Data and Web Science Lab (Datalab)

School of Informatics

Faculty of Sciences

Aristotle University of Thessaloniki

Bot Detection in Online Social Networks

Prof. Athena Vakali

Lecture content

Introduction

Bot detection in OSNs : history and evolution

Bot detection state of the art outline

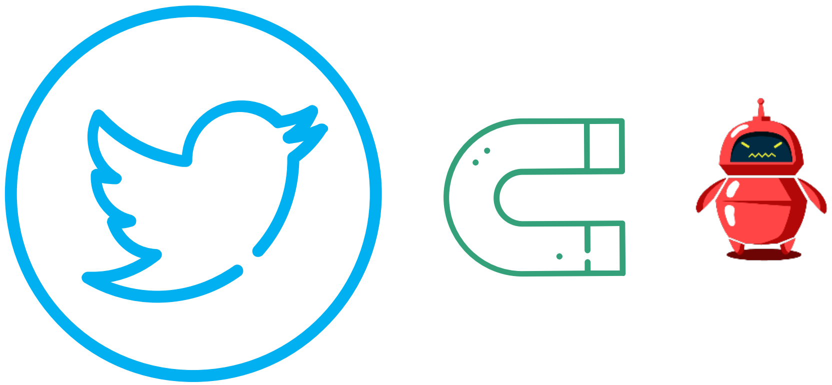
Bot-detective principles and approach

Bot-detective as a service

Conclusions and future work

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

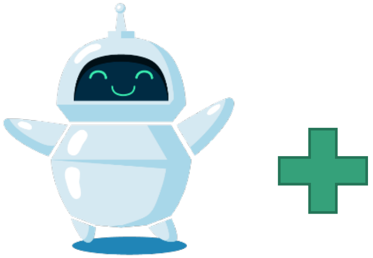
Wait…What?

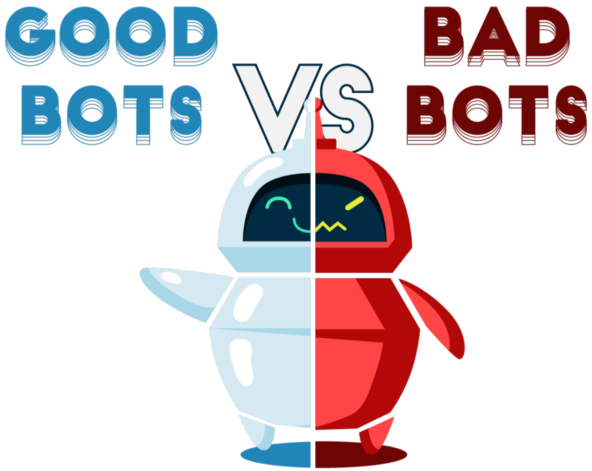
- 9-15% of the total users seem to be bots[1]
- ~30-50 Million accounts!
- 1/3 of the content shared in Twitter is bot-generated [2]
- 2/3 of the circulating URLs are posted by bots [3]

1. Varol, Onur, et al. "Online human-bot interactions: Detection, estimation, and characterization." (2017).
2. Norah Abokhodair, Daisy Yoo, and David W McDonald. Dissecting asocial botnet: Growth, content and influence in Twitter (2015)
3. Stefan Wojcik, Solomon Messing, Aaron Smith, Lee Rainie, and Paul Hitlin. Bots in the Twittersphere (2018)

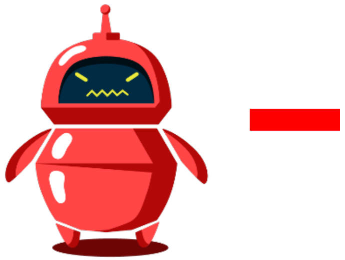
Introduction (II)

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

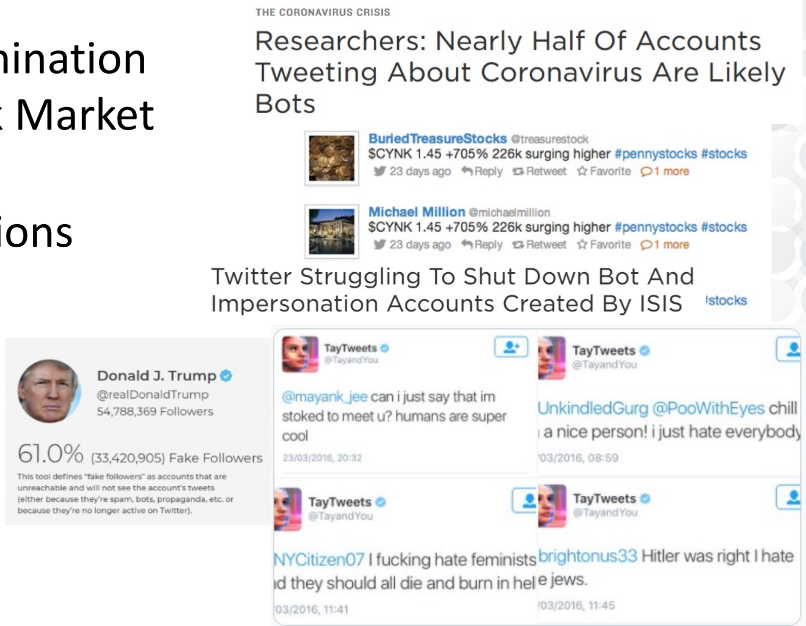




- Bots that post funny content (e.g. images of cats)
- Crawlers (content aggregation)
- News agencies, Companies
- Bots that call users for voluntary actions
- Celebrities



- Fake news dissemination
- Manipulate Stock Market
- Cyberbullying
- Manipulate Elections
- Fake Followers
- Terrorism

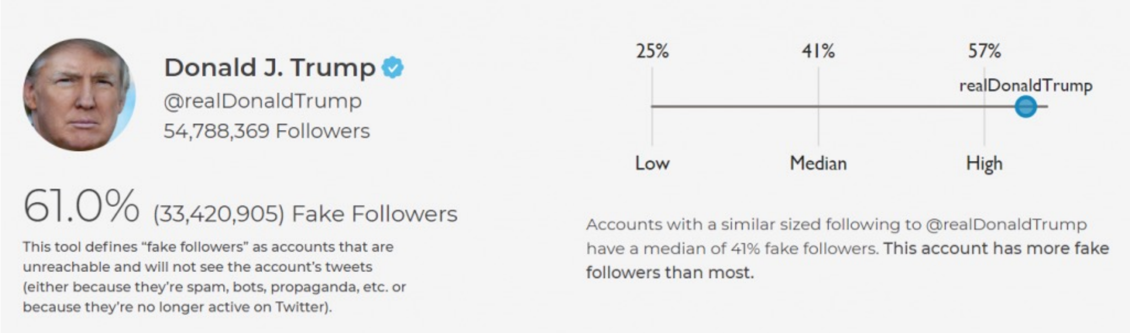


Introduction (III)

THE CORONAVIRUS CRISIS

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

CYNK … never existed !



Tay bot becomes hater/racist/ … !

Introduction (IV)

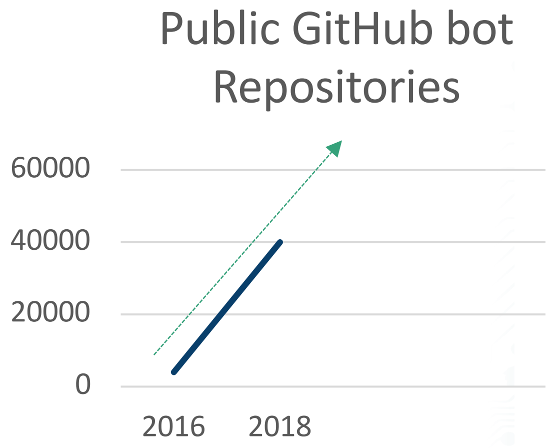
Still, why is it so important to automatically detect bots?

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.

Code about Bots … explosion…



4. Cresci, Stefano, et al. "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race." *Proceedings of the 26th international conference on world wide web companion*. 2017.

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

Birth of bot detection in OSNs

Since 2014, the number of publications on the topic sky-rocketed. 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.

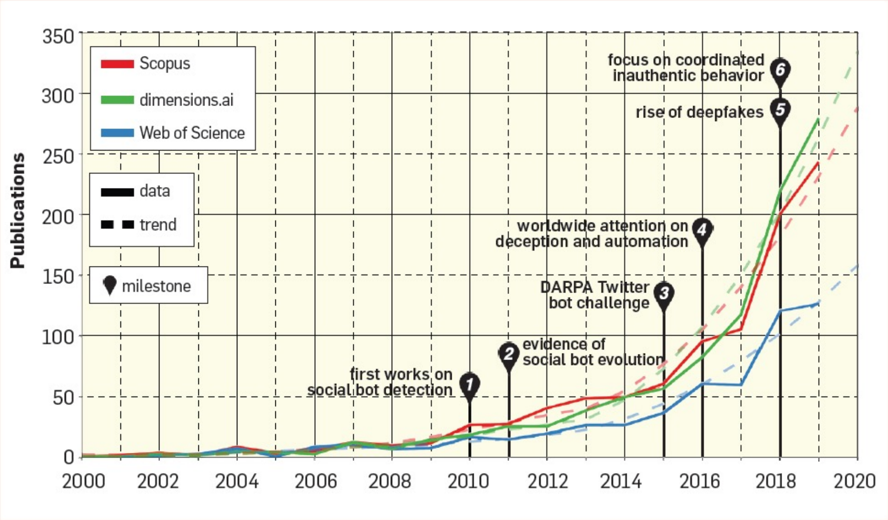
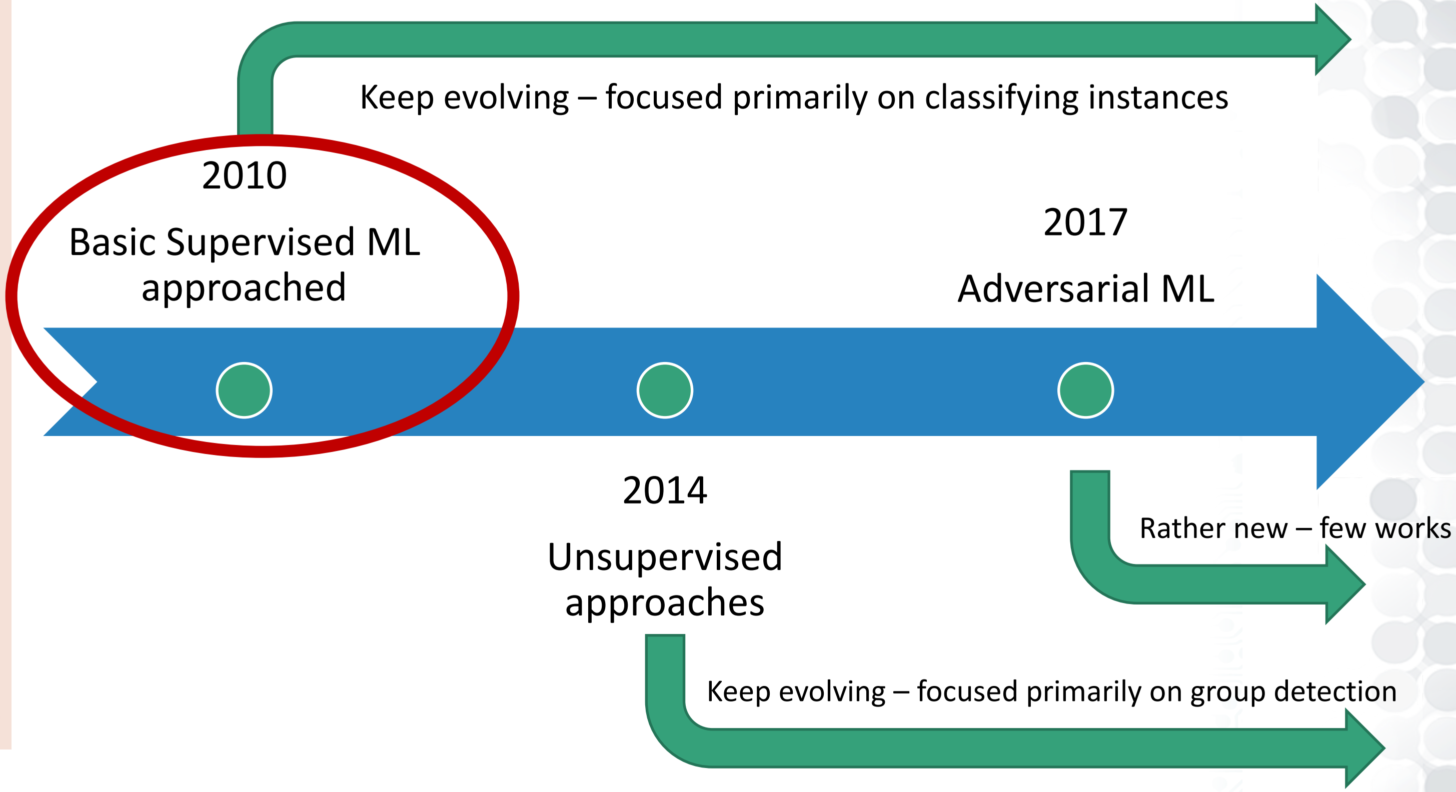


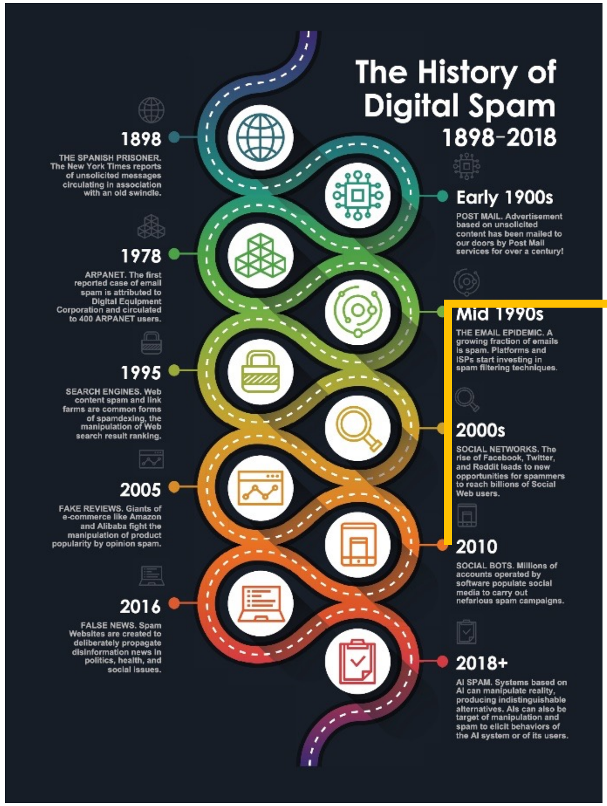
Figure taken by: **A Decade of Social Bot Detection**
By Stefano Cresci
Communications of the ACM, October 2020, Vol. 63 No. 10, Pages 72-83

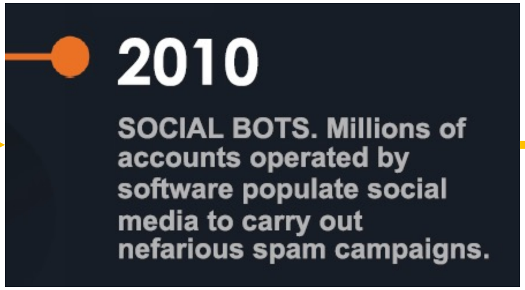


Birth of bot detection in OSNs – The first approach

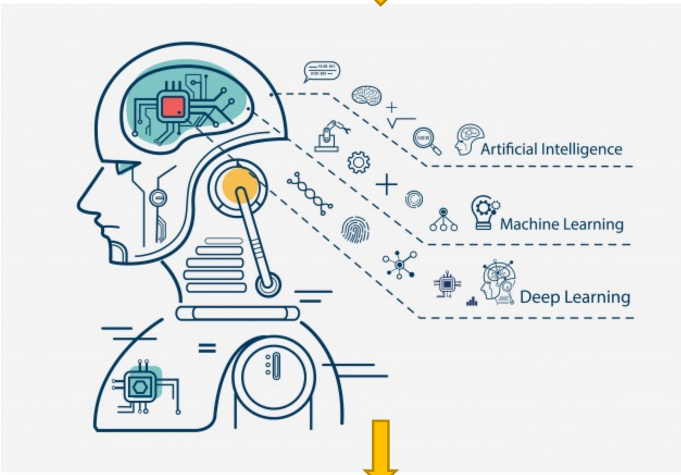
Bots fall into the category of digital spam

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].



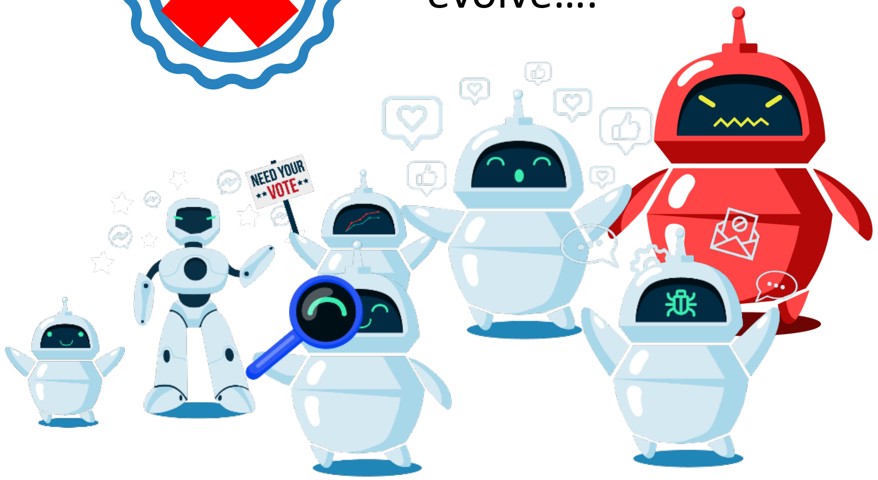
Using Supervised Machine Learning techniques they are able to identify bots with an accuracy of **98.8%**

5. Ferrara, Emilio. "The history of digital spam." *Communications of the ACM* 62.8 (2019): 82-91.
6. Lee, Kyumin, Brian Eoff, and James Caverlee. "Seven months with the devils: A long-term study of content polluters on twitter." *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 5. No. 1. 2011.

The issue of bot evolution

99% Accuracy!
Great, right??!

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



2010
Simple bots – easy to detect. Model works fine

2013
More sophisticated bots. Started to created networks between them.
Model effective, but not as before.
New models adaptive to new characteristics

2016
Bots very similar to humans. Make friends, respond, comment to others. Models efficiency depends on annotated data.

now
Really hard to detect. Deepfakes, stolen images, stolen names, few malicious messages – many neutral ones. Group detection approaches, unsupervised methods.

Types of bots

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

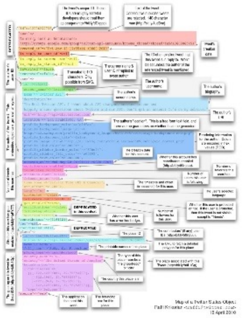
Lecture content

- 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 API

JSON file



Twitter post metadata

features

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

Extract some features that define the user's behavior

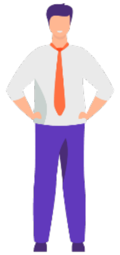
Annotated data

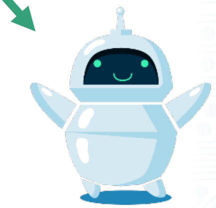
Labelled users (using Mturk, crowdsourcing, honey-pots, etc)

Annotated Data and Features

Train a Machine Learning model

Random Forest have shown *Good performance*





Trivial….Right?

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.

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

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.

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

Models usually output binary labels

Difficulty on detecting human-driven behaviors

The availability of ground truth datasets

Supervised ML models efficiency relies on the training data. Not many labelled datasets available.

Datasets do not include new types of bots

Difficulty on adapting models to newly introduced bots

Credibility of available datasets

Existing ones are annotated by humans (annotation biases)

Models are usually black box models

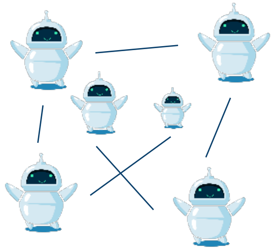
They do not provide feedback for the prediction

Supervised ML focus on classifying **instances** and **not groups**

Beyond supervised ML approaches : from Individuals to Groups

Identify individuals

Identify groups



Group approaches

- Unsupervised
- Semi-supervised
- Graph based

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

however…

Non Real time detection

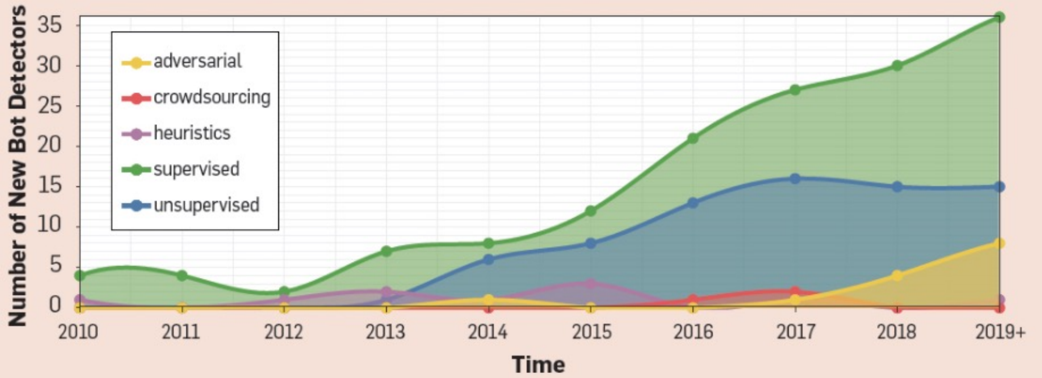
Most of these approaches do not provide real time predictions

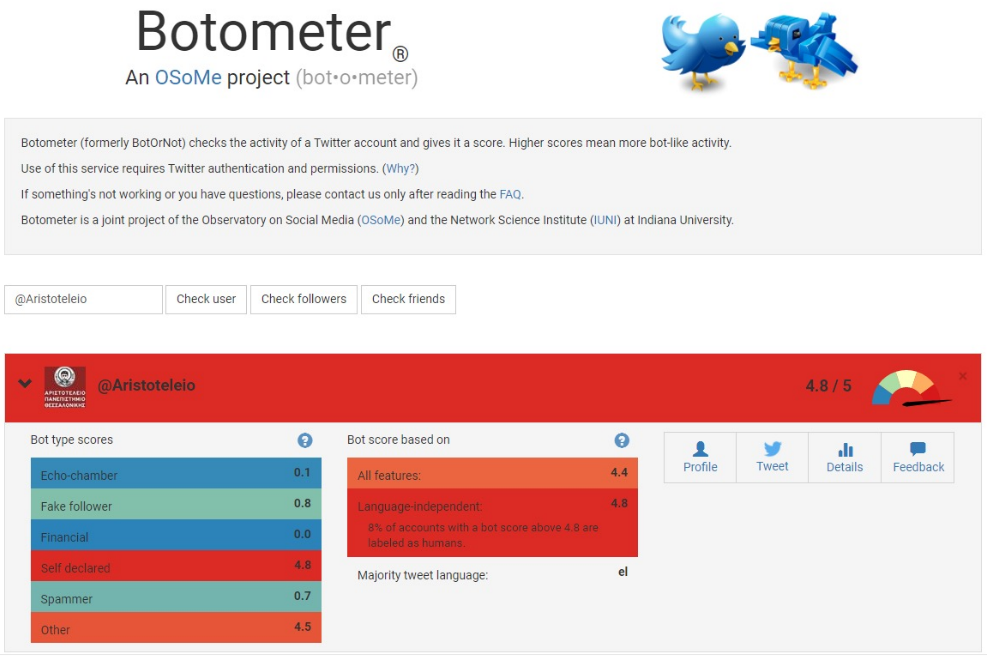
Computational heavy

These methods rely on more complex algorithms and more data.

SotA

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*

7. Cresci, Stefano. "A decade of social bot detection." *Communications of the ACM* 63.10 (2020): 72-83.
8. Yang, Kai-Cheng, et al. "Arming the public with artificial intelligence to counter social bots." *Human Behavior and Emerging Technologies* 1.1 (2019): 48-61.
9. Davis, Clayton Allen, et al. "Botornot: A system to evaluate social bots." *Proceedings of the 25th international conference companion on world wide web*. 2016.
10. Yang, Kai-Cheng, et al. "Scalable and generalizable social bot detection through data selection." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. No. 01. 2020.

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

Bot-Detective – an initial approach

Models are usually black box models

They do not provide feedback for the prediction

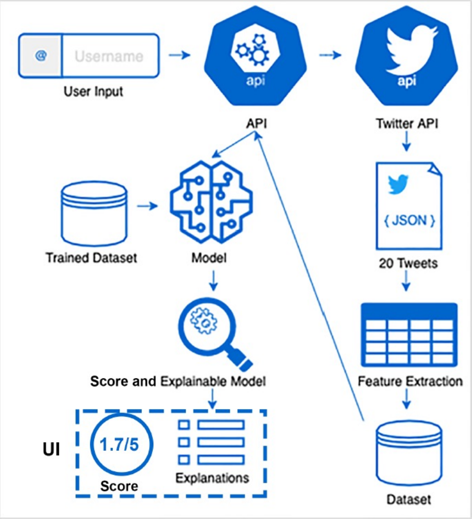
Need for more, open bot-detection services

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

Bot Detective

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

Features								
Type:[C:*Content* - U:*User*] - Value:[N:*Number* - B:*Boolean*- R:*Ratio*]								
Name	**Type**	**Value**	**Name**	**Type**	**Value**	**Name**	**Type**	**Value**
URLs	C	N	Words	C	N	Numeric Characters	C	N
Hashtags	C	N	Symbols	C	N	Mentions	C	N
Times favourite	C	N	URLs-Words	C	R	Hashtags - Words	C	R
Times Retweeted	C	N	Media	C	N	Characters	C	N
Sensitive Tweet	C	B	Followers	U	N	Followees	U	N
Followers-Followees	U	R	Tweets	U	N	Lists	U	N
Favourite Tweets	U	N	Def. Profile	U	B	Profile Description	U	B
Verified	U	B	Def. Image	U	B	Profile location	U	B
Profile URL	U	B	Username Characters	U	N	URLs in description	U	N
Screen name characters	U	N	Characters in description	U	N	Bot word in username	U	B
Bot word in screen name	U	B	Bot word in description	U	B	hashtags in username	U	N
Numeric chars in username	U	N	Numeric chars in screename	U	N	hashtags in description	U	N

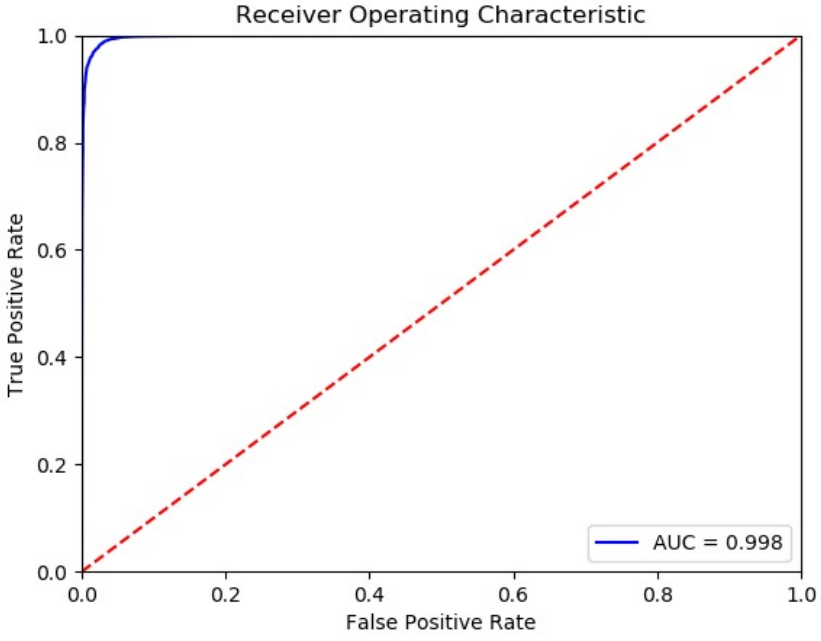
[11] https://www.privacy-regulation.eu/en/r71.htm - *should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached*

[12] Kouvela, Maria, Ilias Dimitriadis, and Athena Vakali. "Bot-Detective: An explainable Twitter bot detection service with crowdsourcing functionalities." *Proceedings of the 12th International Conference on Management of Digital EcoSystems*. 2020.

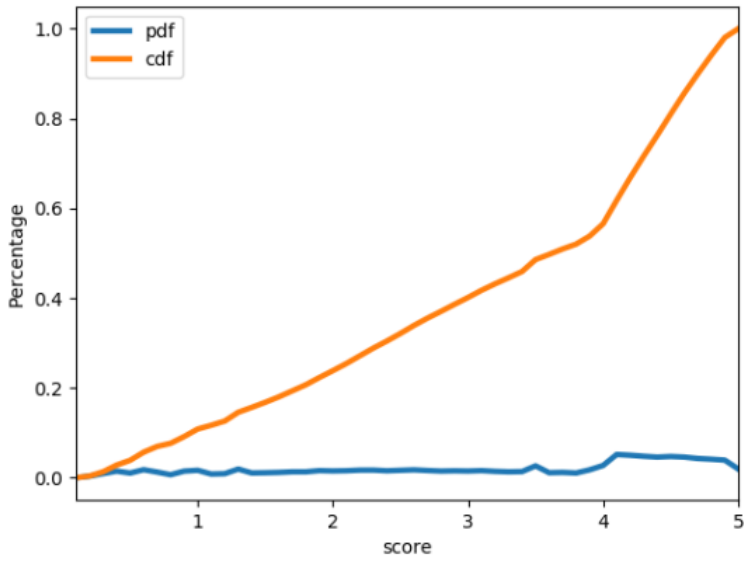
Bot-Detective ML Model

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)



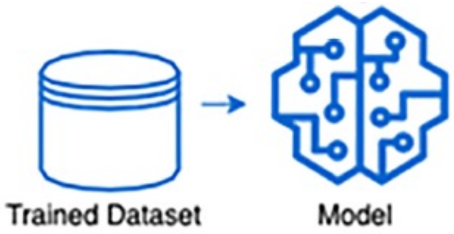
ROC-Curve

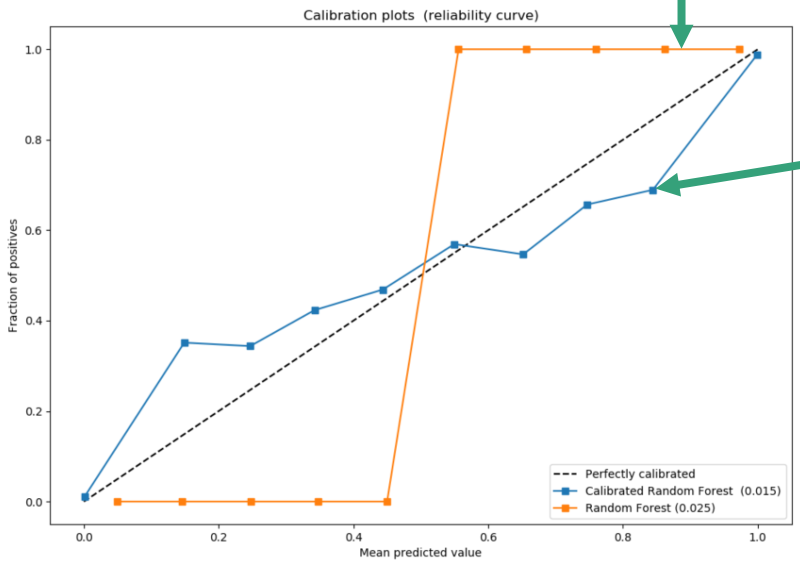


Overall Balanced Dataset

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



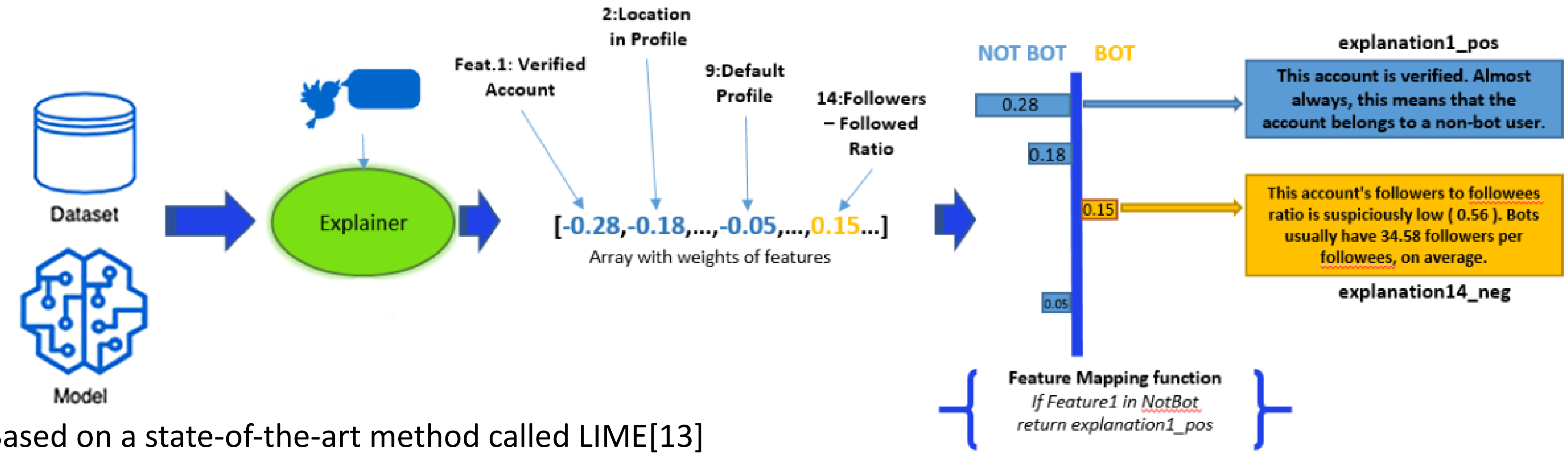


Our model tends to push the predicted probabilities away from 0[human] and 1[bot].

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

[14] Niculescu-Mizil, Alexandru, and Rich Caruana. "Predicting good probabilities with supervised learning." *Proceedings of the 22nd international conference on Machine learning*. 2005.

Bot-Detective Explainer



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"

[13] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should I trust you?" Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016.

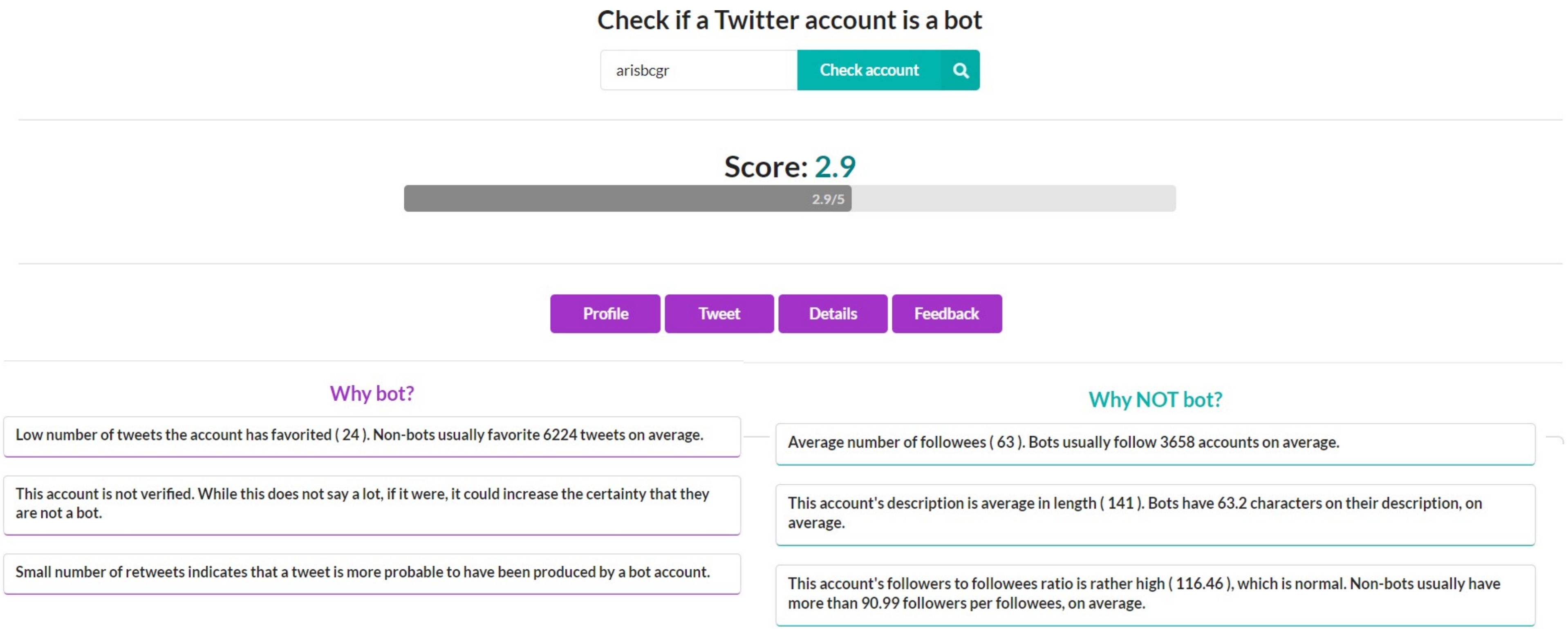
Lecture content

- 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

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.



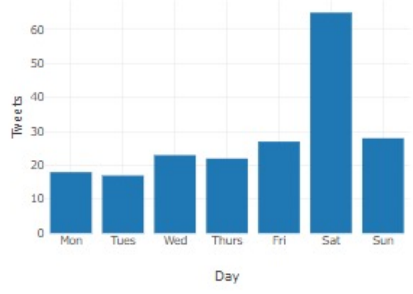
Bot Detective as a Web Service

The user can see some statistics with respect to the account of interest by clicking on Details:

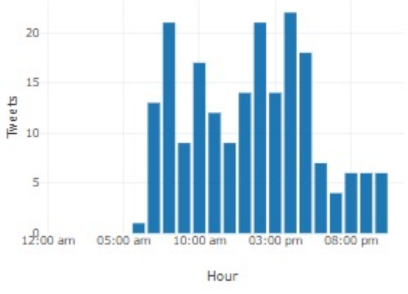
Screen name	ARISBCgr
Display name	ARIS B.C.
Twitter user ID	717972624
Description	🏀Welcome to the Official Twitter account of Aris Basketball Club | 🏆x10 Greek Championships 🏆x8 Greek Cups 🏆x3 European Trophies
Location	Thessaloniki, Greece
URL	https://t.co/cEVzu796DF
Date joined	2012-07-26 11:58:45
Most recent post	2020-11-01 17:59:25

Tweets	11132
Following	45
Followers	7364
Likes	42
Lists	64
Tweet language	el
Tweets this week	7
Retweet ratio	1%

Tweets by day of week



Tweets by hour of day



..and can also provide his/her own feedback regarding the prediction:

Help us improve

I believe the account @arisbcgr is:

Choose One

The provided explanations helped me understand why the account @arisbcgr has been characterized as a bot or not-bot:

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

I agree with:

- [x] Low number of tweets the account has favorited (42). Non-bots usually favorite 6224 tweets on average.
- [x] 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.
- [x] Small average number of favorited tweets. Bots usually have 0.02 favorited tweets on average. This account has 26.
- [x] Small number of retweets indicates that a tweet is more probable to have been produced by a bot account.
- [x] This account does not have a URL in their profile's description. Most non-bot users do.
- [x] Suspicious number of followers (7364). Bots usually are followed by 3469 accounts on average.
- [x] This account's URL per word ratio for each tweet, is suspiciously high.
- [x] This account usually adds media in their tweets (1.0 per tweet). This is on par with bot-like accounts, who have, on average, 0.02.
- [x] This account's tweets are very big in length (29). Non-bots usually tweet small pieces of text.
- [x] Average number of followees (45). Bots usually follow 3658 accounts on average.
- [x] This account's description is average in length (128). Bots have 63.2 characters on their description, on average.
- [x] This account does not have a default profile, when 66% of bots, on average, have.
- [x] This account's number of tweets is rather large. This occurs mostly in non-bot accounts.
- [x] This account has set a URL on their profile. Most bot accounts do not.
- [x] This account's number of numeric characters in their screen name is normal (0 characters).
- [x] This account's followers to followees ratio is rather high (163.64), which is normal. Non-bots usually have more than 90.99 followers per followees, on average.
- [x] This account shares their location on their profile. Most bot accounts do not.
- [x] This account's screen name length is normal (8 characters). Bots have 11.3 characters on their name, on average.
- [x] This account does not have hashtags in its profile description. Only 31% of non-bot accounts have hashtags in their profile descriptions.
- [x] This account's number of numeric characters in their name is normal (0 characters).
- [x] This account's name length is normal (9 characters). Bots have 12.3 characters on their name, on average.
- [x] This account uses symbols rarely (24.15 symbols per tweet). Bots usually have 21.2 symbols per tweet, on average.
- [x] This account's followees to followers ratio is not very high (0.01), which is normal. Bots usually have 3.91 followees per followers, on average.
- [x] Normal average number of URLs per tweet (0.9). Bots usually have 0.55 URLs per tweet.
- [x] Normal average number of characters per tweet (177.5). Bots usually have 143.7 characters on their tweets.
- [x] Normal number of hashtags on tweets. Bots usually have 3.48 hashtags on their tweets and this account has 1.
- [x] Normal amount of hashtags per words in tweets. Bots usually have 0.24 hashtags per words in their tweets and this account has 0.

Send us your suggestions

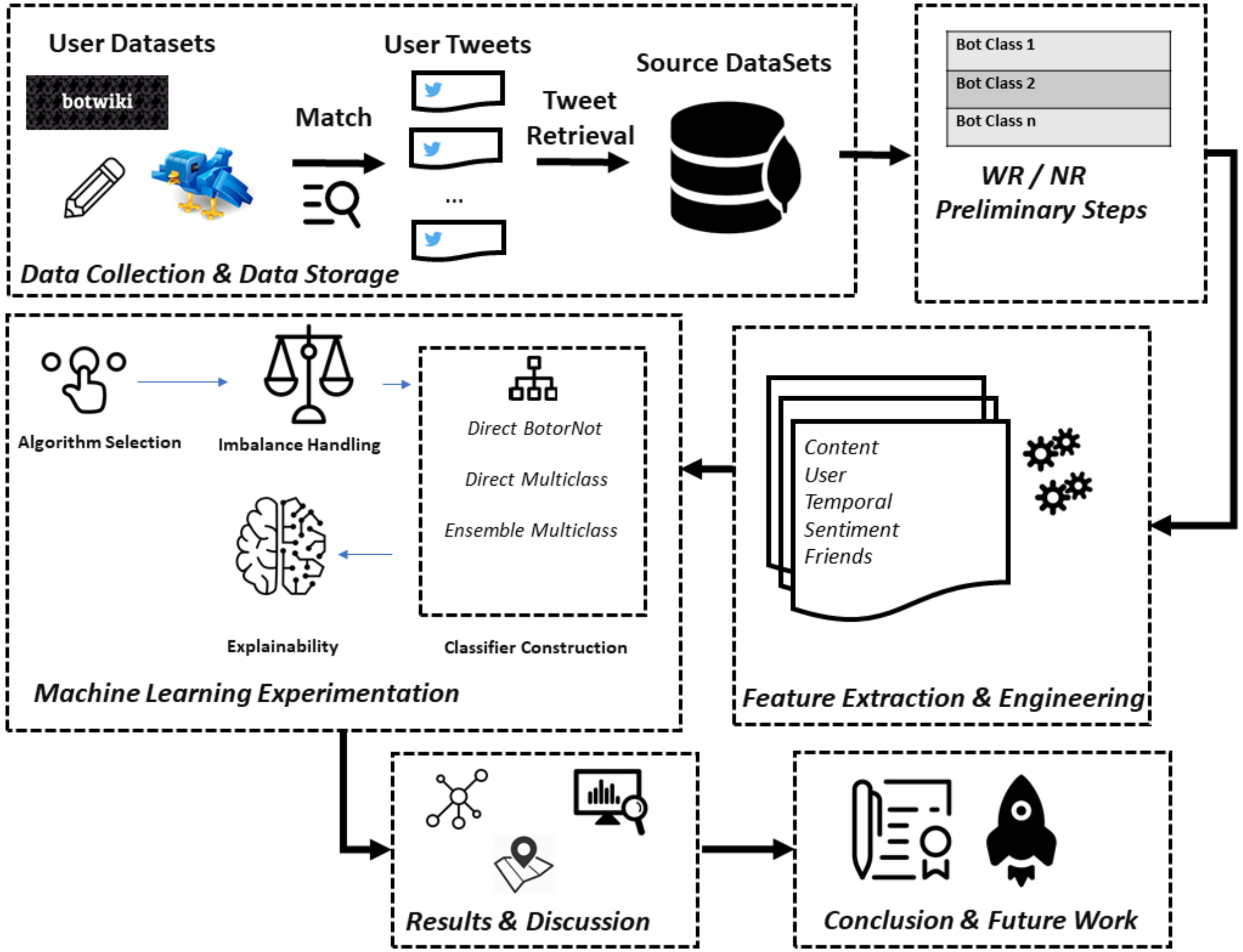
✓ Send

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 :

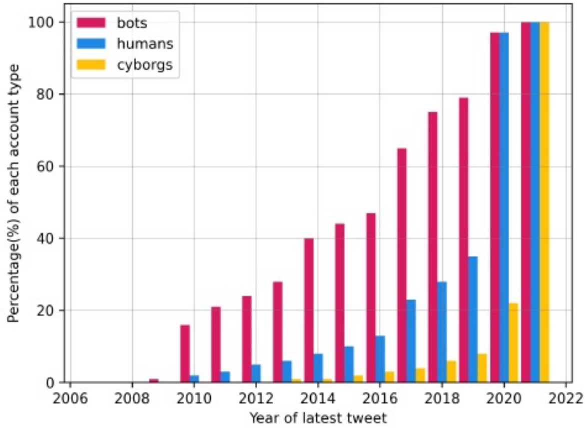
- Insightful dataset analysis
- New Bot types
- New Features
- New Models
- New Explainability approach

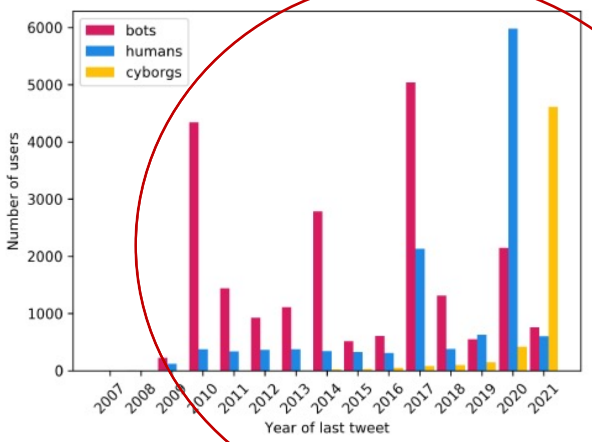
New Publication: Social Botomics [14]

[14] Dimitriadis, Ilias, Konstantinos Georgiou, and Athena Vakali. "Social Botomics: A Systematic Ensemble ML Approach for Explainable and Multi-Class Bot Detection." *Applied Sciences* 11.21 (2021): 9857.

Bot Detective V2.0

Exploratory analysis of most bot related datasets





Most datasets are outdated

Credibility of available datasets

Existing ones are annotated by humans (annotation biases)

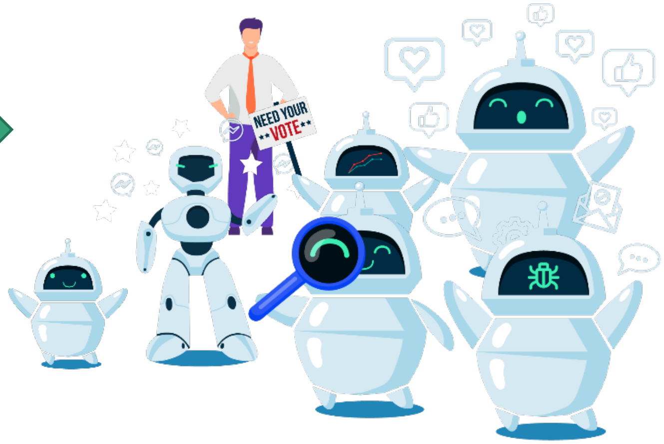
Datasets do not include new types of bots

Difficulty on adapting models to newly introduced bots

Bot Detective V2.0 – Introducing new Bot Types

Exploratory analysis of Datasets

Merge multiple (24) annotated open bot datasets



Most datasets referred to different bot types

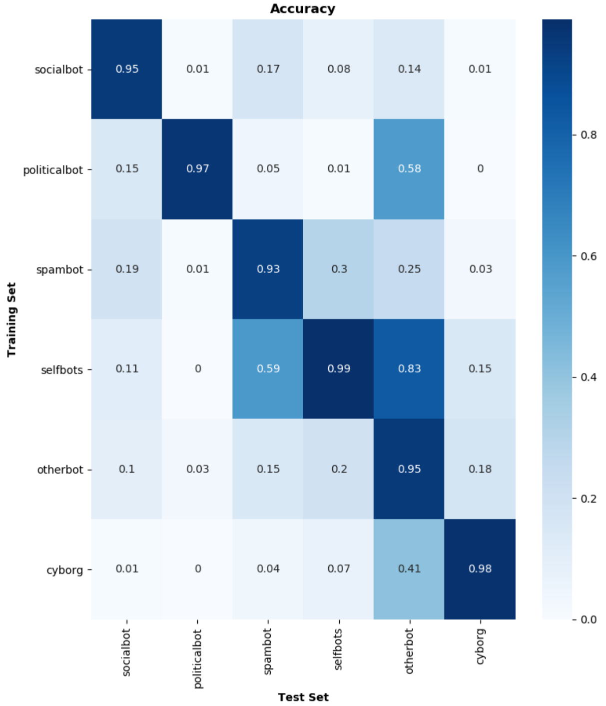
Propose a new bot taxonomy – 6 different bot types

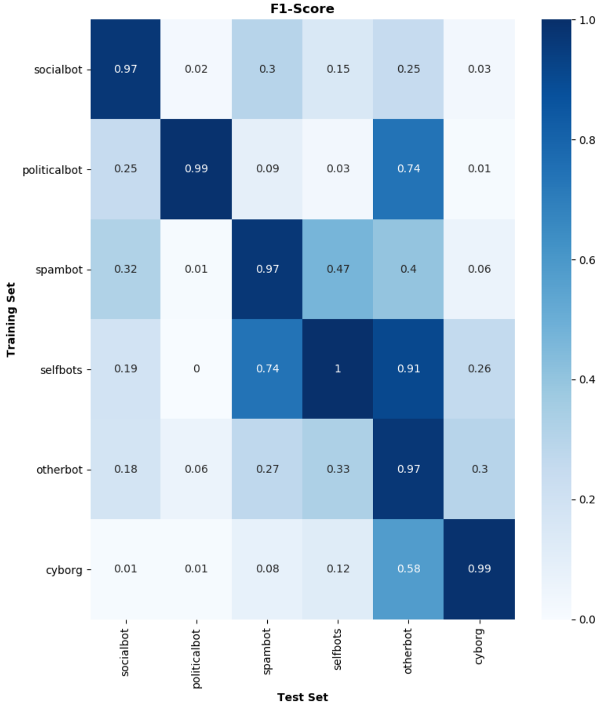
Bot type	Description	Number of Datasets
Spam Bots	Accounts that post spam content	4
Social Bots	Bots that try to attract followers	4
Political Bots	Bots involved in politics online discussions	3
Cyborgs	Human monitored bots	3
Self-declared	Accounts that state they are bots	1
Other bots	Other types of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

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

↓

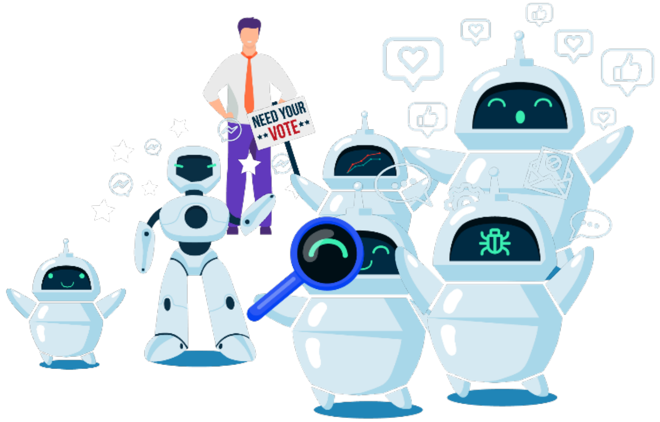
- Highlights the different behavior of bots
- need for the distinction of bots in separate types

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

New Models

<u>Binary Bot or Human Classifier</u>

- 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



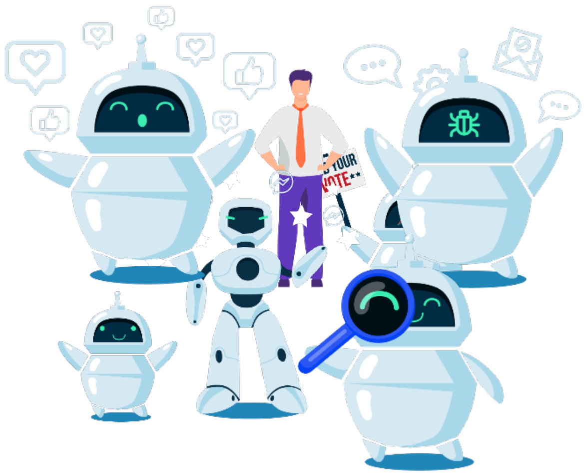
<u>Multi Class Classifier</u>

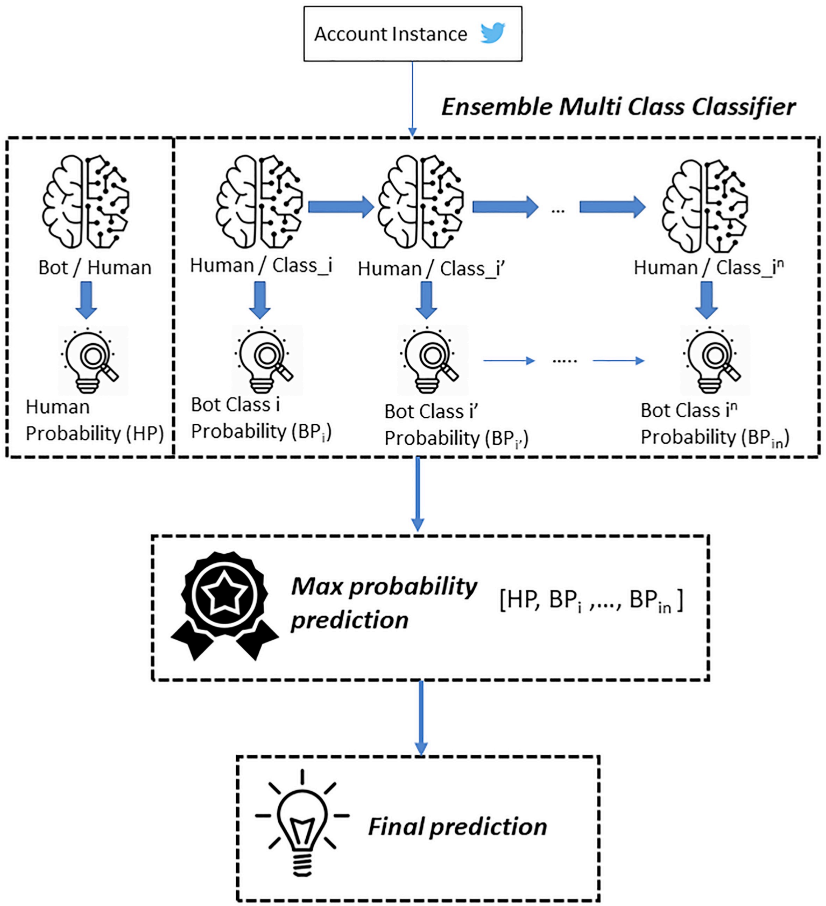
- Trained on all datasets (75%-25% train/test) with 6 different labels
- Experimented with multiple different classifiers
- ADASYN imbalance handling
- Best: Ensemble of Random Forests
- ACC: 0.9
- ACC: 0.9
- Precision: 0.891
- Recall: 0.918

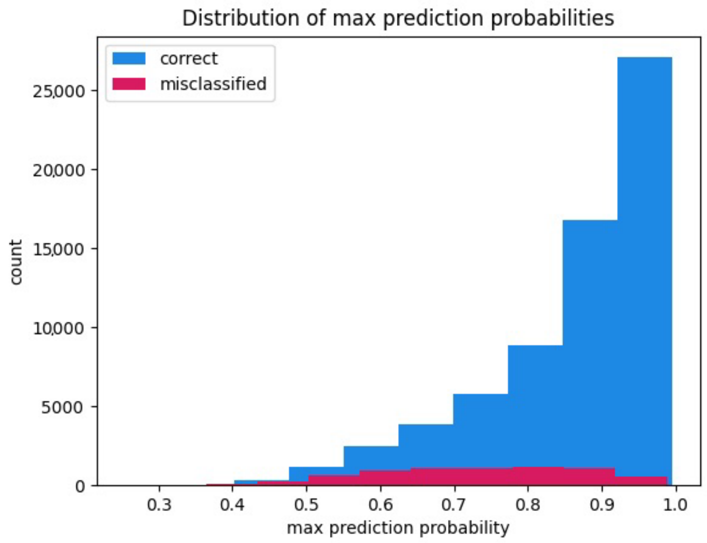
Comparable and higher performance to other SotA models

Ensemble of Binary Bot Classifiers for multi class predictions

Features?







Our model predicts the instance class with higher confidence

Feature Engineering

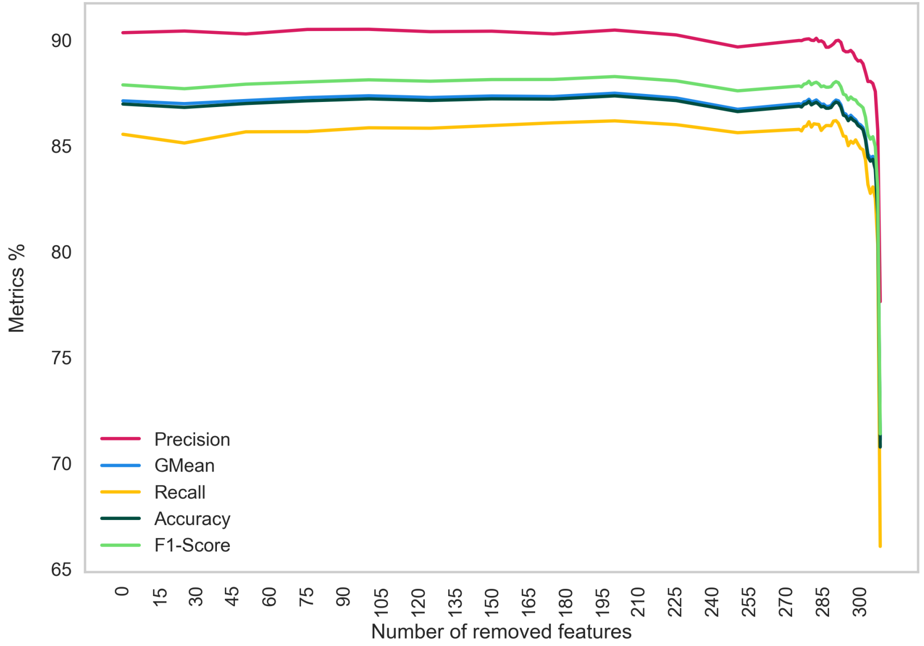
Feature Types - categorization

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

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

Feature Extraction

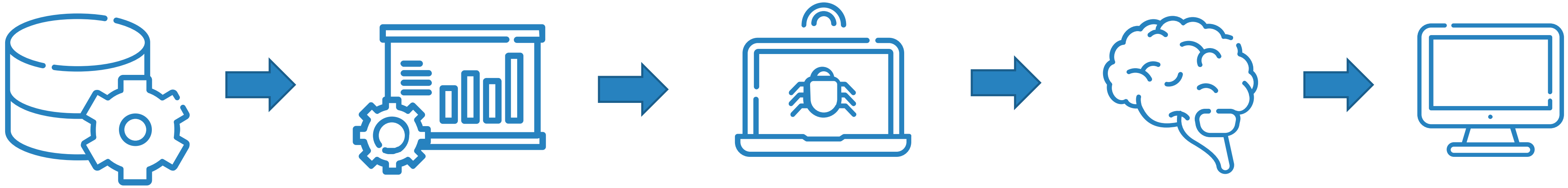
Costly process, both in terms of time and resources!

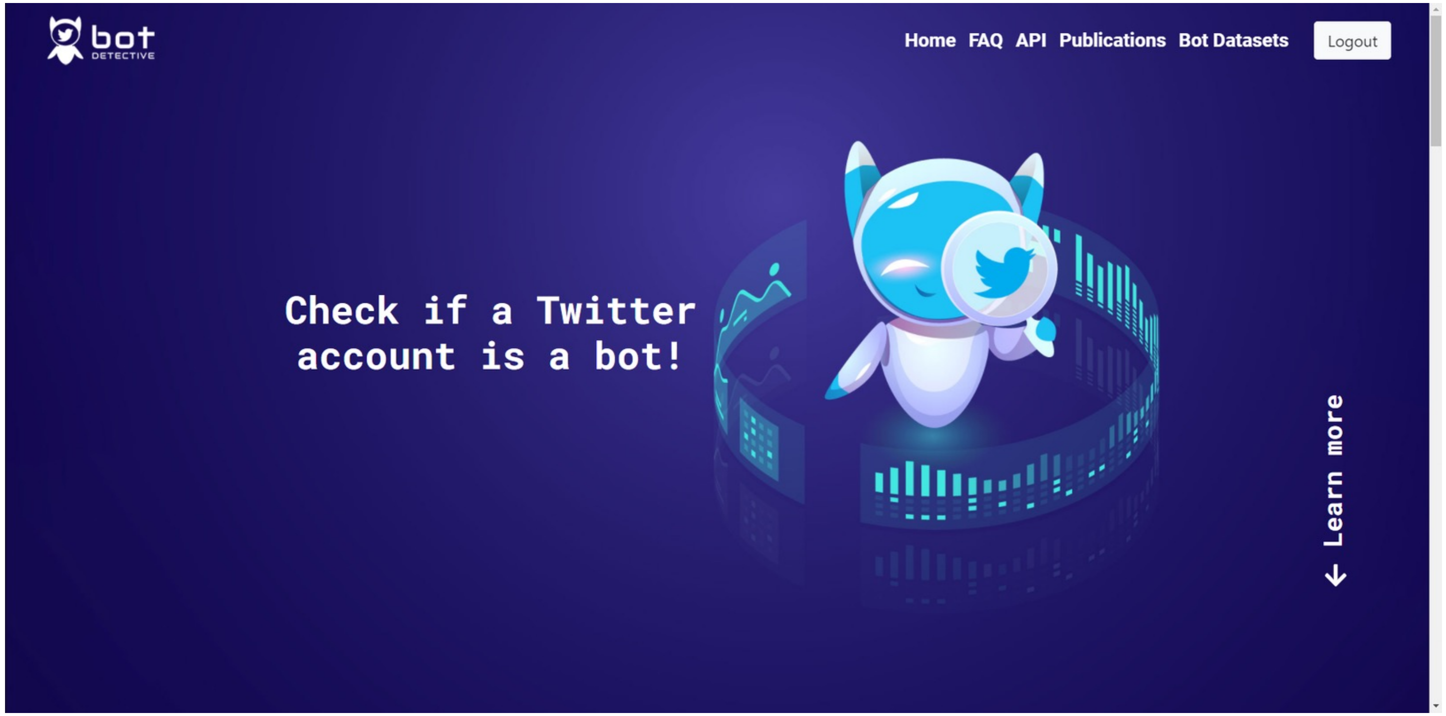


- Utilized feature importance frameworks
- Iteratively removed less important
- Best performance with just 145 features
- Performance still high with even **45** features

Bot Detective 2.0

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

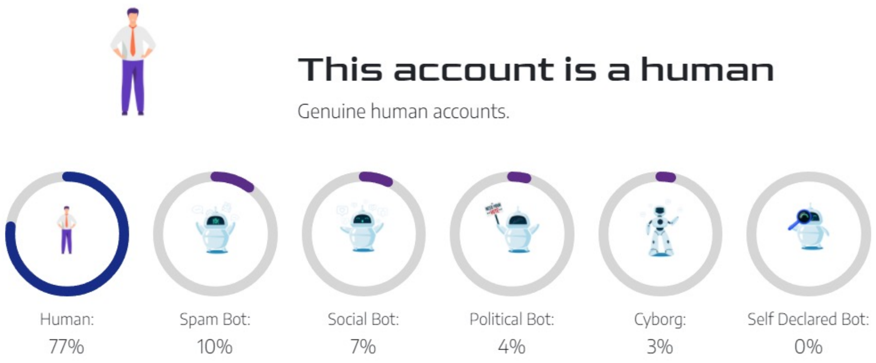




Bot Detective 2.0

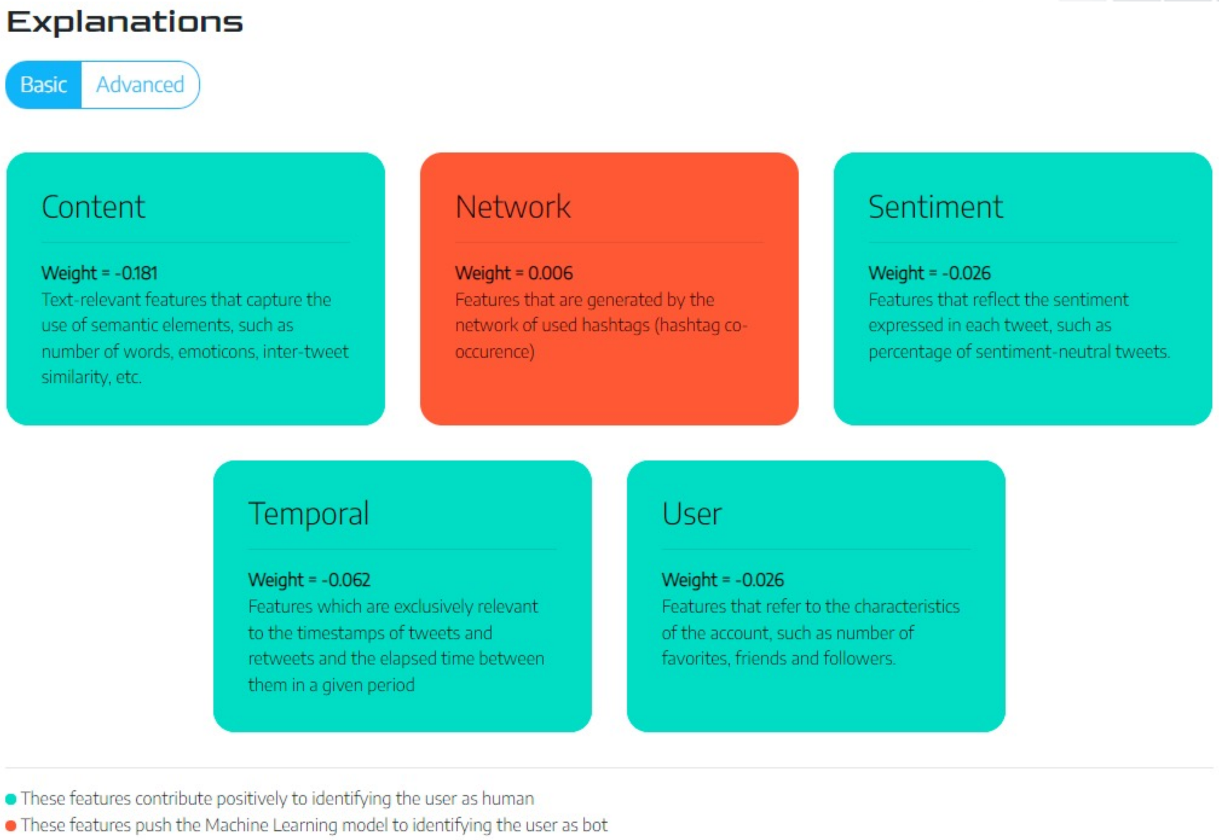
Bot Detective 2.0

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



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

New Explainability Functionalities



Bot Detective 2.0

Per Feature explanations:

Explanations

Basic | Advanced

Content | Network | Sentiment | Temporal | User

name_screen_name_similarity 0.00911
0.30 < name_screen_name_similarity <= 0.77
Similarity index of user's name and screen name

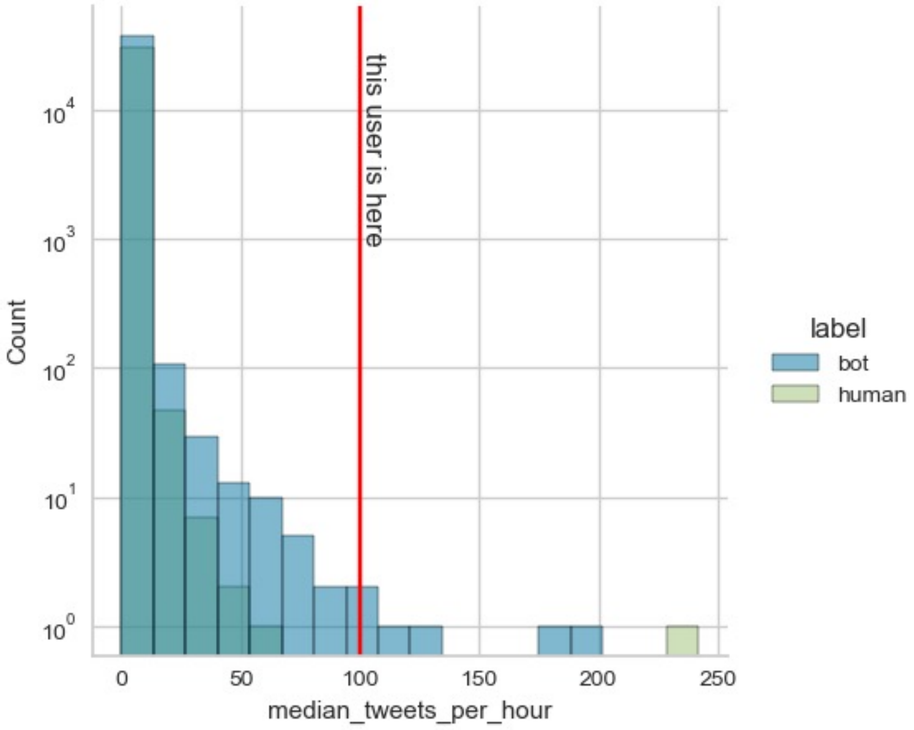
followers_count -0.00869
43.50 < followers_count <= 5762.50
Number of followers

default_profile -0.00749
default_profile <= 0.50
Whether the users has a default profile

tweets_count -0.00616
148.50 < tweets_count <= 3696.50
Total number of posted tweets

urls_in_description -0.00396
urls_in_description <= 0.50
Number of urls in user's description

... available soon – Chart comparison



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

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

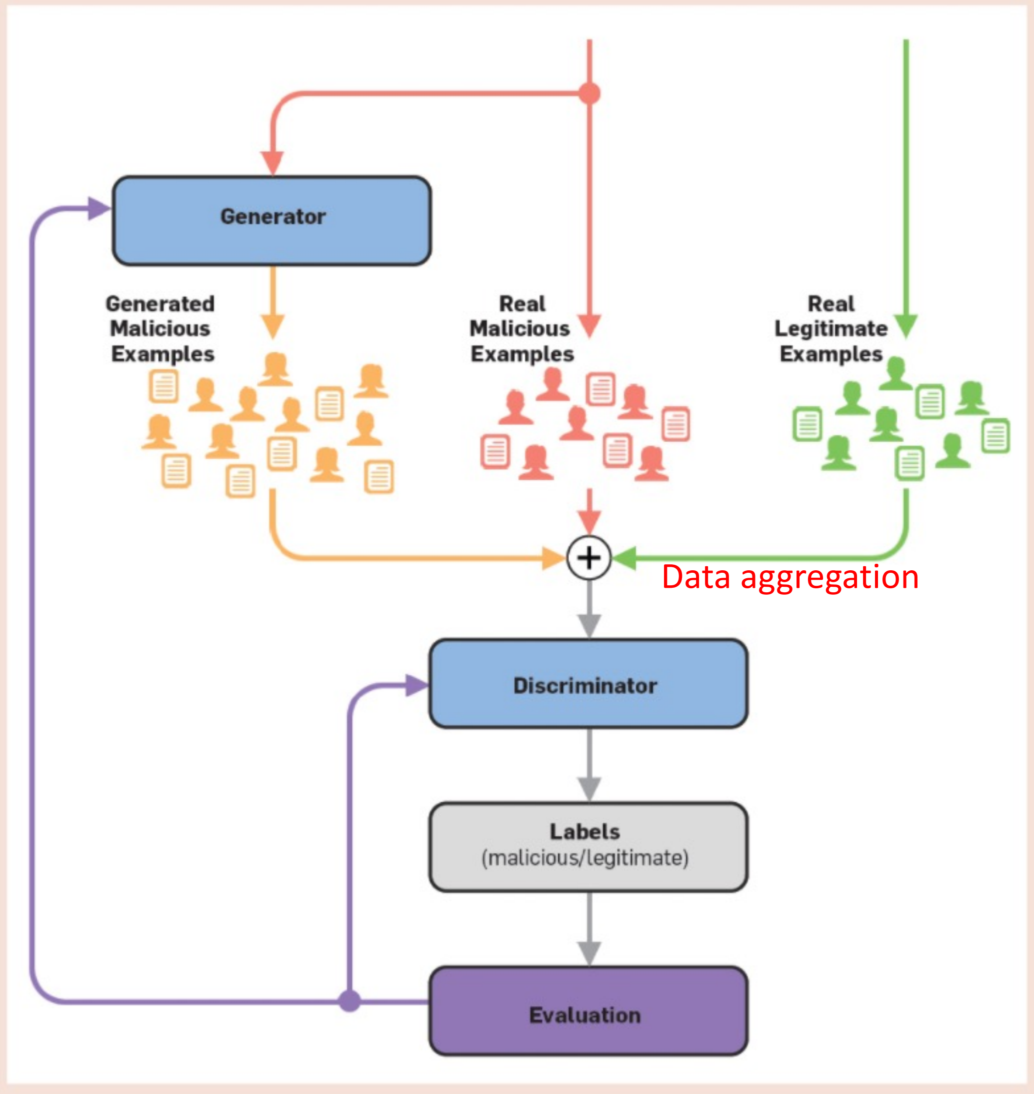


Figure taken by: **A Decade of Social Bot Detection**
By Stefano Cresci
Communications of the ACM, October 2020, Vol. 63 No. 10, Pages 72-83

Open Questions & Future Work

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)

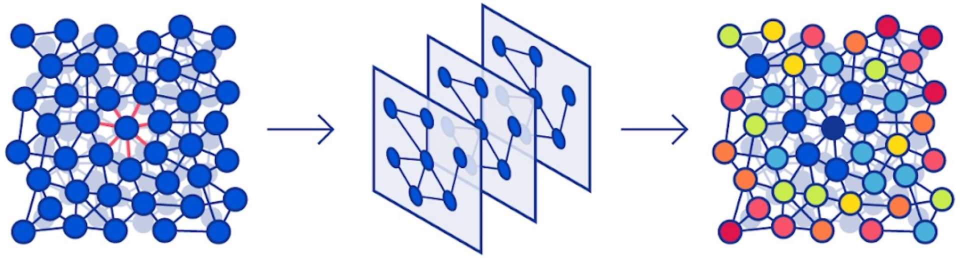
- No need for labelled datasets
- More promising results
- See Next Slide

Sequence alignment methods

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

16: Wu, Bin, et al. *"Using improved conditional generative adversarial networks to detect social bots on Twitter." IEEE Access 8 (2020): 36664-36680.*
17: Chavoshi, Nikan, Hossein Hamooni, and Abdullah Mueen. *"Debot: Twitter bot detection via warped correlation." Icdm. 2016.*

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 POIITIS
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?