Noname manuscript No.
(will be inserted by the editor)

FauxWard: A Graph Neural Network Approach to Fauxtography Detection Using Social Media Comments

Lanyu Shang · Yang Zhang · Daniel Zhang · Dong Wang

the date of receipt and acceptance should be inserted later

Abstract Online social media has been a popular source for people to consume and share news content. More recently, the spread of misinformation online has caused widespread concerns. In this work, we focus on a critical task of detecting fauxtography on social media where the image and associated text together convey misleading information. Many efforts have been made to mitigate misinformation online, but we found that the fauxtography problem has not been fully addressed by existing work. Solutions focusing on detecting fake images or misinformed texts alone on social media often fail to identify the misinformation delivered together by the image and the associated text of a fauxtography post. In this paper, we develop FauxWard, a novel graph convolutional neural network framework that explicitly explores the complex information extracted from a user comment network of a social media post to effectively identify fauxtography. FauxWard is content-free in the sense that it does not analyze the visual or textual contents of the post itself, which makes it robust against sophisticated fauxtography uploaders who intentionally craft image-centric posts by editing either the text or image content. We evaluate FauxWard on two real-world datasets collected from mainstream social media platforms (i.e., Reddit and Twitter). The results show that FauxWard is both effective and efficient in identifying fauxtography posts on social media.

Keywords Fauxtography, Misinformation, Social Media, Fake News, Graph Neural Network

1 Introduction

In recent years, social media has become a popular channel for people to consume and share news content [19, 41]. However, the spread of misinformation on social media platforms has raised many concerns, and a significant amount of efforts have been made to reduce the diffusion of misinformation online [60, 31]. For

Lanyu Shang · Yang Zhang · Daniel Zhang · Dong Wang
Department of Computer Science and Engineering
University of Notre Dame, Notre Dame, IN 46556
E-mail: {lshang, yzhang42, yzhang40, dwang5}@nd.edu

example, leading social media platforms (e.g., Facebook and Google) have stepped up to tackle and prevent the spread of fake news [1]. Many solutions have been developed to combat misinformation propagation on online social media, including the analysis of news content [34], the assessment of news source credibility [56], and a set of fact-checking techniques [33]. In this paper, we focus on an important but largely unsolved problem of detecting "fauxtography" where the image(s) and the associated text of a social media post conveys a questionable or outright false sense of the events it seems to depict [9].

The increasing popularity of visual content on online social media [5, 30] has motivated our study of detecting fauxtography. For example, photos have been recognized as the primary type of content on social media. A social media post accompanied by an image is ten times more likely to attract engagements (e.g., click, like, or share) [3]. In particular, on Twitter, tweets with image content could attract 18% more clicks, 89% more likes, and 150% more retweets than tweets without images [2].

With the growing presence of image-centric content, social media has become a rich playground for the propagation of misinformation [39, 42, 28, 50]. For example, fake images about sightings of creepy killer clowns have caused national hysteria in 2016 in the USA[1]. In this paper, we investigate a unique type of misinformation on social media, namely fauxtography, where an image and its context (e.g., the associated text of the image-centric post) jointly convey misleading information to the viewers of the content. For example, all images in Figure 1 fall under our definition of fauxtography. In particular, the text of image (a) claims that a seven-headed snake was found in Honduras, while in fact, the image was manipulated (i.e., photoshopped) and the claim itself was false. Image (b) claims a baby elephant lost her mother to poachers. While the image itself was a genuine photo (i.e., unedited), the baby elephant did not lose her mother to poachers and the photo was taken when the baby elephant was playing with her keeper at Munich zoo. Image (c) claims at least one person was killed after part of a bridge in China's Zhengzhou city collapsed. Although the claim itself is a true event[2], the image is misleading because the collapsed bridge in the image was intentionally manipulated (i.e., photoshopped) to deliver a false sense that the collapsed bridge was a huge bridge crossing a wide river while the actual one was an overpass in the city. Last, image (d) accompanies a donation post for refugees during the recent bushfire in Australia. While both the image and text are real, it is misleading because the image was taken from an earlier Australian bushfire in 2013[3] and was used to exaggerate the severity of the fire. In short, all the above cases will be considered as fauxtography because the images and the associated texts together convey misleading information.

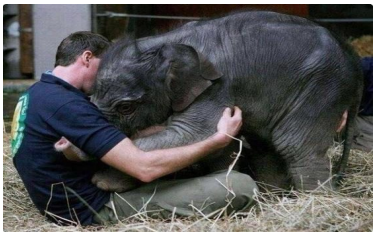
The nature of the fauxtography detection problem requires a joint consideration of not only the truthfulness of the image or its associated text, but also the relation between them. Any content-based method that asserts the veracity of post content (e.g., the image or the text of the post) will be insufficient to address this problem. For example, the "image forgery detection" solutions were devel-

[1] https://www.theverge.com/2016/10/7/13191788/clown-attack-threats-2016-panic-hoax-debunked

[2] http://www.globaltimes.cn/content/759679.shtml

[3] https://www.independent.co.uk/news/world/australasia/family-took-refuge-in-a-lake-to-escape-the-aussie-bushfires-8444881.html

(a) Fake Image, False Text



(b) Real Image, False Text

(c) Fake Image, True Text

(d) Real Image, True Text

Image (a) was titled "A rare seven-headed snake found in Honduras."
Image (b) was titled "A baby elephant lost her mother to poachers."
Image (c) was titled "At least one person was killed after part of a bridge in China's Zhengzhou city collapsed."
Image (d) was titled "Team UMA (Utsav Melbourne Association) is looking for your support and donations towards helping fellow Aussies who have lost everything in the prevailing bushfires."

Fig. 1: Examples of Fauxtography on Social Media

oped to detect image manipulation, such as copy-and-move [12], splicing [16], and image-retouch [57]. However, they only focus on detecting the verity of an image (i.e., fake image) but ignore the necessary context (e.g., the associated text) in an image-centric post. Hence, such kind of solutions cannot be directly applied to solve the fauxtography problem. For example, we observe that real images can convey misleading information that cannot be easily detected (e.g., images (b) and (d) in Figure 1). Furthermore, fact-checking solutions that are focusing on inferring the truthfulness of textual claims on social media [33, 46] are also insufficient to fully address the fauxtography problem, especially when the associated text is true but the image is fake (e.g., images (c) in Figure 1). More recently, a few fake news detection solutions were proposed to leverage both the visual features extracted from images and text information in news articles to identify fake news [44]. However, it is also insufficient to address the fauxtography problem where a falsified association between a real image and true text together convey misleading information (e.g., image (d) in Figure 1). Therefore, it is very difficult for these content-based solutions to effectively detect fauxtography.

In this paper, we develop FauxWard, a novel graph convolutional neural network based approach that can effectively track down fauxtography posts on online social media. To overcome the limitation of content-based solutions that can be misled by posts with real image and/or true text, the proposed FauxWard framework is content-free in that it approaches the fauxtography detection problem without analyzing the content (i.e., both text and image) of the post. The content-free

nature of FauxWard makes it robust against sophisticated content crafters who can intentionally modify the presentation and the description of the images [55, 53]. In particular, it leverages the user comments of the corresponding post and learns valuable information (e.g., textual content and replying pattern) from the comments to identify the fauxtography post. For example, social media users often discuss more on the image verity of the fauxtography posts, and comments that directly debunk the fauxtography post usually receive more endorsements from other users. In contrast, topics discussed in the comments of non-fauxtography posts appear to be more diversified and users tend to have less debunk and endorsement behavior in their comments for non-fauxtography. Current solutions leveraging user comments only focus on the textual contents but ignore the replying pattern of user comments [27, 10]. Previous work [54] adopted the random walk based algorithms to extract the topological features of the user comment network. We observe that the topological feature of the comment structure identified by such an approach is often insufficient and over-simplified, which leads to suboptimal performance in detecting fauxtography posts with complex user comment networks. In FauxWard, we develop a principled framework to extract a diversified set of valuable features (e.g., linguistic features, semantic features, and metadata features) from user comments to systematically characterize fauxtography. FauxWard then aggregates the extracted comment features from the user comment network of various sizes and structures through a graph convolutional neural network framework to track down fauxtography effectively.

To the best of our knowledge, the Fauxward is the first graph neural network based approach to address the fauxtography detection problem on online social media. The graph convolutional neural network design allows FauxWard to effectively learn graph-level representations of the user comment networks that vary in sizes and topological structures. We evaluate the performance of FauxWard on two real-world datasets collected from two mainstream social media platforms, Reddit and Twitter. The results show that our scheme significantly outperforms state-of-the-art fauxtography detection baselines in terms of both detection effectiveness and efficiency.

A preliminary version of this work has been published in [54] to investigate the *fauxtography detection* problem on online social media. This paper is a significant extension of the previous work (i.e., FauxBuster) in the following aspects. First, we identified a new challenge in effectively capturing the underlying topological structure of the user comment network where the size and the structure of the network differ in each post. We re-formulate the fauxtography detection problem under this new challenge. Second, we developed a new graph convolutional neural network approach, FauxWard, to address the above challenge by modeling the fauxtography detection task as a graph classification problem and jointly leveraging the linguistic and semantic attributes of the comments and the topological characteristics of the user comment network to identify fauxtography posts (Section 4). Third, we collected two new datasets from Reddit and Twitter that include more recent fauxtography posts (until 2019) to evaluate the performance and robustness of the proposed scheme in a more realistic scenario (Section 5). Fourth, we compared the FauxWard scheme with two additional state-of-the-art baselines on fauxtography and fake news detection to comprehensively study the effectiveness and efficiency of all compared schemes (Section 6). Fifth, we extended

the related work by reviewing recent works on graph neural networks (Section 2).

2 Related Work

2.1 Fauxtography

The phenomenon of "Fauxtography" first appeared in the 2006 Lebanon War when digitally manipulated photographs were used in news articles [9]. Cooper *et al.* defined fauxtography as "visual images, especially news photographs, which convey a questionable (or outright false) sense of the events they seem to depict" [9]. Examples of fauxtography include taking photos of a staged event, using images from another irrelevant event, using digital editing tools (e.g., Photoshop) to manipulate the image, applying special photography technique (e.g., wide-angle close-ups) to take photos to exaggerate the event, and generating fake images with advanced computer vision technology (e.g., Deepfake). The fauxtography phenomenon has also been observed in social science, but no practical solution has been developed [45, 13]. In this paper, we develop the FauxWard, a graph neural network approach dedicated to addressing the fauxtography detection problem on online social media.

2.2 Image Forgery Detection

Image forgery is closely related to the fauxtography detection problem. A significant amount of efforts have been made to address the image forgery problem. For example, Huynh-Kha *et al.* proposed an algorithm to detect image forgery where images are manipulated by copy-move, splicing, or both in the same image [16]. Pun *et al.* proposed a segmentation-based framework to identify image copy-move forgery [26]. Bayar *et al.* developed a convolutional neural networks based framework to suppress image contents and automatically detect image manipulations [7]. Matern *et al.* proposed a gradient-based scheme to detect image forgery by validating the consistency of illumination between pairs of objects on the image [22]. Gupta *et al.* characterized the phenomenon of fake image propagation on Twitter during a disaster event and developed a supervised detection scheme [13]. However, these schemes only focus on the visual content of the images while ignoring the associated context (e.g., text). Therefore, they cannot address the fauxtography problem when the uploaders leverage real images to convey misleading information. In contrast, FauxWard assumes the fauxtography detection must consider both images and their contexts under a holistic analytical framework.

2.3 Misinformation Detection

Misinformation has emerged as a critical issue on online social media and several solutions have been developed to mitigate the spread of misinformation [52, 38, 36, 37, 40]. For example, Yin *et al.* proposed the first fact-checking scheme *Truth Finder* that uses a Bayesian-based heuristic algorithm to combat misinformation

by finding true facts from a large amount of conflicting information [46]. Wang *et al.* developed an estimation-maximization algorithm that identifies truthful online social media posts by explicitly considering the reliability of data sources [35]. Zhang *et al.* developed a dynamic truth discovery model to incorporate physical constraints and temporal dependencies into the detection of evolving truth [51]. Vo *et al.* developed a fake news detection scheme that leverages the users who actively debunk fake information on social media, and recommends fact-checking URLs posted from these users [33]. Pérez-Rosas *et al.* proposed a natural language processing based scheme to automatically identify fake content in online news media [24]. Yang *et al.* developed a convolutional neural network framework to detect fake news by leveraging textual and visual features extracted from news articles [44]. However, these content-based solutions cannot fully address the fauxtography problem and are insufficient to capture sophisticated fauxtography posts that covey misinformation using real images and true texts. In contrast, FauxWard leverages the "wisdom of the crowd" and explores useful clues in the user comments to effectively identify misinformation of image-centric posts on social media.

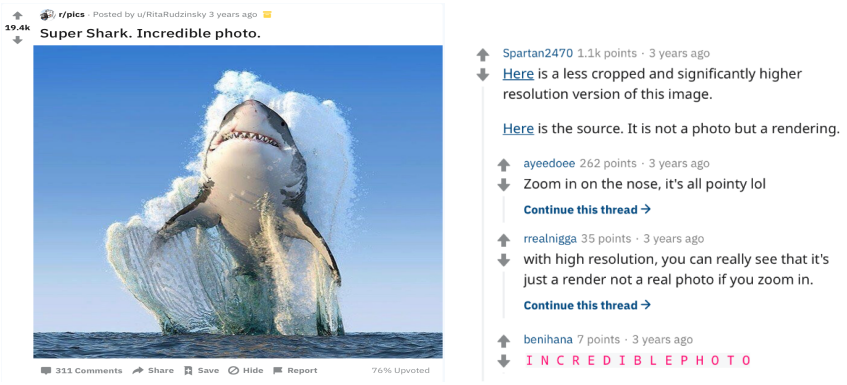
2.4 Graph Neural Network

Our work is related to Graph Neural Network (GNN) [59, 43]. GNN is a deep learning based method that can be applied to complex graph-structured data in the non-Euclidean domain, including social networks, protein-protein interaction networks, and knowledge graphs [58, 15, 11, 14]. For example, Ying *et al.* developed a random walk based graph convolutional network solution to generate high-quality recommendations in large scale recommender systems [47]. Li *et al.* proposed an adaptive GNN framework that predicts toxicological effects of chemical compounds by taking arbitrary graph-structured molecular data as input [21]. Schlichtkrull *et al.* developed a relational GNN scheme to effectively model the multi-relational data in knowledge bases [29]. Chen *et al.* proposed a batched training scheme to classify research topics on citation networks by efficiently training GNN models on large and dense graphs [8]. Nguyen *et al.* developed an argument-aware graph convolutional neural network model to detect events of interest in news articles [23]. Current GNN-based approaches often assume a homogeneous set of nodes in the input graph and ignores the complex information embedded in the nodes. In this paper, we propose a graph convolutional neural network framework that leverages key features captured from user comments and effectively classify user comment networks of various sizes and structures with a cluster-based pooling strategy. To the best of our knowledge, FauxWard is the first GNN-based approach to detect fauxtography on social media.

3 Problem Statement

In this section, we present the fauxtography detection problem on online social media. We first define a few key terms that will be used in the problem formulation.

Definition 1 Image-centric Post (P): an image-centric post (Figure 2) is a social media post that depicts an event, object, or topic with image(s), the context (i.e., text associated with the image), and the comment section.



(a) Claim and Image (b) Sample of Comments

Fig. 2: Example of an Image-centric Post on Reddit

Definition 2 Fauxtography (labeled as "True"): a post that conveys a misleading message to the viewers of the post. In particular, a post is a fauxtography if the image of the post i) directly supports a false claim, or ii) conveys misinformation of a true claim.

Definition 3 Non-Fauxtography (labeled as "False"): images that do not fall under "fauxtography".

To formulate our problem, we assume a set of N posts $\mathcal{P} = \{P_1, P_2, ..., P_N\}$ from online social media. A post P_n, $1 \leq n \leq N$, is defined as a tuple: $P_n = (\mathcal{T}_n, \mathcal{I}_n, \mathcal{C}_n, y_n)$ where $\mathcal{T}_n$ and $\mathcal{I}_n$ refer to the text and the image part of the post, respectively. $\mathcal{C}_n$ represents the comments (including shares and replies) of the post and y_n is the ground truth label on the fauxtography of P_n.

Given the above definitions, the goal of fauxtography detection is to classify each image-centric post into one of the two categories (i.e., fauxtography or not). Formally, for $P_n, 1 \leq n \leq N$, our goal is to find:

$$\arg\max_{\tilde{y_n}} Pr(\tilde{y_n} = y_n | P_n), \ \forall 1 \leq n \leq N \tag{1}$$

where $\tilde{y_n}$ denotes the estimated label for P_n.

Please note that the fauxtography detection problem is *not equivalent to "fake image"* detection [13, 16], which only asserts whether the visual content of the image is manipulated or not. For example, Figure 3 shows two identical images. The image itself is fake (i.e., it is created with photoshop), and should be classified as *fake* by the "fake image" detection algorithm. However, in the problem of fauxtography detection, posts with the same image could be classified into completely different categories when the image is accompanied by different claims as shown in 3(a) and 3(b). Also, fauxtography detection *is not equivalent to "false claim"* detection, which only focuses on checking the truthfulness of textual claims [35, 33]. The fauxtography detection requires a holistic analysis of the image and its associated context, which is a new research problem that has not been well addressed by current solutions.

(a) Fauxtography (b) Non-Fauxtography

Fig. 3: Example of Fauxtography and Non-fauxtography

4 Solution

In this section, we present the FauxWard framework to address the fauxtography detection problem formulated above. The FauxWard scheme is a graph convolutional neural network approach that leverages i) the topological characteristics underlying the user comment network of a social media post, and ii) the linguistic and semantic comment information extracted from the user comments. An overview of the FauxWard framework is shown in Figure 4. The FauxWard framework contains three major components: i) a *User Comment Network Construction* module that constructs the user comment network from the reply relationship of the comments associated with a post; ii) a *Comment Node Attribute Extraction* module that extracts the complex information of the comment node with a set of linguistic and semantic attributes from each user comment; iii) a *GCNN Detection* module that jointly leverages the topological characteristics of the user comment network and the comment node information to classify the fauxtography posts through a principled graph convolutional neural network (GCNN) framework. We will discuss the details below.

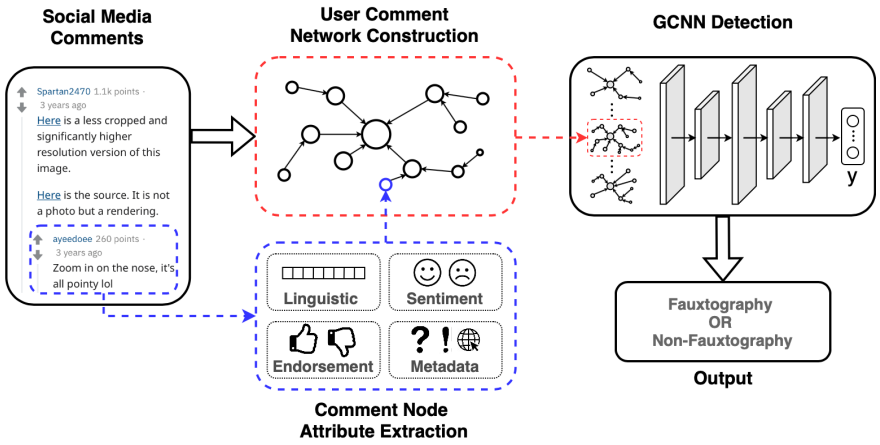


Fig. 4: Overview of the FauxWard Framework

4.1 User Comment Network Construction

We first observe that fauxtography and non-fauxtography posts are different in terms of the topological structure of the user comment network (e.g., the length of a comment thread, the number of replies) and semantic features of user comments (e.g., emotion and polarity of user feedback). For example, we found that users are more likely to use comments to show their negative attitude towards fauxtography posts (e.g., "Aka, fake", "wimpy"). These comments appear to be less attractive to other users for discussion, which often result in a large amount of single-comment threads. In contrast, non-fauxtography posts often get more engagement from social media users. To effectively capture the topological characteristics and semantic features of user comments, we model the comments of a social media post as a directed graph. Specifically, we first define a few key terms in our model.

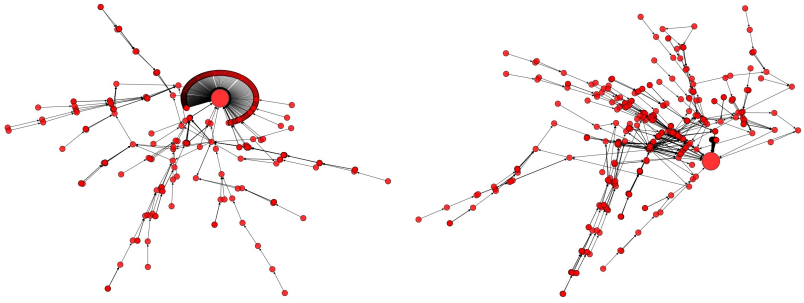
Definition 4 User Comment Network G: the user comment network $\mathbf{G}$ of an image-centric post is constructed as a directed graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, where $\mathbf{V}$ is a set of nodes, and $\mathbf{E}$ is a set of edges indicating the reply relationship between each pair of nodes. In particular, we define a *source node* $v_0 \in \mathbf{V}$ to denote the content of the original social media post and other *comment nodes* (i.e., $v_i \in \mathbf{V}, i \neq 0$) in the user comment network to represent the comments a post receives. We also define the *edge* e_{v_i,v_j} between two nodes v_i and v_j to denote the reply direction from comment v_j to comment v_i.

Definition 5 Adjacency Matrix A**:** we also define an adjacency matrix $A \in \mathbb{R}^{V \times V}$ to record the edges between any pair of the comment nodes. Specifically, for all pair of nodes $v_i, v_j \in \mathbf{V}$, $A_{i,j} = 1$ if there is an edge e_{v_i,v_j} between node v_i and v_j, otherwise $A_{i,j} = 0$.

Figure 5 shows an example of the user comment network of a fauxtography and a non-fauxtography post. We observe that fauxtography posts often receive a large number of comments that directly reply to the post. In contrast, non-fauxtography posts often attract more subsequent discussion in the form of replies to a comment.

4.2 Comment Node Attribute Extraction

We observe that the user comments often contain valuable information (e.g., the vocabulary used, the emotion and polarity reflected, and the endorsement or feedback from other users) in distinguishing fauxtography and non-fauxtography posts. For example, the fauxtography post in Figure 1(a) is likely to be debunked by a comment, "Fake image! It is super easy to photoshop", and such a debunking comment is also likely to be appreciated and endorsed by other users in the form of like/dislike or retweets. Such a debunking comment can be captured by a comment node with negative polarity and high endorsement in the user comment network. Therefore, we extract a set of key features based on the empirical observation of user comments, and incorporate them into the structured user comment network constructed in Section 4.1 to identify fauxtography posts. In particular, we



(a) Fauxtography (b) Non-Fauxtography

Fig. 5: Examples of the User Comment Network for Fauxtography and Non-fauxtography Posts. The source node is denoted with large size.

focus on a set of diversified comment node attributes (i.e., linguistic (L_v), sentiment (S_v), endorsement (E_v), and metadata (M_v)) in order to learn and represent the complex information embedded in each comment node v. We elaborate each comment node attribute below.

Definition 6 Linguistic Attribute L_v: we define the linguistic attribute $L_v \in \mathbb{R}^{1\times K_L}$ of a comment node v as a vector representation to represent the vocabulary used in each comment network.

An example of the the linguistic attribute is shown in Figure 6. We observe that vocabulary used in the comments of fauxtography and non-fauxtography posts are different to some extent. In particular, (a) and (b) show the word clouds of comments in each post category. We note that image verity related words (e.g., "photoshop", "photo", "fake") appears more frequently in fauxtography posts. In contrast, comments in non-fauxtography posts contain more general news topics (e.g., "treason", "Christmas", "ISIS").

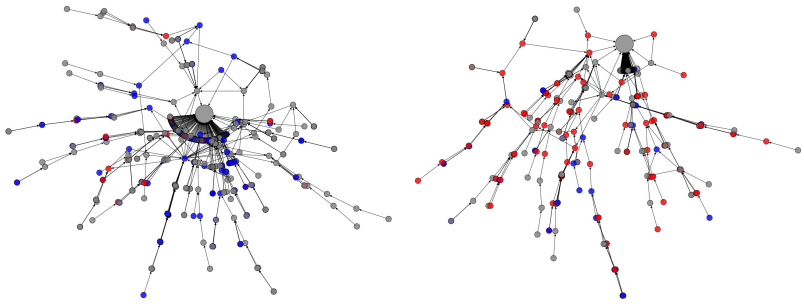
(a) Fauxtography (b) Non-Fauxtography

Fig. 6: Word Cloud

Definition 7 Sentiment Attribute S_v: we define the sentiment attribute $S_v \in [-1.0, 1.0]$ of each comment node v to be the polarity score to indicate the sentiment

in each comment network. Specifically, a positive polarity score (i.e., $S_v > 0$) indicates a positive sentiment, and vice versa.

Figure 7 shows an example of the sentiment attribute in the user comment network of two social media posts. We observe that comments in the fauxtography posts often contain more negative "echo chambers" (i.e., consecutive comments of negative sentiment) that indicate debunk and arguments between users, while the comment sentiments in non-fauxtography posts often appear to be more positive that reflect agreements from users.



(a) Fauxtography (b) Non-Fauxtography

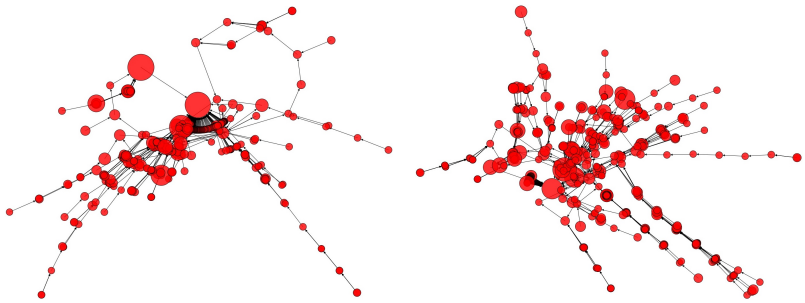
Fig. 7: Illustration of the Sentiment Attribute. The color of each comment node indicates the sentiment attribute of the corresponding comment, i.e., red - positive sentiment ($polarity \geq 0.5$), blue - negative sentiment ($polarity \leq -0.5$), grey - neutral sentiment ($0.5 < polarity < 0.5$).

Definition 8 Endorsement Attribute E_v: we define the endorsement attribute $E_v \in \mathbb{R}$ as the number of aggregated endorsement a comment receives from other users. Specifically, E_v equals to the number of likes - the number of dislikes for Reddit, and E_v equals to the sum of the number of likes and the number of retweets for Twitter.

Figure 8 shows an example of the endorsement attribute in the user comment network. We observe that there are a few "hub" comments in the fauxtography post that receives a large amount of support (i.e., endorsement) from other users. Such "hub" comments are often the ones that directly debunk the fauxtography in the post and thus receives support from users sharing similar points of view. In contrast, the endorsement attribute of comments in non-fauxtography posts appears to be more diversified as users often pay more attention to the content beyond the truthfulness of the image in those scenarios.

Definition 9 Metadata Attribute M_v: we define the metadata attribute $M_v \in \mathbb{R}^{1 \times K_M}$ as a set of metadata features extracted from each comment node v.

These metadata features are often shown direct correlations to characterizing the fauxtography posts. For example, we observe that fauxtography posts often debunked by comments containing URLs that link to the original image or the true



(a) Fauxtography (b) Non-Fauxtography

Fig. 8: Illustration of the Endorsement Attribute. The size of each comment node indicates the endorsement attribute (i.e., the number of aggregated likes) of the corresponding comment.

story associated with the image. We also observe that comments of fauxtography posts often contain many verity-related (e.g., "fake", "false alarm") or image-related words (e.g., "photo", "photoshop"). A summary of the extracted metadata features is listed in Table 1.

Table 1: Metadata Attribute

Feature	**Description**
`Word Count`	Number of words in a comment
`Verity Terms`	Number of verity-related terms in a comment
`Image Terms`	Number of image-related terms in a comment
`Question Marks`	Number of question marks in a comment
`Exclamation Marks`	Number of exclamation mark in a comment
`URLs`	Number of URLs in a comment

Finally, we define the node feature vector to represent the comment node attributes that contain the key characteristics of each user comment.

Definition 10 Node Feature Vector F_v: the node feature vector F_v for a comment node v is defined as $F_v = [L_v, S_v, E_v, M_v]$, s.t. $F_v \in \mathbb{R}^{1\times K}$ $\forall$ $v \in \mathbf{V}$ and K is the sum of the dimensions of node attributes. We denote the node feature matrix $F \in \mathbb{R}^{V\times K}$ as the matrix that stores the feature vectors for all nodes in the user comment networks.

4.3 GCNN Detection

In FauxWard, we model the fauxtography detection task as a graph classification problem and develop a novel graph convolutional neural network approach to solve it. A key challenge of our graph convolutional neural network design lies in effectively characterizing and encoding the user comment networks defined in

Section 4.1 and 4.2. On one hand, the user comment networks (Definition 4) are different in terms of their topological features (e.g., different sizes and structures of the user comment networks). On the other hand, each comment node (Definition 10) consists of diversified node attributes with distinct linguistic and semantic representations. In light of such a challenge, we design a graph convolution neural network model that jointly leverages i) the topological characteristic of the user comment networks, and ii) the rich linguistic and semantic attributes of the comments to detect fauxtography posts. To address the topological challenge in the user comment networks, we design a cluster-based pooling layer in the GCNN framework that first clusters neighboring nodes of various sizes based on their node embeddings, and reconstructs the input graphs to the next graph convolutional layer in our model. In addition, we take advantage of the vector representation of each comment node and encode the diversified comment node attributes into the user comment network to classify fauxtography posts. We present the details of our approach below.

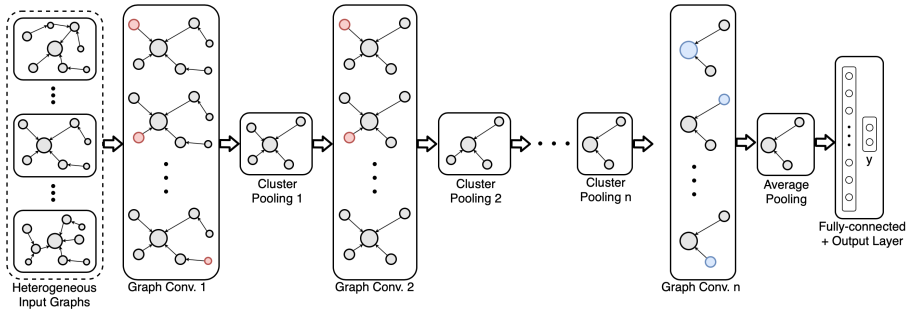


Fig. 9: The architecture of the GCNN Detection Module

An overview of the architecture of the GCNN detection framework is summarized in Figure 9. Let $\mathcal{G} = \{\mathbf{G_1}, \mathbf{G_2}, \cdots, \mathbf{G_N}\}$ be a collection of image-centric social media posts, where $\mathbf{G_n}$, $1 \leq n \leq N$, is the user comment network of post P_n (as defined in Section 4.1). y_n and $\tilde{y_n}$ are the corresponding ground truth and estimated labels of the post, respectively. For a given set of N social media posts $\mathcal{P} = \{(\mathbf{G_1}, y_1), (\mathbf{G_2}, y_2), \cdots, (\mathbf{G_N}, y_N)\}$, we aim to find:

$$\arg\max_{\tilde{y_n}} Pr(\tilde{y_n} = y_n | \mathbf{G_n}), \ \forall \ 1 \leq n \leq N \tag{2}$$

A key challenge in our problem is to effectively extract topological features of user comment network with various sizes (i.e., different number of nodes and edges). To address the challenge, we first define a key concept as follows.

Definition 11 Unified User Comment Node Space $\mathcal{V}$: the union of all comment nodes in the collection of posts. Formally, $\mathcal{V} = \bigcup_{n=1}^{N} \mathbf{V_n} \ \forall \ 1 \leq n \leq N$, where $\mathbf{V_n}$ is the set of nodes in user comment network $\mathbf{G_n}$. The size of $\mathcal{V}$ is denoted as V. An edge between any pair of comment nodes in $\mathcal{V}$ can be recorded using the corresponding adjacency matrix as defined in Definition 5.

Then, any user comment network $\mathbf{G} \in \mathcal{G}$ can be represented as (A, F) where $A \in \mathbb{R}^{V \times V}$ is the adjacency matrix and $F \in \mathbb{R}^{V \times K}$ is the node feature matrix with respect to the unified user comment node space $\mathcal{V}$. However, a direct representation of the user comment network $\mathbf{G}$ with respect to $\mathcal{V}$ will result in a large and sparse adjacency matrix A. Therefore, we adopt the "block diagonal adjacency matrix" strategy to effectively handle the sparse and various sized user comment networks. Formally, let $\bar{A}_1, \bar{A}_2, \cdots, \bar{A}_M$ be the non-empty adjacency matrices (i.e., adjacency matrices without empty rows and columns) for input graphs $\mathbf{G_1}, \mathbf{G_2}, \cdots, \mathbf{G_M}$, the block diagonal adjacency matrix A_{diag} for M input graphs is defined as:

$$A_{diag} = \begin{bmatrix} \bar{A}_1 & & & \\ & \bar{A}_2 & & \\ & & \ddots & \\ & & & \bar{A}_M \end{bmatrix} \tag{3}$$

We then perform sparse matrix multiplication with respect to the A_{diag} to efficiently train the model with batch-wise training.

Next, we adopt the recursive neighborhood aggregation (or "message-passing") strategy in the graph convolutional layer as follows:

$$H^{(k)} = f(A^{(k-1)}, H^{(k-1)}, W^{k-1}) \tag{4}$$

where $A^{(k-1)}$ and $H^{(k-1)}$ are the input node adjacency matrix and node feature matrix at the k^{th} layer of the GNN, respectively. W^k is the trainable weighting parameters and $f(\cdot)$ is the message propagation function. In particular, we initialize the graph convolutional neural network with the user comment networks we constructed in Section 4.1 (i.e., $A^{(0)} = A$), and the set of comment node attributes we extracted in Section 4.2 (i.e., $H^{(0)} = F$). Formally,

$$H^{(1)} = f(A, F, W^0) \tag{5}$$

To aggregate node information in the GCNN framework, we apply graph convolutional layer to the neighbor nodes $\mathcal{N}(v)$ of each node v in the graph, and use the rectified linear unit (ReLU) as the activation function $\sigma(\cdot)$. However, the user comment networks for social media posts often appear to be large and various sized in terms of the number of nodes and edges, which will result in a large and sparse adjacency matrix and cause the potential gradient vanishing problem. To this end, we applied the first-order approximation of localized spectral filters on graph convolutional layer [18] with added self-connection of the adjacency matrix. Formally, the updated adjacency matrix in each graph convolutional layer is formulated as:

$$\tilde{A}^{(k)} = \hat{D}^{(k)^{-\frac{1}{2}}} \hat{A}^{(k)} \hat{D}^{(k)^{-\frac{1}{2}}} \tag{6}$$

where $\hat{A}^{(k)} = I + A^{(k)}$ is the adjacency matrix with added self-loops and I is the identity matrix. $\hat{D}$ is the diagonal degree matrix where $\hat{D}_{ii} = \sum_j \hat{A}_{ij}$. Formally, the k^{th} graph convolutional layer is defined as:

$$\begin{aligned} H^{(k)} &= f\left(\tilde{A}^{(k-1)}, H^{(k-1)}, W^{k-1}\right) \\ &= \sigma\left(\hat{D}^{(k-1)^{-\frac{1}{2}}} \hat{A}^{(k-1)} \hat{D}^{(k-1)^{-\frac{1}{2}}} H^{(k-1)} W^{(k-1)}\right) \end{aligned} \tag{7}$$

In addition, we add a cluster-based pooling layer between the graph convolutional layers to coarsen the graph and efficiently learn the graph representation through the GCNN framework [48]. The cluster-based pooling layer first assigns neighboring nodes into clusters according to node embeddings learned from the previous graph convolutional layer and learns a representation for each cluster that is the input of the next graph convolutional layer. Let $C^{(k)}$ be the clustering matrix after the k^{th} graph convolutional layer. We update the adjacency matrix $\tilde{A}^{(k)}$ and node feature matrix $H^{(k)}$ as follows:

$$\tilde{A}^{(k)} = C^{(k-1)^T} \tilde{A}^{(k-1)} C^{(k-1)} \tag{8}$$

$$H^{(k)} = C^{(k-1)^T} f\left(\tilde{A}^{(k-1)}, H^{(k-1)}, W^{k-1}\right) \tag{9}$$

In this way, we can efficiently extract and preserve the topological features of local substructure (i.e., clusters) in the user comment network. Moreover, such a clustering design of the GCNN can also help to effectively extract and aggregate node information in the user comment network and high-level graph representations, especially for the posts with a large number of comments [48].

Finally, we use mean pooling as the readout layer to summarize the hidden graph representation before the fully-connected layer. A softmax layer is the last layer to output the binary classification results. We adopt the Adaptive Moment Estimation (Adam) optimizer [17] to train the graph neural network and minimize the cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{n=1}^{N} \left(y_n \log \tilde{y_n} + (1 - y_n) \log(1 - \tilde{y_n})\right) \tag{10}$$

5 Data

In this section, we describe the real-world dataset collected from the leading online social media platform Reddit[4] and Twitter [5]. Reddit, self-described as "front page of the Internet", is a popular news aggregation site [25] where massive fresh internet content is constantly shared and commented on by its users. As of October 2019, Reddit has 430 million monthly active users, 199 million posts, and 1.7 billion comments [6]. Twitter is a global micro-blogging platform hosting 330 million active users and 500 million visitors each month [4].

We observe that both Reddit and Twitter have a huge amount of posts that are image-based. It is challenging to collect ground-truth labels for fauxtography posts on these media platforms. To address such a challenge, we first collect verified fauxtography images from 3 independent fact-checkers (i.e., snopes.com, factcheck.org, truthorfiction.com) in a similar way as [20]. The ground truth labels are initially decided based on the majority vote of these fact-checkers. We then assign three independent annotators to manually verify the label of each post using databases

[4]https://www.reddit.com/

[5]https://www.twitter.com/

of historical facts and Google search. The dates of the fack-checked fauxtography images range from January 2010 to October 2019.

Given the labeled images, we perform a reverse search using the Google Vision API [6] to identify the original web URLs that contain the image. If the URL points to a social media post on Reddit or Twitter, we crawl the post and its comment threads using a crawler script we developed. We summarize the real-world dataset used for evaluation in Table 2. We observe that there is a non-trivial amount of the fauxtography posts (10.6% in Reddit and 11.3% in Twitter) actually contain real images. This observation validates the unique challenge of fauxtography detection, where real images can also be leveraged to convey misleading messages.

Table 2: Data Trace Statistics

Data Trace	**Reddit**	**Twitter**
Number of Fact-checked Posts	220	438
Number of Fauxtography	179	378
Number of Fauxtography with Real Images	19	43
Number of Comments	64,183	1,125,622
Number of Distinct Users	40,806	447,897

We observe that the social media posts collected from the fact-checking websites are often biased (e.g., there are more fauxtography posts than non-fauxtography ones). To mitigate this issue, we design a new data collection strategy. In particular, for each fact-checked post found on Reddit, we collect 20 posts immediately ahead and behind the post in the same subreddit (i.e., sub-forum under the same topic) on the same day so that the collected posts reflect the actual ratio of fauxtography on that subreddit. Similarly, for each fact-checked post found on Twitter, we randomly sample 20 tweets that published within the same day as the fact-checked post. Removing invalid posts that do not contain image content, we finally obtain the datasets of 2780 and 2875 posts for Reddit and Twitter respectively, and assume all the posts are not fauxtography.

Table 3: Supplementary Data Trace Statistics

Data Trace	**Reddit**	**Twitter**
Number of Posts	2,780	2,875
Number of Comments	395,964	2,205,635
Number of Distinct Users	141,034	912,956

6 Evaluation

In this section, we evaluate the FauxWard scheme using the real-world online social media datasets described in the previous section. We compare the detection

[6] https://cloud.google.com/vision/

performance of FauxWard with state-of-the-art baseline solutions as well as the FauxBuster solution in our previous work. The results show that the FauxWard scheme significantly outperforms all compared baselines in terms of detection accuracy and efficiency.

6.1 Baselines

We compare the FauxWard with state-of-the-art baselines in fake image detection and fake claim detection.

- *FauxBuster:* A random walk based network embedding solution particularly designed to detect fauxtography posts on social media using user comments [54].
- *Fake Image:* A feature engineering based approach to detect fake images on social media using a decision tree classifier [13].
- *SAME:* A deep learning based framework to detect multimodal fake news by leveraging features extracted from the news content and the sentiment of user comments [10].
- *Truth Discovery:* A representative fact-checking scheme to detect misinformation among conflicting text-based claims on social media [49].
- *Fake News:* A linguistic-based approach to identify fake news by extracting lexical and syntactic features from the news statement [32].

Please note that we carefully tune parameters in each baseline model to achieve its optimal performance for a fair comparison with the proposed scheme. In particular, for all of the compared methods, we use 80% of the evaluation dataset as the training set and tune parameters based on the 5-fold cross-validation performance on the training set.

6.2 Detection Effectiveness

In the first set of experiments, we evaluate the detection effectiveness of FauxWard and the aforementioned baseline solutions. In particular, we adopt the commonly used metrics for binary classification evaluation, including *Accuracy*, *Precision*, *Recall*, and *F1-score*. The results are summarized in Table 4 and Table 5. We observe that FauxWard significantly outperforms all the baseline schemes. In particular, on the Reddit dataset, FauxWard achieves a performance gain of 15.1%, 9.1%, 14.1%, 18.4%, and 32.5% in terms of F1 score compared to the *FauxBuster*, *Fake Image*, *SAME*, *Truth Discovery*, and *Fake News* baselines, respectively. On the Twitter dataset, FauxWard outperforms the *FauxBuster*, *Fake Image*, *SAME*, *Truth Discovery*, and *Fake News* baselines by 6.7%, 18.9%, 9.39%, 19.3%, and 36.9% in terms of F1 score, respectively.

We observe that our FauxWard scheme did outperform the previous FauxBuster scheme. This is because FauxBuster takes users' comments as a whole document to extract the linguistic features (i.e., document embedding), which underexplores the topological patterns underlying the user comment network during the representation learning process. In contrast, FauxWard aggregates the linguistic attribute as well as semantic attributes of each comment through a GCNN framework to preserve such topological patterns. Moreover, the Fake Image baseline

also fails to detect fauxtography posts effectively because it only focuses on image features but does not put them into the context of the textual claims. Therefore, it is not robust against the fauxtography posts containing real images. In addition, the Truth Discovery and Fake News schemes only consider whether the textual claims are truthful or not. This leads to nontrivial false negatives in the results (i.e., fauxtography with fake images but truthful textual claims). In contrast, FauxWard is explicitly developed to detect the fauxtography posts by considering both the image and textual claim together with the message that they collectively express. The results again demonstrate that existing image forgery detectors and fact-checkers cannot effectively solve the fauxtography detection problem.

Table 4: Classification Accuracy for All Schemes (Reddit)

Algorithm	Accuracy	Precision	Recall	F1-Score
FauxWard	**0.7536**	**0.7895**	**0.7692**	**0.7793**
FauxBuster	0.6812	0.6216	0.7419	0.6765
Fake Image	0.6522	0.6667	0.7692	0.7143
SAME	0.6232	0.6364	0.7368	0.6829
Truth Discovery	0.6087	0.6047	0.7222	0.6582
Fake News	0.5942	0.6061	0.5714	0.5882

Table 5: Classification Accuracy for All Schemes (Twitter)

Algorithm	Accuracy	Precision	Recall	F1-Score
FauxWard	**0.7109**	**0.7015**	**0.7344**	**0.7176**
FauxBuster	0.6797	0.6885	0.6562	0.6720
Fake Image	0.6094	0.6129	0.5938	0.6032
SAME	0.6641	0.6721	0.6406	0.6560
Truth Discovery	0.6042	0.6056	0.5972	0.6014
Fake News	0.5312	0.5323	0.5156	0.5238

We also plot the Receiver Operating Characteristics (ROC) curve of all methods in Figure 10 and 11. The ROC curve focuses on the trade-off between the False Positive Rate (FPR) and the True Positive Rate (TPR) by adjusting the classification threshold of each method. We observe that the FauxWard scheme continues to outperform all baselines in terms of the Area Under the Curve (AUC) score on both the Reddit and Twitter datasets. This demonstrates that FauxWard is also robust against the classification threshold.

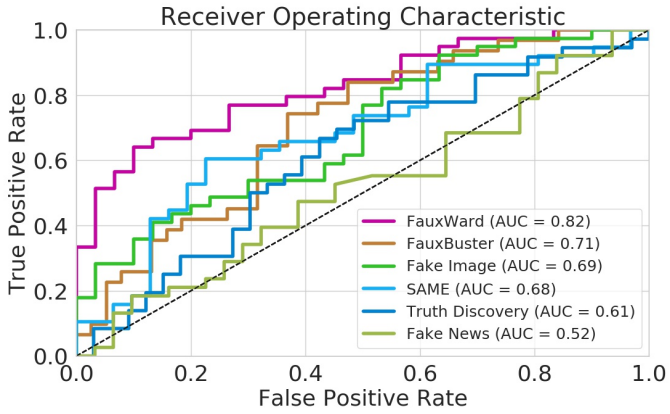


Fig. 10: ROC Curve of All Schemes (Reddit)

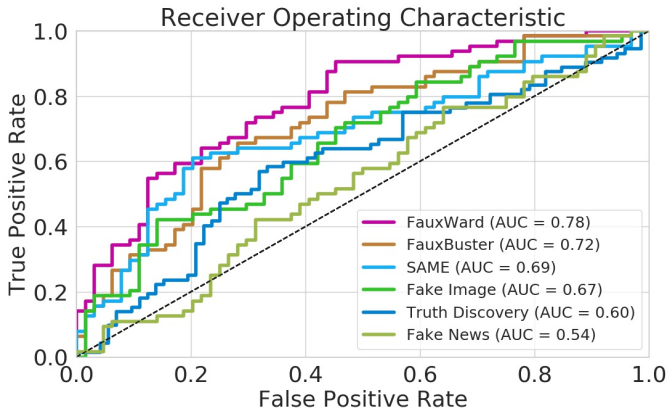


Fig. 11: ROC Curve of All Schemes (Twitter)

6.3 FauxWard versus Humans

In the second set of experiments, we compare the performance of FauxWard with humans. We invite three independent human annotators (denoted as A1, A2, and A3) to manually annotate whether they believe the image is misleading or not. We randomly pick a total of 70 image-based social media posts (38 of which are fauxtography) from the test dataset for them to annotate. Please note that these human annotators are different from the ground-truth annotators in that they have not seen those posts before and are *not allowed* to have access to any external data source (e.g., Google Search, fact-checking websites) to validate their annotations. Furthermore, the annotators were asked to skip the posts that they happen to know the ground truth.

First, we asked these participants to annotate image-centric posts by only showing them the image and the text of a post, which contains the same informa-

tion a user receives from the social media feed. Next, we asked the participants to annotate the same set of posts but also showed them the comments of each post. We design such an experiment process to evaluate whether the user comments from social media would assist humans in identifying fauxtography posts. Table 6 shows the performance of each individual annotator and their aggregated results based on the majority voting (i.e., "overall without comments" and "overall with comments"). We observe that FauxWard consistently outperforms the human annotators even if they are allowed to access the comments from social media users. A possible reason is that humans are often easily affected by their subjectivity and emotions. For example, we found all of the three human annotators fail to identify a fauxtography that shows an injured koala was rescued from the Australian bushfire in 2020 (the fact is that the koala was rescued from another event in 2015). In addition, we also observe that human performance is boosted significantly when the user comments are available to the annotators. Such an observation verifies our assumption on the usefulness of user comments in detecting fauxtography posts. Moreover, we also observe that the fauxtography posts with real images are more likely to convince the human annotators to believe the content of the post. This again demonstrates that the fauxtography detection problem is more challenging than merely detecting "fake images".

Table 6: FauxWard vs. Human Performance

	Accuracy	F1	FPR	FNR
FauxWard	**0.7571**	**0.7733**	**0.2500**	**0.2368**
A1 without comments	0.3714	0.2667	0.4193	0.7949
A1 with comments	0.5857	0.5915	0.3548	0.4615
A2 without comments	0.3571	0.2105	0.3871	0.8462
A2 with comments	0.5714	0.5588	0.3637	0.4864
A3 without comments	0.4143	0.3051	0.3548	0.7692
A3 with comments	0.6286	0.6176	0.3030	0.4324
Overall without comments	0.3857	0.2456	0.3548	0.8205
Overall with comments	0.6143	0.6197	0.3226	0.4359

* "Overall" denotes the majority vote of the three annotators.

6.4 Detection Time

In the last set of experiments, we evaluate the detection performance of the FauxWard scheme against the time after a social media post is originally published. In particular, we limit the time window of the data collected from 1 hour to 5 days and only include user comments posted within the specific time window. The results are shown in Figure 12 and Figure 13. We observe that the performance of the FauxWard scheme improves as the time increases, and more input data is available. In the meantime, FauxWard consistently outperforms all the baselines

on both datasets. More importantly, FauxWard achieves a significant performance gain when the time window is short (e.g., 1 hour), which is necessary to curb the spread of misinformation on social media in a timely manner.

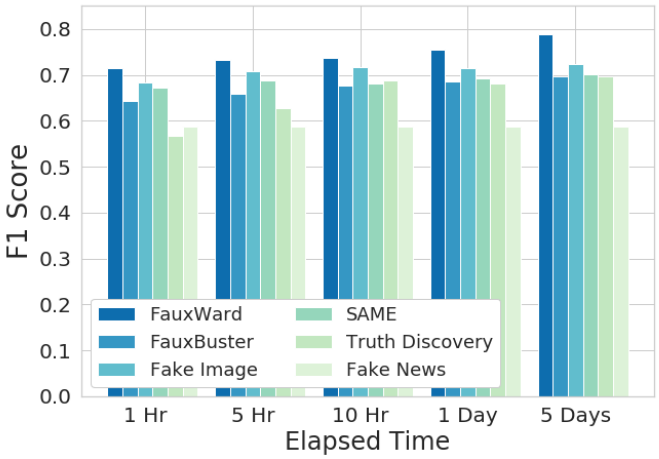


Fig. 12: Elapsed Time vs. Performance (Reddit)

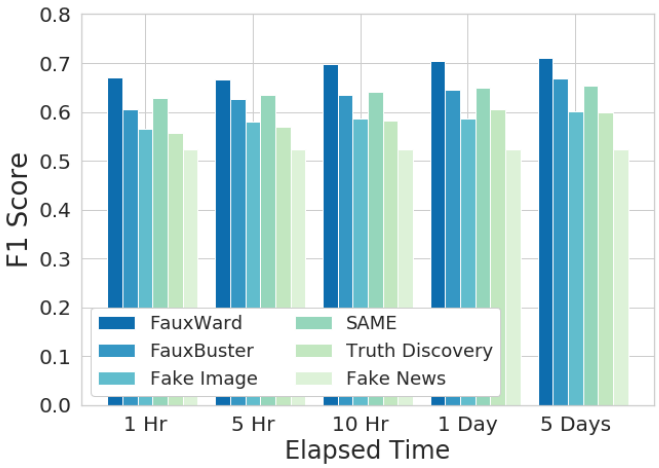


Fig. 13: Elapsed Time vs. Performance (Twitter)

7 Conclusion

In this paper, we develop a graph convolutional neural network approach, FauxWard, to address the fauxtography detection problem in image-based social media posts. FauxWard leverages the "wisdom of the crowd" by exploring the valuable information from the user comments on social media and encoding the linguistic, sentiment, endorsement, and metadata attributes into a graph neural network framework. The FauxWard scheme does not directly analyze the content of image-centric posts and is robust against sophisticated content creators who are good at crafting and spreading the misleading fauxtography content on social media. We evaluate the FauxWard scheme using two real-world datasets collected from Reddit and Twitter. The results demonstrate that FauxWard can effectively detect the fauxtography posts on social media and outperforms the state-of-the-art baselines and human annotators in terms of accuracy and F1 score.

Acknowledgment

This research is supported in part by the National Science Foundation under Grant No. CNS-1845639, CNS-1831669, Army Research Office under Grant W911NF-17-1-0409. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

1. (Accessed: 2018-08-07) Facebook and google's war with fake news heats up. https://moneyish.com/ish/google-joins-facebook-in-declaring-war-on-fake-news/
2. (Accessed: 2018-08-07) How twitter's expanded images increase clicks, retweets and favorites. https://www.fastcompany.com/3022116/what-twitters-expanded-images-mean-for-clicks-retweets-and-favorites/
3. (Accessed: 2018-08-07) Social media engagement – statistics and trends. https://www.invespcro.com/blog/social-media-engagement/
4. (Accessed: 2020-01-02) 60 incredible and interesting twitter stats and statistics. https://www.brandwatch.com/blog/twitter-stats-and-statistics/
5. (Accessed: 2020-01-02) Photo fuels spread of fake news. https://www.wired.com/2016/12/photos-fuel-spread-fake-news/
6. (Accessed: 2020-01-02) Reddit's 2019 year in review. https://redditblog.com/2019/12/04/reddits-2019-year-in-review/
7. Bayar B, Stamm MC (2016) A deep learning approach to universal image manipulation detection using a new convolutional layer. In: Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security, pp 5–10
8. Chen J, Ma T, Xiao C (2018) Fastgcn: fast learning with graph convolutional networks via importance sampling. arXiv preprint arXiv:180110247

9. Cooper SD (2007) A concise history of the fauxtography blogstorm in the 2006 lebanon war. The American Communication Journal 9
10. Cui L, Wang S, Lee D (2019) Same: sentiment-aware multi-modal embedding for detecting fake news. In: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp 41–48
11. Fout A, Byrd J, Shariat B, Ben-Hur A (2017) Protein interface prediction using graph convolutional networks. In: Advances in Neural Information Processing Systems, pp 6530–6539
12. Fridrich AJ, Soukal BD, Lukáš AJ (2003) Detection of copy-move forgery in digital images. In: in Proceedings of Digital Forensic Research Workshop, Citeseer
13. Gupta A, Lamba H, Kumaraguru P, Joshi A (2013) Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In: Proceedings of the 22nd international conference on World Wide Web, pp 729–736
14. Hamaguchi T, Oiwa H, Shimbo M, Matsumoto Y (2017) Knowledge transfer for out-of-knowledge-base entities: A graph neural network approach. arXiv preprint arXiv:170605674
15. Hamilton W, Ying Z, Leskovec J (2017) Inductive representation learning on large graphs. In: Advances in Neural Information Processing Systems, pp 1024–1034
16. Huynh-Kha T, Le-Tien T, Ha-Viet-Uyen S, Huynh-Van K, Luong M (2016) A robust algorithm of forgery detection in copy-move and spliced images. IJACSA) International Journal of Advanced Computer Science and Applications 7(3)
17. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980
18. Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:160902907
19. Kwak H, Lee C, Park H, Moon S (2010) What is twitter, a social network or a news media? In: Proceedings of the 19th international conference on World wide web, pp 591–600
20. Lazer DM, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, Metzger MJ, Nyhan B, Pennycook G, Rothschild D, et al. (2018) The science of fake news. Science 359(6380):1094–1096
21. Li R, Wang S, Zhu F, Huang J (2018) Adaptive graph convolutional neural networks. In: Thirty-Second AAAI Conference on Artificial Intelligence
22. Matern F, Riess C, Stamminger M (2019) Gradient-based illumination description for image forgery detection. IEEE Transactions on Information Forensics and Security 15:1303–1317
23. Nguyen TH, Grishman R (2018) Graph convolutional networks with argument-aware pooling for event detection. In: Thirty-second AAAI conference on artificial intelligence
24. Pérez-Rosas V, Kleinberg B, Lefevre A, Mihalcea R (2017) Automatic detection of fake news. arXiv preprint arXiv:170807104
25. Priya S, Sequeira R, Chandra J, Dandapat SK (2019) Where should one get news updates: Twitter or reddit. Online Social Networks and Media 9:17–29
26. Pun CM, Yuan XC, Bi XL (2015) Image forgery detection using adaptive oversegmentation and feature point matching. IEEE Transactions on Information

Forensics and Security 10(8):1705–1716
27. Qian F, Gong C, Sharma K, Liu Y (2018) Neural user response generator: Fake news detection with collective user intelligence. In: IJCAI, vol 18, pp 3834–3840
28. Rashid MT, Wang D (2020) Covidsens: a vision on reliable social sensing for covid-19. Artificial Intelligence Review pp 1–25
29. Schlichtkrull M, Kipf TN, Bloem P, Van Den Berg R, Titov I, Welling M (2018) Modeling relational data with graph convolutional networks. In: European Semantic Web Conference, Springer, pp 593–607
30. Shang L, Zhang DY, Wang M, Lai S, Wang D (2019) Towards reliable online clickbait video detection: A content-agnostic approach. Knowledge-Based Systems 182:104851
31. Shang L, Zhang DY, Wang M, Wang D (2019) Vulnercheck: a content-agnostic detector for online hatred-vulnerable videos. In: 2019 IEEE International Conference on Big Data (Big Data), IEEE, pp 573–582
32. Shu K, Sliva A, Wang S, Tang J, Liu H (2017) Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter 19(1):22–36
33. Vo N, Lee K (2018) The rise of guardians: Fact-checking url recommendation to combat fake news. arXiv preprint arXiv:180607516
34. Volkova S, Shaffer K, Jang JY, Hodas N (2017) Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp 647–653
35. Wang D, Kaplan L, Le H, Abdelzaher T (2012) On truth discovery in social sensing: A maximum likelihood estimation approach. In: Proc. ACM/IEEE 11th Int Information Processing in Sensor Networks (IPSN) Conf, pp 233–244, DOI 10.1109/IPSN.2012.6920960
36. Wang D, Abdelzaher T, Kaplan L, Aggarwal CC (2013) Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing applications. In: 2013 IEEE 33rd International Conference on Distributed Computing Systems, IEEE, pp 530–539
37. Wang D, Abdelzaher T, Kaplan L, Ganti R, Hu S, Liu H (2013) Exploitation of physical constraints for reliable social sensing. In: Real-Time Systems Symposium (RTSS), 2013 IEEE 34th, IEEE, pp 212–223
38. Wang D, Amin MT, Li S, Abdelzaher T, Kaplan L, Gu S, Pan C, Liu H, Aggarwal CC, Ganti R, et al. (2014) Using humans as sensors: an estimation-theoretic perspective. In: Information Processing in Sensor Networks, IPSN-14 Proceedings of the 13th International Symposium on, IEEE, pp 35–46
39. Wang D, Abdelzaher T, Kaplan L (2015) Social sensing: building reliable systems on unreliable data. Morgan Kaufmann
40. Wang D, Abdelzaher T, Kaplan L, Ganti R, Hu S, Liu H (2015) Reliable social sensing with physical constraints: analytic bounds and performance evaluation. Real-Time Systems 51(6):724–762
41. Wang D, Szymanski BK, Abdelzaher T, Ji H, Kaplan L (2018) The age of social sensing. arXiv preprint arXiv:180109116
42. Wang D, Szymanski BK, Abdelzaher T, Ji H, Kaplan L (2019) The age of social sensing. Computer 52(1):36–45

43. Wu Z, Pan S, Chen F, Long G, Zhang C, Yu PS (2019) A comprehensive survey on graph neural networks. arXiv preprint arXiv:190100596
44. Yang Y, Zheng L, Zhang J, Cui Q, Li Z, Yu PS (2018) Ti-cnn: Convolutional neural networks for fake news detection. arXiv preprint arXiv:180600749
45. Yao QQ, Perlmutter DD, Liu JZ (2017) What are shaping the ethical bottom line?: Identifying factors influencing young readers' acceptance of digital news photo alteration. Telematics and Informatics 34(1):124–132
46. Yin X, Han J, Yu PS (2008) Truth discovery with multiple conflicting information providers on the web. IEEE Transactions on Knowledge and Data Engineering 20(6):796–808, DOI 10.1109/TKDE.2007.190745
47. Ying R, He R, Chen K, Eksombatchai P, Hamilton WL, Leskovec J (2018) Graph convolutional neural networks for web-scale recommender systems. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, pp 974–983
48. Ying Z, You J, Morris C, Ren X, Hamilton W, Leskovec J (2018) Hierarchical graph representation learning with differentiable pooling. In: Advances in Neural Information Processing Systems, pp 4800–4810
49. Zhang D, Wang D, Vance N, Zhang Y, Mike S (2018) On scalable and robust truth discovery in big data social media sensing applications. IEEE Transactions on Big Data 5(2):195–208
50. Zhang D, Vance N, Wang D (2019) When social sensing meets edge computing: Vision and challenges. In: 2019 28th International Conference on Computer Communication and Networks (ICCCN), IEEE, pp 1–9
51. Zhang DY, Wang D, Zhang Y (2017) Constraint-aware dynamic truth discovery in big data social media sensing. In: Big Data (Big Data), 2017 IEEE International Conference on, IEEE, pp 57–66
52. Zhang DY, Badilla J, Zhang Y, Wang D (2018) Towards reliable missing truth discovery in online social media sensing applications. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, pp 143–150
53. Zhang DY, Li Q, Tong H, Badilla J, Zhang Y, Wang D (2018) Crowdsourcing-based copyright infringement detection in live video streams. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, pp 367–374
54. Zhang DY, Shang L, Geng B, Lai S, Li K, Zhu H, Amin MT, Wang D (2018) Fauxbuster: A content-free fauxtography detector using social media comments. In: 2018 IEEE International Conference on Big Data (Big Data), IEEE, pp 891–900
55. Zhang DY, Song L, Li Q, Zhang Y, Wang D (2018) Streamguard: A bayesian network approach to copyright infringement detection problem in large-scale live video sharing systems. In: 2018 IEEE International Conference on Big Data (Big Data), IEEE, pp 901–910
56. Zhang J, Cui L, Fu Y, Gouza FB (2018) Fake news detection with deep diffusive network model. arXiv preprint arXiv:180508751
57. Zhang Y, Dong X, Rashid MT, Shang L, Han J, Zhang D, Wang D (2020) Pqa-cnn: Towards perceptual quality assured single-image super-resolution in remote sensing. In: 2020 IEEE/ACM International Symposium on Quality of Service, IEEE

58. Zhang Y, Dong X, Shang L, Zhang D, Wang D (2020) A multi-modal graph neural network approach to traffic risk forecasting in smart urban sensing. In: The 17th Annual IEEE International Conference on Sensing, Communication and Networking, IEEE
59. Zhou J, Cui G, Zhang Z, Yang C, Liu Z, Wang L, Li C, Sun M (2018) Graph neural networks: A review of methods and applications. arXiv preprint arXiv:181208434
60. Zhou X, Zafarani R (2018) Fake news: A survey of research, detection methods, and opportunities. arXiv preprint arXiv:181200315