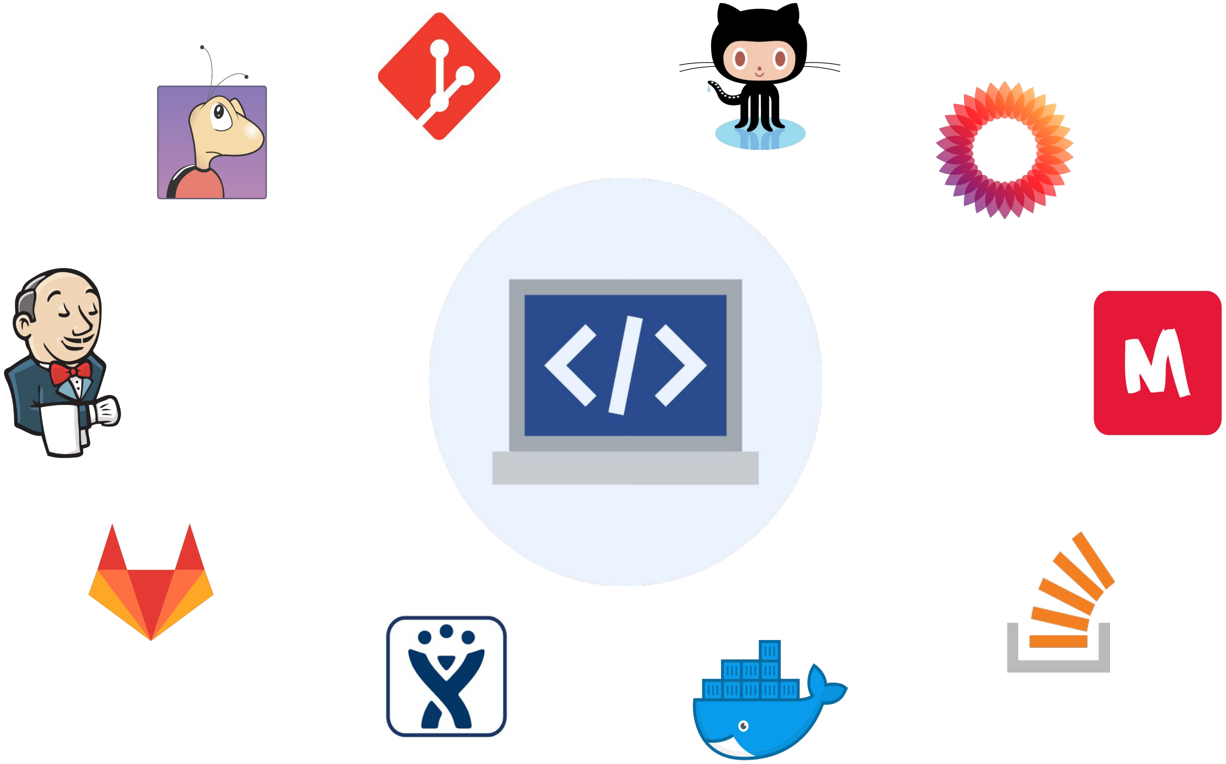
Automatic identification of bot accounts in *Open-Source projects*

Miguel Ángel Fernández Sánchez
MSc on Data Science
Tutor: José Felipe Ortega Soto, PhD.

April 20th, 2023

Universidad Rey Juan Carlos

/motivation



/motivation

Accounts

git	jane.doe@urjc.es janedoe@libresoft.com
GitHub	janedoe
stackoverflow	doe.jane
Mattermost	janedoe

Organizations

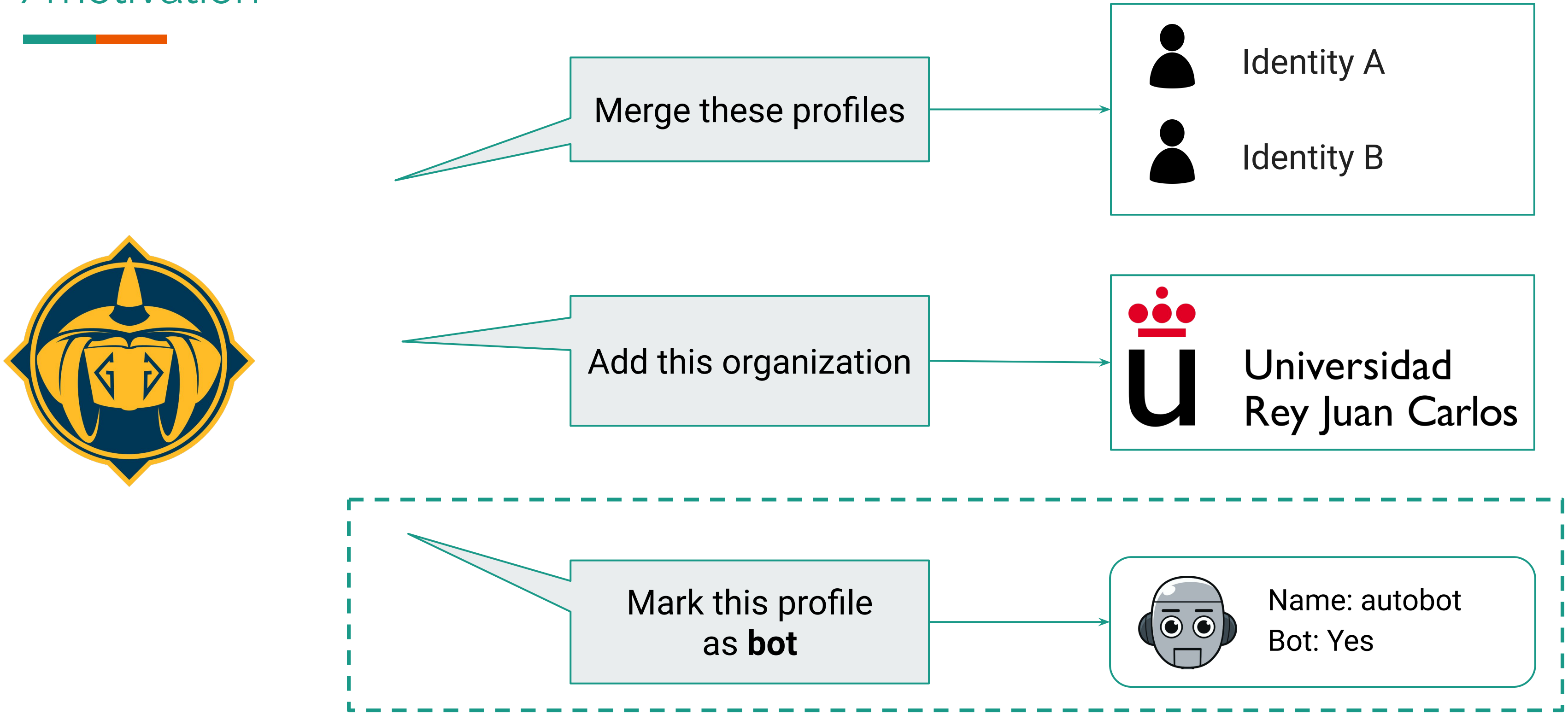
Universidad
Rey Juan Carlos

From: September 1st, 2015
To: January 31st, 2017

GSyC LibreSoft we study libre software

From: February 1st, 2017
To: Present day

/motivation



Merge these profiles

Identity A

Identity B

Add this organization

Universidad Rey Juan Carlos

Mark this profile as **bot**

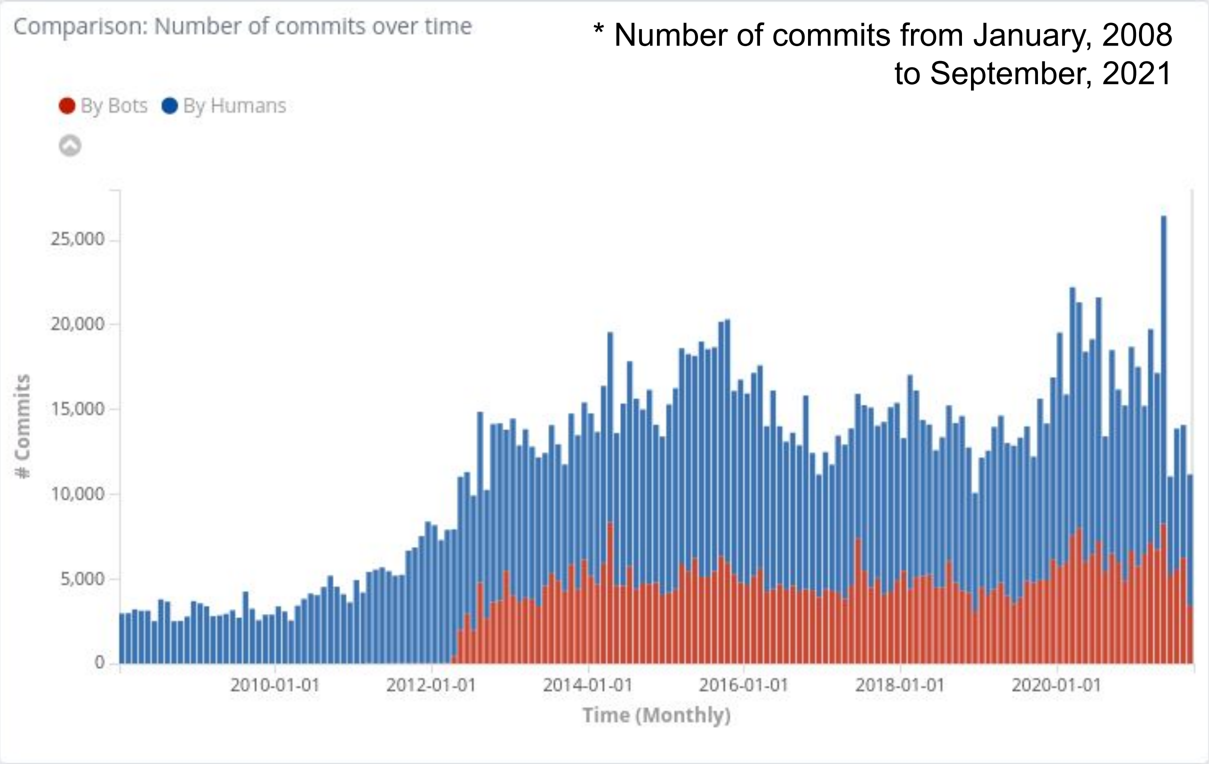
Name: autobot

Bot: Yes

/problem

🔵 Commits from humans

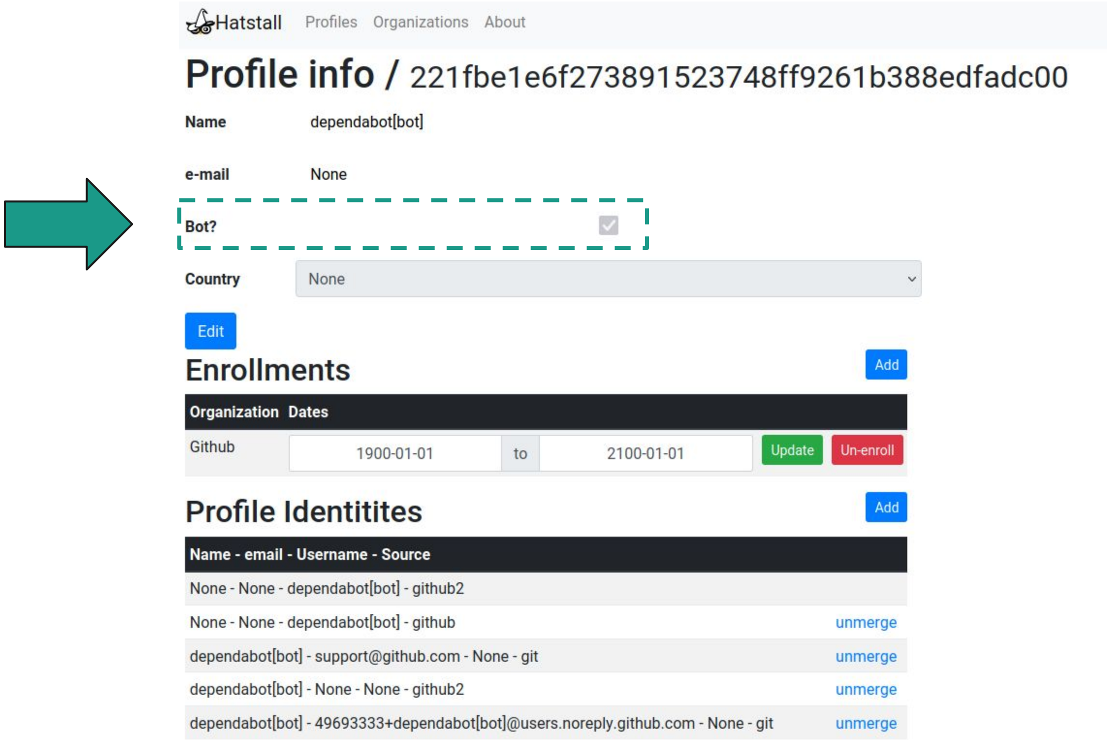
🔴 Commits from bots



Comparison: Number of commits over time

* Number of commits from January, 2008 to September, 2021

/problem



Profile info / 221fbe1e6f273891523748ff9261b388edfadc00

Name	dependabot[bot]
e-mail	None
Bot?	☑
Country	None

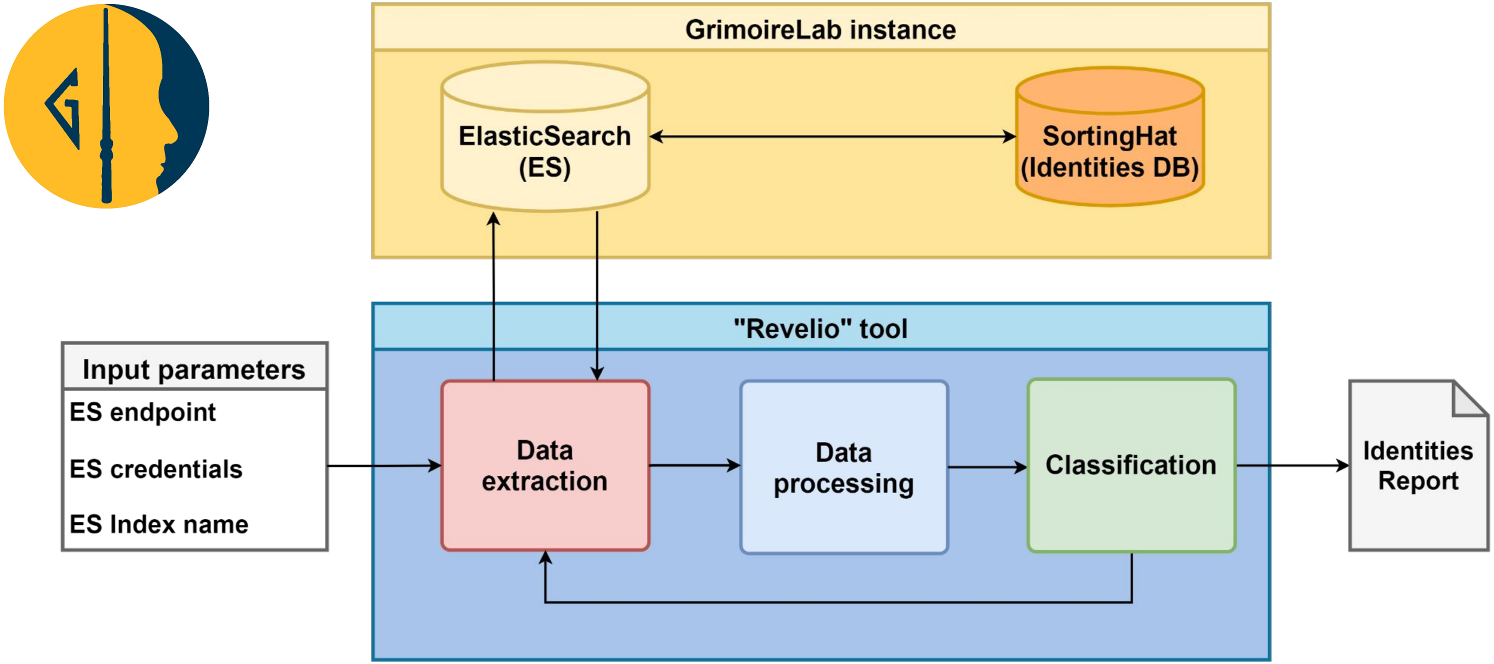
Enrollments

Organization	Dates			
Github	1900-01-01	to	2100-01-01	Update Un-enroll

Profile Identitites

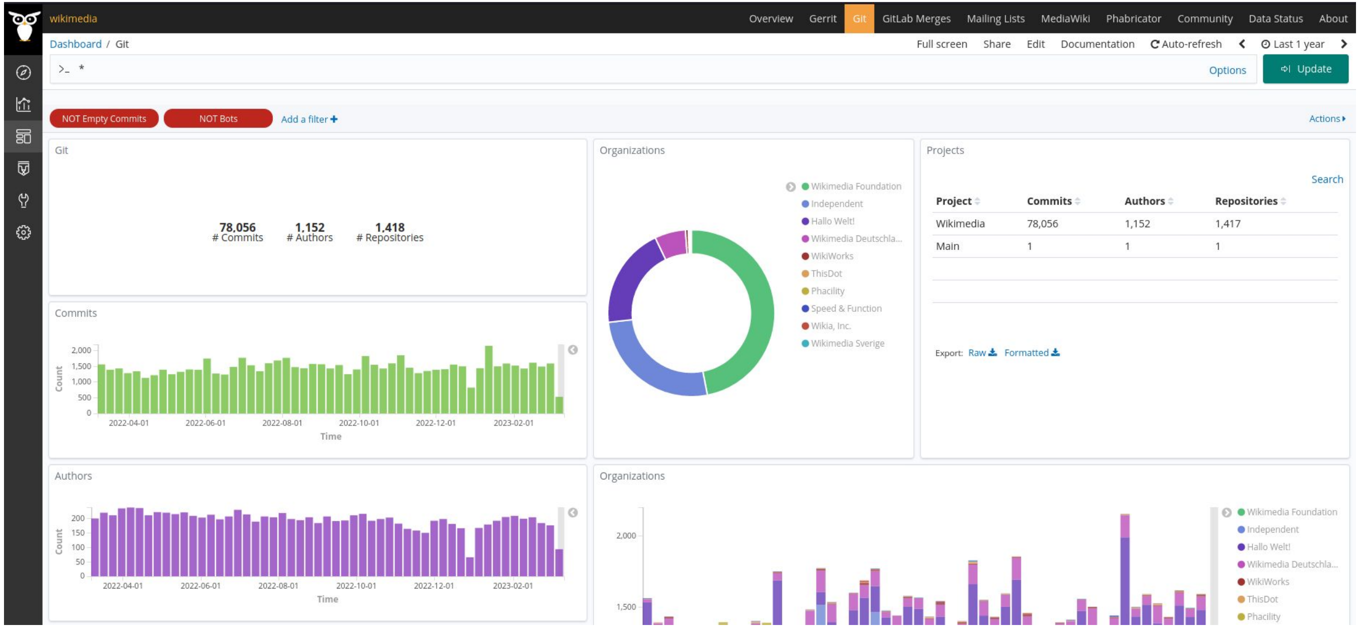
Name - email - Username - Source	
None - None - dependabot[bot] - github2	
None - None - dependabot[bot] - github	unmerge
dependabot[bot] - support@github.com - None - git	unmerge
dependabot[bot] - None - None - github2	unmerge
dependabot[bot] - 49693333+dependabot[bot]@users.noreply.github.com - None - git	unmerge

/proposal


GrimoireLab instance
ElasticSearch (ES)
SortingHat (Identities DB)
"Revelio" tool
Input parameters
ES endpoint
ES credentials
ES Index name
Data extraction
Data processing
Classification
Identities Report

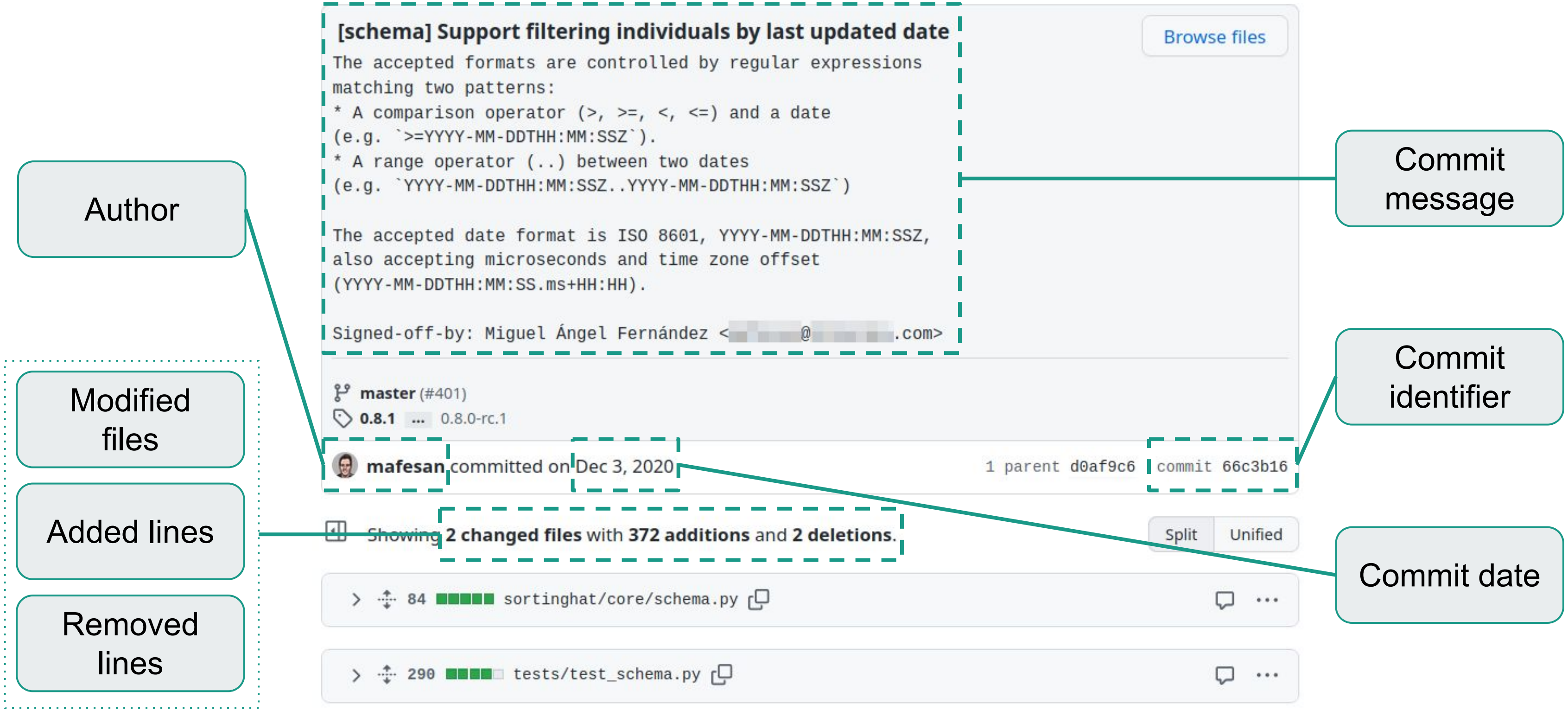
/choosing the community to analyse

WIKIMEDIA
FOUNDATION



Project	Commits	Authors	Repositories
Wikimedia	78,056	1,152	1,417
Main	1	1	1

/details from a Git commit



[schema] Support filtering individuals by last updated date

Browse files

```
The accepted formats are controlled by regular expressions
matching two patterns:
* A comparison operator (>, >=, <, <=) and a date
(e.g. `>=YYYY-MM-DDTHH:MM:SSZ`).
* A range operator (..) between two dates
(e.g. `YYYY-MM-DDTHH:MM:SSZ..YYYY-MM-DDTHH:MM:SSZ`)

The accepted date format is ISO 8601, YYYY-MM-DDTHH:MM:SSZ,
also accepting microseconds and time zone offset
(YYYY-MM-DDTHH:MM:SS.ms+HH:HH).

Signed-off-by: Miguel Ángel Fernández <[redacted]@[redacted].com>
```

master (#401)

0.8.1 ... 0.8.0-rc.1

mafesan committed on Dec 3, 2020

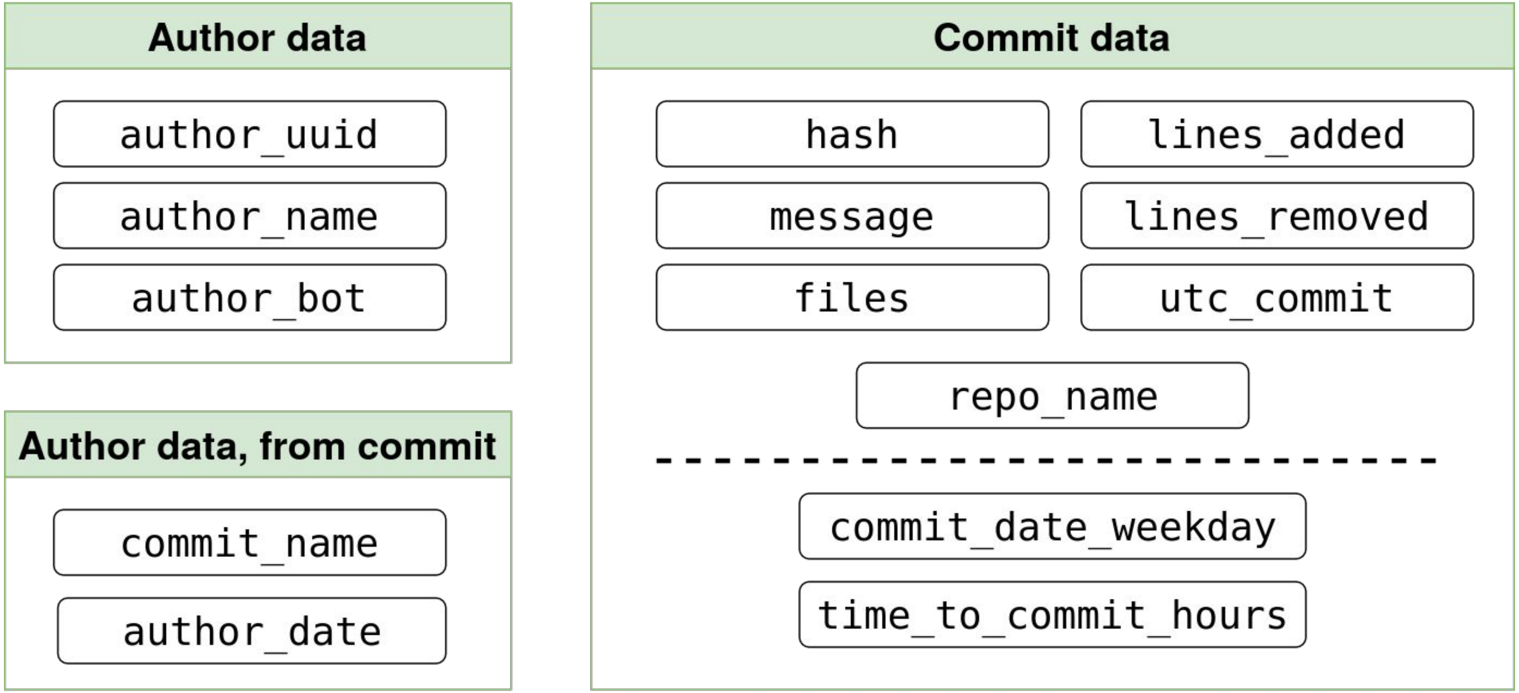
1 parent d0af9c6 commit 66c3b16

Showing **2 changed files** with **372 additions** and **2 deletions**.

Split Unified

```
84 ▇▇▇▇▇ sortinghat/core/schema.py
290 ▇▇▇▇▢ tests/test_schema.py
```

/choosing fields using *GQM methodology*



Author data

- author_uuid
- author_name
- author_bot

Author data, from commit

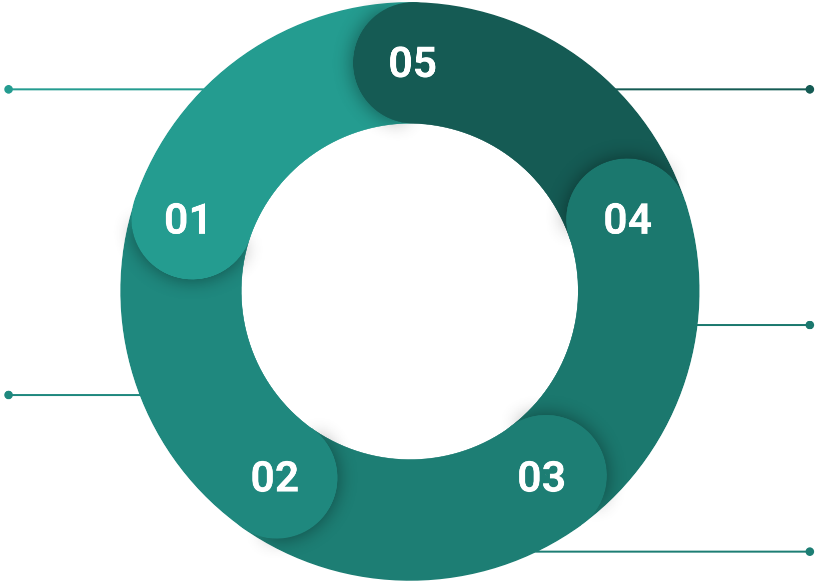
- commit_name
- author_date

Commit data

- hash
- lines_added
- message
- lines_removed
- files
- utc_commit
- repo_name

- commit_date_weekday
- time_to_commit_hours

/experiments phase



Pre-processing of the data

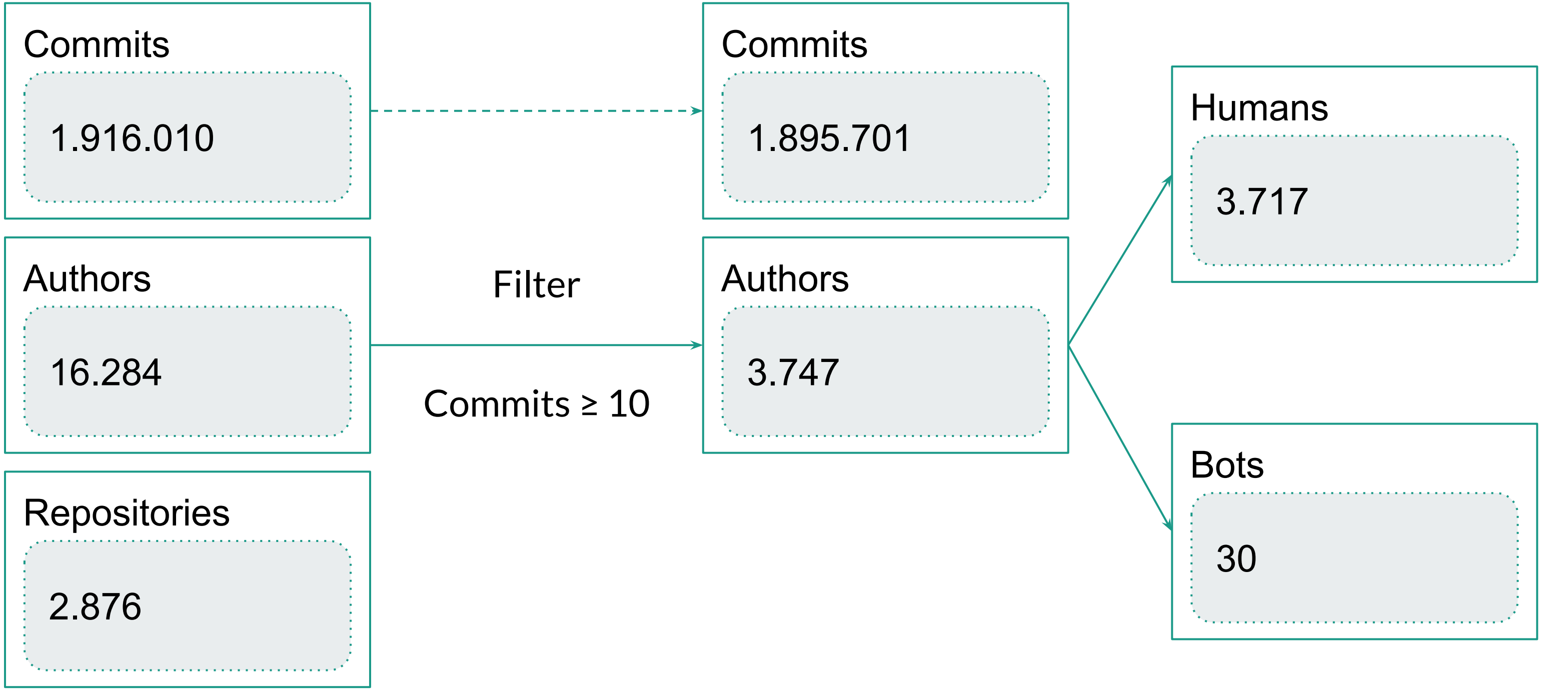
Clasification models

Validation of results

Choosing the classifier

Hyperparameter adjustment

/exploratory data analysis



Commits

1.916.010

Authors

16.284

Repositories

2.876

Filter

Commits ≥ 10

Commits

1.895.701

Authors

3.747

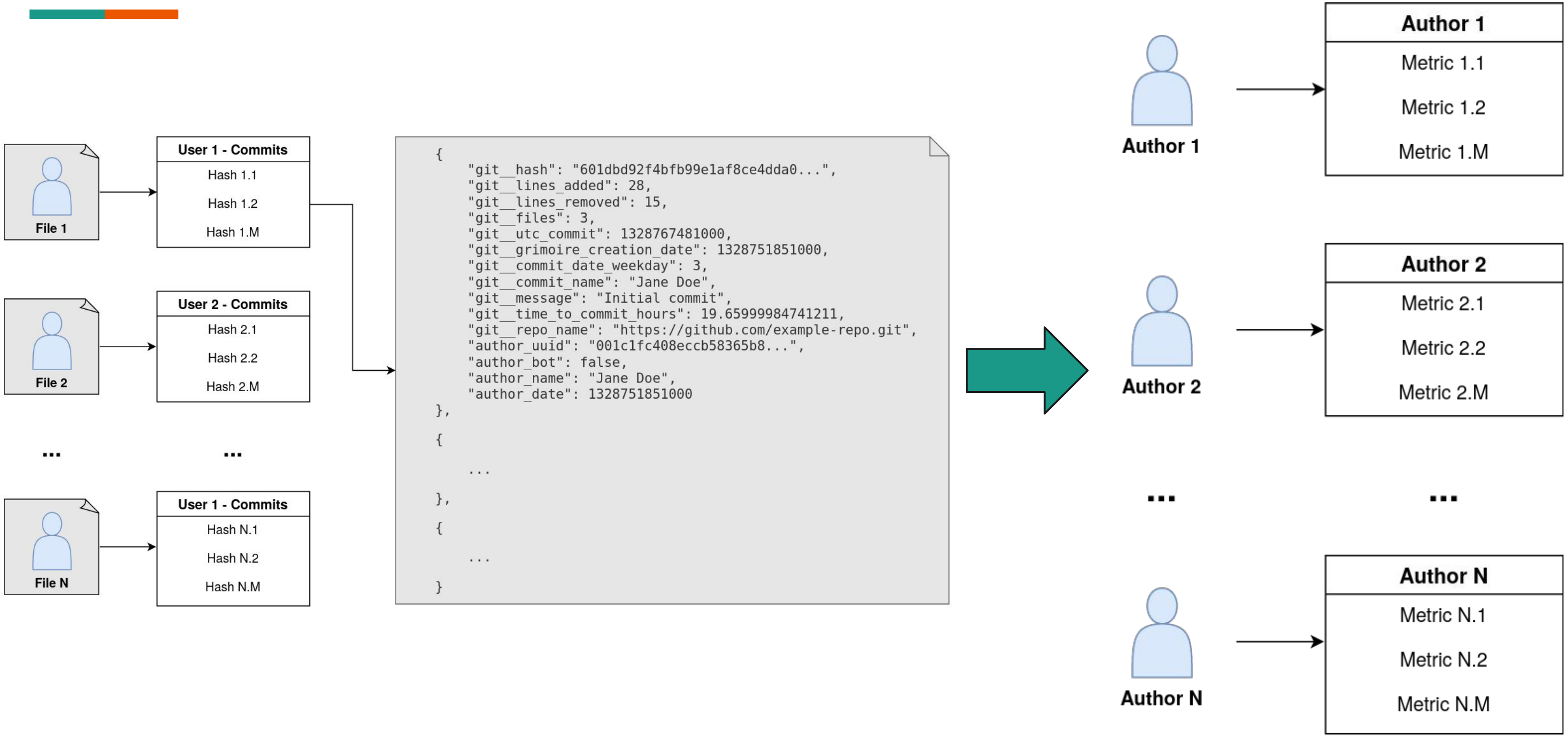
Humans

3.717

Bots

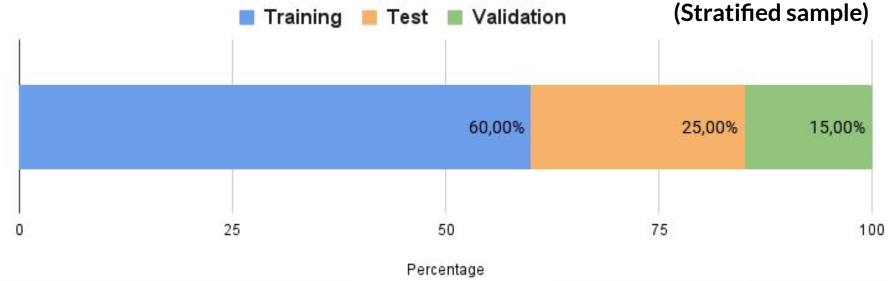
30

/input dataset

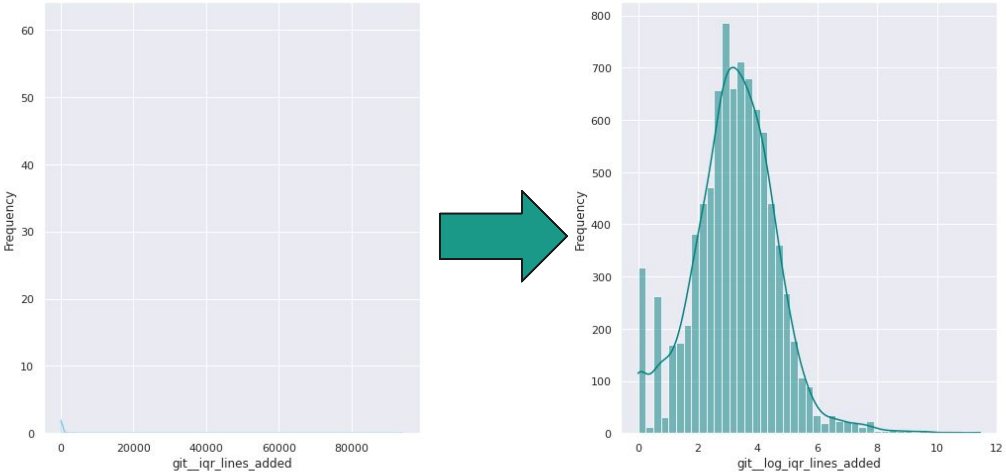


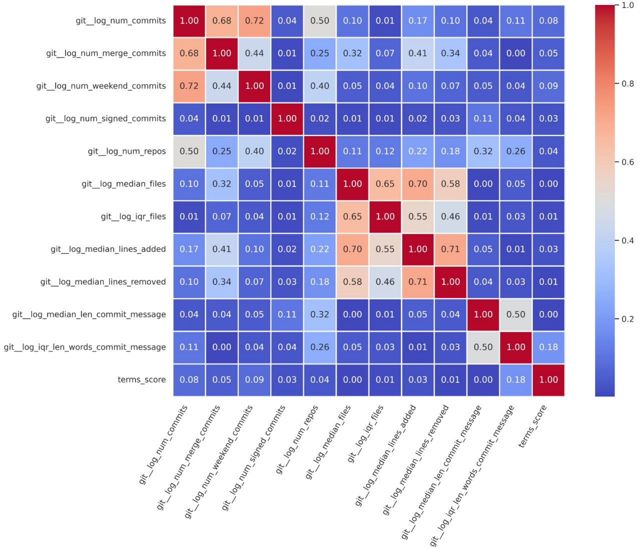
```
{
    "git__hash": "601dbd92f4bfb99e1af8ce4dda0...",
    "git__lines_added": 28,
    "git__lines_removed": 15,
    "git__files": 3,
    "git__utc_commit": 1328767481000,
    "git__grimoire_creation_date": 1328751851000,
    "git__commit_date_weekday": 3,
    "git__commit_name": "Jane Doe",
    "git__message": "Initial commit",
    "git__time_to_commit_hours": 19.65999984741211,
    "git__repo_name": "https://github.com/example-repo.git",
    "author_uuid": "001c1fc408eccb58365b8...",
    "author_bot": false,
    "author_name": "Jane Doe",
    "author_date": 1328751851000
},
{

    ...

},
{

    ...

}
```

/pre-processing of the data and feature transformation



(Stratified sample)





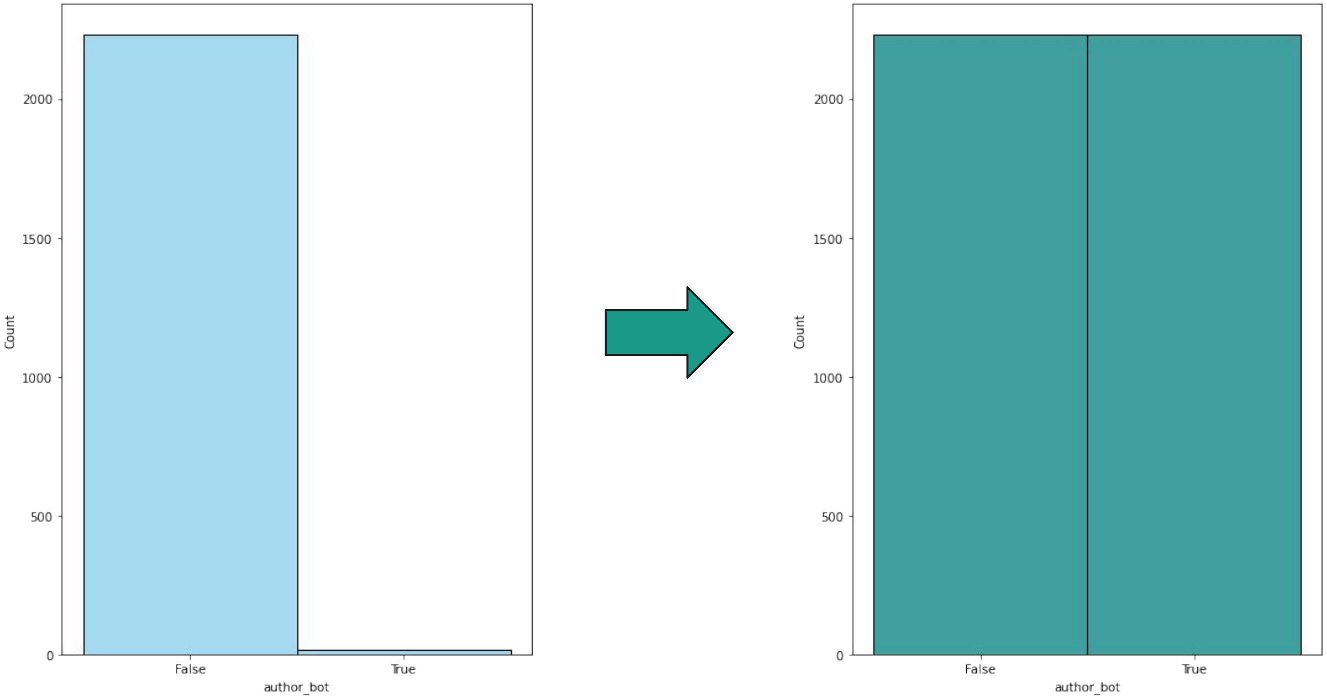
	git__log_num_commits	git__log_num_merge_commits	git__log_num_weekend_commits	git__log_num_signed_commits	git__log_num_repos	git__log_median_files	git__log_iqr_files	git__log_median_lines_added	git__log_median_lines_removed	git__log_median_len_commit_message	git__log_iqr_len_words_commit_message	terms_score
git__log_num_commits	1.00	0.68	0.72	0.04	0.50	0.10	0.01	0.17	0.10	0.04	0.11	0.08
git__log_num_merge_commits	0.68	1.00	0.44	0.01	0.25	0.32	0.07	0.41	0.34	0.04	0.00	0.05
git__log_num_weekend_commits	0.72	0.44	1.00	0.01	0.40	0.05	0.04	0.10	0.07	0.05	0.04	0.09
git__log_num_signed_commits	0.04	0.01	0.01	1.00	0.02	0.01	0.01	0.02	0.03	0.11	0.04	0.03
git__log_num_repos	0.50	0.25	0.40	0.02	1.00	0.11	0.12	0.22	0.18	0.32	0.26	0.04
git__log_median_files	0.10	0.32	0.05	0.01	0.11	1.00	0.65	0.70	0.58	0.00	0.05	0.00
git__log_iqr_files	0.01	0.07	0.04	0.01	0.12	0.65	1.00	0.55	0.46	0.01	0.03	0.01
git__log_median_lines_added	0.17	0.41	0.10	0.02	0.22	0.70	0.55	1.00	0.71	0.05	0.01	0.03
git__log_median_lines_removed	0.10	0.34	0.07	0.03	0.18	0.58	0.46	0.71	1.00	0.04	0.03	0.01
git__log_median_len_commit_message	0.04	0.04	0.05	0.11	0.32	0.00	0.01	0.05	0.04	1.00	0.50	0.00
git__log_iqr_len_words_commit_message	0.11	0.00	0.04	0.04	0.26	0.05	0.03	0.01	0.03	0.50	1.00	0.18
terms_score	0.08	0.05	0.09	0.03	0.04	0.00	0.01	0.03	0.01	0.00	0.18	1.00

/feature generation

Level	Weight	Heuristic terms
1	60	`bot, dependency, fix, integration, merge`
2	30	`auto, build, commit, copy, issue, release, request, review, sync, template, tool, travis`
3	10	`cd, ci, code, patrol, pr, pull`

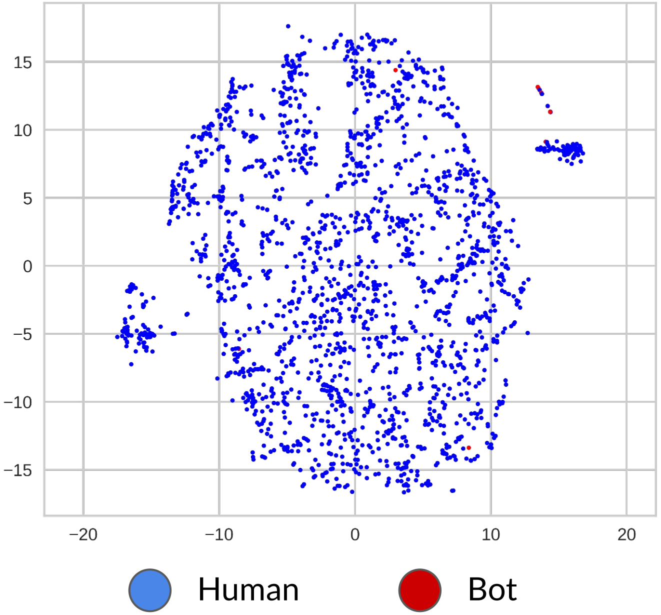
$$\boxed{\mathrm{Ts}(\mathrm{term}) = 60\mathrm{Nl}_1 + 30\mathrm{Nl}_2 + 10\mathrm{Nl}_3}$$

/unbalanced data: SMOTE

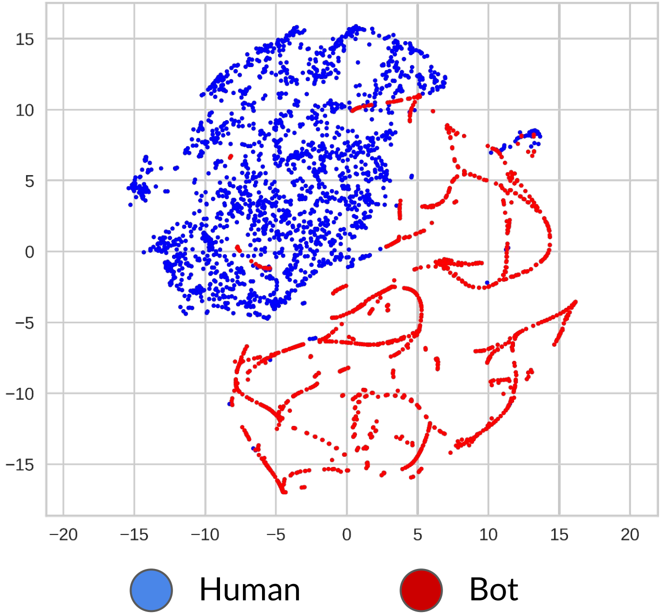


/class distribution using t-SNE

Original



After applying SMOTE



/classifiers: evaluation

	Predicted	
Real	TN	FP
	FN	TP

Precision

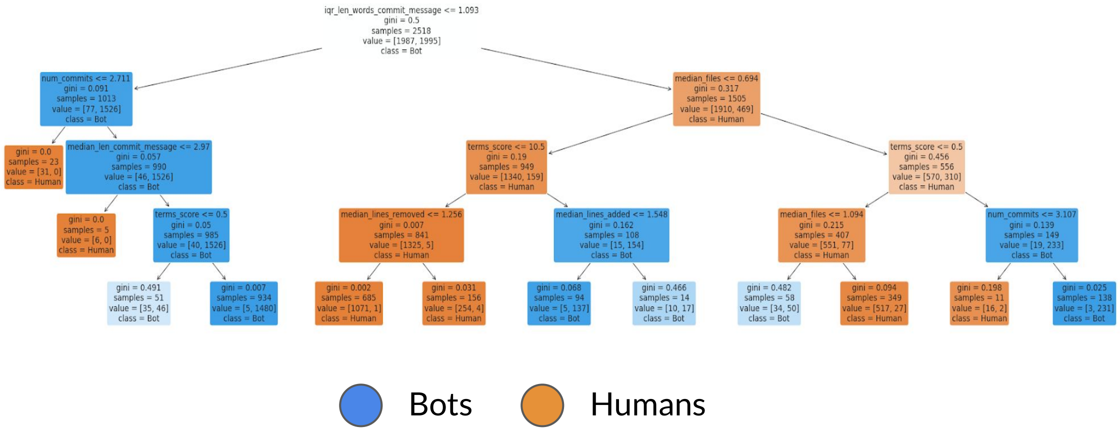
Recall

Fβ-score

Classification model
Gaussian Naive-Bayes
Complement Naive-Bayes
LinearSVC
KNN
Decision Tree
Random Forest
XGBoost

/classifiers: Random Forest

Number of estimators	300
Split criterion	*Gini* impurity
Maximum depth	4 levels


iqr_len_words_commit_message <= 1.093 gini = 0.5 samples = 2518 value = [1987, 1995] class = Bot
num_commits <= 2.711 gini = 0.091 samples = 1013 value = [77, 1526] class = Bot
median_files <= 0.694 gini = 0.317 samples = 1505 value = [1910, 469] class = Human
gini = 0.0 samples = 23 value = [31, 0] class = Human
median_len_commit_message <= 2.97 gini = 0.057 samples = 990 value = [46, 1526] class = Bot
terms_score <= 10.5 gini = 0.19 samples = 949 value = [1340, 159] class = Human
terms_score <= 0.5 gini = 0.456 samples = 556 value = [570, 310] class = Human
gini = 0.0 samples = 5 value = [6, 0] class = Human
terms_score <= 0.5 gini = 0.05 samples = 985 value = [40, 1526] class = Bot
median_lines_removed <= 1.256 gini = 0.007 samples = 841 value = [1325, 5] class = Human
median_lines_added <= 1.548 gini = 0.162 samples = 108 value = [15, 154] class = Bot
median_files <= 1.094 gini = 0.215 samples = 407 value = [551, 77] class = Human
num_commits <= 3.107 gini = 0.139 samples = 149 value = [19, 233] class = Bot
gini = 0.491 samples = 51 value = [35, 46] class = Bot
gini = 0.007 samples = 934 value = [5, 1480] class = Bot
gini = 0.002 samples = 685 value = [1071, 1] class = Human
gini = 0.031 samples = 156 value = [254, 4] class = Human
gini = 0.068 samples = 94 value = [5, 137] class = Bot
gini = 0.466 samples = 14 value = [10, 17] class = Bot
gini = 0.482 samples = 58 value = [34, 50] class = Bot
gini = 0.094 samples = 349 value = [517, 27] class = Human
gini = 0.198 samples = 11 value = [16, 2] class = Human
gini = 0.025 samples = 138 value = [3, 231] class = Bot
Bots
Humans

/classifiers: test

		Predicted	
		Human	Bot
Real	Human	826	3
	Bot	1	6

Model name	***Precision***	***Recall***	***Fβ-score***
Random Forest	0.667	0.857	0.811

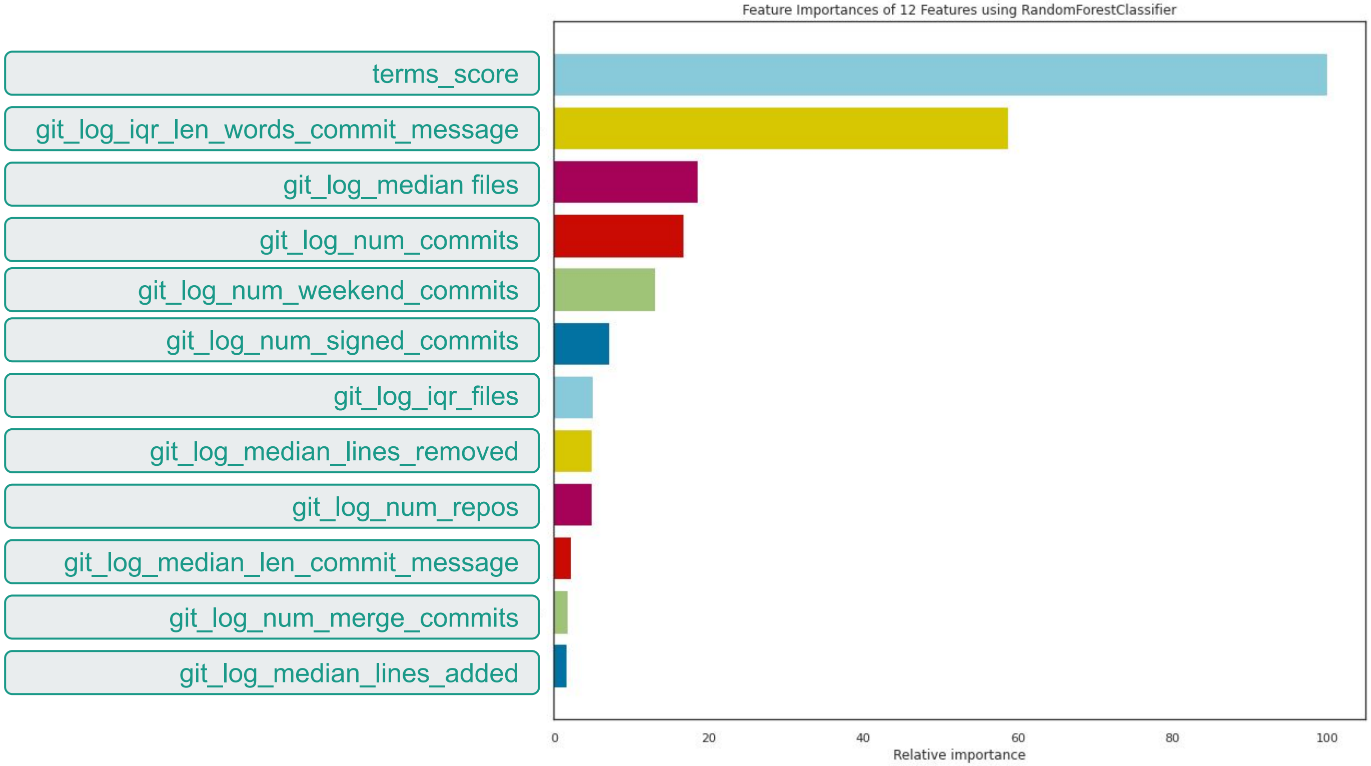
+ 5-fold cross-validation

/classifiers: validation

		Predicted	
		Human	Bot
Real	Human	493	6
	Bot	1	3

Model name	***Precision***	***Recall***	***Fβ-score***
Random Forest	0.333	0.75	0.6

/classifiers: most relevant features



/future work

Improvement of the classification model

- Other projects / Other communities
- Text metrics / Digital footprints
- Summary of the user / *Concept Learning*
- Mixed accounts / Multi-class classification

Integration with SortingHat/GrimoireLab

- Classification report
- Recommendation engine
- Feedback to the classification model after the user's decision

/end

Questions

mafesan.github.io/Memoria-TFM