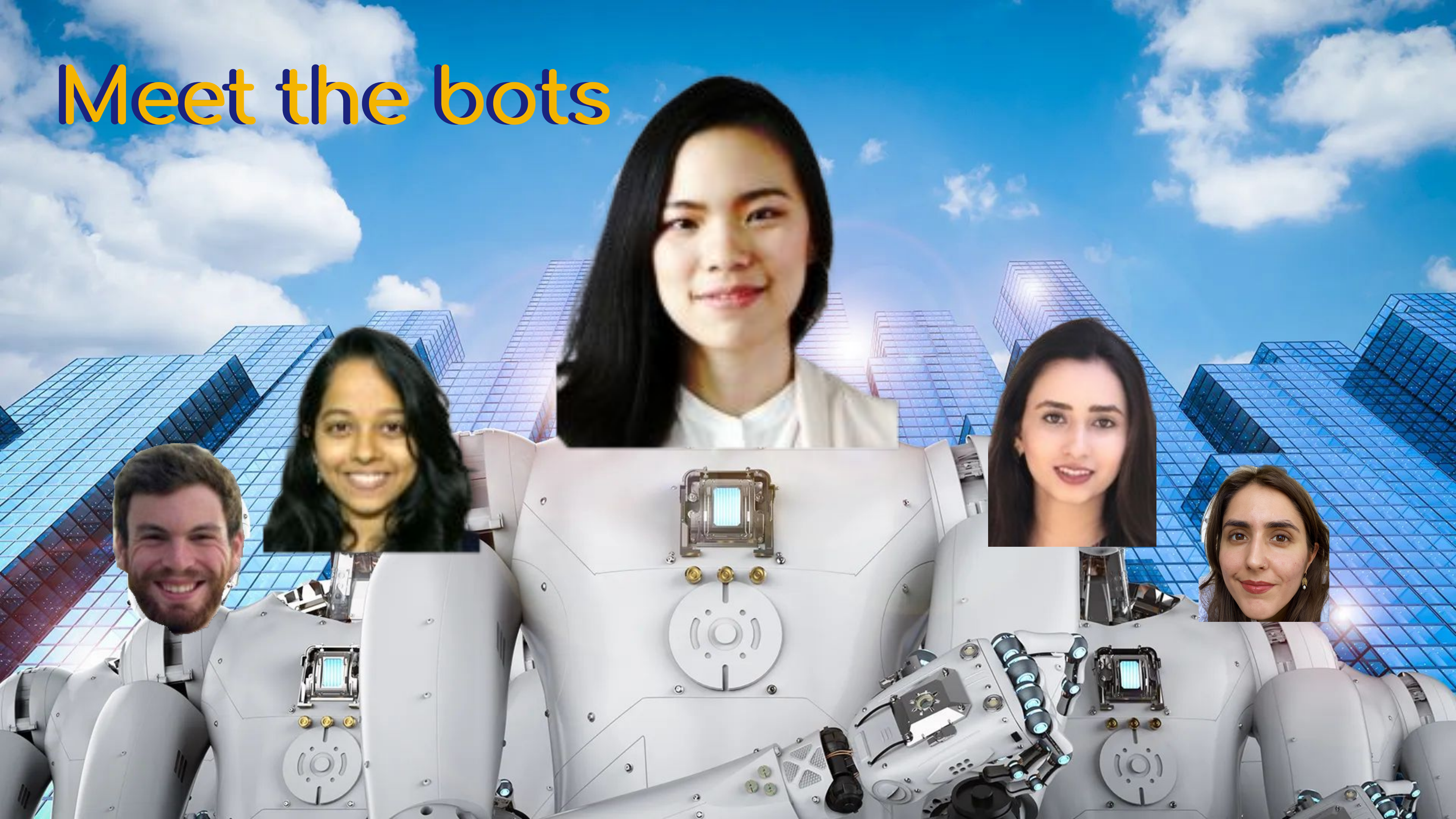
Spot the bot: Developing a bot detection algorithm

Mahnoor Ayub, Aishwarya Kura, Oravee Smithiphol, Matt Lampl, and Sara Maillacheruvu

Meet the bots



Context setting

Problem definition

- Bots rapidly repost content
- Variety of uses:
  - Terrorism, hate speech, harassment, political propaganda, spamming, civic engagement, commerce
- Huge threats to national, international security
  - Election manipulation, conspiracy theories
- Interconnected systems, global impact

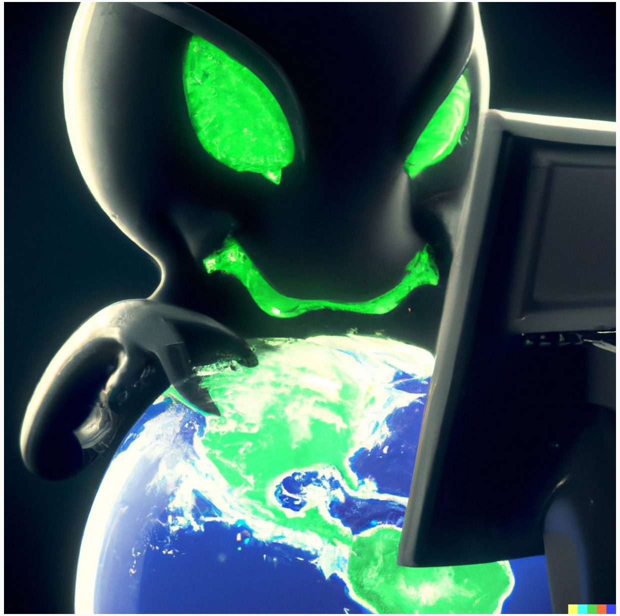


Image generated by DallE, given prompt: "a globe with an evil computer bot trying to control it, digital art"

Existing approaches

- 2011: UT Austin: Honey pots
  - Used bots to generate senseless content that was only meant to attract other bots.
- 2014: IU, USC: Botometer, supervised ML
  - Uses machine learning algorithms to extract over 1000 predictive features that identify suspicious behaviors,
  - Produces an ensemble classification score on a normalized scale that indicates the likelihood that a Twitter account is a bot.
  - Scores closer to 1 indicate a higher probability of being a bot, while scores closer to 0 are more likely to belong to humans.

Existing approaches

- 2018: Bot-hunter: Tiered approach combining many techniques
    - Uses machine learning and manual verification.
    - Bot accounts: often have randomly generated, alphanumeric 15-character string handle
    - 60% have profile pics, many of which are the same
    - Random Forest model performed best and achieved AUC = 0.994 with tuning.

Data overview

Surprises, challenges, paths not taken

- Originally aimed to develop fake news detector
  - Pivoted based on conversation with professor
  - Text-based approaches tested here really faltered
  - Set us back, timeline-wise
- Initial data access issues
  - Large files
  - Data infrastructure can be a big roadblock
- If doing again:
  - Test models on other datasets, especially on non-Twitter data
  - Allocate more time for data infrastructure issues
  - Start with bot detection first

Data Summary

Dataset

Indiana University
Network Science Institute

OSoMe

LUDDY SCHOOL OF INFORMATICS, COMPUTING, AND ENGINEERING
INDIANA UNIVERSITY

- **Indiana University, The Observatory on Social Media**
  - Tweet data & Network Metadata from **2020**
    - Accessed on request
  - **JSON Nested Dictionaries**
    - **1.42 Million** Data points analyzed
    - Very large

Features Analyzed

Profile Features

ID, Location, joining date, activity

Tweets

About 200-300 tweets within the years, tags, likes, favourites

Network Info

#Following, #Followers, Retweets, Timing of retweet etc.

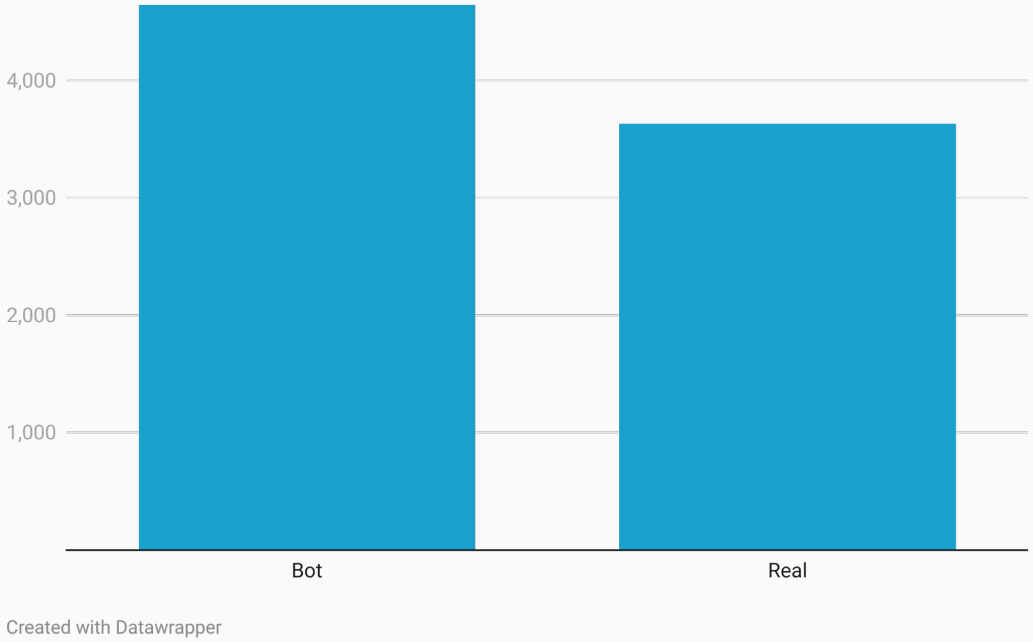
Hashtags, Labels

Tweet activity, RT count, hashtags,

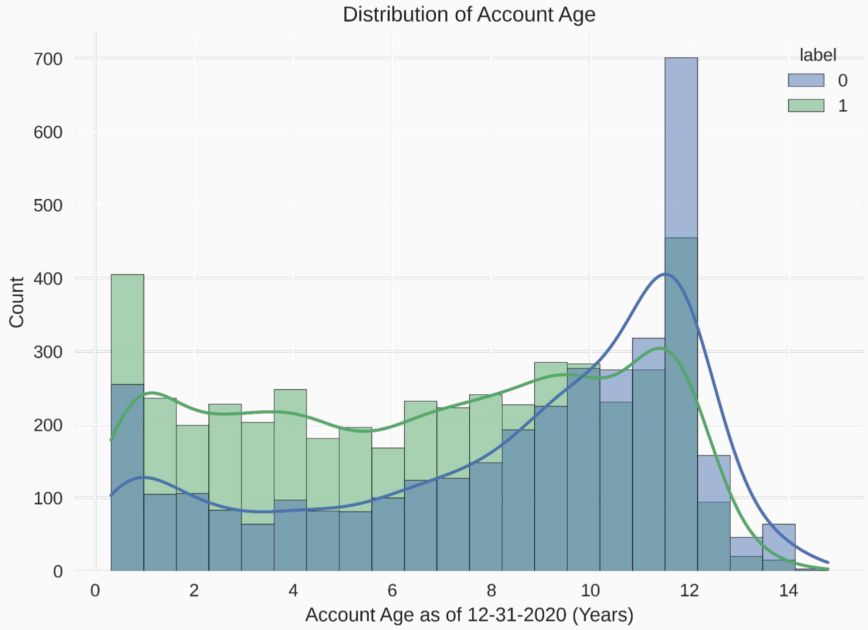
```
Columns in dataset:
id
id_str
name
screen_name
location
profile_location
description
url
entities
protected
followers_count
friends_count
listed_count
created_at
favourites_count
utc_offset
time_zone
geo_enabled
verified
statuses_count
lang
contributors_enabled
is_translator
is_translation_enabled
profile_background_color
```

EDA, data overview

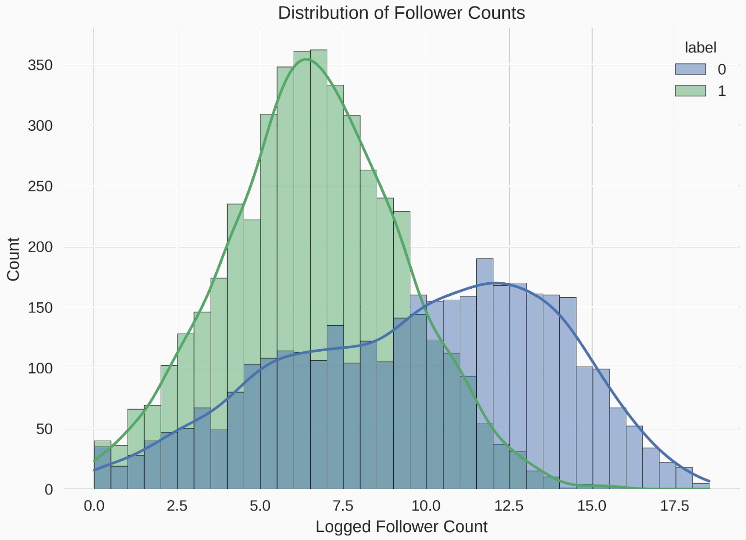
Count of Real vs. Bot Tweets

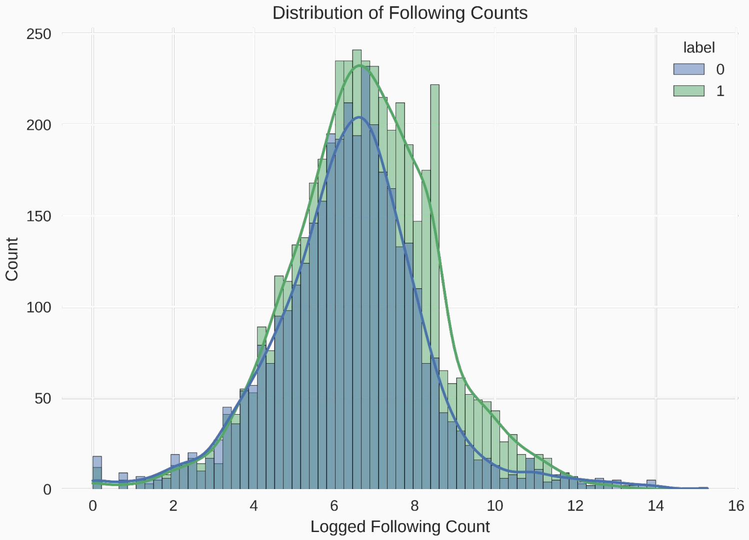


Bot accounts tend to skew younger

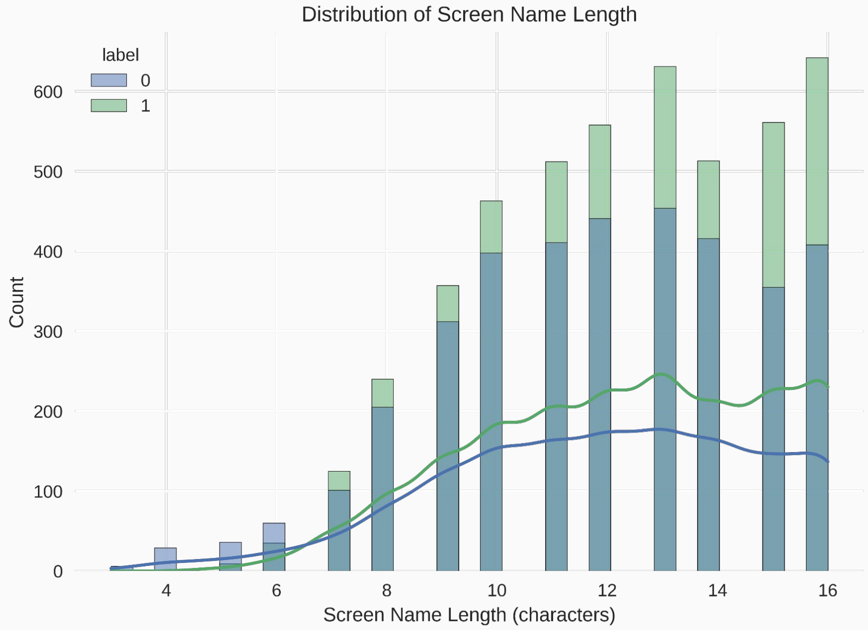


Bot accounts tend to have fewer followers





Screen Name Length similar among bots and real accounts



Our approach and results

Our approach

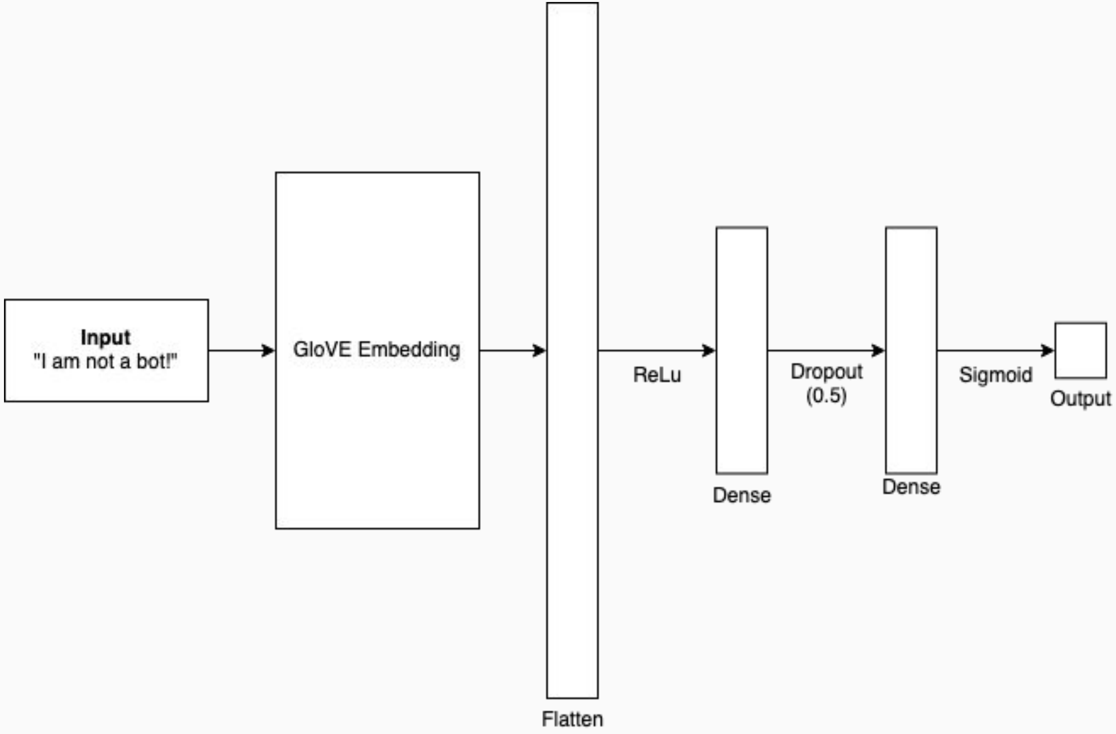
ML approaches we tested:

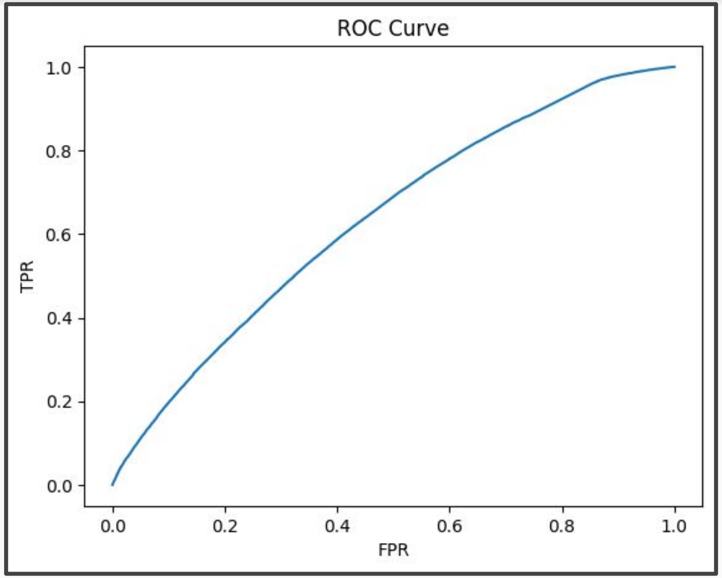
- GloVE Embedding Neural Net
- ML Classification (Random Forest) using Graph features

Data inputs we tested:

1. Tweet text only
2. User metadata only
3. Graph network only
4. Combinations of user metadata and graph networks

Lv.1 : Neural Net Approach (Tweet-based) (acc = 0.61)



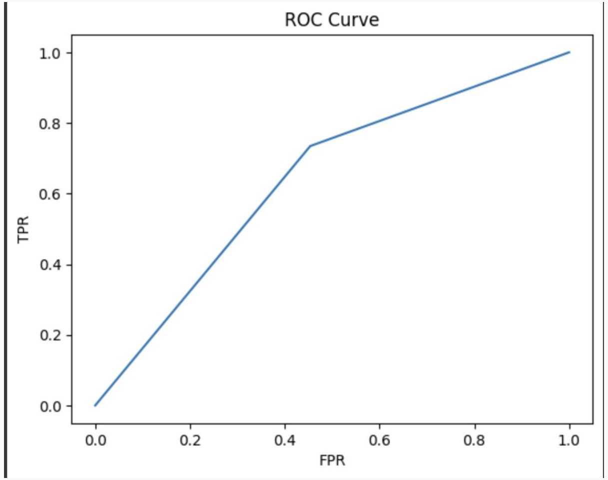


Accuracy	0.61
Precision	0.60
Recall	0.88

Lv. 2: User account-based (acc = 0.65)

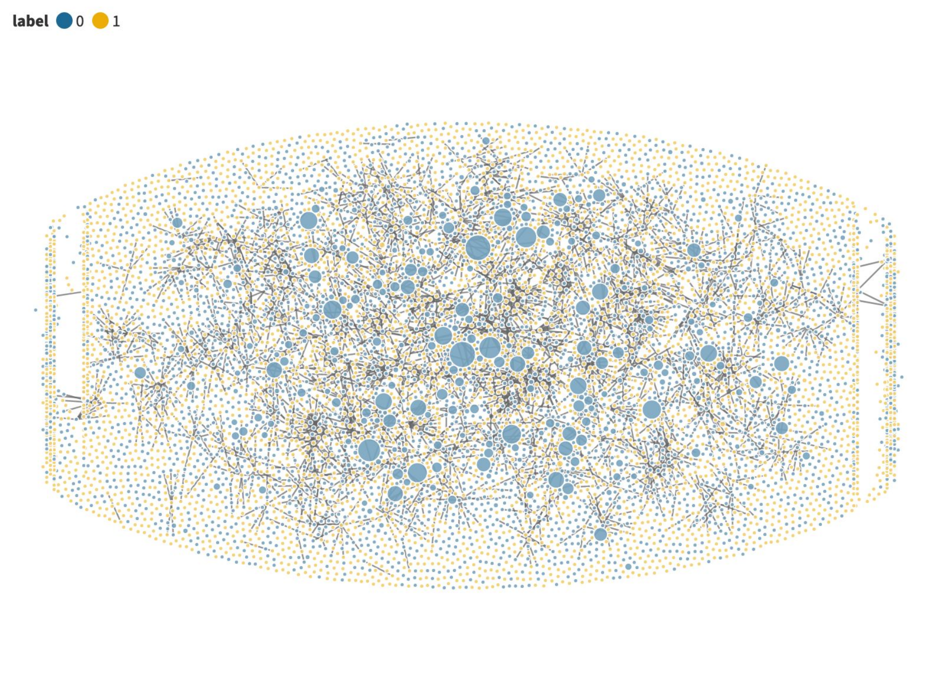
Features used: account age (corr = -0.2), name length, statuses count, favorites count

Accuracy	0.65
Precision	0.69
Recall	0.73

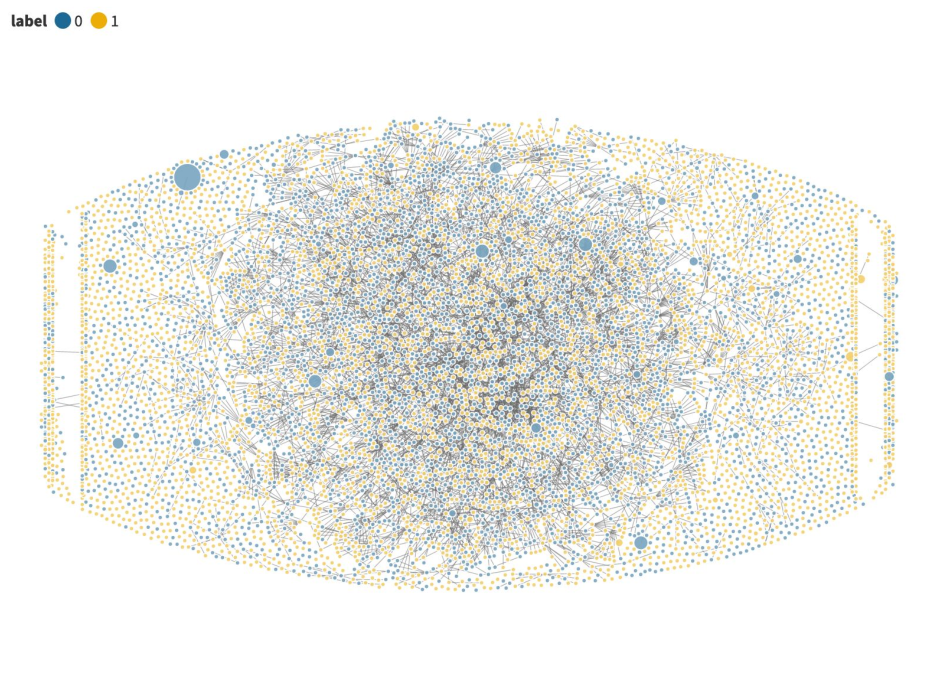


Lv. 3: Network & graph-based (acc = 0.71)

Network of Followers



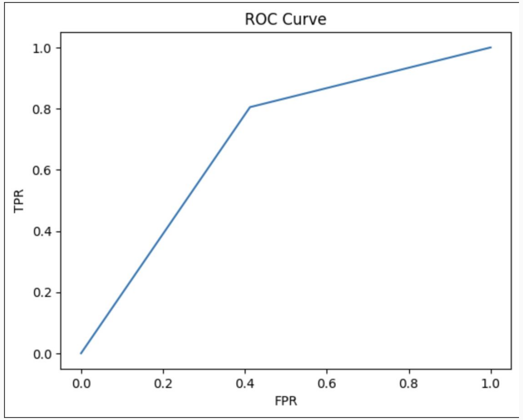
Network of Following



Lv. 3: Network & graph-based (acc = 0.71)

Features: following_count, followers_count (corr = -0.17)
Closeness centrality, Degree centrality, Eigenvector centrality (corr = +0.22)

Accuracy	0.71
Precision	0.73
Recall	0.8



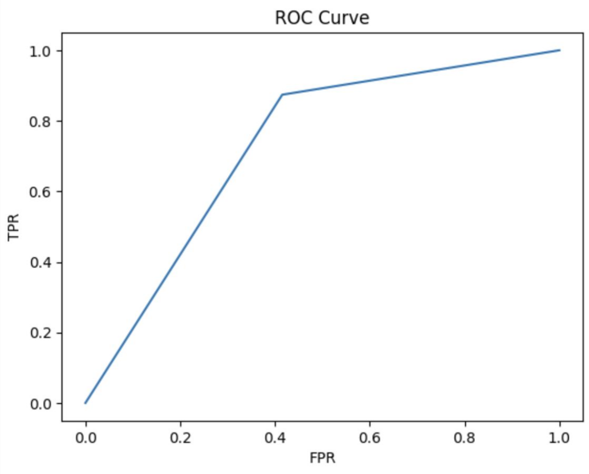
- Degree
  -> Bots follow more people
- Eigen
  -> Accounts bots follow, follow many people
- Closeness
  -> Close to all other nodes throughout the network (follow variety of people)

Lv. 4: user+graph-based (acc =0.75)

Features:

Tier 1: account age, name lengths, statuses_count, favorites_count
Tier 2: friends_count, followers_count, degree, eigenvector, closeness

Accuracy	0.75
Precision	0.74
Recall	0.87



Lessons learned

- Cannot use text alone to identify bots
- Metadata, network data crucial
- Advances in generative AI
    - Reinforce need to go beyond text-based approaches

Image generated by DallE, given prompt: "a globe with an evil computer bot trying to control it, digital art"

Next steps, future work

- Gather more metadata, network data
  - Network measures with high computational requirements, e.g. betweenness, community detection, etc.
- Test additional AI models
  - Models tested: Random Forest (presented) & Logistic regression
- Generalizability
  - Test on other platforms, not only Twitter data
  - Country-level analysis

Any questions?

Thank you!