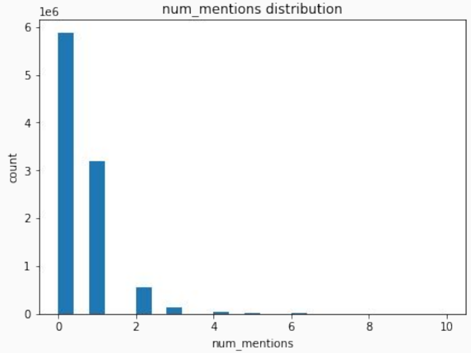
Data Mining project

A.Y. 22/23

Group 12
Leonardo Vona

Data understanding & preparation

- Most of the features have a very high skewed distribution.



- Two tweets were considered equal if they share the same value for *user_id*, *created_at*, and *text*

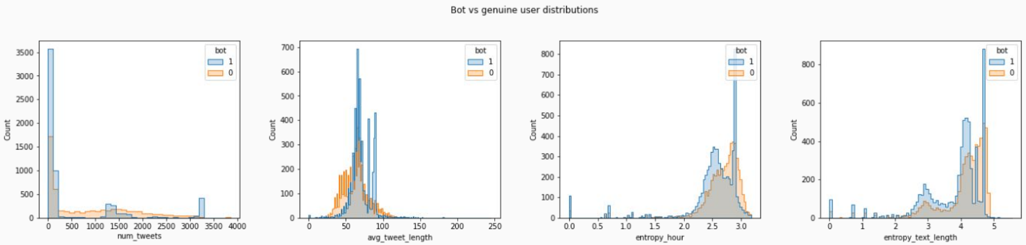
- The features were checked for correctness by also using external sources

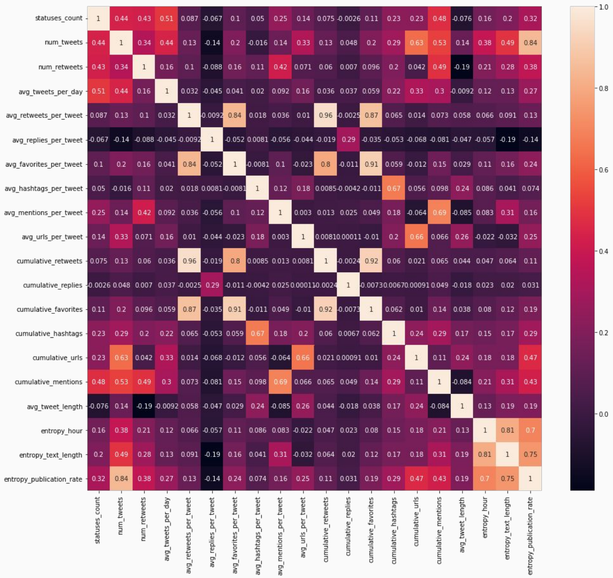
- For most cases, the missing values have been substituted after data cleaning with the median

- The tweets data frame has been separated into tweets and retweets

Data understanding & preparation: new features

The new features have been extracted by joining the users and tweets data frames. For each user, indicators were extracted about its behavior in terms of activity and entropy.


Bot vs genuine user distributions



	statuses_count	num_tweets	num_retweets	avg_tweets_per_day	avg_retweets_per_tweet	avg_replies_per_tweet	avg_favorites_per_tweet	avg_hashtags_per_tweet	avg_mentions_per_tweet	avg_urls_per_tweet	cumulative_retweets	cumulative_replies	cumulative_favorites	cumulative_hashtags	cumulative_urls	cumulative_mentions	avg_tweet_length	entropy_hour	entropy_text_length	entropy_publication_rate
statuses_count	1	0.44	0.43	0.51	0.087	-0.067	0.1	0.05	0.25	0.14	0.075	-0.0026	0.11	0.23	0.23	0.48	-0.076	0.16	0.2	0.32
num_tweets	0.44	1	0.34	0.44	0.13	-0.14	0.2	-0.016	0.14	0.33	0.13	0.048	0.2	0.29	0.63	0.53	0.14	0.38	0.49	0.84
num_retweets	0.43	0.34	1	0.16	0.1	-0.088	0.16	0.11	0.42	0.071	0.06	0.007	0.096	0.2	0.042	0.49	-0.19	0.21	0.28	0.38
avg_tweets_per_day	0.51	0.44	0.16	1	0.032	-0.045	0.041	0.02	0.092	0.16	0.036	0.037	0.059	0.22	0.33	0.3	-0.0092	0.12	0.13	0.27
avg_retweets_per_tweet	0.087	0.13	0.1	0.032	1	-0.0092	0.84	0.018	0.036	0.01	0.96	-0.0025	0.87	0.065	0.014	0.073	0.058	0.066	0.091	0.13
avg_replies_per_tweet	-0.067	-0.14	-0.088	-0.045	-0.0092	1	-0.052	0.0081	-0.056	-0.044	-0.019	0.29	-0.035	-0.053	-0.068	-0.081	-0.047	-0.057	-0.19	-0.14
avg_favorites_per_tweet	0.1	0.2	0.16	0.041	0.84	-0.052	1	-0.0081	0.1	-0.023	0.8	-0.011	0.91	0.059	-0.012	0.15	0.029	0.11	0.16	0.24
avg_hashtags_per_tweet	0.05	-0.016	0.11	0.02	0.018	0.0081	-0.0081	1	0.12	0.18	0.0085	-0.0042	-0.011	0.67	0.056	0.098	0.24	0.086	0.041	0.074
avg_mentions_per_tweet	0.25	0.14	0.42	0.092	0.036	-0.056	0.1	0.12	1	0.003	0.013	0.025	0.049	0.18	-0.064	0.69	-0.085	0.083	0.31	0.16
avg_urls_per_tweet	0.14	0.33	0.071	0.16	0.01	-0.044	-0.023	0.18	0.003	1	0.0081	0.00011	-0.01	0.2	0.66	0.066	0.26	-0.022	-0.032	0.25
cumulative_retweets	0.075	0.13	0.06	0.036	0.96	-0.019	0.8	0.0085	0.013	0.0081	1	-0.0024	0.92	0.06	0.021	0.065	0.044	0.047	0.064	0.11
cumulative_replies	-0.0026	0.048	0.007	0.037	-0.0025	0.29	-0.011	-0.0042	0.025	0.00011	-0.0024	1	-0.0073	0.0067	0.00091	0.049	-0.018	0.023	0.02	0.031
cumulative_favorites	0.11	0.2	0.096	0.059	0.87	-0.035	0.91	-0.011	0.049	-0.01	0.92	-0.0073	1	0.062	0.01	0.14	0.038	0.08	0.12	0.19
cumulative_hashtags	0.23	0.29	0.2	0.22	0.065	-0.053	0.059	0.67	0.18	0.2	0.06	0.0067	0.062	1	0.24	0.29	0.17	0.15	0.17	0.29
cumulative_urls	0.23	0.63	0.042	0.33	0.014	-0.068	-0.012	0.056	-0.064	0.66	0.021	0.00091	0.01	0.24	1	0.11	0.24	0.18	0.18	0.47
cumulative_mentions	0.48	0.53	0.49	0.3	0.073	-0.081	0.15	0.098	0.69	0.066	0.065	0.049	0.14	0.29	0.11	1	-0.084	0.21	0.31	0.43
avg_tweet_length	-0.076	0.14	-0.19	-0.0092	0.058	-0.047	0.029	0.24	-0.085	0.26	0.044	-0.018	0.038	0.17	0.24	-0.084	1	0.13	0.19	0.19
entropy_hour	0.16	0.38	0.21	0.12	0.066	-0.057	0.11	0.086	0.083	-0.022	0.047	0.023	0.08	0.15	0.18	0.21	0.13	1	0.81	0.7
entropy_text_length	0.2	0.49	0.28	0.13	0.091	-0.19	0.16	0.041	0.31	-0.032	0.064	0.02	0.12	0.17	0.18	0.31	0.19	0.81	1	0.75
entropy_publication_rate	0.32	0.84	0.38	0.27	0.13	-0.14	0.24	0.074	0.16	0.25	0.11	0.031	0.19	0.29	0.47	0.43	0.19	0.7	0.75	1

Clustering

Clustering techniques

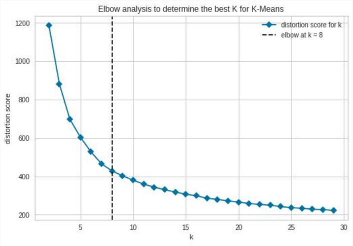
- KMeans
- Density-based
- Hierarchical
- XMeans (pyclustering)
- BSAS (pyclustering)

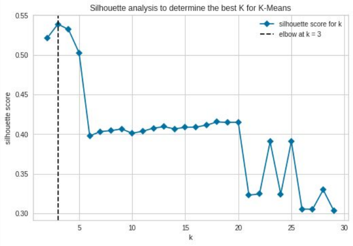
Preprocessing

- Extract numerical features from the dataset describing the user behaviour
- Remove highly correlated features
- Apply normalization

Clustering: KMeans

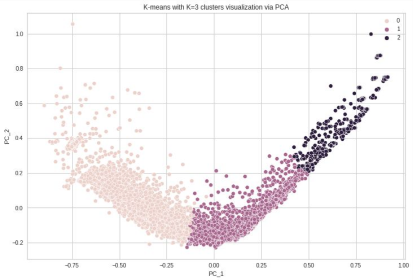
The best K has been determined using the elbow and silhouette methods.

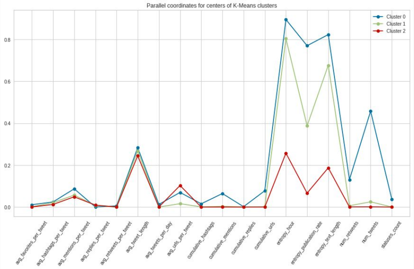




By using SSE, Davies-Bouldin, Silhouette and Calinski-Harabasz scores, the K value of 3 has been chosen from the candidates.

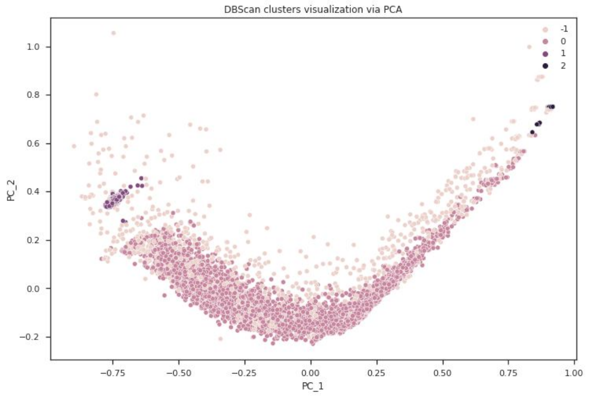
- Cluster 0 groups the most active users
- Cluster 1 contains medium active users
- Cluster 2 contains users with a scarce activity

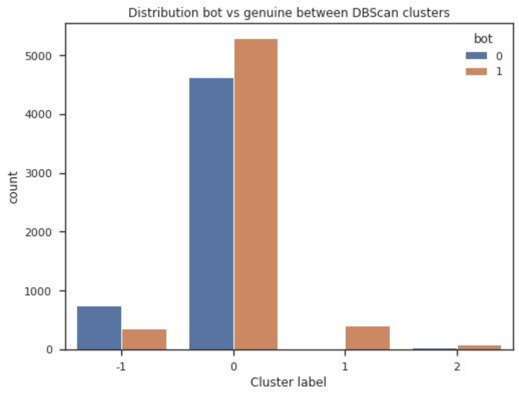




Clustering: Density-based

A grid search is performed to establish the best combination of *minPts* and *eps* parameters.

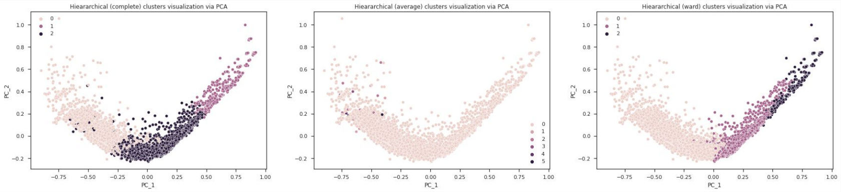


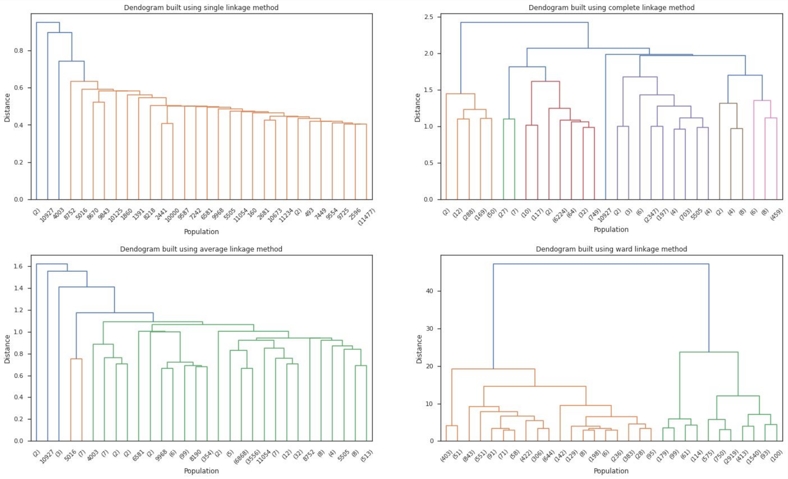


Density based clustering is not particularly suited for our dataset.

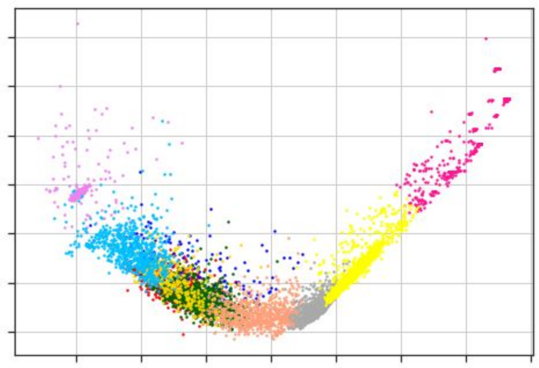
Clustering: Hierarchical

The Complete and Ward linkage methods seem the best options.



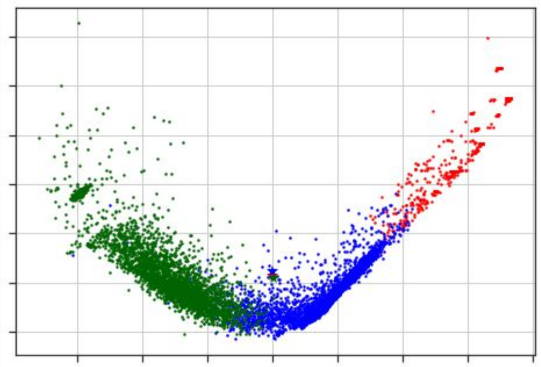


Clustering: pyclustering



X-Means clustering tends to split the data in the maximum number of clusters, because it gives the best BIC score.

BSAS clustering makes a good separation of the data, and is particularly suited when the data is given as a stream.



Classification

Preprocessing

- Drop *id* and *name*, exclude *bot* attribute
- Group small sized languages together
- Extract new features from *created_at*

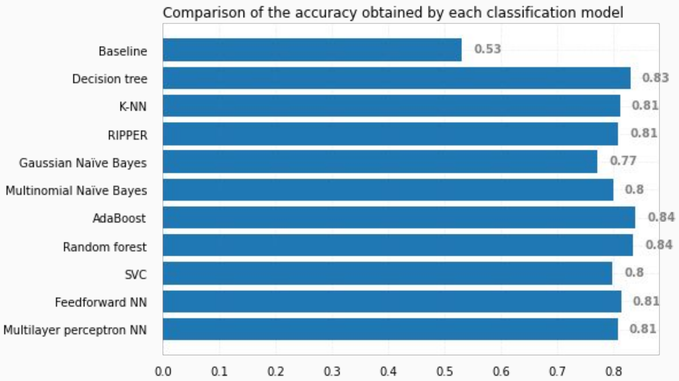
Models
Decision Tree, K-NN, Rule-based, Naive Bayes, AdaBoost, Random Forest, SVM, Neural Network

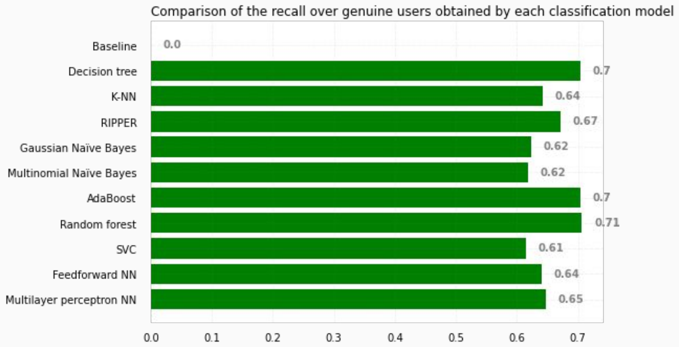
Approach

- Apply specific preprocessing for the model (e.g. normalization, reduce dimensionality, convert features)
- Split in train and test set
- Grid (or randomized) search with CV to select hyperparameters
- Train and evaluate

Classification: comparison

The best classifiers are the Ensemble Methods, reaching an accuracy of 84%.


Comparison of the accuracy obtained by each classification model


Comparison of the recall over genuine users obtained by each classification model

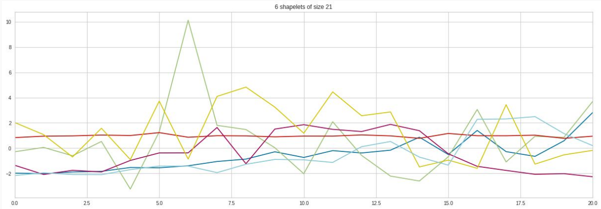
All the models are unable to discriminate well the genuine users, with a general high rate of false positives.

Time series analysis

Preprocessing

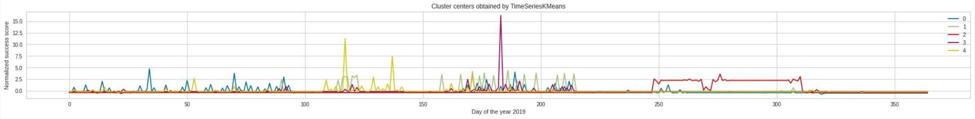
- Compute Success Scores of 2019 for each user
- Fill missing values
- Apply amplitude scaling

Shapelet extraction

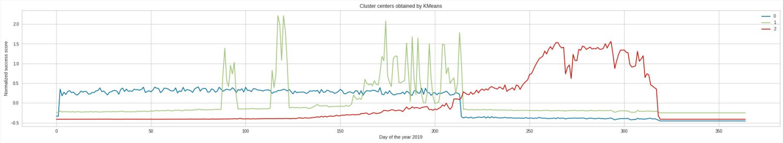


Clustering

- Partitional



- Feature-based



Thanks!