Characterizing and Identifying Shills in Social Media

Anonymized

No Institute Given

Abstract. In addition to the lively discussion by genuine social media users, there is often an effort by duplicitous users to sway the conversation in a particular direction. Traditionally, these efforts are carried out by bots and crowdturfers who are programmed or recruited in mass to spread posts about a product, candidate or cause. In this work, we propose to study a new class of duplicitous user, "shills." Shills are professionals who carry out targeted replies to users with opposing views, often at the pleasure of a campaign. Shills are real users hired by an organization in order to promote an idea or agenda by targeted combating of any opposing opinions. We investigate whether we can effectively characterize shills and further verify if it is possible to automatically detect them. This work represents the first effort in defining social media shills, characterizing them, and suggesting a data mining framework to explore this new, challenging problem.

1 Introduction

Social media acts as a powerful tool for political campaigns. Politicians and political activists have employed many different social media techniques to spread their message. Most recently, both major parties' presidential campaigns used social media to spread information as well as to lambaste political opponents. What these techniques all have in common is that they are overt; users of a social media site know who the political players are and can watch and participate in their dialogue and messaging.

On the other hand, there is another class of political discourse carried out on social media, *covert* techniques to sway opinion. There are many ways for this to manifest online. One is through the presence of bots, automated accounts who push certain hashtags or opinions to the top of trending lists [8]. Because these accounts are automated, they do not contain rich original text like normal users, a bottleneck that can be leveraged to detect them in the real world. Another covert strategy is the deployment of "crowdturfers" [5]. These are real users who are paid by campaigns to write text on their behalf. Because they are not experts on the topic they discuss, they can often be detected due to their novice nature [10].

Recently, a new covert class of user has been employed to sway opinion in social media, "shills." The general term "shill" denotes an enthusiastic accomplice [4]. This phenomenon has made its way into the online world, with modern shills acting as accomplices to their employers to encourage certain opinions in social media. Some political campaigns hire shills. Shill accounts spread information about their campaign by directly engaging with those holding opposing opinions. Instead of creating new posts, these accounts usually direct replies to their candidate's opposition. This is done to correct users who are spreading information that is contrary to the goals of their campaign.

Because shills directly engage with other users, they are not bots. Also, because they need to directly address the points of a contrary opinion, these are not crowdsourced workers such as those often employed in crowdturfing operations [5]. Shills are professional users employed by the campaign organizers, who seed these users with talking points and facts and then ask them to go and engage with users holding differing opinions on social media sites. The requirement that shills be paid professionals means that there are very few of them in the wild; however, these users produce a lot of posts.[1] Following the 1:10:89 rule for content creation on social media[2] we know that it takes very few users to affect change online. Thus, it is critical that we find these shill users in order to ensure that the discourse that is carried out on social media is organic, free of paid promotional content for any idea or candidate.

In this work we propose a framework to categorize shills in online social media. Focusing on one prominent social media site, Reddit, we develop a process that can describe the behavior of shills. We demonstrate its efficacy by showing that it can be used to differentiate shills from real users using a real-world dataset.

2 Collecting Data to Study Shills

The data we collect pertaining to shill accounts comes from Reddit, where we crawled and labeled users posting in a politically-active forum on the site. We describe the process by which we collected the data, and annotated the users as shills. Finally, we provide some analysis on our labels to show that while we may never be certain if the user is a shill, that they are exhibiting the properties that we are looking for in shill detection.

Reddit is a large social media site where users share and comment on links. On Reddit, posts are submitted to “subreddits”, which are groups of posts organized around a common theme or interest. We selected the **/r/politics** subreddit due to its generality. Because this subreddit encompasses multiple political views, we hypothesize that it is the most likely subreddit to invite shill activity: since people from all backgrounds are there they are more likely to reach audiences outside of the ones that the shill agrees with; and we are also more likely to find diverse opinion to trigger the shills into posting responses.

To collect the data, we collected posts from the beginning of April, 2016 through mid-June 2016. For all of the links submitted to the subreddit, we gathered all of the users who commented on or replied to any of them and collected the most recent 1,000 comments and replies from each user. This data was crawled by using Reddit’s REST API.[3] The full dataset will be shared upon acceptance of this publication.

After collecting the data, we assigned ground truth labels for the users. We employed three human annotators to label the user accounts. We took a random sample of 185 users to annotate. For a given user, the annotator was provided with a text file containing the most recent 1,000 replies made by that user, with each reply occupying a single line of the file. Subreddit information, timestamps, and the user’s screen name

[1] The shills in our dataset post 23% more content than regular users.

[2] `https://www.theguardian.com/technology/2006/jul/20/guardianweeklytechnologysection2`

[3] By adding “.json” to the end of any URL, you can get the JSON representation of that page.

are not provided. After reading all of the 1,000 replies by the user, the human then made the assignment based on the following criteria: (1) “Did the user’s replies entirely, or almost entirely support one candidate?”; (2) “Did the user’s posts generally contain claims to support their arguments?”; and (3) “Did the user explicitly mention a tie to any campaign?” For criterion 2, the veracity of the claims purported in the replies was not evaluated. All that was required was that the user’s reply be supported by claims. If the annotator could answer “yes” to the first two criteria, and “no” to the third, then the annotator would mark this user as a shill. We take the intersection of the three sets of users identified as shills by each annotator. The three annotators agreed on a total of 17 shills out of 185 users in the sample, or 9% of all labeled users. This number is strikingly high. We attribute this high number to the nature of the subreddit we crawled.

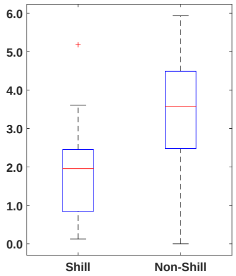

Fig. 1. Entropy of subreddit usage by shills and non-shills. Lower entropy indicates less diversity in subreddit usage. Shills are significantly less likely to participate in multiple subreddits.

While we cannot verify that these users are in fact receiving money from an organization that employs shills, we see that they behave differently than regular users. For example, they are significantly less likely to engage in other subreddits than other users, as shown in Figure 1.

3 Characterizing Shills

We introduce a preliminary approach to characterize shills based upon their topical engagement and user characteristics. We extract features to help a machine learning classifier differentiate shill accounts from real users.

Many of the features we use are extracted using Latent Dirichlet Allocation (LDA) [1]. We hypothesize that modeling topics will be useful to identifying shills as measuring the extent to which users focus on certain topics may help to reveal shills. We run LDA over the dataset, treating users as documents. First, LDA learns topics. We represent these as $\mathbf{T}$, where $\mathbf{T}_j^i$ is the probability of word i occurring in topic j. Second, it learns a document × Topic affinity matrix, represented as $\mathbf{D}_j^u$, which is the affinity of user u for topic j. We represent each user u by the following features:

1. **Content Features.** We extract features pertaining to the user’s posts.
   (a) *Topic Affinities [K features].* The $\mathbf{D}^u$ topic affinity features from LDA.
   (b) *Affinity Entropy [1 feature].* Shills are likely to focus on a particular topic. Entropy [3] is a measure of the randomness of a probability distribution, and may be useful for measuring focus.
2. **Account Features** These represent the user’s activity on the site, specifically:

(a) *Daily Post Rate [1 feature]*. The number of days in which the user has posted on the site, out of the total days they are present in the last 1,000 comments.
(b) *Subreddit Entropy [1 feature]*. We measure the frequency that each user posts in each subreddit, and then compute the entropy of the probability distribution.
(c) *Hour Probability [24 features]*. One feature for the frequency of each hour of the day, normalized.
(d) *Hour Entropy [1 feature]*. It is possible that the user may only log in for a brief period of time each day to shill. To capture this, we measure the focus of the user around any hour of the day.

3. **Community-Based Content Features** We extract a set of *Focus Topic Significance* features *[K features]*. These are the z-score [2] of the user's focus on a topic.

4 Identifying Shills

We apply the data and feature extraction approach to build a classifier that can differentiate shills. We then investigate what features the classifier determines as most important. We test various approaches in a standard machine learning framework. We use 10-fold cross validation. We report the average across the 10 runs. We compare several classifiers, organized into three groups:

- **Baselines.**
  i. **Always-Shill:** Assigns the "shill" label to every user.
  ii. **Random:** Randomly assigns a label to each instance following the distribution of labels seen in the training data.
  iii. **Clinton:** This approach labels any user that mentions "Clinton", case insensitive, in over 10% of their posts as a shill.
- **Activists.** A classifier that strives to find activists in political campaigns [11].
- **Shill Classifiers.** Classifiers trained with the features described in the "Characterizing Shills" section. We choose algorithms which produce interpretable models in order to estimate the importance of different features.

Using the classification setup described, we test several classifiers for their ability to differentiate the two classes. The results are shown in Table 1. These results can be interpreted in several ways. First we see that the best classifier is Logistic Regression. When this classifier identifies a shill, it is correct slightly under half of the time. The recall is 45%, meaning that we detect just under half of the shills present in the data. This elucidates the difficulty of the problem. Because shills are people, and they write original text, it is challenging to differentiate them from political enthusiasts.

Another result is the performance of the activist classifier proposed by [11]. While their proposed approach is very strong for identifying advocates, it does not perform well at identifying shills. This demonstrates that finding advocates is a fundamentally different problem; advocates display different behavioral patterns.

4.1 Feature Importance

In the above experiment, we evaluate several classification algorithms for their ability to discern between the two classes. In the Section 3 we provide the features we used to

Table 1. Classification Results.

Classifier	Precision	Recall	F_1
Random	0.050	0.050	0.050
Always-Shill	0.110	1.000	0.190
Clinton	0.112	1.000	0.201
Activists [11]	0.225	0.500	0.310
SVM	0.325	0.350	0.337
Decision Tree	0.335	0.550	0.417
Logistic Regression	0.475	0.450	**0.462**

make these classifications along with their justification. We hypothesized that the extent to which users focus on certain topics may help to reveal shills. In this section we will rank the features based on their predictive power to understand what is contributing to the classification.

To rank the features we perform a *feature ablation* test, where we run the classification algorithm once for each feature to test its importance. In each run, we test the feature by assessing the F_1 performance of the classifier with all features (p_1) and then re-assessing the performance with that feature removed (p_2). The difference in performance, $p_1 - p_2$, is used as that feature's predictive power.

The most important features identified by this experiment can be seen in Table 2. These results show two important observations about shill behavior. First, shills tend to focus on topics differently than regular users. We know this because the top features used to identify them are topic focus significance scores, as well as the topic affinity entropy. This means that not only do shills focus on different topics than non-shills, but they also are focused on just a handful of topics. We know this because the topic entropy score helped to separate the two classes.

Table 2. Most predictive features (affected performance by at least 3%).

Feature	Top 5 Topic Words	$p_1 - p_2$
Subreddit Entropy	*(Not Applicable)*	0.194
Topic 38 Affinity	politicaldiscussion, sanders, clinton, bernie, vote	0.056
Topic 27 Affinity	politics, trump, hillary, bernie, people	0.039

5 Related Work

The problem of detecting shills online falls closely to three problems: crowdturfing detection, bot detection, and detecting online activists.

Crowdturfing is the process of maliciously using crowdsourcing systems in order to spread fake reviews. This is close to our problem as it involves the mass proliferation of opinion by paid workers. In [12], the authors performed an analysis on different crowdsourcing platforms that harbor malicious tasks such as crowdsourcing fake reviews, and spreading disinformation about competitive brands. Lee et al. [5] proposed a machine learning framework to detect workers on crowdsourcing sites. Other researchers try to find evidence of crowdturfing reviews by analyzing the text of the social media sites [9, 10]. The key difference is that these algorithms focus on review spam and the nature of the unskilled workers in order to make these classifications.

The detection of shills is similar to the identification of central political activists [7] in that both partially rely on the content expressed by key individuals in topics pertaining to their political representatives' respective agendas [11]. Promoters are a different type of user in the sense that they do not hide their identity to be seen as regular users. [6] find hidden campaigns on social media using network information.

6 Conclusion

We present the novel problem of shills in online social media. We discussed how they can harm the authenticity of social media sites by injecting paid content that favors their cause. To back these observations, we collected a dataset of users from /r/politics, a political Reddit forum, and labeled a selection of the users as shills or as people. We find that 9% of all of the users on this forum are shills, a strikingly high number. Based on our observations, we define a strategy to characterize these users in a way that allows for automated approaches to identify them. Next, we demonstrate the efficacy of this strategy by showing that we can build a classifier to separate these instances. We found that subreddit entropy is a major factor to identify shill users. Even though shills are real people, they are constrained to posting ion a few subreddits where their message can have real impact. Areas of future work include monitoring shills to better understand the leanings of a political campaign. Also, shill detection may be more effective on social media sites that offer richer features, such as a social network.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. JMLR 3, 993–1022 (2003)
2. Casella, G., Berger, R.L.: Statistical Inference, vol. 2. Duxbury Pacific Grove, CA (2002)
3. Gray, R.M.: Entropy and information. In: Entropy and Information Theory, pp. 21–55. Springer (1990)
4. Green, J.: Crooked Talk: Five Hundred Years of the Language of Crime. Random House (2011)
5. Lee, K., Tamilarasan, P., Caverlee, J.: Crowdturfers, Campaigns, and Social Media: Tracking and Revealing Crowdsourced Manipulation of Social Media. In: ICWSM (2013)
6. Li, H., Mukherjee, A., Liu, B., Kornfield, R., Emery, S.: Detecting campaign promoters on twitter using markov random fields. In: ICDM. pp. 290–299. IEEE (2014)
7. Morales, A.J., Borondo, J., Losada, J.C., Benito, R.M.: Measuring political polarization: Twitter shows the two sides of Venezuela. Chaos: An Interdisciplinary Journal of Nonlinear Science 25(3), 33114 (2015)
8. Morstatter, F., Wu, L., Nazer, T.H., Carley, K.M., Liu, H.: A New Approach to Bot Detection: Striking the Balance Between Precision and Recall. In: ASONAM (2016)
9. Ott, M., Cardie, C., Hancock, J.: Estimating the prevalence of deception in online review communities. In: WWW. pp. 201–210. ACM (2012)
10. Ott, M., Cardie, C., Hancock, J.T.: Negative Deceptive Opinion Spam. In: HLT-NAACL. pp. 497–501 (2013)
11. Ranganath, S., Hu, X., Tang, J., Liu, H.: Understanding and identifying advocates for political campaigns on social media. In: WSDM. pp. 43–52. ACM (2016)
12. Wang, G., Wilson, C., Zhao, X., Zhu, Y., Mohanlal, M., Zheng, H., Zhao, B.Y.: Serf and turf: crowdturfing for fun and profit. In: WWW. pp. 679–688. ACM (2012)