Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v13n2p375

Empowering detection of malicious social bots and content spammers on Twitter by sentiment analysis

By Tavazoee et al.

Published: 14 October 2020

Electronic Journal of Applied Statistical Analysis
Vol. 13, Issue 02, October 2020, 375-389
DOI: 10.1285/i20705948v13n2p375

Empowering detection of malicious social bots and content spammers on Twitter by sentiment analysis

Farideh Tavazoee[a], Davide Buscaldi[b], Francesco Mola[c], and Claudio Conversano*[d]

[a] *Fidelity Investments Inc, Bostong, Massachusetts*
[b] *Université Sorbonne Paris Cité, Bureau B211, LIPN Université Paris XIII, Campus de Villetaneuse, 99 Avenue Jean-Baptiste Clément, 93430 Villetaneuse FRANCE*
[a, c,d] *University of Cagliari, Department of Business and Economics, Viale S. Ignazio 17, 09123, Cagliari, ITALY*

Published: 14 October 2020

The role of Twitter as a platform to share opinions has been growing in the recent years especially since it has been widely used by public personae such as politicians, personalities of the show business, and other influencers to communicate with the public. For these reasons, the use of social bots to manipulate information and influence people's opinions is also growing. In this paper, we use a supervised classification model to distinguish bots from legitimate users on Twitter. More specifically, we show the importance of sentiment features in bot-human account detection. Moreover, we evaluate our detection model by testing on Russian bot accounts which are a recent set of social bots that appeared on Twitter to show that these techniques may be easily adapted to work on new, unseen types of social bots.

keywords: Bot Account Detection, Sentiment Analysis, Random Forest, Social Bot, Russian Bot.

1 Introduction

Nowadays, social media are very popular and have become a part of our daily life. Twitter has been placed among the top twenty popular sites and ranked as the second

*Corresponding author: conversa@unica.it.

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

top social media after Facebook, according to the Alexa ranking report[1]. Furthermore, Twitter is a popular source of news and constitutes a virtual place that connects people and allows them to share and discuss publicly their opinions. At the same time, these features attract not only legitimate users that post millions of tweets every day, but also automated software programs (a.k.a. *bots*) that can generate either harmful/malicious content or act as automated advertisers in a more innocuous way (Chu et al., 2010; Yang et al., 2013; Stringhini et al., 2010; Wang, 2010). Moreover, among the malicious bots, it is possible to find different kinds of malicious activities; besides the direct propagation of viruses to infect users' devices, they can also influence public opinion by spreading fake news and posting rumors. Last but not least, another malicious activity is to post Spam content with the objective to steal information and money from legitimate users (Ferrara et al., 2016).

Twitter has been improving its techniques to detect and remove the inappropriate bots and their content from the platform. However, new generations of bots are increasingly performing in a more intelligent way (Yang et al., 2013). The newest "social" bots are able to imitate the behavior of human users and to intelligently evade the detection techniques utilized by Twitter (Boshmaf et al., 2011; Haustein et al., 2014; Cresci et al., 2017; Abokhodair et al., 2016). A particular and noteworthy case is represented by accounts that have been involved in malicious activities to influence elections, such as the 2016 presidential elections in the US[2], or the ones in France in 2017[3]. Classifying all these accounts as bots is probably a simplification: there are also some human-controlled accounts. However, their behavior is similar: they spread inflammatory comments and false news to influence the outcome of elections, while trying to look like common citizens expressing their opinions.

Therefore, it is necessary to revise the existing bot detection approaches to take into account these malicious behaviors. Recently, many techniques have been proposed by scholars to automatically detect social bots, based on the stylistic differences between social bots and humans (Chavoshi et al., 2016; Chu et al., 2010; Ferrara et al., 2016; Cresci et al., 2017). As an example, Chu et al. (2010) have shown how a human replies to a tweet more exhaustively while the reply of bots are more simple and concise. Moreover, the devices used by bots to post their tweets are different. According to the authors, humans usually use web or mobile device to post their tweets while bots usually access Twitter via APIs to make the automation easier. Finally, they proposed a classifier system to identify bots and humans based on a set of extracted features from account properties. In another example, Ferrara et al. (2016) introduced a public social bot detection system called Botometer. Their supervised learning method is built around a set of features that include networks, users, friends, contents, sentiments, and timing based features.

Since novel social bots mimic the human behavior in social media to escape the detec-

[1] https://www.alexa.com/topsites

[2] https://www.nbcnews.com/tech/social-media/now-available-more-200-000-deleted-russian-troll-tweets-n844731

[3] https://blog.0day.rocks/uncovering-foreign-trolls-trying-to-influence-french-elections-on-twitter-a78a8c12953

tion techniques, analyzing their behavior might give very significant knowledge to improve the detection methods. However, social bots take advantage of modern language models to appear human-like (Cresci et al., 2017) and also post in different languages to attract more people. Therefore, models should combine both language-dependent features that allow identifying those language regularities that are expected from an automated system, and language-independent features to be able to detect these bots in different languages.

In this paper, we use a supervised classification model to discriminate humans from bots on Twitter, based on the combined use of language-independent features, like profile based features, and language-dependent features using sentiment dictionaries. In this way, we are able to demonstrate the importance of the various feature groups as well as how sentiment features can be an influential alternative for other language-dependent features to adapt the detection model and make it more accurate and robust. This model is then applied to the malicious accounts that were used to influence the 2016 US elections (Popken, 2017). We show that the model can be easily extended by including a small number of observations into the training set.

The remaining of the paper is organized as follows. Section 2 introduces the datasets and the different types of features used in the Bot Account Detection model. Section 3 describes the classification model and shows how it has been used for Social Bots detection. Section 4 extends the model to the Russian bots, whilst Section 5 ends the paper with some concluding remarks.

2 Datasets and Features

We rely on two different datasets, one containing both genuine accounts and social bots, while the second contains the accounts that have been labelled by media as “Russian bots”. These accounts appeared during the U.S. presidential elections in 2016.

The first dataset has been introduced by Cresci et al. (2017). It consists of manually selected genuine accounts corresponding to human operated accounts, hereafter called Genuine (G), and social bots (SB) that advertise the hashtag *#Talnts* on Twitter. This hashtag corresponds to a smartphone application, named Talnts, that enables users to discover, connect, and collaborate with professional talent in their localities. The second dataset has been published by Popken (2017) and includes tweet from the so-called ‘Russian bot’ (RB) accounts published during the U.S presidential elections in 2016.

We have cleaned the datasets after the extraction of the English tweets by removing punctuation and lemmatizing the content of tweets. The removal of all non-English tweets was dictated by the use of sentiment dictionaries that support the English language only. Table 1 summarizes the main information about the datasets used in the statistical analysis. After data cleaning, for the first dataset we have 237 genuine accounts with $440,999$ tweets and 234 social bot accounts with $77,256$ tweets. So, in total, there are 471 accounts and $518,255$ tweets. Thus, the probability to observe a genuine tweet is $P(G) = 0.851$ whilst that of observing a social bot is $P(B) = 0.149$. As for the RB (Russian bots) dataset, there are 356 RB users with $192,473$ tweets.

Table 1: Number of users and tweets for each of the categories of users in the datasets.

Name	N-Users	N-Tweets
Genuine (G)	237	440,999
Social bot (SB)	234	77,256
Russian bot (RB)	356	192,473

In the following, we provide the main information about the features included either in the genuine accounts dataset or in the Russian bot dataset. We consider a set of 465 different features. They are also listed in Table 2 where they have been grouped into four main categories:

1. **Statistical features (Profile)**: This set of features includes the information extracted from the profile of a user and has been widely used in previous studies (Lee et al., 2011; Ferrara et al., 2016; Chu et al., 2010; Singh et al., 2016). Effective features in this category are: the number of characters of the username (*name_len*), of digits in the username (*name_digit*), of followers (*n_followers*), of followings (*n_followings*); of favorites (*n_favorites*); the age of the account in days (*account_age*), the account reputation (*reputation*), the FOFO ratio (*fofo_ratio*), the Twitter activity ratio (*tweet_activity*), statistics about favorites and replies (average, standard deviation, minimum, maximum and median).

2. **Statistical features (Content)**: The impact of content-based features has been shown by many studies (Yang et al., 2013; Chu et al., 2010; Ferrara et al., 2016; Pasca et al., 2018). We derive some statistical features (average, standard deviation, maximum, minimum, and entropy) from the original features extracted from the content of tweets. For example, we consider the number hashtags and, for each user, we compute the above-mentioned statistics of interest. Likewise, the same computations are carried out for the number of url contained in a tweet (*url*), the number of times a user mentioned someone in a tweet (*mention*), the number of punctuation marks (*exclamation* and *question*), the number of characters (*n_char*) and words (*n_words*).

3. **Word Frequency features (Content)**: Frequent words are also representative of the content of tweets. We select the 400 most frequent words (excluding stopwords), occurring more than $1,000$ times in the dataset. We then compute the number of times each frequent word appears in the history of a given account.

4. **Sentiment features (Content)**: Sentiment analysis is the process of discovering people's opinions, attitude or mood on different topics. It is growing rapidly since social media are widely used to share opinion and feelings about a product, news or a service (Kolchyna et al., 2015) and has been applied several times for classification experiments on Twitter's data (see, for example, Corallo et al. (2019) and Tavazoee et al. (2020)). The main motivation behind using sentiment features as predictors

in a classification model is that usually humans express more emotion to interact with their friends in social media than bots (Chu et al., 2010). For this purpose, we exploit three different well-known lexicons, namely AFFIN11 (Nielsen, 2011), Hu-liu (Hu and Liu, 2004) and LabMt (Dodds et al., 2011) to derive some polarity-based features. In particular, in this set of features we include the user specific average number of emoticons and its standard deviation, as well as other statistics obtained from the three considered dictionaries.We exploit sentiments using the following Lexicon-Based Approaches.

The AFFIN11 sentiment lexicon is a list of $2,461$ labeled words with a value ranging from -5 (the most negative) to $+5$ (the most positive). To extract the polarity of each single tweet, we sum the sentiment score of each single word. For each individual account, we compute the average and standard deviation of the scores obtained for its tweets.

The Hu-Liu polarity lexicon provides $2,006$ positive and $4,797$ negative manually labeled words. For each single tweet, similarly to previous lexicon, we count the number of positive and negative words and then compute their average and standard deviation.

LabMt is a lexical source of $5,000$ words designed for sentiment analysis. The words' scores range from 1.30 for the word *terrorist* (the most negative) to 8.50 for the word *laughter* (the most positive). Likewise other lexicons, for each tweet, we sum the score of single words and compute the statistical attributes. The sentiment score of a tweet t is computed as follows:

$$s_L(t) = \sum_{w \in t} L[w]$$

where L is a sentiment lexicon, w a word in the tweet and $L[w]$ is the score of w according to L. For each lexicon, we compute the average μ and the standard deviation σ of tweet scores for each user history T_u.

$$\mu_L(T_u) = \frac{\sum_{t \in T_u} s_L(t)}{|(T_u)|}$$

$$\sigma_L(T_u) = \sqrt{\frac{\sum_{t \in T_u} (s_L(t) - \mu_L(T_u))^2}{|(T_u)|}}$$

In the case of the Hu-Liu polarity lexicon we calculate a separate score for the positive and the negative part of the lexicon, as if they were two separate lexicons.

Table 2: List of Features

Profile	screen name length
	n followers
	n followings
	n favorites
	name length
	n digit in screen name
	account age
	account Reputation
	fofo ratio(following ÷ followers)
	tweeting activity
	favorite(avg,std,max,min)
	reply(avg,std,max,min)
Content	Statistical
	retweet (avg,std,max,min)
	hashtag (avg,std,max,min)
	url (avg,std,max,min)
	mention (avg,std,max,min)
	punctuation (avg,std)
	n character in a tweet (avg,std,max,min)
	n words in a tweet (avg,std,max,min)
	Word Frequency
	400 Frequent words
	Sentiment
	emoticons (avg,std)
	Affin dictionary (avg,std)
	HL dictionary (avg,std)
	LabMT dictionary (avg,std)

3 Detecting Social Bots

As we aim to evaluate whether a Twitter account is a bot or a human, this problem corresponds to a binary classification task. We begin our analysis restricting the attention on the first dataset only (genuine accounts dataset) and evaluate three different classification approaches using the features introduced in the previous section: Support Vector Machines (SVM, e.g.: Chang and Lin (2011)), Random Forests (RF, Breiman (2001)) and Neural Network (NN). SVM is used with maximum margin classifiers (e.g.: Tzelepis et al. (2018)) that perform usually well on binary classification tasks. Random Forests (RF) is considered for classifying the Twitter accounts using 465 features because of its generally acknowledged good performance. It has been evaluated on many domains and has been proved to outperform other classifiers (Fernández-Delgado et al., 2014; Carpita et al., 2019). The NN implementation is a Multi-Layer Perceptron (MLP) with back-propagation (e.g. Bridle (1990)) and one hidden layer composed by 32 units.

In Table 3 we show the results obtained for the three different classifiers on the genuine accounts dataset: human accounts are labeled with H whilst bot accounts are labeled with B. The ground truth for the genuine dataset is $P(H) = 0.503$ and $P(B) = 0.497$, according to the numbers reported in Table 1. We evaluate the results in terms of the F1 score (also F-score or F-measure) that is among the gold standard measures in binary classification to evaluate accuracy. It considers both precision and recall, where the former is the number of correctly identified positive results divided by the number of all positive results, including those not identified correctly, and the latter is the number of correctly identified positive results divided by the number of all samples that should have been identified as positive. F-measure is thus the harmonic mean of precision and recall. The highest possible value of F is one, indicating perfect precision and recall, whilst the lowest possible value is zero, if either the precision or the recall is zero.

Table 3: Results on the genuine accounts dataset (Ten-fold cross validation) with the three classification methods.

Method	F-measure	F-measure (H class)	F-measure (B class)
SVM	0.660	0.655	0.665
NN	0.955	0.956	0.955
RF	0.970	0.971	0.969

Given the results obtained from the classifiers in the first step of our analysis, we chose the RF classifier to carry out the remaining experiments, in particular those needed to answer the following research questions:

- How important are the sentiment-based features for bot vs. human account detection systems when they are merged with other effective features?
- Is there a way to blur the effect of language-dependent features in order to facilitate the portability of the detection system across different languages?

3.1 Feature Importance Analysis

To address the first research question, we hereby highlight the contribution of the features based on the mean decrease in impurity (Ahmadi et al., 2016) as well as on an index expressing how they affect the accuracy of the classification model. This is done in order to give a better insight into the relevance of sentiment features.

To understand the most influential features in the model, we extract the features' importance of the RF classifier. The top 30 features are represented in Figure 1. As the figure shows, account profile based features have the highest impact on decreasing the mean impurity of the tree classifiers followed by sentiment-based features and, in the end, content-based statistical features. Interestingly, most of the sentiment features proposed in this study are listed among the top important features. This result proves the importance of sentiment-based information extracted from the content of tweets.

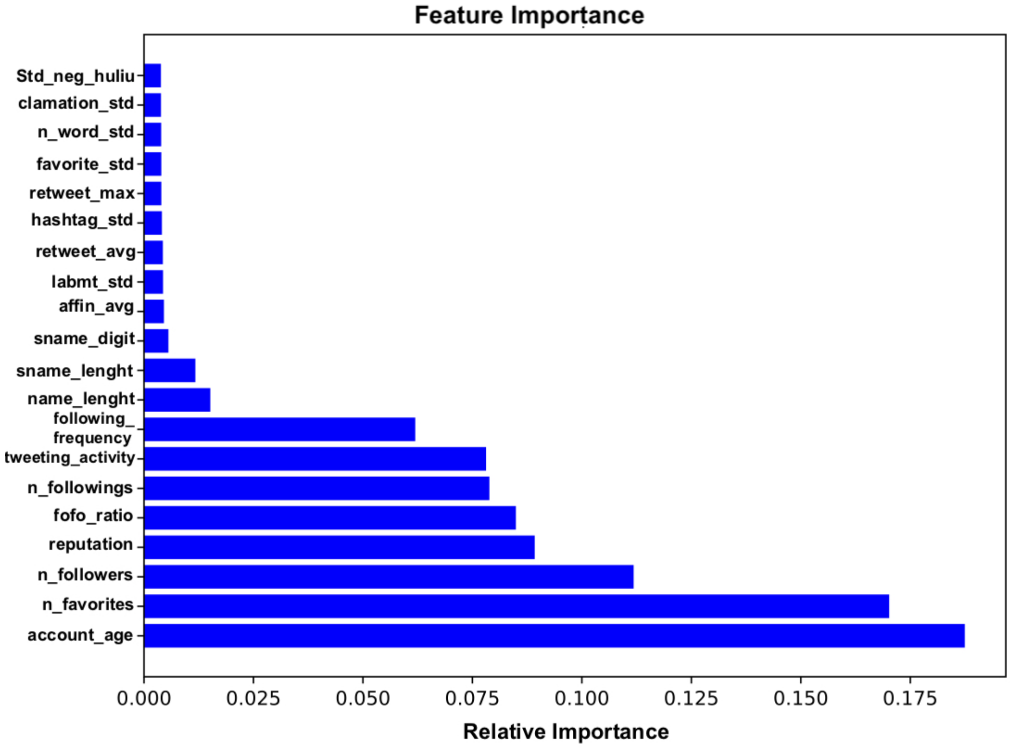

Figure 1: The most influential features (mean decrease impurity). Variable description reported in Table 1 and Section 2

As an additional experiment, we extract the most effective features influencing the accuracy of the model to investigate groups of features having the strongest impact on model accuracy. As Table 4 shows, profile and sentiment-based features are listed as the most effective features.

Table 4: The most Influential features (accuracy)

Feature's Name	**Impact on Accuracy**
Account Age	0.1556
Tweet Activity	0.0667
N Followings	0.0444
N Favorites	0.0444
LabmtAvg	0.0222
AvgNegHuliu	0.0222
Reputation	0.0222

3.2 Results and Discussion

To address our second research question it is possible to argue, as outlined in Section 1, that the new generation of social bots behaves intelligently and posts in different languages in social media platforms to attract more followers. In this respect, if a classification model is designed on less language-dependent features, it can be effectively employed for other languages with minimal efforts.

As Table 5 shows, comparing Random Forest accuracy obtained with 10-fold cross-validation for the two language-dependent groups of content based features, namely word frequency features and sentiment-based ones (see Sec.2), it can be noticed that, although the number of sentiment features (8) is much smaller than the number of frequent word features (400), the model using the combination of sentiment-based features along with the set of base features lead us to identify the most accurate classifier.

Table 5: Accuracy of Random Forest based on different features group

Feature Groups	F-measure	N.of features
Base: Statistical (Profile & Content)	0.968478	57
Base + Sentiment	**0.970652**	**57+8**
Base + Word Frequency	0.961598	57+400
Base + Sentiment + Word Frequency	0.957387	57+8+400

It is clear that sentiment features can outperform word frequency features with a relatively high margin (almost 1%) if one considers that there are more than 300 million active Twitter accounts and 1% means predicting correctly the purpose of more than 3 million accounts. These results demonstrate that word frequency features can be reliably replaced by dictionary-based sentiment features, as the latter not only allow to build a more parsimonious (i.e., with less features) classifier, but also allows us to improve model accuracy and are more robust across languages.

Moreover, results reported in Table 5 are remarkable because the decrease in computational complexity caused by limiting the number of possible features for each split of the trees grown during the creation of a random forest is considerably large. In fact, it is well known that building a random forest with B trees induces a complexity of $O\,(B \times \sqrt{m} \times n \log n)$, with m being the number of variables ($\sqrt{m}$ that used for node spitting), n the number of instances and $\log n$ the tree depth. In our case, moving to 57+8 features from 57(+8)+400 consistently decreases the $\sqrt{m}$ factor influencing computational complexity (Louppe, 2014).

4 Extending the model to detect the 'Russian bots'

In the following, we evaluate the effectiveness of our classification model focusing on the detection of Russian bots (RB). Specifically, we aim to answer three additional research questions:

- How well the model behaves when classifying unseen RB?
- How many RBs are needed to improve the RB detection rate of the model? What is the role of sentiment features in the detection of RBs?
- How the Russian bots behave differently from humans?

To answer the three aforementioned research questions, we performed three experiments.

For the first research question, we built a Random forest model on the genuine accounts dataset, that includes genuine (G) and social bots (SB) accounts and test the model on the Russian bot (RB) accounts dataset. Following results presented in Table 5 (Sec. 3.2) we discard the word frequency features and concentrate on the base features and the sentiment features only. As the first row of Table 6 shows, only 29% of RB can be detected if the training set does not contain any sample of the new bots (that is, the model built on SBs is not effective enough to detect RBs).

To address the second and third research questions, we gradually include a few number of randomly selected RB users to our training set. This inclusion simulates the manual recognition of a finite set of RB bots. Next, we build a Random Forest model with and without sentiment features to assess if including a few positive samples to the training set can influence the detection rate as well as to verify whether the set of chosen sentiment features is able to improve model fitting. To this end, we consider four different designs by randomly appending some RB users. The number of appended RB users varies from 5 to 20 with an increment of 5. The aim is to find out how many additional RBs are needed to improve the model accuracy. In order to extract the misclassification of our prediction model, we computed False Positive (FP) (i.e., the number of humans that are incorrectly predicted as bots) and True Negative (TN) (i.e., the number of bots that correctly predicted), as well as F-score which considers both recall and precision from the confusion matrix of RF classifier for each experiment. Results of these additional experiments are reported in columns 2-5 of Table 6. In the FP column of Table 6, in the

Table 6: Accuracy with and without using sentiment features for each experiment. The first row addresses the first research question. The model is built on combination of Genuine and Social bots (SB) dataset and then tested the Russian bot (RB) accounts dataset. A few RB users are randomly appended to the training set and then tested on a combination of Genuine and RBs dataset. False positive rate (FP) and True Negative Rate (TN) are computed on the dataset including the base and sentiment features (B+S) or the base features (B) only

Training set			Test set		F−score		FP(%)		TN(%)	
G	SB	RB	G	RB	B+S	B only	B+S	B only	B+S	B only
237	234	0	0	333	0.35	0.33	0.0	0.0	29.0	25.0
170	150	5	80	75	0.77	0.73	1.0	1.0	52.0	49.0
170	145	10	80	70	0.85	0.80	1.0	1.0	72.0	68.0
170	140	15	80	65	0.94	0.90	1.0	1.0	89.0	84.0
170	135	20	80	60	0.96	0.94	1.0	1.0	91.0	86.0

first experiment, there is no FP, because there is just one class (i.e., RB) in the test set. For other experiments, the value of FP is fixed (i.e., 1) and the FP rate is 1% because we only add randomly selected RBs to the training set which means the number of genuine account observations is intact. In other words, there is no change in modeling genuine accounts and there is only one account that is misclassified for all experiments. As Table 6 shows, the set of sentiment features can relatively affect the True Negative rate. In addition, even with a very few positive samples, we can dramatically improve the bot detection rate. Moreover, the more samples we add to the training set, the better is the bot detection rate we achieve.

In the third experiment, we remove all the SB users from the dataset and merge the G users with RB ones. Next, we run random forest on this modified dataset and estimate the 10 fold cross validation error. The performance of the model is pretty good since it is able to detect the RB users with 98% accuracy. Furthermore, to prove the usefulness of the RF classifier in our experiments, we compare the performance of RF classifier in the third experiment with Naive Bayes as well as Support Vector Machine (SVM) classifiers. The RF classifier outperforms other classifiers w.r.t. accuracy: results show that accuracy of RF is 98% whilst that of SVM (Naive Bayes) is 93%(85%).

Next, we observe the performance of features individually in order to understand how they affect the RB accounts. As Figure 2 shows, Russian bots follow some simple rules as they need to advertise themselves and spread their malicious contents reaching a greatest audience. Figure 2 shows that these bots use a higher number of hashtags and urls on their tweets compared to humans whilst tweeting less in comparison to humans in their lifetime. It is also interesting to note that these bots tend to produce content that is slightly more negative than humans (bottom panel of Figure 2). This seems to support the findings of Klenner (2018) that extreme right discourse is particularly polarizing and contains concepts related to fear to cause an emotive reaction in people.

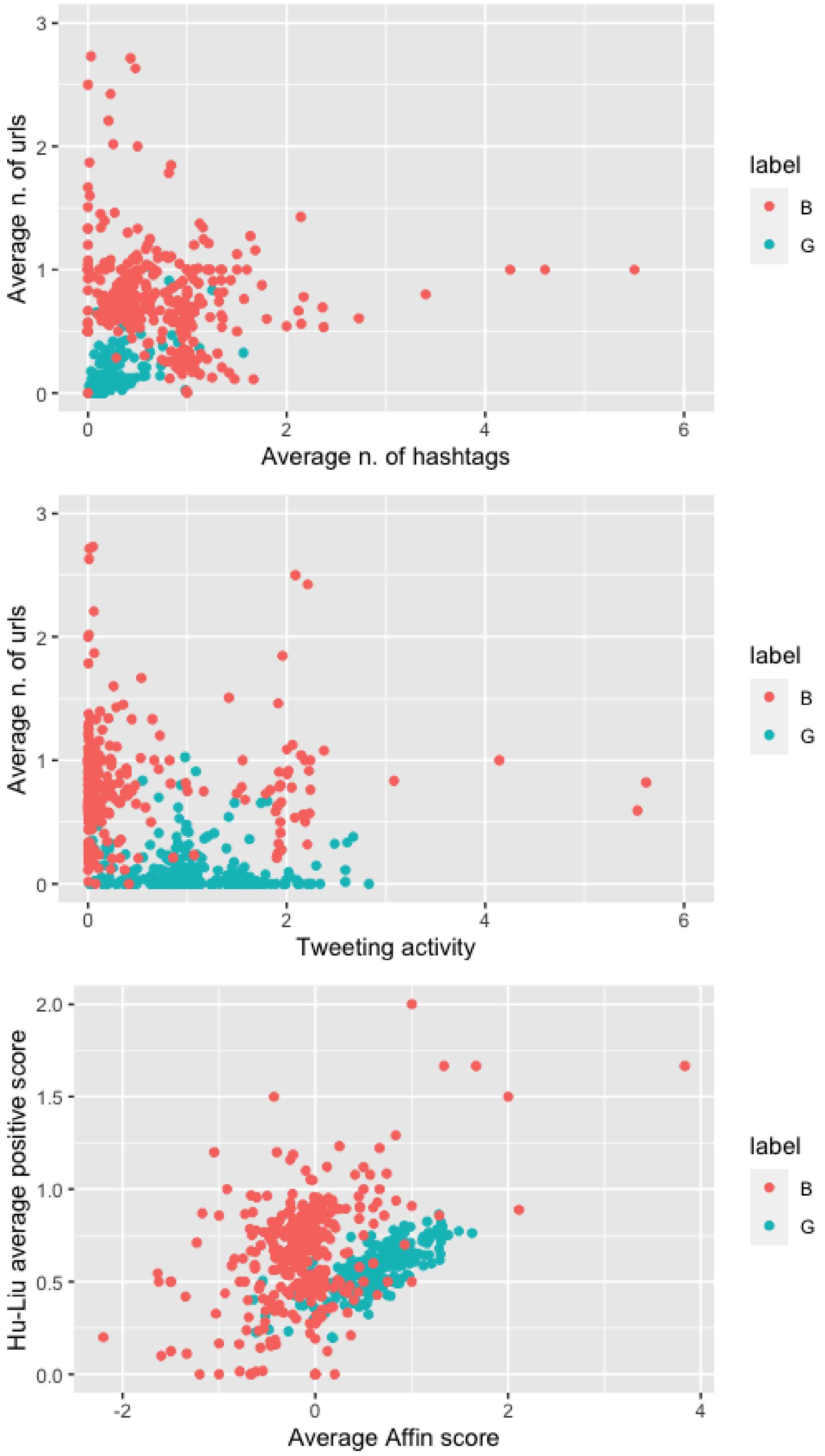

Figure 2: Comparison of the distribution of the values for some features for the "Russian bots" in Popken (2017) and human users. Blue are human users, red are Russian bot accounts.

5 Concluding Remarks

We have considered a Random Forest model for social bot detection on Twitter using four different categories of features. Our approach achieves very high accuracy by relying on a set of dictionary-based sentiment features that have not been used in previous researches on this topic besides a set of previously known features.

First, we have shown the effectiveness of sentiment features, that can be used in place of other language-dependent features, such as bag-of-words or word embeddings, with the result of simplifying the classification model without reducing its effectiveness. In this respect, sentiment features proved to be high-level representatives of the content of the tweets, at least for the bot detection purpose.

Second, we have examined the detection rate on the RB dataset and discovered, interestingly, that if we add only few RB observations to the training set our, model is able to accurately predict the rest of the RB. In this respect, if Twitter identifies a few number of new set of bots manually and add it to the training set, the model is able to detect the rest of that campaign of bots.

In addition, as Table 6 shows, we approved the impact of our novel sentiment features on predicting the RB users.

Finally, we have analyzed the RB accounts behavior to extract the characteristic of RBs. We concluded that the new generation of bots still follow some simple rules to catch their victims in Twitter.

These results can be exploited to ease the task of mapping the model to other languages by building a model around sentiment lexicons. We aim to perform additional experiments on other languages in the future.

References

Abokhodair, N., Yoo, D., and McDonald, D. W. (2016). Dissecting a social botnet: Growth, content and influence in twitter. *CoRR*, abs/1604.03627.

Ahmadi, M., Ulyanov, D., Semenov, S., Trofimov, M., and Giacinto, G. (2016). Novel feature extraction, selection and fusion for effective malware family classification. In *Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy*, CODASPY '16, pages 183–194, New York, NY, USA. ACM.

Boshmaf, Y., Muslukhov, I., Beznosov, K., and Ripeanu, M. (2011). The socialbot network: when bots socialize for fame and money. pages 93–102.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Bridle, J. S. (1990). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In Soulié, F. F. and Hérault, J., editors, *Neurocomputing*, pages 227–236, Berlin, Heidelberg. Springer Berlin Heidelberg.

Carpita, M., Ciavolino, E., and Pasca, P. (2019). Exploring and modelling team performances of the kaggle european soccer database. *Statistical Modelling*, 19(1):74–101.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27.

Chavoshi, N., Hamooni, H., and Mueen, A. (2016). Debot: Twitter bot detection via warped correlation. In Bonchi, F., Domingo-Ferrer, J., Baeza-Yates, R. A., Zhou, Z., and Wu, X., editors, *IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain*, pages 817–822. IEEE.

Chu, Z., Gianvecchio, S., Wang, H., and Jajodia, S. (2010). Who is tweeting on twitter: Human, bot, or cyborg? In *Proceedings of the 26th Annual Computer Security Applications Conference*, ACSAC '10, pages 21–30, New York, NY, USA. ACM.

Corallo, A., Fortunato, L., Massafra, A., Pasca, P., Angelelli, M., Hobbs, M., Al-Nasser, A. D., Al-Omari, A. I., and Ciavolino, E. (2019). Sentiment analysis of expectation and perception of MILANO EXPO2015 in Twitter data: a generalized cross entropy approach. *Soft Computing*, pages 1–11.

Cresci, S., Pietro, R. D., Petrocchi, M., Spognardi, A., and Tesconi, M. (2017). The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. *CoRR*, abs/1701.03017.

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., and Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: hedonometrics and twitter. *PLOS ONE*, 6(12):e26752.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.*, 15(1):3133–3181.

Ferrara, E., Varol, O., Davis, C., Menczer, F., and Flammini, A. (2016). The rise of social bots. *Commun. ACM*, 59(7):96–104.

Haustein, S., Bowman, T. D., Holmberg, K., Tsou, A., Sugimoto, C. R., and Larivière, V. (2014). Tweets as impact indicators: Examining the implications of automated bot accounts on twitter. *CoRR*, abs/1410.4139.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Klenner, M. (2018). What does it mean to be a wutbürger? a first exploration. In *Proceedings of 4th Workshop on Sentic Computing, Sentiment Analysis, Opinion Mining, and Emotion Detection (EMSASW 2018) Co-located with the 15th Extended Semantic Web Conference 2018 (ESWC 2018)*, pages 32–37.

Kolchyna, O., Souza, T. T. P., Treleaven, P. C., and Aste, T. (2015). Twitter sentiment analysis. *CoRR*, abs/1507.00955.

Lee, K., Eoff, B. D., and Caverlee, J. (2011). Seven months with the devils: a long-term study of content polluters on twitter. In *In AAAI Int'l Conference on Weblogs and Social Media (ICWSM*.

Louppe, G. (2014). Understanding Random Forests: From Theory to Practice. *ArXiv e-prints.*

Nielsen, F. Å. (2011). A new ANEW: evaluation of a word list for sentiment analysis in microblogs. *CoRR*, abs/1103.2903.

Pasca, P., Ciavolino, E., and Boyd, R. (2018). A data-mining approach to the parkour discipline. In *Proceedings of the* 49th *Annual Meeting of the Italian Statistical Society*, Palermo, Italy. SIS.

Popken, B. (2017). Twitter deleted 200,000 russian troll tweets. https://www.nbcnews.com/tech/social-media/now-available-more-200-000-deleted-russian-troll-tweets-n844731.

Singh, M., Bansal, D., and Sofat, S. (2016). A novel technique to characterize social network users: Comparative study. In *Proceedings of the 6th International Conference on Communication and Network Security*, ICCNS '16, pages 75–79, New York, NY, USA. ACM.

Stringhini, G., Kruegel, C., and Vigna, G. (2010). Detecting spammers on social networks. In *Proceedings of the 26th Annual Computer Security Applications Conference*, ACSAC '10, pages 1–9, New York, NY, USA. ACM.

Tavazoee, F., Conversano, C., and Mola, F. (2020). Recurrent random forest for the assessment of popularity in social media. *Knowledge and Information Systems*, pages 1847–1879.

Tzelepis, C., Mezaris, V., and Patras, I. (2018). Linear maximum margin classifier for learning from uncertain data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2948–2962.

Wang, A. H. (2010). Don't follow me: Spam detection in twitter. *2010 International Conference on Security and Cryptography (SECRYPT)*, pages 1–10.

Yang, C., Harkreader, R., and Gu, G. (2013). Empirical evaluation and new design for fighting evolving twitter spammers. 8(8):1280–1293.