FauxBuster: A Content-free Fauxtography Detector Using Social Media Comments

Daniel (Yue) Zhang*, Lanyu Shang*, Biao Geng*, Shuyue Lai*, Ke Li*, Hongmin Zhu*, Md Tanvir Amin†, Dong Wang*
*Department of Computer Science and Engineering
University of Notre Dame, Notre Dame, IN, USA
{yzhang40, lshang, bgeng, slai1, kli5, hzhu6, dwang5}@nd.edu
†Google LLC, Mountain View, CA, USA
tanviramin@google.com

***Abstract*—With the increasing popularity of online social media (e.g., Facebook, Twitter, Reddit), the detection of misleading content on social media has become a critical undertaking. This paper focuses on an important but largely unsolved problem: detecting fauxtography (i.e., social media posts with misleading images). We found that the existing literature falls short in solving this problem. In particular, current solutions either focus on the detection of fake images or misinformed texts of a social media post. However, they cannot solve our problem because the detection of fauxtography depends not only on the truthfulness of the images and the texts but also on the information they *deliver together* on the posts. In this paper, we develop the FauxBuster, an end-to-end supervised learning scheme that can effectively track down fauxtography by exploring the valuable clues from user's comments of a post on social media. The FauxBuster is *content-free* in that it does not rely on the analysis of the actual content of the images, and hence is robust against malicious uploaders who can intentionally modify the presentation and description of the images. We evaluate FauxBuster on real-world data collected from two mainstream social media platforms - Reddit and Twitter. Results show that our scheme is both effective and efficient in addressing the fauxtography problem.**

I. Introduction

With the increasing acceptance of using social media as a daily source of news [1], [2], misinformation spread on social media has become a critical issue in recent years [3]. For example, the analysis and detection of falsified facts and rumors on online social media has been a hot topic in the past decade and various fact-checking schemes have been developed [4], [5]. More recently, major online sites (e.g., Facebook and Google) have launched worldwide campaigns to curb the spread of fake news [6]. 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 [7].

The fauxtography detection problem is motivated by the recent trend of image-centric content on social media [8]. For example, photos are found to be the most engaging type of content on Facebook where 87% of the posted photos have been clicked, liked or shared by its users [9]. Similarly on Twitter, tweets with images get 18% more clicks, 89% more likes, and 150% more retweets than tweets without images [10].

The prevalence of image-centric content on social media also opens the door for the severe propagation of misinformation [7], [8]. For example, fake images about sightings of creepy killer clowns have caused national hysteria in 2016 in the USA [1]. In this paper, we focus on a unique type of misinformation called "fauxtography" - an image together with its context (often the text associated with the image) that conveys the 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 Putin is pulling Obama's tie while in fact the image was edited and the claim itself is false. Image (b) claims a guy was waiting at the finish line to propose to his girlfriend who died during the Boston Marathon Bombing event. Though the image itself is a real photo (not edited), it is from a different event and the claim itself is false. Image (c) claims sea creatures are falling from the sky in China during a tornado event. While the claim itself is truthful [2], the image is misleading because it is edited to convey the wrong message that an octopus was falling from the sky, which never happened. Finally, image (d) claims a wildfire is happening in Tennessee. While both the image and text are real, it is misleading because the image was taken from an earlier event (wildfire in the Bitterroot National Forest) and used to exaggerate the severity of the fire [3]. In short, we treat all the above cases as fauxtography because the images and the associated texts together convey the misleading information.

Many solutions have been developed to fight against the misinformation spread from image-centric content on social media [11], [7]. A representative solution is called "image forgery detection" that can detect image editing including copy-and-move [12], splicing [13], and image-retouch [14]. However, this solution only focuses on the detection of "fake" images without considering the context (e.g., the texts associated with the image). Thus, it cannot be directly applied to our problem. For example, we observe that real

[1]https://www.theverge.com/2016/10/7/13191788/clown-attack-threats-2016-panic-hoax-debunked
[2]https://www.snopes.com/fact-check/octopus-fall-sky-china/
[3]http://www.westernhunter.com/Pages/Vol02Issue30/firecorrect.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 "Putin pulling Obama's tie.". Image (b) was titled " At the Boston bombing, a girl was running and her boyfriend was at the finish line waiting to propose but she died." Image (c) was titled "Sea creatures fall from the sky during powerful storm in China". Image (d) was titled "TENNESSEE: PHOTOS: National Guard is being brought in to city of Gatlinburg and beyond as massive wildfires force mandatory evacuations."

Figure 1: Examples of Fauxtography on Social Media

images can also convey misleading information that can not be easily detected (e.g., images (b) and (d) in Figure 1). Additionally, advances in photo editing and manipulation techniques have made it significantly easier to create fake imagery that can bypass the current detection systems. For example, the recent AI technique can automatically generate high-resolution "photographs" of humans and objects that are almost indistinguishable from the real ones [15]. Therefore, it becomes an increasingly challenging problem for content-based methods to detect fauxtography.

Several fact-checking (or "truth discovery") techniques have also been developed to assert the truthfulness of textual claims on social media and can effectively track down misinformation such as fake news and rumors [3], [5]. However, these techniques only focus on identifying the truthfulness of *texts* of the social media posts which is insufficient to address our problem. For example, we found many social media posts are composed of a truthful text but an exaggerating or irrelevant image to convey wrong messages and guide the viewers to misinterpret the event (e.g., images (c) and (d) in Figure 1) [8]. The nature of fauxtography detection problem indicates that any technique that asserts the truthfulness of *image or text alone* will not suffice to address this problem. A system that can effectively address the fauxtography detection problem of image-based content on social media has yet to be developed.

In this paper, we develop the FauxBuster, an end-to-end supervised learning scheme that can effectively track down fauxtography on online social media. FauxBuster adopts a content-free approach that does not analyze the content of the image itself but explores the characteristics of the comments from social media users on a post of interest. For example, users often express anger emotions or make sarcastic jokes when they observe fauxtography and focus on the topic of the post itself when they observe non-fauxtography. We also observe that the comments of a non-fauxtography post often receive more positive feedback (e.g., likes) and have more diversified discussion threads than the comments of a fauxtography post. In FauxBuster, we develop a principled framework to extract a set of valuable clues (e.g., network characteristics, linguistic cues, and metadata) from user's comments to characterize fauxtography using deep autoencoding and neural word embedding techniques. FauxBuster then integrates the extracted clues into a supervised learning framework to track down fauxtography effectively.

To the best of our knowledge, the FauxBuster is the first solution to address the fauxtography detection problem on online social media. The content-free nature of FauxBuster makes it robust against sophisticated uploaders who can intentionally modify the presentation and the description of the images because FauxBuster does not rely on the analysis of the actual content of the images (i.e., content-free). We evaluate the performance of FauxBuster on two mainstream social media platforms - Reddit and Twitter. The results show that our scheme is effective (with 25.6% higher F1 score than state-of-the-art image forgery detection baselines) and efficient (reaching 86.1% detection accuracy within one hour of the original post).

II. Related Work

A. Fauxtography

The term "fauxtography" was coined by Cooper *et al.* in the context of fake image spread during the 2006 Lebanon War [7]. Fauxtography was defined as "visual images, especially news photographs, which convey a questionable (or outright false) sense of the events they seem to depict". 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, and using special photography technique (e.g., wide-angle close-ups) to take images to exaggerate the event. The phenomenon of "fauxtography" has also been observed in social science but no practical solution has been developed [16], [11]. In this paper, we develop the FauxBuster, the first content-free solution dedicated to addressing the fauxtography detection problem on online social media.

B. Image Forgery Detection

Image forgery is closely related to our problem. A set of tools have been developed to detect image forgeries. For example, Huynh-Kha *et al.* developed an image forgery detection scheme that can detect whether an image is manually

edited by copy-move, splicing or both in the same image [13]. Bayar *et al.* developed a deep learning approach to detect the image manipulation using Convolutional Neural Networks [17]. Gupta *et al.* characterized the phenomenon of fake image propagation on Twitter during a disaster event and developed a supervised detection scheme [11]. 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, FauxBuster assumes the fauxtography detection must consider both images and their contexts under a holistic analytical framework.

C. Misinformation Detection

The spread of misinformation on online social media has received a significant amount of attention in recent years [5]. Yin *et al.* proposed the first fact-checking scheme *Truth Finder* that uses a Bayesian-based heuristic algorithm to combat misinformation from multiple conflicting data sources [3]. Wang *et al.* developed an estimation-maximization algorithm that identifies truthful online social media posts by explicitly considering the reliability of data sources [18]. Zhang *et al.* developed a dynamic truth discovery model to incorporate physical constraints and temporal dependencies into the detection of evolving truth [19]. 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 [5]. However, these solutions cannot apply to our problem because they only focus on the textual claims and cannot capture sophisticated fauxtography posts that covey misinformation using images. In contrast, FauxBuster effectively captures the misinformation from the images and their contexts by exploring the useful clues from the "wisdom of the crowd" (i.e., user comments on posts).

III. Problem Definition

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: 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.

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".

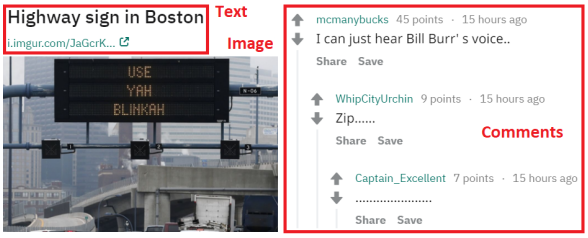


Figure 2: Example of an Image-centric Post on Reddit

Please note that the fauxtography detection problem is *not equivalent to "fake image"* detection [11], [13], which only asserts whether the visual content of the image is manipulated or not. Also, fauxtography detection *is not equivalent to "false claim"* detection, which only focuses on checking the truthfulness of textual claims [18], [5]. 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.

To formulate our problem, we assume a set of N posts $Post = \{P_1, P_2, ..., P_N\}$ from online social media. A post P_n, $1 \leq n \leq N$ is defined as a tuple: $P_n = (\mathcal{T}_n, \mathcal{I}_n, \mathcal{C}_n, z_n)$ 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 z_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-based 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:

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

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

IV. Solution

In this section, we present the FauxBuster scheme to address the fauxtography problem formulated above. FauxBuster consists of four major components: 1) a Comment Network Feature Extraction module to extract semantic and topological features from user's interactions on comments of a post, 2) a Linguistic Feature Extraction module to extract linguistic features in terms of paragraph embeddings from the comments, 3) a Metadata Feature Extraction module to extract auxiliary metadata information of the posts, and 4) a supervised classification algorithm to effectively identify the fauxtography posts given the above features. We discuss these components in detail below.

A. Comment Network Feature Extraction

The goal of the comment network feature extraction component is to extract the key features from the user's comments that are relevant to the fauxtography of a social media post. We have observed that the comments of users in fauxtography and non-fauxtography posts have very different

topological (e.g., depth of comment threads, number of replies) and semantic characteristics (e.g., emotions and polarity of user feedback). FauxBuster effectively captures both the topology and semantic features of fauxtography posts from the user's comments.

1) Building Comment Networks of Social Media Posts: We first define a *Comment Network* $\mathbf{G}$ for *each social media post* as a directed graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ where $\mathbf{V}$ is the set of users and $\mathbf{E}$ is the set of directed edges between users. We define a node $s \in \mathbf{V}$ to denote the user of the original social media post and other nodes (i.e., $v \in \mathbf{V}, v \neq s$) represent the users who comment on the post. Each edge $e_{v,v'} \in \mathbf{E}$ denotes the comment from user v' to user v and is associated with several semantic attributes as follows.

Emotion Attribute $\rho_{em}(v, v')$*:* we obtain emotion scores of each comment using IBM Watson Natural Understanding API [4]. We extract five types of emotions: *anger*, *disgust*, *sadness*, *joy*, and *fear*, each of which is a score in $[0, 1]$. We initialize $\rho_{em}(v, v')$ as the emotion with highest score (i.e., dominant emotion) for $e_{v,v'}$.

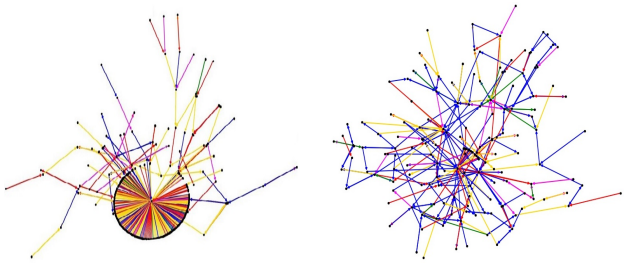
Attitude Attribute $\rho_{at}(v, v')$*:* we also obtain an "attitude" score of each comment. In particular, we consider three types of attitude scores - "debunking" (score of -1), "endorsing" (score of 1) and "neutral" (score of 0). The debunking attitude is derived based on whether a comment i) contains a set of debunking keywords such as "false alarm, fake, lie, not true"; or ii) has an extremely low polarity score ($\leq$ -0.8). The polarity score is extracted from the TextBlob tool [5]. The endorsing attitude is derived based on whether the comment is a "share" (e.g., repost/retweet), which represents an implicit endorsement. We observe that social media users are rarely explicit in acknowledging truthful contents (e.g., if the image is truthful, users seldomly use explicit terms like "real, authentic" to endorse it.).

Feedback Attribute $\rho_{fb}(v, v')$*:* we also obtain a "feedback" score of each comment which is defined as (# of likes - # of dislikes) of a comment.

2) Feature Extraction with Random Walk: After FauxBuster generates a comment network $\mathbf{G}$ for each post, it extracts the key characteristics from the network. In particular, we target two types of characteristics: i) topological features of the network, and ii) semantic features of the edges (comments). Random walk (RW) is a commonly used technique to extract topological information of a network [20]. Formally, a random walk $RW(M, K)$ randomly traverses a graph M times and each traversal visits at most K edges. RW records the visited nodes and the depth of each traversal to represent the topological feature of a network [21]. In FauxBuster, we found that both topological and semantic features of the comment network are closely related to the detection of fauxtography. For example, the "echo chambers" phenomena [22] that are often observed in misinformation spread can be characterized by a long path (topological feature) of the random walk in $\mathbf{G}$ and consecutive "debunking" behavior of the users (semantic feature). We extend the RW algorithm to jointly capture both the topological and semantic features during each traversal of $\mathbf{G}$. In particular, we define three types of semantic random walk paths. Each of these paths represents a particular semantic feature of the comment network.

DEFINITION 4. ***Emotion Path*** **($\mathbf{RW_{em}}$)***: the random walk traverses the graph* $\mathbf{G}$ *from source* s *and records the emotion attributes of each edge on its path. Formally,* $\mathbf{RW_{em}} = \{\mathbf{RW_{em}}(1), \mathbf{RW_{em}}(2), ..., \mathbf{RW_{em}}(K)\}$ *where* $\mathbf{RW_{em}}(k) = \rho_{em}(v_k, v_{k'})$ *represents the emotional attribute of the* k^{th} *edge* $e_{v_k,v_{k'}}$ *on the path. We set* $\mathbf{RW_{em}}(k) = 0$ *if the random walk has already stopped at a vertex with no incoming edges.*

$\mathbf{RW_{em}}$ captures the emotion features of each post. We observe that users often show different emotions in their comments on fauxtography and non-fauxtography posts. Figure 3 shows an example of emotion paths of two social media posts. Each network is traversed randomly 100 times with $K = 10$. We observe that users often express "joy" on fauxtography posts by making sarcastic comments and jokes or debunk the fauxtography posts and curse the uploaders with "anger". In contrast, the emotions of users' comments on the non-fauxtography posts are more diversified.



(a) Emotions of Fauxtography (b) Emotions of Non-Fauxtography

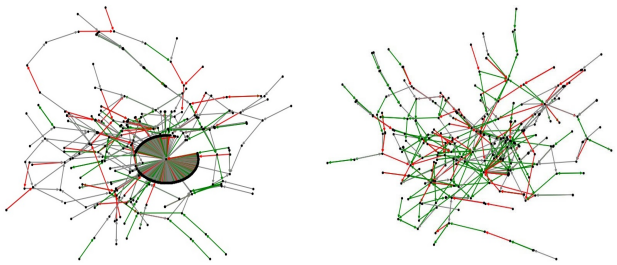
Figure 3: Illustration of emotion features. We use colors to denote the dominant emotion of each comment - "yellow-joy, red-anger, pink-disgust, green-fear, blue-sadness".

DEFINITION 5. ***Attitude Path*** **($\mathbf{RW_{at}}$)***: the random walk traverses the graph* $\mathbf{G}$ *from source* s *and records the attitude attribute of each edge on its path. Formally,* $\mathbf{RW_{at}} = \{\mathbf{RW_{at}}(1), \mathbf{RW_{at}}(2), ..., \mathbf{RW_{at}}(K)\}$ *where* $\mathbf{RW_{at}}(k) = \rho_{at}(v_k, v_{k'})$.

RW_{at} captures the "echo chambers" [23] of user comments - a chain of comments that either represents a public debate (consecutive "debunking" attitudes) or endorsement (consecutive "endorsement" attitudes). We observe that comments of fauxtography posts contain more debunking echo chambers than non-fauxtography posts (Figure 4).

[4]https://www.ibm.com/watson/services/natural-language-understanding/

[5]https://textblob.readthedocs.io/

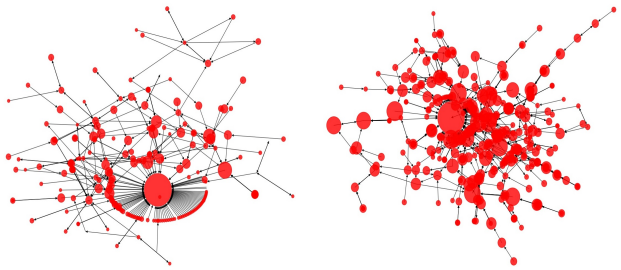


(a) Attitudes of Fauxtography (b) Attitudes of Non-Fauxtography

Figure 4: Illustration of attitude features. We use colors to denote the attitude of each comment - "red-debunk, green-endorse, black-neutral".

DEFINITION 6. ***Feedback Path*** **($\mathbf{RW_{fb}}$)***: the random walk traverses the graph* $\mathbf{G}$ *from source* s *and records the feedback attribute of each edge on its path. Formally,* $\mathbf{RW_{fb}} = \{\mathbf{RW_{fb}}(1), \mathbf{RW_{fb}}(2), ..., \mathbf{RW_{fb}}(K)\}$ *where* $\mathbf{RW_{fb}}(k) = \rho_{fb}(v_k, v_{k'})$.

$\mathbf{RW_{fb}}$ captures the feedback from the users on a post. We found there are usually several "hub" comments for fauxtography posts (e.g., "the image is fake, check original [URL]") that generate the "concentrated" supports. In contrast, users tend to post diverse opinions on non-fauxtography posts, which generate "dispersed" supports (Figure 5).



(a) Feedback of Fauxtography (b) Feedback of Non-Fauxtography

Figure 5: Illustration of feedback features. The size of a vertice indicates the aggregated feedback score (i.e., # of likes - # of dislikes) of all comments from a user.

The random walk is repeated M times for each semantic attribute. We use $\mathbf{RW^m_{em}}, \mathbf{RW^m_{at}}, \mathbf{RW^m_{fb}}$ to denote the semantic paths captured from the m^{th} random walk, $1 \leq m \leq M$. The recorded scores of each M random walk paths is further stored into a feature vector. In particular, we define a *emotion feature vector* $\mathbf{FV_{em}}$ as a $M \times K$ vector with $\mathbf{FV_{em}}(m,k) = \mathbf{RW^m_{em}}(k)$, an *attitude feature vector* $\mathbf{FV_{at}}$ as a $M \times K$ vector with $FV_{at}(m,k) = \mathbf{RW^m_{at}}(k)$, and a *feedback feature vector* $\mathbf{FV_{fb}}$ as a $M \times K$ vector with $\mathbf{FV_{fb}}(m,k) = \mathbf{RW^m_{fb}}(k), 1 \leq k \leq K, 1 \leq m \leq M$.

3) Network Representation Learning via Stacked Autoencoder: Given the feature vectors extracted from the random walk (i.e., $\mathbf{FV_{em}}$, $\mathbf{FV_{at}}$, $\mathbf{FV_{fb}}$), we now derive the signature of the comment network $\mathbf{G}$ using a deep autoencoding technique. An autoencoder is an artificial neural network technique for learning abstract features of high dimensional data using an unsupervised approach [24]. It consists of an *encoder* that maps an input vector $\mathbf{X}$ into a latent subspace $\mathbf{Z}$ and a *decoder* uses the latent representation $\mathbf{Z}$ to recover the original input. We adopt autoencoders in FauxBuster because i) it can reduce the complex and high-dimensional input data into a small number of high quality features [24]; ii) it can capture the latent factors (i.e., $\mathbf{Z}$) that are often shown to be more effective than directly using the original input features in supervised classification tasks [25].

In FauxBuster, we develop a set of stacked autoencoders to extract the latent representation of $\mathbf{G}$ that preserves its characteristics when the dimension of the input data is reduced. FauxBuster develops three six-layer stacked autoencoders to independently encode $\mathbf{FV_{em}}$, $\mathbf{FV_{at}}$, and $\mathbf{FV_{fb}}$. Taking the stacked autoencoder for $\mathbf{FV_{em}}$ (denoted as SAE_{em}) as an example, the representation of the l^{th} layer of the SAE_{em} is derived as:

$$\mathbf{Z}^l = \xi(\mathbf{W}^l \cdot \mathbf{X}^l + \mathbf{b}^l) \tag{2}$$

where $\xi(.)$ is a rectified linear unit (ReLU) activation function. $\mathbf{W}^l$ and $\mathbf{b}^l$ are weighting factor and bias of the l^{th} layer, which are the parameters of SAE_{em} to be learned. $\mathbf{X}^l$ denotes the input to the l^{th} layer which is the latent feature of the previous layer (i.e., $\mathbf{Z}^{l-1}$). We define the input to SAE_{em} as $\mathbf{X}^1 = \mathbf{FV_{em}}$.

To train the autoencoder and derive the latent representation for $\mathbf{FV_{em}}$, we define a customized feature reconstruction loss function as:

$$\mathcal{L}_{em} = ||(\mathbf{FV_{em}} - \widehat{\mathbf{FV_{em}}}) \odot \alpha_{em}||_2^2 \tag{3}$$

where $\odot$ denotes the Hadamard product and $\widehat{\mathbf{FV_{em}}}$ is the reconstructed feature vector. $\alpha(m,k)$ is defined as:

$$\alpha_{em} = \begin{cases} 1, & \mathbf{FV_{em}}(m,k) = 0,\ 1 \leq m \leq M, 1 \leq k \leq K \\ \lambda_{em}, & \mathbf{FV_{em}}(m,k) \neq 0,\ 1 \leq m \leq M, 1 \leq k \leq K \end{cases} \tag{4}$$

where $\lambda_{em} > 1$ is a weighting factor that imposes more penalty to the reconstruction error of non-zero elements in $\mathbf{FV_{em}}$ than that of zero elements. This is because a non-zero $\mathbf{FV_{em}}(m,k)$ carries more explicit emotion information than zero values. The stacked autoencoders are trained by minimizing $\mathcal{L}_{em}$ via layer-wise pre-training [26]. The encoded result of $\mathbf{FV_{em}}$ is denoted as $\mathbf{Z_{em}}$.

Similar loss functions are defined for $\mathbf{FV_{at}}$ and $\mathbf{FV_{fb}}$ and are omitted here due to space limit. After the encoded features vectors are calculated ($\mathbf{Z_{em}}$ for $\mathbf{FV_{em}}$, $\mathbf{Z_{at}}$ for $\mathbf{FV_{at}}$, $\mathbf{Z_{fb}}$ for $\mathbf{FV_{fb}}$), we apply softmax normalization to each of the encoded features. Finally, we apply an aggregation function $\phi(.)$ to combine these normalized encoded features. Common aggregation functions include Concatenation [25], Max pooling [27], and Principal Component Analysis [28].

We pragmatically choose the Concatenation function (i.e., $\mathbf{Z_{all}} = < \mathbf{Z_{em}}, \mathbf{Z_{at}}, \mathbf{Z_{fb}} >$) since it consistently achieves the best performance in our experiment.

B. Linguistic Feature Extraction

Next, we extract linguistic features from the user comments. We observe the usage of words are quite different in fauxtography and non-fauxtography posts. An example is shown in Figure 6. We can observe that image-related words (e.g. "picture", "photo", "photoshop") and verity-related words (e.g., "fake", "real", "original") appear more frequently in fauxtography posts while general news topics (e.g., "climate change", "ban", "China") are more commonly used in non-fauxtography posts.

(a) Fauxtography (b) Non-Fauxtography

Figure 6: Word Cloud

We use the state-of-the-art text embedding technique - Doc2vec to extract the linguistic features. Doc2vec is an unsupervised neural embedding framework that learns fixed-length vector representations from various-length word sequences from a document. It has been shown to be suitable for the unstructured social media data [29]. The representation learned from Doc2vec captures both the syntactic (e.g., word frequency) and the semantic (e.g., context and meaning) features of a social media post. We use the recommended hyper-parameter settings for short texts [30] to train the Doc2vec model and extract the vector representation of each social media post of interest.

C. Metadata Feature Extraction

We further extracted a few metadata features from the user comments. These features are mainly selected based on empirical observations. For example, we observe that many comments under fake images contain URLs of the original image. We also observe that many comments that contain image-related words such as "pixel" and "photo" if the corresponding post is misleading (e.g., "I can tell by the pixels", "Fake pics of Gatlinburg wildfires floating around"). We summarize all metadata features in Table I. For Reddit, we also add a unique feature called "average number of comments per thread" where the thread refers to the conversation under each top-level comment.

Table I: Metadata Features

Feature	Description
`total comments`	Total # of comments in each post
`average comments`	Avg. # of comments under each thread (Reddit)
`average verity`	Avg. # of verity related words in each comment
`average image`	Avg. # of image related words in each comment
`average question`	Avg. # of question marks in each comment
`average exclamation`	Avg. # of exclamation mark in each comment
`total url`	Total # of URLs
`average url`	Avg. # of comments contain URLs
`average word count`	Avg. # of words in each comment

D. Supervised Classifier

Using the network, linguistic and metadata features extracted from the collected data as discussed above, FauxBuster performs binary classification to decide whether a social media post is fauxtography or not. We leverage a set of state-of-the-art supervised machine learning models in the FauxBuster scheme, which includes neural networks, support vector machine, and ensemble methods. These classifiers serve as plug-ins to our FauxBuster scheme and we pragmatically select the one with the best performance from the evaluation of training data. We present the detailed performance evaluation of FauxBuster using different classifiers in Section VI.

V. Data

In this section, we describe our datasets and the data collection process. We choose two mainstream online social media platforms as our experiment playground - Reddit and Twitter [6]. Reddit, self-described as "front page of the Internet", is a large internet community where massive fresh internet content is constantly shared and commented on by its users. As of February 2018, Reddit had 542 million monthly visitors. Twitter is a global micro-blogging platform with 335 million active monthly users worldwide.

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 - (snopes.com, factcheck.org, truthorfiction.com) in a similar way as [31]. The ground-truth labels are initially decided based on majority vote of these fact checkers. We then assign three independent annotators to manually verify the label of each post using databases of historical facts and Google search.

Given the labeled images, we perform a reverse search using the Google Vision API [7] to identify the original web URLs that contain the image. If the URLs point to a social media post on Twitter or Reddit, we crawl the post and its comment thread using a crawler script we developed. We summarize the two real-world datasets used for evaluation in Table II. For Twitter, we also crawl the retweets and

[6]Reddit: https://www.reddit.com/ Twitter: https://twitter.com/
[7]https://cloud.google.com/vision/

replies to the original post. We observe that: i) the number of image-based posts from Twitter is much larger than that from Reddit; ii) a non-trivial amount of the fauxtography posts (13.2% in Reddit and 10.3% in Twitter) actually contain real images. This second observation validates the unique challenge of fauxtography detection where real images can also be leveraged to convey misleading messages.

Table II: Data Trace Statistics

Data Trace	Reddit	Twitter
Number of Posts	196	721
Number of Fauxtography	91	390
Number of Fauxtography with Real Images	12	40
Number of Comments	60,168	1,928,325
Number of Distinct Users	39,702	582,281

VI. Evaluation

In this section, we evaluate the FauxBuster scheme using the two real-world online social media datasets described in the previous section.

A. Evaluation Setup

We choose a few state-of-the-art supervised classifiers [32] that can be integrated with the FauxBuster scheme, including *Naïve Bayes (NB)*, *XGBoost*, *Random Forest (RF)*, *Linear Support Vector Machine (SVM)*, and *Multi-layer Perceptron (MLP)*.

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

- *Fake Image:* A feature engineering based approach to detect fake images on social media using a decision tree classifier [11].
- *Truth Discovery:* A representative fact-checking scheme to detect misinformation among conflicting text-based claims on social media [4].

B. Detection Effectiveness

In the first set of experiments, we evaluate the detection effectiveness of FauxBuster when it is coupled with different classifiers and identify the best-performed classifier for FauxBuster. The detection effectiveness is evaluated using common metrics for binary classification: *Accuracy*, *Precision*, *Recall* and *F1-Score*. For all the supervised classifiers, we use 70% of data as training set and perform 10-fold cross validation for parameter tuning using the training data. For feature extraction, we set M=100 and K=10 for the random walk algorithm and set the dimension of each stacked autoencoder's layer as 512, 128, 50 (hidden representation layer), 128, 512, and 1000 (output layer). We set the length of the Doc2vec embedding as 50.

The results are reported in Table III. We observe the XGBoost consistently outperforms other baseline classifiers. The reason is that the boosting technique employed by XGBoost can effectively aggregate weak decision trees as a powerful classifier [33]. We also found the artificial neural network baseline (i.e., MLP) performs poorly in the evaluation. We attribute it to the fact of limited training data. We use XGBoost as the default classifier for Fauxtography.

We further observe that FauxBuster has significant performance gains compared to Fake Image and Truth Discovery schemes. In particular, FauxBuster outperforms Fake Image and Truth Discovery by 8% and 33% respectively on Reddit and 25.6% and 30.9% respectively on Twitter in terms of F1-scores. This is because Fake Image baseline 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 with real images. On the other hand, Truth Discovery only considers whether the textual claims are truthful or not. This leads to false negatives in the results (i.e., fauxtography with fake images but truthful textual claims). In contrast, FauxBuster is explicitly designed to solve the problem of fauxtography that considers both the image and text 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 problem.

We further plot the ROC curves of all schemes in Figure 7 and 8. The ROC curve is commonly used to visualize the performance of binary classifiers. We observe that FauxBuser (XGBoost) continues to outperform other baseline classifiers as well as Fake Image and Truth Discovery schemes when we tune the classification thresholds.

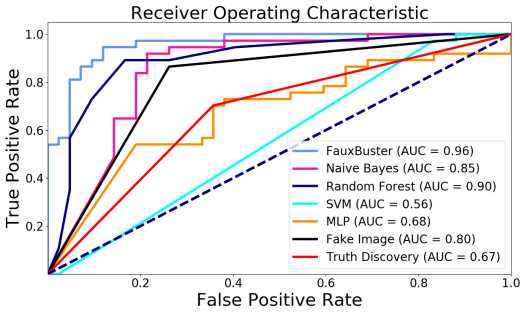


Figure 7: ROC Curve of All Schemes (Reddit)

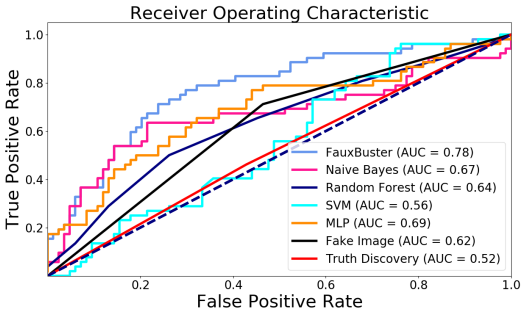


Figure 8: ROC Curve of All Schemes (Twitter)

Table III: Classification Accuracy for All Schemes

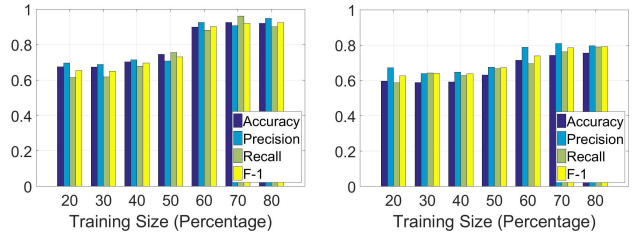
	Reddit				Twitter			
Algorithms	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
XGBoost(FauxBuster)	**0.918**	**0.903**	**0.933**	**0.915**	**0.743**	**0.81**	**0.762**	**0.785**
NB	0.747	0.704	0.905	0.792	0.639	0.746	0.631	0.684
RF	0.864	0.897	0.833	0.864	0.647	0.704	0.738	0.721
SVM	0.559	0.559	0.633	0.594	0.5	0.603	0.559	0.58
MLP	0.608	0.6	0.786	0.68	0.647	0.757	0.631	0.688
Fake Image	0.835	0.837	0.857	0.847	0.603	0.75	0.536	0.625
Truth Discovery	0.683	0.73	0.643	0.683	0.529	0.632	0.571	0.601

C. Feature Analysis

In addition to the holistic evaluation of FauxBuster system, we also investigate the importance of each type of features in FauxBuster. The results are summarized in Table IV. We observe that the addition of each feature increases the overall performance. FauxBuster achieves the best performance when all features are incorporated. Such results demonstrate the necessity of incorporating the network, linguistic and metadata features into the FauxBuster.

D. Influence of Training Size on FauxBuster

We then evaluate the influence of the size of the training set on the performance of FauxBuster. In our experiment, we vary the size of the training set from 20% to 80% of the whole dataset and report the Accuracy, Precision, Recall and F1-scores of FauxBuster. The results are shown in Figure 9. We also observe the performance of FauxBuster is stable when the size of the training set changes and stays reasonably high when the training size is larger than 60%.



(a) Reddit (b) Twitter

Figure 9: Training Size vs. Performance

E. FauxBuster versus Humans

We find it is also interesting to compare the performance of FauxBuster scheme 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 100 image-based social media posts (45 of which are fauxtography) from the two datasets for them to annotate. 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, etc.). Also, the annotators were asked to skip the posts that they happen to know the ground-truth.

In the first experiment, the annotators are allowed only to access the image and the text of a post. In the second experiment, they are allowed access to the entire post including comments of the post. Such an experiment design aims to evaluate if the comments from social media users would help humans to detect the fauxtography. The results are shown in Table V. We observe that FauxBuster significantly outperforms the human annotators even if they are allowed to view the comments of the post. In addition, we also observe that (i) the performance of annotators did improve significantly when they have access to the user's comments, which supports our assumption on the usefulness of comments on detecting fauxtography; (ii) humans are more likely to believe the fauxtography posts with real images than the posts with manually edited images. This again demonstrates that the fauxtography detection problem is more challenging than merely detecting "fake images".

F. Detection Time

Finally, we evaluate the detection time of FauxBuster. The detection time is defined as the amount of time a scheme takes to detect the fauxtography post after it has been originally posted. In the experiment, we tune the time window of the data collected from 1 hour to 5 days and only use the user comments within the specified time window for the tested schemes. The results are reported in Figure 10. We observe that FauxBuster outperforms the baselines (Fake Image and Truth Discovery) consistently on both datasets. We also observe that FauxBuster achieves a high performance quickly (F-1 Score of 0.88 for Reddit, and 0.73 for Twitter within one day). This is because most comments on social media appear at the early stage of the information spread. The above results suggest the FauxBuster scheme can catch fauxtography not only accurately but also timely.

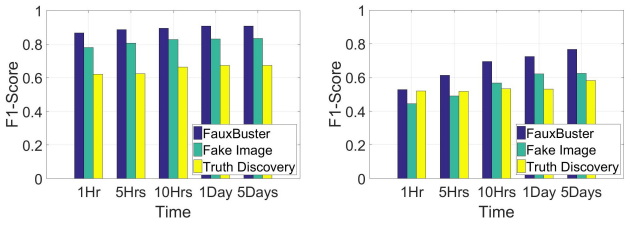
Table IV: Feature Analysis for FauxBuster

	Reddit				Twitter			
Feature Sets	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
All	**0.918**	**0.903**	**0.933**	**0.915**	**0.743**	**0.81**	**0.762**	**0.785**
Network Only	0.711	0.709	0.762	0.736	0.7	0.759	0.717	0.737
Linguistic Only	0.747	0.806	0.691	0.744	0.684	0.747	0.738	0.742
Metadata Only	0.823	0.937	0.714	0.811	0.566	0.658	0.619	0.638
Network + Linguistic	0.772	0.853	0.691	0.763	0.692	0.735	0.735	0.735
Network + Metadata	0.899	0.925	0.881	0.902	0.654	0.761	0.643	0.697
Linguistic + Metadata	0.899	0.947	0.857	0.9	0.639	0.673	0.735	0.702

Table V: FauxBuster vs. Human Performance

	Accuracy	F1	FPR	FNR
FauxBuster	**0.92**	**0.915**	**0.058**	**0.104**
A1	0.44	0.391	0.422	0.672
A1+comment	0.71	0.713	0.222	0.345
A2	0.46	0.413	0.4	0.654
A2+comment	0.7	0.737	0.378	0.236
A3	0.39	0.408	0.6	0.618
A3+comment	0.63	0.648	0.356	0.382
Overall	0.44	0.404	0.444	0.654
Overall+comment	0.74	0.764	0.289	0.236

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



(a) Reddit (b) Twitter

Figure 10: Elapsed Time vs. Performance

VII. Conclusion

In this paper, we develop the first content-free solution (i.e., FauxBuster) to address the fauxtography detection problem in image-based social media posts. The FauxBuster is robust against sophisticated image manipulation by leveraging the valuable clues from the unstructured and noisy social media comments. Using two real-world social media datasets from Reddit and Twitter, we demonstrated that FauxBuster can effectively track down fauxtography on social media and outperform existing baselines in terms of both accuracy and detection time.

Acknowledgement

This research is supported in part by the National Science Foundation under Grant No. CNS-1831669, CBET-1637251, CNS-1566465, and IIS-1447795, Army Research Office under Grant W911NF-17-1-0409, Google 2017 Faculty Research Award. 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] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *Proceedings of the 19th international conference on World wide web*. AcM, 2010.

[2] D. Wang, B. K. Szymanski, T. Abdelzaher, H. Ji, and L. Kaplan, "The age of social sensing," *arXiv preprint arXiv:1801.09116*, 2018.

[3] X. Yin, J. Han, and P. S. Yu, "Truth discovery with multiple conflicting information providers on the web," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 6, pp. 796–808, Jun. 2008.

[4] D. Y. Zhang, R. Han, D. Wang, and C. Huang, "On robust truth discovery in sparse social media sensing," in *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1076–1081.

[5] N. Vo and K. Lee, "The rise of guardians: Fact-checking url recommendation to combat fake news," *arXiv preprint arXiv:1806.07516*, 2018.

[6] "Facebook and google's war with fake news heats up," https://moneyish.com/ish/google-joins-facebook-in-declaring-war-on-fake-news/, accessed: 2018-08-07.

[7] S. D. Cooper, "A concise history of the fauxtography blogstorm in the 2006 lebanon war," 2007.

[8] "Photo fuels spread of fake news," https://www.wired.com/2016/12/photos-fuel-spread-fake-news/.

[9] "Social media engagement – statistics and trends," https://www.invespcro.com/blog/social-media-engagement/, accessed: 2018-08-07.

[10] "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/, accessed: 2018-08-07.

[11] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy," in *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013.

[12] A. J. Fridrich, B. D. Soukal, and A. J. Lukáš, "Detection of copy-move forgery in digital images," in *in Proceedings of Digital Forensic Research Workshop*. Citeseer, 2003.

[13] T. Huynh-Kha, T. Le-Tien, S. Ha-Viet-Uyen, K. Huynh-Van, and M. Luong, "A robust algorithm of forgery detection in copy-move and spliced images," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 7, no. 3, 2016.

[14] P. Korus and J. Huang, "Multi-scale fusion for improved localization of malicious tampering in digital images," *IEEE Transactions on Image Processing*, vol. 25, no. 3, 2016.

[15] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[16] Q. Q. Yao, D. D. Perlmutter, and J. Z. Liu, "What are shaping the ethical bottom line?: Identifying factors influencing young readers' acceptance of digital news photo alteration," *Telematics and Informatics*, vol. 34, no. 1, pp. 124–132, 2017.

[17] B. Bayar and M. C. Stamm, "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*. ACM, 2016.

[18] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher, "On truth discovery in social sensing: A maximum likelihood estimation approach," in *Proc. ACM/IEEE 11th Int Information Processing in Sensor Networks (IPSN) Conf*, Apr. 2012, pp. 233–244.

[19] D. Y. Zhang, D. Wang, and Y. Zhang, "Constraint-aware dynamic truth discovery in big data social media sensing," in *Big Data (Big Data), 2017 IEEE International Conference on*. IEEE, 2017, pp. 57–66.

[20] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2016, pp. 855–864.

[21] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 701–710.

[22] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, 2017.

[23] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–36, 2017.

[24] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 1096–1103.

[25] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "Line: Large-scale information network embedding." International World Wide Web Conferences Steering Committee, 2015, pp. 1067–1077.

[26] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Advances in neural information processing systems*, 2007, pp. 153–160.

[27] Y.-L. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in visual recognition," in *Proceedings of the 27th international conference on machine learning*, 2010.

[28] I. Jolliffe, "Principal component analysis," in *International encyclopedia of statistical science*. Springer, 2011.

[29] L. Q. Trieu, H. Q. Tran, and M.-T. Tran, "News classification from social media using twitter-based doc2vec model and automatic query expansion," in *Proceedings of the Eighth International Symposium on Information and Communication Technology*. ACM, 2017, pp. 460–467.

[30] J. H. Lau and T. Baldwin, "An empirical evaluation of doc2vec with practical insights into document embedding generation," *CoRR*, vol. abs/1607.05368, 2016.

[31] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild *et al.*, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[32] E. Alpaydin, *Introduction to machine learning*. MIT press, 2009.

[33] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016, pp. 785–794.