Understanding the Use of Fauxtography on Social Media

Yuping Wang[♣], Fatemeh Tahmasbi[♦], Jeremy Blackburn[♦], Barry Bradlyn[♠],
Emiliano De Cristofaro[♥], David Magerman[∂], Savvas Zannettou[★], and Gianluca Stringhini[♣]
[♣]Boston University, [♦]Binghamton University, [♠]University of Illinois at Urbana–Champaign,
[♥]University College London, [∂] Differential Venture Partners, [★]Max Planck Institute for Informatics

Abstract

Despite the influence that image-based communication has on online discourse, the role played by images in disinformation is still not well understood. In this paper, we present the first large-scale study of *fauxtography*, analyzing the use of manipulated or misleading images in news discussion on online communities. First, we develop a computational pipeline geared to detect fauxtography, and identify over 61k instances of fauxtography discussed on Twitter, 4chan, and Reddit. Then, we study how posting fauxtography affects engagement of posts on social media, finding that posts containing it receive more interactions in the form of re-shares, likes, and comments. Finally, we show that fauxtography images are often turned into memes by Web communities. Our findings show that effective mitigation against disinformation need to take images into account, and highlight a number of challenges in dealing with image-based disinformation.

1 Introduction

Recent years have seen an increase in false information published online and spread through social media [14]. An important aspect of news consumption is that users not only pay attention to text, but also to the accompanying images in the article. In fact, research in psychology shows that images play a crucial role in both how readers perceive certain issues [37] and in which articles individuals choose to read [38]. Therefore, it is not surprising that images may be manipulated or misrepresented to mislead users.

In this paper, we focus on *fauxtography* [7], i.e., news images that have been modified or miscaptioned to change their intent, often with the goal of spreading a false sense of the events they purport to depict. Although previous research efforts have proposed detection tools for fauxtography [36, 39], to the best our knowledge, the *impact* of fauxtography on news discussion has not been studied. In particular, we set out to investigate two research questions:

- RQ1: Does sharing fauxtography increase engagement on social media?
- RQ2: Do fauxtography images have a life beyond their questionable verisimilitude (their appearance of being being real)? I.e., do new variants and memes using them appear on social media?

To answer these questions, we develop a computational analysis pipeline geared to identify posts containing fauxtography at scale, measure the engagement of users sharing and viewing such posts, and understand how these images are used on different social media platforms. First, we gather 2.6 billion posts from three social media platforms (Twitter, Reddit, and 4chan) as well as 32M news articles published by over 1,000 news websites. Then, we extract all images appearing in these posts and articles, and use perceptual hashing [19] to match them to images labeled as fauxtography by the fact-checking site Snopes. In total, we identify 61K posts containing fauxtography shared by users over the two year period from 2016 to 2018.

To address RQ1, we analyze the reactions to posts containing fauxtography on social media, compared with the reaction to posts by the same users with no image or with images characterized as non-fauxtography. We find that including fauxtography in posts does increase user engagement on social media. On the other hand, posting links to news articles that contain fauxtography (rather than posting images directly) does not increase engagement on Twitter, while it does yield more interactions on Reddit. Surprisingly, the extent to which an image is misleading – e.g., whether it is completely false or just partially true – does not significantly affect engagement, suggesting that the increased engagement is driven by the inflammatory and controversial nature of fauxtography more than its verisimilitude.

For RQ2, we search for variants of fauxtography images that each appear on all of the social media platforms. Our intiution is that instances of fauxtography are likely have some sort of inherent exploitability making them suitable as a base for new memes. Visual memes have become important to the spread of racist and political ideology [9, 35, 32] and have been used by state-sponsored actors to wage information warfare [1, 33]. We find evidence of fauxtography images being turned into memes and being manipulated in ways not related to their original verisimilitude.

Finally, by focusing on three selected case studies of fauxtography which spawned new variants, we will discuss implications for dealing with fauxtography in the wild, considering the current environment of social media moderation.

(a) Manipulated

(b) Original

Figure 1: This picture originally depicted a UK protester holding the "Black Lives Matters" sign. It was manipulated so that the sign says "Lincoln was Racist" and the person has been mischaracterized as being a Missouri State University student. See https://www.snopes.com/fact-check/abe-lincoln-racist-protest-sign/

Figure 2: Miscaptioned image used to falsely claim that people in the migrant caravan burnt an American flag. See https://www.snopes.com/fact-check/caravan-burning-flag/

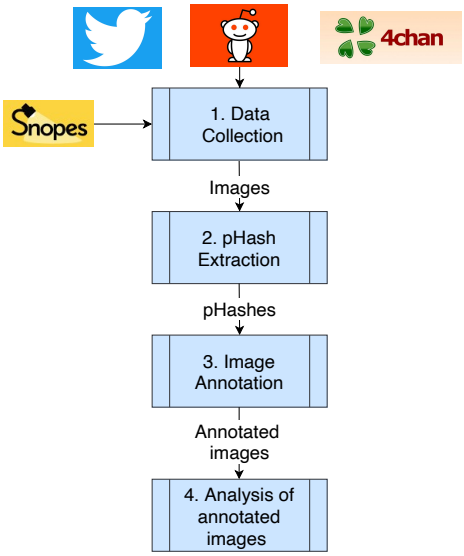

Figure 3: Overview of our computational analysis pipeline.

Platform	#Posts	#Image URLs	#Images Obtained
Twitter	2,213,019,239	701,806,921	435,244,799
Reddit	295,460,914	78,682,398	61,703,316
4chan	99,614,382	27,044,132	23,379,630
News Articles	32,200,604	28,654,146	27,360,218
Snopes	2,286	16,206	7,835

Table 1: Overview of our datasets.

2 Fauxtography

The term fauxtography was first coined by [7] in the context of the 2006 Lebanon war, as combination of the word *faux* (French for false) and *photography*. Cooper defines fauxtography as "visual images, especially news photographs, which convey a questionable (or outright false) sense of the events they seem to depict." Fauxtography usually involves manipulated images aiming to influence the emotions of viewers. Therefore, it involves deception, often realized by directly manipulating the images, captions, or overall the narrative associated with the image.

To better explain what fauxtography is, we provide two examples. Figure 1 shows a picture of a protester in the UK holding a sign reading "Black Lives Matter," which was manipulated to instead read "Lincoln Was Racist." Online sources also erroneously claimed that the person holding the sign was a Missouri State University student at a US protest. Figure 2 shows an image that was not manipulated, but that has often been used out of context and miscaptioned to imply that migrants on a caravan to the US in 2018 had burned the American flag. In reality, this photo was taken at an anti-Trump protest in the US and the flag is actually a Trump banner, not a US flag. These examples demonstrate two important characteristics of fauxtography, distinguishing them from "simple" fake images: 1) they are related to news or public affairs, and 2) users who see them can be fooled relatively easily if the images are not fact-checked.

3 Methodology & Dataset

In this section, we present our computational pipeline, as well as the dataset used in our study. As depicted in Figure 3, the pipeline consists of four components: 1) data collection, 2) pHash extraction, 3) image annotation, 4) image analysis.

3.1 Data Collection

Our study relies on two types of data sources: 1) *images* from Web communities and news articles posted on them; and 2) *annotation sources* to identify which images are fauxtography. For the former, we use Twitter, Reddit, and 4chan, and in particular images shared between July 1, 2016 and October 31, 2018; basic statistics are reported in Table 1. As an annotation source, we use Snopes.com, and specifically images posted on its fauxtography section. This allows us to identify all images that Snopes classified as fauxtography between the early 2000s and October 2019. Note that our analysis pipeline supports any Web community and any annotation source; however, in the following, we provide details of the sources used in this paper.

Images shared on Web communities. First, we collect images posted publicly on Twitter, 4chan, and Reddit. For Twitter, we collect data using the 1% Streaming API, with tweets stored as they were posted, in real time. In total, we parse 2.2B tweets, 702M of which contain at least one image. Note that the Twitter API does not return images directly, but rather a URL pointing to the image. We download the images in March

2020 and are able to retrieve 435M of them. The remaining images are unavailable, either because the image URL had changed, the tweet was deleted, or because the account that posted it was suspended.

For Reddit, we use the Pushshift dataset [2]. We obtain 295M posts, 79M of which contain images. Of these, we successfully retrieve 62M images, with the rest having been deleted. For 4chan, we use the dataset from [22] and obtain 100M posts from 4chan's Politically Incorrect board (/pol/). The dataset does not include the images posted on /pol/ (only an md5 checksum), hence we use 4plebs.org, an archival service, to collect the images. Overall, we collect 27M image URLs from 4plebs, from which we are able to download 23M images.

Images from news articles posted on Web communities. On most social networks, when a user shares a news article, the platform often automatically generates a preview for it. Typically, this includes the main image of the article (i.e., the one appearing on the top). The preview is important with respect to users' image sharing behavior, thus, we complement our image data collection with images included on news articles shared on Twitter, Reddit, and 4chan.

To do so, we use a systematic approach to create a list of news outlets; we start from the top 30K Majestic [17] websites released as of February 2019, and use the VirusTotal API [27], a domain categorization service, to get domains categorized as "news" and "news and media." Note that the news outlet labeling given by VirusTotal is not always accurate, e.g., domains like adbusters.org are incorrectly classified as news outlets. To solve this problem, we use the NewsGuard API [20] to refine the which domains are actually news outlets, and only select those listed in NewsGuard as of February 2019. In total, we identify 1,037 news outlets.

We then collect posts containing URLs to the 1,037 news outlets posted on Twitter, Reddit, and 4chan, gathering a total of 32M news articles, with approximately 29M including image URLs. Note that we only consider the top image from each article, which is the image that appears on top of the article. To collect the images, we use the Newspaper3k Library [21] to parse the HTML of the 32M news articles, and then extract the URL of the top image identified by Newspaper3k. We are able to download 27M images from the 29M image URLs in the news articles.

Snopes. As mentioned, we annotate images using Snopes, a website dedicated to fact-checking news, which has a special section dedicated to fauxtography.[1] Each entry in this section consists of a topic and a claim associated to an image, which is rated by Snopes using ten possible labels, listed in Table 2. For our analysis, we merge these labels into two groups: Merged True (True and Mostly True) and Merged False (Mostly False, False, and Miscaptioned), as illustrated by Table 2. The former category includes cases where although part of the claim might be inaccurate, the usage of the image is still correct, whereas, the latter indicates that the usage of an image for a given claim is problematic.

[1] https://www.snopes.com/fact-check/category/photos/

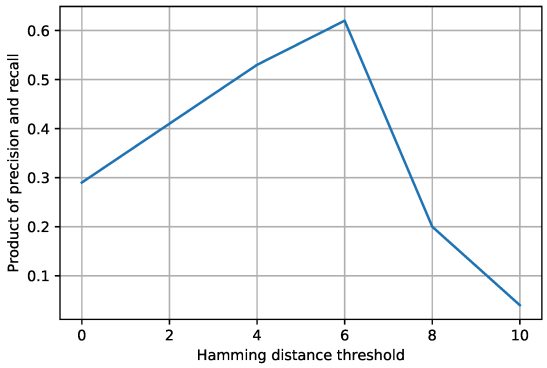

Figure 4: Product of precision and recall at different pHash Hamming distance thresholds in the image annotation process.

We collect data from the Snopes fauxtography category from the very beginning of the site (early 2000s) to October 2019, obtaining 2,286 articles. These include 16K URLs to images, out of which we successfully download 7.8K (the rest of the URLs are no longer available).

3.2 pHash Extraction

Having collected all images from our data sources, the next step in our pipeline is to convert the images to a format that we can easily work with. To do so, we apply the Perceptual Hashing (pHash) algorithm [19] using the ImageHash library [5], which generates a hash for each image in such a way that visually similar images have minor differences in their hashes. The algorithm is robust to image transformations (e.g., slight rotation, skew).

3.3 Image Annotation

Next, we annotate and identify the images that relate to fauxtography. To do this, we perform pairwise comparisons between the pHashes of images obtained from the various Web communities (including news articles) and images obtained from image annotation sites, such as Snopes. We calculate the Hamming distance between a pair of pHashes (i.e., an image from Snopes and an image shared on Twitter) and we assume that an image is related to fauxtography if the Hamming distance is less than or equal to a pre-defined threshold, which we set below. Previous work [32] shows that pHash is ineffective when dealing with images that are dominated by a single background color (e.g., screenshots on a white background), thus, we remove from our dataset images from annotation sites dominated by a single background color (i.e., screenshots, images of sky, etc.). Overall, this leaves is with 5,789 Snopes images for subsequent analysis.

Setting the pHash threshold. We consider two images to be visually similar if the Hamming distance is below a certain threshold. We vary the threshold from 0 to 10 and perform manual inspection of the matched images between the top images of news articles shared on all three Web communities and the corresponding Snopes images.[2] We consider a match to

[2] Empirically, we find that any pair of images with Hamming distance above 10 consists of extremely dissimilar images.

Original Labels	True	Mostly True	Mostly False	False	Miscaptioned	Legend	Outdated	Satire	Unproven	Mixture
Our Labels	Merged True		Merged False			*Not considered*				

Table 2: Overview of the fauxtography labels assigned by Snopes and of the grouping that we use for the analysis in this paper.

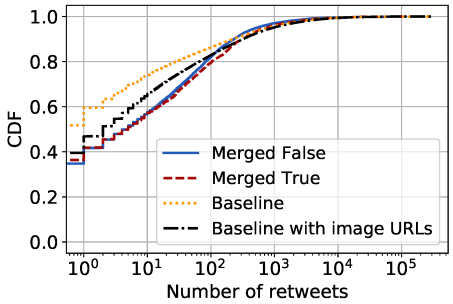

Figure 5: CDF of number of retweets on tweets sharing directly images.

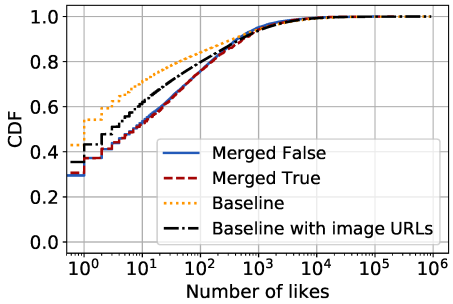

Figure 6: CDF of number of likes on tweets sharing directly images.

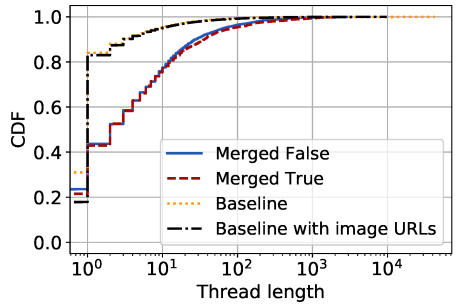

Figure 7: CDF of Reddit submission thread length on submissions sharing directly images.

be correct if a human annotator considers them visually similar. For each value of the Hamming Distance, we calculate the product of precision and recall for all pairs. In total, we manually check 76,067 pairs of matched images.

The result of the pHash threshold selection process is shown in Figure 4: the maximum product of precision and recall is obtained at Hamming distance 6 (0.89 precision and 0.69 recall), hence, we use 6 as the threshold to determine if two images are similar. At this threshold, we find that 2,129 fauxtography images from Snopes have at least one match in posts on one of the social networks or in our dataset. In total, we find 45,567 tweets, 10,916 submissions and comments from Reddit, 2,987 posts from 4chan, and 1,633 news articles that include fauxtography.

4 RQ1: Impact on Engagement

To understand if including fauxtography in social media posts increases engagement, we first look at whether posts on Twitter containing fauxtography produce more retweets and likes than other posts. We next look at submissions on Reddit, where we use the scores that a submission receives and the length of threads as engagement metrics. Finally, we look at posts on Twitter and Reddit that do not include fauxtography directly, but that rather include links to news articles containing fauxtography. Note that we do not analyze the 4chan data here because the small number of data points makes statistical analysis unsuitable (301 threads fauxtography images in total for images that are shared directly, and 38 images for news articles containing fauxtography).

4.1 Twitter

As we use the Twitter streaming API for data collection, our data contains real time activity, i.e., tweets are gathered as soon as they are posted. This makes the dataset less than ideal to assess the engagement received by tweets, because the number of retweets and likes reported by the API represents short-term behavior. To gain a better view of the long-term engagement, we leverage a process called *hydration*[3]: given a tweet ID, we retrieve the latest version of the current number of retweets and likes for that tweet. We hydrate the tweets in our dataset between June and July 2020.

Tweets can be classified as original tweets, retweets, and quote tweets. After hydration, we find that we cannot retrieve the actual retweet and likes count of regular retweets.[4] Therefore, to assess engagement for retweets, we retrieve the latest version of the original tweet that generated the retweet.

As discussed earlier, Snopes provides detailed labels to characterize fauxtography. For our experiments, we combine similar ratings together and form a binary system with two classes, Merged True and Merged False (see Table 2).

To understand whether posts containing fauxtography produce more engagement on Twitter, we extract two baselines: a set of random tweets and a set of tweets containing images that are not labeled as fauxtography. We then compare the engagement distribution of these tweets to posts containing faux-

[3] https://developer.twitter.com/en/docs/twitter-api/v1/tweets/post-and-engage/api-reference/get-statuses-lookup

[4] The "retweet_count" field in the metadata of the retweet, representing how many times a tweet is retweeted, is always equal to the "retweet_count" field in the metadata of the corresponding original tweet. In addition, the field "favorite_count," i.e., how many times a tweet is liked, is always 0 in the metadata of a retweet even if users press "like" on the retweet instead than on the original tweet, and the "favorite_count" of the original tweet is increased instead.

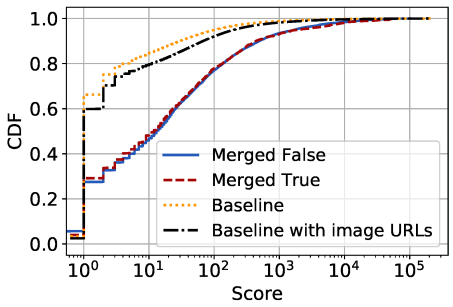

Figure 8: CDF of Reddit submission score on submissions sharing directly images.

tography. We identify 9,858 Twitter users that shared tweets containing Merged True and Merged False fauxtography. We collect 9,771 tweets containing fauxtography rated as Merged False, 2,183 tweets containing fauxtography rated as Merged True. Then, we construct two baselines deriving from all the tweets shared from these 9,858 Twitter users: 1) 1,720,197 tweets that do not include images; and 2) 782,391 tweets that include non-fauxtography images.

Figures 5 and 6 show the cumulative distribution functions (CDFs) of the retweets and likes received by the four types of tweets, respectively. We observe that tweets containing an image from our fauxtography dataset (whether true or false) are more likely to produce more retweets and likes than our baseline tweets: 42% tweets containing Merged False fauxtography and 43% tweets containing Merged True fauxtography have been retweeted more than 10 time, while only 26% tweets from the generic baseline of random tweets have been retweeted more than 10 times. Similarly, 46% tweets containing Merged False fauxtography and 47% tweets containing Merged True fauxtography have been liked more than 10 times, while only 29% tweets from the generic baseline have been liked more than 10 times.

To assess differences between these distributions, we run two sample Kolmogorov-Smirnov tests (K-S test) [16]. We first compare to the baseline set of random tweets. We find that the differences between the following distributions are statistically significant at the $p < 0.05$ level: Merged False tweets compared to the baseline (D=0.182), and Merged True tweets compared to the baseline (D=0.185) when examining retweets. As for likes, we have statistically significant differences between the distribution of Merged False tweets compared to the baseline (D=0.187), and for Merged True tweets compared to the baseline (D=0.192). In all cases, $p \ll 0.001$ We thus reject the null hypothesis that tweets with fauxtography images receive the same level of engagement as random tweets.

There is reason to believe that tweets containing images get more engagement overall [15]. To lend further evidence to the observation that our images in our fauxtography dataset are likely to receive more engagement than random images, we next compare the fauxtography distributions to the baseline of tweets with images in Figures 5 and 6. Again, we observe that tweets with images from our fauxtography dataset are more likely to be retweeted and liked than those with other images: only 34% of tweets containing non-fauxtography image have been retweeted more than 10 times, and only 38% have been liked more than 10 times. Using 2-sample K-S tests, we reject the null hypothesis that tweets containing non-fauxtography images and those with fauxtography have the same probability of receiving engagement ($p \ll 0.001$ in all cases). For retweets, we have D=0.0811 for Merged False tweets compared to non-fauxtography image baseline, and D=0.0896 for Merged True tweets compared to non-fauxtography image baseline. For likes we have D=0.0854 for Merged False tweets compared to the non-fauxtography image baseline, and D=0.0968 for Merged True tweets compared to the non-fauxtography image baseline.

Finally, a question remains as to whether or not the verisimilitude of a fauxtography image affects its engagement. We compare the distribution of engagement between tweets with Merged True and Merged False images. In this case, we reject the null hypothesis that there is a difference with respect to retweets (D=0.0380, p = 0.011), however we are *unable* to reject the null hypothesis of differences with respect to likes (D=0.0219, p = 0.36). One explanation for this result is that images in our fauxtography dataset are usually quite controversial, with a sensationalist tone. We speculate that this tends to drive engagement, regardless of the underlying verisimilitudeof the image itself.

4.2 Reddit

For Reddit, we run analogous experiments using the length of a thread and the score of a submission as engagement metrics. Reddit calculates the score of a post as the difference between the number of upvotes and downvotes that it receives. On Reddit, the initial post in a thread is the "submission," and other posts in that thread are "comments." The length of a thread is obtained from the "num_comments" metadata field, and the score (i.e., the number of upvotes minus the number of downvotes) is obtained from the "score" field in submission metadata. Note that the "score" field is a precise value[5] while upvote and downvote values are fuzzed.

First, we identify 4,883 users that shared submissions containing fauxtography. These users shared 5,444 submissions containing Merged False fauxtography and 1,522 submissions containing Merged True fauxtography, respectively. Then, we construct two baselines based on the same set of Reddit users: 1) 7,248,595 submissions that do not include images; and 2) 3,367,222 submissions that include non-fauxtography images.

Figures 7 and 8 plot the CDF of thread length and score (respectively) for each of the four sets of submissions just described. From the plots, we note that 23% of submissions containing Merged False or Merged True fauxtography resulted in threads with more than 10 comments, while this is true for only 4.6% of non-image submissions and 4.8% submissions containing non-fauxtography images. Similarly, 53% submissions containing Merged False fauxtography and 53% submissions

[5] https://www.reddit.com/wiki/faq#wiki_how_is_a_submission.27s_score_determined.3F

containing Merged True fauxtography have scores higher than 10, but only 15% of generic non-image submissions and 20% submissions containing non-fauxtography images have a score above 10. This suggests that fauxtography images produce more engagement than the baseline, regardless of whether the random post contains an image or not.

The differences in these distributions are again statistically significant as confirmed via 2-sample K-S tests. For the length of threads, we have D=0.404 for Merged False submissions compared to the no-image baseline, and D=0.411 for Merged True submissions compared to the no-image baseline. For likes, we have D=0.426 for Merged False submissions compared to the no-image baseline, and D=0.414 for Merged True submissions compared to the no-image baseline. When comparing to the non-fauxtography image baseline, we have D=0.394 for Merged False submissions and D=0.401 for Merged True submissions when looking at the length of threads. For likes, we have D=0.381 for Merged False submissions compared to the non-fauxtography image baseline, and D=0.368 for Merged True submissions compared to the non-fauxtography image baseline. In all cases, we find $p \ll 0.001$

Similar to Twitter, we are *unable* to reject the null hypothesis that there is no difference in engagement between true and false fauxtography images on Reddit. In the case of Reddit it is important to note that we are unable to reject the null hypothesis for both types of engagement; $D = 0.0220$, $p = 0.61$ for thread length and $D = 0.0225$, $p = 0.51$ for submission score. Again, this suggests that the engagement generated by fauxtography images is independent of the verisimilitude of the image.

4.3 News URLs

We now look at the engagement generated by posts that have links to news articles that include fauxtography rather than directly including fauxtography. On Twitter, we identify 100 tweets with links to news articles that contain Merged False fauxtography images and 94 tweets with links to articles that contain Merged True fauxtography images. On Reddit, we identify 431 submissions with links to articles that contain Merged False fauxtography and 272 submissions with links to articles that contain Merged True fauxtography.

Once again, we compare the engagement of posts containing links to news articles containing fauxtography to a generic baseline of 492,604 tweets on Twitter and 19,704,911 Reddit submissions, respectively, and to a baseline of 239,079 tweets and 9,554,016 posts containing generic news URLs. The baselines are constructed by collecting all non-fauxtography posts posted by the users who made at least one fauxtography related submission on Twitter or Reddit. Figures 9 and 10 show the retweets and likes of tweets containing fauxtography news URLs. Contrary to what observed previously, these tweets do not receive more engagement than baselines. More precisely, on Twitter, 32% tweets containing Merged False fauxtography and 26% tweets containing Merged True fauxtography have been retweeted more than 10 times, while 82% generic tweets and 85% tweets containing non-fauxtography news URLs been retweeted more than 10 times. Furthermore, only 33% tweets containing fauxtography rated as Merged False and 29% tweets containining fauxtography rated as Merged True have been liked more than 10 times, while 83% of generic tweets and 86% tweets contain generic non-fauxtography news URLs have been liked more than 10 times.

On Reddit, on the other hand, we find that posts containing links to fauxtography news articles still receive more engagement. As Figures 11 and 12 show, 16% of submissions containing Merged False fauxtography and 14% of submissions containing Merged True fauxtography have thread lengths longer than 10, while the same is true only for 1.3% of generic submissions and 1.6% non-fauxtography news URL submissions.

On Twitter, we confirm differences in these distributions via the 2-sample K-S test for fauxtography submissions compared to baseline submissions, where for the number of retweets we have D=0.506 for Merged False tweets compared to the non-fauxtography baseline, and D=0.584,for Merged True tweets compared to the generic baseline. For likes, we have D=0.509 for Merged False tweets compared to the generic baseline, and D=0.545 for Merged True tweets compared to the generic baseline. Looking at the non-fauxtography news article baseline, we have D=0.536 for Merged False tweets and D=0.614, for Merged True tweets when looking at retweets. For likes, we have D=0.534 for Merged False tweets, and D=0.571 for Merged True tweets. In all cases, $p \ll 0.001$ leading us to reject the null hypothesis that there are no differences between these distributions.

On Reddit, when looking at the length of threads we have D=0.275 for Merged False submissions compared to the generic baseline, and D=0.358 for Merged True submissions compared to the generic baseline. For Scores, we have D=0.260 for Merged False submissions compared to the generic baseline, and D=0.339 for Merged True submissions compared to the generic baseline. When looking at the non-fauxtography news article baseline, we have D=0.271 for Merged False submissions and D=0.354 for Merged True submissions for the length of threads. For Scores, we have D=0.282 for Merged False submissions compared to the non-fauxtography image baseline, and D=0.362 for Merged True submissions compared to the non-fauxtography image baseline. In all cases, $p \ll 0.001$ leading us to reject the null hypothesis that there are no differences between these distributions.

Note, however, we are unable to reject the null hypothesis that there are differences in engagement between Merged True and Merged False tweets and submissions. On Twitter, a KS test gives us D=0.0998 ($p = 0.7$) for retweets and D=0.0562 ($p \approx 1.0$) for likes. On Reddit, we obtain D=0.0826 ($p = 0.17$) for the length of threads and D=0.0791 ($p = 0.21$) for scores.

While most of the results for this experiment are consistent with what we previously found with regards to directly sharing fauxtography, interestingly, tweets containing links to news articles with fauxtography attract less engagement than other news links. One possible reason is that when sharing news URLs, many confounding factors can come into play with re-

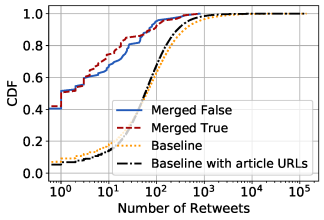

Figure 9: CDf of number of retweets on tweets sharing news articles.

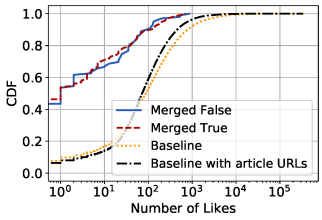

Figure 10: CDf of number of likes on tweets sharing news articles.

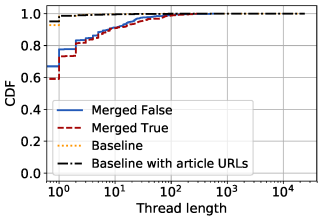

Figure 11: CDF of Reddit submission thread length on submissions sharing news articles.

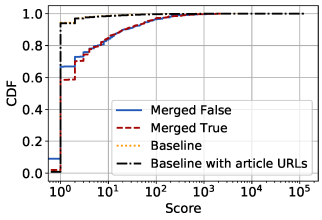

Figure 12: CDF of Reddit submission score on submissions sharing news articles.

gards to enticing users into interacting with the tweet, for example clickbait titles and the content of the article. For Reddit, the results show that using news articles to share fauxtography can increase engagement, which is consistent with the results found sharing images directly.

4.4 Takeaways

Our analysis provides evidence that posts directly containing fauxtography images do indeed generate higher engagement on both Twitter and Reddit. However, when it comes to sharing links to news articles that make use of fauxtography, we find that they generate significantly *less* engagement on Twitter but significantly *more* engagement on Reddit. Further, except in the case of retweets on Twitter, we are unable to reject the null hypothesis that Merged True and Merged False posts containing fauxtography images or links to news stories using fauxtography receive the same levels of engagement. Twitter users seem more resistant to engaging with links to news stories that use fauxtography, but more likely to engage with tweets containing fauxtography images themselves. Reddit users were more likely to engage with any fauxtography related content.

These differences pose interesting problems for social media platforms; for example, fact-checking efforts that focus on links to news articles (some of which have been implemented by Twitter) are likely to have little effect on the spread of fauxtography in general as the images themselves still achieve relatively high engagement.

5 RQ2: Fauxtography's Evolutionary Nature

Previous work has indicated that memes exhibit some evolutionary properties, with new variants frequently emerging. Since, by definition, our fauxtography dataset includes images that have spread wide enough to warrant fact checking, we posit that some might have found life beyond fauxtography. Thus we ask: do fauxtography images become memes with different variants?

To answer this question, we relax our distance threshold used to detect instances of fauxtography images from 6 to 8 and examine the resulting images matches. We further focus on fauxtography images that were labeled only False, to focus on the role of fauxtography in spreading false information. We find 238 source images labeled "False" from our Snopes dataset that appear at least once on all Twitter, Reddit, and 4chan. We note that although measuring engagement on 4chan is problematic enough that we do not include details in Section 4 , 4chan is a key player in the meme ecosystem; thus we include it in this analysis.

For each source image, we manually determine whether each image within distance 8 is a variant. Table 3 provides details on the number of instances of variants across each platform. We observe that, of the 238 source images appearing on all three platforms, there were an additional 162 images on Twitter within distance 8 that we confirmed were indeed a match for a source image. Of these 162, 86 were sufficiently different from the source image to be deemed a variant, while 76 were essentially the same as the source image (i.e., they can be considered false negatives due to our threshold selection). An additional 1,291 images with distance 8 were completely unrelated (i.e., true negatives). For each of the three platforms, we see relatively similar numbers.

5.1 Case Studies

We find that 13 source images have variants that appear at least once on all three platforms we study (although not necessarily the same variant). A manual inspection shows that variants of these 13 source images correspond to memes. We examine three representative and particularly well-known cases in Figure 13. Our intuition is that particularly powerful fauxtography images are likely to take on a life of their own and become memes.

The first source image (Figure 13(a)) is a picture of Al Franken inappropriately touching Leeann Tweeden's breasts while she slept. The image is real, and was taken in 2006 on a C-17 cargo plane on their return from a USO tour in Afghanistan. This source image played a crucial role in then US Senator Al Franken's retirement from politics. The image was particularly controversial due it coming to light at the height of the #MeToo movement [11] as well as claims that it was related to a sketch that had been performed on the USO tour. The image is labeled as a false instance of fauxtography due to a widely circulated claim that the photographer that took the picture said it was staged. However, Franken fully admitted to the picture to be be real and not staged, accepted responsibility for what was ultimately irresponsible behavior, and resigned.

The variant of this image on Twitter (Figure 13(b)) replaces Franken's face with that of Roy Moore, an Alabama political

Platform	#CommonFalse-NoRating Images	#FalseImages with variation	#FalseImages w/o any variation	#FalseImages w/o variation-RandomImages	#FalseImages with variation-SameImage
Twitter	162	86	76	1291	70
Reddit	145	70	75	625	63
4chan	58	25	33	269	117

Table 3: Statistics for false images variations in Twitter, Reddit, and 4chan.

(a) Al Franken source image (b) Al Franken Twitter variant (c) Al Franken Reddit variant (d) Al Franken 4chan variant

(e) Trump Marine 1 original (f) Marine 1 Twitter variant (g) Marine 1 Reddit variant (h) Marine 1 4chan variant

(i) GWB reading book original (j) GWB Twitter variant (k) GWB Reddit variant (l) GWB 4chan variant

Figure 13: Variations of three common False images on all three platforms.

figure that lost a hotly contested race against Democrat Doug Jones for Jeff Sessions's US Senate seat after he was appointed US Attorney General. The text added to the image is related to allegations of sexual assault and pedophilia by Roy Moore, and Gloria Aldred's involvement in the incident. The variation on Reddit (Figure 13(c)) is much less political, merely replacing Franken's head with that of 80s sitcom character Alf.

On 4chan (Figure 13(d), the variant replaces Franken's face with with that of Donald Trump, replaces Leeann Tweeden's head with Stephen Colbert's head, and places two scoops of ice cream over Tweeden's breasts. This is likely in reference to Stephen Colbert's comments on sexual harassment [26] and Trump's alleged routine of receiving two scoops of ice cream for desert when everyone else at the table receives only one [18] (e.g., Colbert's nickname for Trump, "Donnie Two Scoops").

The second source image Figure 13(e) shows a rear view of Donald Trump entering Marine One. Based on Snopes this image is a "slightly manipulated" image (Trump's posterior has been enhanced) originally taken by a Reuters' photographer while president Trump boarded Marine One at Joint Base Andrews in Maryland. This photo was generally considered unflattering for Trump, as can be seen in the Twitter variant (Figure 13(f)), which uses Trump's buttocks to replace the two "Os" in Google's logo. This is indicative of some of the derision expressed online towards Trump's physical appearance. The Reddit variant (Figure 13(g)) introduces Vladimir Putin embracing Trump by "grabbing his butt." The 4chan variant (Figure 13(h)) is somewhat different, and replaces the saluting Marine guard with a saluting North Korean soldier with a rifle slung over his shoulder. The end of the rifle barrel is depicted as being inserted into Trump's buttocks.

The final source image we examine (Figure 13(i)) shows George W. Bush at a book reading at school in Houston in 2002. Snopes labels it as false because a manipulated version showing Bush holding the book upside down with a false caption was being spread on the Web. On Twitter (Figure 13(j)), we see a variant that has a non-manipulated version of the image, but has added text that implies Bush is telling the student about how right when the world needed it, "Q" (from the

Qanon conspiracy) appeared to save us all. The variant that appears on Reddit (Figure 13(k)) is the manipulated variant where it appears Bush is holding the book upside down. Finally, we see a variant on 4chan (Figure 13(l)) that uses the manipulated version with the upside down book, and adds large arrows pointing to the stars behind the students along with the text "WTF!" We are not entirely sure what this variant is trying to express, but based on our understanding of 4chan, we suspect it is conspiracy theory related.

5.2 Takeaways

Fauxtography is a complicated issue, in large part due to its visual nature and the Web's propensity for not just spreading visual information, but modifying it. The Franken image, which is not altered in any way and has a known provenance, is easily exploited for uses completely unrelated to its use in fauxtography. Similarly, the Bush image shows that even relatively innocuous pictures manipulated in subtle ways can become further manipulated to politicize them. The Trump image shows how even slight manipulations of real photos can elicit numerous meme variants.

This raises serious concerns about how to mitigate the relatively low-tech problem of fauxtography. For example, none of the variants we found were particularly convincing in terms of being real photos; the majority were very clearly manipulated, as is common for memes. What is there to fact check about a fictional TV alien groping a sleeping woman, after all? However, these variants tend to carry the same fundamental idea as the source image that *was* fact checked, and thus can still cause damage. Although issues like this warrant future exploration, at minimum, they calls into question the efficacy of fact checking *visual* mis/disinformation.

6 Related work

In this section, we review previous work on approaches to study and counter broad disinformation efforts and, more closely, on the use of images in mis/disinformation.

(Textual) Disinformation. A large body of research has studied disinformation on social media, with a specific focus on textual content. [28] show that fake news spread faster than true news on Twitter. By investigating the discussions on mass shooting events on Twitter, [25] reveals that alternative news outlets actively propagate alternative narratives, while [30] study information operations through the lens of the "Aleppo Boy" narrative, and show that some news media collaborate to spread alternative narratives. Also, [34] analyze disinformation campaigns carried out by state-sponsored actors, characterizing their influence on social networks, while [13] analyze user comments to characterize the public's (dis)belief towards news items. [10] survey users consuming news on social network and find that both sources and content play key roles in how they evaluate news veracity. Aiming to detect and counter disinformation, researchers have often relying on machine learning classification [23, 31, 24, 6, 29]. Also, [4] introduce a framework to evaluate the performance of different fake news classification models. For a comprehensive review of work in this space, please refer to [14].

Image-based Disinformation. More recently, the research community has begun to look at the interplay between images and disinformation. [12] collect and analyze disinformation images in India from WhatsApp, while [8] present a pipeline to extract themes and sentiments conveyed in images, and highlight several instances where images were used to share disinformation. [32] study image memes, showing that they are often used to spread political and hateful content. [9] also focus on memes containing both images and text and find that a third of them are related to politics, also confirming how memes are shared to spread disinformation as well as conspiracy theories. Finally, [33] show that Russian-sponsored trolls actively shared politically charged images on Twitter, and that these also influence other social platforms like Reddit, 4chan, and Gab.

Prior work has also studied fauxtography, aiming to detect false images. [36] build a fauxtography detector called "FauxBuster" based on machine learning techniques, while [3] use deep learning to detect manipulated images. Similarly, [39] extract various features from images and text, and use machine learning to assess the authenticity of specific claims. Furthermore, they describe which features are the most effective in verifying the authenticity of the claims.

Remarks. To the best of our knowledge, our paper is the first to study the effect that fauxtography images have on user engagement on social media, as well as to measure how these images are discussed and shared on different online services.

7 Discussion & Conclusion

In this paper, we presented a data-driven study of fauxtography on social media. We found that including fauxtography in social media posts increases user engagement, irrespective of the verisimilitude of the fauxtography image. This highlights the need to take images into account when developing disinformation mitigations. At the same time, we showed that fauxtography images are often taken out of context and turned into memes, which highlights the challenges faced in automatically identifying image-based disinformation.

Next, we discuss the implications of our findings and highlight some limitations of our study.

Implications of our findings. The fact that sharing fauxtography on social media increases user engagement highlights how image-based disinformation cannot be overlooked, and that any effort to curb the problem should take not only text into account, but also images. At the same time, we showed that fauxtography images are often used as memes on social media, blurring the line between the intention to mislead and satire. This opens up a number of problems when moderating fauxtography, since it is challenging to automatically determine the intention with which an image is posted, which is often context specific. Crucially, our study also highlighted the fact that the verisimilitude of fauxtography images does not have an impact on the engagement that they receive. This suggests that the "clickbait" power of these images is what drives

engagement, and raises questions on the effectiveness of mitigations based on fact-checking labels and user warnings.

Limitations. Naturally, our study is not without limitations. First, our image analysis pipeline allows us to identify images that are very similar to fauxtography images, but is unable to verify if the image is used in the misleading setting flagged by Snopes. For example, we are unable to tell if miscaptioned images are being used in a miscaptioned context. Similarly, for manipulated photos, it is possible that our analysis pipeline identifies the unmodified picture as a fauxtography one. This motivates future work combining our analysis pipeline with semantic analysis techniques to study the context in which fauxtography is used. Third, our identification of news outlets using the top 30K Majestic websites excludes many small local news outlets. Since we expect that local news outlets have less fastidious fact-checking as compared to larger venues, this suggests our analysis will tend to underestimate the spread of fauxtography on the Web.

Additionally, collecting images at scale from the Web present challenges. In particular, we found that many images were no longer available when we attempted to download them. Still, we believe that the scale of our dataset is large enough to allow us to gain a comprehensive view of the use of fauxtography on social networks.

Acknowledgements This work was partially supported by the Natural Science Foundation under grant CNS-1942610 and by a BU College of Engineering Dean's Catalyst Award. B.B. acknowledges the support of the National Science Foundation under grant no. DMR-1945058.

References

[1] C. Abidin. Meme factory cultures and content pivoting in Singapore and Malaysia during COVID-19. *Harvard Kennedy School Misinformation Review*, July 2020.

[2] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn. The Pushshift Reddit Dataset. In *ICWSM*, 2020.

[3] B. Bayar and M. C. Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *ACM IH*, 2016.

[4] L. Bozarth and C. Budak. Toward a better performance evaluation framework for fake news classification. In *ICWSM*, 2020.

[5] J. Buchner. A python perceptual image hashing module: Imagehash, 2020. https://github.com/JohannesBuchner/imagehash.

[6] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *WWW*, 2011.

[7] S. Cooper. A Concise History of the Fauxtography Blogstorm in the 2006 Lebanon War. *American Communication Journal*, 9, 2007.

[8] P. Dewan, A. Suri, V. Bharadhwaj, A. Mithal, and P. Kumaraguru. Towards Understanding Crisis Events On Online Social Networks Through Pictures. In *ASONAM*, 2017.

[9] Y. Du, M. A. Masood, and K. Joseph. Understanding visual memes: An empirical analysis of text superimposed on memes shared on twitter. In *ICWSM*, 2020.

[10] M. Flintham, C. Karner, K. Bachour, H. Creswick, N. Gupta, and S. Moran. Falling for fake news: investigating the consumption of news via social media. In *ACM CHI*, 2018.

[11] M. Garber. Al Franken, That Photo, and Trusting the Women. https://www.theatlantic.com/entertainment/archive/2017/11/al-franken-thatc-and-trusting-the-women/545954/, 2017.

[12] K. Garimella and D. Eckles. Images and misinformation in political groups: Evidence from whatsapp in india. *arXiv:2005.09784*, 2020.

[13] S. Jiang, M. Metzger, A. Flanagin, and C. Wilson. Modeling and measuring expressed (dis) belief in (mis) information. In *ICWSM*, 2020.

[14] S. Kumar and N. Shah. False information on web and social media: A survey. In *arXiv:1804.08559*, 2018.

[15] Y. Li and Y. Xie. Is a picture worth a thousand words? an empirical study of image content and social media engagement. *Journal of Marketing Research*, 57(1):1–19, 2020.

[16] B. Lindgren. *Statistical Theory*, volume 22. 1993.

[17] Majestic. The Majestic Million List. https://majestic.com/reports/majestic-million, 2019.

[18] D. Mercia. Trump gets 2 scoops of ice cream, everyone else gets 1 – and other top lines from his Time interview. https://www.cnn.com/2017/05/11/politics/trump-time-magazine-ice-cream/index.html, 2017.

[19] V. Monga and B. L. Evans. Perceptual image hashing via feature points: performance evaluation and tradeoffs. *IEEE transactions on Image Processing*, 15(11), 2006.

[20] NewsGuard. The Internet Trust Tool. https://www.newsguardtech.com/, 2019.

[21] Newspaper3k. Article scraping & curation. https://newspaper.readthedocs.io/en/latest/, 2013.

[22] A. Papasavva, S. Zannettou, E. De Cristofaro, G. Stringhini, and J. Blackburn. Raiders of the Lost Kek: 3.5 Years of Augmented 4chan Posts from the Politically Incorrect Board. In *ICWSM*, 2020.

[23] K. Shu, D. Mahudeswaran, S. Wang, and H. Liu. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *ICWSM*, 2020.

[24] 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*, 19(1), 2017.

[25] K. Starbird. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *ICWSM*, 2017.

[26] S. Van Hoozer and S. Peuchaud. "Speaking of Sexual Harassers Who Should Resign Tomorrow... Donald Trump": A Feminist Rhetorical Analysis of Stephen Colbert's Late Show Monologues. *The Journal of Popular Culture*, 53(1):34–57, 2020.

[27] Virus Total. https://www.virustotal.com, 2020.

[28] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359(6380), 2018.

[29] W. Y. Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *arXiv:1705.00648*, 2017.

[30] T. Wilson, K. Zhou, and K. Starbird. Assembling strategic narratives: Information operations as collaborative work within an online community. In *CSCW*, 2018.

[31] L. Wu and H. Liu. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *WSDM*, 2018.

[32] S. Zannettou, T. Caulfield, J. Blackburn, E. De Cristofaro, M. Sirivianos, G. Stringhini, and G. Suarez-Tangil. On the origins of memes by means of fringe web communities. In *ACM IMC*, 2018.

[33] S. Zannettou, T. Caulfield, B. Bradlyn, E. De Cristofaro, G. Stringhini, and J. Blackburn. Characterizing the use of images in state-sponsored information warfare operations by russian trolls on twitter. *arXiv:1901.05997*, 2019.

[34] S. Zannettou, T. Caulfield, W. Setzer, M. Sirivianos, G. Stringhini, and J. Blackburn. Who Let The Trolls Out?: Towards Understanding State-Sponsored Trolls. In *WebSci*, 2019.

[35] S. Zannettou, J. Finkelstein, B. Bradlyn, and J. Blackburn. A quantitative approach to understanding online antisemitism. In *ICWSM*, 2020.

[36] D. Y. Zhang, L. Shang, B. Geng, S. Lai, K. Li, H. Zhu, M. T. Amin, and D. Wang. Fauxbuster: A content-free fauxtography detector using social media comments. In *IEEE Big Data*, 2018.

[37] D. Zillmann, R. Gibson, and S. L. Sargent. Effects of photographs in news-magazine reports on issue perception. *Media Psychology*, 1(3), 1999.

[38] D. Zillmann, S. Knobloch, and H.-s. Yu. Effects of photographs on the selective reading of news reports. *Media Psychology*, 3(4), 2001.

[39] D. Zlatkova, P. Nakov, and I. Koychev. Fact-checking meets fauxtography: Verifying claims about images. In *EMNLP-IJCNLP*, 2019.