Information Sciences 467 (2018) 312–322

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Deep neural networks for bot detection

Sneha Kudugunta[a], Emilio Ferrara[b,*]

[a] Indian Institute of Technology, Hyderabad, Hyderabad, India
[b] USC Information Sciences Institute, Marina Del Rey, CA, USA

ARTICLE INFO

Article history:
Received 27 February 2018
Revised 4 August 2018
Accepted 8 August 2018
Available online 9 August 2018

Keywords:
Social media networks
Web and social media
Social bots
Deep learning
Deep neural networks

ABSTRACT

The problem of detecting bots, automated social media accounts governed by software but disguising as human users, has strong implications. For example, bots have been used to sway political elections by distorting online discourse, to manipulate the stock market, or to push anti-vaccine conspiracy theories that may have caused health epidemics. Most techniques proposed to date detect bots at the account level, by processing large amounts of social media posts, and leveraging information from network structure, temporal dynamics, sentiment analysis, etc. In this paper, we propose a deep neural network based on contextual long short-term memory (LSTM) architecture that exploits both content and metadata to detect bots at the tweet level: contextual features are extracted from user metadata and fed as auxiliary input to LSTM deep nets processing the tweet text. Another contribution that we make is proposing a technique based on synthetic minority oversampling to generate a large labeled dataset, suitable for deep nets training, from a minimal amount of labeled data (roughly 3000 examples of sophisticated Twitter bots). We demonstrate that, from just one single tweet, our architecture can achieve high classification accuracy (AUC $>$ 96%) in separating bots from humans. We apply the same architecture to account-level bot detection, achieving nearly perfect classification accuracy (AUC $>$ 99%). Our system outperforms previous state of the art while leveraging a small and interpretable set of features, yet requiring minimal training data.

1. Introduction

During the past decade, social media like Twitter and Facebook emerged as a widespread tool for massive-scale and real-time communication. These platforms have been promptly praised by some researchers for their power to democratize discussions [31], for example by allowing citizens of countries with oppressing regimes to openly discuss social and political issues. However, due to many recent reports of social media manipulation, including political propaganda, extremism, disinformation, etc., concerns about their abuse are mounting [19].

One example of social media manipulation is the use of bots (a.k.a. social bots, or *sybils*), user accounts controlled by software algorithms rather than human users. Bots have been extensively used for disingenuous purposes, ranging from swaying political opinions to perpetuating scams. Existing social media bots vary in sophistication. Some bots are very simple and merely retweet specific posts (based on some rules), whereas others are sophisticated and have the capability to even interact with human users.

* Corresponding author.
E-mail addresses: cs14btech11020@iith.ac.in (S. Kudugunta), emiliofe@usc.edu (E. Ferrara).

https://doi.org/10.1016/j.ins.2018.08.019
0020-0255/

The challenge of bot detection has been taken seriously by our research community [43]. Different approaches have been proposed to detect social media bots [17]. Supervised learning, in particular, exhibited promising results [13,38,45]: examples of activity of human users and bots, labeled as such, can be fed to machine learning algorithms; trained models are then used to classify unforeseen accounts, leveraging available data, e.g., obtained via the Twitter API. Alternatives based on unsupervised learning aimed at identify large-scale behavioral anomalies and associate them to bot accounts.

However, most, if not all, of the successful methods introduced so far detect bots at the account level. This means that, given a record of activity (e.g., a few hundred tweets posted by a user), the algorithm would determine whether the scrutinized account is likely a bot or not. This type of approaches tend to focus on the overall account's activity, e.g., the content and sentiment of user posts, the network structure, and the temporal usage patterns.

Though quite successful, account-level bot detection approaches are expensive as they require significant amounts of data for each user to be scrutinized, as well as large labeled datasets for training purposes. In contrast, most available labeled datasets have at most a few hundreds examples of tweets posted by a few thousands bots. For a comprehensive survey of bot detection methods, we direct the reader to our recent review [17].

1.1. Research questions & contributions

These fundamental limitations pose two research questions, that we try to address in this paper:

RQ1: Is it possible to accurately predict whether a given tweet has been posted by a bot or human account?
RQ2: Is it possible to enhance existing labeled datasets to produce more examples of bot and human accounts without the additional (and very expensive) data collection and annotation steps?

The contributions we provide here aim to address these challenges:

1. We advance the problem of classifying individual social media accounts from single observations, i.e., determining whether a single tweet comes from a Twitter bot or from a human user. We demonstrate that tweet-level bot detection is possible and can be very accurate: by exploiting both textual features and tweet metadata, we detect bots from single tweets and even exceed the performance of earlier works that make use of a given user's entire profile and recent posting history.
2. As a technical contribution, we introduce the concept of a *Contextual LSTM* (Long Short-Term Memory) deep neural network [23], an architecture that takes both the tweet text and the tweet metadata as an input. Related architectures that use side-information to enhance recurrent model representations have been alluded to by some authors before, primarily in the context of language models, but has never been used in the context of social media classification—to the best of our knowledge. The proposed architecture allows us to reach state-of-the-art performance in bot detection (i.e., over 96% AUC scores).
3. Finally, we introduce a technique based on the usage of *synthetic minority oversampling* [9] to enhance existing datasets by generating additional labeled examples. This will allow us to achieve near perfect classification performance on the account-level bot detection task, by leveraging only a minimal number of features and very small training datasets.

1.2. Impact of this work

A successful tweet-level bot detection approach would potentially overcome the limitations presented above, namely the need for computationally expensive models that require large numbers of features, large labeled datasets for training purposes, and access to the recent history of activity of the user profile to scrutinize.

Given the same pool of users, a tweet-based bot detection approach would have significantly more labeled examples to exploit. For instance, in the dataset we use in this paper (discussed in the next section), we have labels for 3474 human users, who overall generated 8,377,522 tweets; we also have labels for 4912 social bot accounts, who generated 3,457,344 tweets.

A tweet-level detection approach would be capable of leveraging nearly 12 million labeled datapoints, while an account-level detection system would only be able to exploit about eight thousands examples of bots and human accounts, while using those millions of tweets to learn patterns associated with the originating accounts.

Shifting to tweet-level bot detection, and thus having training data orders of magnitude larger than otherwise, makes the problem of bot detection far more amenable to the usage of deep learning models. Such techniques benefit greatly from vast amounts of labeled data, showing extremely high performance in many contexts where such resources are available [29], from image classification [28] to mastering games [35,40].

Traditional deep learning techniques used for text classification purposes (as well as in the broader context of language models) rely solely on textual features (e.g., characters or n-grams) [25]. A straightforward implementation of such techniques to tweet-level bot detection could be based exclusively on tweet texts as inputs for the deep neural network of choice. However, prior results in bot detection suggested that tweet text alone is not highly predictive of bot accounts [17]. Exploiting additional features such as account metadata, network structure information, or temporal activity patterns, have been found to yield more robust and accurate results [17].