Key User Profile:

Data to collect:
Name
Screen name
Description
Twitter verification label
Profile picture
Number of followers
Number of friends
Date of creation
Number of Tweets
Number of Favorites

Preparation of data:
Count the number of characters in the name
Count the number of characters in the screen name
Try to find the substring “bot” in the name or the screen name
Remove all the space and underscore in both name and screen name
Compare the similarity of the name and the screen name
Count the number of characters in the description
Make a ratio of friends number on followers number
Transform the creation date to the age in day of the account
Calculate the number of tweets posted by day

List of subindex:
- Similarity of name and screen name
- Number of digits in the screen name
- Length of the name
- Length of the screen name
- Length of the description
- Age
- Tweets by day
- Favorite
- Profile picture
- Ratio of friends/followers

Calcul of the probability:
- If the account is veriried by Twitter, the final result will simply be always 0
- Base value of each subindex : 0,15
- Similarity score is calculated by comparing the number of letters and each letters in common, range is between 0 and 1. If the name or screen name contains the substring “bot”, this value will be 1
- Number of digits : If it is more than 2, the number of digit is multiplacted by 0,12, maximum is 1, if not, the base value is kept
- Length of name : If it is more than 15, the length is multiplacted by 0,009, maximum is 1, if not, the base value is kept
- Length of screen name : If it is more than 10, the length is multiplacted by 0,012, maximum is 1, if not, the base value is kept
- Length of description : If it is less than 10, the length is multiplacted by 0,1 and removed from 1, minimum is 0, if not, the base value is kept

- Age : If it is more than 90 (3 months), the age is multiplicated by 0,001 and removed from 1, maximum is 0, if not, 1 is used
- Tweets by day : Number of tweets by day multiplicated by 0,01 (no maximum limit)
- Favorite : The number of favorites is multiplated by 0,01 and removed from 1, minimum at 0
- Profile picture : If a profile picture exist, the score is the base score, if not, it is 10
Ratio of friends/followers : We remove the ratio from 1, minimum at 0

Then, an average of all the subindex is calculated and used for the User score.

More explanations:
A Twitter verified account (blue label) means that the account was verified by an human agent, who confirmed that the account is authentic. The accout had to provide a lot of information and validate it with a phone number. We can assume that an expert for evaluate an accout will always be better than an algorithm for spot if an account is an human or a bot.

The base value could be 0, but it is means that we start by considering all profiles as humans. This is not really good. It could be 1, so we start by assuming all account checked are bot, it is worst. The official number from Twitter is 15% of accounts are bot. So we use this base value (0,15) for our calculations

Most of the time, humans will create a screen name similar to its name. A bot can, sometimes, generate two names totally different for this. There have some official bot that put “bot” in their name. If it is true, Similarity will be equal to 1.

A long name or screen name is suspicious, an human does not need so much character usually. But if the bot generate a random name, it can be long sometimes.

For the description, this is opposite, a bot will be more likely to make an empty description, or really short, than a long one

The usual number of digits (if used) in a name, is usually two for an human, that can represent the two firsts number of a zip code, an age, or a birth year. More than 2, it becomes suspicious, and can let think that the name was randomly generated

If you need to check a new Twitter account (less than 3 month) we can consider that account suspicious. So we use the value of 1 (surely a bot) for the average, after 3 months. The score start to decreases, with a minimal of 0, for each day more in the age.

Tweets by day is the only score that have no maximal, if an user can post more than 5 000 tweets in one day, it is almost surely a bot, so the maximal score is not limited for influence more the average

Most of bot don’t have a profile picture, so the value of 1 is used in case of no image.

There have many kind of bots, sometimes they are working alone and just try to follow a lot of people to get a lot of followers in return and be more influent. Sometimes, the bots are automatically followed by all the others bot of the same network. A normal human user have usually a number of followers close to the number of friends. The most the ratio is far from 1, the most the account is considerate as bot

Test / Examples:

Profile	Our result	Botometer user index result	Reality
@sunneversets100	68%	61%	Bot
@Betelgeuse_3	42%	47%	Bot
@infinite_scream	41%	44%	Bot
@tinycarebot	45%	35%	Bot
@EarthquakesSF	28%	35%	Bot
@KookyScrit	18%	19%	Bot
@factbot1	38%	38%	Bot
@IvanDuque	0%	48%	Human
@leorugeles	23%	16%	N/A
@roiberhol	39%	56%	N/A
@jocamacho10	21%	32%	N/A
@LaGallo92	15%	15%	N/A
@lolitotani	25%	19%	N/A
@Ricardo57517052	60%	58%	N/A
@AnavortizV	41%	56%	N/A
@tavo1283	44%	21%	N/A
@Andresf24091961	60%	56%	N/A
@seperbupe	32%	52%	N/A
@Jguarin1014	28%	29%	N/A
@OmarLop18383024	62%	59%	N/A

Most of Bot tested are known bot from some official page, Their profile are was, normally, completed for an human.

Improving possibilities:
- Improve the calculation of the average by including a coefficient to each subindex, for this, it is required to choose which one is more important than another for determinate if a profile is a bot or not
- Using external API for try to find the profile picture of an user on Internet, It can be a picture of a public place (for example Eiffel Tower), from an image database (for example Shutterstock), or from another existing profile on social network
- Using external API for image recognition, if the profile picture is a woman, but the account have a male name/screen name. It looks more like a bot. It is common from a bot who generate everything to find this.
- More user information to analyze
- Improving the calculation of each subindex