InterSpot: Interactive Spammer Detection in Social Media

Kaize Ding, Jundong Li, Shivam Dhar, Shreyash Devan and Huan Liu
Arizona State University, Tempe, AZ 85281, USA
{kaize.ding, jundong.li, sdhar3, srdevan, huan.liu}@asu.edu

Abstract

Spammer detection in social media has recently received increasing attention due to the rocketing growth of user-generated data. Despite the empirical success of existing systems, spammers may continuously evolve to evade the current detection system, leading to the fact that a built system will gradually lose its efficacy in spotting spammers. To address this issue, grounded on the contextual bandit model, we present a novel system for conducting interactive spammer detection. We demonstrate our system by showcasing the interactive learning process, which allows the detection model to keep optimizing its detection strategy through incorporating the feedback information from human experts.

1 Introduction

Social media services (e.g., Facebook, Youtube) have emerged as popular platforms for content sharing and information dissemination. The rapid growth of social media also provides malicious users a new and convenient medium to spread spamming contents for their noxious intentions. Those malicious users, also known as social spammers [Lee et al., 2010; Webb et al., 2008; Hu et al., 2014], are able to perform various attacks such as spreading fake news [Shu et al., 2017], disseminating phishing links [Hu et al., 2014], and promoting or even sabotaging the reputation of targeted products [Mukherjee et al., 2012]. The massive spamming contents generated by social spammers may have an adverse effect on the user experience on these social media platforms. Therefore, detecting social spammers is a vital research problem that has significant implications on keeping social media users from unwanted information that is generated by malicious attacks.

To counter these severe threats, extensive research efforts have been devoted to detecting social spammers with disruptive behaviors. Generally, a vast majority of existing methods can be classified into two categories. One family of methods is based on supervised learning techniques [Lee et al., 2010; Benevenuto et al., 2010; Hu et al., 2013]. For instance, [Benevenuto et al., 2010] proposed to adopt both content and behaviors of each user as its attributes and apply SVM to train a social spammer classifier. [Lee et al., 2010] proposed to deploy honeypots in social networks for collecting training data, and learn a spam detector using extracted feature vectors. The training of spammer detection classifiers heavily relies on the assumption that a sufficient amount of labeled samples are available. Nevertheless, due to the prohibitive cost for accessing the ground truth social spammers, it is unrealistic to collect a large amount of annotated data. As alternative solutions, unsupervised methods [Uemura et al., 2008; Bouguessa, 2011; Tan et al., 2013; Ding et al., 2019a] have received a surge of research interests and achieved extensive success by characterizing the difference between legitimate users and social spammers. However, as spammers can quickly evolve and new types of spammers may also arise, a trained model will lose its power owing to the inability of capturing such environment changes. Fortunately, advanced research in human-in-the-loop machine learning [Holzinger, 2016; Huang et al., 2018; Ding et al., 2019b] show that by interactively incorporating expert knowledge in the learning process, the model is able to sense the environment changes and the performance can be remarkably improved. As such, there is an urgent need for developing a system that supports us to spot spammers in social media in an interactive fashion.

Contribution. In this study, on the basis of the contextual multi-armed bandit algorithm, we present a novel system: InterSpot1, which facilitates the detection of social spammer in an interactive manner. By continuously incorporating the expert knowledge about social spammers into the learning process, our system is able to constantly optimize the detection strategy for tracing the environment changes and thus achieve superior detection performance in practical usage.

2 System Overview

In this section, we carefully illustrate the overview of our proposed system on three aspects: (1) studied dataset; (2) the proposed algorithm; and (3) system interface.

2.1 Studied Dataset

We showcase our system on a real-world spammer detection dataset: YelpCHI. This dataset is collected from Yelp.com and has significant implications on keeping social media users from unwanted information that is generated by malicious attacks.

A demo video can be found at https://youtu.be/oW1pOD6zc1g
has been widely used in previous research [Mukherjee et al., 2013; Rayana and Akoglu, 2015]. The dataset includes reviews by 38,063 reviewers on 201 different hotels and restaurants. According to the results from Yelp anti-fraud filter, we are able to divide the reviewers into two classes: authors of fake reviews (social spammers), and authors of real reviews (legitimate users). We create a reviewer-reviewer network following the way of [Kaghazgaran et al., 2018]. Additionally, we apply the bag-of-words model on the whole reviews to extract the feature vector of each user for model learning.

2.2 Algorithm Description

Our system possesses a contextual multi-armed bandit (CMAB) backbone which attempts to address the problem of spammer detection in an interactive manner. In many real-world applications (e.g., recommender systems [Li et al., 2010a; Bouneffouf et al., 2012] and display advertising [Li et al., 2010b; Chapelle and Li, 2011]), we often need to tackle the so-called exploration-exploitation dilemma where it is important to make a trade-off between exploiting the current accumulated knowledge and exploring new knowledge by trying out the unknown space. In our scenario, we also need to address the dilemma between exploiting existing known types of spammers and exploring new types of spammers, to achieve superior detection performance. Therefore, contextual multi-armed bandit algorithm [Chu et al., 2011; Li et al., 2010a; Lu et al., 2010] is a principled tool that we can resort to for conducting interactive learning.

To formulate our social spammer detection problem within the $K$-armed contextual bandit framework, we first partition the $N$ users into $K$ different clusters. The reason is that the users in one cluster can be considered as samples drawn from the distribution behind a bandit arm. Thus for each user, we can regard the cluster it belongs to as an arm to pull, and when we pull that particular arm, we consider its features as the contextual feature vector. With the contextual feature vectors of all the users $\{x_i\}_{i=1}^N$ at each trial $t \in \{1, \ldots, T\}$, our system selects one suspicious user $i_t$ and queries the human expert if it is considered as social spammer or not. In order to model both user features and network structure information, follow the framework of LinUCB [Li et al., 2010a], the expected payoff of selecting user $i$ can be defined as:

$$r_i = x_i^T \theta_{a(i)} + \alpha y_i^T \phi_{a(i)}$$

s.t. $y_i = \text{AGGR}(\{x_j, \forall j \in N(i)\})$, (1)

where $\alpha$ is a controlling parameter to balance the impact between two information modalities and $a(i)$ represents the arm that user $i$ belongs to. $\theta$ and $\phi$ are the coefficient vectors for modeling user features and network structure, respectively. $\text{AGGR}()$ is a predefined aggregator function which aggregates the features from neighbors, and one prevalent choice is to use the mean operator [Xu et al., 2018]. Once the human expert provides his feedback information, our system will incorporate the feedback and can update its selection strategy according to the observed reward $r_i \in \{0, 1\}$. We repeat the whole process until we run out of the $T$ queries budget.

2.3 System Demonstration

As visually depicted in Figure 1, once the input dataset is uploaded, the system will enter into the interactive detection process. In each round, the system will present one candidate spammer along with its auxiliary information to the human expert. To facilitate the human expert to assess the abnormality of each candidate, in our demo system, we provide four classes of auxiliary information. Part (a) shows some statistics (e.g., top words) about the reviews of the spammer candidate (up) and its neighbors (down). Part (b) shows the original reviews of the spammer candidate. Additionally, the ego-network of the spammer candidate is shown in Part (c). When the human expert clicks one of the neighboring nodes, the original reviews of this neighbor will be displayed in Part (d). Note that other auxiliary information can also be exploited for further extension. After the human expert clicks the button according to his domain knowledge, the feedback information will be integrated back into the spammer detection model to update its selection strategy at the next iteration. This interactive process will iterate until the human expert stops the algorithm or the budget is used up.
References


