Police Bot: Enhancing Social Media Governance with Policing Bots

Milestone 6 Presentation

Group Members:

Students:

- Gabriel Silva
- Cody Manning
- Liam Dumbell
- Nickolas Falco

Faculty Advisor / Project Client:

- Khaled Slhoub

Computer Science Project Instructor:

- Philip Chan

Overview:

- Discussion of Task Completion:
  - Code Improvements
  - Testing Metrics
  - Decide Module
  - Demo/Commercial
  - Update on Poster and Ebook Page
- Milestone Completion Task Matrix
- Advisor Feedback
- Lessons Learned

Testing Method

Using List of Known Humans:

- Compiled using the moderator lists of the top subreddits (where bot moderators are declared as bots, eg. AutoModerator)
- Used to test the percentage of false positives (Humans identified as Bots)

  List of Known Bots

Using List of Known Bots:

- Compiled using various public list of Reddit Bot accounts
- Used to test the percentage of false negatives (Bots identified as Humans)

Testing Metrics

Positive Assessments (Known Bot List)

- Objective: 70%-80% correct bot detection
- Results: 320/424 (bots) - - - - - 76% Bots identified as Bots

Negative Assessments (Known Human List)

- Objective: <20% false positives
- Results: 102/651 (humans) - - 15% of Humans identified as Bots

Code Improvements

- Enhanced organization
- Enhanced maintainability
- Enhanced reusability
- Enhanced Scalability

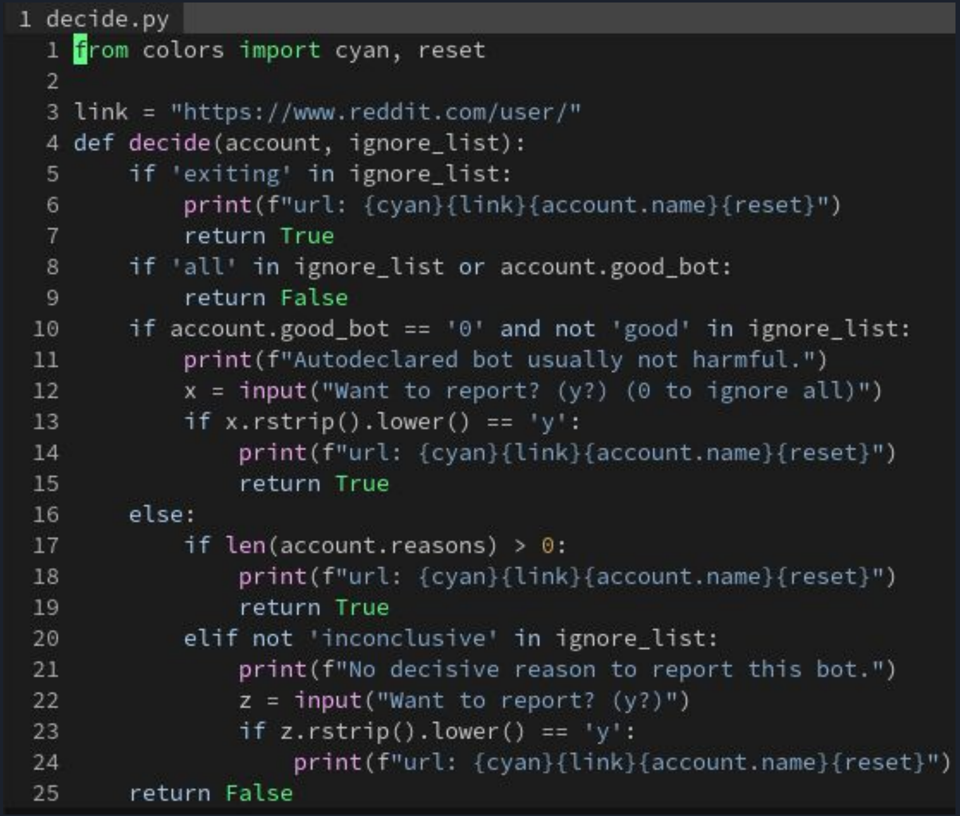
Decide Module

Decide to report or not a scanned bot:

- Scamming Bots or Harassing Bots (Bad)
  - Are provided with a link to the report page and a tutorial.
- Auto Declared Bots or Known Bots (Good)
  - Ask the user whether to offer the option to report good bots.

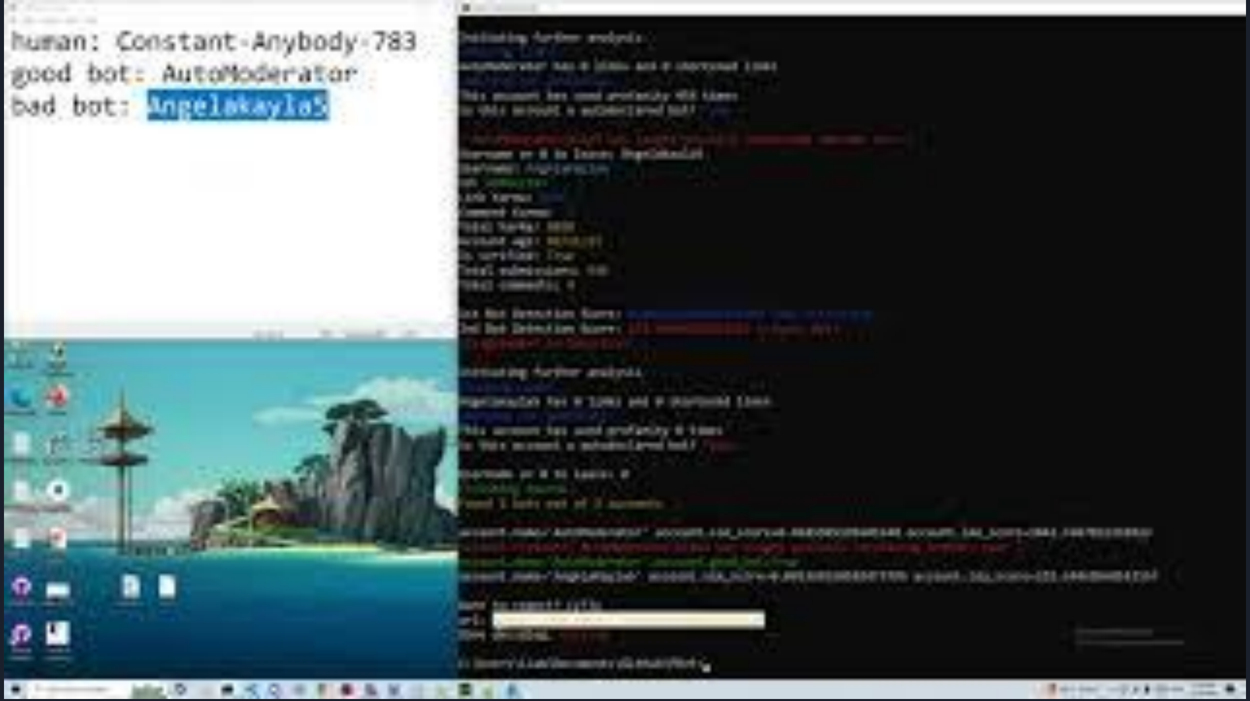
Decide Module

○



```python
from colors import cyan, reset

link = "https://www.reddit.com/user/"
def decide(account, ignore_list):
    if 'exiting' in ignore_list:
        print(f"url: {cyan}{link}{account.name}{reset}")
        return True
    if 'all' in ignore_list or account.good_bot:
        return False
    if account.good_bot == '0' and not 'good' in ignore_list:
        print(f"Autodeclared bot usually not harmful.")
        x = input("Want to report? (y?) (0 to ignore all)")
        if x.rstrip().lower() == 'y':
            print(f"url: {cyan}{link}{account.name}{reset}")
            return True
    else:
        if len(account.reasons) > 0:
            print(f"url: {cyan}{link}{account.name}{reset}")
            return True
        elif not 'inconclusive' in ignore_list:
            print(f"No decisive reason to report this bot.")
            z = input("Want to report? (y?)")
            if z.rstrip().lower() == 'y':
                print(f"url: {cyan}{link}{account.name}{reset}")
    return False
```

Demo/Commercial

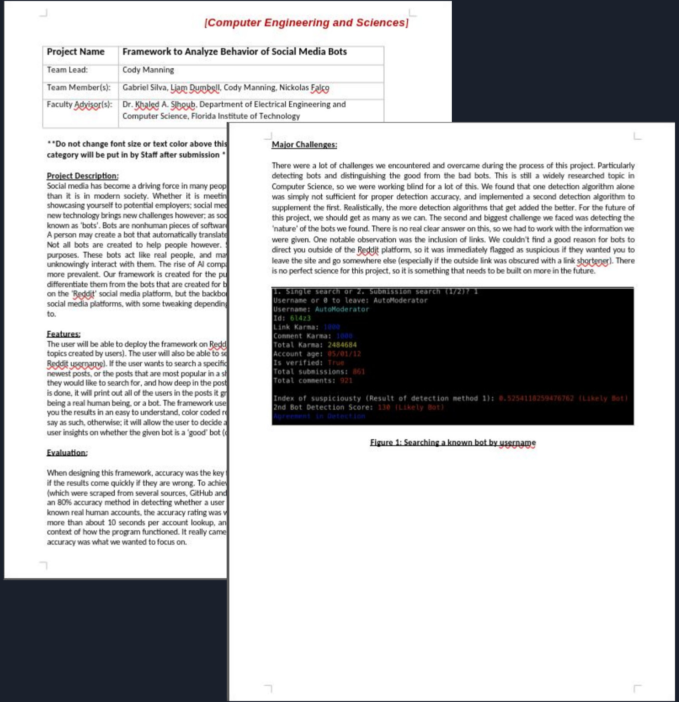


```
human: Constant-Anybody-783
good bot: AutoModerator
bad bot: Angelakayla5
```

```
Initiating further analysis:
[illegible]
AutoModerator has 0 links and 0 shortened links
[illegible]
This account has used profanity 458 times
Is this account a autodeclared bot? [illegible]

[illegible]
Username or 0 to leave: Angelakayla5
Username: Angelakayla5
[illegible]
Link karma: [illegible]
Comment Karma: [illegible]
Total karma: [illegible]
Account age: [illegible]
Is verified: True
Total submissions: [illegible]
Total comments: 0

1st Bot Detection Score: [illegible]
2nd Bot Detection Score: [illegible]
[illegible]

Initiating further analysis:
[illegible]
Angelakayla5 has 0 links and 0 shortened links
[illegible]
This account has used profanity 0 times
Is this account a autodeclared bot? [illegible]

Username or 0 to leave: 0
[illegible]
Found 1 bots out of 2 accounts

account_name='AutoModerator' account_1st_score=[illegible] account_2nd_score=[illegible]
[illegible]
account_name='AutoModerator' account_good_bot=True
account_name='Angelakayla5' account_1st_score=[illegible] account_2nd_score=[illegible]

Name to report? [illegible]
url: [illegible]
[illegible]

C:\Users\[illegible]\Documents\[illegible]>
```

https://youtu.be/LVJUa5dlSaI

EBook Page



[Computer Engineering and Sciences]

Project Name	Framework to Analyze Behavior of Social Media Bots
Team Lead:	Cody Manning
Team Member(s):	Gabriel Silva, Liam Dumbell, Cody Manning, Nickolas Falco
Faculty Advisor(s):	Dr. Khaled A. Slhoub, Department of Electrical Engineering and Computer Science, Florida Institute of Technology

****Do not change font size or text color above this
category will be put in by Staff after submission ***

Project Description:
Social media has become a driving force in many peop
than it is in modern society. Whether it is meetin
showcasing yourself to potential employers; social med
new technology brings new challenges however; as soc
known as 'bots'. Bots are nonhuman pieces of software
A person may create a bot that automatically translate
Not all bots are created to help people however.
purposes. These bots act like real people, and ma
unknowingly interact with them. The rise of AI compa
more prevalent. Our framework is created for the pu
differentiate them from the bots that are created for b
on the 'Reddit' social media platform, but the backbo
social media platforms, with some tweaking dependin
to.

Features:
The user will be able to deploy the framework on Redd
topics created by users). The user will also be able to se
Reddit username). If the user wants to search a specific
newest posts, or the posts that are most popular in a sh
they would like to search for, and how deep in the post
is done, it will print out all of the users in the posts it gr
being a real human being, or a bot. The framework use
you the results in an easy to understand, color coded re
say as such, otherwise; it will allow the user to decide a
user insights on whether the given bot is a 'good' bot (

Evaluation:

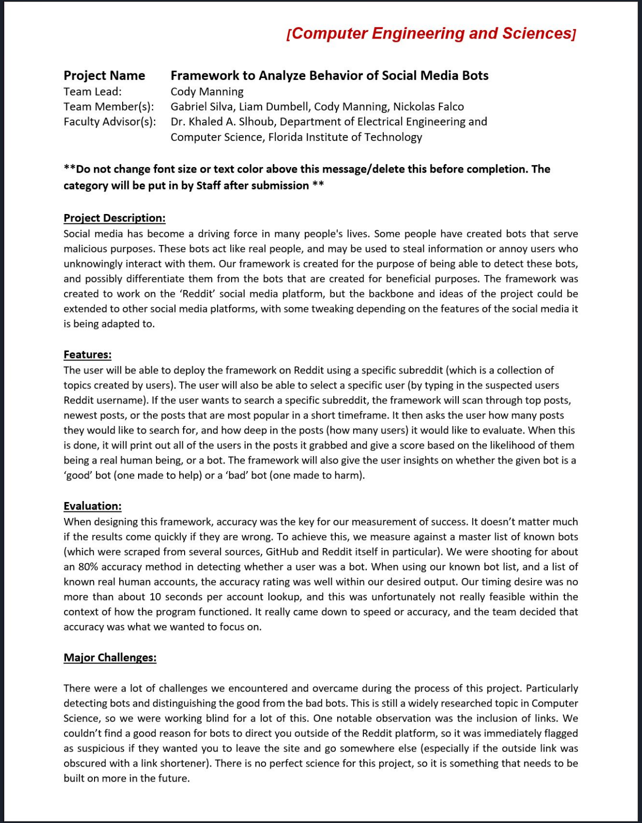
When designing this framework, accuracy was the key
if the results come quickly if they are wrong. To achie
(which were scraped from several sources, GitHub and
an 80% accuracy method in detecting whether a user
known real human accounts, the accuracy rating was
more than about 10 seconds per account lookup, an
context of how the program functioned. It really came
accuracy was what we wanted to focus on.

Major Challenges:

There were a lot of challenges we encountered and overcame during the process of this project. Particularly detecting bots and distinguishing the good from the bad bots. This is still a widely researched topic in Computer Science, so we were working blind for a lot of this. We found that one detection algorithm alone was simply not sufficient for proper detection accuracy, and implemented a second detection algorithm to supplement the first. Realistically, the more detection algorithms that get added the better. For the future of this project, we should get as many as we can. The second and biggest challenge we faced was detecting the 'nature' of the bots we found. There is no real clear answer on this, so we had to work with the information we were given. One notable observation was the inclusion of links. We couldn't find a good reason for bots to direct you outside of the Reddit platform, so it was immediately flagged as suspicious if they wanted you to leave the site and go somewhere else (especially if the outside link was obscured with a link shortener). There is no perfect science for this project, so it is something that needs to be built on more in the future.

```
1. Single search or 2. Submission search (1/2)? 1
Username or 0 to leave: AutoModerator
Username: AutoModerator
Id: 6l4z3
Link Karma: 1000
Comment Karma: 1000
Total Karma: 2484684
Account age: 05/01/12
Is verified: True
Total submissions: 861
Total comments: 921

Index of suspiciousty (Result of detection method 1): 0.5254118259476762 (Likely Bot)
2nd Bot Detection Score: 130 (Likely Bot)
Agreement in Detection
```

Figure 1: Searching a known bot by username



[Computer Engineering and Sciences]

Project Name **Framework to Analyze Behavior of Social Media Bots**
Team Lead: Cody Manning
Team Member(s): Gabriel Silva, Liam Dumbell, Cody Manning, Nickolas Falco
Faculty Advisor(s): Dr. Khaled A. Slhoub, Department of Electrical Engineering and Computer Science, Florida Institute of Technology

****Do not change font size or text color above this message/delete this before completion. The category will be put in by Staff after submission ****

Project Description:
Social media has become a driving force in many people's lives. Some people have created bots that serve malicious purposes. These bots act like real people, and may be used to steal information or annoy users who unknowingly interact with them. Our framework is created for the purpose of being able to detect these bots, and possibly differentiate them from the bots that are created for beneficial purposes. The framework was created to work on the 'Reddit' social media platform, but the backbone and ideas of the project could be extended to other social media platforms, with some tweaking depending on the features of the social media it is being adapted to.

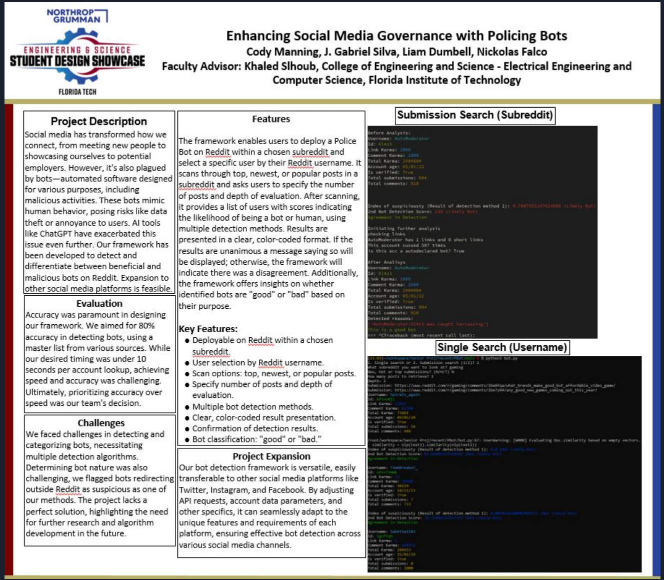
Features:
The user will be able to deploy the framework on Reddit using a specific subreddit (which is a collection of topics created by users). The user will also be able to select a specific user (by typing in the suspected users Reddit username). If the user wants to search a specific subreddit, the framework will scan through top posts, newest posts, or the posts that are most popular in a short timeframe. It then asks the user how many posts they would like to search for, and how deep in the posts (how many users) it would like to evaluate. When this is done, it will print out all of the users in the posts it grabbed and give a score based on the likelihood of them being a real human being, or a bot. The framework will also give the user insights on whether the given bot is a 'good' bot (one made to help) or a 'bad' bot (one made to harm).

Evaluation:
When designing this framework, accuracy was the key for our measurement of success. It doesn't matter much if the results come quickly if they are wrong. To achieve this, we measure against a master list of known bots (which were scraped from several sources, GitHub and Reddit itself in particular). We were shooting for about an 80% accuracy method in detecting whether a user was a bot. When using our known bot list, and a list of known real human accounts, the accuracy rating was well within our desired output. Our timing desire was no more than about 10 seconds per account lookup, and this was unfortunately not really feasible within the context of how the program functioned. It really came down to speed or accuracy, and the team decided that accuracy was what we wanted to focus on.

Major Challenges:

There were a lot of challenges we encountered and overcame during the process of this project. Particularly detecting bots and distinguishing the good from the bad bots. This is still a widely researched topic in Computer Science, so we were working blind for a lot of this. One notable observation was the inclusion of links. We couldn't find a good reason for bots to direct you outside of the Reddit platform, so it was immediately flagged as suspicious if they wanted you to leave the site and go somewhere else (especially if the outside link was obscured with a link shortener). There is no perfect science for this project, so it is something that needs to be built on more in the future.

Poster



NORTHROP GRUMMAN
ENGINEERING & SCIENCE
STUDENT DESIGN SHOWCASE
FLORIDA TECH

Enhancing Social Media Governance with Policing Bots

Cody Manning, J. Gabriel Silva, Liam Dumbell, Nickolas Falco
Faculty Advisor: Khaled Slhoub, College of Engineering and Science - Electrical Engineering and Computer Science, Florida Institute of Technology

Project Description

Social media has transformed how we connect, from meeting new people to showcasing ourselves to potential employers. However, it's also plagued by bots—automated software designed for various purposes, including malicious activities. These bots mimic human behavior, posing risks like data theft or annoyance to users. AI tools like ChatGPT have exacerbated this issue even further. Our framework has been developed to detect and differentiate between beneficial and malicious bots on Reddit. Expansion to other social media platforms is feasible.

Evaluation

Accuracy was paramount in designing our framework. We aimed for 80% accuracy in detecting bots, using a master list from various sources. While our desired timing was under 10 seconds per account lookup, achieving speed and accuracy was challenging. Ultimately, prioritizing accuracy over speed was our team's decision.

Challenges

We faced challenges in detecting and categorizing bots, necessitating multiple detection algorithms. Determining bot nature was also challenging, we flagged bots redirecting outside Reddit as suspicious as one of our methods. The project lacks a perfect solution, highlighting the need for further research and algorithm development in the future.

Features

The framework enables users to deploy a Police Bot on Reddit within a chosen subreddit and select a specific user by their Reddit username. It scans through top, newest, or popular posts in a subreddit and asks users to specify the number of posts and depth of evaluation. After scanning, it provides a list of users with scores indicating the likelihood of being a bot or human, using multiple detection methods. Results are presented in a clear, color-coded format. If the results are unanimous a message saying so will be displayed; otherwise, the framework will indicate there was a disagreement. Additionally, the framework offers insights on whether identified bots are "good" or "bad" based on their purpose.

Key Features:

- Deployable on Reddit within a chosen subreddit.
- User selection by Reddit username.
- Scan options: top, newest, or popular posts.
- Specify number of posts and depth of evaluation.
- Multiple bot detection methods.
- Clear, color-coded result presentation.
- Confirmation of detection results.
- Bot classification: "good" or "bad."

Project Expansion

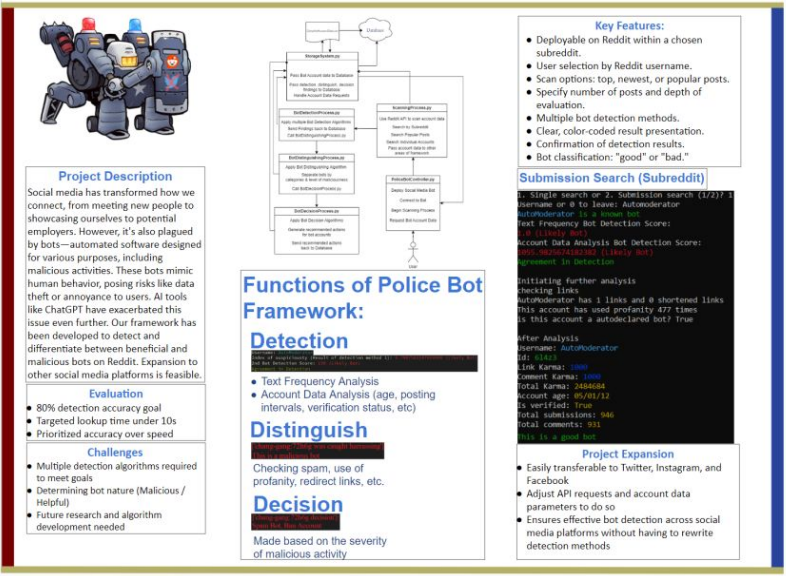
Our bot detection framework is versatile, easily transferable to other social media platforms like Twitter, Instagram, and Facebook. By adjusting API requests, account data parameters, and other specifics, it can seamlessly adapt to the unique features and requirements of each platform, ensuring effective bot detection across various social media channels.

Submission Search (Subreddit)

```
Before Analysis:
Username: AutoModerator
Id: 614z3
Link Karma: 1000
Comment Karma: 1000
Total Karma: 2484684
Account age: 05/01/12
Is verified: True
Total submissions: 904
Total comments: 918



Index of suspiciousty (Result of detection method 1): [illegible] (Likely Bot)
2nd Bot Detection Score: 130 (Likely Bot)
Agreement in Detection

Initiating further analysis
checking links
AutoModerator has 1 links and 0 short links
This account cussed 507 times
is this acc a autodeclared bot? True

After Analisys
Username: AutoModerator
Id: 614z3
Link Karma: 1000
Comment Karma: 1000
Total Karma: 2484684
Account age: 05/01/12
Is verified: True
Total submissions: 904
Total comments: 918
Detected reasons:
['AutoModerator:614z3 was caught harrassing']
This is a good bot
>>> ^CTraceback (most recent call last):
```

Single Search (Username)

```
[21:01]-~/workspace/Senior Proj/recent/PBot:main ✗ $ python3 bot.py
1. Single search or 2. Submission search (1/2)? 2
What subreddit you want to look at? gaming
New, Hot or top submissions? (N/H/T) N
How many posts to retrieve? 2
Depth: 2
Submission: https://www.reddit.com/r/gaming/comments/1be0fqa/what_brands_make_good_but_affordable_video_game/
Submission: https://www.reddit.com/r/gaming/comments/1be7y94/any_good_new_games_coming_out_this_year/
Username: Spirals_again
Id: [illegible]
Link Karma: [illegible]
Comment Karma: [illegible]
Total Karma: 75683
Account age: 09/05/20
Is verified: True
Total submissions: 58
Total comments: 986

/root/workspace/Senior Proj/recent/PBot/bot.py:67: UserWarning: [W008] Evaluating Doc.similarity based on empty vectors.
  similarity = nlp(text1).similarity(nlp(text2))
Index of suspiciousty (Result of detection method 1): 0.0 (Not Likely Bot)
2nd Bot Detection Score: [illegible] (Not Likely Bot)
Agreement in Detection

Username: TimeBreaker_
Id: [illegible]
Link Karma: 11
Comment Karma: [illegible]
Total Karma: 10229
Account age: 10/11/23
Is verified: True
Total submissions: 7
Total comments: 733

Index of suspiciousty (Result of detection method 1): [illegible] (Not Likely Bot)
2nd Bot Detection Score: [illegible] (Not Likely Bot)
Agreement in Detection

Username: Sabetha1183
Id: [illegible]
Link Karma: 1
Comment Karma: [illegible]
Total Karma: 284433
Account age: 25/02/19
Is verified: True
Total submissions: 0
Total comments: 1000
```



Project Description

Social media has transformed how we connect, from meeting new people to showcasing ourselves to potential employers. However, it's also plagued by bots—automated software designed for various purposes, including malicious activities. These bots mimic human behavior, posing risks like data theft or annoyance to users. AI tools like ChatGPT have exacerbated this issue even further. Our framework has been developed to detect and differentiate between beneficial and malicious bots on Reddit. Expansion to other social media platforms is feasible.

Evaluation

- 80% detection accuracy goal
- Targeted lookup time under 10s
- Prioritized accuracy over speed

Challenges

- Multiple detection algorithms required to meet goals
- Determining bot nature (Malicious / Helpful)
- Future research and algorithm development needed

Database
StorageSystem.py
Pass Bot Account data to Database
Pass detection, distinguish, decision findings to Database
Handle Account Data Requests
ScanningProcess.py
Use Reddit API to scan account data
Search by Subreddit
Search Popular Posts
Search Individual Accounts
Pass account data to other areas of framework
BotDetectionProcess.py
Apply multiple Bot Detection Algorithms
Send Findings back to Database
Call BotDistinguishingProcess.py
BotDistinguishingProcess.py
Apply Bot Distinguishing Algorithm
Separate bots by categories & level of maliciousness
Call BotDecisionProcess.py
BotDecisionProcess.py
Apply Bot Decision Algorithms
Generate recommended actions for bot accounts
Send recommended actions back to Database
PoliceBotController.py
Deploy Social Media Bot
Connect to Bot
Begin Scanning Process
Request Bot Account Data
User

Functions of Police Bot Framework:

Detection

```
Username: AutoModerator
Index of suspiciousty (Result of detection method 1): [illegible] (Likely Bot)
2nd Bot Detection Score: 130 (Likely Bot)
Agreement in Detection
```

- Text Frequency Analysis
- Account Data Analysis (age, posting intervals, verification status, etc)

Distinguish

```
['chang-gang:72bfg was caught harrassing']
This is a malicious bot
```

Checking spam, use of profanity, redirect links, etc.

Decision

```
['chang-gang:72bfg decision']
Spam Bot, Ban Account
```

Made based on the severity of malicious activity

Key Features:

- Deployable on Reddit within a chosen subreddit.
- User selection by Reddit username.
- Scan options: top, newest, or popular posts.
- Specify number of posts and depth of evaluation.
- Multiple bot detection methods.
- Clear, color-coded result presentation.
- Confirmation of detection results.
- Bot classification: "good" or "bad."

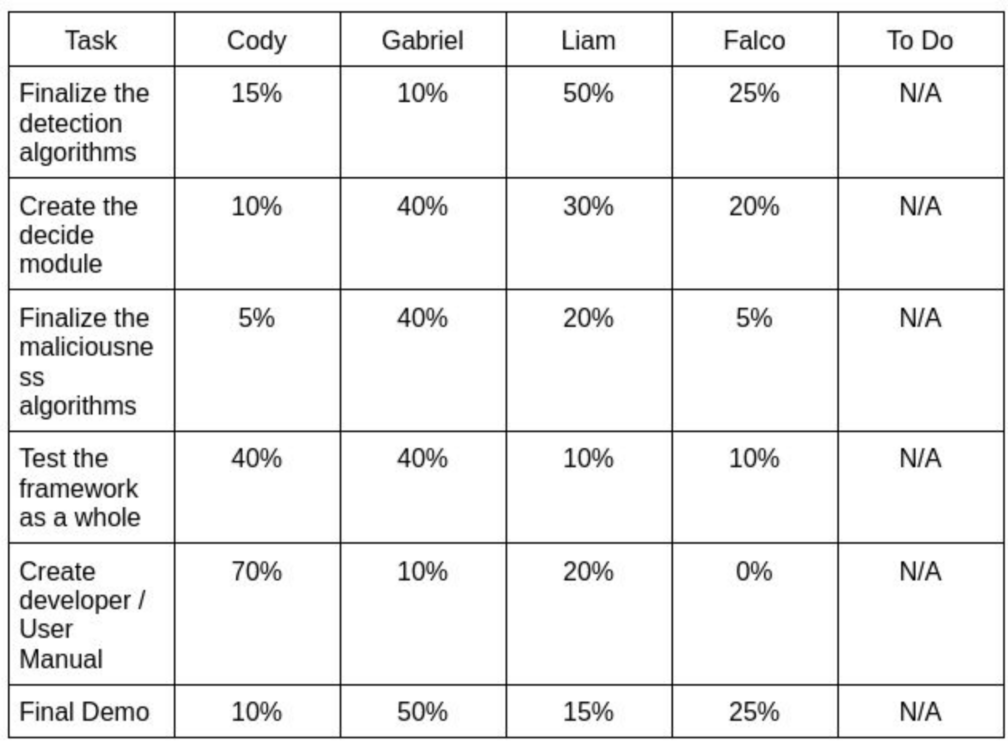
Submission Search (Subreddit)

```
1. Single search or 2. Submission search (1/2)? 1
Username or 0 to leave: Automoderator
AutoModerator is a known bot
Text Frequency Bot Detection Score:
1.0 (Likely Bot)
Account Data Analysis Bot Detection Score:
1055.9825674182382 (Likely Bot)
Agreement in Detection

Initiating further analysis
checking links
AutoModerator has 1 links and 0 shortened links
This account has used profanity 477 times
is this account a autodeclared bot? True

After Analysis
Username: AutoModerator
Id: 614z3
Link Karma: 1000
Comment Karma: 1000
Total Karma: 2484684
Account age: 05/01/12
Is verified: True
Total submissions: 946
Total comments: 931
This is a good bot
```

Project Expansion

- Easily transferable to Twitter, Instagram, and Facebook
- Adjust API requests and account data parameters to do so
- Ensures effective bot detection across social media platforms without having to rewrite detection methods

Milestone Completion Task Matrix



Task	Cody	Gabriel	Liam	Falco	To Do
Finalize the detection algorithms	15%	10%	50%	25%	N/A
Create the decide module	10%	40%	30%	20%	N/A
Finalize the maliciousness algorithms	5%	40%	20%	5%	N/A
Test the framework as a whole	40%	40%	10%	10%	N/A
Create developer / User Manual	70%	10%	20%	0%	N/A
Final Demo	10%	50%	15%	25%	N/A

Advisor Feedback

- Satisfied with our current progress.
    - As aforementioned in early presentations the project's goal was to build from the ground up a framework that polices a social media looking for bots
- Invited us to continue working on this project even after we are done with the class.
- Mentioned writing and publishing a paper with our findings,
- Or making the project available to students in the future.

Lessons Learned

This project has been quite an ordeal, as aforementioned in early presentations the project's goal was to build from the base up a framework that can be improved in the future.

- Have a clear plan, and stick to it
- Understand API and Platform Limitations
- Start early
- Test early

This concludes our presentation, Thank You