Utilizing Natural Honeypots for Efficiently Labeling Astroturfer Profiles

Jonathan Schler[1], Elisheva Bonchek-Dokow[2], Tomer Vainstein[1], Moshe Gotam[1], and Mike Teplitsky[1]

[1] Holon Institute of Technology, Israel schler@hit.ac.il
[2] Ashkelon Academic College, Israel
elishevabd@edu.aac.ac.il

Abstract. Astroturfing is the practice of using a fake online social media (OSM) profile in order to influence public opinion, while giving the impression that the profile belongs to an authentic human user. In attempting to train a classifier for discriminating between authentic users and astroturfers, a labeled dataset must first be arranged. The labeling is generally done manually, by human judges, on a collection of profiles garnered from the social media network. However, the fact that any randomly collected set of profiles will statistically contain a small proportion of astroturfers, renders this process inefficient: a lot of time and effort is invested on manually labeling lots of data, while producing only a small set of astroturfer profiles. We present here a method for quickly and efficiently collecting a data set for manual labeling, with a high percent of astroturfers.

Keywords: Astroturfing · Facebook · Efficient Labeling.

1 Introduction

Along with the growing use of social networks, so has the realization of its dangers grown, albeit arguably at a slower pace. Adversarial use of online social media (OSM) profiles comes forth in various scenarios, such as social ("Cyber Bullying"), financial, health (disseminating anti-vax pseudo scientific claims). We focus here on the political scenario. The practice of attempting to create a false pretense of wide public support of a specific candidate in the political field, by using a fake OSM profile, is known as astroturfing. [1] defines astroturfing as the process of seeking electoral victory or legislative relief for grievances by helping political actors find and mobilize a sympathetic public, and is designed to create the image of public consensus where it does not necessarily exist. [2] analyzes the phenomenon of astroturfing as it appears in its digital form on OSM platforms. They define digital astroturfing as a form of manufactured, deceptive

and strategic top-down activity on the Internet initiated by political actors that mimics bottom-up activity by autonomous individuals. The phenomenon of fake profiles involved in political discussions on OSM has attracted the attention of several research groups. Some of these are funded by democratic governments, due to the growing awareness that such profiles pose a threat to the cornerstones of our democratic society, namely, trust and faith. A recent survey [4] on controlling astroturfing on the Internet refers to it as one of the most impactful threats online today. A preliminary step in the process of building an automatic classifier for identifying astroturfers, as in any machine learning problem, is that of creating a training set. The labeling itself is generally done manually, by human judges trained for the task. However, experience has taught us that the proportion of astroturfers found in any random collection of profiles is very low. Three judges employed for the task of labeling a random set of profiles, received detailed guidelines with criteria for labeling a profile as astroturfer. When presented with a set of 400 randomly selected profiles, their labeling resulted in only 9 (2.25%) profiles being labeled as astroturfers. Such a low figure creates an unbalanced dataset. This, in and of itself, could be tolerated and handled by the algorithms in a satisfactory way, and by starting out with a large enough dataset to begin with. However, the amount of time and effort expended for this task is unreasonable, rendering the process highly inefficient. What is needed here is a method which allows for a higher ratio of labeled astroturfers produced.

One approach for solving this problem is to harvest suspicious accounts by using what are known as "social honeypots" [3] [5]. These are fake profiles, set up by the research team with characteristics that are expected to lure the type of profiles which are being targeted. The profiles which political astroturfers are attracted to are naturally those that belong to political candidates. Obviously, setting up a honeypot in the form of a fake political candidate would not be a viable tactic. We introduce here a simple and straightforward technique, arising from insight gained over several months of collecting and analyzing profiles, posts and comments, in the Israeli political Facebook scene.

Our dataset consists of close to half a million Facebook profiles collected over a span of 15 months, from close to four million comments on some twenty political candidates' posts. During those 15 months, the political system in Israel went through one upheaval after another, with three rounds of elections attempting to reach a decree that would enable the formation of a government. This rare situation created for us a rich dataset, with many novel attributes useful for identifying astroturfers.

2 Utilizing the Innate Honeypot Nature of Political Profiles

As mentioned, creating artificial honeypots in order to attract political astroturfers would not work. However, we came to realize that the existing profiles of the political candidates are in fact natural honeypots, in that they attract, by their very nature, the astroturfer profiles. This in itself is not enough, since au-

thentic users are also attracted to these honeypots, and as mentioned, the ratio of labeled astroturfers to authentic users who interact with political profiles, is very low. This is where one of the many features we studied presented itself as useful for the task at hand. We noticed that astroturfers are quick to pounce on fresh posts, rendering the pool of the first several commentors rich with astroturfers. The method we suggest consists of collecting profiles which are among the first several commentors on the posts, and presenting these profiles to the human judges for labeling. We posited that the percent of astroturfers labeled would be significantly higher than for a random collection of profiles. We present here data to support this thesis.

We described above the motivation for this study: having only 2.25% of the training set labeled as astroturfers, which is not much to show for the huge amount of effort which went into the task. Our first attempt was based on the realization that the proportion of comments astroturfers produce is much higher than their own proportion in the population. Therefore, instead of sampling the profiles, we sampled the comments, and chose those profiles which produced the sampled comments. This approach indeed brought forth a higher proportion of astroturfers labeled—46 of 400 profiles (11.5%). This is better, but not enough.

Our next attempt was to apply a preliminary manual sifting of profiles, creating two sets: one of suspected profiles and one of innocent looking ones. These two sets were then presented to the judges for labeling, using the same guidelines. The preliminary sifting proved useful—of 364 profiles, 76 (20.88%) were labeled as astroturfers. This is much better, however it required extra effort expended for the preliminary sifting.

However, the significant growth in the proportion of labeled astroturfers was achieved by what we present here as our method: instead of choosing from all comments, we chose only from the top 10 comments. Apparently, the tendency to comment as quickly as possible is driven by astroturfers' high motivation to disseminate and promote their agenda immediately, whenever the opportunity presents itself, in the form of a new post. This insight reveals the innate honeypot nature of political profiles—there is no need to create fake profiles in order to lure the astroturfers. These profiles already exist, they just need to be utilized in the right way. We tested this hypothesis by targeting those profiles responsible for the first 10 comments on 50 randomly chosen posts. The profiles were targeted by randomly sampling with returns from the pool of 500 comments (top 10 from 50 posts). It is worth noting that these 500 comments belonged to some 300 commentors—a fact which can be attributed to the nature of such quick-to-comment users, who also have a tendency to comment more than once. Labeling this pool of profiles uncovered 25% of them as astroturfers.

Figure 1 summarizes and compares all four methods—Random Baseline (where profiles were selected randomly from among all profiles), Chosen Posts (where profiles were selected by choosing randomly from comments rather than from profiles), Past Selection (after preliminary sifting, as done in our past research) and Top 10 (which is the proposed method of choosing from among the first 10 comments). Clearly the Top 10 has the highest ratio of labeled astroturfers. Not

far behind is the Past Selection, however recall that the manual labour invested in this result was far greater, so that the ratio of profiles produced to effort expended is even more pronounced than what the figure shows.

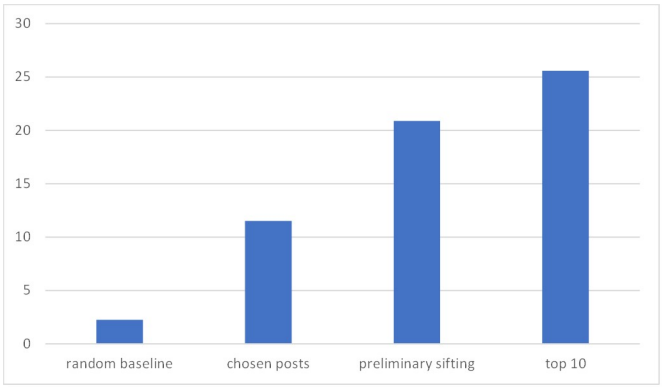


Fig. 1. Percent of Labeled Astroturfers by Method

3 Discussion and Future Work

These are only preliminary results. The number of first comments we targeted in this study was 10, however this is no magic number. Further studies should compare results for different values of K, using the first K comments. It is important to stress again that the human annotators received training with detailed criteria and guidelines *before* they began labeling the data. These guidelines did not change, throughout the various profile sets presented for labeling. In addition, it should be kept in mind that this method is pertinent only to the first stage of preparing training data for the classification algorithms. The correct choice of algorithms and parameters can then be found, applied and analyzed, in order to create a high performance classifier. A large set of labeled data is a critical resource, and a method for achieving such a set without wasting precious time and manual effort is valuable. We intend to make this valuable repository of validated astroturfer profiles available for the use of the research community.

References

1. Howard, P.N.: New Media Campaigns and the Managed Citizen. Cambridge University Press, New York, NY (2005)
2. Kovic, M., Rauchfleisch, A., Sele, M., Caspar, C.: Digital astroturfing in politics: Definition, typology, and countermeasures. Studies in Communication Sciences (1) (2018)

3. Kyumin, L., Eoff, B.D., Caverlee, J.: Seven months with the devils: A long-term study of content polluters on twitter (2011)
4. Mahbub, S., Pardede, E., Kayes, A., Rahayu, W.: Controlling astroturfing on the internet: a survey on detection techniques and research challenges. International Journal of Web and Grid Services **15**(2), 139–158 (2019)
5. Stringhini, G., Kruegel, C., Vigna, G.: Detecting spammers on social networks. In: Proceedings of the 26th annual computer security applications conference. pp. 1–9 (2010)

Utilizing Natural Honeypots for Efficiently Labeling Astroturfer Profiles

Jonathan Schler[1], Elisheva Bonchek Dokow[2], Tomer Vainstein[1], Moshe Gotam[1], and Mike Teplitsky[1]

[1] Holon Institute of Technology, Israel, schler@hit.ac.il [2] Ashkelon Academic College, Israel, elishevabd@edu.aac.ac.il

Motivation

Astroturfers are fake online social media profiles, posing as authentic, with the aim of deceitfully influencing public opinion.

In attempting to train a classifier for discriminating between authentic users and astroturfers, a labeled dataset must first be arranged. The labeling is generally done manually, by human judges, on a collection of profiles garnered from the social media network.

However, the fact that any randomly collected set of profiles will statistically contain only a small proportion of astroturfers, renders this process inefficient: much time and effort is invested in manually labeling lots of data, while producing only a small set of astroturfer profiles.

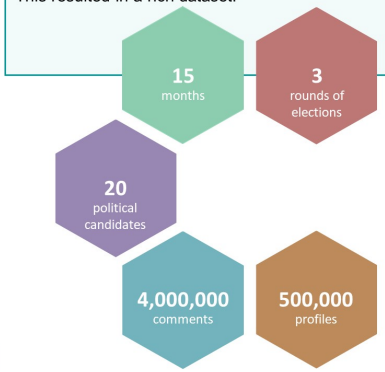
Objective

To quickly and efficiently generate a data set for manual labeling, with a high percentage of astroturfers.

The Dataset

Unique political circumstances in Israel, of three contiguous elections held within a time span of 15 months, provided an opportunity for collecting a vast amount of data from Facebook: comments to posts of political candidates were harvested, and profiles who generated those comments were reeled in.

This resulted in a rich dataset:



Method

Noting that astroturfers are quick to pounce on new posts, we propose to exploit this fact and collect profiles for labeling only from the top K comments on posts, where K is to be determined experimentally.

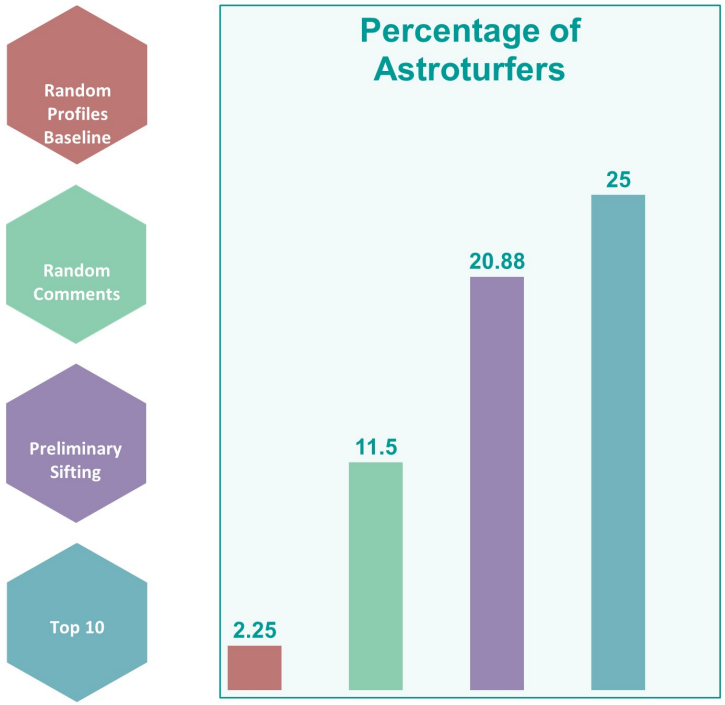
Experiments

We tried four methods of collecting profiles for labeling. For each method, we chose 400 profiles from the dataset, and presented them to our human judges. The judges then labeled them, following preset guidelines, which were the same for all methods.

- **Random Profiles-** sampling the set of profiles randomly. This serves as our baseline.
- **Random Comments-** noting that the percentage of astroturfer **comments** is higher than the percentage of the astroturfers themselves, we sampled the set of comments, and then took the profiles which generated these comments.
- **Preliminary Sifting-** manually sorting the profiles into suspicious and non-suspicious, and then sampling from the suspicious subset. Note that this requires extra effort.
- **Top 10-** sampling only from those profiles which were one of the first ten profiles to comment on a post. This is our proposed method.

Results

The Top 10 method showed significant increase in the percentage of labeled astroturfers, as compared to the other methods. The posts of the political candidates prove to act as naturally occurring honeypots, attracting astroturfers as first commenters.



Future Work

The best value for K remains to be determined, by experimentation.

The rich and unique dataset is intended to be made public, for the benefit of the research community.