Twitter Bot Detection

Twitter bot detection aims to automatically detect automated accounts that are operated to achieve malicious purposes, such as spreading misinformation, promoting hate speech, and manipulating public opinion.

Existing models could be categorized as:

- **Feature-based**: where feature engineering is conducted with user metadata and tweets while combined with simple classifiers.
- **Text-based**: where word embeddings and language models are adopted to identify bots based on user tweets and descriptions.
- **Graph-based**: where network and graph mining models are adopted to analyze the structure of Twitter to identify bots.

The TwiBot-22 Benchmark

To address these issues, we propose TwiBot-22, which:

- establishes the largest benchmark to date with ~1M users.
- provides diversified entities and relations on the Twitter network.
- has considerably improved annotation quality through weak supervision.

Table 1: Statistics of the 9 datasets. TwiBot-20 contains unlabelled users so that User ≠ Human ≠ Bot. C-15, G-17, C-17, M-18, C-18, C-19, B-19, F-19 are short for cres-2015, glam-2017, cres-2017, midnight-18, cres-stock-2018, cresriobou-2019, bomsonere-feedback-2019. C-17 contains only “post” edges between users and tweets, which is not a graph-based dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>User</th>
<th>Human</th>
<th>Bot</th>
<th>C-15</th>
<th>G-17</th>
<th>C-17</th>
<th>M-18</th>
<th>C-18</th>
<th>C-19</th>
<th>B-19</th>
<th>F-19</th>
<th>TwiBot-20</th>
<th>TwiBot-22</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-15</td>
<td>1,999</td>
<td>1,968</td>
<td>31</td>
<td>3,672</td>
<td>5,957</td>
<td>4,774</td>
<td>548</td>
<td>548</td>
<td>548</td>
<td>550</td>
<td>548</td>
<td>1,065,077</td>
<td>1,094,921</td>
</tr>
<tr>
<td>F-15</td>
<td>3,581</td>
<td>1,090</td>
<td>184</td>
<td>10,894</td>
<td>43,484</td>
<td>7,192</td>
<td>353</td>
<td>136</td>
<td>6,249</td>
<td>17,942</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C-17</td>
<td>5,581</td>
<td>2,484</td>
<td>624</td>
<td>10,358</td>
<td>13,379</td>
<td>660</td>
<td>156</td>
<td>209,580</td>
<td>1,060,490</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># Tweet</td>
<td>2,631,737</td>
<td>0</td>
<td>0</td>
<td>8,677,625</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>88,217,437</td>
<td></td>
<td></td>
</tr>
<tr>
<td># Mention</td>
<td>2,631,737</td>
<td>0</td>
<td>0</td>
<td>2,859,584</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>88,217,437</td>
<td></td>
<td></td>
</tr>
<tr>
<td># Edge</td>
<td>7,396,134</td>
<td>0</td>
<td>0</td>
<td>8,677,625</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>170,082,937</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

However, graph-based models are poorly supported by existing datasets!

- Only 2 out of the 18 datasets provide the Twitter network structure.
- Existing datasets suffer from limited dataset scale, incomplete graph structure, and low annotation quality.

Benchmarks Twitter Bot Detection

Armed with TwiBot-22, we provide a comprehensive benchmarking of Twitter bot detection to enable a rethinking of research progress.

- **35 bot detection models**, covering both the classic and the advanced.
- **9 representative and publicly available bot detection datasets**.

Table 2: Average bot detection accuracy and standard deviation of five runs of 35 baseline methods on 9 datasets. Bold and underline indicate the highest and second highest performance. The F, T, and G in the “Type” column indicate whether a baseline is feature-based, text-based, or graph-based. Cresci et al. and Botometer are deterministic methods or APIs without standard deviation. “*” indicates that the dataset does not contain enough user information to support the baseline. “**” indicates that the baseline is not scalable to the largest TwiBot-22 dataset.

Limited Dataset Scale

Existing datasets contain at most 10k users, while online conversations and discussions about heated topics often involve hundreds of thousands of users.

Incomplete Graph Structure

Only users and follow relationships are provided, while Twitter is a heterogeneous information network with many types of entities and relations.

Low Annotation Quality

Crowdsourcing is often adopted for data annotation, while whether crowd workers could be trusted to identify advanced and evasive bots remain debated.

Graph-based Twitter bot detection approaches are the most advanced, achieves state-of-the-art performance, and helps to tackle the many challenges in bot detection such as bot evolution and generalization.

Results and Discussion

- **Graph-based approaches consistently outperform other models.**

Implication: future research on the network structure of Twitter

- **TwiBot-22 establishes the largest benchmark while exposing the scalability issue of certain approaches.**

Implication: emphasize scalability in bot detection research

- **Model performance on TwiBot-22 is on average 2.7% lower on TwiBot-20 than TwiBot-20, which features older generation of bots.**

Implication: combating the everlasting bot evolution

Generalization Test

- Graph-based methods are better at generalizing to new communities.
- Good performance does not guarantee good generalization.

The TwiBot-22 Evaluation Framework

We consolidate all implemented codes and datasets into the TwiBot-22 evaluation framework, which provides a one-stop shop for future research in Twitter bot detection.