Bot Detection in Online Social Networks

Prof. Athena Vakali
Introduction

Bot detection in OSNs: history and evolution
Bot detection state of the art outline
Bot-detective principles and approach
Bot-detective as a service
Conclusions and future work
Introduction

• Use of social media has skyrocketed during the past 15 years.
• In 2005 only 5% of US adults reported using a social media platform. Today this number is around 70%.
• Facebook is the market leader with around 2.8 billion active users.
• Twitter though, remains one of the most popular ones with ~350 million active users.
• Twitter has radically transformed various sectors (journalism, politics, economy, etc.)
Introduction (I)

Huge Popularity

Fertile ground for “malicious” activities

The rise of Bots!

What is a bot?
• Online account that is at least partially automated
• Social media accounts that mimic humans
• Really easy to develop one or thousands of them
• Actually fake accounts that have taken over OSNs

Wait...What?

• 9-15% of the total users seem to be bots[1]
• ~30-50 Million accounts!
• 1/3 of the content shared in Twitter is bot-generated [2]
• 2/3 of the circulating URLs are posted by bots [3]

---

There are benevolent bots & malevolent bots. The problem lies in the bots’ intentions!

- Bots that post funny content (e.g. images of cats)
- Crawlers (content aggregation)
- News agencies, Companies
- Bots that call users for voluntary actions
- Celebrities
- Fake news dissemination
- Manipulate Stock Market
- Cyberbullying
- Manipulate Elections
- Fake Followers
- Terrorism
Introduction (III)

Researchers: Nearly Half Of Accounts Tweeting About Coronavirus Are Likely Bots

CYNK ... never existed!

Tay bot becomes hater/racist/ ...!
Still, why is it so important to automatically detect bots?

People have difficulties discriminating bot accounts from humans.

According to recent research[4]...

- Tech-savvy users are able to tell apart new bots from legitimate users only 24% of the times.
- Although social platforms try their best to remove bots, only 5% of the newly introduced ones are detected.

Introduction

Bot detection in OSNs: history and evolution
Bot detection state of the art outline
Bot-detective principles and approach
Bot-detective as a service
Conclusions and future work
Birth of bot detection in OSNs

Since 2014, the number of publications on the topicskyrocketed. We forecast that from 2021 there will be more than one new paper published per day on social bots, which poses a heavy burden on those trying to keep pace with the evolution of this thriving field. Efforts aimed at reviewing and organizing this growing body of work are needed in order to capitalize on previous results.

- **2010**: Basic Supervised ML approached
  - Focus on coordinated fraudulent behavior
  - Rise of deceivers

- **2014**: Unsupervised approaches
  - Evidence of social bot evolution
  - First works on social bot detection
  - Worldwide attention on deception and automation

- **2017**: Adversarial ML
  - Rather new — few works
  - Keep evolving — focused primarily on group detection

---

Figure taken by: A Decade of Social Bot Detection
By Stefano Cresci
Communications of the ACM, October 2020, Vol. 63 No. 10, Pages 72-83
Bots fall into the category of digital spam
Digital spam and human activities coexist for more than a century [5]

Researchers set traps on Twitter to “catch” bots by creating Twitter accounts (bots) whose role was solely to create nonsensical tweets. These accounts attracted many followers. The suspicious followers were indeed social bots [6].

Using Supervised Machine Learning techniques they are able to identify bots with an accuracy of 98.8%
The issue of bot evolution

99% Accuracy! Great, right??!

This model is effectively detecting simple bots. But bots evolve....

Simple bots – easy to detect. Model works fine

2010

2013

2016

now

Bots very similar to humans. Make friends, respond, comment to others. Models efficiency depends on annotated data.

More sophisticated bots. Started to create networks between them. Model effective, but not as before. New models adaptive to new characteristics

Really hard to detect. Deepfakes, stolen images, stolen names, few malicious messages – many neutral ones. Group detection approaches, unsupervised methods.
Based on many research efforts, we identify the next Bots types:

- **Spam Bots**: encapsulate every type of automated account related to continuously posting spam content.
- **Social bots**: automated accounts related to impersonators, influence bots and pay-bots (attract likes, follows, ...).
- **Self-declared bots**: refer to automated accounts that identify themselves as bots.
- **Cyborgs**: human accounts with bot behavior mostly celebrities, news agencies and organizations.
- **Political Bots**: a rather unique class, including automated accounts that have been used for political purposes.
- **Other Bots**: any type of automated accounts that do not fit in any of the previous categories.
Lecture content

Introduction
Bot detection in OSNs: history and evolution
Bot detection state of the art outline
Bot-detective principles and approach
Bot-detective as a service
Conclusions and future work
Supervised ML as a baseline

Twitter post metadata

Extract some features that define the user’s behavior

- Friends to followers ratio
- Text length
- Number of tweets per minute, hour, day
- Intra-tweet similarity

Annotated data

- Labelled users (using Mturk, crowdsourcing, honey-pots, etc)

Annotated Data and Features

Train a Machine Learning model

Random Forest have shown Good performance

Trivial….Right?

Twitter API

Twitter post metadata

JSON file
**Supervised ML KnowHow**

**Well...not actually...**

**Key assumption:** bots and humans are clearly separable and malicious accounts have individual features that make it distinguishable from legitimate ones.

**Features:** As bots adapt ..., researchers needed to discover new features that, up to that point, were unnecessary.

**Multiple fragmented approaches:** by several researchers with different set of features, improved performance, more models, but same methodology.

---

**Not quite true...**The models’ performance was really good on specific trained data, but gradually decreases while newly added bots reform and adapt accordingly...

**Example**

- The first bot versions continuously posted tweets during all day and all night, really easy to spot them by measuring the intra-tweet gap duration per day.
- New versions mimic human behavior (eg. inactive during night).
- The intra-tweet gap duration feature lost its “importance”.
- Need for new updated & adaptive models !!
supervised ML critical issues

Lack of data due to OSNs restrictions
Most OSNs have closed their APIs and do not provide data, even for research purposes

The availability of ground truth datasets
Supervised ML models efficiency relies on the training data. Not many labelled datasets available.

Credibility of available datasets
Existing ones are annotated by humans (annotation biases)

Models usually output binary labels
Difficulty on detecting human-driven behaviors

Datasets do not include new types of bots
Difficulty on adapting models to newly introduced bots

Models are usually black box models
They do not provide feedback for the prediction
Beyond supervised ML approaches: from Individuals to Groups

The availability of ground truth datasets
Unsupervised models and graph based models do not necessarily need labelled data

Credibility of available datasets
Since data doesn’t have to be labelled we overcome the issue of annotation bias

Datasets include new types of bots
Analyzing large groups of accounts, means more data. More data -> higher probability of including multiple types of bots

Non-Real time detection
Most of these approaches do not provide real-time predictions

Computational heavy
These methods rely on more complex algorithms and more data.

however...
Despite the disadvantages of supervised ML, many researchers still focus on such approaches[7].

State of the Art at the moment ... : **Botometer** which covers
+ Wide research on bots [8,9,10]
+ Online Tool
+ Multiple Bot types
  - *Explainability*
  - questionable ... accuracy

---

Lecture content

Introduction
Bot detection in OSNs: history and evolution
Bot detection state of the art outline
Bot-detective principles and approach
Bot-detective as a service
Conclusions and future work
Models are usually black box models. They do not provide feedback for the prediction.

Need for more, open bot-detection services

Which should offer explainable results [11] and should allow people to share their own opinion – declare their objections.

To that end we introduced Bot-Detective [12]

- An online social bot detection service
- Explainable results
- Crowdsourcing functionalities
- New dataset
- New model

We relied on previous research to collect a short but efficient set of features:

[11] https://www.privacy-regulation.eu/en/71.htm - should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached.

Although we experimented with various ML algorithms, we finally used Random Forest which provided the best results.

Newly created real labelled dataset of ~2M tweets about cryptocurrency (known place for scams)

All the users have been annotated with the use of Botometer and those that were deleted by Twitter were labelled as bots. Score equal to 0 means human, score equal to 5 means bots.
Model Calibration

Our model tends to push the predicted probabilities away from 0[human] and 1[bot].

Platt’s calibration methodology provided a solution to this issue[14].

Based on a state-of-the-art method called LIME[13]

**Input**
- Trained dataset instances and their scores
- labels of the features
- indexes of categorical features

**Output**
- Array with weights of features
- negative values: affects the model in predicting low bot score
- positive values: high bot scores

**Explanations**
- Manually generated sentences
- Mapping function “Features:Explanations”

---

Introduction
Bot detection in OSNs: history and evolution
Bot detection state of the art outline
Bot-detective principles and approach
Bot-detective as a service
Conclusions and future work
Bot Detective as a Web Service

Available in: bot-detective.csd.auth.gr

- The architecture of the developed service follows the client-server model.
- The user logs in with his Twitter credentials, accepting the Bot-Detective terms of service.
- The user fills in the screen name or user id of the Twitter account he/she wants to check and gets a prediction score along with a set of explanations.

Check if a Twitter account is a bot

Score: 2.9

Why bot?
- Low number of tweets the account has favorited (24). Non-bots usually favorite 6224 tweets on average.
- This account is not verified. While this does not say a lot, if it were, it could increase the certainty that they are not a bot.
- Small number of retweets indicates that a tweet is more probable to have been produced by a bot account.

Why NOT bot?
- Average number of followees (63). Bots usually follow 3658 accounts on average.
- This account’s description is average in length (141). Bots have 63.2 characters on their description, on average.
- This account’s followers to followees ratio is rather high (116.46), which is normal. Non-bots usually have more than 90.99 followers per followees, on average.
The user can see some statistics with respect to the account of interest by clicking on Details: ..and can also provide his/her own feedback regarding the prediction:

Feedback helps:
- Retrain our models
- Evaluate the performance
- Improve explainability
Bot Detective V2.0 – refined approach

Approach the Bot Detection – classification problem based on previous research and all available data

Contributions / extensions:
• Insightful dataset analysis
• New Bot types
• New Features
• New Models
• New Explainability approach

New Publication: Social Botomics [14]

Bot Detective V2.0

Most datasets are outdated

**Credibility of available datasets**
Existing ones are annotated by humans (annotation biases)

**Datasets do not include new types of bots**
Difficulty on adapting models to newly introduced bots
Bot Detective V2.0 – Introducing new Bot Types

- Merge multiple (24) annotated open bot datasets
- Most datasets referred to different bot types

Exploratory analysis of Datasets

Propose a new bot taxonomy – 6 different bot types

<table>
<thead>
<tr>
<th>Bot type</th>
<th>Description</th>
<th>Number of Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spam Bots</td>
<td>Accounts that post spam content</td>
<td>4</td>
</tr>
<tr>
<td>Social Bots</td>
<td>Bots that try to attract followers</td>
<td>4</td>
</tr>
<tr>
<td>Political Bots</td>
<td>Bots involved in politics online discussions</td>
<td>3</td>
</tr>
<tr>
<td>Cyborgs</td>
<td>Human monitored bots</td>
<td>3</td>
</tr>
<tr>
<td>Self-declared</td>
<td>Accounts that state they are bots</td>
<td>1</td>
</tr>
<tr>
<td>Other bots</td>
<td>Other types of bots</td>
<td>5</td>
</tr>
<tr>
<td>Human</td>
<td>Genuine human accounts</td>
<td>11</td>
</tr>
</tbody>
</table>

Is this dataset categorization valid?
Bot types validity check

- Train Binary Classifiers for each type of bot
- Test each classifier on other bot types
- Cross-type performance of each classifier

• In-type performance is strong for all bot types
• Cross-type performance is really low

• Highlights the different behavior of bots
• Need for the distinction of bots in separate types

Exception: the other bots category!
Reasoning: Contains instances of the rest bot types!
New Models

**Binary Bot or Human Classifier**
- Trained on all datasets (75%-25% train/test)
- ADASYN imbalance handling
- Random Forest
- Parameters tuned with GridSearch
- ACC: 0.861
- F1-Score: 0.87
- Precision: 0.895
- Recall: 0.85

**Multi Class Classifier**
- Trained on all datasets (75%-25% train/test) with 6 different labels
- Experimented with multiple different classifiers
- ADASYN imbalance handling
- Best: Ensemble of Random Forests
- ACC: 0.9
- F1-Score: 0.87
- Precision: 0.891
- Recall: 0.918

Comparable and higher performance to other SotA models
Ensemble of Binary Bot Classifiers for multi class predictions

Our model predicts the instance class with higher confidence.
Feature Engineering

Feature Types - categorization

- User Related
- Temporal Features (Activity)
- Friends Features (Retweeters)
- Content
- Sentiment
- Hashtag Network

Feature Extraction

Costly process, both in terms of time and resources!

- Utilized feature importance frameworks
- Iteratively removed less important
- Best performance with just 145 features
- Performance still high with even 45 features

- Related Research: Totally more than 1000 features (not explicitly mentioned)
- Our work: 297 features
Bot Detective 2.0

New Data → New Features → New Bot types → New Models → New Web App

Check if a Twitter account is a bot!
Bot Detective 2.0

http://bot-detectivev2.csd.auth.gr/

- New enhanced UI
- Multi Class Models
- Faster Real Time prediction
- Improved Explainability

New Explainability Functionalities

Content
- Weight = -0.81
  Text-relevant features that capture the use of semantic elements, such as number of words, emoticons, inter-tweet similarity, etc.

Network
- Weight = 0.005
  Features that are generated by the network of used hashtags (hashtag co-occurrence)

Sentiment
- Weight = -0.025
  Features that reflect the sentiment expressed in each tweet, such as percentage of sentiment-neutral tweets.

Temporal
- Weight = -0.062
  Features which are exclusively relevant to the timestamps of tweets and retweets, and the elapsed time between them in a given period

User
- Weight = -0.026
  Features that refer to the characteristics of the account, such as number of favourites, friends and followers.

These features contribute positively to identifying the user as human.
These features push the Machine Learning model to identifying the user as bot.
Bot Detective 2.0

Per Feature explanations:

Explanations

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Content</td>
<td>Network</td>
<td>Sentiment</td>
</tr>
<tr>
<td>Temporal</td>
<td>User</td>
<td></td>
</tr>
</tbody>
</table>

**name_screen_name_similarity**

Similarity index of user's name and screen name

0.30 < name_screen_name_similarity <= 0.77

**followers_count**

Number of followers

43.50 < followers_count <= 572.50

**default_profile**

Whether the user has a default profile

default_profile <= 0.50

**tweets_count**

Total number of posted tweets

143.50 < tweets_count <= 3696.50

**urls_in_description**

Number of urls in user's description

urls_in_description <= 0.50

... available soon – Chart comparison
Introduction
Bot detection in OSNs: history and evolution
Bot detection state of the art outline
Bot-detective principles and approach
Bot-detective as a service
Conclusions and future work
Ongoing and future extensions

Adversarial Machine Learning (GANs)

- Create plausible adversarial examples using GANs
- Overcome the scarcity of labelled datasets
- Improve imbalanced datasets [16]
- Use GANs to test the classifiers on adversarial bots.

- Conditional GANs
- Controllable GANs
- Synthetic Data Generation
- GANs for multi-class

Data aggregation

Figure taken by: A Decade of Social Bot Detection
By Stefano Cresci
Communications of the ACM, October 2020, Vol. 63 No. 10, Pages 72-83
Main Issues still remain:

1. **Bot Evolution**: New type of bots constantly appear. How can we adapt our models to them?
2. **Lack of labelled Datasets**: Human annotation is biased. Current datasets are outdated.

- **Adversarial Machine Learning (GANs)**
  - Create plausible adversarial examples using GANs
  - Overcome the scarcity of labelled datasets
  - Improve imbalanced datasets [16]
  - Use GANs to test the classifiers on adversarial bots.

- **Unsupervised / Semi-supervised ML (GNNs)**
  - No need for labelled datasets
  - More promising results
  - See Next Slide

- **Sequence alignment methods**
  - Current solution is considered SotA [17]
  - Unlabeled data – Not Real Time
  - Works Great if tweets have already been collected

---

Open Questions & Future Work - GNNs

Currently experimenting with GNNs, issues posed by low connectivity in available datasets.

Use the expressive power of **Graph Neural Networks (GNNs)** to capture bots:

- Create meaningful user and graph **representations** in an **automated** manner and feed them to classic ML algorithms for bot prediction. **Superior results**
- Create **end-to-end** models for bot prediction by combining multiple GNNs together and adjusting their behavior to capture bot dynamics. Better **modeling** and **expressiveness** of bot behavior

Requirements/Limitations:

- Datasets: **Graph structure** and connectivity information is required. Labels are always a plus.
- Models: Current models are not **fine-tuned** towards capturing bot dynamics
Datalab Team for BotDetective

ILIAS DIMITRIADIS
PHD CANDIDATE, DATA SCIENTIST & RESEARCH ASSISTANT

MARINOS POITIS
PHD CANDIDATE, DATA SCIENTIST, RESEARCH ASSISTANT

PAVLOS SERMPEZIS
PHD IN COMPUTER SCIENCE, ELECTRICAL & COMPUTER ENGINEER

https://datalab.csd.auth.gr/

Bot Detective Contact Person:
Ilias Dimitriadis idimitriad@csd.auth.gr

... any questions?