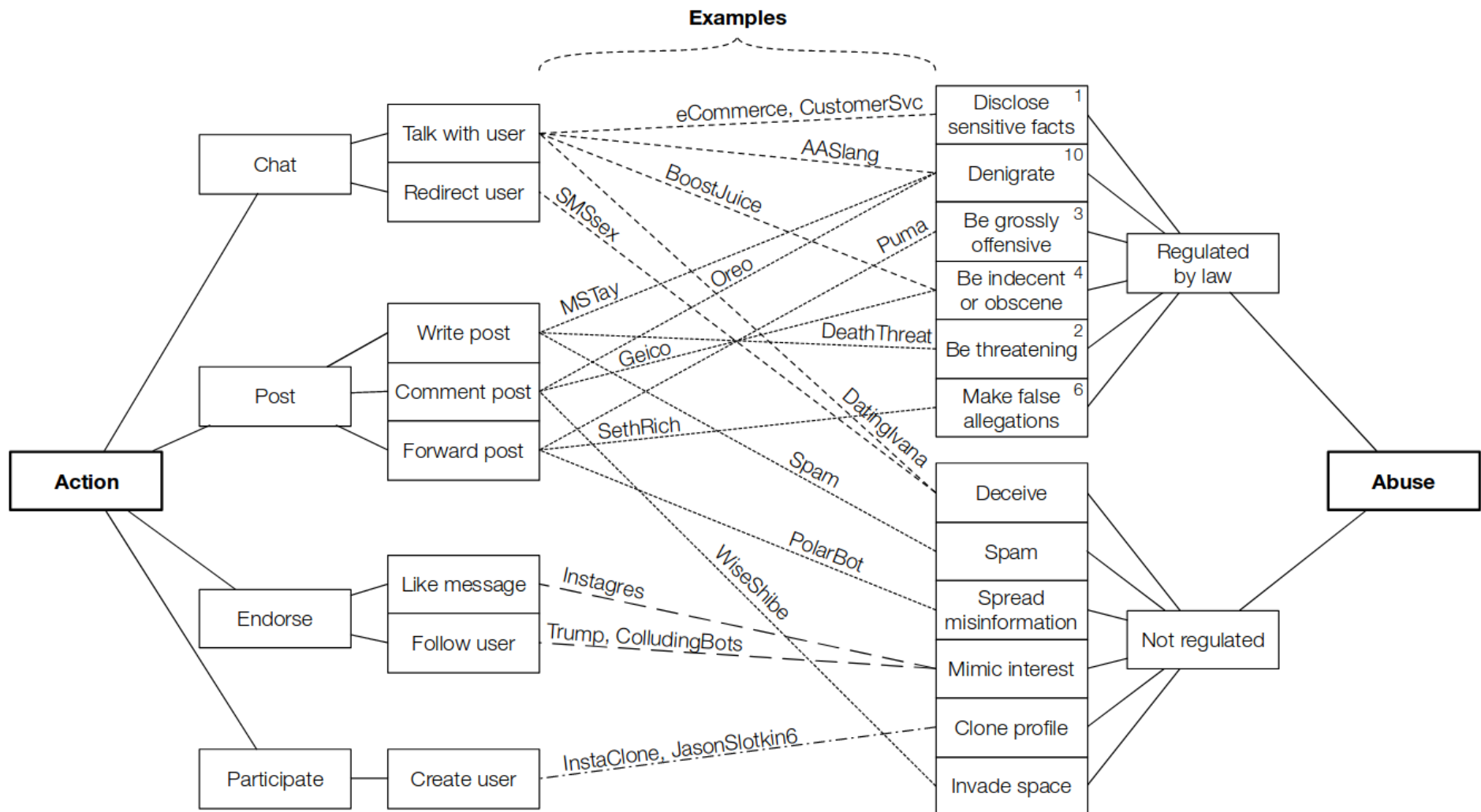




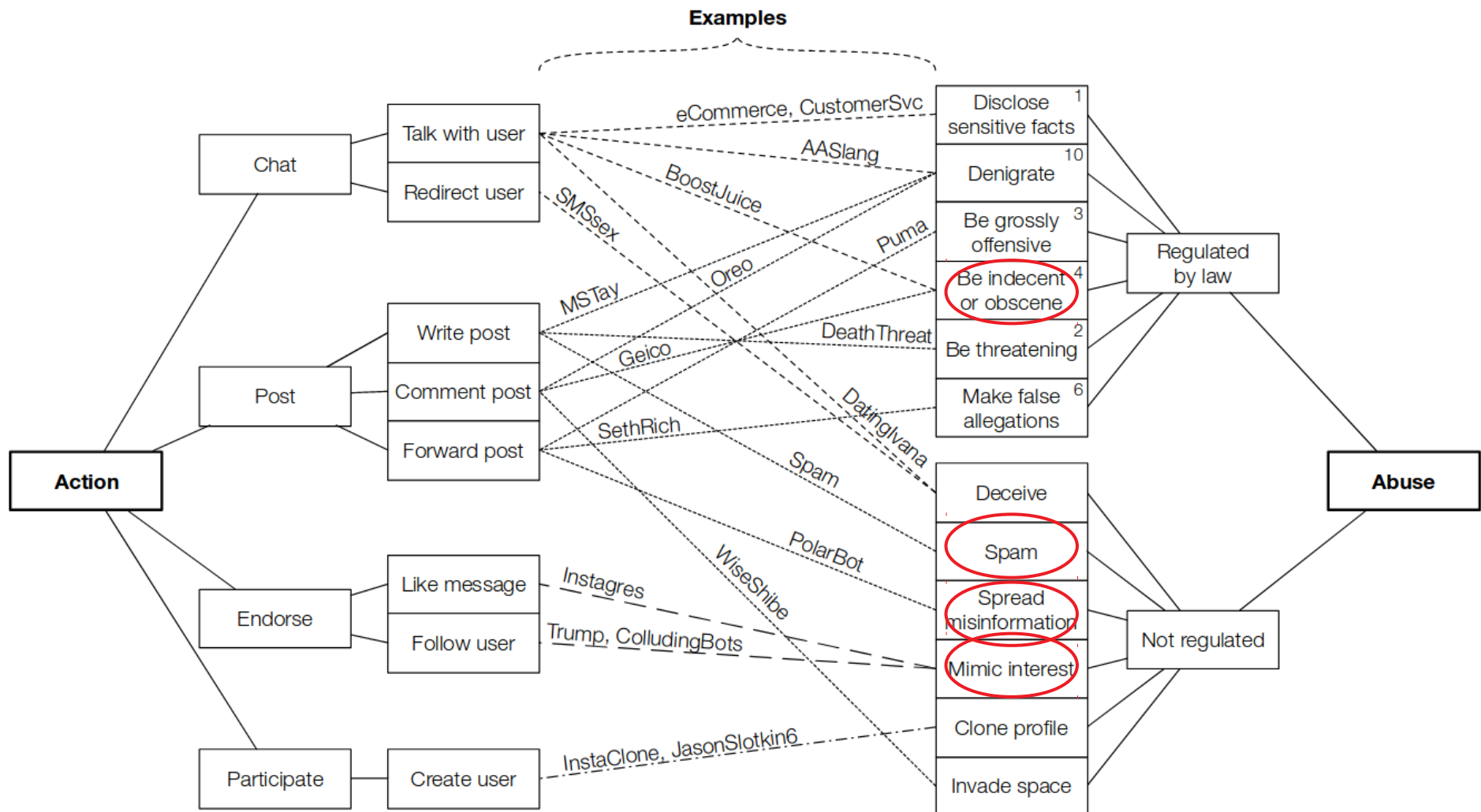
POLITECNICO
MILANO 1863

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

Social bots



Twitter bots



Twitter bots

- **NSFW**
- **News-spreaders**
- **Spambots**
- **Fake-followers**
- ✓ **Genuine**

NSFW

The image shows a Twitter profile for Elizabeth Little (@ns_mia). The profile picture is a circular image of a woman in a red shirt and denim shorts, holding a phone. The header has a pink background. The bio says "Do you like fast? Come in!" with three heart-eye emojis and a link to go9to.pro/ns_mia4AuR. The location is Novi Sad and the join date is May 2010. The tweet section shows a pinned tweet from 1 March with the text "Hello! How are you? Rate my site in bio!" and three smiley face with hearts emojis. The tweet has 79 replies, 16 followers, and 375 likes.

Elizabeth Little
@ns_mia

Do you like fast? Come in! 🥰🥰🥰
go9to.pro/ns_mia4AuR

📍 Novi Sad

📅 Iscrizione a maggio 2010


Tweet **79** **Follower 16** **Mi piace 375**

Tweet **Tweet e risposte** **Contenuti**

Elizabeth Little @ns_mia · 1 mar
Hello! How are you? Rate my site in bio! 😊😊😊

🌐 Traduci il Tweet







News-spreaders



mary susan evans
@msevans034
Iscrizione a novembre 2016

Invia Tweet

12 foto e video



Tweet140.000

Following5.001

Follower1.104


MI place106.000

Tweet

Tweet e risposte

Contenuti


mary susan evans ha ritwittato

**Steve Vladeck** @steve_vladeck · 18 lug
Here's Fifth Circuit Judge Jim Ho's thoughtful @GB2d article on the subject:
[gibsondunn.com/wp-content/upl...](#)
Traduci il Tweet

11340

Mostra questa discussione


mary susan evans ha ritwittato

**Steve Vladeck** @steve_vladeck · 18 lug
One can argue that #SCOTUS got it wrong, but it would take the Court overruling itself or a constitutional amendment to change that rule; doing it by Executive Order would be unconstitutional.
Traduci il Tweet

51472

Mostra questa discussione

mary susan evans ha ritwittato

**Ryan Knight** @ProudResister · 11 h
The mayor asked Trump not to visit while they are "burying the dead" and Jewish leaders told him to stay out until he stops "endangering minorities."
[@realDonaldTrump](#) is not respecting their wishes. He is traveling to Pittsburgh to visit the scene of the crime that he incited.
Traduci il Tweet

Spambot

English Job Portugal
@PortugalEnglish
English Speaking Jobs in Portugal
Portugal
toplanguagejobs.com.pt/en/language/En...
Iscrizione a marzo 2011

Tweet 4.020 **Followers** 32

Tweet e risposte

English Job Portugal @PortugalEnglish · 2 nov 2016
English Customer Care Representative
Traduci il Tweet

English Customer Care Representative
Do you like to help people? Do you see yourself working in a big multicultural company? We are currently recruiting Customer Service Specialists for our project...
toplanguagejobs.co.uk

English Job Portugal @PortugalEnglish · 2 nov 2016
Technical Customer Care with Spanish and Portuguese bit.ly/2f9QPMY
Traduci il Tweet

English Job Portugal @PortugalEnglish · 1 nov 2016
Veritas Sales Executive - Dutch bit.ly/2f91Ex3
Traduci il Tweet

English Job Portugal @PortugalEnglish · 1 nov 2016
Social Media Support Specialist - Russian
Traduci il Tweet

Social Media Support Specialist - Russian
Social Media Support Specialist - Russian Speaking
Location: Belfast, United Kingdom About the Role: The Social Media Support Specialist will engage with cust...

Fake-followers

A screenshot of a Twitter profile for a user named Savinraj (@savinthegreat). The profile header has a blue background. The profile picture is a circular image of a lake reflecting mountains and a cloudy sky. To the right of the profile picture, it shows 'Following 331' and 'Follower 5'. Below the profile picture, the name 'Savinraj' is displayed in bold, followed by the handle '@savinthegreat'. Below the handle, it says 'Iscrizione a giugno 2009' with a calendar icon. A blue button labeled 'Invia Tweet' is positioned below the bio. To the right of the bio, a message states '@savinthegreat non ha twittato' (Savinraj has not tweeted) with a subtext 'Quando lo farà, i suoi Tweet verranno mostrati qui.' (When he/she does, his/her tweets will be shown here).

Following
331

Follower
5

Savinraj
@savinthegreat

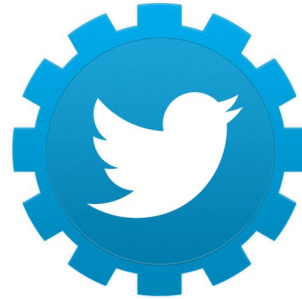
Iscrizione a giugno 2009

Invia Tweet

@savinthegreat non ha twittato
Quando lo farà, i suoi Tweet verranno mostrati qui.

Tools

→ **Twitter API**



→ **Botometer API**



→ **Hoaxy API**



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

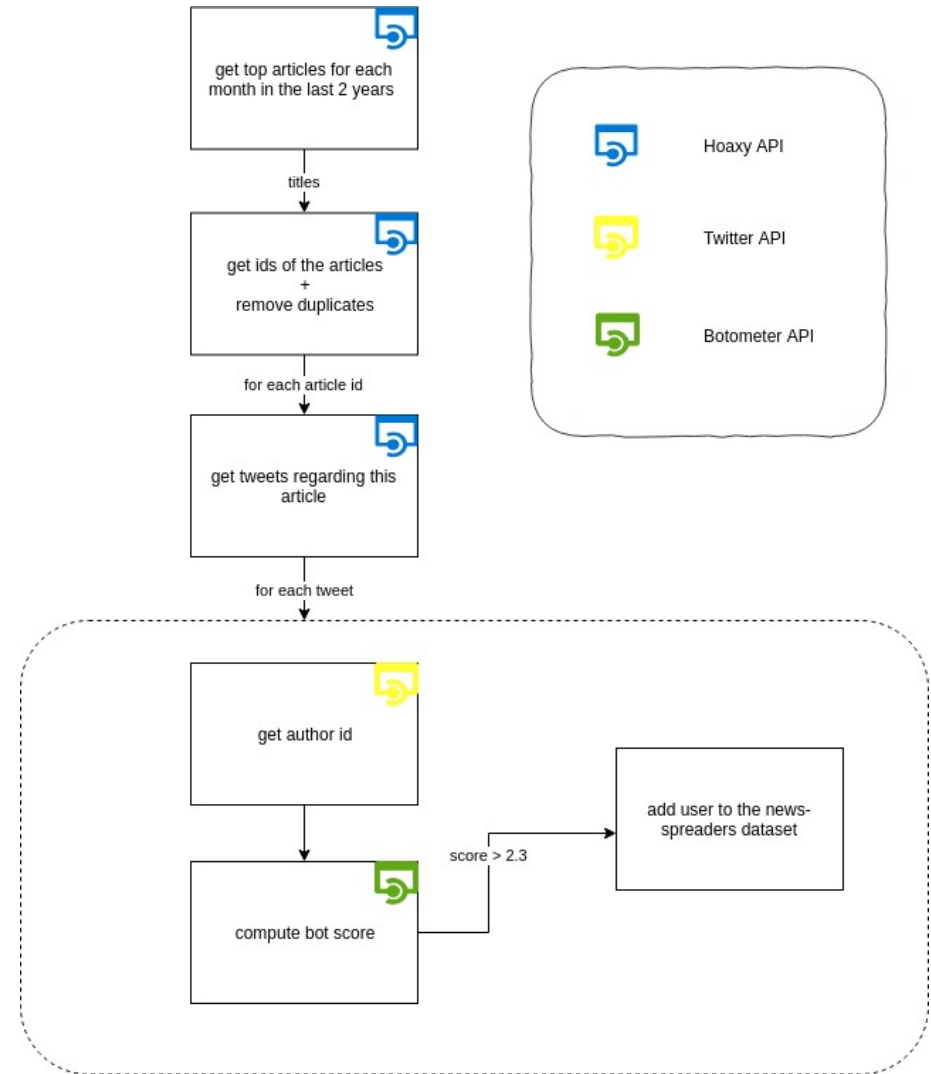
From Cresci dataset:

→ Genuine

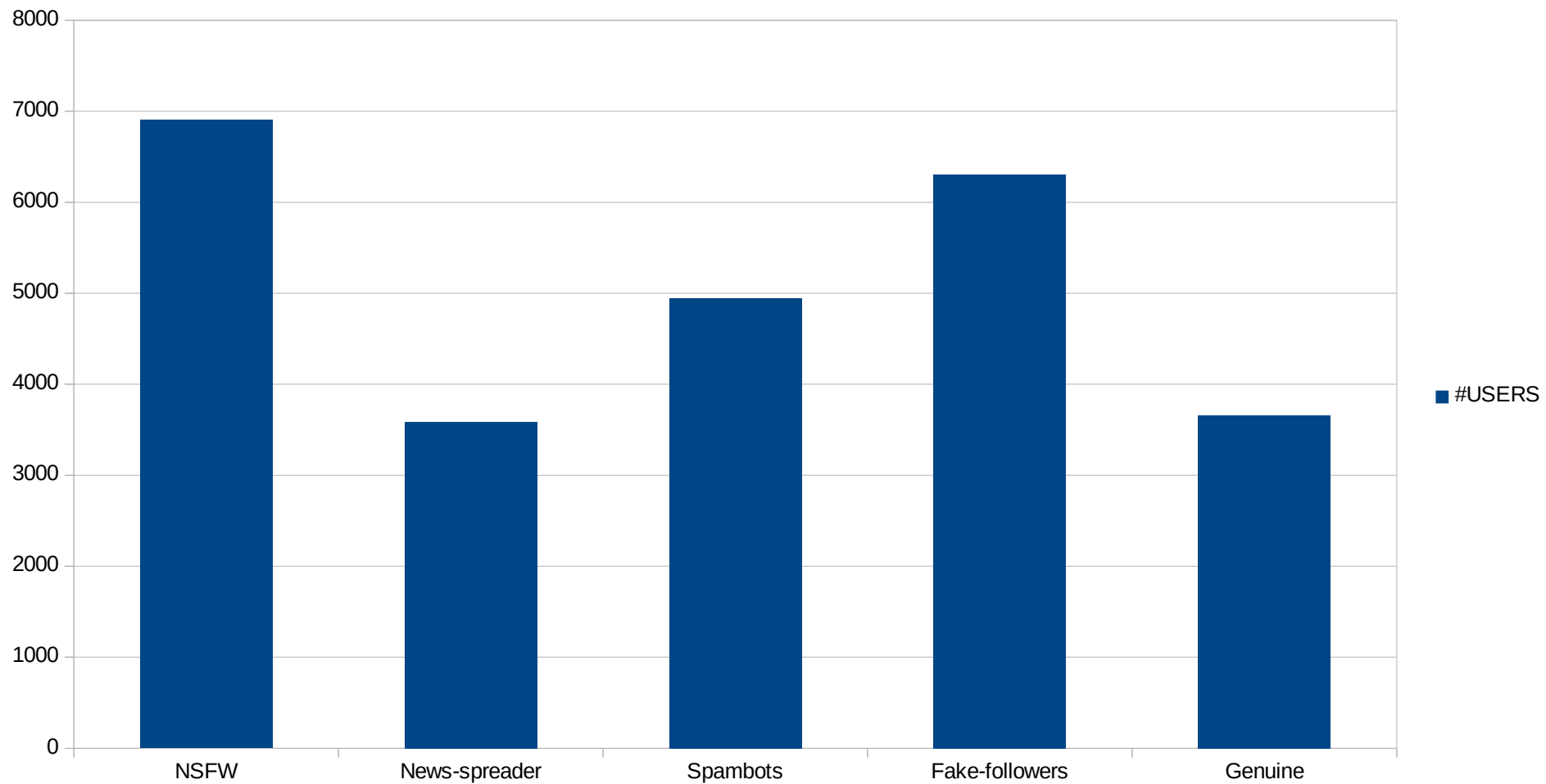
Data Collection

News-spreaders

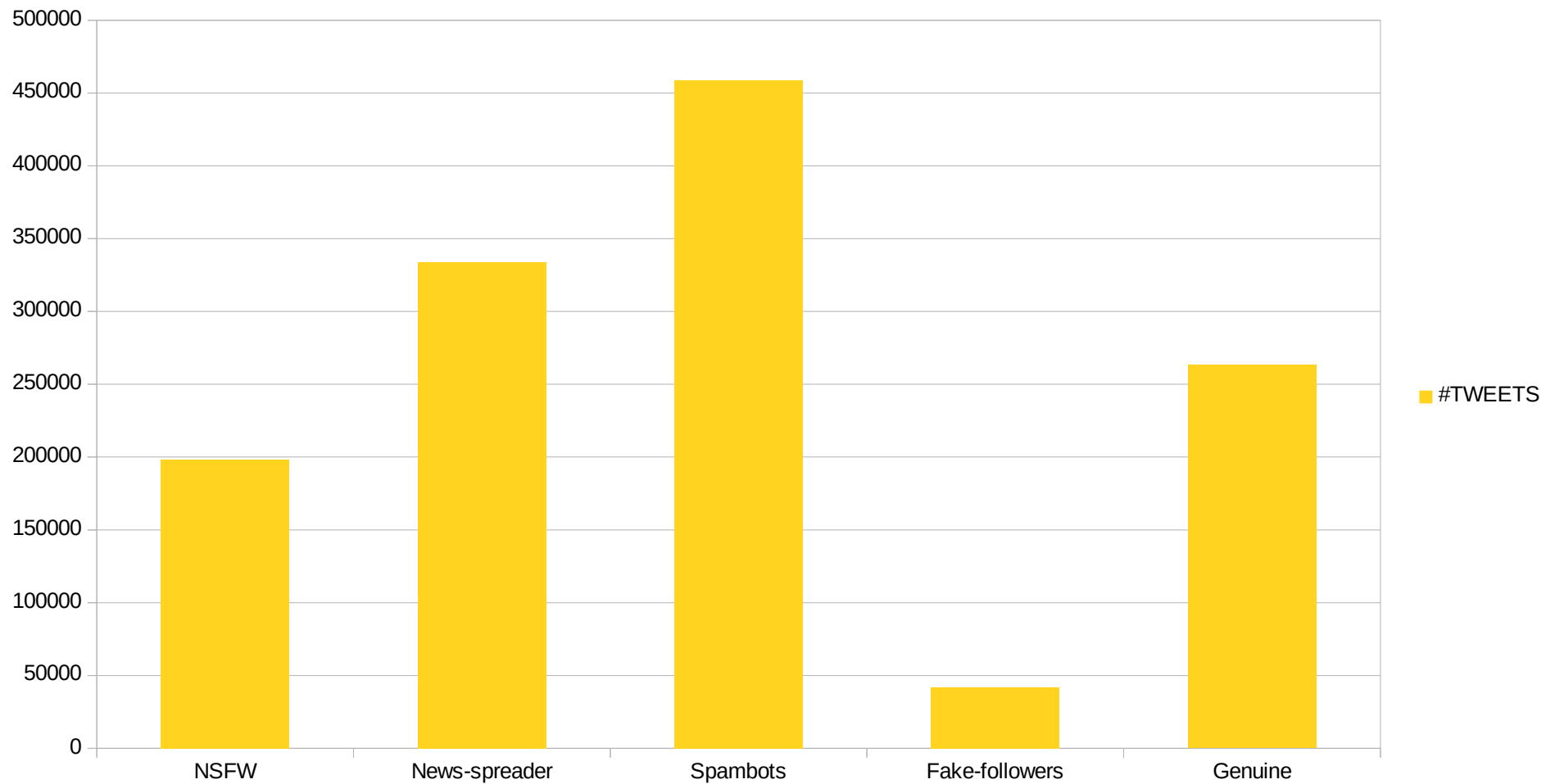
- Find fake-news tweets (Hoaxy)
- Find authors information (Twitter API)
- Check if authors are bots (Botometer)



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 lengths	percentage of retweets, made by user, over its tweets
Max, min, avg number of retweets	percentage of media content incorporated 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 spreaders 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 identify 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 intrinsic features
- 5 extrinsic features
- 2 image features

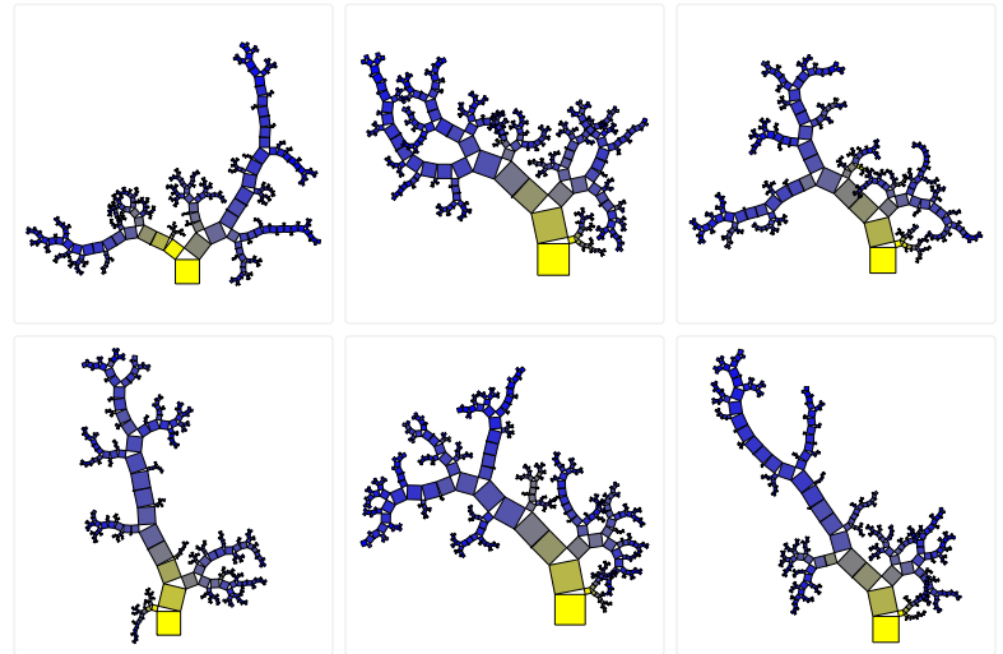
Models

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

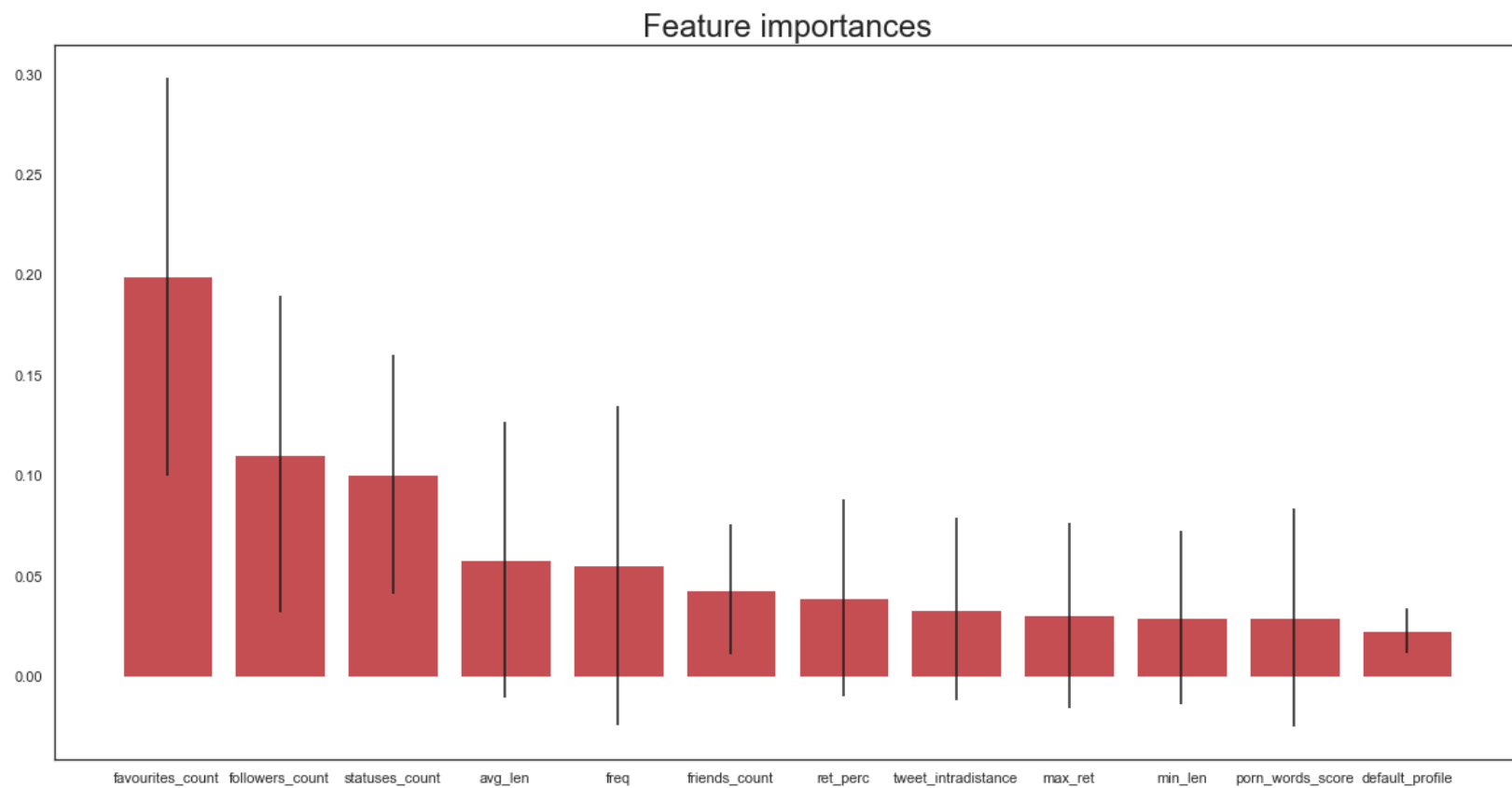
- Random Forest multiclass classifier for users
- Naive Bayes text classifier for tweets
- Random Forest binary classifier (bot or not)

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
 $f1 = 0.946$



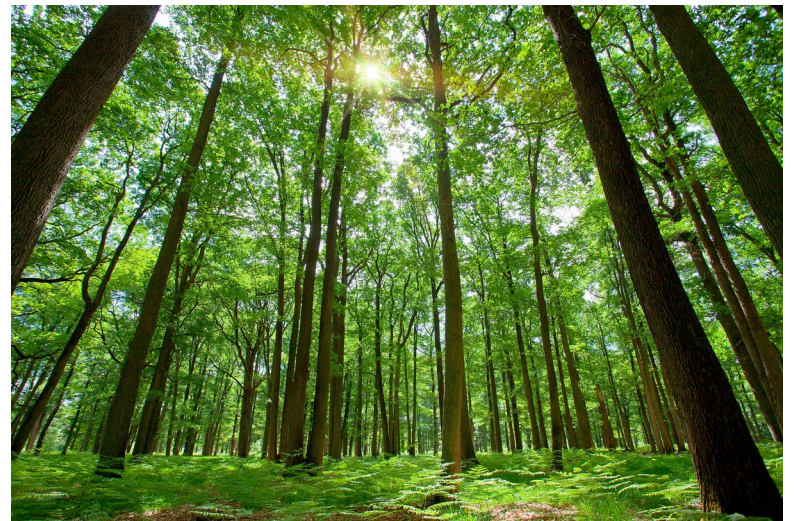
Features Ranking



Binary classifier

- RandomForestClassifier
- Considers the whole features vector except image and extrinsic features
- Parameters tuning performed with Grid Search
- `max_depth = None`
- `n_estimators = 250`
- `criterion = "entropy"`

- Cross validation score
AUC = 0.936



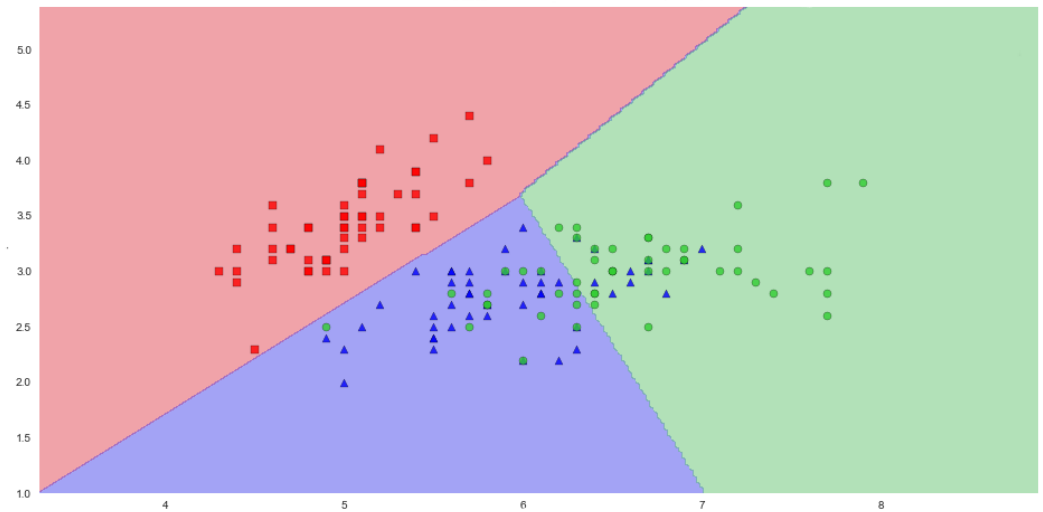
Text classifier

- Every tweet is labeled with the target of its author
- Considers only tweet texts
- Based on unigrams
- Pipeline
 - Stemming
 - TF-IDF encoding
 - MultinomialNB
- Final prediction over user is computed with the average of the predictions of his tweets
- Cross validation score
 - f1 = 0.71

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

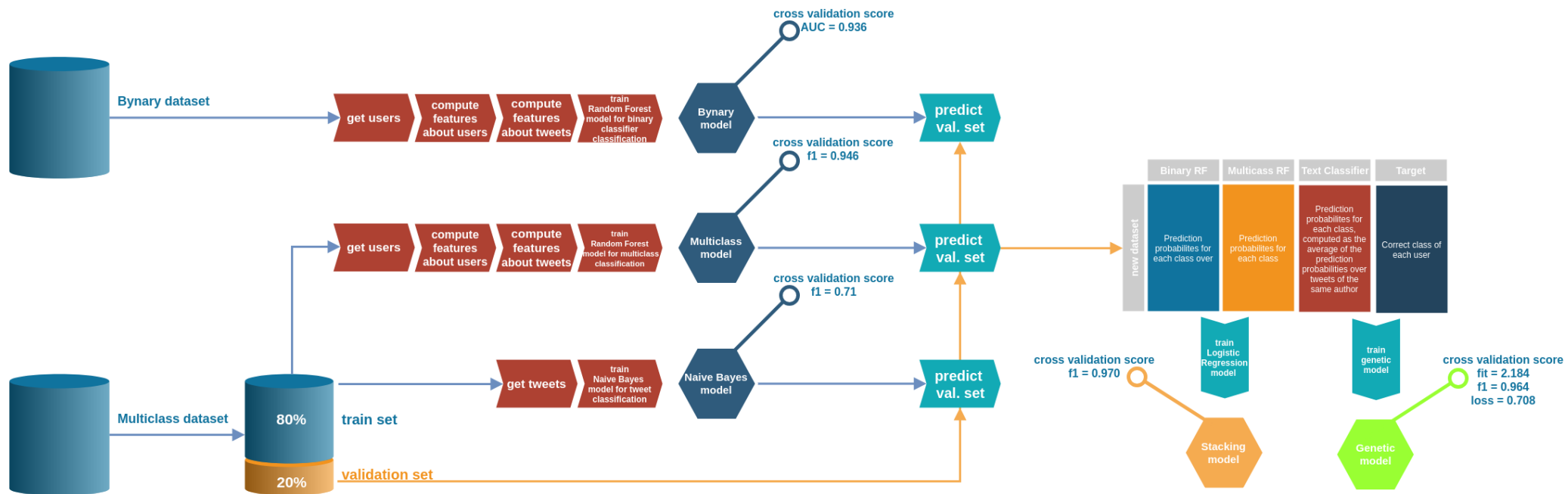
Ensemble

- 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"`
- L1 penalty

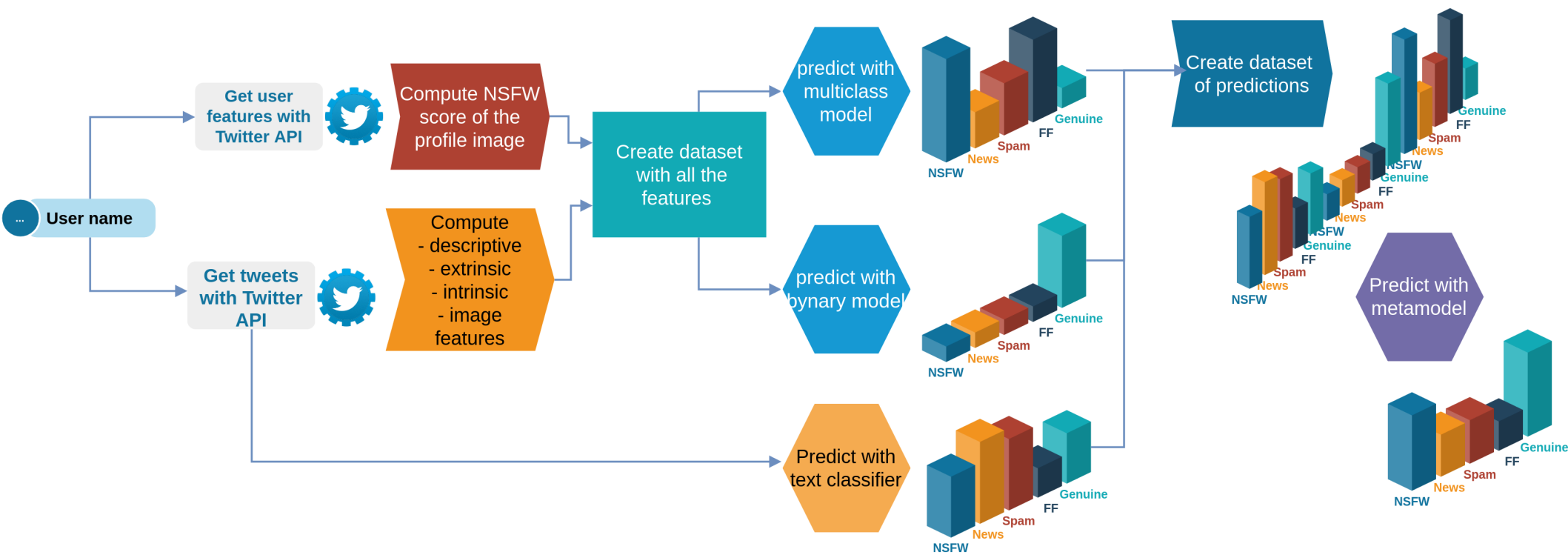


- cross validation score
 $f1 = 0.9734$

Stacking



Prediction



Web App

Web service that allows users to classify twitter accounts

Frontend development

→ AngularJS



Backend development

→ Flask



Flask

BotBuster

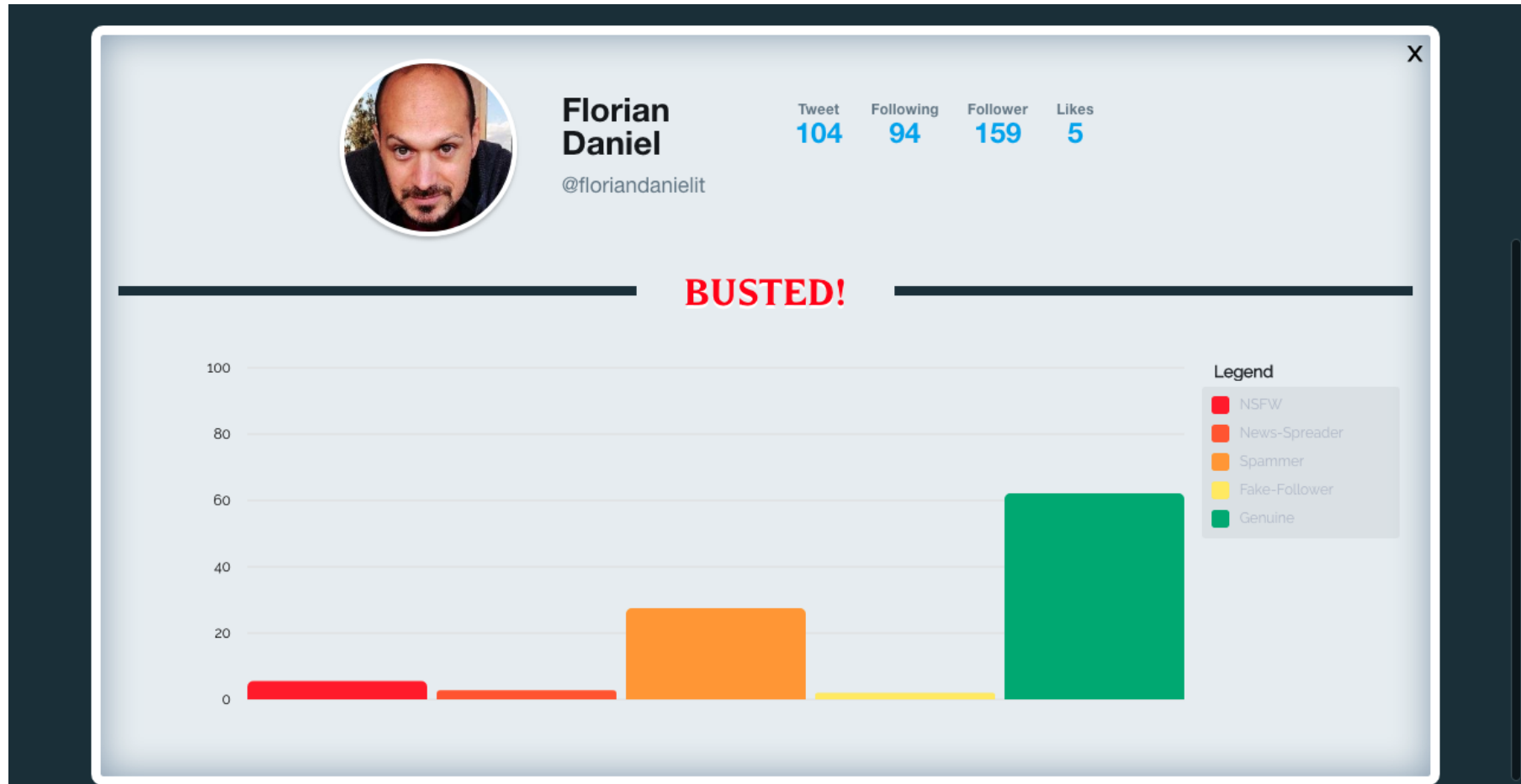


BOTBUSTER

Enter Twitter username

BUST IT

BotBuster



Thanks for your attention