BotHawk: An Ensemble Learning-Based for Bot Detection in Open Source Software Projects

Fenglin Bi

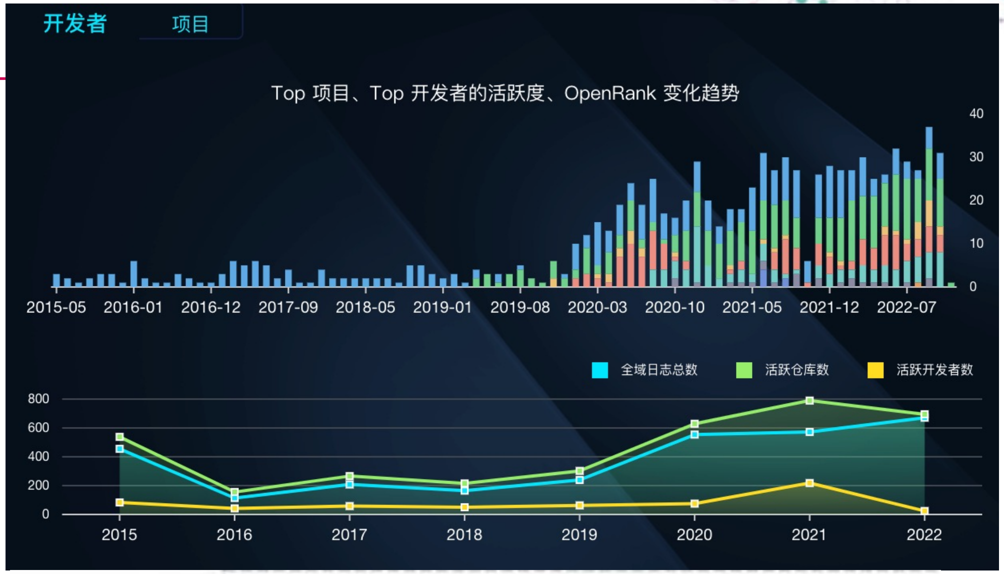
Ecnu

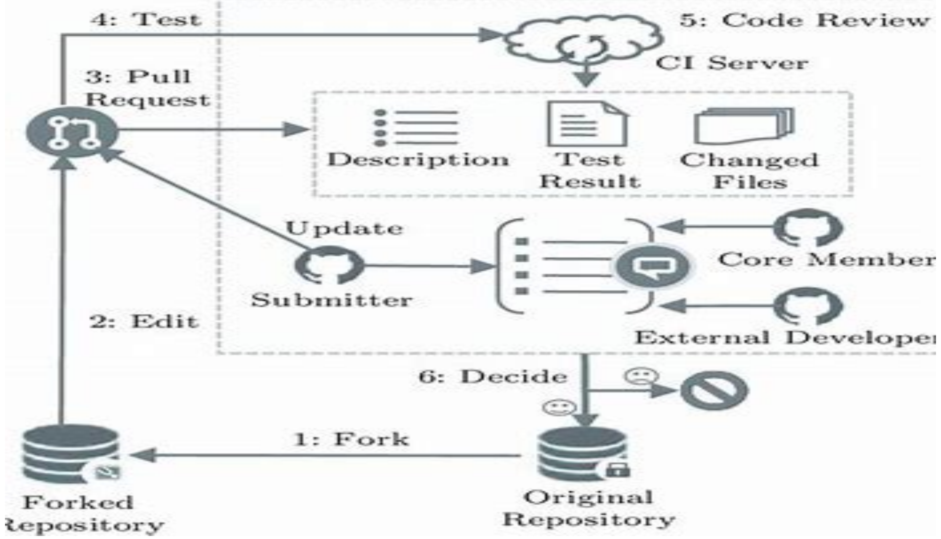
X-lab

INTRODUCTION

- **Open-source Software Collaboration：**
  - Share knowledge
  - Identify and fix bugs
  - Deliver promptly

- **Workload Increased Massively ：**
  - Manage communication with contributors
  - Review source code
  - Handle contributor license
  - Discuss issues
  - Explain project guidelines
  - Run tests and build code
  - Merge pull requests





INTRODUCTION

- **What is Open-Source Software Bots?**
  - Serving various roles in social coding platforms, are crucial in automating tasks and facilitating interactions.[1]
  - A task-oriented bot responsible for automating well-defined tasks on GitHub repositories. A GitHub bot behaves similarly to a human user, serving as an interface between users and services.[2]
  - e.g., Googlebot(ensuring license agreement signing),pdf.js test(running automated tests)

1. Schueller W, Wachs J, Servedio V D P, et al. Evolving collaboration, dependencies, and use in the rust open source software ecosystem[J]. Scientific Data, 2022, 9(1): 703.
2. Wessel M, De Souza B M, Steinmacher I, et al. The power of bots: Characterizing and understanding bots in oss projects[J]. Proceedings of the ACM on Human-Computer Interaction, 2018, 2(CSCW): 1-19.

INTRODUCTION

- **Open-source Software Bots Problem :**
  - Impersonation
  - Information overload
  - Bias
  - e.g., Maldeniya investigated the composition and operation of virtual, loosely-knit teams. They excluded the activities of automated accounts.
- **Identifying Bots Challenges :**
  - Trigger through a platform's API or directly on the platform's website
  - Complexity of their functions and dual roles : social characteristics and collaborate in software development
  - Accounts may display characteristics indicative of both automated bot behavior and human behavior.

INTRODUCTION

- **Evaluating OSS Bot Detection Datasets And Models Problems：**
  - Dataset lack of currency:Bodegha dataset original 5000 accounts included, only 2976 could be located via GitHub search. 128 bot accounts
  - Different datasets may lack sufficient evaluations
- **Motivation：**
  - Data Cleaning
  - Expanded Bot Research
  - Platform Maintenance

INTRODUCTION

- **BotHawk: An Approach for Bots Detection in Open-Source Software Projects**
  - A ground truth dataset:19779 rows, 17 features
  - Categorized OSS bots according to their behaviors
  - Ensenble Model:State-of-the-art OSS bot detection methods
  - OSS Bot Detection Tool and Service

- **Solve the Problem：**
  - How to created a standard groud truth dataset for bot detection.
  - What are the categories of behavior patterns for bot accounts
  - How effective is our approach compared to the state-of-the-art?
  - What features are the best indicators of bot accounts detection?

RELATED WORK

- **Taxonomy：**
  - Lebeuf
    - 3 dimensions, 22 aspects.
    - Include the bot's environment, internal properties exhibited, the interaction between the bot and its environment.
    - Problem: their taxonomy is relatively complex for bots in the open-source domain
  - Erlenhov
    - Identified the characteristics of DevBots (robots that support software development) by applying an aspect-based taxonomy.
    - Problem： limited to bots that support software development and does not extend to the entire domain of open-source software robots.
  - Wessel
    - Acquired 351 popular open-source projects and detected 93 of them (26%)
    - Categorized into various functions, such as "Ensuring License Agreement Signing" and "Reporting Continuous Integration Failures."
    - Problem: their classification method is less useful for identifying bots using automated tools

RELATED WORK

- **Datasets for Bot Detection and Feature extraction：**
  - Golzadeh
    - 36K software package registries
    - 5,000 GitHub accounts, with 4,473 pertaining to human accounts and 527 to robot accounts
    - Problem: their features are limited, primarily using comment data from issues.
  - Zhao - BIMAN
    - □ 461 robot accounts and 13,762,430submissions.
    - □ includes submission metadata, account names, and email addresses
    - □ Problem： the account login names for Github accounts are absent from the dataset and they lack time-series-related features.

- Other datasets
  - □ BotHunter: An Approach to Detect Software Bots in GitHub
  - □ Effects of Adopting Code Review Bots on Pull Requests to OSS Projects
  - □ Problem: not have public tool or model

RELATED WORK

- **Algorithms for bot detection：**
  - BIMAN： studied three machine learning classifiers to recognize commit profile and commit comments submitted.
  - BoDeGHa： a machine learning-based approach that identifies software robots posting comments on issues and pull requests on GitHub by analyzing comment-related features like repetitive comment patterns.
  - BotHunter： a machine learning-based method to distinguish robot accounts based on 19 pre-selected features.

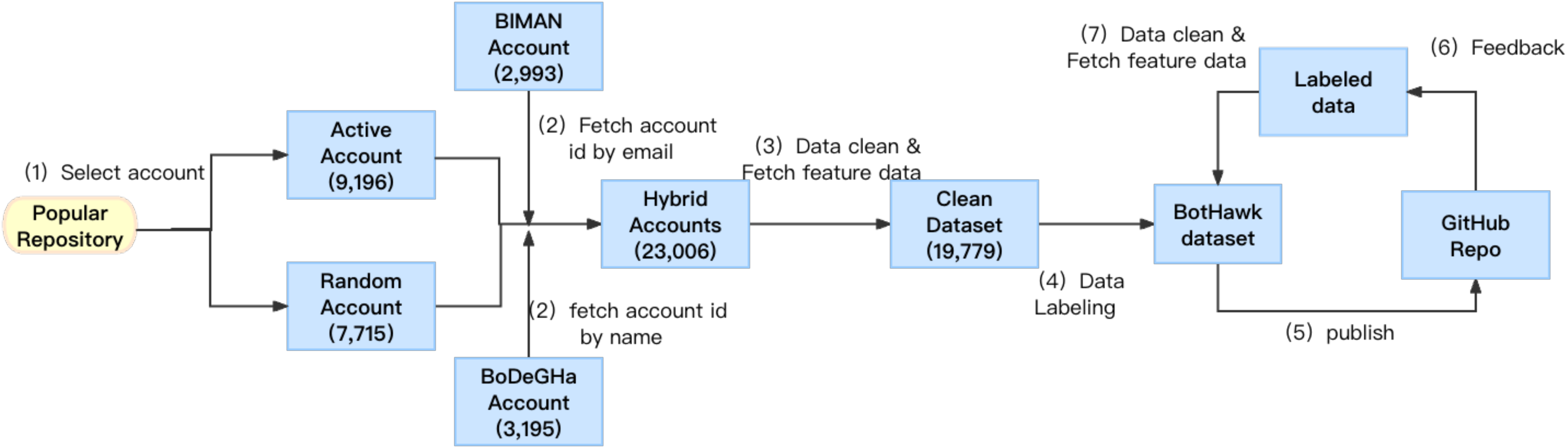
Rusult:

model	F1-score	AUC
BoDeGHa	98%	
BotHunter	92.4%	98.7%
BIMAN		90%

Bot detection is challenging?

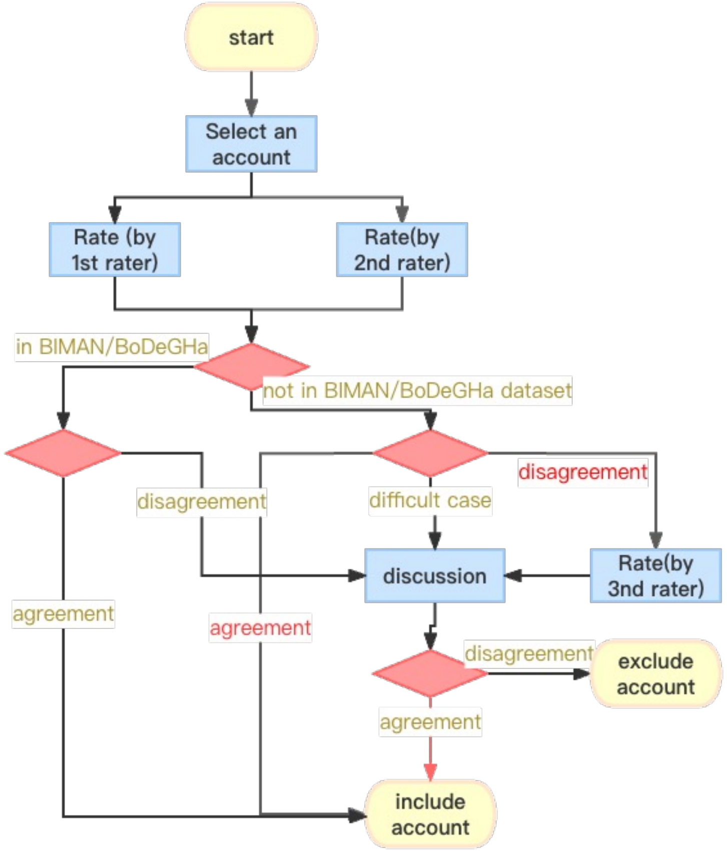
GROUND TRUTH DATASET

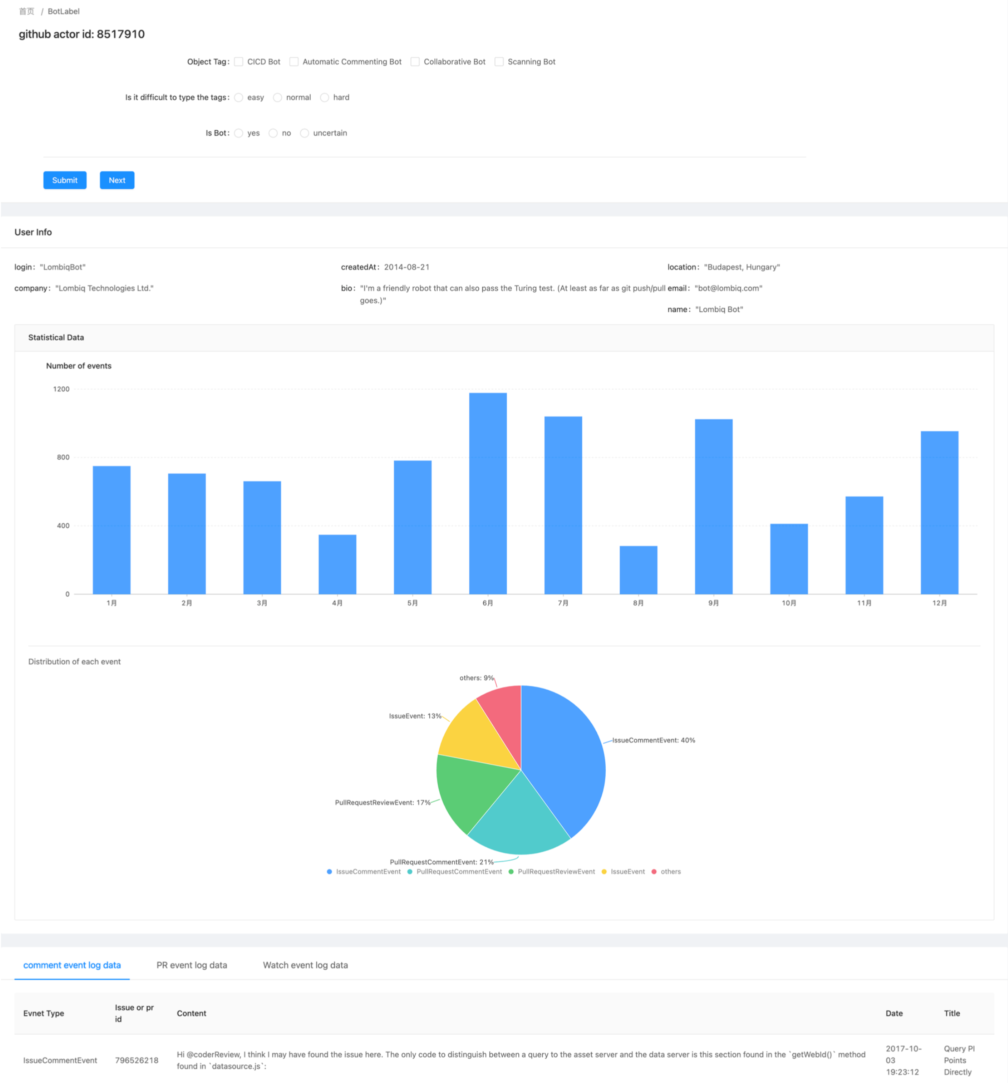
- **Criterion of Datasets for Bot Detection ：**
  - Generalization ability : 4 Dataset. 17 Relevant features
  - Data extendibility: Seamlessly incorporate new data by GitHub
  - Timeliness： Update the labeled data from Open-digger repo



GROUND TRUTH DATASET

- **Criterion of Datasets for Bot Detection ：**
  - Accuracy: Labeling processes, Kappa sore 0.871





github actor id: 8517910

Object Tag： □ CICD Bot □ Automatic Commenting Bot □ Collaborative Bot □ Scanning Bot

Is it difficult to type the tags： ○ easy ○ normal ○ hard

Is Bot： ○ yes ○ no ○ uncertain

User Info

login： "LombiqBot"

createdAt： 2014-08-21

location： "Budapest, Hungary"

company： "Lombiq Technologies Ltd."

bio： "I'm a friendly robot that can also pass the Turing test. (At least as far as git push/pull goes.)"

email： "bot@lombiq.com"

name： "Lombiq Bot"

Statistical Data

Number of events

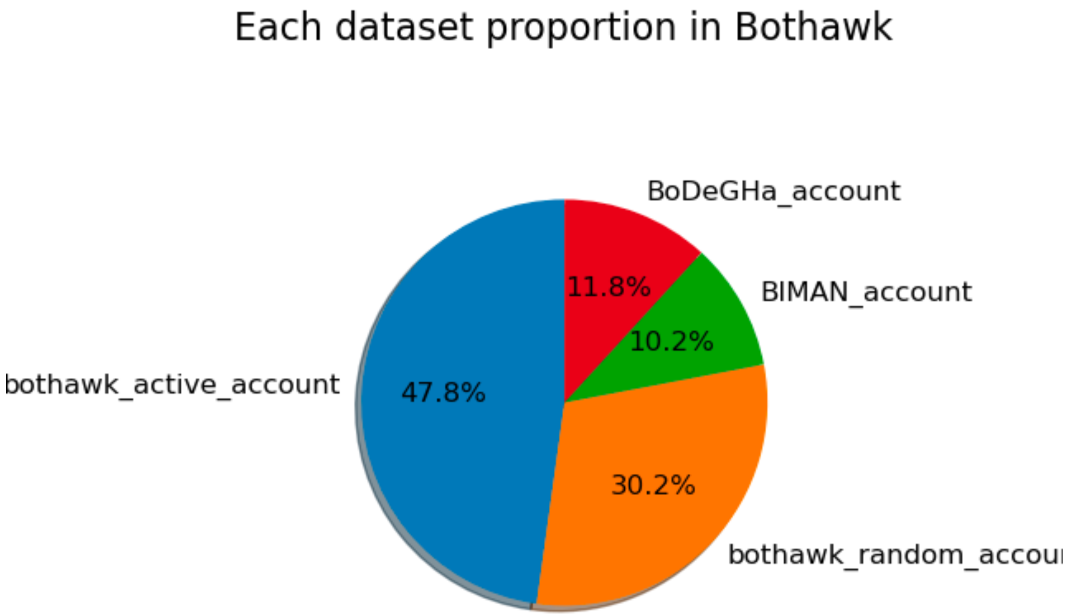
Distribution of each event

others: 9%
IssueEvent: 13%
IssueCommentEvent: 40%
PullRequestReviewEvent: 17%
PullRequestCommentEvent: 21%
● IssueCommentEvent ● PullRequestCommentEvent ● PullRequestReviewEvent ● IssueEvent ● others

Evnet Type	Issue or pr id	Content	Date	Title
IssueCommentEvent	796526218	Hi @coderReview, I think I may have found the issue here. The only code to distinguish between a query to the asset server and the data server is this section found in the `getWebId()` method found in `datasource.js`:	2017-10-03 19:23:12	Query PI Points Directly

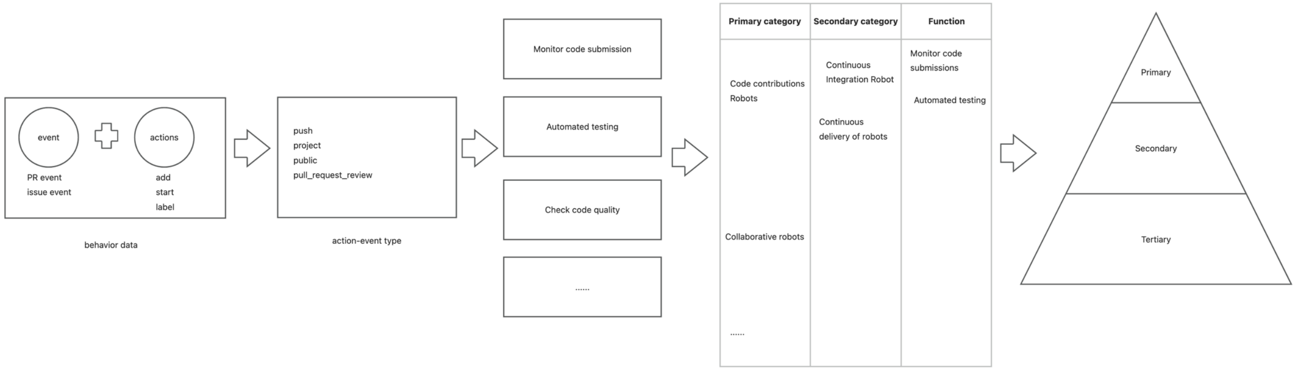
GROUND TRUTH DATASET

- **BotHawk Dataset：**
  - 19,779 rows
  - Bot label 756 rows
  - 17 features

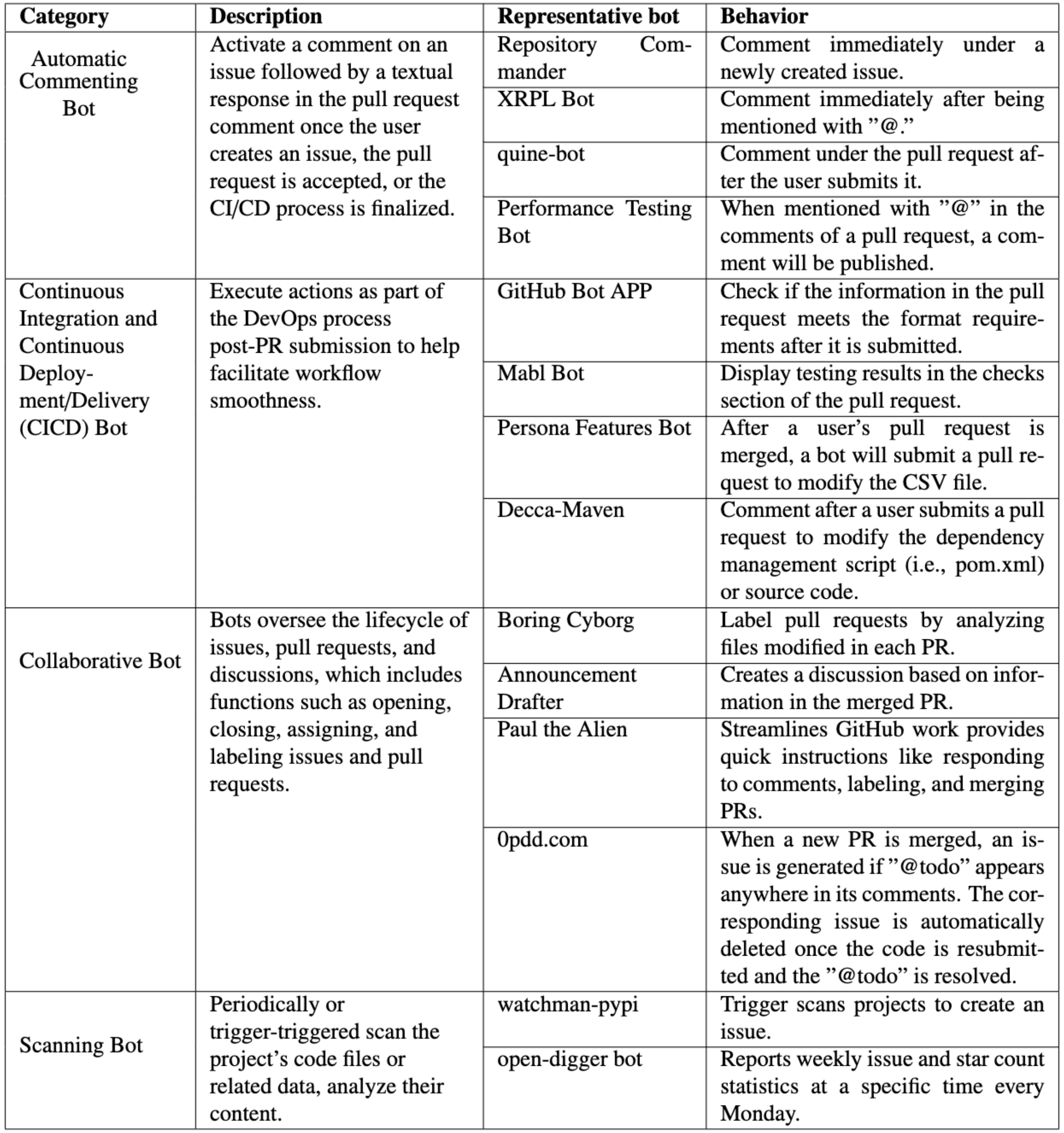


FEATURE SELECTION

- **Bot Behavior Activity Analysis：**
  - 721 GitHub Apps on the GitHub Marketplace as of June 2023
  - Behavioral encoding
  - Expert validation
- **OSS Bot Taxonomy：**
  - 754 Bot Account + 721 GitHub Apps

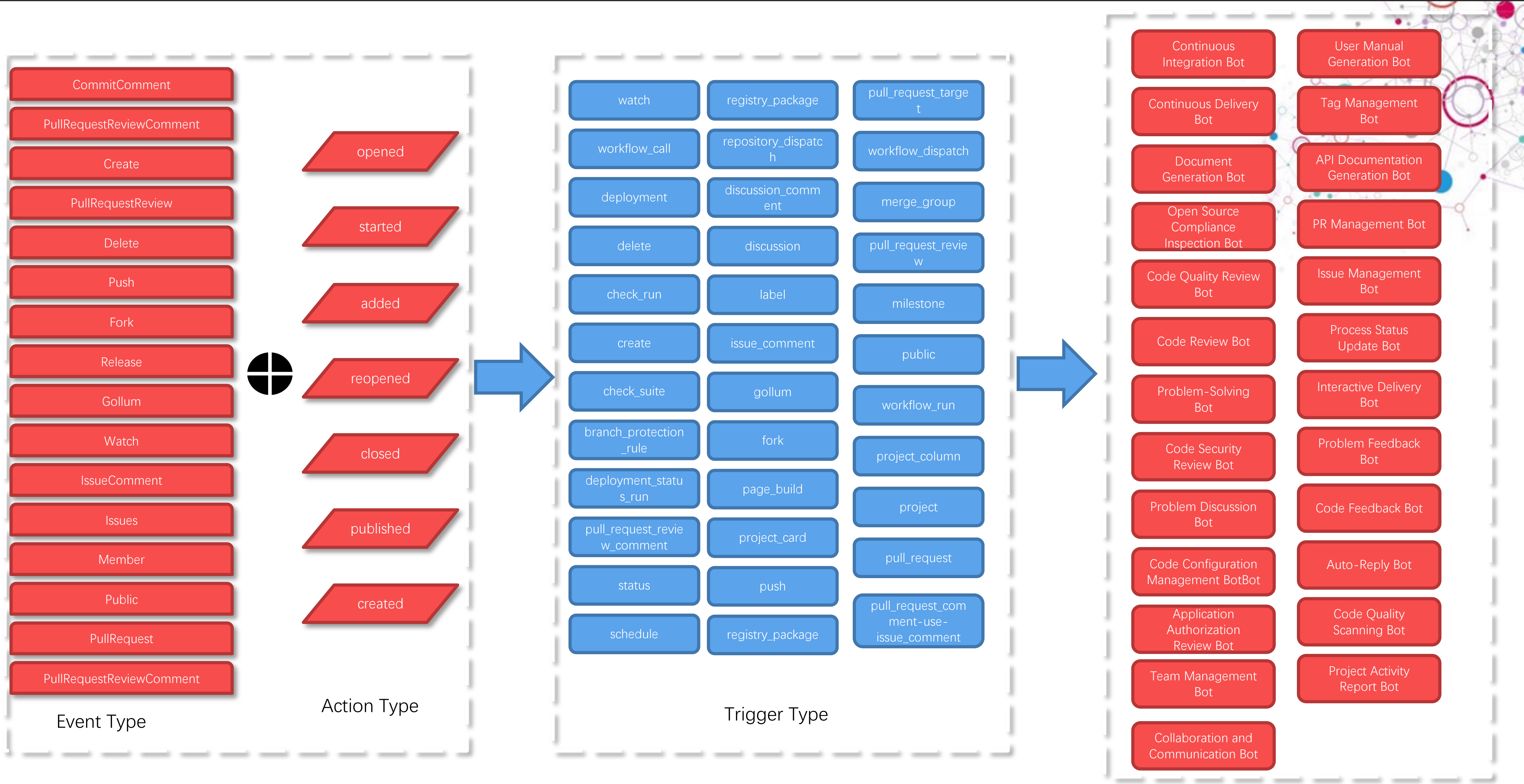


Primary category	Secondary category	Function
Code contributions Robots	Continuous Integration Robot	Monitor code submissions Automated testing
	Continuous delivery of robots	
Collaborative robots		
......		



Category	Description	Representative bot	Behavior
Automatic Commenting Bot	Activate a comment on an issue followed by a textual response in the pull request comment once the user creates an issue, the pull request is accepted, or the CI/CD process is finalized.	Repository Commander	Comment immediately under a newly created issue.
		XRPL Bot	Comment immediately after being mentioned with ”@.”
		quine-bot	Comment under the pull request after the user submits it.
		Performance Testing Bot	When mentioned with ”@” in the comments of a pull request, a comment will be published.
Continuous Integration and Continuous Deployment/Delivery (CICD) Bot	Execute actions as part of the DevOps process post-PR submission to help facilitate workflow smoothness.	GitHub Bot APP	Check if the information in the pull request meets the format requirements after it is submitted.
		Mabl Bot	Display testing results in the checks section of the pull request.
		Persona Features Bot	After a user's pull request is merged, a bot will submit a pull request to modify the CSV file.
		Decca-Maven	Comment after a user submits a pull request to modify the dependency management script (i.e., pom.xml) or source code.
Collaborative Bot	Bots oversee the lifecycle of issues, pull requests, and discussions, which includes functions such as opening, closing, assigning, and labeling issues and pull requests.	Boring Cyborg	Label pull requests by analyzing files modified in each PR.
		Announcement Drafter	Creates a discussion based on information in the merged PR.
		Paul the Alien	Streamlines GitHub work provides quick instructions like responding to comments, labeling, and merging PRs.
		0pdd.com	When a new PR is merged, an issue is generated if ”@todo” appears anywhere in its comments. The corresponding issue is automatically deleted once the code is resubmitted and the ”@todo” is resolved.
Scanning Bot	Periodically or trigger-triggered scan the project's code files or related data, analyze their content.	watchman-pypi	Trigger scans projects to create an issue.
		open-digger bot	Reports weekly issue and star count statistics at a specific time every Monday.

Table 1. GitHub Apps behavior Category



CommitComment

PullRequestReviewComment

Create

PullRequestReview

Delete

Push

Fork

Release

Gollum

Watch

IssueComment

Issues

Member

Public

PullRequest

PullRequestReviewComment

Event Type

opened

started

added

reopened

closed

published

created

Action Type

watch

registry_package

pull_request_target

workflow_call

repository_dispatch

workflow_dispatch

deployment

discussion_comment

merge_group

delete

discussion

pull_request_review

check_run

label

milestone

create

issue_comment

public

check_suite

gollum

workflow_run

branch_protection_rule

fork

project_column

deployment_status_run

page_build

project

pull_request_review_comment

project_card

pull_request

status

push

pull_request_comment-use-issue_comment

schedule

registry_package

Trigger Type

Continuous Integration Bot

User Manual Generation Bot

Continuous Delivery Bot

Tag Management Bot

Document Generation Bot

API Documentation Generation Bot

Open Source Compliance Inspection Bot

PR Management Bot

Code Quality Review Bot

Issue Management Bot

Code Review Bot

Process Status Update Bot

Problem-Solving Bot

Interactive Delivery Bot

Code Security Review Bot

Problem Feedback Bot

Problem Discussion Bot

Code Feedback Bot

Code Configuration Management BotBot

Auto-Reply Bot

Application Authorization Review Bot

Code Quality Scanning Bot

Team Management Bot

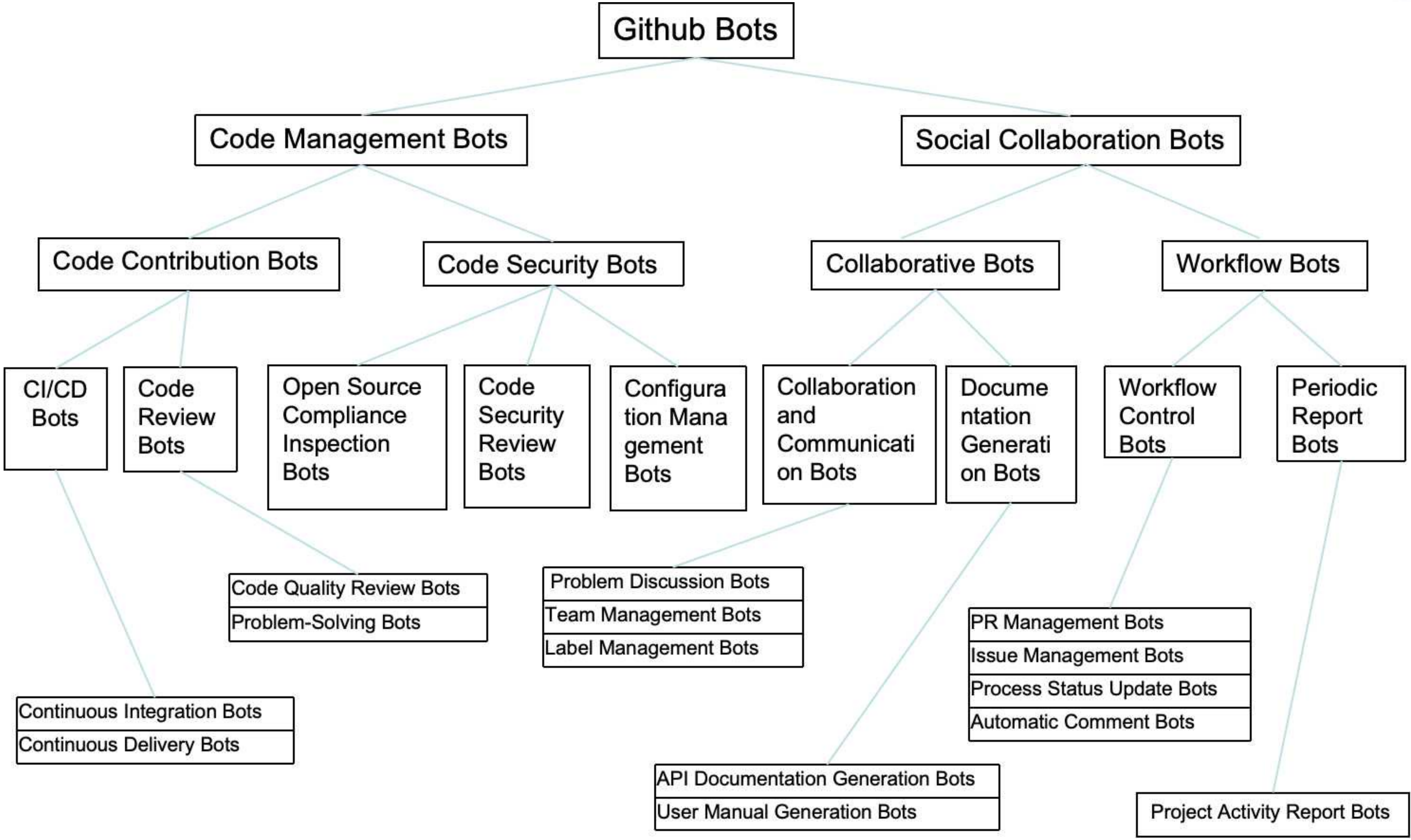
Project Activity Report Bot

Collaboration and Communication Bot

Find Behavior

Encoding to Trigger

Encoding to Bot



Github Bots

Code Management Bots

Social Collaboration Bots

Code Contribution Bots

Code Security Bots

Collaborative Bots

Workflow Bots

CI/CD Bots

Code Review Bots

Open Source Compliance Inspection Bots

Code Security Review Bots

Configuration Management Bots

Collaboration and Communication Bots

Documentation Generation Bots

Workflow Control Bots

Periodic Report Bots

Continuous Integration Bots
Continuous Delivery Bots

Code Quality Review Bots
Problem-Solving Bots

Problem Discussion Bots
Team Management Bots
Label Management Bots

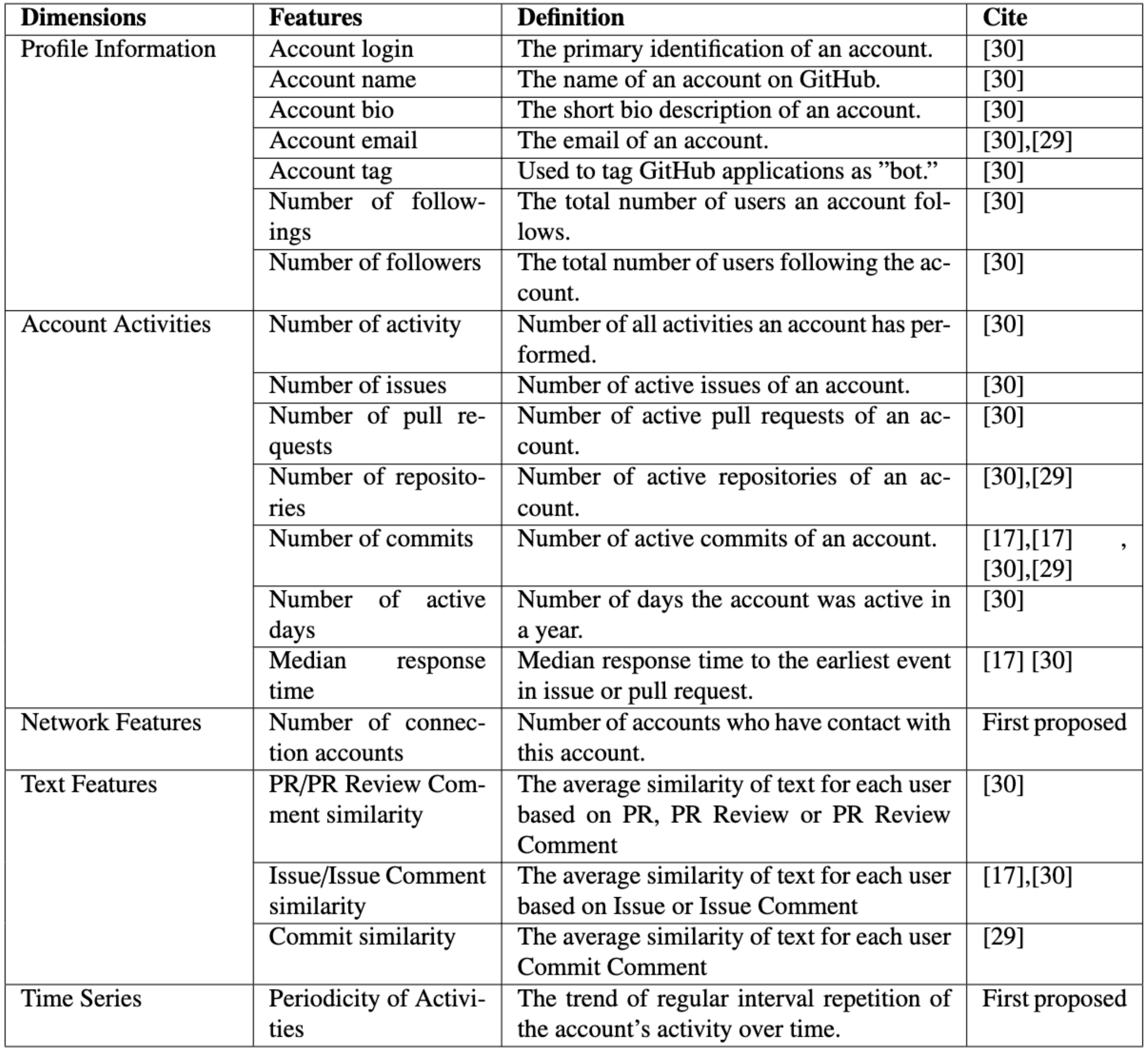
API Documentation Generation Bots
User Manual Generation Bots

PR Management Bots
Issue Management Bots
Process Status Update Bots
Automatic Comment Bots

Project Activity Report Bots

FEATURE SELECTION

- **BotHawk dataset feature：**
    - 17 Features
    - 5 Dimensions



Dimensions	Features	Definition	Cite
Profile Information	Account login	The primary identification of an account.	[30]
	Account name	The name of an account on GitHub.	[30]
	Account bio	The short bio description of an account.	[30]
	Account email	The email of an account.	[30],[29]
	Account tag	Used to tag GitHub applications as ”bot.”	[30]
	Number of followings	The total number of users an account follows.	[30]
	Number of followers	The total number of users following the account.	[30]
Account Activities	Number of activity	Number of all activities an account has performed.	[30]
	Number of issues	Number of active issues of an account.	[30]
	Number of pull requests	Number of active pull requests of an account.	[30]
	Number of repositories	Number of active repositories of an account.	[30],[29]
	Number of commits	Number of active commits of an account.	[17],[17] , [30],[29]
	Number of active days	Number of days the account was active in a year.	[30]
	Median response time	Median response time to the earliest event in issue or pull request.	[17] [30]
Network Features	Number of connection accounts	Number of accounts who have contact with this account.	First proposed
Text Features	PR/PR Review Comment similarity	The average similarity of text for each user based on PR, PR Review or PR Review Comment	[30]
	Issue/Issue Comment similarity	The average similarity of text for each user based on Issue or Issue Comment	[17],[30]
	Commit similarity	The average similarity of text for each user Commit Comment	[29]
Time Series	Periodicity of Activities	The trend of regular interval repetition of the account’s activity over time.	First proposed

Table 2. An overview of features used to identify account type

FEATURE SELECTION

- **BotHawk dataset feature：**
  - Account login、Account name、Account bio、Account email、Account tag

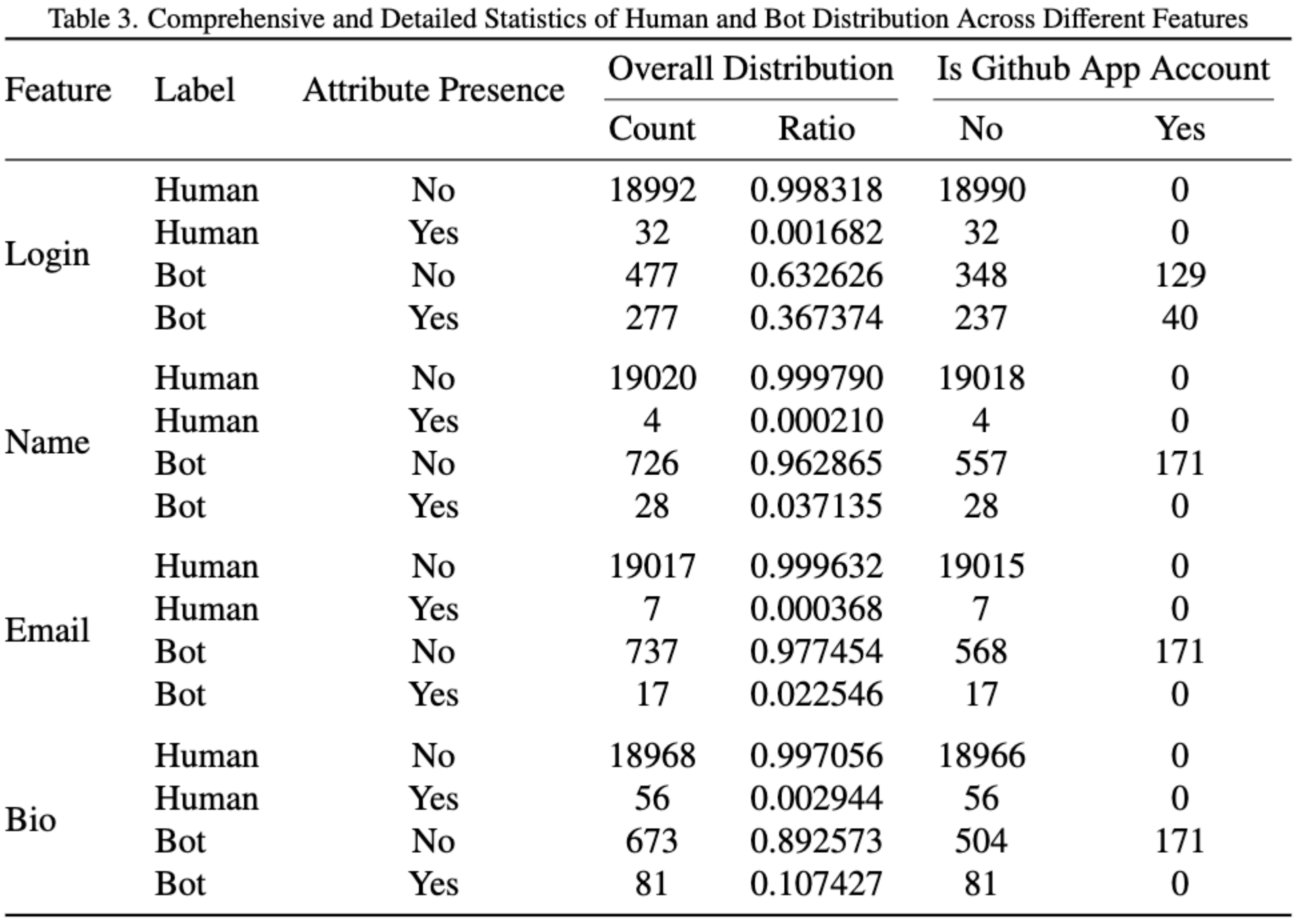


Table 3. Comprehensive and Detailed Statistics of Human and Bot Distribution Across Different Features

Feature	Label	Attribute Presence	Overall Distribution		Is Github App Account	
			Count	Ratio	No	Yes
Login	Human	No	18992	0.998318	18990	0
	Human	Yes	32	0.001682	32	0
	Bot	No	477	0.632626	348	129
	Bot	Yes	277	0.367374	237	40
Name	Human	No	19020	0.999790	19018	0
	Human	Yes	4	0.000210	4	0
	Bot	No	726	0.962865	557	171
	Bot	Yes	28	0.037135	28	0
Email	Human	No	19017	0.999632	19015	0
	Human	Yes	7	0.000368	7	0
	Bot	No	737	0.977454	568	171
	Bot	Yes	17	0.022546	17	0
Bio	Human	No	18968	0.997056	18966	0
	Human	Yes	56	0.002944	56	0
	Bot	No	673	0.892573	504	171
	Bot	Yes	81	0.107427	81	0

$$Feature_{login,name,bio,email} = \begin{cases} 1, & \text{if account contains 'bot', 'auto', 'ci', 'cla', 'io', et.} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

FEATURE SELECTION

- **BotHawk dataset feature：**
  - Number of following、Number of follower
  - Counts of activity、Counts of issue、Counts of pull request、Counts of repository、Counts of commit

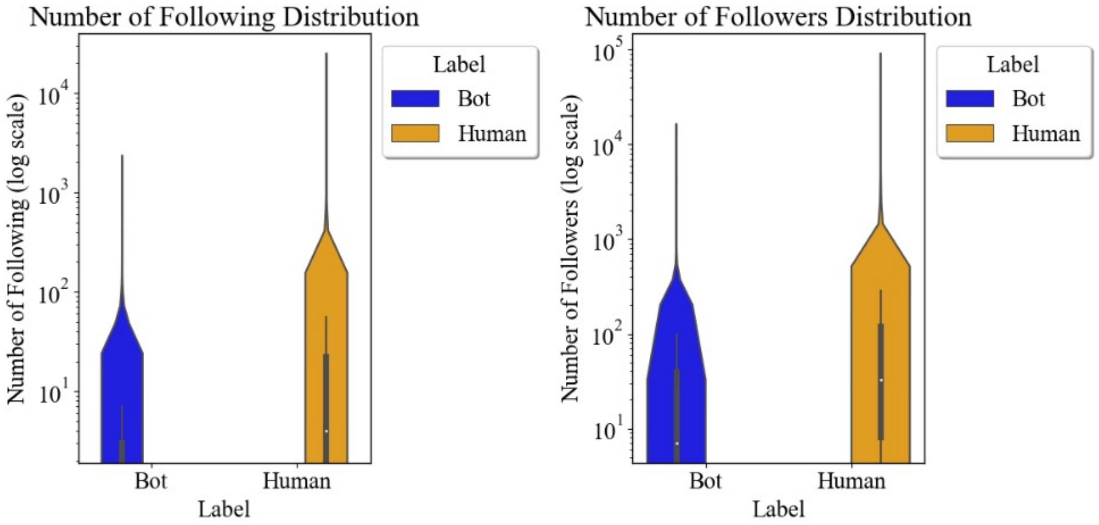


Figure 6. Number of Following and Number of Followers Distribution

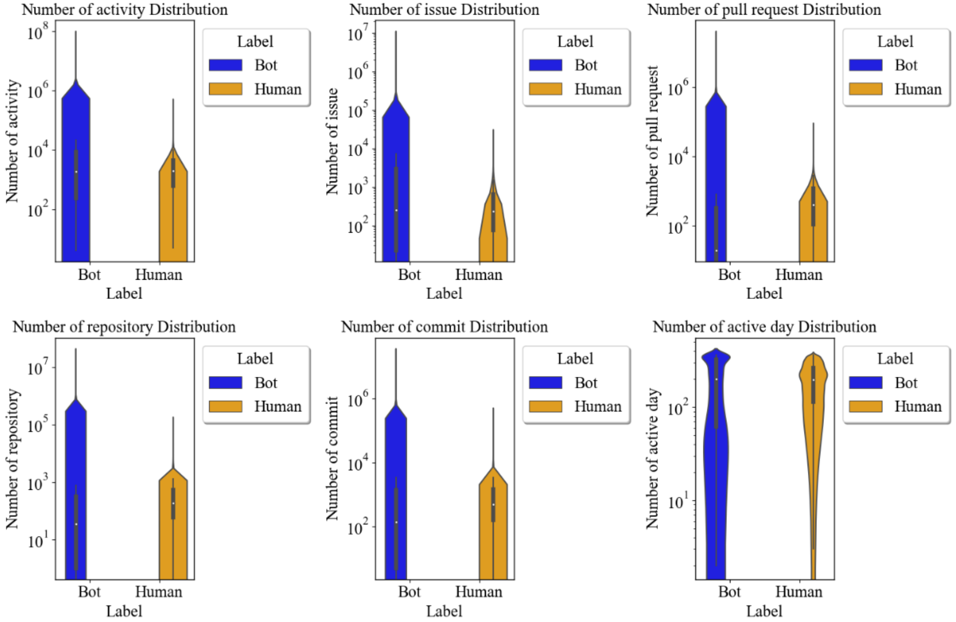
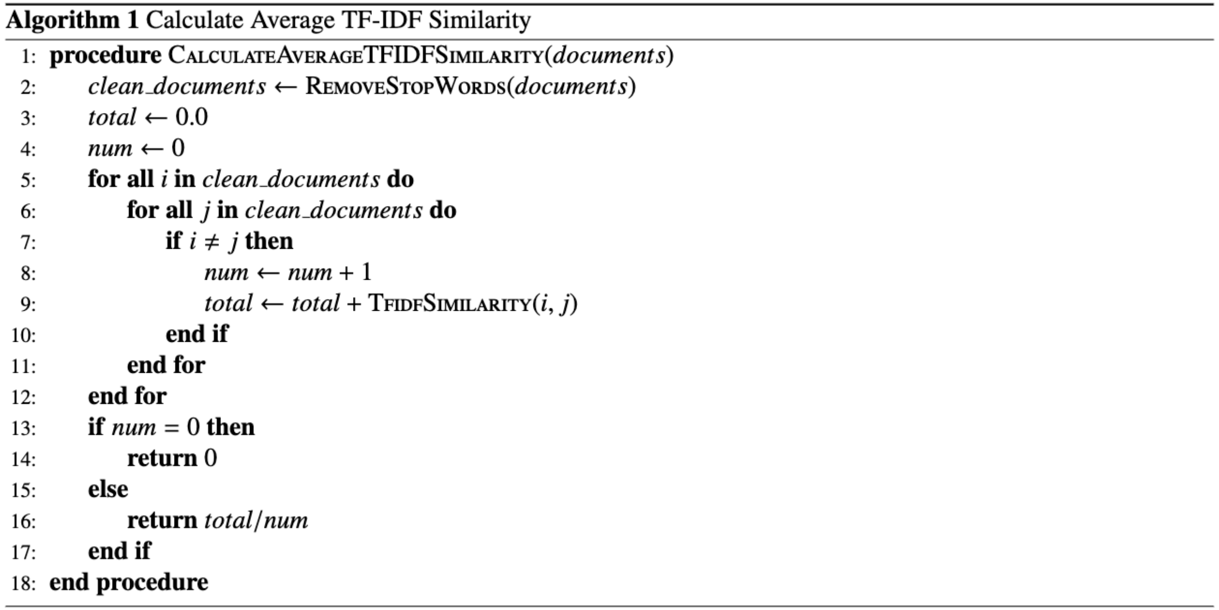


Figure 7. Acticity Issue PR Repository Commit Activity per day Distribution

FEATURE SELECTION

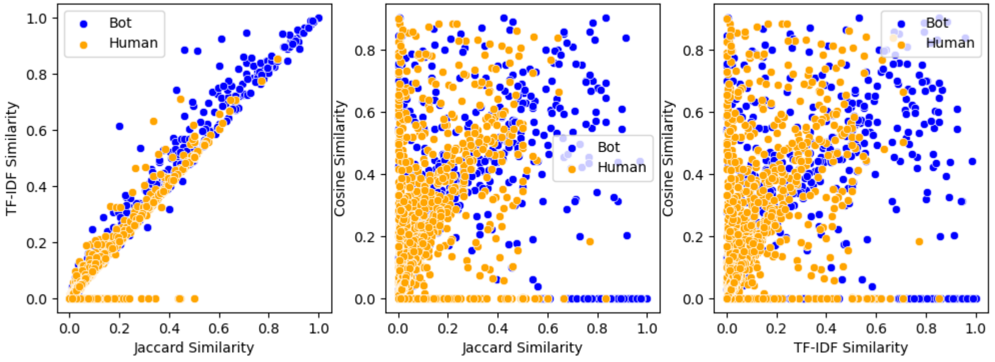
- **BotHawk dataset feature:**
  - Text Similarity
    - □ Jaccard Similarity
    - □ Cosin Similarity
    - □ TF-IDF Similarity



Algorithm 1 Calculate Average TF-IDF Similarity

```
procedure CalculateAverageTFIDFSimilarity(documents)
    clean_documents ← RemoveStopWords(documents)
    total ← 0.0
    num ← 0
    for all i in clean_documents do
        for all j in clean_documents do
            if i ≠ j then
                num ← num + 1
                total ← total + TfidfSimilarity(i, j)
            end if
        end for
    end for
    if num = 0 then
        return 0
    else
        return total/num
    end if
end procedure
```

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \log\left(\frac{N}{df(t)}\right)$$



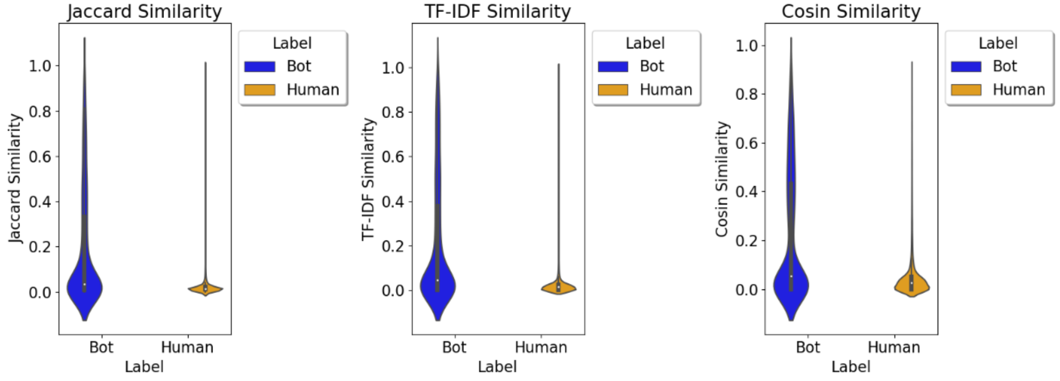


Figure 10. Jaccard, TF-IDF, Cosin Smilarity Distribution

FEATURE SELECTION

- **BotHawk dataset feature：**
  - Counts of connection account:
  - Median response time
  - Periodicity of Activities

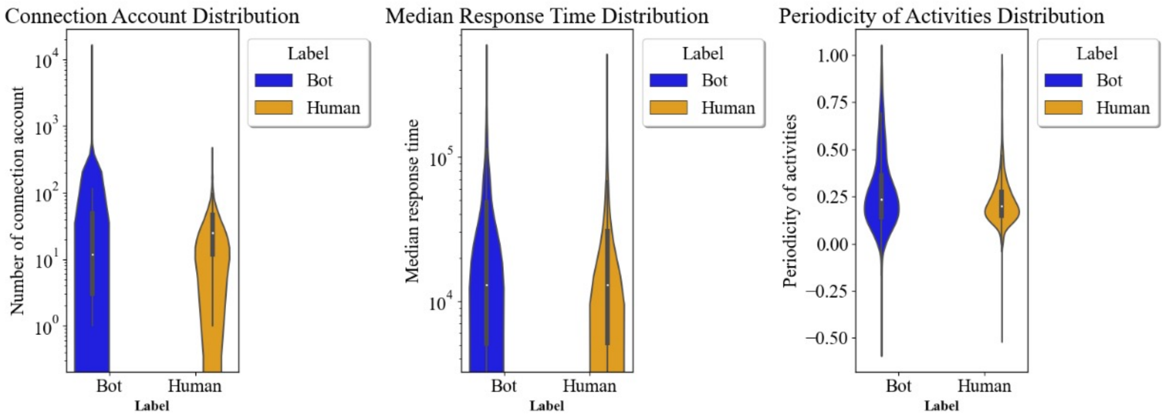


Figure 8. Number of connection account, Median response time Distribution, Periodicity of Activities Distribution

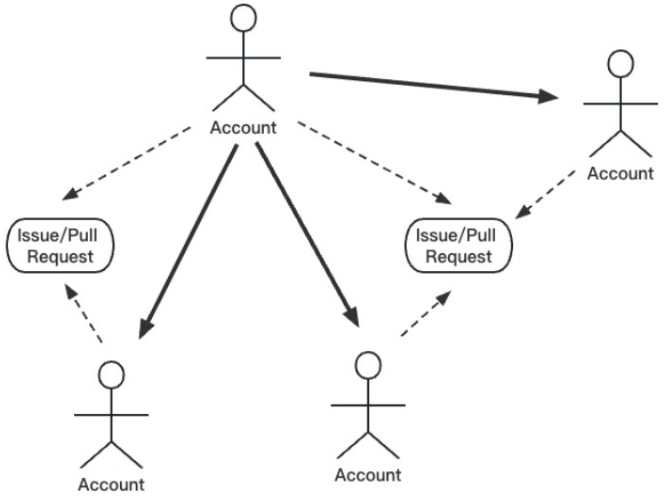


Figure 9. Network of interconnected accounts

$$F(f) = \int^{\infty} X(t)e^{-2\pi ift}dt$$

$$f_{max} = \arg\max_{f} |F(f)|$$

$$T = \frac{1}{f_{max}}$$

CLASSIFICATION MODEL

- **Introduction:**

- A common approach to binary classification is to use a decision function g(X) that maps the feature space to a real number and then applies a threshold T to determine the class label

$$Y_{\text{pred}} = \begin{cases} 1 & \text{if } g(X) \geq T \\ 0 & \text{otherwise} \end{cases}$$

- **Basic Model: Dataset Imbalance**

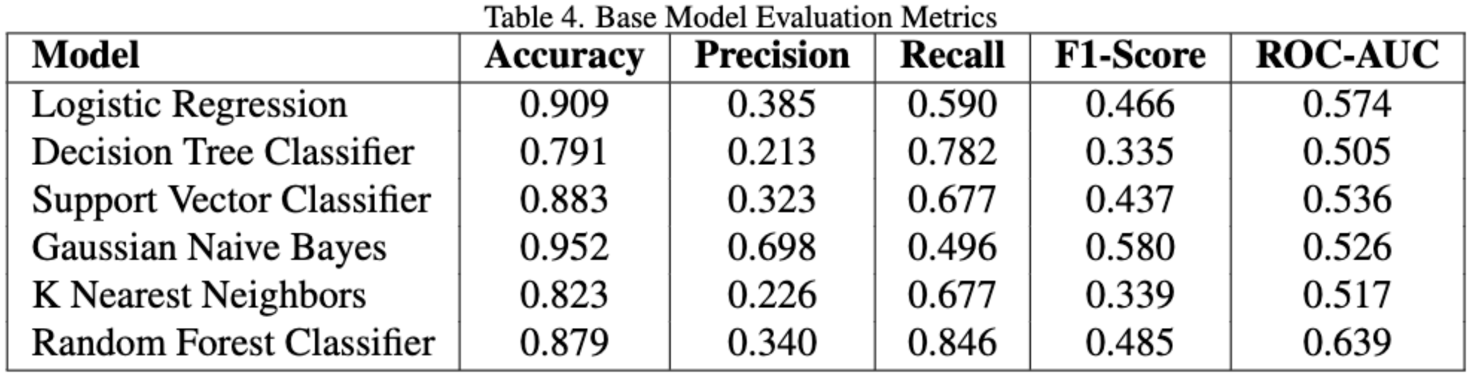


Table 4. Base Model Evaluation Metrics

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.909	0.385	0.590	0.466	0.574
Decision Tree Classifier	0.791	0.213	0.782	0.335	0.505
Support Vector Classifier	0.883	0.323	0.677	0.437	0.536
Gaussian Naive Bayes	0.952	0.698	0.496	0.580	0.526
K Nearest Neighbors	0.823	0.226	0.677	0.339	0.517
Random Forest Classifier	0.879	0.340	0.846	0.485	0.639

CLASSIFICATION MODEL

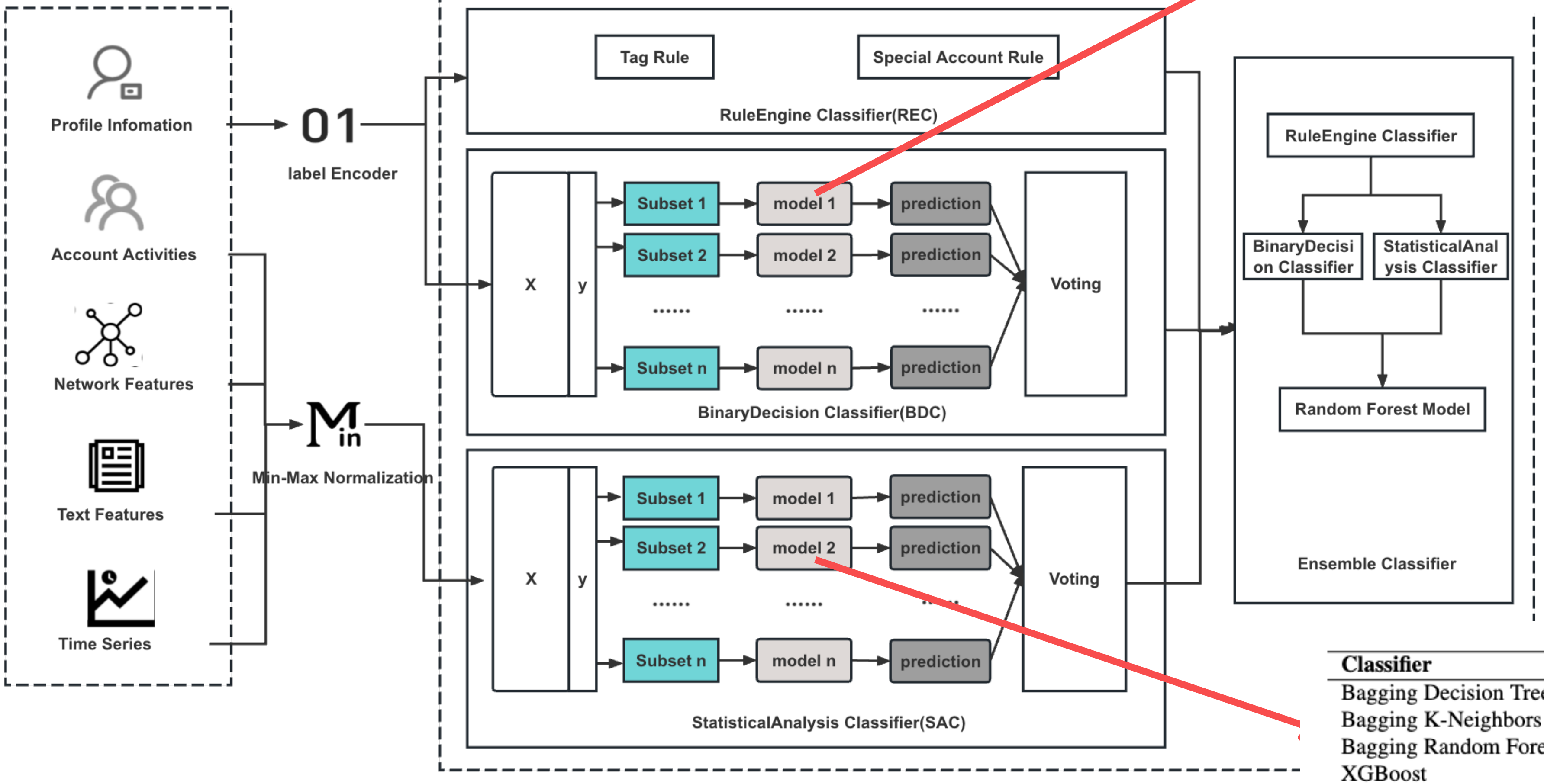


Table 5. Performance Metrics of BDC Classifiers

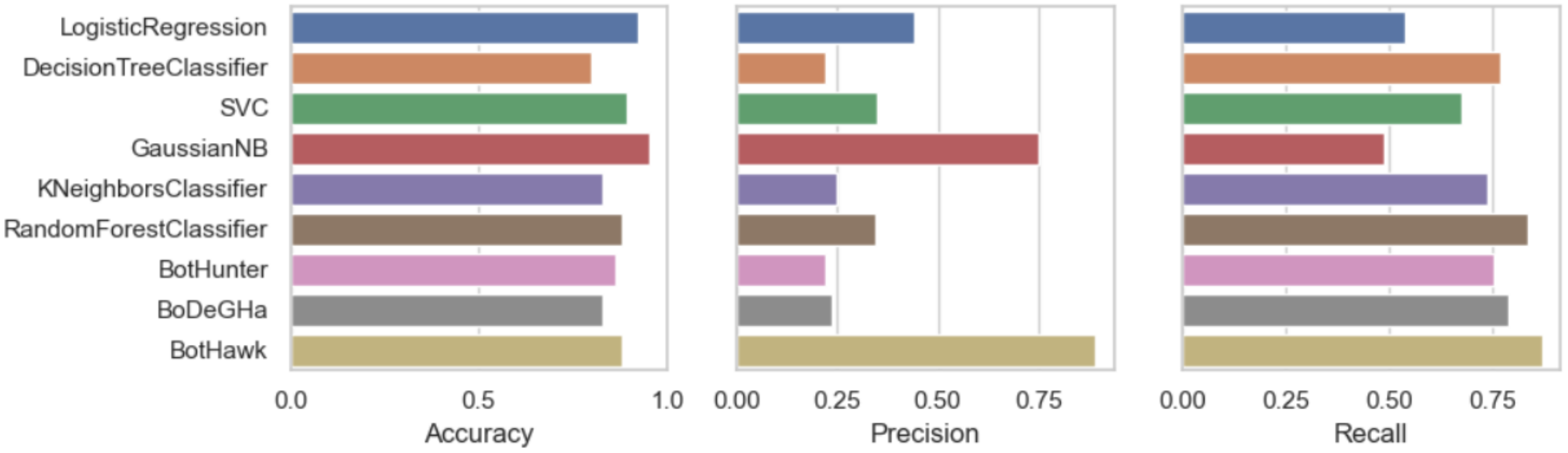
Classifier	Accuracy	Precision	Recall	F1-score	AUC
Bagging Decision Tree	0.715	0.98	0.43	0.60	0.85
Bagging KNeighbors	0.708	0.97	0.42	0.58	0.84
Bagging Random Forest	0.720	0.99	0.44	0.61	0.86
Bagging XGBoost	0.725	0.95	0.45	0.62	0.87
Bagging Logistic Regression	0.710	0.96	0.41	0.57	0.83
Bagging SVC	0.712	0.97	0.40	0.57	0.82
Bagging GaussianNB	0.707	0.94	0.39	0.55	0.81

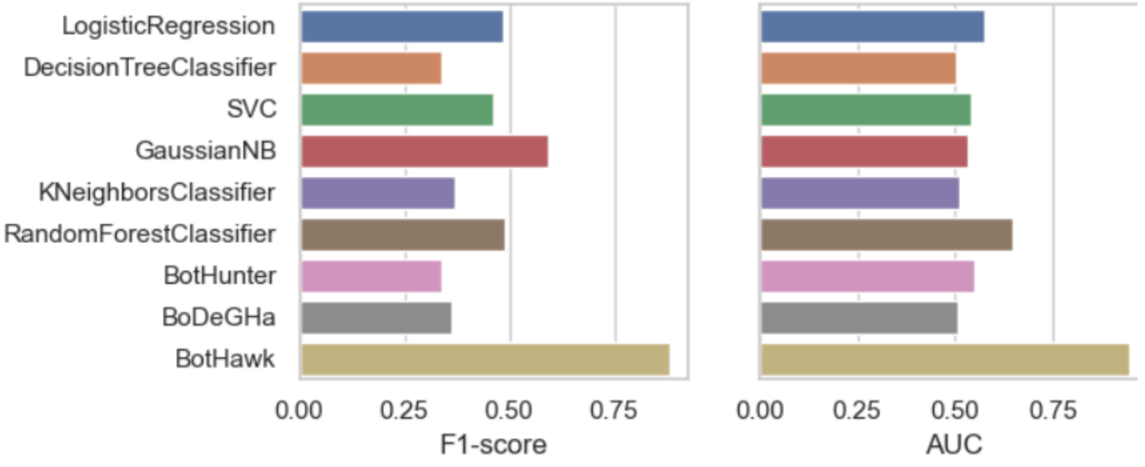
Table 6. Performance Metrics of SAC Classifiers

Classifier	Accuracy	Precision	Recall	F1 Score	AUC
Bagging Decision Tree	0.9470	0.9396	0.9524	0.9059	0.9446
Bagging K-Neighbors	0.9007	0.9606	0.8299	0.8505	0.9063
Bagging Random Forest	0.9636	0.9722	0.9524	0.9122	0.9483
XGBoost	0.9470	0.9338	0.9592	0.9063	0.9873
Bagging Logistic Regression	0.9040	0.9758	0.7831	0.8930	0.9153
Bagging SVC	0.8940	0.9832	0.7959	0.8397	0.9067
Bagging Gaussian NB	0.8675	0.9908	0.7347	0.8038	0.9323

RESULT

- Assessment indicators
    - Accuracy
    - Precision
    - Recall
    - F1-Score
    - AUC
- BotHawk（0.947）、RandomForestClassifier（0.639）、SVC（0.536）





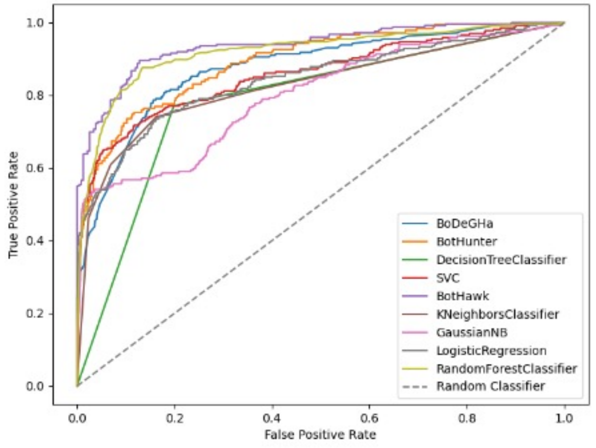
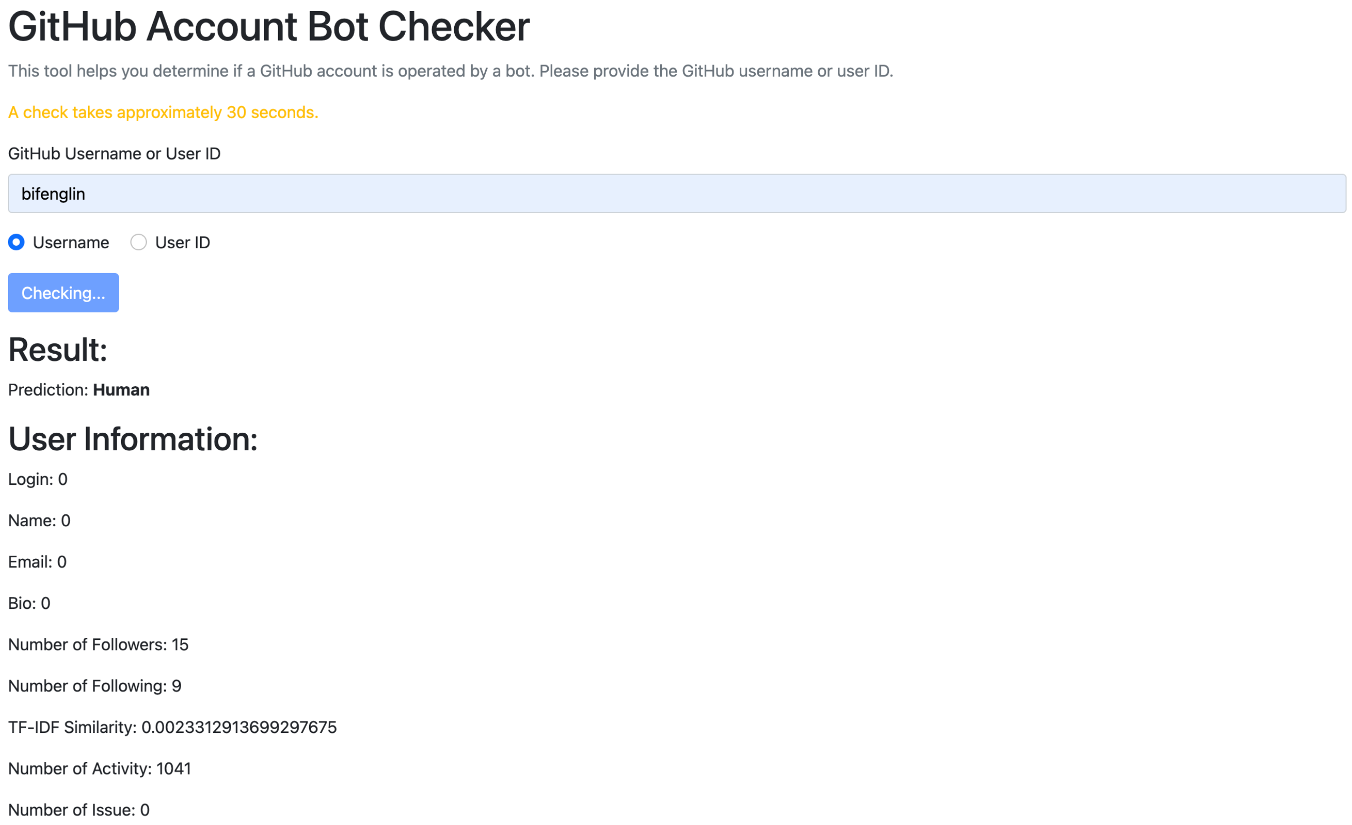


Figure 14. ROC Curve for Different Models

Figure 12. Comparison of Classification Models

Tool And Service

- Website and Service:
  - http://139.224.63.134:8000/
  - RESTFUL API
- Model and tool:
  - https://github.com/bifenglin/BotHawk



GitHub Account Bot Checker

This tool helps you determine if a GitHub account is operated by a bot. Please provide the GitHub username or user ID.

A check takes approximately 30 seconds.

GitHub Username or User ID

bifenglin

● Username ○ User ID

Result:

Prediction: **Human**

User Information:

Login: 0

Name: 0

Email: 0

Bio: 0

Number of Followers: 15

Number of Following: 9

TF-IDF Similarity: 0.0023312913699297675

Number of Activity: 1041

Number of Issue: 0

DESCUSSION

- **Identifying Bot Accounts:**
  - **BotHawk:** Trained on a dataset that includes a wide variety of bot account types, providing a more realistic portrayal of bot-related scenarios, and performs exceptionally well.
  - **BoDeGHa:** Excels in identifying bot accounts that exhibit comment-related features but is limited to assessing bot behavior within a specific repository, lacking a comprehensive perspective.
  - **BotHunter:** Focuses on simplistic features and fails to explore the comprehensive behavioral characteristics associated with bots.

BotHawk exhibits outstanding performance in handling datasets that closely emulate real-world scenarios, particularly in recognizing CICD and Scanning bots.

DESCUSSION

- **Importance of Features:**
  - 'Tag', 'Number of followers', and 'Number of Issues' show higher levels of importance, suggesting a strong positive correlation with the identification performance of OSS bots.

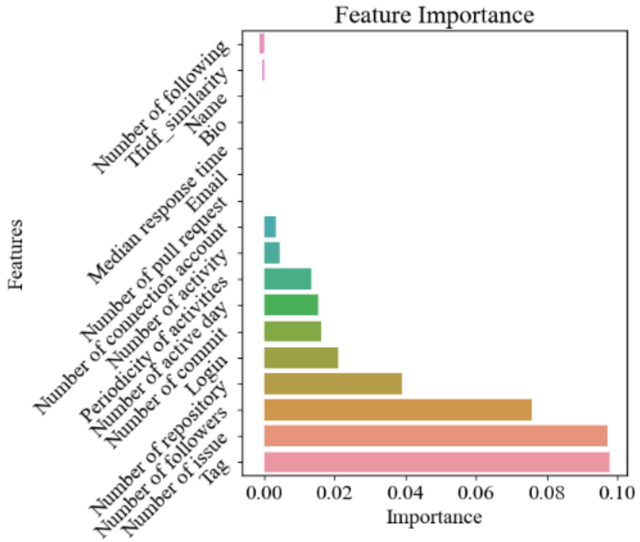


Figure 14. Feature Importance for BotHawk model

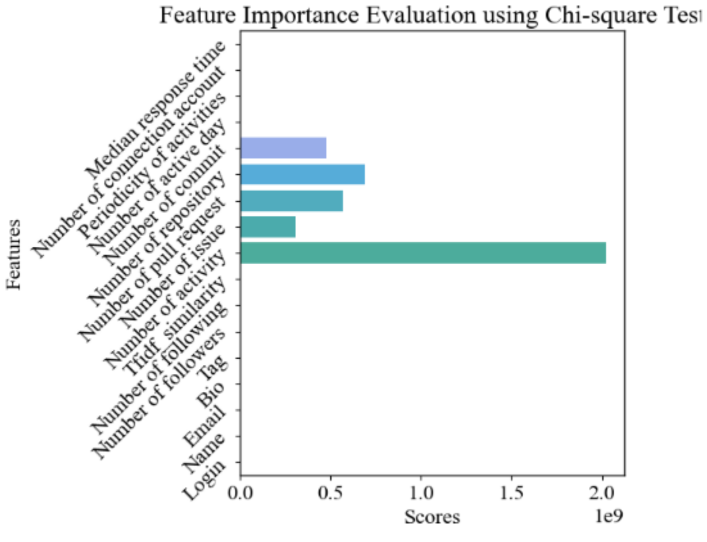
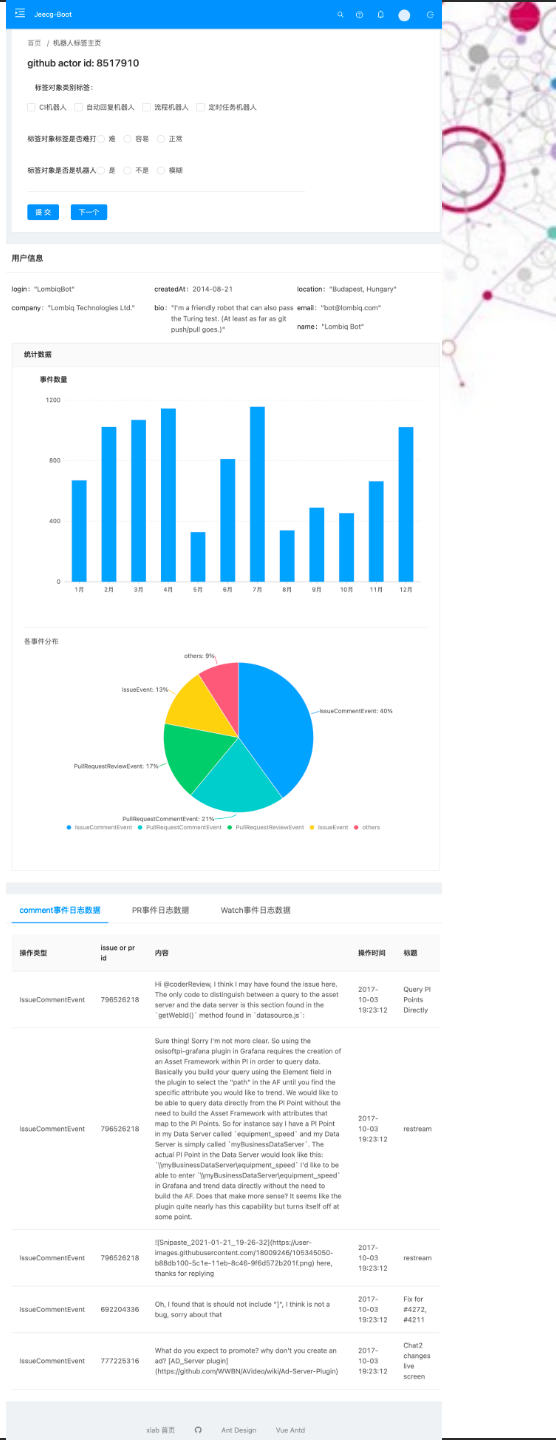


Figure 15. Feature Importance Evaluation using Chi-square Test

CONCLUSION

- **Work：**
  - **A more extensive open dataset on open source bot detection**
  - **Study and categoriy about behavior patterns of OSS bot.**
  - **Find best indicators of bot detection**
  - **A state-of-the-art model of OSS bot detection**
  - **Bot detection tool and service**
- **Future work:**
  - **Add more features： Graph feature**
  - **Consider more models: GNN**
  - **Multi-label classification task**
  - **More fine-grained recognition tasks： behavior level recognition**

THANKS!



github actor id: 8517910

标签对象类别标签：

□ CI机器人 □ 自动回复机器人 □ 流程机器人 □ 定时任务机器人

标签对象标签是否难打 ○ 难 ○ 容易 ○ 正常

标签对象是否是机器人 ○ 是 ○ 不是 ○ 模糊

用户信息

login："LombiqBot"

createdAt：2014-08-21

location："Budapest, Hungary"

company："Lombiq Technologies Ltd."

bio："I'm a friendly robot that can also pass the Turing test. (At least as far as git push/pull goes.)"

email："bot@lombiq.com"

name："Lombiq Bot"

统计数据

操作类型	issue or pr id	内容	操作时间	标题
IssueCommentEvent	796526218	Hi @coderReview, I think I may have found the issue here. The only code to distinguish between a query to the asset server and the data server is this section found in the `getWebId()` method found in `datasource.js`:	2017-10-03 19:23:12	Query PI Points Directly
IssueCommentEvent	796526218	Sure thing! Sorry I'm not more clear. So using the osisoftpi-grafana plugin in Grafana requires the creation of an Asset Framework within PI in order to query data. Basically you build your query using the Element field in the plugin to select the "path" in the AF until you find the specific attribute you would like to trend. We would like to be able to query data directly from the PI Point without the need to build the Asset Framework with attributes that map to the PI Points. So for instance say I have a PI Point in my Data Server called `equipment_speed` and my Data Server is simply called `myBusinessDataServer`. The actual PI Point in the Data Server would look like this: `\\myBusinessDataServer\equipment_speed` I'd like to be able to enter `\\myBusinessDataServer\equipment_speed` in Grafana and trend data directly without the need to build the AF. Does that make more sense? It seems like the plugin quite nearly has this capability but turns itself off at some point.	2017-10-03 19:23:12	restream
IssueCommentEvent	796526218	![Snipaste_2021-01-21_19-26-32](https://user-images.githubusercontent.com/18009246/105345050-b88db100-5c1e-11eb-8c46-9f6d572b201f.png) here, thanks for replying	2017-10-03 19:23:12	restream
IssueCommentEvent	692204336	Oh, I found that is should not include "]", I think is not a bug, sorry about that	2017-10-03 19:23:12	Fix for #4272, #4211
IssueCommentEvent	777225316	What do you expect to promote? why don't you create an ad? [AD_Server plugin](https://github.com/WWBN/AVideo/wiki/Ad-Server-Plugin)	2017-10-03 19:23:12	Chat2 changes live screen