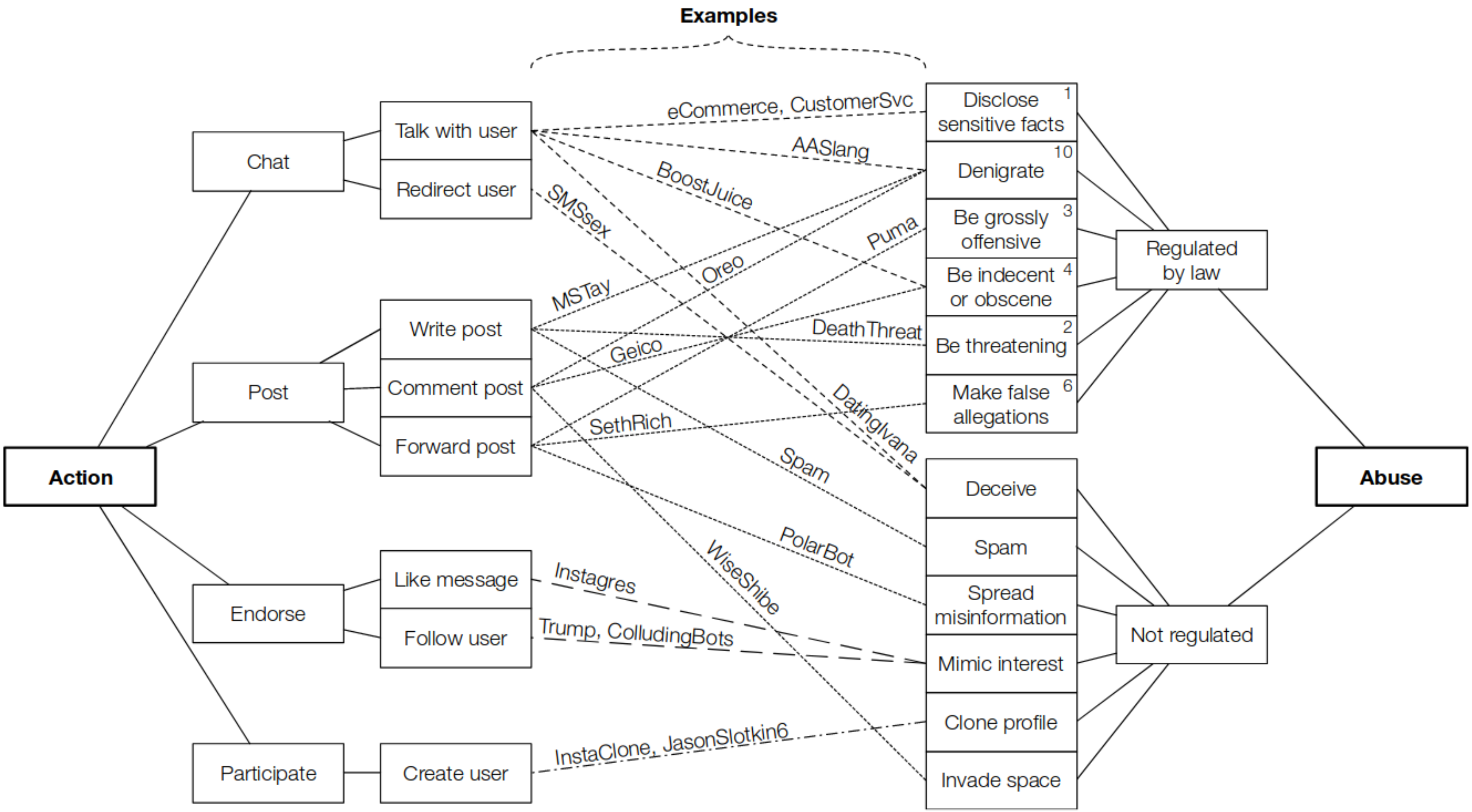
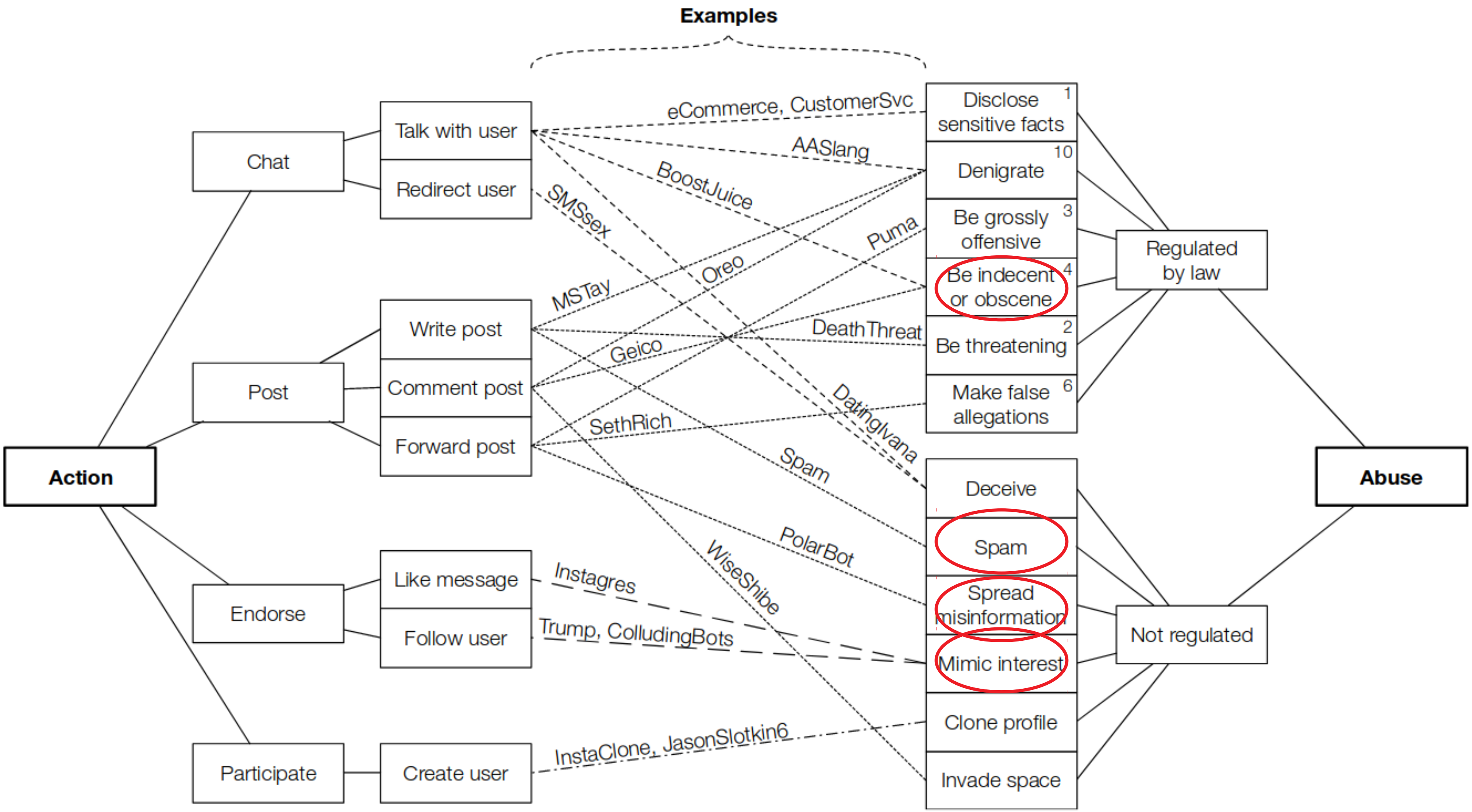
POLITECNICO
MILANO 1863

Detection and Classification of Harmful Bots in Social Human-Bot Interaction

Social bots


Examples
Action
Chat
Post
Endorse
Participate
Talk with user
Redirect user
Write post
Comment post
Forward post
Like message
Follow user
Create user
eCommerce, CustomerSvc
AASlang
BoostJuice
SMSsex
MSTay
Oreo
Puma
Geico
DeathThreat
SethRich
DatingIvana
Spam
PolarBot
WiseShibe
Instagres
Trump, ColludingBots
InstaClone, JasonSlotkin6
Disclose sensitive facts 1
Denigrate 10
Be grossly offensive 3
Be indecent or obscene 4
Be threatening 2
Make false allegations 6
Deceive
Spam
Spread misinformation
Mimic interest
Clone profile
Invade space
Regulated by law
Not regulated
Abuse

Twitter bots



Twitter bots

- NSFW
- News-spreaders
- Spambots
- Fake-followers

- Genuine

Tools

- **Twitter API**
- **Botometer API**
- **Hoaxy API**

HOAXY beta

Visualize the spread of claims and fact checking

Datasets

- **Caverlee-2011**
  - Bots
  - Humans
- **Cresci-2017**
  - spambots (job offers)
  - spambots (mobile app)
  - Spambots (Amazon products)
  - fake followers
  - Humans
- **Varol-2017**
  - Bots
  - Humans
- **BotBlock**
  - NSFW bots (adult contents)

Data Collection

NSFW

- ➔Get ids from BotBlock list
- ➔ Scrape users and tweets information with Twitter API

Spambots

From Cresci dataset:

- ➔ traditional spambots 1 (generic Spabots)
- ➔social spambots 2 (mobile app)
- ➔social spambots 3 (Amazon products)

Data Collection

Fake-followers

From Cresci dataset:

➔ Fake-followers

followers bought from.

➔instakipci.com/

➔rantic.com/buy-legit-twitter-followers

Genuine

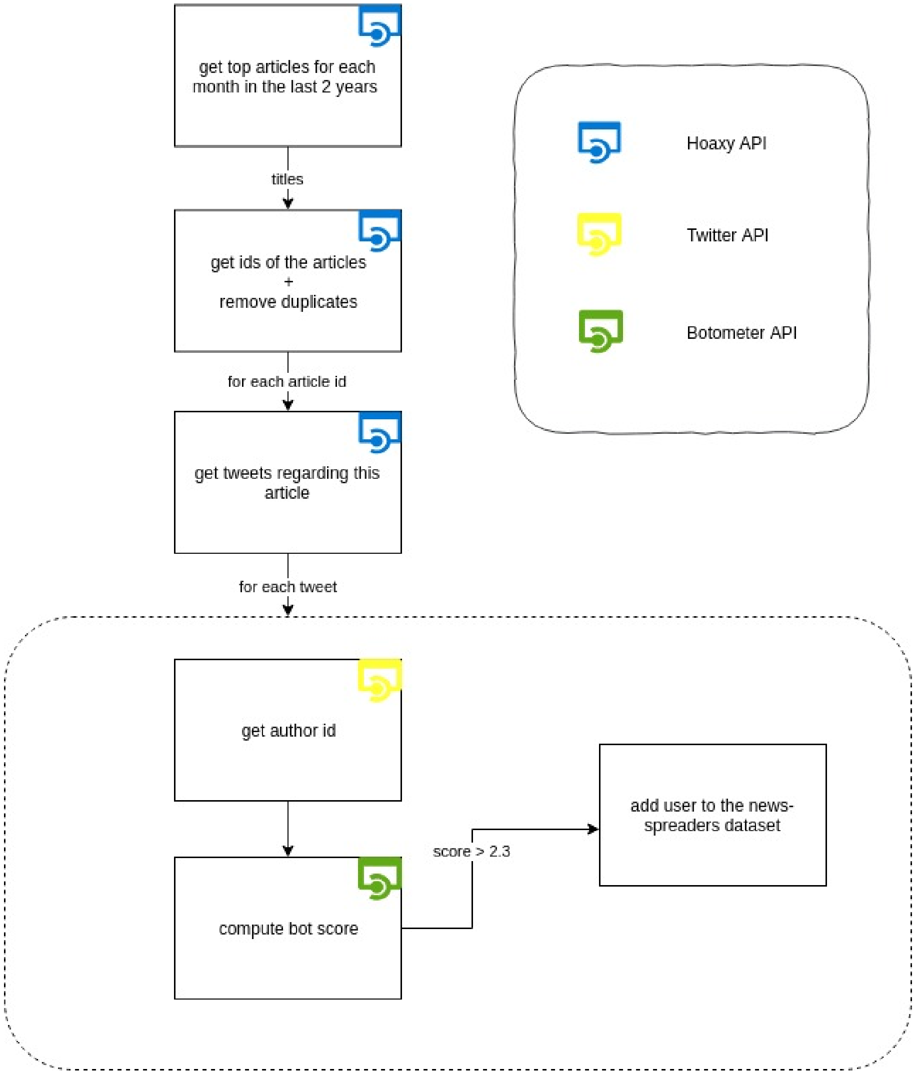
From Cresci dataset:

➔ Genuine

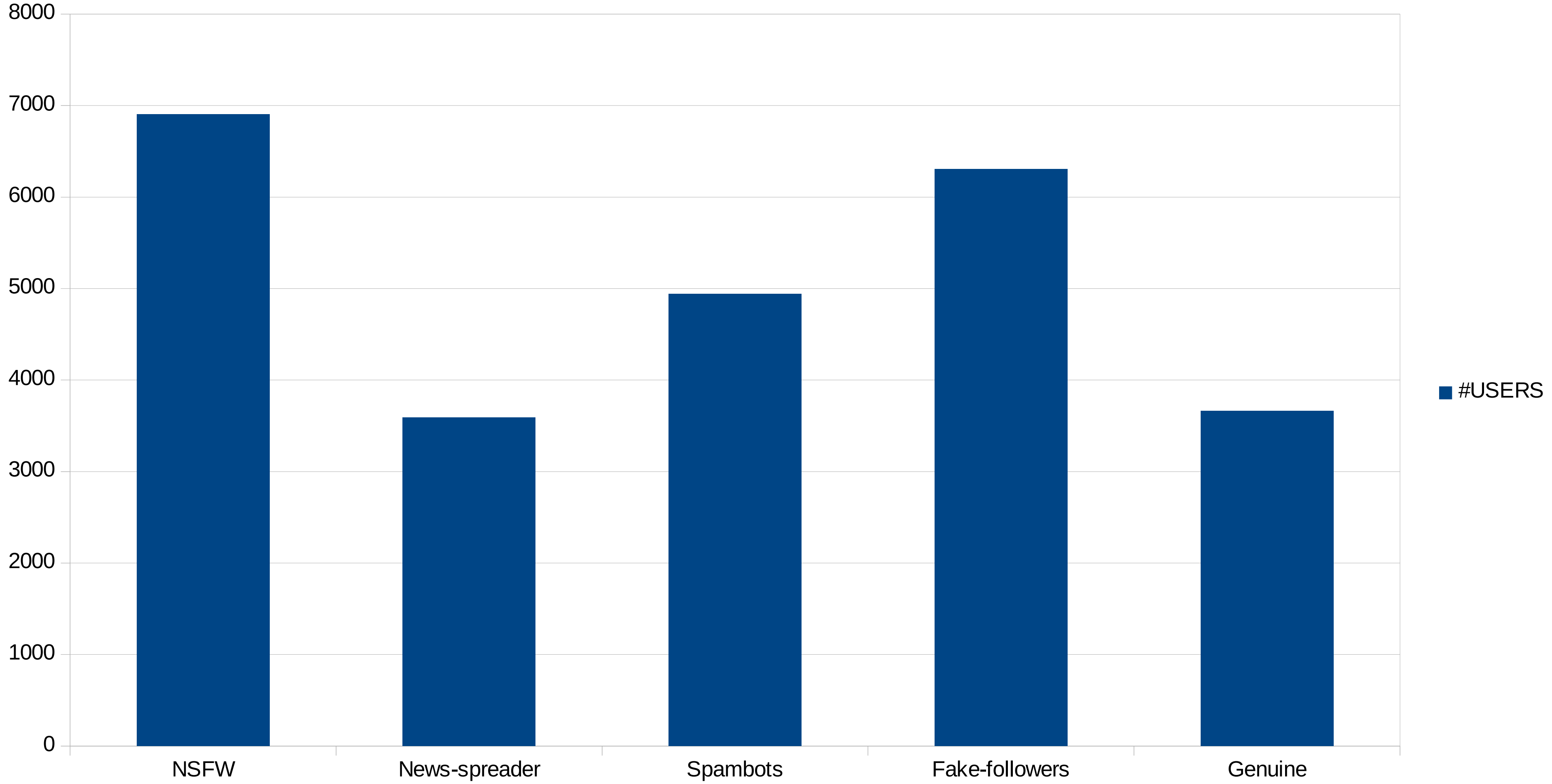
Data Collection

News-spreaders

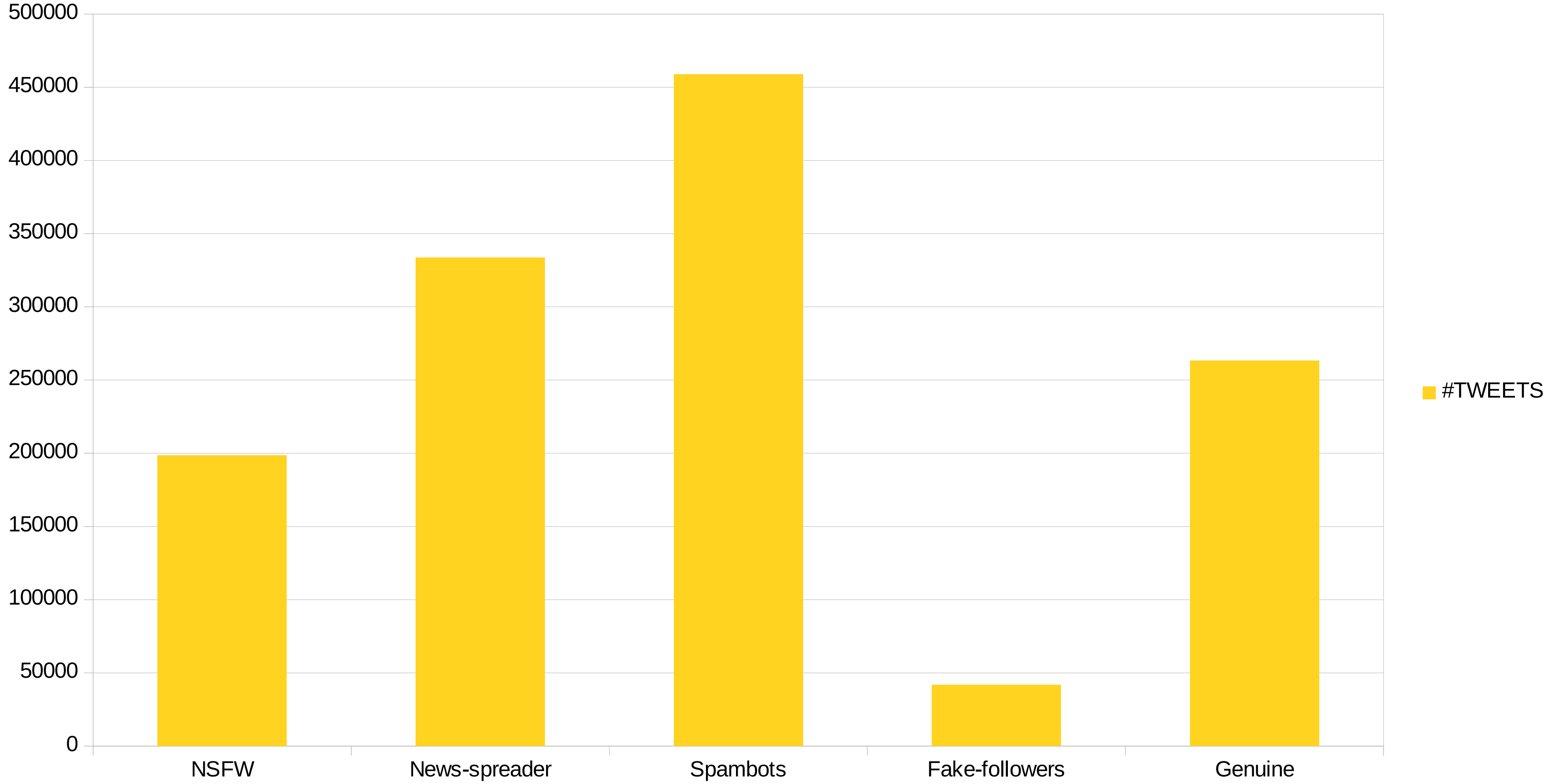
- ➔ Find fake-news tweets (Hoaxy)
- ➔ Find authors information (Twitter API)
- ➔ Check if authors are bots (Botometer)



Final dataset



Final dataset



Original Features

User Features

- ➔ Personal data
- ➔ Metadata
- ➔ Setting preferences

Tweet Features

- ➔ Text
- ➔ Media attached
- ➔ Mentions
- ➔ Source

Features Engineering

4 categories of features added

- ➔ Descriptive
- ➔ Intrinsic
- ➔ Extrinsic
- ➔ Image

Descriptive Features

- ”Meta features” related to tweets
- Synthesis statistics and counters

Max, min, avg of tweet lenghts	percentage of retweets, made by user, over its tweets
Max, min, avg number of retweets	percentage of media content incorpored in tweets
Max, min, avg number of likes	percentage of URL links placed inside tweets
Max, min, avg number of hashtag used	percentage of quotes, made by the user, over its tweets
amount of tweets per day (up to 100)	

Intrinsic Features

- ➔ **Related to monotony of the user**
- ➔ **euclidean distance from a TF-IDF-encoded centroid**

Tweets intradistance	Monotony coefficient about tweet texts
URL intradistance	Monotony coefficient about domain promoted

Extrinsic Features

- ➔ **Features related to common behaviours**
- ➔ **Creation of 5 different dictionaries**
- ➔ **Weighted words in dictionary**
- ➔ **Scores computed according to the use of words belonging to dictionaries**

NSFW words score
news spreders words score
spam bots words score
fake followers words score
genuine words score

Image Features

- ➔ **Related to profile image and attached media**
- ➔ **Scores computed with a CNN trained to indentify NSFW contents**

NSFW profile	NSFW evaluation of the profile picture
NSFW average	Average of NSFW scores of the last 10 tweeted media

Features Vector

The final feature vector contains 38 features

- ➔12 default user features
- ➔17 descriptive features
- ➔2 instrinsic features
- ➔5 extrinsic features
- ➔2 image features

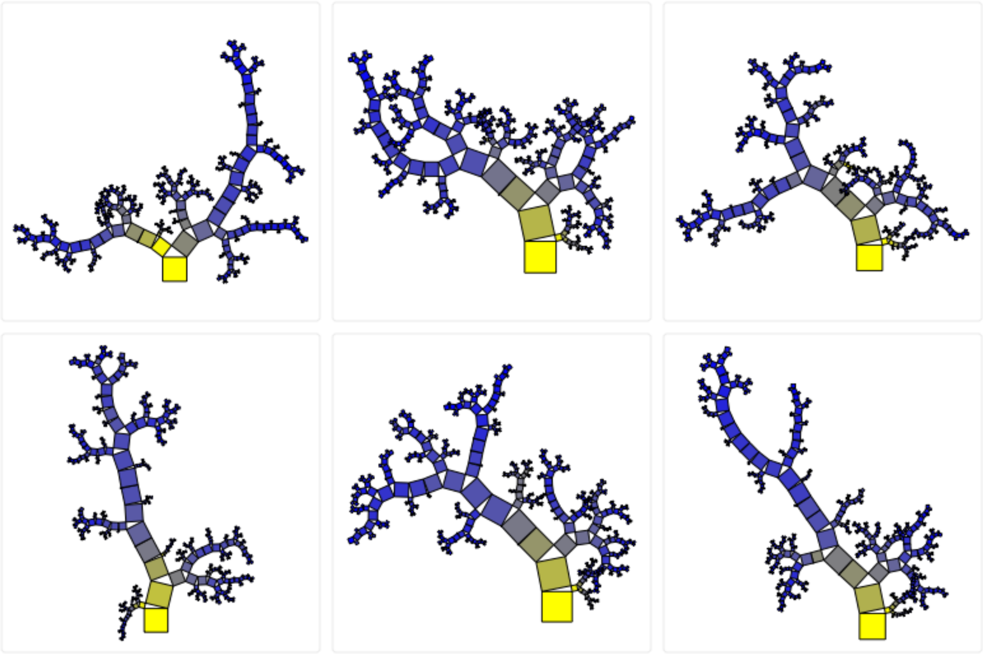
Models

The final solution is composed by a stacking ensemble of three different models

- ➔ Random Forest multiclass classifier for users
- ➔ Naive Bayes text clasifier for tweets
- ➔ Random Forest binary classifier (bot or not)

Multiclass classifier

- ➔ RandomForestClassifier
- ➔ Considers the whole features vector
- ➔ Parameters tuning performed with Grid Search
- ➔ max_depth = None
- ➔ n_estimators = 250
- ➔ criterion = “entropy”



- ➔ Cross validation score

  f1 = 0.946

Text classifier

- Every tweet is labeled with the target of its author
- Considers only tweet texts
- Based on unigrams
- Pipeline
  - Stemming
  - TF-IDF encoding
  - MultinomialNB

$$p(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k)p(C_k)}{p(\mathbf{x})}$$

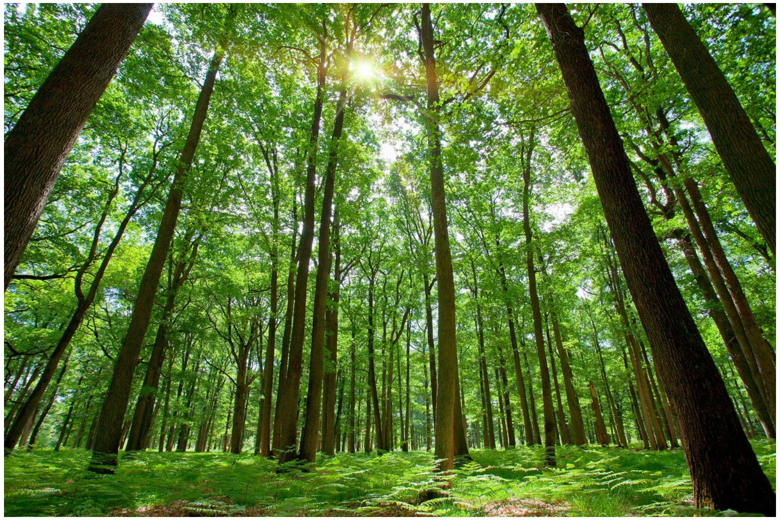
- Final prediction over user is computed with the average of the predictions of his tweets

- Cross validation score

  f1 = 0.71

Binary classifier

- ➔ RandomForestClassifier
- ➔ Considers the whole features vector except image and extrinsic features
- ➔ Parameters tuning performed with Grid Search
- ➔ max_depth = 26
- ➔ n_estimators = 175
- ➔ criterion = “entropy”

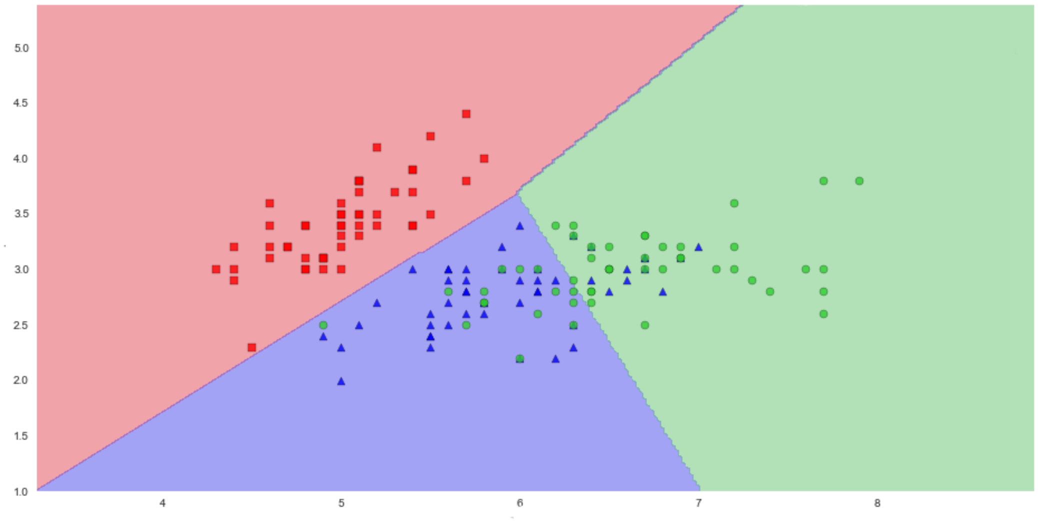


- ➔ Cross validation score

  AUC = 0.936

Ensemble

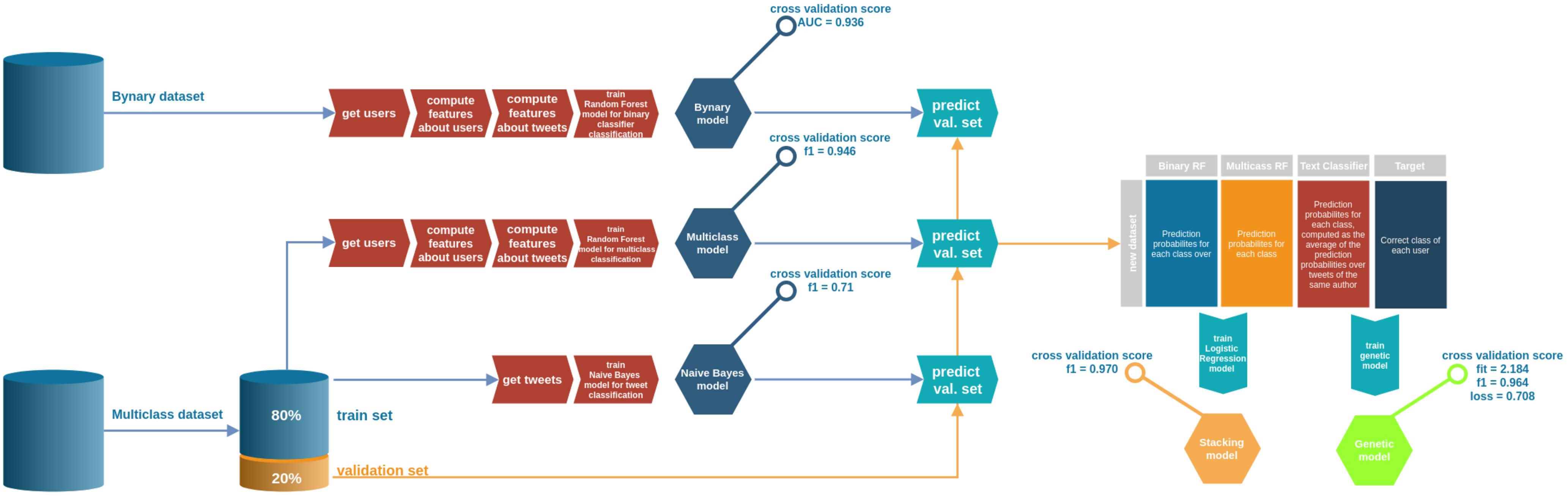
- Stacking with LogisticRegression
- Input data consist in the output probabilities of the other models
- max_iter = 100
- solver = “saga”
- class_weight = “balanced”
- multi_class = “multinomial”



- cross validation score

  f1 = 0.9734

Stacking



Bynary dataset
get users
compute features about users
compute features about tweets
train Random Forest model for binary classifier classification
Bynary model
cross validation score AUC = 0.936
predict val. set
Multiclass dataset
80%
train set
20%
validation set
train Random Forest model for multiclass classification
Multiclass model
cross validation score f1 = 0.946
get tweets
train Naive Bayes model for tweet classification
Naive Bayes model
cross validation score f1 = 0.71
Binary RF
Multicass RF
Text Classifier
Target
new dataset
Prediction probabilites for each class over
Prediction probabilites for each class
Prediction probabilites for each class, computed as the average of the prediction probabilities over tweets of the same author
Correct class of each user
train Logistic Regression model
train genetic model
cross validation score f1 = 0.970
Stacking model
cross validation score fit = 2.184 f1 = 0.964 loss = 0.708
Genetic model

Web App

Web service that allows users to classify twitter accounts

Frontend development

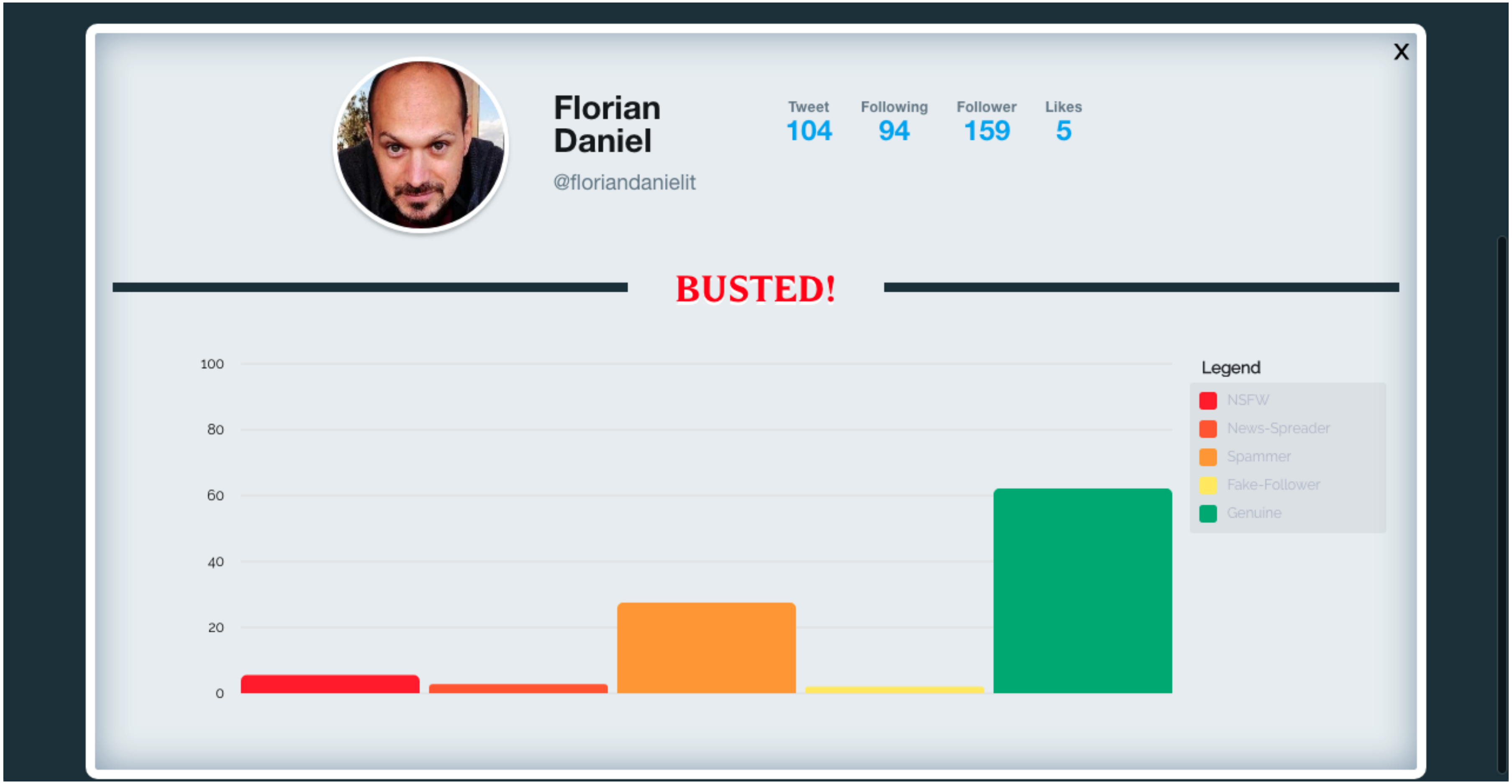
- ➔ AngularJS

Backend development

- ➔ Flask

BotBuster

BotBuster


X
Florian Daniel
@floriandanielit
Tweet 104
Following 94
Follower 159
Likes 5
BUSTED!
100
80
60
40
20
0
Legend
NSFW
News-Spreader
Spammer
Fake-Follower
Genuine

Thanks for your attention