Facebook. Because studying events and spammers on Twitter is too mainstream.

Prateek Dewan, Ponnurangam Kumaraguru
Indraprastha Institute of Information Technology - Delhi (IIITD)
{prateek, pk}@iiitd.ac.in

ABSTRACT

Facebook and Twitter are two of the biggest social networks in the world, with a combined audience covering over 20% of the world's total population. While multiple studies in the past have studied Twitter during real-world events, very little work has been done on analyzing Facebook. Given the recent introduction of features like hashtags and searchable public posts on Facebook, more content on Facebook is becoming public, and hence luring from spammers' point of view. In this work, we study the most active users, and the content posted by them on Twitter and Facebook during 12 real-world events, and present a comparative analysis between these two social media platforms during the 12 events. We analyzed 186 Twitter users and 153 Facebook users who were most active during these events, and found that Twitter had more spammers among the most active users than Facebook. Further, we show that spam on Facebook is highly unrelated to the real world events during which it is posted. We believe that these findings will motivate researchers to dig deeper into Facebook's public side in the future.

1. INTRODUCTION

With more Facebook users than newspaper readers in most major countries of the world, online social media has stamped it's authority as one of the largest information and news propagators on the Internet, and perhaps, the entire world. [1] [2] Today, people across the globe resort to social media platforms like Twitter and Facebook when it comes to spreading or learning about breaking news like floods, fire, bomb blasts, earthquakes, public shootouts, terror attacks etc. Twitter, in particular, has been widely studied by researchers as a news medium during real-world events [1, 5, 6, 7, 8]. However, few studies have looked at social media platforms other than Twitter to study news and real-world events [4]. With more than double the number of monthly active users as Twitter, Facebook can be conjectured as an important platform for news and information dissemination during real-world events [3].

Facebook is currently, the largest online social network in the world, having more than 1 billion monthly active users. However, unlike Twitter, Facebook's fine-grained privacy settings make majority of its content private, and publicly inaccessible. The private nature of Facebook has been a major challenge in collecting and analyzing its users, content and network in the computer science research community. But even with a small percentage of content being public, the mere volume of this publicly available content makes Facebook a rich source of information. Recent introduction of features like hashtag support and Graph search for posts, have largely increased the level of visibility of public content on Facebook, either directly or indirectly. [3] [4] Users can now *search* for topics and hashtags to look for content, in a fashion highly similar to Twitter; thus making the public Facebook content more visible and consumable by it's users. This increasing public visibility, and an enormous user-base potentially makes Facebook one of the largest and most widespread sources of real time news and information on the Internet.

In this work, we highlight the aforementioned capability of Facebook as a news medium, and present a comparative analysis between the most active users on Facebook and Twitter during 12 real-world events. In particular, we analyze the malicious and spam content posted by the most active users during these events. We then characterize malicious users who post spam and non-relevant content during such events to deteriorate the quality of information.

2. METHODOLOGY

We used the MultiOSN framework [2] for collecting event specific data from multiple social media platforms. MultiOSN uses REST based, keyword search API for collecting public Facebook posts, and Twitter's search and streaming APIs for collecting public tweets. The 12 events analyzed were: Indian Premier League - IPL, Creation of a new Indian state (Telangana), Kashmir Earthquake, Washington Navy Yard Shooting, Mothers' Day Shooting, Boston Blasts, London Terror Attack, Birth of the Royal Baby, Champions Trophy cricket tournament, Iran Earthquake, Floods in Uttarakhand (Northern India), and Oklahoma Tornado.

Ideally, analyzing this entire data would have been a good approach to compare the content of both the social network during the 12 events. However, such amount of data would have been extremely difficult to clean, manage, annotate, and analyze together. We thus decided to analyze content posted by only the most active users during these events.

[1] http://www.mapsofworld.com/world-top-ten/countries-with-highest-newspapers-map.html
[2] http://en.wikipedia.org/wiki/Facebook_statistics

[3] http://newsroom.fb.com/News/728/Graph-Search-Now-Includes-Posts-and-Status-Updates
[4] https://www.facebook.com/help/587836257914341

.

Category	Facebook	Twitter
Users	**23**	**180**
Pages	**122**	-
Verified	5	6
Suspended	-	**5**
Deleted / Not found	**8**	**1**
Spammers	11	23
False positives	21	28
Total	**153**	**186**

Table 1: Detailed statistics of the most active users on Facebook and Twitter captured during 12 real-world events.

Extracting most active users.

To extract the most active users, we calculated the *information gain* associated with each user present in our dataset. The *information gain* associated with a user was calculated by the fraction of information added by the user through the number of posts he / she made during the event. For each event, and for each social media, we sorted the users in decreasing order of the number of posts by made them during the event. If the percentage of information contributed by a given user U_i, with respect to the information already contributed by users who are more active than U_i, is more than a given threshold; we pick U_i as one of the most active users. Table 1 represents the detailed statistics of the most active users extracted from this data.

Identifying spammers.

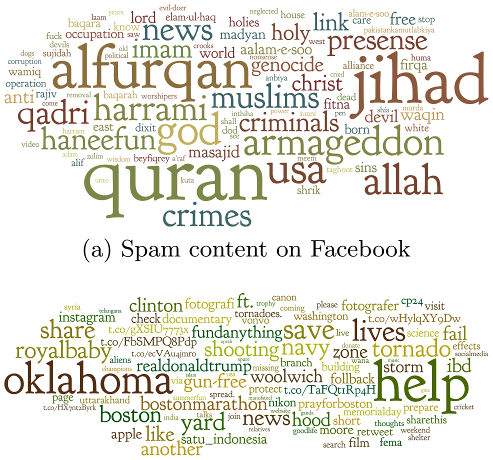
We manually went through all the content posted by the most active users identified from the previous step, and marked each user as spam or non-spam. Users posting highly irrelevant and / or repetitive content were marked as spammers. We also encountered numerous false positive accounts in our data. For example, Twitter accounts posting about traffic updates got captured during the Royal Baby Birth event, since Prince William is also the name of a street in the USA. Similarly, posts from some confession pages, and pages offering jobs were captured during the Telangana event, due to the presence of locations like Hyderabad, and Telangana. Such accounts were marked as false positive, and were not considered as spam.

3. SPAM ANALYSIS AND DISCUSSION

From the most active users, we identified 11 Facebook accounts, and 23 Twitter accounts as spammers (Table 1). Out of the 11 Facebook accounts, 5 were user profiles, 4 were pages, and 2 accounts were deleted / not found.

Figure 1 shows a tag cloud of the top 100 most frequently occurring terms in the content posted by spammers on Facebook and Twitter separately. As evident, spam on Facebook is very different from Twitter spam. Facebook spam reflects radicalization and propaganda, especially with respect to Islam; and is totally unrelated to the events during which it is being talked about. However, Twitter spam is highly related to the events during which it is posted, and hence, hard to differentiate from genuine content via automated means.

On Facebook, Royal Baby birth was the most spammed event, with 6 out of the 11 users spamming during the event. Telangana, and Champions Trophy events saw 3 and 2 spammers respectively. On Twitter however, Boston Blasts was the most spammed event, with 7 spammers. London Terror Attacks followed next, with 6 spammers. Interestingly, 2 Twitter users were found to be spamming across more than one event. Both these accounts are still active. Apart from the IPL and two earthquake events, all other events saw spammers on Twitter.



(a) Spam content on Facebook

(b) Spam content on Twitter

Figure 1: Spam content on Facebook and Twitter posted by most active users during 12 real-world events.

This is an ongoing project, results presented in this paper are highly primitive and exploratory. Our goal is to get deeper insights into such malicious content and users on multiple online social media services during real world events, and come up with techniques to automatically segregate good quality content and users from the bad ones. We intend to leverage and aggregate content and features from multiple social media services in order to achieve our objectives.

4. REFERENCES

[1] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on twitter. In *ICWSM*, 2011.

[2] P. Dewan, M. Gupta, K. Goyal, and P. Kumaraguru. Multiosn: realtime monitoring of real world events on multiple online social media. In *5th IBM ICARE*. ACM, 2013.

[3] Diffen.com. Facebook vs twitter. *http://www.diffen.com/difference/Facebook_vs_Twitter*, 2013.

[4] S. Hille and P. Bakker. I like news: Searching for the holy grail of social media: The use of facebook by dutch news media and their audiences. *European Journal of Communication*, page 0267323113497435, 2013.

[5] M. Hu, S. Liu, F. Wei, Y. Wu, J. Stasko, and K.-L. Ma. Breaking news on twitter. In *CHI*, pages 2751–2754. ACM, 2012.

[6] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW*, pages 591–600. ACM, 2010.

[7] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW*, pages 851–860. ACM, 2010.

[8] J. Weng and B.-S. Lee. Event detection in twitter. In *ICWSM*, 2011.