Vers une analyse des rumeurs dans les réseaux sociaux basée sur la véracité des images : état de l'art

Abderrazek Azri, Cécile Favre, Nouria Harbi, Jerome Darmont

Vers une analyse des rumeurs dans les réseaux sociaux basée sur la véracité des images : état de l'art

Abderrazek Azri, Cécile Favre, Nouria Harbi, Jérôme Darmont

Université de Lyon, Lyon 2, ERIC EA 3083
5 avenue Pierre Mendès France, F69676 Bron Cedex, France
{a.azri, cecile.favre, nouria.harbi, jerome.darmont}@univ-lyon2.fr

Résumé. Le développement rapide des réseaux sociaux a favorisé l'échange d'une masse de données importante, mais aussi la propagation de fausses informations. De nombreux travaux se sont intéressés à la détection des rumeurs, basés principalement sur l'analyse du contenu textuel des messages. Cependant, le contenu visuel, notamment les images, demeure ignoré ou peu exploité. Or, les données visuelles sont très répandues sur les médias sociaux et leur exploitation s'avère être importante pour analyser les rumeurs. Dans cet article, nous présentons une synthèse de l'état de l'art des travaux relatifs à la classification des rumeurs et résumons les tâches principales de ce processus, ainsi que les approches suivies pour analyser ce phénomène. Nous nous focalisons plus particulièrement sur les techniques adoptées pour vérifier la véracité des images. Nous discutons également les jeux de données utilisés pour l'analyse des rumeurs et présentons les pistes de recherche que nous comptons explorer.

1 Introduction

Grâce à leur capacité à nous tenir informés de divers événements, les réseaux sociaux tiennent une place de plus en plus importante dans nos vies professionnelles et personnelles. Les utilisateurs de ces plateformes génèrent une quantité massive d'informations par la création, l'annotation et le partage de contenu multimédia. Les réseaux sociaux sont aussi un terrain propice pour la propagation des rumeurs. En effet, plus d'un tiers des sujets d'actualité sur les sites de *microblogging* contiennent de fausses informations (Zhao et al., 2015).

Pour traquer et référencer les rumeurs, des méthodes manuelles ont été développées dans un premier temps. Basées sur l'annotation des données, elles posent toutefois des problèmes liés à la qualité de l'annotation, qui est par ailleurs une tâche difficile et coûteuse demandant des compétences spécifiques (Cao et al., 2018). Face à ces contraintes, la communauté scientifique propose maintenant des méthodes semi-automatiques qui réduisent la charge de travail en ne nécessitant pas l'étiquetage de toutes les données. De plus, les résultats sont plus précis que ceux obtenus par les méthodes manuelles.

La majorité des travaux relatifs à l'analyse des rumeurs dans les réseaux sociaux est basée sur l'exploitation du contenu textuel des messages et vise à prédire la véracité du contenu en ligne. L'exploitation du contenu visuel n'est pas abordée, alors qu'elle pourrait avoir un rôle important.

C'est dans cette perspective que nous présentons et discutons dans cet article un état de l'art qui porte sur : 1) des travaux qui traitent le problème de la classification des rumeurs dans les réseaux sociaux ; 2) des méthodes d'analyse de la véracité des images dans les réseaux sociaux. Cet état de l'art nous conduit à proposer une cartographie des tâches et des familles de méthodes pour prédire la véracité des rumeurs, ainsi qu'une typologie des approches de vérification de la véracité d'une image. Ceci nous permet de

développer alors les travaux que nous prévoyons pour associer les contenus textuel et visuel des messages de réseaux sociaux en vue d'analyser la véracité des rumeurs.

Le reste de l'article est structuré comme suit. Nous présentons tout d'abord les différentes définitions d'une rumeur (Section 2). Nous présentons et discutons ensuite les tâches principales liées à la classification des rumeurs (Section 3), puis les différentes techniques utilisées pour statuer sur la véracité des images dans les réseaux sociaux (Section 4), nous faisons ensuite un état des lieux des jeux de données utilisés dans les travaux liés à la classification des rumeurs (Section 5), jeux de données nécessaires pour envisager les possibilités de développement des pistes de recherche que nous souhaitons explorer dans ce contexte et que nous présentons ici (Section 6). Nous concluons finalement cet article (Section 7).

2 Définitions d'une rumeur

Les rumeurs sont le sujet de recherches dans plusieurs disciplines, comme la philosophie (DiFonzo et Bordia, 2006; Donovan, 2007), la psychologie sociale (Allport et Postman, 1965; Jaeger et al., 1980; Rosnow et Foster, 2005), les sciences politiques (Allport et Postman, 1946; Berinsky, 2017), les sciences de gestion (DiFonzo et al., 1994; Kimmel, 2013) et, plus récemment, l'informatique, notamment avec l'expansion des réseaux sociaux.

De nombreuses définitions différentes de la rumeur ont été proposées dans la littérature. Zhao et al. (2015) définissent une rumeur comme étant une déclaration controversée et vérifiable, Zubiaga et al. (2018) comme une information en circulation dont le statut de véracité n'a pas encore été vérifié au moment de la publication. Pour Hamidian et Diab (2015), une rumeur peut être à la fois vraie et fausse. C'est une affirmation dont la véracité est mise en doute et n'a pas de source claire, même si ses origines et ses intentions idéologiques ou partisanes sont claires.

Cao et al. (2018) vont plus loin en proposant des familles de rumeurs. Les rumeurs générales ont valeur de vérité non vérifiée. Les rumeurs générales englobent deux sous-familles de rumeurs : les rumeurs objectives, dont la valeur de vérité est confirmée par une source fiable ou crédible ; et les rumeurs subjectives, dont la valeur de vérité est déterminée par les jugements subjectifs des utilisateurs des réseaux sociaux.

Ces définitions partagent généralement deux idées communes sur la nature des rumeurs. Primo, une des caractéristiques des rumeurs est qu'elles apparaissent dans un contexte d'ambiguïté. Par conséquent, leur véracité est incertaine. Secundo, bien que sa valeur de véracité soit incertaine, une rumeur n'implique pas forcément de fausses informations.

À notre sens, la définition la plus utilisée par la communauté scientifique est celle évoquée par DiFonzo et Bordia (2006) et Qazvinian et al. (2011), où une rumeur est définie comme une information qui émerge et se propage, et dont la valeur de vérité est non vérifiée ou délibérément fausse. Nous l'adoptons donc dans cet article.

3 Tâches impliquées dans l'analyse de rumeurs

L'analyse de la crédibilité des rumeurs issues des réseaux sociaux est une tâche difficile qui exige, d'une part, la compréhension du contenu des rumeurs et le contexte social associé à la diffusion de ces rumeurs et, d'autre part, l'élaboration de méthodes de détection efficaces. Nous identifions dans cette section trois tâches qui permettent d'aboutir à une classification des rumeurs.

3.1 Détection des rumeurs

La détection d'une rumeur consiste, à partir d'un ensemble de messages d'un média social, en une classification binaire où chaque message est labellisé {Rumeur, Non Rumeur}. L'objectif est donc la

détection précoce de la rumeur comme étant une information non encore vérifiée, sans pour autant statuer sur sa valeur de véracité.

Pour cela, Zhao et al. (2015) partent de l'idée que les rumeurs provoquent des messages d'utilisateurs sceptiques, qui posent des questions ou se renseignent sur la véracité de la rumeur. Un certain nombre de messages interrogatifs implique alors que l'information est une rumeur. Pour identifier les messages interrogatifs, les auteurs ont créé manuellement une liste de cinq expressions régulières (par exemple, *is (that | this | it) true*). Les messages sont ensuite regroupés en utilisant la similarité de Jaccard et chaque groupe constitue une rumeur potentielle.

Zubiaga et al. (2016b) proposent une autre approche pour classer une information comme rumeur ou non, en s'appuyant sur l'hypothèse selon laquelle un tweet à lui seul ne suffit peut-être pas, en raison de l'absence de contexte. Ils utilisent donc une approche d'apprentissage de contexte exploitant les champs conditionnels aléatoires (CRF) comme un classifieur séquentiel. Les expériences effectuées sur cinq jeux de données démontrent que cette approche a une meilleure précision que celle de Zhao et al. (2015).

Enfin, pour identifier les rumeurs dans les médias sociaux pendant les périodes de catastrophes et d'urgence, McCreadie et al. (2015) utilisent des plateformes du *crowdsourcing*. En analysant trois jeux de données, ils concluent que l'étiquetage des rumeurs par ces plateformes a un potentiel similaire aux méthodes automatiques de détection des rumeurs. Les auteurs catégorisent aussi les rumeurs en six types : message non fondé, message contesté, désinformation, message racontant un événement, message s'opposant à une rumeur et message exprimant une opinion.

D'autres travaux sont consacrés à la détection de la ou des sources d'une rumeur. Les techniques développées sont inspirées des modèles épidémiologiques utilisés pour décrire les processus d'infection et de récupération des nœuds d'un réseau (Hethcote, 2000), notamment, *Susceptible-Infected*, *Susceptible-Infected-Recovery* et *Susceptible-Infected-Susceptible*. Les modèles, *Independent Cascad* (Goldenberg et al., 2001) et *Linear Threshold* (Granovetter, 1978), sont utilisés pour traiter le problème de la propagation de l'influence dans les réseaux sociaux (Kempe et al., 2003).

Puisque ces travaux explorent le problème de la diffusion de l'information et que nous nous focalisons sur les approches de détection et de prédiction de la véracité des rumeurs, nous ne les détaillons pas plus avant.

3.2 Classification de la position

La classification de la position dans les réseaux sociaux en ligne consiste à déterminer le type d'orientation que chaque message individuel exprime à l'égard de la véracité contestée d'une rumeur. Un classifieur de position prend en entrée un ensemble $R = \{r_1, r_2, ..., r_n\}$ de rumeurs, où chaque rumeur r_i est composée d'une collection de taille variable de messages $M = \{m_1, m_2, ..., m_k\}$ discutant de cette rumeur. L'objectif consiste à déterminer la position (pour ou contre, par exemple) de chaque message m_j traitant de r_i.

Mendoza et al. (2010) visent à comprendre les positions des utilisateurs de Twitter (*twittos*) via une analyse manuelle. Ils constatent que la majorité des tweets qui sont liés à de vraies rumeurs supportent ces rumeurs, alors que la moitié des tweets qui sont associés à de fausses rumeurs questionnent ou s'opposent à ces rumeurs. Cela tend à montrer que les positions exprimées vis-à-vis des rumeurs sont une indication précieuse pour déterminer la véracité d'une rumeur.

Qazvinian et al. (2011) proposent une méthode de classification supervisée pour classer des messages (tweets) dans les catégories « supporter », « s'opposer », « questionner » et « neutre », en utilisant un grand ensemble de caractéristiques issu du contenu des tweets et des propriétés de la propagation. De même, Zeng et al. (2016) utilisent diverses approches supervisées pour classer des messages : régression logistique, classification naïve Bayésienne et forêts aléatoires. Toutefois, ils ne classent la position qu'en deux classes : « affirmer » et « nier ». Les meilleurs résultats de classification sont obtenus avec les forêts aléatoires. Afin d'obtenir une classification plus fine de la position (en quatre types : « supporter »,

« s'opposer », « questionner » et « commenter »), Zubiaga et al. (2016a) utilisent les champs aléatoires conditionnels (CRF) comme classifieur séquentiel.

Finalement, d'autres travaux s'appuient sur l'apprentissage profond pour la classification séquentielle de la position de tweets en utilisant des *Long/Short-Term Memory Networks* (LSTM) (Kochkina et al., 2017) ou des *Bidirectionnal-LSTM* (Augenstein et al., 2016). Chen et al. (2017) adoptent, eux, les réseaux de neurones convolutionnels (CNN) pour représenter les tweets, puis attribuent des probabilités aux différentes classes auxquelles un tweet pourrait appartenir grâce à un classifieur *softmax*.

3.3 Classification de la véracité d'une rumeur

La classification de la véracité d'une rumeur est la finalité et l'étape cruciale du processus d'analyse de la rumeur. Cette tâche consiste à déterminer une valeur de véracité. Formellement, une rumeur r est définie comme un ensemble de messages $M = \{m_1, m_2, ..., m_n\}$. Détecter la véracité de la rumeur consiste à déterminer si la rumeur r est confirmée comme vraie, prouvée fausse ou que sa valeur de véracité demeure non vérifiée, par une fonction de prédiction : $f(r) \rightarrow \{\text{vraie, fausse, non vérifiée}\}$, telle que :

$$f(r) = \begin{cases} vraie & \text{si } r \text{ est confirmée comme vraie,} \\ fausse & \text{si } r \text{ est prouvée fausse,} \\ non\ v\acute{e}rifi\acute{e}e & \text{sinon.} \end{cases}$$

Castillo et al. (2011) évaluent la véracité d'un ensemble de messages (tweets) en les classant comme crédibles ou non avec des arbres de décision J48, sur la base d'un ensemble de caractéristiques issu du contenu des tweets, de leur diffusion et des *twittos* qui les ont diffusés. De manière similaire, Kwon et al. (2013) utilisent un ensemble de caractéristiques temporelles, structurelles et linguistiques, ainsi que trois types de classifieurs : arbres de décision, SVM et forêts d'arbres de décision. Les caractéristiques sélectionnées classent les rumeurs avec une grande précision et rappel entre 87 % et 92 %.

Giasemidis et al. (2016) ont mené des expériences sur 100 millions de tweets associés à 72 rumeurs différentes. Ils introduisent d'autres caractéristiques comme le comportement présent et passé des utilisateurs et proposent une méthode pour agréger les propriétés des utilisateurs et des tweets. Les meilleurs résultats de classification sont obtenus avec les arbres de décision avec un taux de précision de 100 % et 94.1 % de rappel.

Finalement, Wu et al. (2015) proposent une nouvelle technique basée sur la structure de propagation pour l'extraction des caractéristiques liées aux utilisateurs et aux messages à partir des arbres de propagation des messages, dans le contexte du site de *microbloging* Sina Weibo. Ils utilisent un classifieur SVM hybride avec les noyaux *random walk kernel* et *RBF kernel*. Le modèle développé atteint un taux de précision de classification de 91,3 % qui est nettement plus élevé par rapport à d'autres méthodes.

Contrairement aux travaux précédents, qui tentent de détecter les rumeurs et de statuer sur leur véracité, Tong et al. (2017) proposent de bloquer les rumeurs. Ils se basent sur l'hypothèse que le premier message lu par un utilisateur influence son opinion. Ils développent un algorithme pour limiter la propagation des rumeurs.

3.4 Discussion

Les travaux existants relatifs à l'analyse des rumeurs dans les réseaux sociaux s'intéressent principalement à trois tâches : la détection d'une rumeur, la classification de la position ou de la posture vis-à-vis d'une rumeur et la classification de la véracité d'une rumeur. Les composantes de ce processus sont étroitement liées aux spécificités du cas étudié. Le Tableau 1 présente quelques exemples de travaux qui ont étudié ces trois tâches. À notre connaissance, aucune contribution n'a étudiée les trois tâches simultanément, par conséquent le nombre de tâches étudiées peut varier d'un travail à l'autre.

Les techniques adoptées pour réaliser ces tâches peuvent être classées en trois familles.

Travaux	Détection de la rumeur	Classification de la position	Classification de la véracité
Zubiaga et al. (2017)	✓		
Zeng et al. (2016)		✓	
Enayet et El-Beltagy (2017)	✓	✓	
Giasemidis et al. (2016)			✓
Ma et al. (2015)			✓

TAB. 1: Tâches de classification d'une rumeur dans la littérature

1. **Basées sur le contenu textuel.** Un ensemble de caractéristiques est tout d'abord extrait du contenu des messages, des propriétés de leur diffusion et des profils des utilisateurs. Ensuite, des algorithmes de classification supervisée permettent de prédire la véracité des messages (Qazvinian et al., 2011; Hamidian et Diab, 2015; Gupta et al., 2014; Castillo et al., 2011). Dans cette approche, la qualité des caractéristiques extraites des rumeurs est une étape cruciale pour obtenir des résultats de classification fiables.
2. **Basées sur la structure de la propagation ou l'optimisation des graphes.** Contrairement aux méthodes de la première famille, qui évaluent chaque message et événement associé individuellement, ces méthodes évaluent la crédibilité des messages et des événements dans leur ensemble. Ils commencent par la création d'un graphe de crédibilité où les entités impliquées dans la détection de la rumeur, comme les messages et les utilisateurs, constituent les nœuds, et les relations entre ces entités les arêtes. Les arêtes sont pondérées par l'intensité de la relation. Chaque entité a une valeur initiale de crédibilité, puis ces valeurs de crédibilité sont propagées dans le graphe jusqu'à convergence et évaluation de la crédibilité finale de chaque entité (Gupta et al., 2012; Jin et al., 2014, 2016b; Zhou et al., 2015). L'inconvénient majeur de cette technique est qu'elle ignore le contenu textuel des messages.
3. **Basées sur l'apprentissage profond.** Ces méthodes automatiques utilisent essentiellement deux structures de réseaux de neurones : les réseaux de neurones récurrents (RNN), qui modélisent les données textuelles des messages comme des données séquentielles (Ma et al., 2016; Chen et al., 2018; Jin et al., 2017), et les CNN, qui peuvent apprendre la représentation textuelle latente des données de la rumeur et améliorer la précision de la classification (Yu et al., 2017; Nguyen et al., 2017). Grâce à leur capacité d'apprendre la représentation profonde des données de la rumeur, ces approches améliorent considérablement les performances de prédiction par rapport aux deux familles d'approches précédentes.

En guise de synthèse, nous proposons dans la Figure 1 une cartographie des tâches et des familles de méthodes utilisées pour prédire la véracité des rumeurs.

Il est à noter que les travaux s'inscrivant dans cette démarche de prédiction de la véracité des rumeurs ne se basent pas sur la véracité des images elles-mêmes. C'est précisément ce point que nous détaillons dans la section suivante.

4 Analyse de la véracité des images dans les réseaux sociaux

Dans cette section, nous étudions les méthodes utilisées pour analyser une composante importante des messages, en l'occurrence leur contenu visuel et plus particulièrement les images, du point de vue de leur véracité.

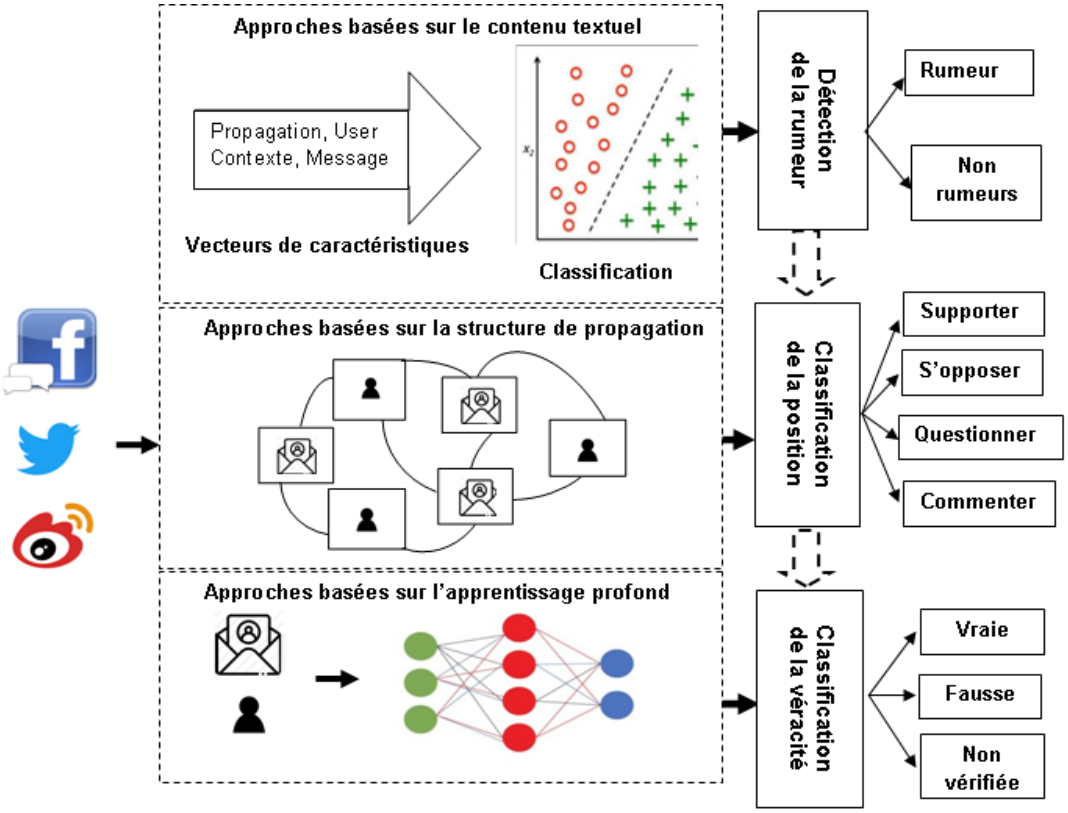

FIG. 1: Cartographie des tâches et des familles de méthodes pour prédire la véracité des rumeurs

4.1 Fausses images dans les réseaux sociaux

Il existe dans la littérature plusieurs définitions d'une fausse image. Pour Boididou et al. (2017), il s'agit de toute image attachée à un message qui ne représente pas d'une manière fidèle l'événement auquel elle fait référence. Pour Marra et al. (2018), c'est une image attachée à un faux message. Jin et al. (2016a) identifient trois types de fausses images sur les réseaux sociaux : des images anciennes utilisées pour décrire des événements récents, des images délibérément manipulées et des images qui sont utilisées d'une manière imprécise pour décrire un faux événement.

L'analyse de la véracité des images est une tâche difficile qui nécessite de relever de multiples défis. Nous classons les approches qui abordent cette problématique en trois catégories (Figure 2) : celles qui analysent le contenu de l'image pour en détecter les altérations (Marra et al., 2018; Zampoglou et al., 2015) ; celles qui analysent les caractéristiques textuelles des messages pour la classification de la véracité des images associées (Gupta et al., 2013; Jin et al., 2015) ; et celles qui utilisent des informations externes à l'image pour déterminer sa véracité (Maigrot et al., 2017). Il s'agit dans ce dernier cas de rechercher dans une base de données ou sur le Web des images similaires ou identiques afin de déterminer si l'image étudiée a été modifiée ou détournée.

4.2 Analyse du contenu de l'image

Les méthodes décrites dans cette section exploitent uniquement le contenu des images. Il existe principalement deux types d'algorithmes de détection des images contrefaites : les algorithmes actifs et passifs

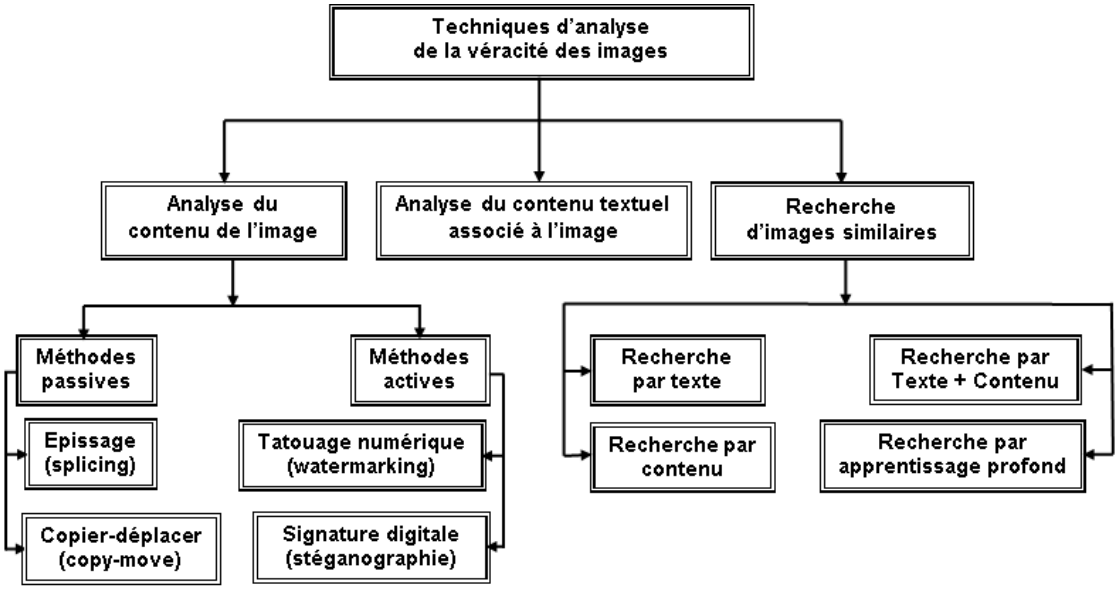

FIG. 2: Approches de vérification de la véracité d'une image

(Muhammad et al., 2014).

Les algorithmes actifs exploitent une signature, comme un filigrane, la double compression des images JPEG (Bianchi et Piva, 2012) ou la signature laissée par les appareils de capture (Goljan et al., 2011). Cette signature est mise en correspondance avec la signature de l'image originale pour détecter toute altération.

La famille des algorithmes actifs se décompose en deux sous-familles : la signature digitale (stéganographie) qui permet de cacher un message dans une image numérique, et les tatouages numériques (*watermarking*), qui incluent des informations de copyright ou d'identification dans une image.

Les algorithmes passifs exploitent le contenu même de l'image. Il existe deux familles principales d'algorithmes passifs : les algorithmes de détection d'épissage (*splicing*) et de copier-déplacer (*copy-move*). Dans une falsification par épissage, des parties de deux images ou plus sont assemblées pour former une nouvelle image. Dans une falsification par copier-déplacer, une partie d'une image est copiée et collée dans une autre partie de la même image.

4.3 Analyse du contenu textuel associé à une image

Gupta et al. (2013) proposent une méthode d'apprentissage supervisé pour la classification des images dans les messages (tweets) relatifs à l'ouragan Sandy. Pour cela, ils développent deux ensembles de caractéristiques extraites du profil des *twittos* et du contenu des tweets, qui sont exploités par classifieurs, des arbres de décision J48 et un classifieur Bayésien naïf. Ils obtiennent les meilleurs résultats de la classification avec les arbres de décision et les caractéristiques basées sur le contenu des tweets.

Par ailleurs, lors d'une investigation relative à la perception des utilisateurs concernant la crédibilité du contenu sur Twitter, Morris et al. (2012) ont découvert que les indicateurs importants sur lesquels les utilisateurs jugent la crédibilité sont les informations visibles au premier regard, notamment celles relatives au profil de l'utilisateur (son nom et son image), qui ont un grand impact sur la crédibilité des messages publiés par cet utilisateur.

4.4 Recherche d'images similaires

Cette approche utilise des données externes à l'image étudiée par la recherche d'images similaires dans une collection d'images de référence répertoriées comme vraies ou fausses. Une image requête dont on cherche la classe reçoit la classe de l'image la plus similaire de la base de données, si elle existe. Sinon, l'image requête reçoit la classe « inconnue ». Plusieurs techniques sont utilisées pour rechercher les images similaires.

Recherche textuelle. Cette technique utilise les données décrivant l'image, comme des métadonnées (nom, taille, format, titre de la page web, ...), des mots clés, des tags ou même un message associé à l'image dans le cas des réseaux sociaux. Les images sont indexées dans une base de données selon ces données textuelles et l'utilisateur formule des requêtes textuelles pour rechercher des images (Karthikram et Parthiban, 2014; Alkhawlani et al., 2015).

Recherche par le contenu. À partir du contenu de l'image étudiée, on calcule une signature et on la stocke dans une base de données indexée *offline*. La recherche d'images similaires implique le calcul *online* de la signature de l'image étudiée et sa comparaison avec toutes les signatures stockées dans la base de données.

La signature est calculée à partir des caractéristiques visuelles de l'image : les pixels ou toute information dérivée de l'image elle-même comme les couleurs, les textures, les formes ou une combinaison de ces caractéristiques. Une fois les caractéristiques extraites, la comparaison consiste à définir diverses distances entre ces caractéristiques et à définir une mesure de similarité globale entre deux images. On peut alors calculer la similarité entre une image requête et celles de la base de données. Les images résultats sont classées selon leur score de similarité (Rashno et Rashno, 2019; Kumari, 2019).

Recherche mixte. L'exploitation simultanée des contenus textuel et visuel d'une image s'avère une bonne alternative. Luo et al. (2003) ont en effet testé trois combinaisons pour concevoir un système de recherche d'images similaires sur le Web : utiliser la recherche textuelle en premier, puis la recherche par le contenu parmi les images résultant de la première recherche ; l'inverse ; et les deux types de recherche simultanément. Les meilleurs résultats, évalués en fonction de la précision et du rappel, sont obtenus avec la première approche. Unar et al. (2019) proposent un système qui prend en charge trois modes de recherche : par image requête, par mot-clé et les deux ensemble. Ils fusionnent les caractéristiques textuelles et visuelles de l'image dans un seul vecteur en utilisant des techniques de sac de mots textuels et visuels, respectivement.

Recherche par apprentissage profond. Afin de prédire automatiquement la véracité d'images à partir de leur contenu seulement, Maigrot et al. (2017) calculent les descripteurs des images en utilisant les CNN. La similarité entre l'image requête et les images de la base de référence est calculée par une mesure de similarité cosinus.

4.5 Discussion

L'analyse des différentes approches de vérification de la véracité des images nous a permis de tirer la synthèse et les conclusions suivantes.

Approche basée sur le contenu textuel Cette approche exploite les caractéristiques traditionnelles du contenu des messages, des propriétés de leur diffusion ou du profil de l'utilisateur. Elle ignore le contenu visuel lorsqu'il faut statuer sur la véracité des images.

Approche basée sur l'analyse du contenu des images

— Les algorithmes actifs de détection de falsification ne sont pas utiles dans notre contexte, car les techniques de *tatouages* numériques ou de *stéganographie* modifient les valeurs de quelques pixels permettant de cacher un message dans l'image sans modifier son aspect visuel. Si une telle image est diffusée sur les réseaux sociaux, elle ne suscite aucun intérêt car les modifications sont invisibles à l'œil nu. De plus, les techniques basées sur un format particulier d'image (par exemple JPEG) ou sur le type de l'appareil de capture ne sont pas applicables dans le contexte des réseaux sociaux. En effet, les formats d'image peuvent subir des altérations lors de la publication du message. De même, l'information relative au type de l'appareil de capture de la photo n'est pas non plus toujours disponible.

— Les techniques de détection des images épissées sont difficilement applicables dans le contexte des réseaux sociaux, en raison des opérations de redimensionnement et de recompression d'images automatiquement appliquées par les plateformes de réseaux sociaux à tous les contenus téléchargés (Zampoglou et al., 2015).

— Les techniques de détection de copier-déplacer, bien qu'elles soient fiables sur des petits jeux de données, ne sont pas susceptibles de passer à l'échelle (Warif et al., 2016). Elles n'ont de fait jamais été utilisées dans le contexte des médias sociaux.

— Un autre défi pour l'analyse de la véracité des images est que les plateformes de médias sociaux ont tendance à effacer les métadonnées, en particulier les données Exif des images, qui sont des informations utiles pour la détection de l'altération [1].

Approche basée sur la recherche d'images similaires

— La recherche d'images similaires via des éléments textuels nécessite la description ou l'annotation manuelle des images, qui est une tâche difficile, notamment quand les données sont volumineuses. Elle est également fortement dépendante de la langue utilisée. De plus, les images étant riches en contenu et présentant différents niveaux de détail, un même annotateur peut donner, de par sa subjectivité, une description différente de deux images avec le même contenu visuel. Par conséquent, les résultats pertinents de recherche pourraient être accompagnés par un grand nombre de résultats non pertinents à cause du niveau de description faible de l'image requête ou/et de l'annotation des images de la base d'images.

— Dans la recherche d'images similaires par le contenu, les caractéristiques visuelles calculées sont dites de bas niveau, car elles sont très proches du signal et ne comportent aucune information sémantique. De ce fait, ces techniques obtiennent des résultats satisfaisants pour certains types de requêtes et certains types de base d'images, toutefois elles peuvent rendre des résultats éloignés de l'objet de la requête soumise par l'utilisateur. De plus, les difficultés liées au calcul des caractéristiques des images et leur stockage peuvent survenir, particulièrement si la recherche des images similaires est effectuée sur le Web qui contient un nombre illimité d'images.

— L'utilisation de techniques d'apprentissage profond pour la recherche d'images similaires pourrait être une solution au problème de l'écart sémantique qui existe entre les caractéristiques visuelles de bas niveau capturées par des machines et les concepts sémantiques de haut niveau perçus par l'être humain. Des expériences poussées menées par Wan et al. (2014) ont donné des résultats encourageants par rapport aux approches précédentes.

Cette section a permis de présenter les travaux relatifs à l'analyse de la véracité des images. Il est à noter que ces travaux ne s'inscrivent pas spécialement dans une démarche liée à la vérification de rumeurs qui constitue notre objectif.

1. http ://www.embeddedmetadata.org/social-media-test-procedure.php

5 Jeux de données

La collecte de données à partir des médias sociaux est une phase importante pour réussir une bonne classification des rumeurs. Elle est réalisable grâce aux interfaces de programmation applicative (API) mises à disposition, notamment, par certaines plateformes de réseaux sociaux. Ces outils, ouverts et documentés, fournissent un ensemble de méthodes bien définies pour la collecte des données et permettent de construire facilement des applications ou des services.

Les trois principaux médias sociaux utilisés comme sources de données pour la classification des rumeurs sont Twitter, Sina Weibo et Facebook. L'API de Twitter est la plus ouverte, ce qui explique sa fréquente utilisation par la communauté scientifique. Cependant, elle est conçue principalement pour collecter des données récentes ou en temps réel. Il est donc difficile de collecter des données anciennes. L'API de Sina Weibo est similaire à celle de Twitter, mais présente plus de restrictions pour la collection de données. L'API de Facebook présente aussi des restrictions pour collecter les données, car la plus grande partie du contenu posté par les utilisateurs de Facebook est privé. De plus, les métadonnées fournies avec chaque message sont plus limitées que celles fournies par l'API de Twitter.

La comparaison entre les travaux de la classification des rumeurs s'avère difficile, en raison des jeux de données (*datasets*) différents utilisés dans leur validation. De plus, le manque de jeux de données accessibles publiquement a constitué jusqu'à récemment une limitation pour le développement de ces recherches. Heureusement, quelques jeux de données publiés récemment, formés de données réelles collectées sur les médias sociaux, sont désormais couramment utilisés pour la classification des rumeurs (Tableau 2).

Dataset	Source	Rumeurs	Non rumeurs	Base	Contenu
PHEME	Twitter	1972	2830	message	textuel
MediaEval	Twitter	7223	10826	message	textuel+visuel
KWON	Twitter	47	45	événement	textuel
RumourEval	Twitter	145	74	événement	textuel
MULTI	Sina Weibo	4749	4779	message	textuel+visuel

TAB. 2: Exemples de jeux de données récents

PHEME [2] est un projet européen qui s'intéresse à la détection des rumeurs sur les réseaux sociaux (Derczynski et Bontcheva, 2014). Le jeu de données contient 1972 rumeurs et 2830 non-rumeurs collectées à partir de Twitter, relatives à cinq événements.

MediaEval est un jeu de données publié par la tâche *Verifying Multimedia Use* de l'atelier *Mediaeval* (Boididou et al., 2014), qui vise à vérifier la véracité du contenu multimédia sur Twitter. Le jeu de données est composé d'un ensemble d'apprentissage constitué de plus de 9000 tweets de fausses rumeurs et plus de 6000 tweets de vraies rumeurs, ainsi que d'un ensemble de test composé de 2200 tweets relatifs à 35 événements. Les données relatives aux tweets sont de natures textuelle et multimédia (images et vidéos).

KWON est composé de 47 événements rumeurs et 55 événements non rumeurs collectés à partir de Twitter (Kwon et al., 2013). Chaque événement contient au moins 60 tweets.

RumourEval est une sous-tâche de SemEval 2017 (*International Workshop on Semantic Evaluation*) qui travaille sur la véracité des rumeurs (Derczynski et al., 2017). Ce jeu de données est composé de 325 rumeurs dont 145 vraies, 74 fausses et 106 non vérifiées.

Finalement, MULTI comprend 4749 messages rumeurs et 4779 messages non rumeurs, de natures textuelle et visuelle, collectés à partir de Sina Weibo (Jin et al., 2017).

2. http ://www.pheme.eu/

Les jeux de données basés sur des événements regroupent un ensemble de messages liés à un événement donné. Ce type de *dataset* est recommandé pour la classification d'événements dans les médias sociaux. *A contrario*, les jeux de données basés sur des messages traitent de messages individuels.

Une limitation de tous ces jeux de données est que leur taille n'est pas suffisamment grande, bien que certains travaux commencent à constituer des jeux de données relativement grands. Une autre difficulté dans la collecte des jeux de données est que le nombre de rumeurs prouvées fausses par des sources fiables est inférieur à celui des non rumeurs, générant ainsi des jeux de données d'apprentissage fortement déséquilibrés.

6 Perspectives et plan de recherche

La détection et la classification des rumeurs dans les réseaux sociaux en ligne est une tâche très difficile, en raison de la quantité massive de données bruitées transitant dans ces canaux de communication et la complexité de l'environnement de diffusion.

L'analyse des rumeurs par les méthodes manuelles basées sur l'avis des experts humains à travers les plateformes de référencement des rumeurs ou l'apport des utilisateurs des réseaux sociaux s'avère peu efficace et coûteuse en temps et en effort. Aussi, la communauté scientifique affiche beaucoup d'intérêt ces dernières années pour des méthodes automatiques d'analyse des données issues des médias sociaux.

La majorité de ces méthodes se concentre sur l'analyse du contenu textuel pour statuer sur la véracité du contenu en ligne, et ignore ou exploite peu le contenu visuel des messages.

Les images véhiculent notamment un contenu riche en information et sont facilement compréhensibles. Elles sont devenues très répandues dans les médias sociaux. À titre d'exemple, les *twittos* étant limités à un texte de 280 caractères, ils utilisent souvent une image jointe pour bien décrire un événement ou publier un texte long. Par conséquent, l'exploitation du contenu visuel est importante pour prédire la véracité du contenu intégral des messages.

Dans ce contexte, et outre de trouver des fonctionnalités et des algorithmes plus efficaces, nous résumons les principaux défis que nous souhaitons traiter dans nos travaux futurs consacrés à l'analyse de la véracité des rumeurs basée sur les images dans les réseaux sociaux.

1. Les rumeurs qui circulent en nombre important via les médias sociaux sont constituées, en plus du contenu textuel, de données multimédia (images et vidéos) qui posent des problèmes aux méthodes traditionnelles de détection (Cao et al., 2018). À notre sens, la recherche et l'analyse des relations potentielles entres ces données multimodales pourrait être enrichie par une classification hybride des composantes visuelles et textuelles des messages, afin d'obtenir une détection plus fine des rumeurs.
2. En plus des caractéristiques textuelles traditionnelles des messages, le développement des caractéristiques similaires pour l'image et l'exploitation du contenu visuel pourraient être une clé pour la vérification de la véracité des rumeurs. À titre d'exemple, les caractéristiques visuelles relatives à l'estimation de la qualité d'image dans les systèmes de recherche d'images similaires, comme le score de clarté, le score de cohérence ou le score de diversité (Jin et al., 2017) devraient être utiles dans la classification des rumeurs.
3. Les travaux relatifs à l'analyse des rumeurs se contentent généralement de donner une valeur booléenne unique de la véracité des rumeurs et ne présentent aucun détail sur les raisons de cette décision. Nous pensons qu'une bonne classification doit proposer à l'utilisateur des preuves ou des sources d'information à l'appui de la décision prise, qui pourraient être utiles pour démystifier la rumeur et empêcher sa propagation. De telles preuves doivent aider l'utilisateur à rendre son propre jugement vis-à-vis de la véracité de la rumeur.
4. L'absence de jeux de données de référence de taille suffisamment grande constitue un handicap pour évaluer l'efficacité de chaque technique et comparer les approches entre elles. En effet, les

jeux de données disponibles sont insuffisants pour : (i) acquérir de nouvelles connaissances sur les propriétés pertinentes des rumeurs et (ii) la construction de modèles capables de fonctionner correctement dans un scénario réel. Les recherches devraient porter sur la construction de jeux de données volumineux et surtout la conception d'un banc d'essai.

5. S'agissant des modèles utilisés, la classification des rumeurs dans les réseaux sociaux est généralement supervisée. À notre sens, l'orientation des efforts de recherche vers des modèles semi-supervisés ou non-supervisés pourrait être une solution. Ces techniques ont en effet données des résultats prometteurs dans le domaine de la vérification des faits (*fact cheking*) (Shi et Weninger, 2016).

7 Conclusion

Dans cet article, nous avons présenté et discuté l'état de l'art des travaux traitant le problème de la classification des rumeurs dans les réseaux sociaux, en identifiant les tâches principales du processus de classification des rumeurs, à savoir la détection des rumeurs, la classification de la position et de la véracité des rumeurs. Nous avons également catégorisé les approches relevant de ces tâches en trois paradigmes : méthodes basées sur le contenu textuel, méthodes basées sur la structure de propagation et méthodes basées sur l'apprentissage profond.

Compte-tenu de l'importance du contenu visuel des messages dans les réseaux sociaux, nous avons présenté et discuté également les techniques d'analyse de la véracité et de l'intégrité des images dans les médias sociaux. Nous avons identifié trois approches : les méthodes basées sur le contenu textuel des messages associés ou décrivant l'image, les méthodes basées sur l'analyse du contenu de l'image pour la détection d'éventuelles manipulations et les approches basées sur la recherche d'images similaires dans une base de données externe.

Nous avons exposé les jeux de données utilisés dans les travaux liés à la détection et à l'analyse des rumeurs, qui pourraient être utilisés dans le cadre du développement de nos propres perspectives de recherche que nous avons présentées et qui visent à exploiter l'analyse de la véracité des images pour détecter la véracité des rumeurs elles-mêmes dans les réseaux sociaux.

Références

Alkhawlani, M., M. Elmogy, et H. El Bakry (2015). Text-based, content-based, and semantic-based image retrievals : A survey. *IJCIT 4*(01).

Allport, G. et L. Postman (1965). The psychology of rumor. russel & russell.

Allport, G. W. et L. Postman (1946). An analysis of rumor. *Public Opinion Quarterly 10*(4), 501–517.

Augenstein, I., T. Rocktäschel, A. Vlachos, et K. Bontcheva (2016). Stance detection with bidirectional conditional encoding. *arXiv preprint arXiv :1606.05464*.

Berinsky, A. J. (2017). Rumors and health care reform : experiments in political misinformation. *BJPS 47*(2), 241–262.

Bianchi, T. et A. Piva (2012). Image forgery localization via block-grained analysis of jpeg artifacts. *IEEE Transactions on IFS 7*(3), 1003–1017.

Boididou, C., S. Papadopoulos, L. Apostolidis, et Y. Kompatsiaris (2017). Learning to detect misleading content on twitter. In *ICMR 2017*, pp. 278–286. ACM.

Boididou, C., S. Papadopoulos, Y. Kompatsiaris, S. Schifferes, et N. Newman (2014). Challenges of computational verification in social multimedia. In *Proceedings of the 23rd ICWWW*, pp. 743–748. ACM.

Cao, J., J. Guo, X. Li, Z. Jin, H. Guo, et J. Li (2018). Automatic rumor detection on microblogs : A survey. *arXiv preprint arXiv :1807.03505*.

Castillo, C., M. Mendoza, et B. Poblete (2011). Information credibility on twitter. In *WWW 2011*, pp. 675–684. ACM.

Chen, T., X. Li, H. Yin, et J. Zhang (2018). Call attention to rumors : Deep attention based recurrent neural networks for early rumor detection. In *PAKDD 2018*, pp. 40–52. Springer.

Chen, Y.-C., Z.-Y. Liu, et H.-Y. Kao (2017). Ikm at semeval-2017 task 8 : Convolutional neural networks for stance detection and rumor verification. In *Workshop SemEval 2017*, pp. 465–469.

Derczynski, L. et K. Bontcheva (2014). Pheme : Veracity in digital social networks. In *UMAP workshops*.

Derczynski, L., K. Bontcheva, M. Liakata, R. Procter, G. W. S. Hoi, et A. Zubiaga (2017). Semeval-2017 task 8 : Rumoureval : Determining rumour veracity and support for rumours. *arXiv preprint arXiv :1704.05972*.

DiFonzo, N. et P. Bordia (2006). Rumor, gossip and urban legends. *Diogenes* (1), 23–45.

DiFonzo, N., P. Bordia, et R. L. Rosnow (1994). Reining in rumors. *Organizational Dynamics 23*(1), 47–62.

Donovan, P. (2007). How idle is idle talk ? one hundred years of rumor research. *Diogenes 54*(1), 59–82.

Enayet, O. et S. R. El-Beltagy (2017). Niletmrg at semeval-2017 task 8 : Determining rumour and veracity support for rumours on twitter. In *Proceedings of the 11th IWSE (SemEval-2017)*, pp. 470–474.

Giasemidis, G., C. Singleton, I. Agrafiotis, J. R. Nurse, A. Pilgrim, C. Willis, et D. V. Greetham (2016). Determining the veracity of rumours on twitter. In *ICSI*, pp. 185–205. Springer.

Goldenberg, J., B. Libai, et E. Muller (2001). Talk of the network : A complex systems look at the underlying process of word-of-mouth. *Marketing letters 12*(3), 211–223.

Goljan, M., J. J. Fridrich, et M. Chen (2011). Defending against fingerprint-copy attack in sensor-based camera identification. *IEEE Transactions on IFS 6*(1), 227–236.

Granovetter, M. (1978). Threshold models of collective behavior. *American journal of sociology 83*(6), 1420–1443.

Gupta, A., P. Kumaraguru, C. Castillo, et P. Meier (2014). Tweetcred : Real-time credibility assessment of content on twitter. In *ICSI*, pp. 228–243. Springer.

Gupta, A., H. Lamba, P. Kumaraguru, et A. Joshi (2013). Faking sandy : characterizing and identifying fake images on twitter during hurricane sandy. In *WWW 2013*, pp. 729–736. ACM.

Gupta, M., P. Zhao, et J. Han (2012). Evaluating event credibility on twitter. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, pp. 153–164. SIAM.

Hamidian, S. et M. Diab (2015). Rumor detection and classification for twitter data. In *SOTICS 2015*, pp. 71–77.

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM review 42*(4), 599–653.

Jaeger, M. E., S. Anthony, et R. L. Rosnow (1980). Who hears what from whom and with what effect : A study of rumor. *Personality and Social Psychology Bulletin 6*(3), 473–478.

Jin, Z., J. Cao, H. Guo, Y. Zhang, et J. Luo (2017). Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *ICM 2017*, pp. 795–816. ACM.

Jin, Z., J. Cao, Y.-G. Jiang, et Y. Zhang (2014). News credibility evaluation on microblog with a hierarchical propagation model. In *ICDM 2014*, pp. 230–239. IEEE.

Jin, Z., J. Cao, J. Luo, et Y. Zhang (2016a). Image credibility analysis with effective domain transferred deep networks. *arXiv preprint arXiv :1611.05328*.

Jin, Z., J. Cao, Y. Zhang, et J. Luo (2016b). News verification by exploiting conflicting social viewpoints in microblogs. In *30th AAAI Conference on AI*.

Jin, Z., J. Cao, Y. Zhang, et Y. Zhang (2015). Mcg-ict at mediaeval 2015 : Verifying multimedia use with a two-level classification model. In *MediaEval*.

Jin, Z., J. Cao, Y. Zhang, J. Zhou, et Q. Tian (2017). Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia 19*(3), 598–608.

Karthikram, G. M. P. et G. Parthiban (2014). Tag based image retrieval (tbir) using automatic image annotation. *IJRET 3*(03).

Kempe, D., J. Kleinberg, et É. Tardos (2003). Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 137–146. ACM.

Kimmel, A. J. (2013). *Rumors and rumor control : A manager's guide to understanding and combatting rumors*. Routledge.

Kochkina, E., M. Liakata, et I. Augenstein (2017). Turing at semeval-2017 task 8 : Sequential approach to rumour stance classification with branch-lstm. *arXiv preprint arXiv :1704.07221*.

Kumari, M. (2019). Content based image retrieval. *Available at SSRN 3371777*.

Kwon, S., M. Cha, K. Jung, W. Chen, et Y. Wang (2013). Prominent features of rumor propagation in online social media. In *ICDM*, pp. 1103–1108. IEEE.

Luo, B., X. Wang, et X. Tang (2003). World wide web based image search engine using text and image content features. In *Electronic Imaging 2003*, Volume 5018, pp. 123–130.

Ma, J., W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, et M. Cha (2016). Detecting rumors from microblogs with recurrent neural networks. In *IJCAI*, pp. 3818–3824.

Ma, J., W. Gao, Z. Wei, Y. Lu, et K.-F. Wong (2015). Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pp. 1751–1754. ACM.

Maigrot, C., E. Kijak, et V. Claveau (2017). Détection de fausses informations dans les réseaux sociaux : l'utilité des fusions de connaissances. In *COnférence Recherche d'Information et Applications*, pp. 107–122.

Marra, F., D. Gragnaniello, D. Cozzolino, et L. Verdoliva (2018). Detection of gan-generated fake images over social networks. In *2018 IEEE Conference on MIPR*, pp. 384–389. IEEE.

McCreadie, R., C. Macdonald, et I. Ounis (2015). Crowdsourced rumour identification during emergencies. In *Proceedings of the 24th International Conference on WWW*, pp. 965–970. ACM.

Mendoza, M., B. Poblete, et C. Castillo (2010). Twitter under crisis : Can we trust what we rt ? In *Proceedings of the first workshop on social media analytics*, pp. 71–79. ACM.

Morris, M. R., S. Counts, A. Roseway, A. Hoff, et J. Schwarz (2012). Tweeting is believing ? : understanding microblog credibility perceptions. In *ACM conference on CSC*, pp. 441–450.

Muhammad, G., M. H. Al-Hammadi, M. Hussain, et G. Bebis (2014). Image forgery detection using steerable pyramid transform and local binary pattern. *machine vision applications 25*(4), 985–995.

Nguyen, T. N., C. Li, et C. Niederée (2017). On early-stage debunking rumors on twitter : Leveraging the wisdom of weak learners. In *ICSI*, pp. 141–158. Springer.

Qazvinian, V., E. Rosengren, D. R. Radev, et Q. Mei (2011). Rumor has it : Identifying misinformation in microblogs. In *Proceedings of the CEMNLP*, pp. 1589–1599. ACL.

Rashno, A. et E. Rashno (2019). Content-based image retrieval system with most relevant features among wavelet and color features. *arXiv preprint arXiv :1902.02059*.

Rosnow, R. L. et E. K. Foster (2005). Rumor and gossip research. *PSA 19*(4), 1–2.

Shi, B. et T. Weninger (2016). Fact checking in heterogeneous information networks. In *Proceedings of the 25th ICCWWW*, pp. 101–102. IWWWC Steering Committee.

Tong, G., W. Wu, L. Guo, D. Li, C. Liu, B. Liu, et D.-Z. Du (2017). An efficient randomized algorithm for rumor blocking in online social networks. *IEEE Transactions on NSE*.

Unar, S., X. Wang, C. Wang, et Y. Wang (2019). A decisive content based image retrieval approach for feature fusion in visual and textual images. *Knowledge-Based Systems 179*, 8–20.

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

Warif, N. B. A., A. W. A. Wahab, M. Y. I. Idris, R. Ramli, R. Salleh, S. Shamshirband, et K.-K. R. Choo (2016). Copy-move forgery detection : Survey, challenges and future directions. *J. NCA 75*, 259–278.

Wu, K., S. Yang, et K. Q. Zhu (2015). False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st international conference on data engineering*, pp. 651–662. IEEE.

Yu, F., Q. Liu, S. Wu, L. Wang, T. Tan, et al. (2017). A convolutional approach for misinformation identification. In *IJCAI*, pp. 3901–3907.

Zampoglou, M., S. Papadopoulos, et Y. Kompatsiaris (2015). Detecting image splicing in the wild (web). In *IEEE IC on Multimedia & Expo Workshops (ICMEW)*, pp. 1–6. IEEE.

Zeng, L., K. Starbird, et E. S. Spiro (2016). # unconfirmed : Classifying rumor stance in crisis-related social media messages. In *International AAAI Conference on Web and Social Media*.

Zhao, Z., P. Resnick, et Q. Mei (2015). Enquiring minds : Early detection of rumors in social media from enquiry posts. In *WWWW 2015*, pp. 1395–1405.

Zhou, X., J. Cao, Z. Jin, F. Xie, Y. Su, D. Chu, X. Cao, et J. Zhang (2015). Real-time news certification system on sina weibo. In *WWW 2015*, pp. 983–988. ACM.

Zubiaga, A., A. Aker, K. Bontcheva, M. Liakata, et R. Procter (2018). Detection and resolution of rumours in social media : A survey. *ACM Computing Surveys (CSUR) 51*(2), 32.

Zubiaga, A., E. Kochkina, M. Liakata, R. Procter, et M. Lukasik (2016a). Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. *arXiv preprint arXiv :1609.09028*.

Zubiaga, A., M. Liakata, et R. Procter (2016b). Learning reporting dynamics during breaking news for rumour detection in social media. *arXiv preprint arXiv :1610.07363*.

Zubiaga, A., M. Liakata, et R. Procter (2017). Exploiting context for rumour detection in social media. In *ICSI*, pp. 109–123. Springer.

Summary

The rapid development of social networks has promoted the exchange of a large amounts of data, but also the spread of false information. Many research works have addressed the detection of rumors, mostly by analyzing the textual content of messages. However, the visual content, especially images, remains ignored or little exploited in the literature. Yet, visual data are very popular on social media and their exploitation proves important for analyzing rumors. In this paper, we present a synthesis of the state of the art about rumor classification and summarize the main tasks of this process, as well and the approaches to analyze this phenomenon. We particularly focus on the techniques adopted to verify the veracity of images. We also discuss the datasets used for rumor analysis and present the research leads we plan to investigate.