ISSN 2007-9737

Organization, Bot, or Human: Towards an Efficient Twitter User Classification

Kheir Eddine Daouadi[1], Rim Zghal Rebaï[2], Ikram Amous[2]

[1] Sfax University, MIRACL-FSEGS, Sfax,
Tunisia

[2] Sfax University, MIRACL-ISIMS, Sfax,
Tunisia

{khairi.informatique, rim.zghal}@gmail.com, ikram.amous@enetcom.usf.tn

Abstract. Today, through Twitter, researchers propose approaches for classifying user accounts. However, they have to face confidence challenges owing to the diversity of the types of data propagated throughout Twitter. In addition, the messages from Twitter are imprecise, very short and even written in many dialects and languages. Moreover, the majority of the related works focus on the overall user's activity, which makes them not suitable at the post-level classification. This paper presents an alternative approach for classifying user accounts as being accounts of bots, humans or organizations. The suggested approach consists in accurately classifying user accounts from one single post by leveraging a minimal number of language-independent features. We performed several experiments over a Twitter datasets and supervised learn-based algorithms. Our results demonstrated that simply using a minimal number of language-independent features extracted from one single post is sufficient to classify user accounts accurately and quickly. Our proposed approach yielded high F1-measure (>95%) and high AUC (>99%) using Random Forest.

Keywords. Social network analysis, twitter user classification, human vs. bot vs. organization, statistical-based approach, content-based approach, hybrid-based approach.

1 Introduction

Nowadays, social networks play a major role in our everyday life. Billions of users exchange a huge amount of data on such social networks. These platforms allow users to register through the creation of an account, follow others and share content. Twitter is one of the leading social networks. This free micro blogging has been widely used not only by humans, but also by organizations and by bots. In [1], the authors showed that the organizations users makes up 9.4% of the accounts on Twitter. In [2], the authors showed that between 9% and 15% of the active Twitter users were bots. The ability to classify the patterns of user accounts is needed for developing of many applications such as recommendation engines and information dissemination platforms. Also, adding to that can help users focus on valuable social accounts, get effective information and avoid the network traps etc.

Today, through Twitter, researchers propose approaches for classifying user accounts. However, they have to face confident challenges owing to the diversity of the types of data propagated throughout Twitter. Moreover, they majority of the related works focus on the overall user's activity ([15] and [2], for example, used a few hundred posts), which make it not suitable at the post-level classification. In addition, the majority of the related works used parameters from Natural Language Processing (NLP). Nevertheless, such approaches still have some drawbacks like the use of language dependent features, the time consumption and the computational cost. Such limits stands as an obstacle to obtaining real-time,

doi: 10.13053/CyS-23-2-3192

multi-language and multidialectal user classification.

In this paper, we present an alternative approach for classifying Twitter user accounts as being accounts of humans, bots or organizations. Our proposed approach is more practical and simpler because it does not have to face problems related to text, as the related works have faced. Moreover, our proposed approach leverages a minimal number of features in order to perform the classification task. In addition, our proposed user descriptors are extracted from one single post, unlike the related works that focus on the overall account's activity. The choice of limiting the number of features was motivated by both model efficiency and interpretability reasons. In this study, we will try to address the following research question:

– Is it possible to accurately classify Twitter user accounts by leveraging a minimal number of language-independent features extracted from one single post?

The remainder of this paper is structured as follows. Section 2 discusses the related works. Section 3 describes our proposed approach: we will discuss its different steps. In Section 4, we present the experimental settings and evaluation results. In Section 5, we will discuss our proposed approach. Finally, Section 6 concludes the paper.

2 Related Works

Twitter user accounts have been analyzed across different dimensions. The main purpose of this analysis is to infer a latent attribute of such user. Several works have been done on a single and specific perspective such as organization detection [1, 3, 4, 5], bot detection [2, 6, 7, 8, 9], political orientation [10], age prediction [11] etc. Twitter user classification approaches involve three major types, mainly statistical-based approaches, content-based approaches and hybrid-based approaches.

First, statistical-based approaches use statistical parameters in order to perform the classification task. Several statistical parameters were proposed such as temporal features and post frequency features. The main advantage from this type of approach is that it takes a shorter time and require less computational cost than the other types of approaches. In addition, they used language independent features (i.e. assures multi-language and multi-dialectal user classification). Authors in [12] classify users from those of bots, humans and cyborgs. They used parameters from the post frequency; this yielded an accuracy result of 88%. In [6], the authors classify users from those of bots and humans. They used parameters from the metadata of user profiles; this yielded an AUC of 92.0%. Authors in [15] classify users from those of humans, bots and organizations. They used the time distribution between posts; this yielded a classification performance of 78.8%.

Second, the content-based approaches that use parameters from NLP to perform the classification task. This type of approaches used language dependent features, which not assures multi-language and multi-dialectal user classification. In addition, they are confronted to the problem inherent to text. In other words, messages from social networks are very short and even written in informal language full of mistakes. Several parameters from NLP were proposed (ex. sentiments, n-grams, emotion expressions, Tweet style). In [4], the authors classify tweets belonging to organizations or individuals. They used parameters from different linguistic aspects of tweet content, this yielded F1-measure result of 89.20% for the English tweets. In [7], the authors classify tweets belonging to humans or bots. They used word2vec, Wordnet and Conceptnet to create a semantic word vector, which is used later by the Long Short Term Memory (LSTM) for the classification task. This yielded F1-measure result of 96.84%.

Third, the hybrid-based approaches that attempt to combine both statistical and content features in the classification task. Authors in [2] classify users from those of humans and bots. They used more than thousand parameters, these included content, sentiment features, friend, user metadata, network and timing features. This yielded AUC result of 95.0%. Authors in [8] classify users from those of bots and humans. They used parameters from content-based, trust-based and user-based features. This yielded accuracy result of 92.1%. In [9], the authors classify users from those of bots and humans. They used parameters from the metadata of user profile and tweeting features. This yielded F1-measure result of 86,625%. In [13], the authors classify users from those of bots and humans. They used parameters from user-demographic, user-friendship networks, user-content and user history features. This yielded F1-measure result of 98.4%.

In summary, the current approaches for Twitter user classification are not sufficient in order to classify user accounts accurately and quickly both at post-level and at user-level. Our work is similar to that of [15], but in this previous work, authors used a large set of posts from the user account. Therefore, their proposed approach cannot be suitable for the post-level classification. In this work, we investigate the possibility of accurately classifying user accounts by leveraging a minimal number of language-independent features extracted from one single post. Therefore, we propose a statistical-based approach for classifying the patterns of user accounts into those of humans, those of bots or those of organizations.

3 Proposed Approach

Following the previously-mentioned challenges and based on the existing works, we propose an efficient statistical-based approach for classifying user accounts on Twitter. In general, the task of supervised user classification begins with a training set U = $(U_1 \ldots . U_N)$ of users that are labeled with a class ci ∊ C (ex. Organization, human or bot). Then, the task is to determine the optimal features set F= $(f_1 \ldots f_n)$ that is able to correctly assign a class (ci) to a new user (U_i) that has the feature vector $F^i = (f_1{}^i .. f_n{}^i)$. Our proposed approach consists of three main steps, namely data acquisition, features extraction and classification.

3.1 Data Acquisition

In this step, our main objective is to use Twitter Timeline (API), which allows collecting the K most recent posts of such user.

Table 1. Our initial ground truth corpus description

Dataset	Description	Active User
Human accounts	Represent 3474 Twitter users that were labelled as human users; this was published in [14].	2702
Bot accounts	Represent 4912 Twitter users that were labelled as social spambots users, and 2661 Twitter users that were labelled as traditional spambots users; this was published in [14].	6465
Organization accounts	Represent 1911 Twitter users that were labelled as organization users; this was published in [1].	1811

The posts come with an information form of User, Tweet and Time attributes. The User attribute contains an information about the user account (e.x. the number of posts issued by the user and the number of followings).

The Tweet attribute contains information about the content of tweet and their statistics (e.x. number of retweets and favorites that this post has). The Time attribute contains information from the date and time that this post was posted on Twitter. To create our ground truth we used the datasets that were published in [14] and in [1].

Table 1 presents an overview about our interested datasets. These datasets contain a user_id with a label as human, bot or organization. We used the user_id with Twitter Timeline (API) in order to collect the most recent post of each labelled user. We observed the users' accounts' statuses via the statuses/user_Timeline API endpoint.

These statuses can take one of four values: Private account, Suspended account, Removed account, Active account.

Table 2. Metadata of user profile features

Feature Description
Whether the user provided a URL in association with their profile (URL).
Whether the user has not altered the background or theme of their user profile (Default Profile).
Whether the user has a verified account (Verified).
Whether the user has provided a description in association with their profile (Description).
Whether the user has enabled the possibility of geo-tagging their posts (Geo).
Whether the user has not uploaded their own image (Image).
The number of public lists that this user is a member of (Lists).
The ratio of the number of followers to the number of followings (Ratio1).
The age of the user account in days (Age).
The total number of posts that were issued by the user (Posts).
Average post per day (Average Posts).
The number of favorites done by the user (Favorites).
The number of favorites done by the user per day (Favorites per day).
The ratio of the number of followings to the sum of both followers and followings (Reputation).
The number of followers per day (Followers per day).

Table 3. Retweet and favorite features

Feature description
The number of favorites that this post has per hour (Retweets per hour). The ratio of the number of retweets of the post to the number of followers. (Retweets per Followers).
The number of favorites that this post has per hour (Favorites per hour).
The ratio of the number of favorites of the post to the number of followers (Favorites per Followers).

Since we could have access to only the information from the active accounts. In this study, we used only the active accounts. Each recent post (P_i) of each active labelled user (U_i) is retrieved, and this way the next step could follow.

3.2 Features Extraction

In this step, our objective was to build the feature vector of the user account. However, we used a minimal number of language-independent features extracted from the collected post of each labelled user. The choice of minimizing the number of features was motivated by both model efficiency and interpretability reasons.

Our proposed parameters include 15 metadata of user profile features (MUP), 2 retweet features and 2 favorite features (ReFa) as described in Table 2 and Table 3. However, our MUP illustrates the activity of the user whereas the ReFa features represent how the others interact with the post of the user.

We created a feature vector F^i= ($f_1{}^i..f_n{}^i$) for each (U_i) from the collected (P_i), where n=19 is the number of our proposed features.

3.3 Classification

In this step, our main objective was to use a supervised learning approach in order to classify the feature vectors F^i=($f_1{}^i..f_n{}^i$) that has ci ∈ C (Organization, Human or Bot). First, we chose the algorithm that performs better with our proposed features. However, we compared several supervised learning algorithms; these include Random Forest (RF), Logit Boost (LB), Rotation Forest (RoF), Bagging (B) and Multi-Layer Perceptron (MLP).

We used each algorithm with its standard parameters in order to ensure the comparability between them. Second, we chose the optimal number of features. However, we built three models: one containing MUP features, one containing ReFa features and one containing all features.

Third, we manually labelled 1983 user accounts; 1352 of them were labelled as ‘human’ and the remaining ones were labelled as ‘organization’. These were used in order to

Table 4. 10-fold cross validation average metrics results

Algorithm	P	R	F1	AUC
RF	**0,959**	**0,958**	**0,958**	**0.994**
LB	0,918	0,920	0,918	0.984
RoF	0,955	0,954	0,954	0.939
MLP	0,885	0,884	0,884	0.963
B	0,951	0,950	0,950	0.992

Table 5. 10-fold cross validation metrics results using different types of features

Features	P	R	F1	AUC
MUP	0,958	**0,958**	**0,958**	**0.994**
ReFa	0,614	0,653	0,613	0.707
MUP+ReFa	**0,959**	**0,958**	**0,958**	**0.994**

Table 6. Test metrics results using MUP features

Class	P	R	F1	AUC
Human	0,997	0,981	0,989	0.992
Organization	0,993	0,976	0,981	0.993

Table 7. 10-fold cross validation metrics results using Random forest for each class using MUP features

Class	P	R	F1	AUC
Human	0,939	0,940	0,939	0.993
Bot	0,986	0,977	0,981	0.995
Organization	0,888	0,919	0,902	0.992

evaluate the overall accuracy of our proposed approach.

Fourth, given the skewed distribution in our ground truth we used the Synthetic Minority Oversampling Technique (SMOTE) [16] in order to enhance previous labelled datasets with more examples of human, bot and organization samples. The SMOTE algorithm aims to generate new samples based on the feature vectors of such classes.

The main advantage of using SMOTE is that it forces the decision region of such classes and enhances previous labelled datasets without any additional and very expensive labelling step. Finally, we chose the optimal model with the optimal features sets.

4 Experiment and Evaluation

In order to evaluate our proposed approach, we used the confusion matrix in order to visualize the performance measurements. Several metrics can be drawn from the confusion matrix were the well-known F1-measure (F1), Precision (P), Recall (R), Accuracy (A) and Area Under the Curve (AUC) metrics. The tests were implemented using the Waikato Environment for Knowledge analysis (WEKA). We used 10-fold cross validation in order to calculate the performance measurements.

Our first set of experiments found the optimal supervised learning algorithm. As can be concluded from the experiment presented in Table 4, the RF algorithm outperformed the other ones by achieving (P=95.9%, R=95.8%, F1 = 95.9% and AUC=99.4%).

The second set of our experiments found the optimal number of features. As shown in Table 5, our proposed ReFa features achieved accuracy results of 61,3% and 70,7% in F1 and AUC metrics, respectively. The main reason is that this type of features are related to the time when the posts was retrieved. In contrast, our proposed MUP features achieved high accuracy metrics (>95.

The third set of our experiments tested our manually-labelled user accounts. Here, we focused on the leading companies in the USA according to Fortune 500 site [17]. These were labelled as organization users. Then, we labelled human accounts by verifying their profiles manually. As shown in Table 6, our proposed MUP features achieved high accuracy results for classifying our labelled humans and organizations users.

As can be concluded from the experiments presented in Table 7 and Table 8, when adding our manually-labelled accounts to our initial ground truth, a significant gain was observed on the F1 with 1.9% and 1.6% for the organization and human classes respectively. Our fourth set of experiments enhanced previous-labelled datasets by using the SMOTE algorithm.

Indeed, a significant gain was observed in F1 when using SMOTE with our proposed MUP features as shown in Table 9. In summary, we developed a highly-accurate model for Twitter user classification.

Experiments showed that our proposed MUP is the optimal set of features for classifying user accounts accurately in terms of AUC and F1 metrics. Our final predictive model leveraged 15 parameters from the metadata of user profiles while achieving high accuracy results of classification (>95). We demonstrated that simply using a minimal number of language-independent features extracted from one single post is sufficient to classify user accounts accurately and quickly.

5 Discussion

Although quite successful, previous Twitter user classification approaches are very expensive, as they require a large number of labeled datasets as well as a huge amount of data for each user to be classified. In other words, most, if not all, of the related works focus on the overall user's activity (ex. use a few hundreds of posts), which makes these works fail on the post-level classification. The scalability issue also faces these approaches as they are faced by the query rate limits imposed by Twitter Timeline API. Besides, this makes it impossible to analyze user accounts in a large population. In contrast, our proposed approach classifies a user account from one single post, which makes it suitable for both the user-level and the post-level.

In addition, our proposed approach leverages a minimal and interpretable language-independent feature set, yet it requires a minimal amount of training data. A limited number of statistical feature sets with a clear meaning, like our proposed MUP, allows producing an interpretable model, especially when combined with deep learning strategies, which are very hard to interpret. Moreover, our proposed approach ensures a multi-language and multi-dialectal user classification. In order to answer our research question, we performed an extensive amount of experiments over our initial ground truth, our labelled datasets, and our proposed features. Experiments showed that our proposed MUP is the minimal number of features in order to accurately classify the patterns of user accounts as those of humans, bots or organizations. Our final predictive model achieved F1 results of 98.6%, 96.2% and 96% for classifying bot, human and organization users.

Table 8. 10-fold accuracy results when adding our labelled user using MUP features

Class	P	R	F1	AUC
Human	**0.952**	**0.958**	**0.955**	**0.994**
Bot	0.986	0.973	0.980	0.995
Organization	**0.911**	**0.931**	**0.921**	**0.993**

Table 9. The effect of SMOTE with our MUP features

Datasets	F1
without SMOTE	0.961
SMOTE (Organization)	0.970
SMOTE (Human)	0.971
SMOTE (Bot)	0.973

Our proposed approach has two limitations. First, although our proposed MUP features were a decisive factor, combining them with content deep-learning strategies could enhance the overall accuracy. Second, we limited our classification to only three classes. However, other types can be added in order to make our approach more generalizable.

6 Conclusion

Nowadays, social networks have become an important part of our everyday life. These platforms are used not only by humans to share content and follow others, but also by bots to spread misinformation and pollute content. Moreover, the organizations used these platforms to spread information and engage their users.

Classifying the patterns of users from those of humans, bots and organizations can help users focus on valuable social accounts, get effective information, ensure their own security and avoid network traps etc. In this paper, we presented an efficient statistical-based approach for classifying user accounts on Twitter. The suggested approach aims to accurately classify user accounts by leveraging a minimal number of language-independent features extracted from one single observation.

Our proposed approach may offer an important advantage by using it both at the user-level and at the post-level with high accuracy results. Experiments demonstrated that our proposed MUP features are sufficient in order to accurately classify user accounts as those of humans, bots or organizations. Our final predictive model yielded high accuracy results (>95%).

As future works, we plan to enrich our model with other labelled user types. We will make our proposed framework an open source in order to ensure the comparability with future works and allow the research community to perform post-level or user-level classification tasks using it. In addition, we will extend our work to be able to detect the topic of the classified post. Our proposed work is very important to both academia and industry. We will use our proposed framework to analyze Twitter conversations in different contexts to determine the extent of the interference of bots, humans and organizations with public discourse, and to understand how their sophistication and capabilities evolve over time.

References

1. **McCorriston, J., Jurgens, D., & Ruths, D. (2015).** Organizations Are Users Too: Characterizing and Detecting the Presence of Organizations on Twitter. *9th International Conference on Web and Social Media ICWSM,* pp. 650–653.
2. **Varol, O., Ferrara, E., Davis, C.A., Menczer, F., & Flammini, A. (2017).** *Human-Bot Interactions: Detection, Estimation, and Characterization.* pp. 280–289.
3. **Oentaryo, R.J., Low, J.W., & Lim, E.P. (2015).** Chalk and Cheese in Twitter: Discriminating Personal and Organization Accounts. *European Conference on Information Retrieval (ECIR´15) Lecture Notes in Computer Science,* pp. 465–476. DOI: 10.1007/978-3-319-16354-3_51.
4. **De Silva, L. & Riloff, E. (2014).** *User Type Classification of Tweets with Implications for Event Recognition*. pp. 28–108.
5. **Daouadi, K.E., Rebaï, R.Z., Amous, I. (2018).** Organization vs. individual: Twitter user classification. *Proceedings of the second conference on language processing and knowledge management,* pp. 1–8.
6. **Ferrara, E. (2017).** Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday*, Vol. 22, No. 8. DOI: 10.2139/ssrn.2995809.
7. **Jain, G., Sharma, M., & Agarwal, B. (2018).** Optimizing semantic LSTM for spam detection. *International Journal of Information Technology*, Vol. 11, No. 2, pp. 239–250. DOI: 10.1007/s41870-018-0157-5.
8. **Singh, M., Bansal, D., & Sofat, S. (2018).** Who is Who on Twitter–Spammer, Fake or Compromised Account? A Tool to Reveal True Identity in Real-Time*. Cybernetics and Systems,* Vol. 49, No. 1, pp. 1–25. DOI: 10.1080/01969722.2017.1412866.
9. **Stukal, D., Sanovich, S., Bonneau, R., & Tucker, J.A. (2017).** Detecting bots on Russian political Twitter. *Big data,* Vol. 5, No. 4, pp. 310–324. DOI: 10.1089/big.2017.0038.
10. **Preoţiuc-Pietro, D., Liu, Y., Hopkins, D., & Ungar, L. (2017).** Beyond binary labels: political ideology prediction of twitter users. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics,* pp. 729–740. DOI: 10.18653/v1/P17-1068.
11. **Guimaraes, R.G., Rosa, R.L., De Gaetano, D., Rodriguez, D.Z., Bressan, G. (2017).** Age Groups Classification in Social Network Using Deep Learning, *IEEE Access,* Vol. 5, pp. 10805–10816. DOI: 10.1109/ACCESS.2017.2706674.
12. **Tavares, G.M., Mastelini, S.M., & Barbon-Jr, S. (2017).** User Classification on Online Social Networks by Post Frequency*. CEP*, pp. 464–471.
13. **Lee, K., Eoff, B.D., & Caverlee, J. (2011).** Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter. *ICWSM,* pp. 185–192.
14. **Cresci, S., Di-Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2017).** The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race*. Proceedings of the 26th International Conference on World Wide Web*, pp. 963–972.
15. **Tavares, G., & Faisal, A. (2013).** Scaling-laws of human broadcast communication enable distinction between human, corporate and robot twitter users. *PloS one*, Vol. 8, No. 7, pp. 1–12. DOI: 10.1371/journal.pone.0065774.
16. **Chawla, N.V., Bowyer, K.W., Hall, L.O., & Kegelmeyer, W.P. (2002).** SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, Vol. 16, pp. 321–357. DOI: /10.1613/jair.953.

Article received on 20/10/2018; accepted on 01/02/2019. Corresponding author is Kheir Eddine Daouadi.