

Detecting Collective Attention Spam

Kyumin Lee, James Caverlee, Krishna Y. Kamath, Zhiyuan Cheng
Department of Computer Science and Engineering
Texas A&M University
College Station, TX 77843
{kyumin, caverlee, kykamath, zcheng}@cse.tamu.edu

ABSTRACT

We examine the problem of *collective attention spam*, in which spammers target social media where user attention quickly coalesces and then collectively focuses around a phenomenon. Compared to many existing spam types, collective attention spam relies on the users themselves to seek out the content – like breaking news, viral videos, and popular memes – where the spam will be encountered, potentially increasing its effectiveness and reach. We study the presence of collective attention spam in one popular service, Twitter, and we develop spam classifiers to detect spam messages generated by collective attention spammers. Since many instances of collective attention are bursty and unexpected, it is difficult to build spam detectors to pre-screen them before they arise; hence, we examine the effectiveness of quickly learning a classifier based on the first moments of a bursting phenomenon. Through initial experiments over a small set of trending topics on Twitter, we find encouraging results, suggesting that collective attention spam may be identified early in its life cycle and shielded from the view of unsuspecting social media users.

Categories and Subject Descriptors: H.3.5 [Online Information Services]: Web-based services; J.4 [Computer Applications]: Social and behavioral sciences

General Terms: Design, Experimentation, Security

Keywords: collective attention, spam, social media

1. INTRODUCTION

With the emergence of global-scale social media, we have seen repeated evidence of breaking news, viral videos, and popular memes captivating the attention of huge numbers of users. For example, during the recent run-up and immediate aftermath of President Obama’s announcement about the raid targeting Osama Bin Laden, Twitter boasted a peak of 5,000 tweets per second (corresponding to 432 million tweets per day) and a sustained average rate of 3,000 tweets per second over several hours (corresponding to 259 million tweets

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WebQuality ’12, April 16, 2012, Lyon, France
Copyright 2012 ACM 978-1-4503-1237-0 ...\$10.00.

Amanda Knox: Murder on Trial VIDEO



Figure 1: Example YouTube video posted to attract interest after the Amanda Knox court decision.

per day).¹ On a lighter note, the ubiquitous “Charlie Bit My Finger” video has attracted over 420 million views on YouTube. Popular memes on sites like Reddit and 4chan have attracted huge audiences and have subsequently propagated throughout the web. Similarly, the death of Michael Jackson prompted huge spikes in search traffic on web and social media services for more information. These and related phenomenon are examples of *collective attention*, as Wu and Huberman have noted, in which “attention to novel items propagates and eventually fades among large populations” [20]. In the typical life cycle, an item (be it a video, news article, image, etc.) catches the interest of a few people, then accumulates a larger following as more people begin paying attention to it, before (in some cases) breaking out across social media to explosive attention, until finally

¹<http://blog.twitter.com/2011/03/numbers.html>

fading away. Naturally, there is a great demand for understanding these dynamics, modeling the life cycle of these phenomenon, and ultimately predicting the future growth of new items [4, 7, 15, 19].

Knowing that interest may quickly coalesce and then collectively focus around a particular phenomenon, we are increasingly seeing threats to the quality of information associated with this collective attention. To illustrate, consider the following three examples of what we refer to as *collective attention spam*:

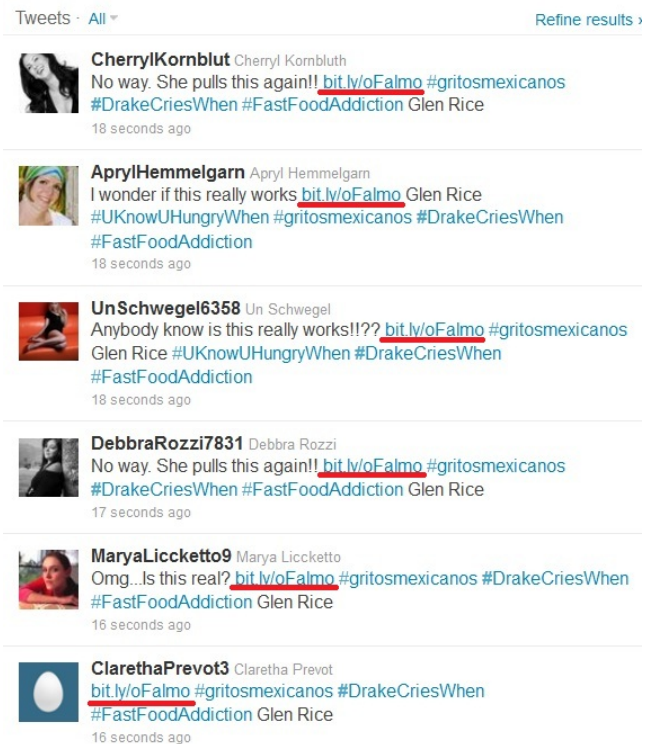
Example 1: YouTube Breaking Videos. When Amanda Knox was freed after appeal in an Italian court on October 3, 2011, many popular mainstream US news outlets posted breaking news updates and provided extensive coverage of the high-interest case. Knowing that interest in Amanda Knox would skyrocket, spammers began posting videos to YouTube that were tagged with keywords associated with Amanda Knox, but that were expressly designed to promote an unrelated spammer-controlled website. In the immediate aftermath of the court announcement, we found that four of the top-six videos returned for the YouTube query “Amanda Knox” were examples of such *collective attention spam*. Figure 1 shows one example, which includes a link ostensibly to view more Amanda Knox related videos.

Example 2: Twitter Trending Topics. To give some insight into the current pulse of the real-time web, Twitter posts a selection of the current trending topics. These popular topics (typically hashtags or keywords) signal to spammers what Twitter users are currently collectively interested in, and so spammers can easily target these popular topics by posting spam messages containing the trending hashtags or keywords. Figure 2 shows a sample search result for the trending hashtag “#DrakeCriesWhen” for which the most recently posted six messages are all spam. Note that the six messages post the same URL, but from multiple accounts, suggesting that spammers are strategically posting to Twitter in an organic-like way to simulate the behavior of regular (non-spam) users.

Example 3: Popular Facebook Profiles. As the most popular social media site, Facebook – with over 800 million users – attracts large interest to the celebrity profiles hosted on the site. Knowing that many users will naturally visit these popular profile pages, particularly when the celebrity is in the news, we have encountered many examples of spam photos being posted to these profiles (since many popular profile pages support photo uploads by fans). For example, Figure 3 shows a Facebook photo associated with the M&M’s candy official page; the photo is clearly unrelated to M&M’s and includes a spam URL.

In contrast to many examples of more traditional spam, *collective attention spam relies on the users themselves to seek out the content where the spam will be encountered*. In this way, users themselves have self-selected for interest in the topic and made themselves susceptible to collective attention spam. As a counterpoint, consider email spam, where the spammer must identify a group of targets, send a spam payload embedded in an email, and then hope that (if the email passes through all spam filters) the user ultimately decides to click on the link. Similarly, in many cases of social media spam, spammers target particular users with friend requests, post spam links on the target’s profile page, or

Results for #DrakeCriesWhen



The image shows a screenshot of a Twitter search result for the hashtag #DrakeCriesWhen. At the top, it says "Tweets · All" and "Refine results". There are six tweets listed, all of which are spam. Each tweet includes a profile picture, the user's name, their bio, the text of the tweet, and the time it was posted. The text of each tweet is identical: "No way. She pulls this again!! [bit.ly/oFalmo](#) #gritosmexicanos #DrakeCriesWhen #FastFoodAddiction". The users are: Cheryl Kornblut, Apryl Hemmelgarn, UnSchwegel6358, DebbraRozzi7831, Marya Liccketto9, and ClarethPrevot3. All tweets were posted 16-18 seconds ago.

Figure 2: Spam messages targeting the Twitter trending topic #DrakeCriesWhen.

send spam messages via the messaging services provided by the social media service. In all cases, the user must cross a fairly high threshold to respond positively toward the spammer (though of course the probability of response may be increased through spear phishing or other targeted attacks). In contrast, collective attention spam targets users who are inherently interested in the topic.

In the rest of the paper, we focus on two related efforts:

- First, we study the presence of collective attention spam in one popular service, Twitter, and examine the properties of collective attention spam including the longevity of spam accounts, the total amount of collective attention spam, and the properties of accounts engaged in such behavior; and
- Second, we develop a machine learning based spam classifier to detect spam messages generated by collective attention spammers. Since many instances of collective attention are bursty and unexpected, it is difficult to build spam detection algorithms to pre-screen them before they arise; hence, we examine the effectiveness of quickly learning a classifier based on the first moments of a bursting phenomenon.

Through initial experiments over a small set of trending topics on Twitter, we find encouraging results, suggesting that collective attention spam may be identified early in its life cycle and shielded from the view of unsuspecting social media users.

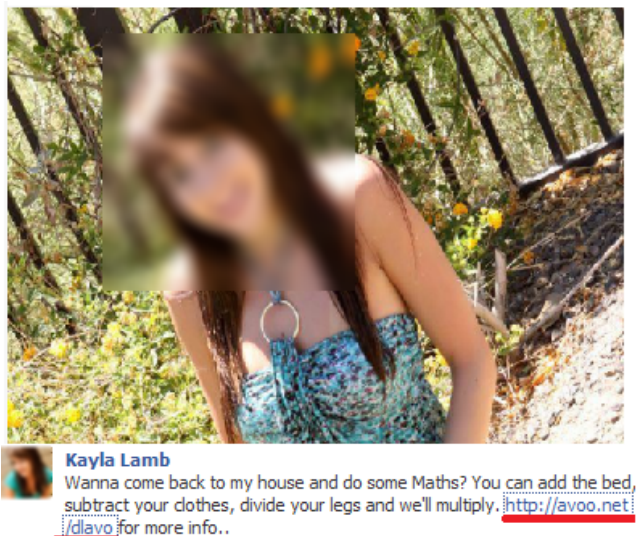


Figure 3: Photo posted to the M&M’s candy Facebook profile, including a spam link.

2. EXAMINING COLLECTIVE ATTENTION SPAM ON TWITTER

In this section, we study a sample of collective attention spam from Twitter, analyzing spammers and their tactics.

For Twitter-based trending topics, users both generate and consume messages around popular hashtags or keywords. Our working hypothesis is that spammers approach collective attention as in Figure 4, where spammers monitor breaking news, trends, and public issues for predicting which issues will attract significant collective attention in the future. Once selected, spammers insert spam messages into the popular topic, while user attention is focused on the topic. By inserting spam URLs into messages associated with these trending topics, the spammers hope is for users to click on these links.

2.1 Dataset

We sampled trending topic search results on Twitter during 11 days between September and October, 2011. We searched the trending topics every 5 minutes and collected the Twitter search engine’s recently posted messages associated with the trending topics, resulting in a dataset consisting of 5.3 million messages posted by 1.5 million users. While we were collecting the messages and the users, we periodically checked whether the users were suspended or not. When we accessed an account profile page, if the account is suspended, Twitter will redirect the page to <http://twitter.com/account/suspended>.

To determine if all suspended accounts could be treated as spam accounts, we sampled messages from each suspended account and labeled them by hand. We randomly sampled 200 messages each from the messages posted by suspended accounts and from those posted by non-suspended accounts. Two human judges manually labeled the 400 messages as either spam or non-spam. 199 out of 200 messages sampled from non-suspended accounts were labeled as non-spam messages, and 187 out of 200 messages sampled from suspended accounts were labeled as spam messages. Overall accuracy is

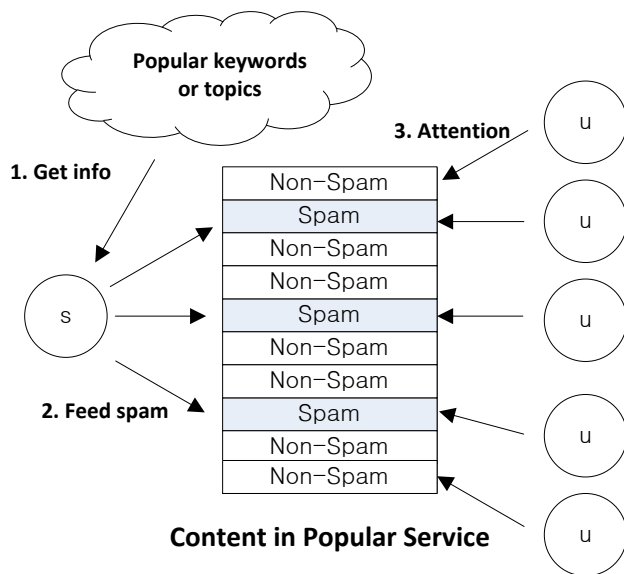


Figure 4: Collective attention spam.

0.965 and even though there are some errors, for the rest of the paper we assume that all messages posted by suspended users are spam.

2.2 Spammers and Their Tactics

From the dataset, we selected six popular topics to see how quickly spammers post spam messages once a topic becomes popular. Figure 5 depicts all messages and spam messages distribution over time in the six popular topics. We can observe that spammers quickly posted spam messages with a popular topic when it becomes popular. Spammers’ intention is to expose the spam messages to larger number of users who are interested in the trending topics, confirming the susceptibility of trending topics to collective attention spam. Overall, we find that an average of 4% of messages across the six popular topics is spam.

Properties of Spam Accounts. Next, we analyze the properties of the spam accounts to better understand the tactics of collective attention spammers. In the trending topic search dataset, 17,411 users were suspended by the Twitter safety team and these spammers posted 136,255 messages. We show in Table 1 the minimum, maximum, average and median value of the number of followings, number of followers, and number of messages posted by the spammers. Also the values of regular users in the dataset are shown in Table 2.

Table 1: Properties of Collective Attention Spammers

	# of Followings	# of Followers	# of Messages
Min	0	0	0
Max	67,579	189,805	228,634
Avg	104	183	567
Median	0	0	9

We see that 86% of spammers have fewer than 50 followings and fewer than 10 followers, posting fewer than 120 messages. These values are lower than the regular users’

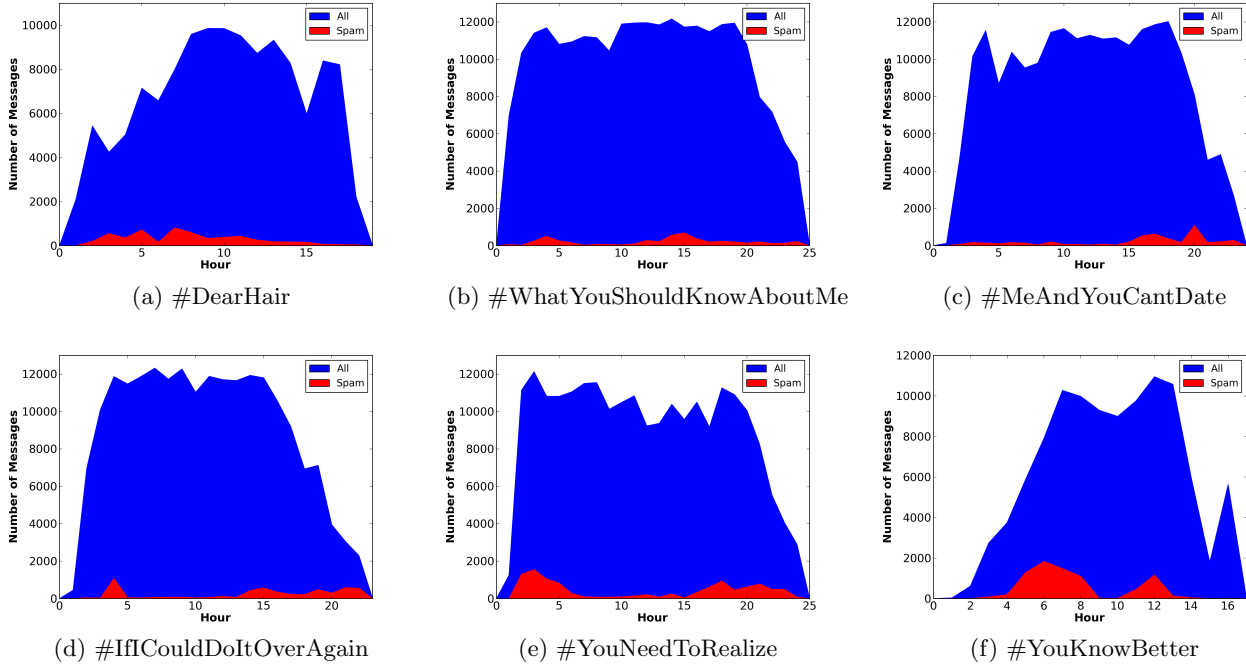


Figure 5: All messages and spam messages associated with each of six popular topics.

Table 2: Properties of Regular Users

	# of Followings	# of Followers	# of Messages
Min	0	0	0
Max	298,287	5,211,919	610,869
Avg	324	506	4,984
Median	156	116	1,625

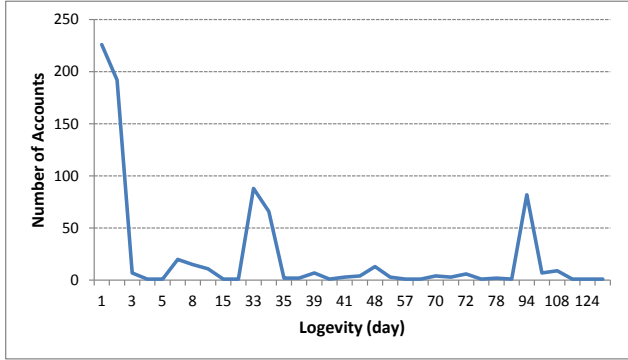


Figure 6: Longevity of spammers.

values. The most interesting part in Table 1 is the median in which number of following and number of followers are 0. More than half of the spammers only focused on posting more messages than making friends since their goal is to pollute the collective attention of Twitter users and not engage in social capital building as has been reported in previous studies of Twitter spam behavior [14].

Longevity. Another interesting property is the longevity of spammers; how long do spammers live? Are they newly-

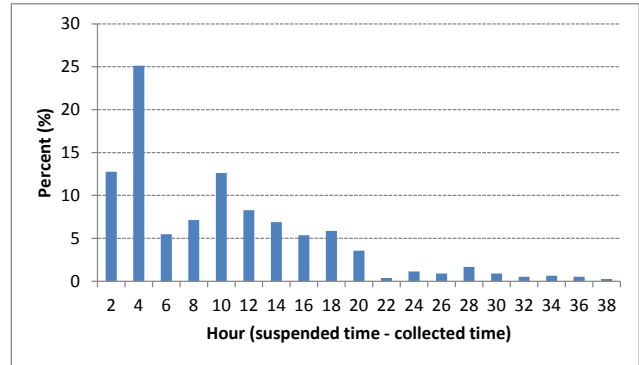


Figure 7: How long does it take for Twitter safety team to suspend spam accounts?

created accounts or long-lived repeated use spam accounts? Using the account creation time accessed via the Twitter API, we additionally record the account suspension time through our repeated checking of each account in the dataset. Once an account is declared suspended, we record the time. Figure 6 shows the longevity (suspended date - creation date). We see that 54% of the spam accounts lived for fewer than 3 days, and that 80% of the spam accounts were alive for at most 34 days. Interestingly we note that 18% of the spam accounts were alive for more than 46 days, with a large number living more than 94 days. These long-lived accounts suggest that spammers create wholesale accounts for strategic future deployment in anticipation of future bursting topics.

Spam Activation. Next, we study how long it takes for the Twitter safety team to suspend these accounts once

Table 4: Seven popular topics and their properties (period, total lifespan and # of messages).

Topic	Period	Total Lifespan	# of messages
#DearHair	2011-09-29 02:19 ~ 2011-09-29 19:25	17 hrs 06 minutes	128,499 (4.3% spam)
#TheyNeedToBringBack	2011-09-30 19:12 ~ 2011-10-02 04:59	33 hrs 47 minutes	222,176 (2.2% spam)
#WhatYouShouldKnowAboutMe	2011-10-01 05:11 ~ 2011-10-02 04:59	23 hrs 48 minutes	249,535 (2.1% spam)
#MeAndYouCantDate	2011-10-02 06:53 ~ 2011-10-03 04:42	21 hrs 49 minutes	209,343 (2.4% spam)
#IfICouldDoItOverAgain	2011-10-03 07:06 ~ 2011-10-04 04:55	21 hrs 49 minutes	202,159 (2.7% spam)
#YouNeedToRealize	2011-10-04 05:31 ~ 2011-10-05 04:55	23 hrs 24 minutes	222,888 (4.9% spam)
#YouKnowBetter	2011-10-05 03:31 ~ 2011-10-05 18:39	15 hrs 08 minutes	104,123 (7.5% spam)

Table 3: Confusion matrix

		Predicted	
		Spam	Non-spam
Actual	Spam	<i>a</i>	<i>b</i>
	Non-spam	<i>c</i>	<i>d</i>

classified non-spam messages. The accuracy means the fraction of correct classifications and is $(a + d)/(a + b + c + d)$. FP denotes $c/(c + d)$, and FN denotes $b/(a + b)$.

Total Spam Detection (TSD) measures the effectiveness of a spam detection approach. When we apply an approach to detect spam messages, we use the total spam detection, that is, how many spam messages have been detected during a topic’s lifespan (from the time when a topic becomes popular to the time when the topic is not popular anymore). In other words, we can see how many messages have been prevented by the approach.

$$TSD_{topic}(\%) = \frac{\text{detected spam messages}}{\text{total \# of spam messages in the topic}}$$

3.2 Experimental Results

For the following experiments, we selected seven popular topics that each became popular on a different day as a case study for collective attention spam detection. The topics are: #DearHair, #TheyNeedToBringBack, #YouKnowBetter, #WhatYouShouldKnowAboutMe, #MeAndYouCantDate, #IfICouldDoItOverAgain and #YouNeedToRealize as shown in Table 4. Together, these topics account for 1,338,723 messages.

Feature Selection. Before building a classifier, finding good features is very important for high accuracy. We build classifiers based on two sets of features: (1) **main features**: 6 features extracted from each message – # of urls, # of hashtags, #of @mentions, is a message retweeted?, the length of a message and the length of a payload, where, given a message, we first remove @mention, urls and hashtags and call the remaining text a payload.; and (2) **the original + bag-of-words features**: the same 6 features as well as bag-of-words features extracted from messages, where each term as a feature is represented by tf-idf value. In spam and non-spam messages analysis, we learned that the main features have power to distinguish between spam and non-spam messages. The reason why we split the features to two sets (the main features and the original + bag-of-words features) is we want to measure not only how the main features are good, but also how bag-of-words features are helpful to improve a classifier’s effectiveness. In addition, when we use

bag-of-words features, there are pros and cons. The pros would be bag-of-words features may be helpful for improving classifier’s effectiveness. But, the cons would be they will increase the number of dimensions, and computation time (building a classifier and predicting a message’s class). We adopt a decision tree based Random Forest classifier as a supervised learning method following previous success reported in [14].

Investigating Collective Attention Spam Detection.

We begin with an exploration of detecting collective attention spam by focusing on one topic – #DearHair. As illustrated in Figure 10(a), we see that 4.3% messages associated with the topic are indeed spam (again, recall the setup in Section 2.1). The first question is whether spam messages detected in the early stages may accurately identify spam that follows as a topic becomes popular. Hence, in Figure 10(b) we report the classification accuracy for training sets of varying time windows. That is, 1 hour in the *x*-axis means that the training set consists of messages posted within 1 hour after the topic became a trending topic (and hence, made available to spammers as a potential target), and a testing set consists of messages posted after 1 hour. The *y*-axis shows the accuracy when we use the training set to build a classifier and predict labels of the messages in the testing set. This experiment emulates a real deployment scenario of such a collective attention spam detector, in which partial data is available for predicting future spam. Notice that as the training set grows in size the classification result becomes better. Of course, the goal is not only to have better classification result, but also to detect more spam messages as early as possible. Hence, we show in Figure 10(c) the total spam detection percent (i.e., how many spam messages out of all spam messages in the topic a classifier detected correctly). When we built a classifier with the first two hour training set, it detected 71.3% spam and achieved 98.7% accuracy using the main features, and detected 54.8% spam and achieved 98.1% accuracy using the original + bag-of-words features.

Figure 10(d) shows the false positive rate – indicating how many real non-spam messages are classified as spam messages by the classifier. Overall, the false positive rate is low. After 2 hours, we find a 0.002 and 0.0002 false positive rate using the main features and the original + bag-of-words features, respectively.

Detection Across Additional Topics. Next, we build classifiers over the remaining six popular topics and evaluate them. Table 5 presents the best classification results of the other six popular topics in the context of the largest TSD and their average result. Each topic’s best training time varies depending on the volume of generated messages and

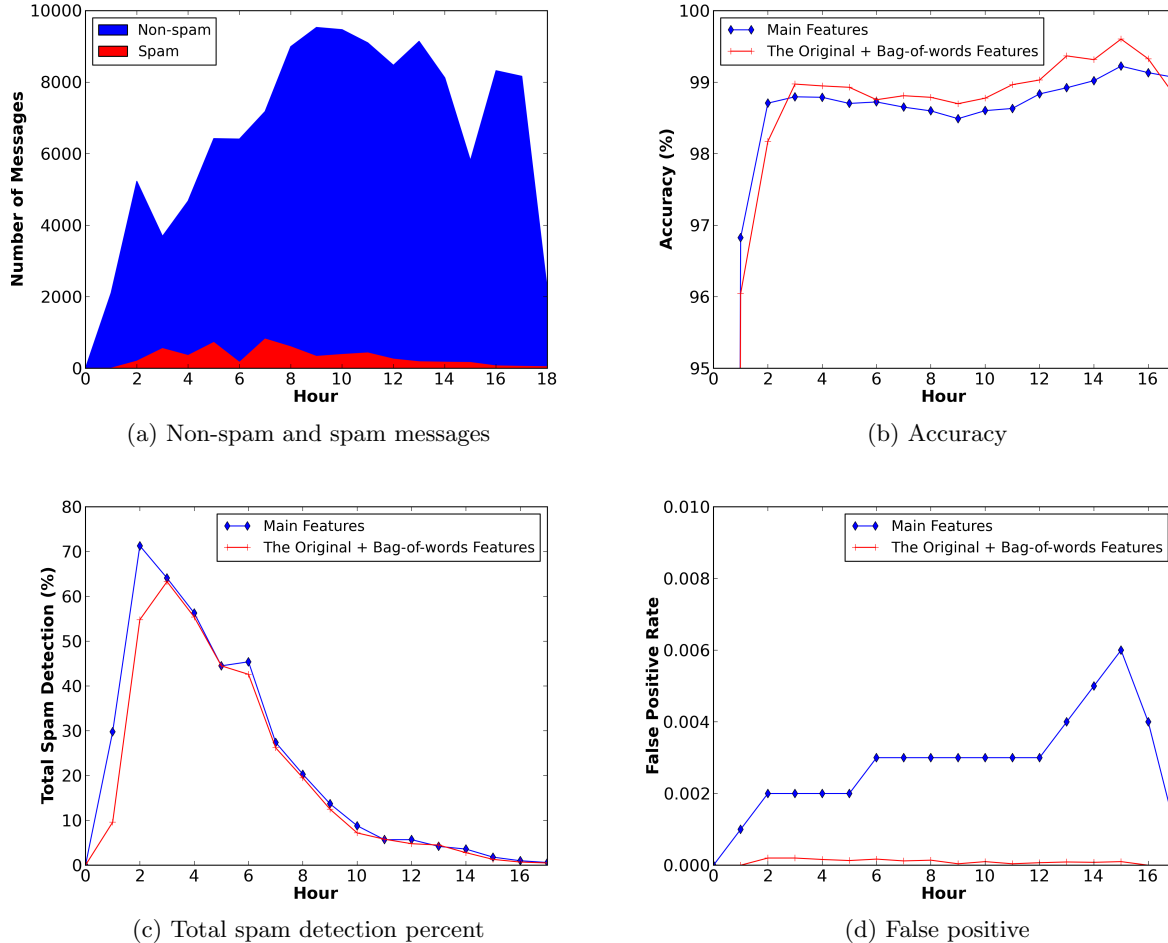


Figure 10: Classification Results for the #DearHair topic.

Table 5: The best classification results of six popular topics and their average result.

Topic	Training Time	TSD (%)	Accuracy	FP	FN
#TheyNeedToBringBack	First 4 hours	74.42%	99.47%	0.002	0.148
#WhatYouShouldKnowAboutMe	First 3 hours	60.25%	99.08%	0.002	0.347
#MeAndYouCantDate	First 5 hours	51.28%	98.79%	0.001	0.409
#IfICouldDoItOverAgain	First 3 hours	68.25%	99.02%	0.001	0.307
#YouNeedToRealize	First 1 hour	80.11%	98.92%	0.002	0.172
#YouKnowBetter	First 3 hours	71.49%	97.67%	0.002	0.272
Average	First 3 hours	68%	98.83%	0.001	0.275

the number of spam messages before the training time. In all of the topics, classifiers based on the main features outperformed classifiers based on the original + bag-of-words features. Overall, building a classifier with the first three hours’ messages gives us 68% TSD, 98.83% accuracy, 0.001 FP and 0.275 FN.

Summary. Through the above experiments, we found that it is possible to detect and prevent collective attention spam messages by learning early-age spam messages in a topic. The most encouraging results are achieving a high TSD and accuracy, and low false positive rate which means a few non-spam messages are detected as spam messages. Our

approach using the main features is lightweight, so we anticipate continued deployment for near real-time spam message detection. An open question is how to verify that the spam messages in the first few hours used to bootstrap the learning approach are indeed spam. We are investigating methods for collaboratively labeling samples of these early messages with a high likelihood of being spam (based on features learned from earlier instances of collective attention spam) using Amazon Mechanical Turk.

4. RELATED WORK

Spam, especially in social media, has received increasing

attention with the commensurate rise in the popularity of services like Facebook and Twitter. Jagatic et al. [10] studied how social phishing is effective. When a friend sends phishing messages, 72% of recipients clicked a phishing link in the messages while when an unknown person sends phishing messages, only 16% of recipients clicked a phishing link. Similarly, Brown et al. [2] showed that context-aware attacks in social systems are very effective. Irani et al. [8] studied how easily users were tempted by manipulated social services such as recommendation system, demographic search, and visitor tracking service in social media sites. Grier et al. [5] showed that many spam URLs posted in Twitter are newly created and that spammers use URL shortening services for obfuscation. Heymann et al. [6] generally summarized three main anti-spam strategies: (i) detection strategy; (ii) demotion strategy; and (iii) prevention strategy. Other researchers proposed domain-specific spam detection solutions. For example, Koutrika et al. [11] studied spam detection problem in social tagging systems. Benevenuto et al. [1] studied the online video spam problem and used supervised learning methods to detect online video spammers. In a different direction, researchers have studied the problem of review spammers [16, 17, 21]. Lee et al. [13] created social honeypots to collect social spammers’ information to understand their behaviors and tactics, and proposed machine learning method to detect spammers. Castillo et al. [3] studied information credibility, especially in newsworthy topics in Twitter and built a classifier to determine whether messages associated with a topic are credible or not. Ratkiewicz et al. [18] built a classifier to detect astroturf political campaigns in Twitter. Lee et al. [12] proposed a content-driven framework to extract various campaigns in Twitter. Most related to the work presented here is a recent study by Irani et al. [9] on “trend-stuffing” in Twitter’s trending topics. They proposed a machine learning based approach trained by text and web page content features to classify tweets associated with trending topics. In contrast, our focus is on understanding and detecting collective attention spam as it evolves.

5. CONCLUSION AND FUTURE WORK

In this paper, we have examined the problem of collective attention spam, studied the presence of it in one popular service – Twitter – and have tested an initial approach for early detection. In our experiments, we found that learning a classifier based on the first moments of a bursting phenomenon is effective to detect spam messages in the future as more attention focuses. In our continuing research, we are moving in two complementary directions: first, we are expanding our study to a larger number of trending topics to better understand the robustness of the proposed approach; and second, we are investigating adaptations of the developed methods for alternative domains (e.g., YouTube videos).

6. REFERENCES

- [1] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonçalves. Detecting spammers and content promoters in online video social networks. In *SIGIR*, 2009.
- [2] G. Brown, T. Howe, M. Ihbe, A. Prakash, and K. Borders. Social networks and context-aware spam. In *CSCW*, 2008.
- [3] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *WWW*, 2011.
- [4] M. Goetz, J. Leskovec, M. McGlohon, and C. Faloutsos. Modeling blog dynamics. In *ICWSM*, 2009.
- [5] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @spam: the underground on 140 characters or less. In *CCS*, 2010.
- [6] P. Heymann, G. Koutrika, and H. Garcia-Molina. Fighting spam on social web sites: A survey of approaches and future challenges. *IEEE Internet Computing*, 11(6):36–45, 2007.
- [7] L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in twitter. In *WWW*, 2011.
- [8] D. Irani, M. Balduzzi, D. Balzarotti, E. Kirida, and C. Pu. Reverse social engineering attacks in online social networks. In *Detection of intrusions and malware, and vulnerability assessment (DIMVA)*, 2011.
- [9] D. Irani, S. Webb, C. Pu, and K. Li. Study of trend-stuffing on twitter through text classification. In *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2010.
- [10] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer. Social phishing. *Commun. ACM*, 50(10):94–100, 2007.
- [11] G. Koutrika, F. A. Effendi, Z. Gyöngyi, P. Heymann, and H. Garcia-Molina. Combating spam in tagging systems: An evaluation. *ACM Trans. Web*, 2(4):1–34, 2008.
- [12] K. Lee, J. Caverlee, Z. Cheng, and D. Z. Sui. Content-driven detection of campaigns in social media. In *CIKM*, 2011.
- [13] K. Lee, J. Caverlee, and S. Webb. Uncovering social spammers: Social honeypots + machine learning. In *SIGIR*, 2010.
- [14] K. Lee, B. D. Eoff, and J. Caverlee. Seven months with the devils: A long-term study of content polluters on twitter. In *ICWSM*, 2011.
- [15] K. Lerman and R. Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *ICWSM*, 2010.
- [16] E. P. Lim, V. A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw. Detecting product review spammers using rating behaviors. In *CIKM*, 2010.
- [17] A. Mukherjee, B. Liu, J. Wang, N. Glance, and N. Jindal. Detecting group review spam. In *WWW*, 2011.
- [18] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer. Detecting and tracking political abuse in social media. In *ICWSM*, 2011.
- [19] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *WWW*, 2011.
- [20] F. Wu and B. A. Huberman. Novelty and collective attention. *Proceedings of the National Academy of Sciences*, 104(45):17599–17601, Nov. 2007.
- [21] G. Wu, D. Greene, B. Smyth, and P. Cunningham. Distortion as a validation criterion in the identification of suspicious reviews. In *SIGKDD Workshop on Social Media Analytics (SOMA 2010)*, 2010.