CEUR-WS.org/Vol-2279/LPKM2018_paper_2.pdf

Organization vs. Individual: Twitter User Classification.

Kheir Eddine Daouadi[1], Rim Zghal Rebaï[2] and Ikram Amous[3]

[1] MIRACL-FSEGS, Sfax University, 3000 Sfax, Tunisia.
[2,3] MIRACL-ISIMS, Sfax University, Tunis Road Km 10, 3021 Sfax, Tunisia.
{khairi.informatique,rim.zghal}@gmail.com
ikram.amous@enetcom.usf.tn

Abstract. This paper presents a novel technique for classifying user accounts on Twitter. The main purpose of our classification is to distinguish the patterns of users from those of individuals and organizations. However, such a task is non-trivial. Classic and consolidated approaches use textual features from Natural Language Processing (NLP) for classification. Nevertheless, such approaches still have some drawbacks like the computational cost and the fact that they depend on a specific language. In this work, we propose a statistical-based approach based on metadata of user profiles, popularity of posts and other statistical features in order to recognize the type of users without using the textual content. We performed a set of experiments over a twitter dataset and learn-based algorithms. This yielded an F-measure of 95.6% using the Random Forest algorithm and synthetic minority oversampling technique.

Keywords: Twitter User Classification, Individual Vs. Organization, Metadata of User Profile, Popularity of Posts Features.

1 Introduction:

Nowadays, social media attract the attention of several parties in the community such as organizations, researchers, politicians etc. This growing phenomenon has become an important part of our everyday life. Billions of users generate a huge amount of data on such social media which allow users to register through the creation of a profile account, following other members and sharing content. Several recent research studies were motivated by the emergence of social media such as author profiling [1, 2, 3], event detection [4], user recommendation [5] etc. Twitter is one of the leading social media, which allow users to post 280 characters known as tweets. This free micro blogging has been used not only by individuals to generate and share various types of content, but also by organizations to spread information and engage users. In [6], the authors showed that the presence of organizations makes up 9.4 % of the accounts on Twitter.

In this paper, we attempt to classify user accounts on Twitter from those of individuals and organizations. The ability to classify the patterns from both user types is needed for developing many applications such as recommendation engines, products opinion mining tools etc. Classical and consolidated approaches of text mining use textual features from NLP for classification. Nevertheless, such approaches still have some drawbacks like the computational cost and the fact that they depend on a specific language. In this context, the main contribution of this work is to demonstrate the importance of using statistical parameters in order to perform the individual Vs. Organization classification task. We illustrated the benefits from using both metadata of user profiles and metadata of posts in order to classify user accounts from those of individuals and organization.

The remainder of this paper is organized as follows. Section 2 presents the related works. In Section 3, our proposed method is detailed. In Section 4, experimental results are discussed. Finally, the conclusion is presented in Section 5.

2 Related works

Several works of author profiling have been done on a single and specific perspective such as organization and individual classification [3, 6, 7], political orientation [8], bot detection [9], age prediction [10]. Twitter user classification approaches involves three major types of research, namely statistical-based approaches, content-based approaches and hybrid approaches.

First, statistical-based approaches that used statistical parameters such as the metadata of user profiles [9], the time distribution between posts [11], post frequency [12]. Second, content-based approaches that used only the textual features from (NLP) such as n-grams [2], n-grams, linguistics, informal language and twitter stylistic features [7], word2vec neural language model with a Convolutional Neural Network (DCCN) [10], the semantic Long Short Term Memory (LSTM) [13]. Third, hybrid approaches that aim to combine the statistical-based and content-based approaches. Different classes of features were used together such as posts content, social and temporal features [3], posts content, stylistic, structural and behavioral features [6], metadata of user profiles and derived tweeting behavior features [1].

Although content-based and hybrid approaches were heavily used in the literature, these approaches also have their drawbacks such as the computational cost and the fact that they depend on a specific language. On the other hand, statistical-based approaches use language-independent features without using the content parameters. Therefore, we proposed a statistical-based approach in order to distinguish Twitter user accounts from those of individuals and organizations. Our work is similar to that of [3, 6, 7, 14, 15], but these previous works used textual content features in the classification task. In fact, messages from social media are imprecise, short, written in informal language etc. In another similar work in [16], the authors used network features. This type of approach is very expensive and more complex than other types of approaches.

To overcome the challenge previously mentioned, we proposed a statistical-based approach in order to classify the patterns of users from those of individuals and organizations. The main advantage from our proposed approach is that using language independent features without using the content features while achieving a high accuracy of classification. We designed a rich set of statistical features which proved their importance during this work. In the following section, we will discuss our proposed method.

3 Proposed method

Our proposed approach contains three main phases, namely data collection, features extraction and classification. The following subsection discusses the progression of our proposed approach.

3.1 Data collection

In order to create our data sets, we used Twitter timeline (API), which allows collecting the K most recent posts of such user. Previous studies have shown that 200 posts are typically sufficient and it is enough for predicting Twitter user characteristics [6]. We used the datasets published in [6][1]. **Table.1,** presents an overview of the used datasets. These datasets contain a user_id with a label like "ind" or "org" (individual or organisation). We used the user_id with the Ttwitter timeline API in order to collect the 200 most recent posts of each user. The posts came with a little attribute; they were Timestamp, User and Tweet. The Timestamp attribute which contain the time and date of posting. The Tweet attribute which contain metadata about the post. The User attribute which contain metadata of the user profile.

We observed the user accounts' statuses via the statuses/user_Timeline API endpoint. These statuses can take on one of four values, namely active account, removed account, private account and suspended account. Since we could not have access to the posts of suspended, removed and private accounts, in this work we used only active user accounts. This way, the next step of the method could follow.

Table 1. An overview of the used datasets.

	Total number labeled accounts	Number of Removed, suspended, or private accounts	Number of Active accounts
Individual	18362	3065	**15297**
Organization	1911	96	**1815**

3.2 Features extraction

The main purpose of this phase is to build a feature vector for the user account. Three types of features were proposed, these included parameters from the metadata of the user profile, the popularity of post features and other statistical features.

Metadata of user profile features. Different parameters from metadata of the user profile are exploited. The 'User' attribute described in (Section 3.1) was used in order

[1] Available at http://networkdynamics.org/software/

to capture information from the metadata of the user profile; these included binary and numerical features. First, the numerical features included the number of followers, the number of followings, the ratio of followers to followings, the number of lists that this user is a member of, the number of tweets that this user has 'liked', the number of all posts per day, and the total number of posts. Second, the binary features include whether the user has enabled the possibility of geo-tagging their Tweets, whether the user has declared their location, whether the user has provided a URL in association with their profile, whether the profile has a description and whether the user has a verified account.

Popularity of post features. In addition to the metadata of user profile features, we used parameters from the metadata of posts; these features represent how much users interact with the posts of the user. 12 popularity of post features were proposed. These are calculated using (Eq. 1), (Eq. 2), (Eq. 3), and (Eq. 4).

$$fav(P) = (\sum_{k=0}^{n} number\ of\ favorites\ (P)) \tag{1}$$

$$ret(P) = (\sum_{k=0}^{n} number\ of\ retweets(P)) \tag{2}$$

$$pop\_fav(P) = \frac{Average\_number\_of\_favorites(P)}{number\ of\ followings} \tag{3}$$

$$pop\_ret(P) = \frac{Average\_number\_of\_retweets(P)}{number\ of\ followings} \tag{4}$$

where: P may be (tweets, retweets, or replies), n represents the number of P, 'number of retweets' is the number of users reposting the post of the user, 'number of favorites' is the number of users favoring the post of the user, 'number of followings' is the number of followings of the user.

Statistical features. Other parameters from the metadata of posts were used; these include the number of posts that contain a mention to another user, the number of quoted tweets (i.e. retweets with a comment), the number of posts that were posted in association with a place, the number of geo-tagged posts, and the average interval between posts.

3.3 Classification

The main purpose of this phase is the classification task. We tested a set of supervised machine-learning algorithms in order to make a decision as to which classifier performs better with our proposed features. Finally, we constructed our predictive model in order to use it in the process of individual vs. organization classification task. The following section discusses the experimental settings and results.

4 Experiment and evaluation

To verify the validity of our proposed framework, we performed several experiments. We used a set of supervised machine-learning classifiers, namely Random Forest (RF), Simple Logistic (SL), Logit Boost (LB), Bagging (B) and Multilayer Perceptron (MLP). We performed a 10-fold cross validation in order to test the performance measurement of each classifier. The tests were implemented using the Waikato Environment for Knowledge Analysis (WEKA). To evaluate the performance of our proposed method, we used four metrics, namely Recall, Precision, F-measure and Accuracy. **Table.2** shows the performance measurement of each classifier. The Random Forest algorithm outperformed the other ones by achieving an F-measure of 93.0%.

Table 2. 10-fold cross validation performance measurement.

	Avg._Precision%	Avg._Recall%	Avg._f-measure%
RF	**92.9**	**93.3**	**93.0**
SL	89.4	90.9	89.3
LB	91.9	92.5	92.1
B	92.6	93.1	92.8
MLP	89.2	90.3	89.6

Table 3. F-measure results for both classes and in two conditions.

	Balanced datasets			Full datasets		
	Ind.	Org.	%f-measure	Ind.	Org.	%f-measure
Majority class	0	100	50	100	0	89.39
Proposed method	**89.2**	**89.5**	**89.3**	**96.3**	**65.1**	**93.0**

Given the skewed distribution of the user account types, we compared the performance of the proposed approach with the majority-class baseline, which classifies all the instances in the class that contains the majority samples (in our case the ‘organization’ class). We evaluated our proposed approach in two conditions, using balanced datasets and imbalanced datasets. As shown in **Table.3,** we outperformed the majority-class baseline in both balanced datasets and imbalanced datasets.

In order to evaluate the importance of features, we also evaluated the F-measure of the corresponding models:

- (MUP): metadata of user profile features F-measure=91.4%.
- (PP): popularity of post features F-measure=90.2%.
- (S): statistical features F-measure=85.9%.
- (MUP) + (PP): metadata of user profile and popularity of post features F-measure=92.8%.
- (MUP)+(S): metadata of user profile and statistical features F-measure=91.9 %.

- (PP)+(S): popularity of posts features and statistical features F-measure=90.8 %.
- (MUP)+(PP)+(S): all proposed features F-measure=93.0%.

We also used Synthetic Minority Oversampling Technique (SMOTE) [17] in order to balance the datasets. (SMOTE) algorithm allows generating a new sample based on the feature vector of the minority class (in our case the 'organization' class) and is a powerful method that has been successful in many domains [18]. Significant performance gains were observed on balancing datasets. As shown in **Table.4,** when we used our proposed features with (SMOTE) technique, the accuracy reached 95.64%. Our proposed framework outperformed previous ones by using language independent features without using the content parameters while achieving similar accuracy of classification than previous works. Moreover, our proposed framework assures multi-dialectal and multi-lingual organization detection, unlike previous proposed frameworks, which were dependent on a specific language. In addition, we designed a minimal number of statistical features that demonstrated their utility in classifying the patterns of users into humans and organizations.

Table 4. Comparison with similar works.

Work	Accuracy%
[6]	95.50
[15]	93.40
Our proposed features	93.10
Our proposed features + SMOTE	**95.64**

5 Conclusion

This paper presented a statistical-based approach for classifying user accounts on online social networks. Our proposed approach differs from the previous ones in terms of the features used in the classification task. Previous works used the textual content of posts as features. However, they had some drawbacks such as time consumption and dependence on a specific language. Our work outperforms the previous ones in two dimensions. First, in previous works, features extraction needed multilingual and multidialectal resources, but in our proposed approach, feature extraction is performed without taking into account the user's language. Second, the user may post various types of content (i.e. text, images, videos etc.); previous works fail with the user who posts multimedia posts. In contrast, our proposed approach does well with the user who uses a multimedia content since it uses only the metadata of the user profile and metadata of the posts' features. Our proposed framework achieve high F-measure result of 95.6%. We demonstrated that using a minimal number of statistical features are sufficient to classify user accounts accurately and quickly. Although we have chosen only the Twitter platform but our proposed approach is generic and can be translated into any social networks such as Facebook and Instagram.

References

1. Kim, A., Miano, T., Chew, R., Eggers, M., Nonnemaker, J.: Classification of Twitter Users Who Tweet About E-Cigarettes. In: JMIR Public Health and Surveillance, vol. 3, no 3, p. e63, (2017).
2. Abbassi, A., Mechti, S., Belguith, L. H., Faiz, R. Author Profiling for Arabic Tweets based on n-grams. The first Conference on Language Processing and Knowledge Management (LPKM), (2017).
3. Oentaryo, R. J., Low, J. W., Lim, E. P.: Chalk and Cheese in Twitter: Discriminating Personal and Organization Accounts. In: Hanbury A., Kazai G., Rauber A., Fuhr N. (Advances in Information Retrieval). European Conference on Information Retrieval ECIR, Lecture Notes in Computer Science, vol 9022, pp. 465-476. Springer, Cham (2015).
4. Troudi, A., Zayani, C. A., Jamoussi, S., & Amor, I. A. B.: A New Mashup Based Method for Event Detection from Social Media. Information Systems Frontiers, 1-12. (2018).
5. Kalaï, A., Zayani, C. A., Amous, I., Abdelghani, W., & Sèdes, F.: Social collaborative service recommendation approach based on user's trust and domain-specific expertise. Future Generation Computer Systems (355-367), (2018).
6. McCorriston, J., Jurgens, D., Ruths, D.: Organizations Are Users Too: Characterizing and Detecting the Presence of Organizations on Twitter. In: 9th International Conference on Web and Social Media ICWSM. pp. 650-653, The AAAI Press, UK, (2015).
7. De Silva, L., Riloff, E.: User Type Classification of Tweets with Implications for Event Recognition. ACL ,98-108 (2014).
8. Preoţiuc-Pietro, D., Liu, Y., Hopkins, D., Ungar, L.: Beyond binary labels: political ideology prediction of twitter users. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 729-740, Canada (2017).
9. Ferrara, E.: Disinformation and social bot operations in the run up to the 2017 French presidential election. First Monday, 22(8). (2017).
10. Guimaraes, R. G., Rosa, R. L., De Gaetano, D., Rodriguez, D. Z., Bressan, G.: Age Groups Classification in Social Network Using Deep Learning, IEEE Access, 1-11 (2017).
11. Tavares, G., Faisal, A.: Scaling-laws of human broadcast communication enable distinction between human, corporate and robot Twitter users. PloS one, vol. 8, no 7, p. e65774, (2013).
12. Tavares, G.M., MASTELINI, S.M., BARBON JR, S.: User Classification on Online Social Networks by Post Frequency. CEP, vol. 86057, pp. 970-977 (2017).
13. Jain, G., Sharma, M., & Agarwal, B. Optimizing semantic LSTM for spam detection. International Journal of Information Technology, 1-12, (2018).
14. Kim, S. M., Paris, C., Power, R., & Wan, S.: Distinguishing Individuals from Organisations on Twitter. In Proceedings of the 26th International Conference on World Wide Web Companion (pp. 805-806), (2017).
15. Wood-Doughty, Z., Mahajan, P., & Dredze, M.: Johns Hopkins or johnny-hopkins: Classifying Individuals versus Organizations on Twitter. In Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (pp. 56-61), (2018).
16. Alzahrani, S., Gore, C., Salehi, A., & Davulcu, H.: Finding Organizational Accounts Based on Structural and Behavioral Factors on Twitter. In International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation (pp. 164-175), (2018).
17. Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P.: SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357. (2002).

18. He, H., & Garcia, E. A. : Learning from imbalanced data. IEEE Transactions on Knowledge & Data Engineering, (9), 1263-1284. (2008).