Fake Tweet Buster: A Webtool to Identify Users Promoting Fake News on Twitter

Diego Saez-Trumper
Universitat Pompeu Fabra
Barcelona, Spain
dsaez-trumper@acm.org

ABSTRACT

We present the "Fake Tweet Buster"[1] (FTB), a web application that identifies tweets with fake images and users that are consistently uploading and/or promoting fake information on Twitter. To do that we mix three techniques: (i) reverse image searching, (ii) user analysis and (iii) a crowd sourcing approach to detected that kind of malicious users on Twitter. Using that information we provide a credibility classification for the tweet and the user.

Categories and Subject Descriptors

H.4.2 [**Types of Systems**]: Decision support; H.3.3 [**Information Storage and Retrieval**]: Information Filtering

General Terms

Experimentation, Human factors

Keywords

Social networks; Fake tweets; Fake Photos; News; Credibility; Webtool; Tools

1. INTRODUCTION

Nowadays social media, and more specifically Twitter, is one of the Battlefields for political confrontations. Although that the majority of the people involved in these discussions are using this tool to express their opinions and support their positions in a good manner, there are also people using dirty tricks to win this battle. Specifically, we focus on users that deliberately post fake photos of news. This phenomeno is increasing in the last years, and the goal of those "fakers" is to delegitimize their opponents or to show that their positions have more support that it really has[2]. The usual way to do this, is to take an old photo (e.g. a photo of a demonstration in country X from two years ago), and present it as new (e.g. say that is a demonstration in country Y, right now). The sophistication of these fake photos can include some retouch or cropping, but often is just the same photo with a new "title". However, without knowing the context (sometimes people in other countries are following conflicts through Twitter) could be difficult to know if those photos are fake or not. Futhermore, sometimes these photos become viral, and even some newspapers can publish them [1]. Therefore, there are two types of people publishing this kind of images: malicious users that intentionally post fake information, and naive users that believes on that information.

The goal of the "Fake Tweet Buster" (FTB) application is to help normal users to check the credibility of a Tweet and also to detected malicious users that are consistently uploading and/or promoting fake information on Twitter. To that, we propose a online application, very easy to use (just copy/paste the tweet url), that returns a credibility score for a given tweet.

[1] `http://grupoweb.upf.edu/fake-buster`

HT'14, September 1–4, 2014, Santiago, Chile.
ACM 978-1-4503-2954-5/14/09.
http://dx.doi.org/10.1145/2631775.2631786.

2. RELATED WORK

In 2010 Mendoza *et al.*[8] study the tweets during the earthquake in Chile, finding that rumors propagates in a different way than the actual news. In 2011 Castillo *et al.* [5] studied tweets' credibility using posting behavior, text characteristics, and the citation of external sources, to classify tweets as credible or not credible.

Some user's characteristics such as are account's age, number of tweets and followers has been also found to be correlated with tweets credibility[6, 7]. Although that the FTB application focus in tweets with photos, exploding information's, we also take in account the users characteristics.

To the best of our knowledge, beyond the research done in this area, there are no online services available to detect users that are intentionally posting fake photos on twitter.

3. HOW IT WORKS?

"Fake Tweet Buster" works as follows:

- **Step 1:** The user copy/paste a Tweet url or and Tweet account URL.
- **Step 2:** FTB returns a panel showing similar images, a small text description of the image (See the "Best guess for this image" description in Reverse image search description), the original twitter account that posted that image, the account history (what

Figure 1: A photo from the student protest in Chile in 2011 that was *tweeted* as a Venezuelan demonstration in 2014.

other FBT users opined about that account) and a credibility score based on those informations.

- **Step 3:** FTB user is asked to tag that tweet as Fake or Legitimate.

We use three approaches to identify whether a tweet is legitimate or fake:

- **Reverse image search:** It is a technique that uses a image as query, and the search engine returns similar images. Reverse image search can be used to find out where a photo comes from. For FTB we use two popular reverse image search services: Google Images [3] and TinEye [4]. Google Images has a feature called "Best guess for this image" that returns the textual query that matches better with the image. TinEye, that has a smaller database than Google (i.e. less images to compare with), allows to sort the results by date, allowing to discover the image's "age". Using the two results (text related and image's age), we can know if the photo is new or old, and in which context it appears. Therefore, we can estimate if the photo seems to be related with the tweet or not. For example, imagine a Tweet with a photo claiming that this image is a demonstration in Country X, in 2014. Now, applying reverse image search, we found that the "Best guess for this image" does not refer to Country X, but to Country Y, and that the photo is from 2012. Hence, that photo seems to be fake.
- **User analysis:** Using the Twitter API we obtain the number of tweets, number followers, and the age of an account to determinate if this is suspicious or not. New accounts with an small amount of tweets and followers tends to be more suspicious that an old account with a lot of tweets and followers.
- **Crowd sourcing:** We include a panel with all the information about the photo and the twitter account that posted it and then ask FTB's users to tag that account as fake or legitimate.

4. FUTURE WORK

We will improve our model using the crowd sourced information, that will be a dataset of manually tagged (as fakers or legitimate) accounts. We also want to implement the methodology described in [5] to use the tweet text and posting behavior to classify tweets as credible or not.

4.1 Acknowledgements

This work has been partially funded by the "Understanding Social Media: An Integrated Data Mining Approach" (TIN2012-38741) project from the Spanish Economy and Competitiveness Ministry.

5. REFERENCES

[1] Austrian Newspaper Apologizes for Fake Syria Photo. `http://www.imediaethics.org/News/3284/Austrian_newspaper_apologizes_for_fake_syria_photo.ph`.

[2] 'Fake' Images Shared on Venezuela Protests. `http://www.bbc.com/news/magazine-26258335`.

[3] Google images. `http://images.google.com`.

[4] Tineye. `http://www.tineye.com`.

[5] C. Castillo, M. Mendoza, and B. Poblete. Information Credibility on Twitter. In *Proceedings of World Wide Web Conference (WWW)*, pages 675–684. ACM Press, Feb. 2011.

[6] A. Gupta and P. Kumaraguru. Credibility ranking of tweets during high impact events. In *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, PSOSM '12, pages 2:2–2:8, New York, NY, USA, 2012. ACM.

[7] A. Gupta, H. Lamba, and P. Kumaraguru. $1.00 per RT #BostonMarathon #PrayForBoston: Analyzing Fake Content on Twitter. In *Eigth IEEE APWG eCrime Research Summit (eCRS)*, page 12. IEEE, 2013.

[8] M. Mendoza, B. Poblete, and C. Castillo. Twitter under crisis: can we trust what we rt? In *Proceedings of the first workshop on social media analytics*, pages 71–79. ACM, 2010.