Detecting Social Bots Using Twitter Data

Yusuf Alptigin Gün, Yavuz Kanber, Advisor: Prof. Dr. Mehmet Keskinöz

Introduction

Besides regular social media users' activity, the amount of big data generated on social media on a daily basis inevitably leads to fabricated and malicious content. This type of content is generated by social media bot accounts. Our project focuses on social bots that are used to spread such content and will devise a machine learning algorithm that will be used to detect them.

Methods

To form our model's basis, we sought past research and found two researches that applied language-agnostic [1] and one-class classification [2] methods to the Cresci-17 [4] dataset. By going through these researches, we were able to create the re-do versions of them. Also, using our intuition, in addition to inspiration from [1], [2] and [3], we were able to create our own bot detection model, which is an hybrid approach that detects social bots using many different feature categories.

A Novel Hybrid Approach

- Using account-based and tweet-based features, 5 different features groups are extracted.

Category	Description	Feature
Metadata	Geo Enabled (GE)	(1)
	Statuses Count (SC)	(2)
	Favourites Count (FC)	(3)
	Friends Count (FRC)	(4)
	Followers Count (FOC)	(5)
Account-Based	Username Length (UL)	(6)
	Screen Name Length (SNL)	(7)
	Description Length (DL)	(8)
	Levenshtein Distance (LD)	(9)
	Account Age (AA)	(10)
	Tweets Count to Age Ratio (TCAR)	(11)
Behavioral	Tweet Time Mean (TTM)	(12)
	Tweet Time Standard Deviation (TTSD)	(13)
	Tweet Time Interval Mean (TTIM)	(14)
	Tweet Time Interval Standard Deviation (TTISD)	(15)
	Retweet Ratio (RR)	(16)
	Average Mentions per Tweet (AMT)	(17)
	Average Hashtags per Tweet (AHT)	(18)
	Average URLs per Tweet (AUT)	(19)
Content	Average Tweet Size (ATS)	(20)
	Average Retweets per Tweet (ART)	(21)
	Average Favourites per Tweet (AFT)	(22)
Graph-Based	Follower to Friends Ratio (FFR)	(23)
	Reputation Score (RS)	(24)

All features of our novel approach

A Language-Agnostic Approach

Bots are fundamentally different from humans in 2 main categories.

- Technical differences
- Purpose Related Differences

Category	Description	Feature
Account-Based	Default Profile	(1)
	Geo Enabled	(2)
	Protected	(3)
	Verified	(4)
	Friends Count	(5)
	Followers Count	(6)
	Listed Count	(7)
	Statuses Count	(8)
	Username Length	(9)
	Screen Name Length	(10)
	Screen Name Digits	(11)
	Levenshtein Distance	(12)
Content-Based	Time Between Tweets	(13)
	Time Between Retweets	(14)
	Emoji Count (Distributional)	(15)
	Tweet Size (Distributional)	(16)
	Number of Hashtags (Distributional)	(17)
	Number of URLs (Distributional)	(18)

Two feature categories are used, namely account-based and content-based features

A One-Class Classification Approach

- This approach is used when the model tries to find the exceptions that can occur in a specific class.
- The model detects the behaviors of legitimate users, in turn detecting any deviation from the legitimate user regardless of what type of bot is creating the deviation.
- The model is tested with both binary classification and one-class classification algorithms.

Characteristics	Description	Type
retweets	Ratio between retweet count and tweet count.	Account Usage
replies	Ratio between reply count.	Account Usage
favoriteC	Ratio between favorited tweet and tweet count.	Account Usage
hashtag	Ratio between hashtag count and tweet count.	Account Usage
url	Ratio between url count and tweet count.	Account Usage
mentions	Ratio between mention count and tweet count.	Account Usage
intertime	Average seconds between postings.	Account Usage
ffratio	Friends-to-followers ratio.	Account Information
favorites	Number of tweets favorited in this account.	Account Usage
listed	Number of listed tweets in the account.	Account Information
uniqueHashtags	Ratio between unique hashtag count and tweet count.	Account Usage
uniqueMentions	Ratio between unique mention count and tweet count.	Account Usage
uniqueURL	Ratio between unique urls count and tweet count.	Account Usage

All features of the one-class classification approach

Data Analysis

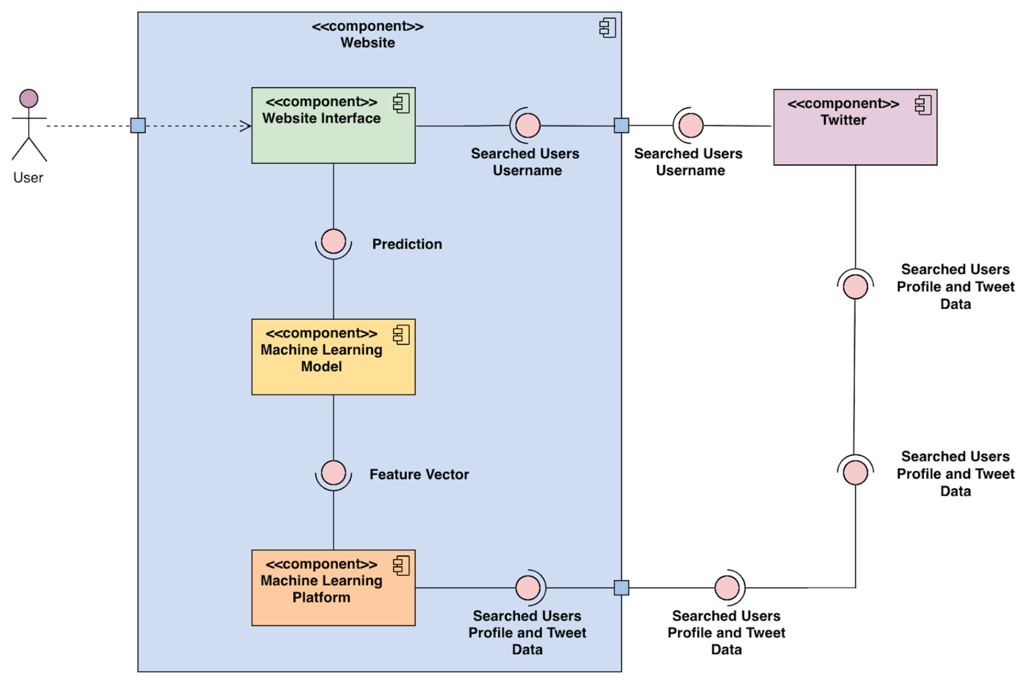
Datasets

We used the Cresci-17 dataset to train and test our models. In addition, we also used a randomly hand-picked live dataset to test our novel approach.

- **Cresci-17 Dataset:** 14.368 Twitter accounts, 3.474 human and 10.894 bot accounts with their tweet data.
- **Live Dataset:** 100 Twitter accounts, 50 human and 50 bot accounts.

Data Model

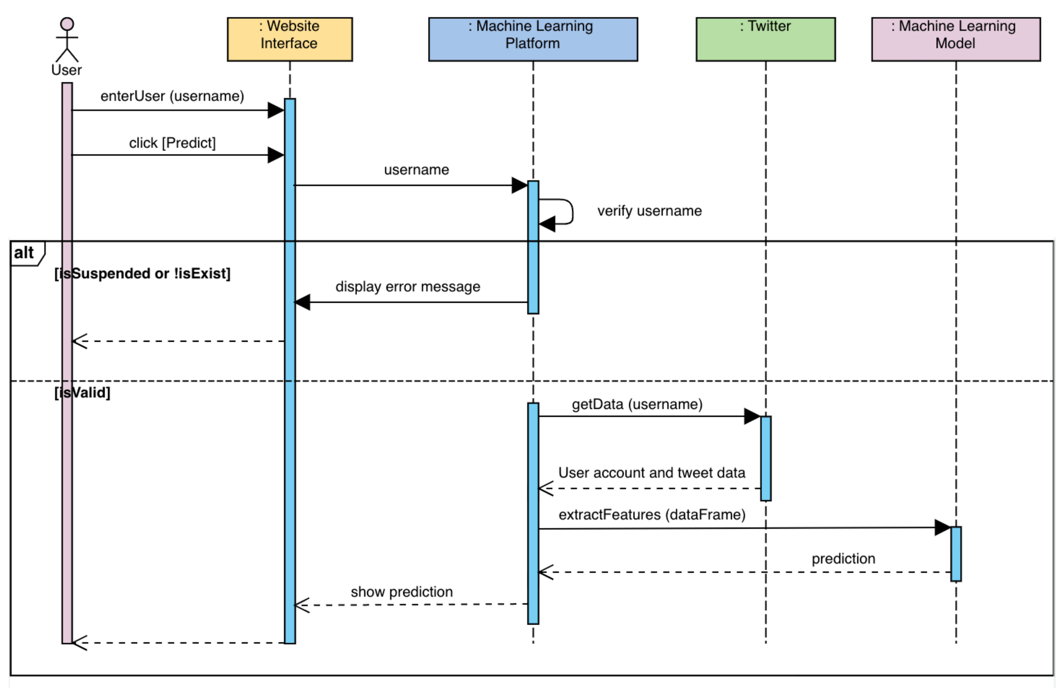
We built a whole system that enables users to check the prediction of our machine learning model on Twitter accounts.



Component Diagram of the System

Dynamic Model

The overall system has one essential use case, which can be called "Check Prediction", that gives an overview of the steps both the user and the system go through. The sequence diagram gives an illustration of these steps.



Sequence Diagram of the System

Results

We experimented all approaches with a basis testing of 6 machine learning models and 5 evaluation metrics. Also, specifically for our novel approach, we compared it with other approaches in literature and constructed live tests for it.

Basis Testing

Language-Agnostic Approach Results

	Acc	Pre	Rec	Spe	F1-S
XGBoost	0.9907	0.9874	0.9788	0.9952	0.9831
AdaBoost	0.9885	0.9796	0.9786	0.9923	0.9791
Random Forest	0.9914	0.9890	0.9890	0.9958	0.9844
Logistic Regression	0.9572	0.8530	0.9851	0.9469	0.9143
Naive-Bayes	0.9103	0.8933	0.8010	0.9581	0.8446
K-NN	0.9782	0.9636	0.9570	0.9863	0.9603

*An 80-20 split was used for testing

One-Class Classification Approach Results

	Acc	Pre	Rec	Spe	F1-S
XGBoost	0.9899	0.9849	0.9782	0.9943	0.9816
AdaBoost	0.9897	0.9849	0.9782	0.9943	0.9816
Random Forest	0.9900	0.9841	0.9796	0.9940	0.9818
Logistic Regression	0.9322	0.7752	0.9746	0.9218	0.8635
Naive-Bayes	0.8660	0.5141	0.9898	0.8455	0.6767
K-NN	0.9684	0.9110	0.9714	0.9873	0.9402

*An 80-20 split was used for testing

Novel Hybrid Approach Results

	Acc	Pre	Rec	Spe	F1-S
XGBoost	0.9946	0.9959	0.9978	0.9744	0.9969
AdaBoost	0.9934	0.9954	0.9969	0.9707	0.9962
Random Forest	0.9936	0.9957	0.9969	0.9725	0.9963
Logistic Regression	0.9762	0.9878	0.9847	0.9198	0.9862
Naive-Bayes	0.9575	0.9855	0.9660	0.8926	0.9757
K-NN	0.9824	0.9885	0.9911	0.9272	0.9898

*20-fold cross validation was used for testing

Comparison with Approaches in Literature

Comparison of our Novel Approach with J. Knauth [1], AdaBoost Classifier

	Acc	Pre	Rec	F1-S
J. Knauth (AdaBoost)	0.9881	**0.9958**	0.9835	0.9896
Our Approach (AdaBoost)	**0.9934**	0.9954	**0.9969**	**0.9962**

Comparison of our Novel Approach with J. Knauth [1], Random Forest Classifier

	Acc	Pre	Rec	F1-S
J. Knauth (AdaBoost)	0.9881	**0.9958**	0.9835	0.9896
Our Approach (Rf)	**0.9936**	0.9957	**0.9969**	**0.9963**

Comparison of our Novel Approach with J. Knauth [1], XGBoost Classifier

	Acc	Pre	Rec	F1-S
J. Knauth (AdaBoost)	0.9881	0.9958	0.9835	0.9896
Our Approach (XGBoost)	**0.9946**	**0.9959**	**0.9978**	**0.9969**

Comparison of our Novel Approach with Rodríguez, J. et al. [2], Auc Score

	AdaBoost	Rf	LR	Naive-Bayes	K-NN
Rodríguez, J. et al.	0.812	0.804	0.903	0.712	0.745
Our Approach	**0.984**	**0.984**	**0.929**	**0.867**	**0.970**

Live Dataset Testing

Live Dataset Results

	Acc	Pre	Rec	Spe	F1-S
XGBoost	0.75	1.0	0.5	1.0	0.6666
AdaBoost	0.78	0.9375	0.6	0.96	0.7317
Random Forest	0.89	0.9756	0.8	0.98	0.8791

Live Dataset Results with Non-Malicious Bots Excluded

	Acc	Pre	Rec	Spe	F1-S
XGBoost	0.7912	1.0	0.5365	1.0	0.6984
AdaBoost	0.8241	0.9310	0.6585	0.96	0.7714
Random Forest	0.9010	0.9705	0.8048	0.98	0.88

Conclusion

- A novel feature extraction approach to bot detection was proposed.
- A whole system that enables users to detect bot accounts on Twitter was created.
- You can also try the system using the QR code!

References

1. J. Knauth, "Language-Agnostic Twitter-Bot Detection," in Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), Sep. 2019, pp. 550–558. doi: 10.26615/978-954-452-056-4_065.
2. Rodríguez, J., Mata Sánchez, J., Monroy, R., Loyola-González, O., & López-Cuevas, A. (2020). A one-class classification approach for bot detection on Twitter. Computers & Security, 91, 101715. Https://doi.org/10.1016/j.cose.2020.101715
3. M. Fazil and M. Abulaish, "A Hybrid Approach for Detecting Automated Spammers in Twitter," IEEE Transactions on Information Forensics and Security, vol. 13, no. 11, pp. 2707–2719, 2018, doi: 10.1109/TIFS.2018.2825958.
4. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2017, April). The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In Proceedings of the 26th International Conference on World Wide Web Companion (pp. 963-972). ACM.

ITÜ ISTANBUL TECHNICAL UNIVERSITY
Department of Computer Engineering

Faculty of Computer and Informatics Engineering
İTÜ Ayazaga Campus, 34469, Maslak/Istanbul
bbf@itu.edu.tr
http://bm.itu.edu.tr