKDE*	0.5	10.00	7.10	19.19	16.67	30.43	29.17
Dense	σ = 0.75	0.1	66.15	46.99	**64.65**	56.14	54.35	52.08
Dense	morpho	0.5	14.45	26.78	10.73	29.82	14.00	43.75
Dense		0.1	27.43	**50.82**	23.66	**65.79**	22.00	**68.75**
Dét.	*gKDE* σ = 0.5	0.1	**75.23**	44.81	60.95	56.14	**57.78**	54.14

Méthode		Comparaison au niveau des pixels					
		Re		**WW**		**Me$_{req}$**	
		FPR	FNR	FPR	FNR	FPR	FNR
Dense	*gKDE* σ = 0.5	5.04	**39.48**	25.10	**16.89**	24.01	22.34
Dense	*gKDE* σ = 0.75	**1.72**	70.17	**10.86**	46.56	**8.07**	50.69
dense	morpho	2,26	62,97	19,57	31,26	14,36	27,52
Dét.	*gKDE* σ = 0.5	3,73	39.83	26.64	19.27	25.28	**22.02**

5.4.3 Comparaison d'approches similaires

[BROGAN et al. 2017] est une étude présentant des approches très similaire à LFM : Cette partie du chapitre s'intéresse à la comparaison du système LFM présenté précédemment avec les meilleures approches de cet article.

Comparaison à l'échelle des pixels (*Pixel-wise comparison*) Parmi les approches présentées dans ces travaux, les deux meilleures approches sont nommées `IRPSNR` et `SSIM` respectivement basées sur *Peak Signal to Noise Ratio* (PSNR) et *Structural Similarity Index Measure* (SSIM). Ces deux approches sont comparées à l'approche `LFM` en se basant sur le code et les paramètres fournis par les auteurs. Concernant `IRPSNR`, la prédiction THM est calculée comme étant l'PSNR au niveau des pixels entre les versions gaussiennes floues des deux images (avec l'écart type σ = 4). Concernant `SSIM`, la THM est le SSIM au niveau des pixels entre les deux images, en utilisant un rayon de voisinage de 32 pixels.

Dans les deux méthodes, les valeurs faibles de la THM indiquent des zones probablement altérées. Ces méthodes étant basées sur des calculs au niveau des pixels, une opération de déformation est nécessaire pour transformer l'une des images. Cette opération a pour but d'obtenir des images de même dimension avec pour même motivation que l'étape de recadrage de l'approche `LFM`. Dans [Brogan et al. 2017], $\mathcal{C}$ est déformée selon l'homographie H' qui apparie les points de $\mathcal{C}$ au système de coordonnées de $\mathcal{R}$. Dans ces expérimentations présentées, de la même manière que pour `LFM`, $\mathcal{R}$ peut être transformée lorsque $\mathcal{C}$ est détecté comme étant une version recadrée de $\mathcal{R}$.

Résultats et analyse

L'évaluation est basée sur les trois méthodes d'évaluation précédemment nommées. Concernant les données, nous utilisons **WW** qui est le jeu de données utilisé par [Zampoglou, Papadopoulos et Kompatsiaris 2015], ainsi que **Re** qui présente une grande variété aussi bien au niveau des types d'attaques réalisées que de la taille des attaques. Enfin, nous utilisons le même protocole d'évaluation que les auteurs de ce même article.

Efficacité des méthodes basées sur le contexte La table 5.4 présente les résultats en terme de nombre de cas correctement détectés par rapport au nombre total de cas considérés avec les trois meilleures performances rapportées en [Zampoglou, Papadopoulos et Kompatsiaris 2015] en comparaison. Les approches utilisant une image de référence performent beaucoup mieux que les images se basant exclusivement sur l'image elle-même [Farid 2009 ; Mahdian et Saic 2009 ; Z. Lin et al. 2009].

Cela confirme l'importance d'utiliser les indices de contexte lorsqu'ils sont disponibles.

TABLE 5.4 – Localisation des modifications dans **WW** : ratio entre le nombre de cas détectés et le nombre total de cas. Utilisation du score de **WW** avec $E > 0.45$

LFM	IRPSNR	SSIM
0.60	0.72	0.79
GHO [FARID 2009]	NOI1 [MAHDIAN et SAIC 2009]	ADQ1 [Z. LIN et al. 2009]
0.35	0.18	0.16

Localisation des modifications Les résultats de l'évaluation au niveau des pixels sont donnés dans la figure 5.18 pour les deux jeux de données. Le seuil de binarisation pour les prédictions THM est contrôlé par un paramètre σ, valeur entre 0 et 100, qui représentent un pourcentage de la valeur possible THM_{max}. Pour LFM, une valeur haute pour σ conduit à une zone prédite plus petite, à l'inverse des approches IRPSNR et SSIM qui auront des zones plus grandes.

LFM fonctionne légèrement mieux que SSIM sur **Re** et IRPSNR est toujours la pire des approches. D'autre part, l'évaluation basée sur la métrique évaluant les composantes connexes, reportée dans la table 5.5, donne une autre perspective.

TABLE 5.5 – Scores F1-mesure sur les jeux de données **WW** et **Re** avec la mesure basée sur les composantes connexes en utilisant la meilleure prédiction THM par image (tous seuils de binarisation confondus).

Dataset	LFM	IRPSNR	SSIM
Reddit	**0.84**	0.62	0.69
WW	0.21	**0.40**	0.21
All images	**0.63**	0.53	0.48

L'approche du *gKDE* utilisée par LFM pour transformer la liste d'*outliers* tend à produire des zones larges et homogènes. À l'inverse, IRPSNR et SSIM tendent à sur-segmenter, ce qui est pénalisé par les scores basés sur les composantes. Ce comportement est illustré dans la figure 5.19 même si la carte binaire peut être augmentée avec des opérations morphologiques. Ce post-traitement est sensible aux éléments

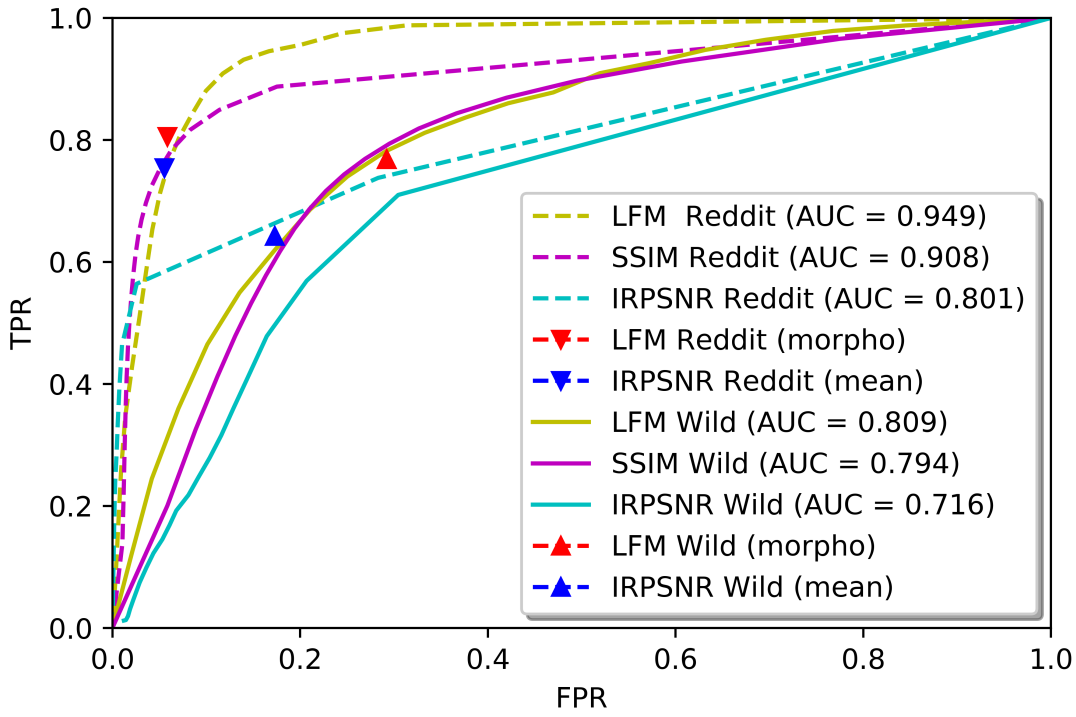


FIGURE 5.18 – Comparaison des courbes ROC générées par les prédictions THM, avec les scores au niveau des pixels sur les jeux de données **WW** and **Re**. Les triangles rouges et bleus correspondent aux scores des approches `LFM` morpho, and `IRPSNR` binarisé avec un seuil fixé par la moyenne des valeurs de la prédiction THM.

structurés utilisés et induit de nouvelles localisations erronées tout en corrigeant certains défauts.

Sensibilité à l'appariement Les approches étudiées fonctionnent relativement bien en dépendant grandement de la qualité de l'estimation de l'homographie H entre $\mathcal{R}$ et $\mathcal{C}$:

— `IRPSNR` et `SSIM` car ces approches reposent sur une comparaison pixel par pixel et donc sur la qualité du recadrement ;

— `LFM` car la prédiction THM est basée sur des valeurs aberrantes par rapport à H.

`IRPSNR` et `SSIM` ont pour particularité d'être des approches particulièrement prudentes lors de l'étape d'appariement, et en utilisant simplement une projection de $\mathcal{R}$ vers $\mathcal{C}$ comme dans [BROGAN et al. 2017] ce qui est souvent insuffisant lorsqu'il s'agit d'images recadrées de différentes manières.

(a) prédiction de LFM, prédiction de IRPSNR, vérité terrain

(b) prédiction de LFM, prédiction de SSIM, vérité terrain

FIGURE 5.19 – Deux exemples d'erreurs des méthodes IRPSNR et SSIM comparées avec la prédiction binaire de LFM. Les zones rouges correspondent aux prédictions et à la vérité terrain.

Les résultats sont très différents entre les jeux de données **WW** et **Re**. Deux cas d'échecs sont remarquées sur **WW**. Le premier cas est lorsque l'image candidate $\mathcal{C}$ est un peu différente de l'image requête $\mathcal{R}$ (*e.g.* même événement mais photo prise d'un autre point de vue ou à une autre période). L'appariement, et donc l'estimation de l'homographie, échouent généralement. La second cas est lorsque l'image $\mathcal{C}$ est l'image donneuse. C'est-à-dire l'image contenant l'élément (habituellement petit) utilisé pour la modification, et non l'image hôte (c'est-à-dire celle qui a reçu la modification, produisant l'image modifiée). Quand cela se produit, la partie comparable entre les images $\mathcal{R}$ et $\mathcal{C}$ correspond à l'élément de la modification, et la falsification sera détectée comme tout le reste de l'image. La prédiction étant cohérente par rapport aux images comparées, mais aura un masque inversé.

5.4.4 Analyse de la chaîne complète

La présentation précédente du module de détection de modifications est présenté comme étant indépendant du module de recherche d'images similaires car il ne nécessite qu'une paire d'images en entrée pour fonctionner. Cependant, il est intéressant d'analyser le fonctionnement des deux modules lorsqu'ils sont utilisés en chaîne et ainsi analyser les conséquences d'une erreur du premier module sur le second. Pour cela, nous testons les jeux de données **Re**, **Me**, **Ho**, **MICC$_{\text{F600}}$** et **Tw**. Le but étant de tester des images représentant les différents types de requête (*i.e.* requêtes positives et négatives), ainsi que images modifiées contenant des types d'attaques de différents types (*i.e.* duplication et insertion) et de différentes tailles.

Recherche d'images similaires

Bien que cette étape soit similaire aux expérimentations présentées dans la section 5.3, l'analyse de la chaîne complète nécessite d'étudier aussi la première étape. En effet, il est important de noter les images présentant des erreurs à ce niveau (faux positifs et faux négatifs) qui auront des répercussions directes sur les performances de la deuxième étape.

Deux cas généraux d'erreurs peuvent survenir :

1. des `faux positifs` surviennent lorsque le système retourne une image candidate mais cette dernière n'est pas celle attendue ou il s'agissait d'une requête négative ;
2. des `faux négatifs` correspondent à des images n'étant associées à aucune image candidate alors qu'une image dans la base est connue comme étant une bonne réponse.

À ce niveau, les jeux de données utilisés peuvent être regroupés en trois catégories :

1. les requêtes positives correspondant à une image modifiée : **Re**, **Me**, **MICC$^{\text{mod}}_{\text{F600}}$** et **Tw**. Ces jeux de données correspondent au coeur des travaux, le but étant ici de trouver les images originales associées aux requêtes. L'ensemble des requêtes permet de tester un large ensemble de types de modifications ;
2. les requête positives correspondant à une image originale : **MICC$^{\text{ori}}_{\text{F600}}$** ;

3. les requêtes négatives : **Ho**. L'ensemble des images diffusées sur les réseaux étant évolutif, il n'est pas envisageable de prétendre à un système de recherche d'images similaires travaillant sur une base d'images possédant obligatoirement une image similaire à notre image requête. Nous nous intéressons donc à tester des requêtes négatives (*i.e.* ne possédant pas d'images similaires dans notre base).

La Table 5.6 présente les performances du système lors cette première étape sur les différents jeux de données. Plusieurs situations peuvent expliquer les erreurs qui apparaissent ici : premièrement, les images issues de **Ho** étant associée à une image présentent souvent une forte ressemblance avec l'image candidate trouvée (*e.g.* les deux images représentent une forêt) ce qui peut expliquer à la fois la ressemblance des vecteurs de contenu et le calcul d'une homographie cohérente par association de descripteurs similaires. Concernant les faux négatifs des autres jeux de données, la présence de trop grandes modifications dans l'image requête $\mathcal{R}$ tend à baisser de trop forte manière le résultat du calcul de la similarité avec l'image $\mathcal{C}$ voulue.

Il est important de noter que classer une image comme n'ayant pas d'images comparables est très pénalisant car cette image est exclue de la chaîne de traitement. C'est pourquoi, il est préférable d'obtenir un faux positif qui sera détecté lors de la prochaine étape plutôt qu'un faux négatif (impossibilité de ce rendre compte de cette erreur par la suite).

Jeu de données	Nombre de requêtes	VP	FP	FN	VN
Re	255	237	1	16	0
$\mathbf{Me}_{\mathbf{req}}$	18	10	1	7	0
$\mathbf{MICC}^{\mathbf{mod}}_{\mathbf{F600}}$	160	160	0	0	0
$\mathbf{Tw}_{\mathbf{req}}$	23	20	0	3	0
$\mathbf{MICC}^{\mathbf{ori}}_{\mathbf{F600}}$	400	400	0	0	0
Ho	1491	0	0	0	1491

TABLE 5.6 – Performance du système de recherche d'images similaires

Localisation des modifications

Les images traitées ici correspondent aux `vrais positifs` et `faux positifs` de l'étape précédente, soit toutes les images pour lesquelles une image candidate a été

trouvée. Le but est alors de comparer les deux images pour trouver les différences et ainsi localiser les modifications apportées à l'image.

La Table 5.7 montre les performances de cette deuxième étape en utilisant les couples d'images (requête / candidat) construits par la recherche d'images comparables.

Nous remarquons 41 erreurs lors de cette étape : deux faux négatifs issus du jeu de données **Re** ; cinq du jeu de données $\mathbf{Me}_{\mathbf{req}}$, ainsi que trois faux négatifs et 31 faux positifs du jeu de données $\mathbf{MICC}_{\mathbf{F600}}$.

Il est à noter que pour toutes les comparaisons issues d'un `faux positif` issus de la première étape, l'homographie sera obligatoirement incohérente. Ces mauvaises homographies ont pour effet direct de rendre de nombreux *outliers* qui seront présent de manière globale dans toute l'image. La prédiction sera ainsi de très grande taille. Cela est directement vérifié avec les trois cas de `faux positifs` qui obtiennent toutes les trois des prédictions très larges en sortie de la deuxième étape. Cela permet de filtrer les faux positifs, produit par le premier module, à ce stade de l'analyse.

Jeu de données	Nombre de requêtes	VP	FP	FN	VN
Re	238	236	0	2	0
$\mathbf{Me}_{\mathbf{req}}$	11	6	0	5	0
$\mathbf{MICC}^{\mathbf{mod}}_{\mathbf{F600}}$	160	157	0	3	0
$\mathbf{Tw}_{\mathbf{req}}$	20	20	0	0	0
$\mathbf{MICC}^{\mathbf{ori}}_{\mathbf{F600}}$	400	0	31	0	369

TABLE 5.7 – Performance du système de localisation

5.5 Caractérisation des modifications

5.5.1 Représentation uniforme des patches

Comme mentionné précédemment, les images peuvent être modifiées de différentes façons. Certaines de ces modifications sont réalisées dans le but de tromper les personnes visualisant l'image (*e.g.* l'insertion d'un élément dans l'image) alors que d'autres modifications sont réalisées dans un but informatif (*e.g.* ajout d'une flèche pour attirer l'attention de la personne). Dans ce troisième module représenté sur la figure 5.20, notre but est de discriminer ces deux types de modifications et d'apporter une connaissance supplémentaire sur la nature des modifications.

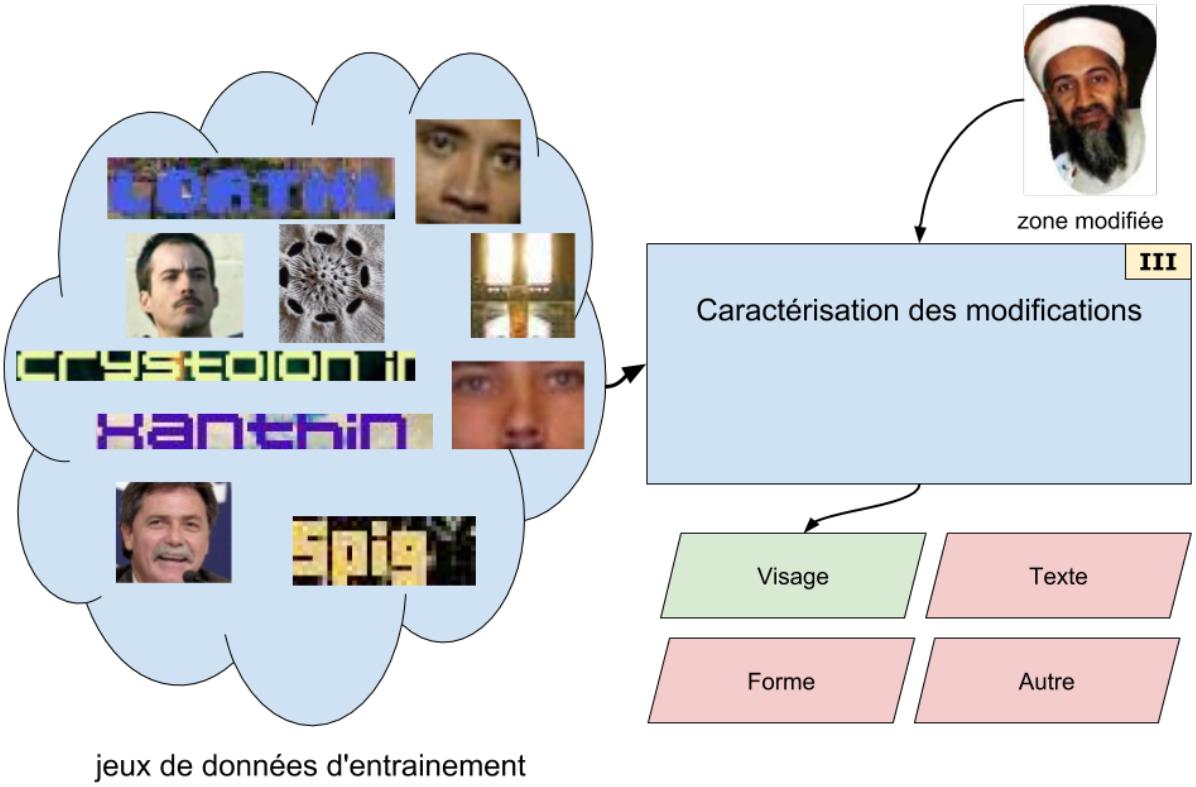


FIGURE 5.20 – Représentation de la troisième étape : Caractérisation des modifications

Étant donné un masque binaire qui peut être soit la prédiction réalisée précédemment, soit le masque de vérité terrain si nous souhaitons tester ce module sans prendre en compte les erreurs des modules précédents, nous extrayons les imagettes associées aux zones actives du masque.

Une image pouvant contenir plusieurs modifications, nous utilisons un algorithme de détection des composantes connexes afin de séparer les différentes modifications prédites et ainsi obtenir plusieurs imagettes dans le cas de plusieurs modifications dans une même image. Chaque modification dans l'image est ainsi traitée individuellement.

Cette étape de caractérisation est abordée comme un problème de classification d'images. Pour cela nous définissons quatre classes : *forme, texte, visage* et *autre*. Les deux premières sont définies comme n'étant pas mal intentionnées, contrairement aux deux autres classes. Une fois qu'une modification est détectée à l'étape précédente, nous définissons une zone à extraire, décrire puis classer.

La tâche étant abordée comme un problème de classification, il est nécessaire de constituer un ensemble d'apprentissage. Pour cela nous avons constitué un jeu de donnée avec 10 000 exemples pour chaque classe.

L'ensemble d'images associées au label *forme* est composé de différentes formes de tailles différentes artificiellement insérées dans des images sélectionnées de manière aléatoire dans le jeu de données `MIRFLICKR-1M` [HUISKES, THOMEE et LEW 2010]. L'ensemble d'images associées au label *texte* contient des instances provenant du jeu de données `COCO-Text` [VEIT et al. 2016]. Ce jeu propose des images de texte avec une écriture manuscrite ou générée par ordinateur dont des écritures considérées comme lisibles ou illisibles pour les deux catégories. D'autres images sont ajoutées pour compléter cet ensemble, qui contiennent du texte ajouté artificiellement dans des images avec des tailles, police d'écriture et couleur différentes. Comme pour l'ensemble précédent, les images dans lesquelles le texte est ajouté sont sélectionnées parmi le jeu de données `MIRFLICKR-1M`. L'ensemble d'images associées au label *visage* contient des instances sélectionnées aléatoirement parmi le jeu de données LFW [G. B. HUANG et al. 2007]. Enfin, l'ensemble d'images associées au label *autre* est composé de *patchs* de tailles différentes extraites d'images sélectionnées aléatoirement dans le jeu de données `MIRFLICKR-1M`.

Pour décrire les imagettes, nous utilisons le même principe de description que pour la recherche d'image similaires. C'est à dire une description basée sur la sortie d'une couche du réseau `VGG-19`. Nous souhaitons ici aussi déterminer la meilleure couche du réseau à utiliser pour décrire les imagettes. Il est nécessaire de réaliser de nouvelles expérimentations pour cela car il n'y a aucune indication quant à la performance de la même couche que lors de l'étape de recherche d'images similaires. Cependant, ces deux étapes partagent la nécessité de se baser sur le contenu des images. C'est pourquoi nous gardons la liste des trois couches $conv_{5\_4}$, fc_6 et fc_7. Ici aussi, il sera nécessaire de déterminer si il est préférable d'utiliser un *max-pooling* ou un *mean-pooling*. Il est à noter que cette tâche de caractérisation est plus proche de la tâche initiale `VGG-19` de description qui a pour but de détecter le contenu d'une l'image.

De la même manière que lors de l'étape de recherche d'images comparables (section 5.3), nous appliquons une normalisation sur l'ensemble des vecteurs de descriptions.

5.5.2 Expérimentations

Afin de tester les capacités de discrimination d'un classifieur entre les classes, nous utilisons une classification en validation croisée sur l'ensemble d'apprentissage. Les descripteurs des couches $conv_{5\_4}$, fc_6 et fc_7 sont normalisés soit par une normalisation ℓ_2 ou *power*. Dans le cas de la couche $conv_{5\_4}$, le tenseur obtenu étant de taille 7 il est nécessaire d'appliquer soit un *pooling* maximum ou moyen. Les résultats de ces expérimentations sont présentés dans la table 5.8. Le nombre de plis pour ces tests est fixé à 10.

		Taux de B.C	Autre	Visage	Forme	Texte
$conv_{5\_4}$ Mean	ℓ_2	98,47	96,84	99,50	**99,51**	98,29
$conv_{5\_4}$ Mean	power	98,44	96,75	99,48	99,22	**98,43**
$conv_{5\_4}$ Max	ℓ_2	97,96	96,36	99,22	98,98	97,56
$conv_{5\_4}$ Max	power	97,81	96,16	99,26	91,18	97,74
fc_6	ℓ_2	98,86	97,76	**99,63**	99,31	98,84
fc_6	power	98,86	**97,82**	99,59	99,19	98,89
fc_7	ℓ_2	98,42	96,81	99,44	99,08	98,45
fc_7	power	**99,38**	96,95	99,43	98,57	98,55

TABLE 5.8 – Résultats de la classification validation-croisée sur les descripteurs produits par les différentes couches de `VGG-19` et une normalisation ℓ_2 ou *power* sur les quatre jeux de données d'entraînement en terme de taux de bonne classification et de F1-score par classe.

Ces expérimentations ayant pour but de tester indépendamment l'étape de caractérisation, nous souhaitons ne pas être influencé par des erreurs provenant des étapes précédentes. C'est pourquoi les imagettes sont identifiées et extraites à partir des masques de vérité terrain pour éviter les erreurs de détection et ainsi avoir les imagettes les plus précises possible.

Les résultats de la classification à partir des imagettes issues des vérités terrain sont présentés dans la Table 5.9. La majorité des erreurs proviennent de confusions entre les classes `Texte` et `Forme` d'une part et `Visage` et `Autre` d'autre part. Cela n'est pas réellement surprenant étant donné la proximité de ces deux classes. Le score `forme` ne peut pas être calculé pour le jeu de données **WW**, aucune imagette de ce label n'étant présente.

Maintenant, nous nous intéressons à une classification binaire : La modification est caractérisée comme mal intentionné si elle classé comme `Visage` ou `Autre`, non mal

intentionné sinon. Nous remarquons alors que cette prise en compte de cette nouvelle règle de classification aide grandement à améliorer les résultats, notamment pour les classes `Texte` et `Visage`. La prédiction binaire permet un taux de bonne classification de 91.1% pour **Re**, 71.4% pour **Me** et 77.9% pour **WW**.

	Taux de B.C	Autre	Visage	Texte	Forme
Re	68.95	77.47	47.37	66.67	40.00
WW	68.38	75.43	48.00	64.62	NaN
Me	61.22	63.83	40.00	70.97	40.00

TABLE 5.9 – Résultats de l'étape de caractérisation de la modification sur les trois jeux de données **Re**, **WW** et **Me** en terme de taux B.C. et de F1-score par classe. Les résultats sont présentés à partir de la vraie zone modifiée (`Vérité terrain`)

5.6 Conclusion

Dans ce chapitre, un système d'analyse d'images est présenté [MAIGROT, KIJAK et CLAVEAU 2018 ; MAIGROT, KIJAK, SICRE et al. 2017]. Le but de ce dernier est de déterminer automatiquement si une image est modifiée ou non. Pour cela, le système est composé de trois parties successives : 1) recherche d'une image similaire ; 2) comparaison des deux images ; 3) analyse des différences.

Le système présenté est capable d'une part de se prononcer de manière binaire quant à la présence probable d'une ou plusieurs modifications et d'autre part de trouver et localiser les modifications dans une image.

Lors de la première étape, le système met en place une recherche d'images similaires étant donné une image requête. Lors des expérimentations présentées dans ce chapitre une base d'image est utilisée, mais une telle structure pourrait être utilisé sur des moteurs de recherche d'images tels que *Google Image*. Le système permettant la recherche d'une image similaire est en deux temps : une recherche par descripteurs globaux et un filtrage spatial en utilisant des descripteurs locaux. Chacune de ces recherches attribut un score à chaque image candidate ce qui permet de retrouver la meilleure image pour réaliser une comparaison entre l'image requête et cette image. Cependant, le système montre aussi plusieurs limites :

— Le système étant basé sur une comparaison d'images, il est nécessaire d'arriver à trouver une image permettant la comparaison par rapport à l'image requête.

(a) image modifiée (b) image cible (c) image donneuse (d) vérité terrain

FIGURE 5.21 – Exemple d'une image requête (gauche) avec les deux images ayant été utilisée pour la produire (centre) ainsi que la vérité terrain de la modification (droite).

Dans le cas d'une fausse information récurrente dans le temps, le système fonctionne parfaitement, retrouvant ainsi les anciennes versions publiées. Si un événement nouveau arrive sur les réseaux sociaux par contre, les bases d'images connues telles que *Google* ou *Hoaxbuster* ne seront d'aucune aide ;

— Les images requêtes ne sont pas forcément des versions modifiées. Dans le cas, d'une image originale mise en entrée du système, un problème peut survenir si il existe une version modifiée de cette image. Le système va ainsi comparer les deux images et relever des différences. La zone modifiée dans l'image retrouvée sera correctement détectée mais dans l'image originale (*i.e.* non modifiée). Une solution à ce problème serait de présenter les deux images à l'utilisateur avec la zone annotée sur les deux images qui décidera d'attribuer la mention *modifiée* à l'une des images, la détection de l'image modifiée entre les deux versions n'étant pas traitée ici ;

— Une erreur similaire peut se produire lorsque la modification prend une grande part de l'image modifiée. Le système peut alors retrouver non pas l'image cible de la modification, mais celle contenant l'élément ayant servi à la modification. Dans ce cas là, la zone prédite comme modifiée correspond globalement à la prédiction inverse que celle attendue. Ce cas est illustré dans la figure 5.21. Dans le cas où les deux images originales (cible et donneuse) sont dans notre base d'images, l'image la plus similaire sera déterminée selon la taille de la modification.

La deuxième étape, qui prend en compte un couple d'images, a pour but de per-

mettre la comparaison de ces deux images. Ce système a pour avantage de pouvoir traiter n'importe quelle paire d'image en ajustant les deux images pour maximiser les capacités de comparaison (*e.g.* détecter le recadrage d'une image et recadrer la seconde de la même manière) et surtout peut prétendre à détecter une multitude de types d'attaques différentes tout en étant robuste à plusieurs transformation (*e.g.* translation, rotation, ...). Cette étape aussi comporte plusieurs limites :

— Les images floues ne permettent pas d'obtenir des descripteurs fiables et engendrent une mauvaise estimation de l'homographie. Cela a pour effet de détecter un très grand nombre d'*outliers*. Lorsque cela se produit, le système prédit une très large proportion de l'image comme étant modifiée ;

— Si l'image contient plusieurs modifications et que plusieurs de ces dernières sont proches, le système a tendance à les confondre en une seule grande zone lors de la mise sous forme binaire de la prédiction.

— Enfin, une modification trop petite sera possiblement considéré comme du bruit et sera supprimée par le système au moment de l'analyse par le deuxième module. Cela peut notamment se produire lorsqu'une autre modification, beaucoup grande que la première, est détectée dans l'image.

CHAPITRE 6

CONCLUSION

Contents

Une conclusion générale du manuscrit est présentée dans ce dernier chapitre. La section 6.1 résume les contributions réalisées dans cette thèse et une discussion quant à ces dernières. La section 6.2 propose des perspectives de travail pour chaque partie présentées précédemment dans ce manuscrit.

6.1 Synthèse des travaux et discussion des contributions

Dans ce manuscrit, nous avons présenté les travaux réalisés durant les trois années consacrées à la détection de fausses informations dans les réseaux sociaux. Les fausses informations ont toujours existé dans les médias, mais depuis l'apparition des réseaux sociaux leur diffusion est accrue de part le pseudo-journalisme (n'importe qui peut utiliser son smartphone pour partager une actualité) et la capacité élevée des réseaux sociaux à partager rapidement une actualité. Cette rapidité de diffusion des fausses informations dans les réseaux sociaux rend obligatoire une vérification de ces dernières de manière le plus automatique possible sous peine de voir la fausse information trop partagée pour être contrée. De plus, il est important de prendre qu'un système répondant seulement *vrai* ou *faux* sera plus difficilement crédible pour un utilisateur qu'un système qui justifie son choix. Il est donc important de prévoir une méthode qui permet d'expliquer le résultat final à l'utilisateur.

L'intérêt publique se ressent aussi au niveau de l'attrait qu'a pu avoir la thèse auprès du grand public et des journalistes. Nous avons eu l'occasion de parler de la détection de fausses informations à de nombreux personnes expertes d'un domaine lié à ces travaux (traitement automatique des langues, traitement de l'image, analyse des réseaux sociaux, etc.) ou non. De plus, nous avons eu été amenés à discuter avec des journalistes et présenter des aspects plus ou moins spécifiques des travaux menés. Ainsi, des reportages ou articles ont été publiés par l'université de Rennes 1 [1], le CNRS [2] [3],

1. https://dossiers.univ-rennes1.fr/index.php/longform/la-cyber/index.htm
2. https://lejournal.cnrs.fr/articles/des-algorithmes-contre-les-images-truquees
3. http://www2.cnrs.fr/presse/communique/5599.htm

le média WebPatron [4], Le Temps [5], Le Monde [6] [7], Sciences et Avenir [8], LCI [9] et l'Espace des Sciences de Rennes [10].

Comme défini dans le chapitre 1, les fausses informations se présentent sous de nombreuses formes sur les réseaux sociaux que ce soit au niveau des intentions de l'auteur que de sa constitution. Toutes les informations disponibles dans une publication sont regroupées dans cinq catégories : le texte, le contenu multimédia, les informations sociales, le cheminement de la publication et l'événement qui lui est associé. Cependant, toutes ces catégories ne sont pas toujours présentes pour toutes les publications à traiter.

Ces différentes informations permettent une grande variété d'approches possibles comme présenté dans le chapitre 2. C'est pourquoi il a été indispensable de se focaliser sur un aspect plus précis. Nous nous sommes orientés plus spécifiquement sur les publications possédant une image et un texte.

6.1.1 Discrimination de médias traditionnels et de réinformation

Après une étude préliminaire de l'état de l'art, nous avons proposé une approche basée sur le contenu textuel de publications issues de *Facebook* et motivée par les descripteurs déjà existants pour des tâches similaires.

Les publications étudiées proviennent de deux types de médias : les médias traditionnels et les médias de réinformation. Les médias *traditionnels* correspondent aux médias connus du grand public et correspondant à une structure journalistique professionnelle. Les médias *de réinformation* correspondent aux groupes de personnes n'étant pas des professionnels de l'information et souhaitant promouvoir une vision des faits en opposition à celle présentée par les médias traditionnels.

4. https://bit.ly/2RhLsPI
5. https://www.letemps.ch/sciences/twitter-mensonge-se-diffuse-plus-vite-plus-loin-verite
6. https://www.lemonde.fr/sciences/video/2018/01/24/comment-la-science-aide-a-reperer-les-fake-news_5246356_1650684.html
7. https://www.lemonde.fr/pixels/article/2018/05/24/loi-sur-les-fausses-informations-les-chercheurs-du-cnrs-sceptiques_5303791_4408996.html
8. https://www.sciencesetavenir.fr/high-tech/informatique/identifier-les-fake-news-et-les-images-truquees-a-l-aide-du-machine-learning-c-est-possible_124289
9. https://bit.ly/2M5EAyZ
10. https://www.espace-sciences.org/sciences-ouest/363/actualite/les-chasseurs-de-fake-news

L'analyse de ces publications s'est basée sur deux types de descripteurs. Les descripteurs de surface décrivent la forme du message (*e.g.* nombre de mots dans le message). Les descripteurs de contenu représentent le fond du message par la description directe des mots utilisés.

L'étude a été réalisée sur une analyse avec chaque type de descripteur, puis l'unification de tous les descripteurs disponibles. Les résultats ont permis de montrer les capacités des deux types de descripteurs à discriminer ces deux types de médias. Les descripteurs de contenu permettent cependant une augmentation des résultats par rapport aux descripteurs de surface, la combinaison des deux types de descripteurs permettant une très légère augmentation des résultats.

Plusieurs pistes d'améliorations restent possibles quant à cette approche. Premièrement, l'analyse des publications est réalisée ici exclusivement sur le contenu textuel des publications. Certaines publications sont courtes et ne permettent pas de produire des descripteurs de contenu représentatifs. Pour éliminer cette limite de l'approche, il est possible de considérer le contenu des liens dans les messages comme étant la suite du texte. Deux conséquences sont à prévoir :

1. Le contenu des liens étant dans leur grande majorité plus long que le contenu textuel des messages, cela aura pour effet de créer une grande différence de taille de messages entre les publications avec ou sans source ;
2. La source citée dans la publication n'est pas obligatoirement un prolongement de la publication. Il est possible que la publication cite un autre article en souhaitant le contredire. Ajouter le contenu textuel de la source dans cette situation inversera le sens de la publication ;
3. Ajouter le contenu des liens sort du contexte initial qui est un utilisateur regardant une publication d'une source inconnue et devant décider en ne se basant que sur cette publication si il peut lui faire confiance ou non.

6.1.2 Analyse des différentes modalités d'une publication

Lors de l'édition 2016 de l'atelier *MediaEval*, nous avons participé à la tâche *Verifying Multimedia Use* de l'atelier *MediaEval*. Cette tâche avait pour but de classer des tweets selon leur véracité (*vrai* ou *faux*). Cela nous a permis de mettre en place trois approches basées sur l'analyse de trois modalités que sont le texte, l'image et les

sources citées. Une fusion est ensuite proposée à partir de la prédiction de ces trois approches.

Ce choix de traiter les modalités séparément permet de proposer une justification de la classification à l'utilisateur, la prédiction par modalité étant accessible par le système. Les approches basées sur le texte et les sources ont permis d'obtenir des scores de performance de l'ordre de 90% de F-Mesure. Quant à l'approche image, les résultats ont été plus pauvres que prévu, cela s'explique par un biais corrigé dans l'approche proposé dans le chapitre 5. Le biais est expliqué dans la section suivante de cette conclusion générale.

Une fois le défi terminé, nous avons eu accès aux prédictions des trois autres équipes participantes à la tâche. Cela nous a permis de tester et proposer de nouvelles fusions en nous basant sur toutes les prédictions. Ce nouveau système peut être vu comme un seul grand système prenant en entrée une publication. Cette dernière est passée dans tous les systèmes en parallèle. Une fois toutes les prédictions réalisées, l'ensemble des prédictions est envoyé dans la dernière partie du système chargée de réaliser la prédiction finale.

Cette nouvelle fusion des connaissances, basée sur le principe que chaque approche possède une capacité à prédire un certain type de fausse information, a pour ambition de tirer le meilleur de chaque approche. On note alors une augmentation des résultats par rapport à la meilleure des approches proposées lors de la tâche.

Une limite à cette méthode est la perte de justification à l'utilisateur quant à la classification réalisée par le système. Nous avons donc aussi une fusion en deux temps où le premier niveau vise à unifier les prédictions basées sur une même modalité (texte, image et source), puis un second niveau pour obtenir une prédiction finale.

6.1.3 Détection de modification dans une image

Suite à notre participation à la tâche *Verifying Multimedia Use* de l'atelier *Mediaeval*, nous avons remarqué un biais dans l'approche image. L'approche proposée se basait simplement sur la recherche d'une image requête dans une base d'image connues et annotées comme étant *originale* ou *modifiée*. Le label de l'image la plus similaire retrouvée était alors propagé à l'image requête.

Cependant, nous étions confronté à un cas d'erreur relativement fréquent où l'image requête et l'image retrouvée correspondent à des quasi-copies, mais ne sont

pas identiques. Le cas le plus défavorable étant celui où l'image requête est une version modifiée de l'image retrouvée (ou inversement). C'est pourquoi nous avons proposé une amélioration de cette approche image dans laquelle l'image retrouvée est comparée à l'image requête.

Bien que possédant plusieurs défauts, cette approche présente l'avantage de ne nécessiter aucune information supplémentaire sur l'image, ni de se restreindre à un type de modification ou à un format particulier d'image.

Cette approche possède cependant plusieurs limites qui reposent principalement sur deux contraintes :

1. Cette approche nécessite de posséder une quasi-copie de l'image requête dans la base d'images connues. Cette base peut être grandement augmentée en utilisant des moteurs de recherche par image tels que *Google Image*[11] ou *TinEye*[12], mais il sera nécessaire de filtrer les images identiques retournées puisque ces services retournent en premier les images identiques à la requête (fonction première de ces services). De plus, si l'image vient d'apparaitre sur les réseaux sociaux ou internet, aucune image ne sera retrouvée (identique ou quasi-copie) ;

2. La qualité de la comparaison dépend directement de la qualité de l'homographie estimée. Plusieurs paramètres peuvent perturber le calcul de l'homographie (*e.g.* application d'un flou sur une des deux images).

D'autres limites, plus mineures ont été remarquées :

1. L'extraction dense des points d'intérêts ne permet pas de détecter correctement les petites modifications. Cela est lié au pas d'échantillonage des points d'intérêts de 10 pixels dans nos expérimentations. Cette valeur pourrait être réduite pour permettre une localisation plus précise des modifications au prix d'un temps de calcul beaucoup plus grand ce qui rend impossible son utilisation par utilisateur en temps réel ;

2. D'un point de vue purement expérimental, les résultats obtenus dépendent pleinement de la manière dont ont été réalisés les masques de vérité terrain.

11. https ://images.google.com/
12. https ://www.tineye.com/

6.2 Perspectives pour les travaux futurs

Plusieurs pistes n'ont pas été traitées durant cette thèse ou l'ont été partiellement. Ces différentes pistes représentent des évolutions possibles .

6.2.1 Étude multimodale

Les publications issues des réseaux sociaux présentent plusieurs modalités possibles (voir définition générique des publications dans le chapitre 2). C'est pourquoi une étude des différentes modalités d'une publication a été réalisée lors de cette thèse (chapitre 4). Cependant, cela ne correspond pas pleinement à une analyse multimodale de la publication.

Comme défini dans la mise en contexte en début de manuscrit, certaines fausses informations, nommées détournement d'information, correspondent à une mauvaise association d'une image et un texte représentant deux contextes différents. La détection de ce type de fausses informations nécessite l'étude simultanée des deux modalités.

Une étude a été commencé en partant du système proposé par [JIN, CAO, GUO et al. 2017] (figure 6.1). Par manque de temps, ces expérimentations sont encore en cours, mais présentent un très bon potentiel quant aux possibilités de traitement du texte et de l'image en un seul système. Le système proposé possède cependant plusieurs limites. Premièrement, ce système ne permet pas d'expliquer la classification finale. Ce problème est un des points faibles de l'utilisation des réseaux de neurones. De plus, le système en l'état ne peut traiter la détection de détournement d'information. En effet, la grande majorité des détournements d'informations nécessite un apport d'information supplémentaire au tweet puisqu'il est impossible de se rendre compte d'un détournement d'information en ne prenant en compte que le texte et l'image d'une publication.

La suite de cette section présente plusieurs possibilités imaginées durant cette thèse.

Recherche d'une réutilisation de l'image

Une première approche possible est de se baser sur le fait qu'une image détournée possède obligatoirement une utilisation originale. En recherchant les autres utilisations

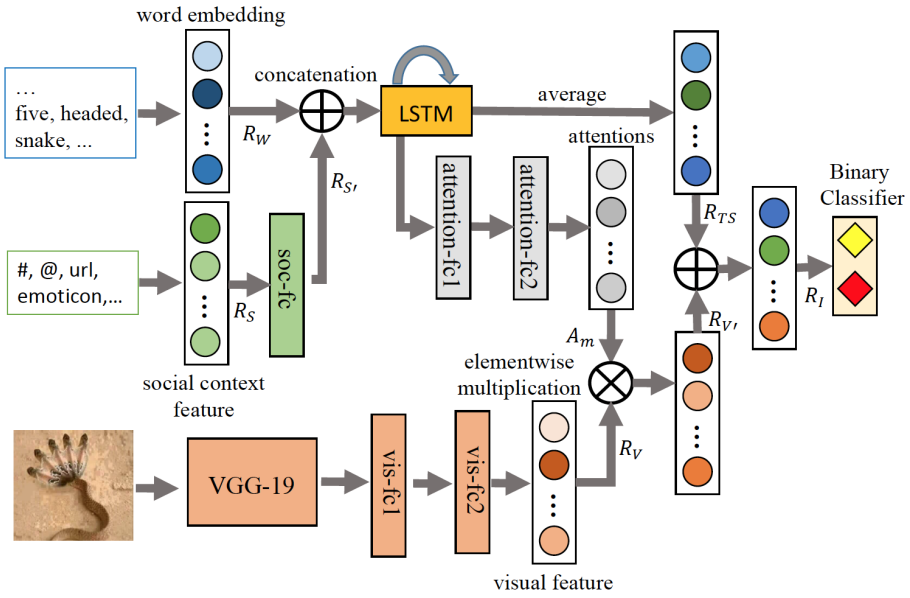


FIGURE 6.1 – Schéma de l'approche proposée par [JIN, CAO, GUO et al. 2017]

d'une image requête à l'aide de moteurs de recherche tels que *Google Image*, nous obtenons en retour les articles ou pages utilisant cette même image. Un exemple d'image détournée est donné dans la figure 6.2. Ce tweet semble montrer Paris le lendemain des attentats du 13 novembre 2015. Or, en recherchant l'image sur *Google Image*, on trouve l'origine de l'image qui date de 2008 [13].

Une fois les autres contextes d'utilisations identifiés, il est nécessaire d'estimer si ces contextes sont identiques à celui analysé. Une solution possible est de comparer le texte de la publication analysée et celui de la page retournée par le moteur de recherche.

Si les contextes sont différents, il existe alors un détournement d'information soit par la publication analysée soit par la page retournée par le moteur de recherche. Afin de distinguer ces deux cas, il est possible soit de comparer les dates de publications soit de compter parmi toutes les pages retournées par le moteur de recherche la proposition de page partageant un contexte similaire avec celui de la publication.

13. http://blog.grainedephotographe.com/silent-world-paris-new-york-plus-vide-que-jamais/

FIGURE 6.2 – Exemple d'une image détournée de son contexte originale

Comparaison des modalités

Une seconde approche serait de comparer sémantiquement les deux modalités. Concernant le texte, cela peut se représenter par une extraction des entités nommées et/ou des termes les plus représentatifs [DROUIN 2003]. Concernant l'image, nous pouvons envisager l'utilisation des systèmes de recherche d'objets et de personnes dans une image qui est une tâche très active [LAVI, SERJ et ULLAH 2018].

En comparant les contextes sémantiques, il serait possible de trouver les cas où une photo représente une mauvaise personne (par exemple, un texte évoquant un fait à propos de Angela Merkel et une photo montrant une autre personne).

6.2.2 Fact Checking

La détection de fausses informations demande des connaissances externes dès que nous voulons mettre en place une analyse avancée de la publication à traiter. Une

des formes les plus avancées d'analyse des informations est la vérification de faits. Cette problématique, considérée comme une tâche en elle-même de part sa complexité, est un domaine de recherche très actif. Cependant, il est difficile d'automatiser pleinement un système de vérification de faits comme le montre [GRAVES 2018]. Les auteurs de ces travaux évoquent pleinement le fait que les systèmes de vérification automatique de faits sont, pour le moment, particulièrement performants lorsqu'il est question d'assister un professionnel, mais beaucoup moins lorsque le système doit prendre une décision par lui-même. Cela est principalement dû à la nécessité d'un jugement et de sensibilité au contexte qu'il n'est actuellement pas possible d'intégrer à un système de vérification de faits entièrement automatisé. Cependant des travaux, comme le projet *Content Check*, s'intéressent à construire des outils de traitement pour faciliter la mise en contexte de sujets abordés dans des articles ou sur les réseaux sociaux [CAZALENS et al. 2018]. Il serait possible de s'en inspirer pour détecter les fausses informations.

6.2.3 Cohérence du résultat de l'approche image

Un grand avantage de l'approche image au niveau de l'acceptabilité du résultat par l'utilisateur est la localisation des modifications. En montrant à l'utilisateur la ou les zones détectées comme étant modifiées, l'utilisateur est plus à même de croire l'analyse automatique. Cependant, deux limites de l'approche posent un problème au niveau de l'acceptabilité du résultat par l'utilisateur.

Premièrement, la détection du contenu d'une modification par le troisième module (voir chapitre 5) évoque brièvement le fait qu'une modification puisse rendre l'image fausse ou non. Un exemple simple d'une modification ne rendant pas l'image fausse est l'ajout d'une flèche. Cet aspect de rendre une image fausse ou non est à approfondir pour ne pas lancer inutilement des alertes auprès de l'utilisateur.

Deuxièmement, l'incapacité du système à différencier l'image originale et l'image modifiée. Si l'image requête $\mathcal{R}$ correspond à une image non modifiée et l'image $\mathcal{C}$, l'image la plus similaire trouvée dans la base d'images, une version modifiée de $\mathcal{R}$, l'approche proposée va comparer ces deux images et compter la zone modifiée dans $\mathcal{C}$ comme étant modifiée dans $\mathcal{R}$.

Sources primaires

MAIGROT, Cédric, Vincent CLAVEAU et Ewa KIJAK (2017), « Détection de fausses informations dans les réseaux sociaux : vers des approches multi-modales », in : *Extraction et Gestion des Connaissances (EGC)*.

— (2018), « Fusion-based multimodal detection of hoaxes in social networks », in : *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, IEEE, page(s): 222-229.

MAIGROT, Cédric, Vincent CLAVEAU, Ewa KIJAK et Ronan SICRE (2016), « MediaEval 2016 : A Multimodal System for the Verifying Multimedia Use Task », in : *MediaEval Workshop*.

MAIGROT, Cédric, Ewa KIJAK et Vincent CLAVEAU (2016), « Médias traditionnels, médias sociaux : caractériser la réinformation », in : *23ème Conférence sur le Traitement Automatique des Langues Naturelles*.

— (2017), « Détection de fausses informations dans les réseaux sociaux : l'utilité des fusions de connaissances », in : *COnférence Recherche d'Information et Applications*.

MAIGROT, Cédric, Ewa KIJAK, Ronan SICRE et al. (2017), « Tampering detection and localization in images from social networks : A CBIR approach », in : *International Conference on Image Analysis and Processing*, Springer, page(s): 750-761.

BIBLIOGRAPHIE

AMERINI, Irene, Lamberto BALLAN, Roberto CALDELLI, Alberto Del BIMBO et al. (2013), « Copy-move forgery detection and localization by means of robust clustering with J-Linkage », in : *Signal Processing : Image Communication (SPIC)*.

AMERINI, Irene, Lamberto BALLAN, Roberto CALDELLI, Alberto DEL BIMBO et al. (2010), « Geometric tampering estimation by means of a SIFT-based forensic analysis », in : *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, TX, USA.

— (2011), « A SIFT-based forensic method for copy-move attack detection and transformation recovery », in : *TIFS*.

BABENKO, Artem et al. (2014), « Neural Codes for Image Retrieval », in : *European Conf. on Computer Vision (ECCV)*.

BAY, Herbert, Tinne TUYTELAARS et Luc VAN GOOL (2006), « Surf : Speeded up robust features », in : *Computer vision–ECCV 2006*, page(s): 404-417.

BENTLEY, Jon Louis (1975), « Multidimensional binary search trees used for associative searching », in : *Communications of the ACM* 18.*9*, page(s): 509-517.

BIANCHI, Tiziano et Alessandro PIVA (2012), « Image forgery localization via block-grained analysis of JPEG artifacts », in : *IEEE Transactions on Information Forensics and Security*.

BIRAJDAR, Gajanan K. et Vijay H. MANKAR (2013), « Digital image forgery detection using passive techniques : A survey », in : *Digital Investigation* 10.*3*, page(s): 226-245.

BOIDIDOU, Christina, Katerina ANDREADOU et al. (2015), « Verifying Multimedia Use at MediaEval 2015 », in : *Proceedings of the MediaEval 2015 Workshop*.

BOIDIDOU, Christina, Stuart MIDDLETON et al. (2016), « The VMU Participation @ Verifying Multimedia Use 2016 », in : *MediaEval 2016 Workshop*.

BOIDIDOU, Christina, Symeon PAPADOPOULOS, Duc-Tien DANG-NGUYEN, Giulia BOATO et Yiannis KOMPATSIARIS (2015), « The CERTH-UNITN Participation@ Verifying Multimedia Use 2015. », in : *MediaEval*.

BOIDIDOU, Christina, Symeon PAPADOPOULOS, Duc-Tien DANG-NGUYEN, Giulia BOATO, Michael RIEGLER et al. (2016), « Verifying multimedia use at MediaEval 2016 », in : *MediaEval 2016 Workshop*.

BREIMAN, Leo et al. (1984), *Classification and Regression Trees*, Statistics/Probability Series, Belmont, California, U.S.A. : Wadsworth Publishing Company.

BROGAN, Joel et al. (2017), « Spotting the Difference : Context Retrieval and Analysis for Improved Forgery Detection and Localization », in : *arXiv preprint arXiv :1705.00604*.

CAO, Juan et al. (2016), « MCG-ICT at MediaEval 2016 : Verifying Tweets From Both Text and Visual Content », in : *MediaEval 2016 Workshop*.

CASTILLO, Carlos, Marcelo MENDOZA et Barbara POBLETE (2011), « Information credibility on twitter », in : *Proceedings of the 20th international conference on World wide web*, ACM, page(s): 675-684.

CAZALENS, Sylvie et al. (2018), « Computational fact checking : a content management perspective », in : *Proceedings of the VLDB Endowment* 11.*12*, page(s): 2110-2113.

CHATFIELD, Ken et al. (2011), « The devil is in the details : an evaluation of recent feature encoding methods. », in : *BMVC*, t. 2, 4, page(s): 8.

CHEN, Tong et al. (2017), « Call Attention to Rumors : Deep Attention Based Recurrent Neural Networks for Early Rumor Detection », in : *arXiv preprint arXiv :1704.05973*.

CHIERCHIA, Giovanni et al. (2014), « A Bayesian-MRF approach for PRNU-based image forgery detection », in : *IEEE Transactions on Information Forensics and Security* 9.*4*, page(s): 554-567.

CHRISTLEIN, Vincent, Christian RIESS et Elli ANGELOPOULOU (2010), « On rotation invariance in copy-move forgery detection », in : *Workshop on Information Forensics and Security*.

CLAVEAU, Vincent et Christian RAYMOND (2017), « Participation de l'IRISA à DeFT2017 : systèmes de classification de complexité croissante », in : *Actes de l'atelier Défi Fouille de texte, DeFT2017*.

COZZOLINO, Davide, Francesco MARRA et al. (2017), « PRNU-Based Forgery Localization in a Blind Scenario », in : *International Conference on Image Analysis and Processing*, Springer, page(s): 569-579.

COZZOLINO, Davide et Luisa VERDOLIVA (2016), « Single-image splicing localization through autoencoder-based anomaly detection », in : *Information Forensics and Security (WIFS), 2016 IEEE International Workshop on*, IEEE, page(s): 1-6.

DAVIES, Mark et Joseph L. FLEISS (1982), « Measuring agreement for multinomial data », in : *Biometrics*, page(s): 1047-1051.

DROUIN, Patrick (2003), « Term extraction using non-technical corpora as a point of leverage », in : *Terminology* 9.*1*, page(s): 99-115.

EVERINGHAM, Mark et al. (2010), « The pascal visual object classes (voc) challenge », in : *International journal of computer vision* 88.*2*, page(s): 303-338.

FARID, Hany (2009), « Exposing Digital Forgeries From JPEG Ghosts », in : *IEEE Transactions on Information Forensics and Security* 4.*1*, page(s): 154-160.

FISCHLER, Martin A et Robert C BOLLES (1981), « Random sample consensus : a paradigm for model fitting with applications to image analysis and automated cartography », in : *Communications of the ACM* 24.*6*, page(s): 381-395.

FLEISS, Joseph L et Jacob COHEN (1973), « The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability », in : *Educational and psychological measurement* 33.*3*, page(s): 613-619.

GOLJAN, Miroslav, Jessica FRIDRICH et Mo CHEN (2011), « Defending against fingerprint-copy attack in sensor-based camera identification », in : *IEEE Transactions on Information Forensics and Security* 6.*1*, page(s): 227-236.

GRAVES, Lucas (2018), *FACTSHEET : Understanding the Promise and Limits of Automated Fact-Checking*, rapp. tech., Tech. Rep.). Reuters Institute for the Study of Journalism, University of Oxford.

GUPTA, Manish, Peixiang ZHAO et Jiawei HAN (2012), « Evaluating Event Credibility on Twitter », in : *2012 SIAM International Conference on Data Mining*.

HALL, Mark et al. (2009), « The WEKA data mining software : an update », in : *ACM SIGKDD explorations newsletter* 11.*1*, page(s): 10-18.

HASHMI, Mohammad Farukh, Vijay ANAND et Avinash G KESKAR (2014), « A copy-move image forgery detection based on speeded up robust feature transform and Wavelet Transforms », in : *Computer and Communication Technology (ICCCT), 2014 International Conference on*, IEEE, page(s): 147-152.

HSU, Yu-Feng et Shih-Fu CHANG (2006), « Detecting image splicing using geometry invariants and camera characteristics consistency », in : *Multimedia and Expo, 2006 IEEE International Conference on*, IEEE, page(s): 549-552.

HUANG, Gary B. et al. (2007), *Labeled Faces in the Wild : A Database for Studying Face Recognition in Unconstrained Environments*, rapp. tech. 07-49, Univ. of Massachusetts.

HUISKES, Mark J., Bart THOMEE et Michael S. LEW (2010), « New Trends and Ideas in Visual Concept Detection : The MIR Flickr Retrieval Evaluation Initiative », in : *MIR '10 : Proceedings of the ACM International Conference on Multimedia Information Retrieval*, page(s): 527-536.

JEGOU, Herve, Matthijs DOUZE et Cordelia SCHMID (2008), « Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search », in : *European Conf. on Computer Vision (ECCV)*.

JIN, Zhiwei, Juan CAO, Han GUO et al. (2017), « Multimodal fusion with recurrent neural networks for rumor detection on microblogs », in : *Proceedings of the 2017 ACM on Multimedia Conference*, ACM, page(s): 795-816.

JIN, Zhiwei, Juan CAO, Yu-Gang JIANG et al. (2014), « News credibility evaluation on microblog with a hierarchical propagation model », in : *Data Mining (ICDM), 2014 IEEE International Conference on*, IEEE, page(s): 230-239.

JIN, Zhiwei, Juan CAO, Yazi ZHANG et al. (2015), « MCG-ICT at MediaEval 2015 : Verifying Multimedia Use with a Two-Level Classification Model. », in : *MediaEval*.

JIN, Zhiwei, Juan CAO, Yongdong ZHANG et al. (2016), « News Verification by Exploiting Conflicting Social Viewpoints in Microblogs. », in : *AAAI*, page(s): 2972-2978.

JOHNSON, Micah K et Hany FARID (2006), « Exposing digital forgeries through chromatic aberration », in : *Proceedings of the 8th workshop on Multimedia and security*, ACM, page(s): 48-55.

KAUR, Amanpreet et Richa SHARMA (2013), « Copy-move forgery detection using DCT and SIFT », in : *International Journal of Computer Applications* 70.*7*.

KRIPPENDORF, Klaus (1980), *Content Analysis : An Introduction to its Methodology*, Sage Publications.

KRIZHEVSKY, Alex, Ilya SUTSKEVER et Geoffrey E. HINTON (2012), « ImageNet Classification with Deep Convolutional Neural Networks », in : *Neural Information Processing Systems 25*.

KWON, Sejeong, Meeyoung CHA et Kyomin JUNG (2017), « Rumor detection over varying time windows », in : *PloS one* 12.*1*, page(s): e0168344.

KWON, Sejeong, Meeyoung CHA, Kyomin JUNG, Wei CHEN et al. (2013), « Prominent features of rumor propagation in online social media », in : *International Conference on Data Mining*, IEEE.

LANDIS, J Richard et Gary G KOCH (1977), « The measurement of observer agreement for categorical data », in : *biometrics*, page(s): 159-174.

LAVI, Bahram, Mehdi Fatan SERJ et Ihsan ULLAH (2018), « Survey on Deep Learning Techniques for Person Re-Identification Task », in : *arXiv preprint arXiv :1807.05284*.

LI, Bin et al. (2017), « A multi-branch convolutional neural network for detecting double JPEG compression », in : *arXiv preprint arXiv :1710.05477*.

LI, Jixian et al. (2018), « Double JPEG compression detection based on block statistics », in : *Multimedia Tools and Applications*, page(s): 1-16.

LI, Weihai, Yuan YUAN et Nenghai YU (2009), « Passive detection of doctored JPEG image via block artifact grid extraction », in : *Signal Processing* 89.*9*, page(s): 1821-1829.

LIN, Shinfeng D et Tszan WU (2011), « An integrated technique for splicing and copy-move forgery image detection », in : *Image and Signal Processing (CISP), 2011 4th International Congress on*, t. 2, IEEE, page(s): 1086-1090.

LIN, Zhouchen et al. (2009), « Fast, Automatic and Fine-grained Tampered JPEG Image Detection via DCT Coefficient Analysis », in : *Pattern Recognition* 42.*11*.

LOWE, David G. (2004), « Distinctive Image Features from Scale-Invariant Keypoints », in : *International Journal of Computer Vision* 60.*2*, page(s): 91-110.

LUO, Weiqi, Jiwu HUANG et Guoping QIU (2006), « Robust detection of region-duplication forgery in digital image », in : *Proceedings of the 18th International Conference on Pattern Recognition-Volume 04*, IEEE Computer Society, page(s): 746-749.

MA, Jing et al. (2016), « Detecting Rumors from Microblogs with Recurrent Neural Networks. », in : *IJCAI*, page(s): 3818-3824.

MAHDIAN, Babak et Stanislav SAIC (2009), « Using noise inconsistencies for blind image forensics », in : *Image and Vision Computing* 27.*10*, page(s): 1497-1503.

MAIGROT, Cédric, Vincent CLAVEAU et Ewa KIJAK (2017), « Détection de fausses informations dans les réseaux sociaux : vers des approches multi-modales », in : *Extraction et Gestion des Connaissances (EGC)*.

MAIGROT, Cédric, Vincent CLAVEAU et Ewa KIJAK (2018), « Fusion-based multimodal detection of hoaxes in social networks », in : *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, IEEE, page(s): 222-229.

MAIGROT, Cédric, Vincent CLAVEAU, Ewa KIJAK et Ronan SICRE (2016), « MediaEval 2016 : A Multimodal System for the Verifying Multimedia Use Task », in : *MediaEval Workshop*.

MAIGROT, Cédric, Ewa KIJAK et Vincent CLAVEAU (2016), « Médias traditionnels, médias sociaux : caractériser la réinformation », in : *23ème Conférence sur le Traitement Automatique des Langues Naturelles*.

— (2017), « Détection de fausses informations dans les réseaux sociaux : l'utilité des fusions de connaissances », in : *COnférence Recherche d'Information et Applications*.

— (2018), « Context-aware forgery localization in social-media images : A feature-based approach evaluation », in : *2018 25th IEEE International Conference on Image Processing (ICIP)*, IEEE, page(s): 545-549.

MAIGROT, Cédric, Ewa KIJAK, Ronan SICRE et al. (2017), « Tampering detection and localization in images from social networks : A CBIR approach », in : *International Conference on Image Analysis and Processing*, Springer, page(s): 750-761.

MIDDLETON, Stuart E. (2015), « Extracting attributed verification and debunking reports from social media : mediaeval-2015 trust and credibility analysis of image and video », in : *Mediaeval 2015 Workshop*.

MIKOLOV, Tomas et al. (2013), « Efficient estimation of word representations in vector space », in : *arXiv preprint arXiv :1301.3781*.

MITCHELL, Thomas M. (1997), *Machine Learning*, 1re éd., New York, NY, USA : McGraw-Hill, Inc., ISBN : 0070428077, 9780070428072.

MORRIS, Meredith Ringel et al. (2012), « Tweeting is believing ? : understanding microblog credibility perceptions », in : *Proceedings of the ACM 2012 conference on computer supported cooperative work*, ACM, page(s): 441-450.

MUSHTAQ, Saba et Ajaz Hussain MIR (2014a), « Digital image forgeries and passive image authentication techniques : A survey », in : *International Journal of Advanced Science and Technology* 73, page(s): 15-32.

— (2014b), « Novel method for image splicing detection », in : *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on*, IEEE, page(s): 2398-2403.

NGUYEN, Tu Ngoc, Cheng LI et Claudia NIEDERÉE (2017), « On early-stage debunking rumors on twitter : Leveraging the wisdom of weak learners », in : *International Conference on Social Informatics*, Springer, page(s): 141-158.

PASQUINI, Cecilia, Fernando PÉREZ-GONZÁLEZ et Giulia BOATO (2014), « A Benford-Fourier JPEG compression detector », in : *IEEE International Conference on Image Processing (ICIP)*.

PHAN, Quoc-Tin et al. (2016), « A hybrid approach for multimedia use verification », in : *MediaEval 2016 Workshop*.

QU, Zhenhua, Guoping QIU et Jiwu HUANG (2009), « Detect digital image splicing with visual cues », in : *International workshop on information hiding*, Springer, page(s): 247-261.

ROBERTSON, Stephen E., Steve WALKER et Micheline HANCOCK-BEAULIEU (1998), « Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive », in : *7th Text Retrieval Conference (TREC)*.

RUCHANSKY, Natali, Sungyong SEO et Yan LIU (2017), « Csi : A hybrid deep model for fake news detection », in : *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, ACM, page(s): 797-806.

SALLOUM, Ronald, Yuzhuo REN et C-C Jay KUO (2018), « Image Splicing Localization Using A Multi-Task Fully Convolutional Network (MFCN) », in : *Journal of Visual Communication and Image Representation* 51, page(s): 201-209.

SCHMID, Helmut (1994), « Probabilistic Part-of-Speech Tagging Using Decision Trees », in : *International Conference on New Methods in Language Processing*, Manchester, UK, page(s): 44-49.

SILVERMAN, Craig (2014), *Verification Handbook : An Ultimate Guideline on Digital Age Sourcing for Emergency Coverage*, sous la dir. de Craig SILVERMAN, The European Journalism Centre (EJC).

SIMONYAN, Karen et Andrew ZISSERMAN (2014), « Very Deep Convolutional Networks for Large-Scale Image Recognition », in : *CoRR* abs/1409.1556.

SPARCK JONES, Karen (1972), « A statistical interpretation of term specificity and its application in retrieval », in : *Journal of documentation* 28.*1*, page(s): 11-21.

TOLIAS, Giorgos, Ronan SICRE et Hervé JÉGOU (2016), « Particular object retrieval with integral max-pooling of CNN activations », in : *4th International Conference on Learning Representations (ICLR)*.

VEIT, Andreas et al. (2016), « COCO-Text : Dataset and Benchmark for Text Detection and Recognition in Natural Images », in : *arXiv preprint arXiv :1601.07140*.

WANG, Qing et Rong ZHANG (2016), « Double JPEG compression forensics based on a convolutional neural network », in : *EURASIP Journal on Information Security* 2016.*1*, page(s): 23.

WANG, Wei, Jing DONG et Tieniu TAN (2009), « Effective image splicing detection based on image chroma », in : *Image Processing (ICIP), 2009 16th IEEE International Conference on*, IEEE, page(s): 1257-1260.

WU, Ke, Song YANG et Kenny Q ZHU (2015), « False rumors detection on sina weibo by propagation structures », in : *Data Engineering (ICDE), 2015 IEEE 31st International Conference on*, IEEE, page(s): 651-662.

YANG, Fan et al. (2012), « Automatic detection of rumor on Sina Weibo », in : *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, ACM, page(s): 13.

YU, Feng et al. (2017), « A convolutional approach for misinformation identification », in : *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, AAAI Press, page(s): 3901-3907.

ZAMPOGLOU, Markos, Symeon PAPADOPOULOS et Yiannis KOMPATSIARIS (2015), « Detecting image splicing in the wild (WEB) », in : *Conference on Multimedia & Expo Workshops (ICMEW), 2015 IEEE International*, IEEE, page(s): 1-6.

ZHAO, Zhe, Paul RESNICK et Qiaozhu MEI (2015), « Enquiring minds : Early detection of rumors in social media from enquiry posts », in : *Proceedings of the 24th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, page(s): 1395-1405.

ZHOU, Denny et al. (2004), « Learning with local and global consistency », in : *Advances in neural information processing systems*, page(s): 321-328.

ZHU, Xiaojin et Zoubin GHAHRAMANI (2002), « Learning from labeled and unlabeled data with label propagation », in :

ZHU, Xiaojin, Zoubin GHAHRAMANI et John D LAFFERTY (2003), « Semi-supervised learning using gaussian fields and harmonic functions », in : *Proceedings of the 20th International conference on Machine learning (ICML-03)*, page(s): 912-919.

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

Titre : détection de fausses informations dans les réseaux sociaux

Mot clés : fausses informations, réseaux sociaux, crédibilité de la source, traitement des langues, analyse du signal

Resumé : Les fausses informations se multiplient et se propagent rapidement sur les réseaux sociaux. Dans cette thèse, nous analysons les publications d'un point de vue multimodal entre le texte et l'image associée. Plusieurs études ont été menées durant cette thèse. La première compare plusieurs types de médias présents sur les réseaux sociaux et vise à les discriminer de manière automatique. La second permet la détection et la localisation de modifications dans une image grâce à la comparaison avec une ancienne version de l'image. Enfin, nous nous sommes intéressés à des fusions de connaissances basées sur les prédictions d'autres équipes de recherche afin de créer un système unique.

Title : detection of false information on social networks

Keywords : false information, social networks, source trustworthiness, natural language processing, signal analysis

Abstract : False information are multiplying and are spreading quickly on social networks. In this thesis, we analyze the publications from a multimodal point of view between the text and the associated image. Several studies were conducted during this thesis. The first compares several types of media present on social networks and aims to discriminate them automatically. The second one allows the detection and the localization of modifications in an image thanks to the comparison with an old version of this image. Finally, we focused on merged knowledge based on the predictions of other research teams to create a single system.