Person, Organization, or Personage: Towards User Account Type Prediction in Microblogs

Ivan Samborskii[1,2], Andrey Filchenkov[1],
Georgiy Korneev[1], and Aleksandr Farseev[2]

[1]ITMO University,
49 Kronverksky Pr., St. Petersburg, Russia, 197101
[2]National University of Singapore
13 Computing Dr., Singapore, 117417
{samborsky@rain.ifmo.ru,afilchenkov@corp.ifmo.ru,kgeorgiy@rain.ifmo.ru,
farseev@u.nus.edu}

Abstract. During the past decade, microblog services have been extensively utilized by millions of business and private users as one of the most powerful information broadcasting tools. For example, Twitter attracted many social science researchers due to its high popularity, constrained format of thought expression, and the ability to reflect actual trends. However, unstructured data from microblogs often suffer from the lack of representativeness due to the tremendous amount of noise. Such noise is often introduced by the activity of organizational and fake user accounts that may not be useful in many application domains. Aiming to tackle the information filtering problem, in this paper, we classify Twitter accounts into three categories: "Personal", "Organization" and "Personage". Specifically, we utilize various text-based data representation approaches to extract features for our proposed microblog account type prediction framework "POP-MAP". To study the problem at a cross-language level, we harvested and learned from a multi-lingual Twitter dataset, which allows us to achieve better classification performance, as compared to various state-of-the-art baselines.

Keywords: Twitter, social media, profile learning, natural language processing, account type classification

1 Introduction

Web scientists use social media as a rich source of information about users' individuality, behavior, and preferences [9, 13, 15, 25]. It is used to recover user profile [3, 10, 12] and make targeted recommendation [11, 19]. The availability of these personal user attributes allows them to compete with traditional sociologists, epidemiologist and political experts in such tasks as voting outcome prediction [14, 24], disease outbreaks prediction [7, 17], or group population visualization [1]. However, the representativeness of the data in most of web science studies is extremely low due to the significant level of noise.

The noise in social media is often related to the fact that not all accounts represent a real human. For example, this can be caused by specific bots that mimic human behavior while being governed by an algorithm or another human. Many works are devoted to detecting such accounts [4–6, 26]. At the same time, some microblog accounts may not represent a person, but be related to something else: accounts of corporations (Adidas[1]), banks (DBS bank[2]), museums (The State Hermitage Museum[3]), animals (Grumpy Cat[4]), or personages (such as Harry Potter[5]). These accounts represent a certain subject that may or may not be equipped with the aforementioned personal user attributes (i.e. demographics). However, most of them are irrelevant to social studies.

Nevertheless, most of the existing social media analysis studies either do not perform irrelevant user account filtering [11, 12], perform it manually [16, 22], or do not utilize openly available user-generated data [23, 20]. For example, Tavares et al. [23] presented a method to classify personal and corporate accounts, which solved the problem with 84.6% accuracy. However, the authors did not use user-generated content, which may result in a sub-optimal performance due to the lack of data representativeness. At the same time, Oentaryo et al. [20] utilized contextual, social, and temporal features, which allowed for achieving 91% account type classification accuracy by gradient boosting algorithm. However, the employed data types are often not available for public use, which constrains the applicability of the proposed approach to real-world scenario.

Indeed, in our study, we perform the task of microblog user account type inference based on textual user-generated content only, which makes it applicable in the real-world settings. We assume that textual data is sufficient for achieving high classification performance and train our-proposed **"POP-MAP"** framework to perform "**P**erson"–"**O**rganization"–"**P**ersonage" **M**icroblog **A**ccount **P**rediction.

2 On Microblog Account Typization

Microblog is a specific type of social media resource, which allows its users to share short status updates to their subscribers. One of the most well-known microblogs is Twitter, where messages (statuses) are publicly accessible in contrast to other big social networks, such as Facebook, and the length of message cannot exceed 140 symbols, which makes its posts standardized and rarely representing more than one topic [28].

According to Barone et al. [2], each Twitter account belongs to one of the following five types:

[1] `http://twitter.com/adidas`

[2] `http://twitter.com/dbsbank`

[3] `http://twitter.com/hermitage_eng`

[4] `http://twitter.com/realgrumpycat`

[5] `http://twitter.com/arrypottah`

1. **Corporate Account**, which is typically a company news feed: Facebook[6], Google[7], Yandex[8], and VKontakte[9].
2. **Corporate-led Persona Account**, which is a corporate account that includes both personal and business sides. For example, an account of online shop Zappos[10] is Tony Hsieh's account, in fact.
3. **Strictly Personal Account** is an account representing an individual microblog user.
4. **Business/Personal Hybrid Account** is a mixture of the personal account and professional account types, where most of the tweets contain information about its user, but also a considerable amount of tweets is dedicated to the users professional interests. Accounts of famous people usually belong to this type, for example, Pavel Durov[11] or Jimmy Wales[12] accounts.
5. **Personage Account**, which is the personage-based account that typically is an animal, plant, or fictional hero.

In this paper, we adopt three most popular accounts types from the above categorization: organization account, personal account, and personage account. The other two hybrid types are considered to be a part of the selected ones, so that all the Corporate-led Persona Accounts are treated as organization accounts, while Business/Personal Accounts are considered to be personal accounts.

3 Feature Extraction

Classification algorithms strongly depend on features which describe objects. Thus, feature engineering is a key step in solving most of the data mining problems. In this section, we define all the features we used to describe a Twitter account.

Words frequency. Individual users typically use everyday vocabulary in their tweets, while organizations may adopt a domain-specific vocabulary that can be a good indicator of the organization account type. In accordance with this assumption, we use the following features:

- average word frequency among all words in tweet;
- average word frequency among all words in all user's tweets.

We utilized Sharov's Frequency Dictionary[13] and Word frequency data[14] for obtaining general usage frequency of Russian and English words respectively.

[6] `http://twitter.com/facebook`

[7] `http://twitter.com/google`

[8] `http://twitter.com/yandex`

[9] `http://twitter.com/vkontakte`

[10] `http://twitter.com/zappos`

[11] `http://twitter.com/durov`

[12] `http://twitter.com/jimmy_wales`

[13] `http://dict.ruslang.ru/freq.php`

[14] `http://www.wordfrequency.info`

Spelling mistakes. It is well-known that individual user accounts tend to post more grammatical mistakes/misspellings as compared to properly-maintained organizational accounts. Inspired by this phenomenon, we utilized Language-Tool[15] to extract the number of mistakes/misspellings per account.

Hashtags. Hashtags are often used for grouping microblog messages and improvement of Twitter search. Personal accounts are characterized by extensive use of hashtags to express their thoughts, feelings, as compared to corporate accounts. We thus extracted the following hashtag-based features:

- average number of unique hashtags per account;
- average number of hashtags per tweet;
- average length of hashtag per tweet.

Users' mentions. Similar to hashtags, user mentions spread in social networks. However, we cannot expect personage accounts to use them often due to the lower number of actual social ties between them and individual Twitter users. To incorporate this aspect, we extracted the following user mention features:

- average number of unique mentions per account;
- average number of mentions per tweet;
- average length of mention per tweet.

Tweet/word length. Many acronyms (i.e. "gotcha" meaning "I got you") widespread among users of social networks. The reason is that they are useful to fit in more information into short twitter message. These acronyms, however, are not popular among organizational twitter accounts. Therefore, we extracted the following features representing text length:

- average length of word per account;
- average length of tweet per account.

Part of speech (POS). To reflect different styles of language use, we included features related to words' POS. The following POS groups have been identified:

- noun;
- verb;
- personal pronoun;
- pronoun (others);
- adjective;
- adverb;
- preposition, conjunction, particle;
- adverb + adjective;
- adverb + adverb.

For each group, we then calculated the following features:

[15] `http://languagetool.org`

- average number of groups per account;
- average number of groups per tweet;
- average number of negative particles per account.

Personal words. Accounts belonging to people or personages can be easily identified by the so-called personal words. Inspired by this fact, we extracted "average number of personal words per account" feature.

Symbols. Similarly to previous studies, for each symbol in Table 1, we calculated the following features:

- average number of signs used per tweet;
- average number of unique signs per tweet;
- average number of tweets with the sign per account;
- average number of signs per tweet;
- average number of tweets with signs per account;
- average number of unique signs per account.

Table 1. Symbols that are used to calculate features

!	@	#	$	%	&	*	(
)	_	+	-	=	˜	‘	,
.	<	>	/	?	\	|	;
:	’	[	]	{	}	№	”

Emoticons. Similar to the symbol features, for each group of emoticons in Table 2, we calculated emotion features:

- average number of emoticons using per tweet;
- average number of tweets with emoticon per account;
- average number of emoticons per tweet;
- average number of unique emoticons per account.

Table 2. Emoticons groups that are used to calculate features

:) :-) =)	:(:-(=(	;) ;-)
8) 8-) %) %-)	:’) :’-) :,) :,-) =’) =,)	:’(:’-(:,(:,-(=’(=,(
:* :-* =*	o_o o_O =O =0 0_0	:-b :-p :b :p =p =b ;b ;p
:D xD =D ;D	:-[=[:3 >_<	

Vocabulary uniqueness. Organization accounts on Twitter are often created to be used for specific applications. For example, Yandex.Taxi[16] is designed

[16] http://twitter.com/yandextaxi

to support taxi services, while Yandex.Market[17] is related to e-commerce services aggregation. Every specific usage domain reduces the diversity of words in organizations' microblog accounts. Based on this assumption, we extracted the following vocabulary-uniqueness features:

- average number of unique words per account;
- average number of words not from a vocabulary per account.

Hyperlinks. Users often post URLs to third-party resources, such as events, pictures, etc. The URL usage can be a good indicator of individual user accounts. Based on this assumption, we extracted the features below:

- average number of links per account;
- average number of tweets with links per account.

Twitter-specific features. Organization accounts are often characterized by a large number of subscribers (followers), but a relatively small number of subscriptions (following). This is also the case of popular personage accounts. Also, it is worth mentioning that corporate accounts are often verified, which often does not hold for personal accounts, while personage accounts are almost never verified.

- number of subscribers;
- number of subscriptions;
- if the account is verified;
- average number of "favorite" tweets.

Overall, there we suggest 136 features for Twitter account type classification. It is worth mentioning that some of them (such as usage of hashtags, hyperlinks, and personal words) were never adopted before and, thus, they are one of the contributions of this study.

4 Experiment setup

4.1 Data collection

Due to the lack of publicly available datasets on Twitter account type inference, we collected our dataset. To do so, we developed a crawler for downloading last $n = 500$ tweets of each specified user, where the list of account names was created manually.

4.2 Utilized machine learning methods

We employed the following commonly-utilized classification baselines that are implemented as part of WEKA[18] machine learning library: k nearest neighbors,

[17] `http://twitter.com/yandexmarket`

[18] `http://www.cs.waikato.ac.nz/ml/weka/`

Naïve Bayes classifier, Support Vector Machines (SVM) classifier, Decision Trees (its C4.5 version), and Random Forest. These algorithms were applied to the profiles represented by our-extracted POP-MAP features that were presented in Section 3.

We used several feature selection (FS) algorithms [27] to select only representative features:

- dependency-based elimination, such as: *CFS-BiS, CFS-GS, CFS-LS, CFS-RS, CFS-SBS, CFS-SFS, CFS-SWS, CFS-TS*;
- consistency-based elimination, such as: *Cons-BiS, Cons-GS, Cons-LS, Cons-RS, Cons-SBS, Cons-SFS, Cons-SWS*;
- *Significant* algorithm, which is based on estimating feature "significance";
- *ReliefF* measures feature importance based on comparison to similar objects of the same class.

Also, we utilized the well-known dimensionality reduction algorithm PCA that is also implemented in WEKA.

To evaluate the prediction performance by using the two well-adopted evaluation measures: accuracy and F-measure. We organized model evaluation using 5-fold cross validation.

5 Experiments on Russian Text Corpora

We have collected the sample consisting of 298 Russian personal accounts, 160 Russian organization accounts and 151 Russian personage accounts by the tool and method, described in the previous section.

5.1 Comparing Baselines

Since there are no existing solutions for the problem of microblog account type inference, we consider standard text classification techniques as our baselines:

Naïve Bayes (*NB*) — simple Naïve Bayes classifier with minor preprocessing (all hyperlinks are removed and letters are changed to lowercase) [8].

Classifier with stemmer (*Stemmer*) — *NB* with Porter's stemmer applied [21].

Classifier with emoticons (*Emoticon*) — the classifier from Lin [18] work, which determines chat users' age and gender based on emoticons in users' posts. To implement this method, we identified 500 different emoticons.

The baseline results are presented in Table 3. As we can see, stemming has expectedly improved *NB* but outperformed *Emoticon*. This is possibly due to organizations use less formal language in Twitter than we expected.

5.2 Comparing approaches trained POP-MAP features

POP-MAP without feature selection

Table 3. Results of baselines for account classification for the Russian language

Classifier	Accuracy	F-measure
NB	0.711	0.678
Stemmer	**0.749**	**0.702**
Emoticon	0.511	0.519

We conducted experiments using the setup described in Section 4 on the collected dataset. The results are presented in Table 4. The best performance was shown by Random Forest, which is consistent with previous study [12] and can be explained by its feature selection ability.

Table 4. Results for account classification for the Russian language without feature selection

Classifier	Accuracy	F-measure
kNN	0.770	0.761
Naïve Bayes	0.645	0.688
SVM	0.490	0.219
Decision Tree	0.792	0.789
Random Forest	**0.862**	**0.858**
Best baseline (Stemmer)	0.749	0.702

POP-MAP with feature selection

To improve classification performance, we applied dimensionality reduction algorithms described in Section 4. First, we applied PCA. As we can see from Table 5, PCA did not improve the classification performance.

Table 5. Results for account classification for the Russian language with PCA

Classifier	Accuracy	F-measure
kNN	0.719	0.708
Naïve Bayes	0.495	0.547
SVM	0.820	0.815
Decision Tree	0.720	0.712
Random Forest	0.806	0.801
Random Forest (no FS)	**0.862**	**0.858**

Then we picked the best feature selection algorithm for each classifier with respect to the resulting performance. The evaluation results are presented in Table 6. As it can be seen, feature selection improved performance of all the models. However, Random Forest kept its position of the best classifier, which can be explained by its additional built-in feature selection ability.

Table 6. Results for account classification for the Russian language with feature selection

Classifier	Accuracy	F-measure	FS alg.	number of features
kNN	0.799	0.792	CFS-RS	29
Naïve Bayes	0.795	0.790	ReliefF	44
SVM	0.639	0,616	Cons-SS	10
Decision Tree	0.813	0.808	CFS-BiS	23
Random Forest	**0.878**	**0.874**	CFS-TS	23
Random Forest	0.862	0.858	—	136

5.3 Results summary

From the Table 6, it can be seen that the best performance was achieved by Random Forest classifier on the CFC-TS-preprocessed data. The contingency matrix is presented in Table 7 shows us that the resulting classifier makes a small number of misclassifications, while the most complex task for it is to distinguish personal accounts and personage accounts. This can be explained by the similar nature of these two types of accounts, which conforms well with manual comparison of such accounts.

Table 7. Contingency table of the best classifier for the Russian language

	Person	Organization	Personage
Person	55	1	6
Organization	1	32	1
Personage	3	0	23

We used mutual information (MI) measure to estimate feature importance. The most valuable features are average number of personal words per account (0.679), average number of personal pronouns per tweet (0.633), average number of personal words per tweet (0.472), average number of links per account (0.402), and a number of subscriptions (0.378). Among other features with MI greater than 0.2, seven are POS features, one is tweets with links per account, two are tweets length features.

As we can see, the most important features are related to personality and references. We may expect the same situation and for the English language.

6 Experiments on English Text Corpora

6.1 Dataset

To perform evaluation on English corpora, we have collected the sample consisting of 281 English personal accounts, 130 English organization accounts and 130 English personage accounts using the tool and method described in Section 3.

6.2 Results

In this setup, we tested only Random Forest since it has shown the ultimate performance for the Russian language. The best-achieved result was after applying Con-GS algorithm selecting 44 features and resulting in 0.894 of accuracy and 0.879 of F-measure. The contingency table is presented in Table 8. The resulting classifier also makes only a small amount of mistakes. As we can see, the classifier for English corpora outperforms the best one for Russian corpora classifier.

Table 8. Contingency table of the best classifier for the English language

	Person	Organization	Personage
Person	52	0	4
Organization	1	23	3
Personage	1	1	24

The most valuable features with respect to the MI are: number of subscriptions (0.709), average number of personal words per account (0.516), if the account is verified (0.479), average number of tweets with links per account (0.290), average number of unique signs per account (0.274). Among other features with MI greater than 0.2, four are symbol features, one is number of subscribers, one is average number of hyperlinks per tweet, and one is average length of tweets.

We can see that personal words are also the strong feature besides Twitter-specific features. However, POS-tagged features are not at the top as in Russia. Instead, symbol-specific features are useful for English.

6.3 Results for binary classification

We also compared our results with results, reported in [23], where authors classified microblog accounts only into personal and corporate types. To do so, we selected only personal and organization accounts from the initial datasets and run the best-built classifiers for English and Russian. The results of the comparison are presented in Table 9. As it can be seen, the POP-MAP results on both the Russian and English corpora are similarly high and significantly surpass the behavior-based approach.

Table 9. Results for classification of personal and organization accounts

Algorithm	Accuracy	F-measure
User's behavior [23]	0.846	—
POP-MAP for English	0.975	0.947
POP-MAP for Russian	0.969	0.966

7 Conclusion

In this paper, we addressed the problem of Twitter account classification. We described 136 features, which we then used in different classification models. We run experiments on corpora of Russian and English tweets and achieve similarly high classification performance for both languages with the Random Forest model. However, we discovered that there is a difference in text feature importance for two languages, while Twitter-specific features have the same importance. The only exception is a strong feature related to personal words that are useful in both English and Russian.

Acknowledgments

This research is supported by the Government of Russian Federation, Grant 074-U01 and the Singapore National Research Foundation under its International Research Center @ Singapore Funding Initiative and administered by the IDM Programme Office.

References

1. Aramaki, E., Maskawa, S., Morita, M.: Twitter catches the flu: detecting influenza epidemics using twitter. In: Proceedings of the conference on empirical methods in natural language processing. pp. 1568–1576. Association for Computational Linguistics (2011)
2. Barone, L.: Which type of twitter account should you create? http://smallbiztrends.com/2010/02/types-of-twitter-accounts.html (2010), [Online; accessed 2016-04-15]
3. Bartunov, S., Korshunov, A., Park, S.T., Ryu, W., Lee, H.: Joint link-attribute user identity resolution in online social networks. In: Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, Workshop on Social Network Mining and Analysis. ACM (2012)
4. Boshmaf, Y., Muslukhov, I., Beznosov, K., Ripeanu, M.: Design and analysis of a social botnet. Computer Networks 57(2), 556–578 (2013)
5. Cao, Q., Sirivianos, M., Yang, X., Pregueiro, T.: Aiding the detection of fake accounts in large scale social online services. In: Presented as part of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12). pp. 197–210 (2012)

6. Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S.: Who is tweeting on twitter: human, bot, or cyborg? In: Proceedings of the 26th annual computer security applications conference. pp. 21–30. ACM (2010)
7. Culotta, A.: Towards detecting influenza epidemics by analyzing twitter messages. In: Proceedings of the first workshop on social media analytics. pp. 115–122. ACM (2010)
8. Deitrick, W., Miller, Z., Valyou, B., Dickinson, B., Munson, T., Hu, W.: Gender identification on twitter using the modified balanced winnow. Communications and Network 4(3), 1–7 (2012)
9. Farseev, A., Akbari, M., Samborskii, I., Chua, T.S.: 360 user profiling: past, future, and applications by aleksandr farseev, mohammad akbari, ivan samborskii and tat-seng chua with martin vesely as coordinator. ACM SIGWEB Newsletter (Summer), 4 (2016)
10. Farseev, A., Chua, T.S.: Tweetfit: Fusing sensors and multiple social media for wellness profile learning. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. AAAI (2017)
11. Farseev, A., Kotkov, D., Semenov, A., Veijalainen, J., Chua, T.S.: Cross-social network collaborative recommendation. In: Proceedings of the ACM Web Science Conference. p. 38. ACM (2015)
12. Farseev, A., Nie, L., Akbari, M., Chua, T.S.: Harvesting multiple sources for user profile learning: a big data study. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. pp. 235–242. ACM (2015)
13. Farseev, A., Samborskii, I., Chua, T.S.: bbridge: A big data platform for social multimedia analytics. In: Proceedings of the 2016 ACM Conference on Multimedia. pp. 759–761. ACM (2016)
14. Filchenkov, A.A., Azarov, A.A., Abramov, M.V.: What is more predictable in social media: Election outcome or protest action? In: Proceedings of the 2014 Conference on Electronic Governance and Open Society: Challenges in Eurasia. pp. 157–161. ACM (2014)
15. Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T., Weitzner, D.: Web science: an interdisciplinary approach to understanding the web. Communications of the ACM 51(7), 60–69 (2008)
16. Kafeza, E., Kanavos, A., Makris, C., Vikatos, P.: T-pice: Twitter personality based influential communities extraction system. In: 2014 IEEE International Congress on Big Data. pp. 212–219. IEEE (2014)
17. Lee, K., Agrawal, A., Choudhary, A.: Real-time disease surveillance using twitter data: demonstration on flu and cancer. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1474–1477. ACM (2013)
18. Lin, J.: Automatic author profiling of online chat logs. Ph.D. thesis, Monterey, California. Naval Postgraduate School (2007)
19. Lin, J., Sugiyama, K., Kan, M.Y., Chua, T.S.: Addressing cold-start in app recommendation: latent user models constructed from twitter followers. In: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. pp. 283–292. ACM (2013)
20. Oentaryo, R.J., Low, J.W., Lim, E.P.: Chalk and cheese in twitter: Discriminating personal and organization accounts. In: European Conference on Information Retrieval. pp. 465–476. Springer (2015)
21. Porter, M.F.: An algorithm for suffix stripping. Program 14(3), 130–137 (1980)

22. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E., et al.: Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one 8(9), e73791 (2013)
23. Tavares, G., Faisal, A.: Scaling-laws of human broadcast communication enable distinction between human, corporate and robot twitter users. PloS one 8(7), e65774 (2013)
24. Tsakalidis, A., Papadopoulos, S., Cristea, A.I., Kompatsiaris, Y.: Predicting elections for multiple countries using twitter and polls. IEEE Intelligent Systems 30(2), 10–17 (2015)
25. Varlamov, M., Turdakov, D.Y.: A survey of methods for the extraction of information from web resources. Programming and Computer Software 42(5), 279–291 (2016)
26. Wang, A.H.: Detecting spam bots in online social networking sites: a machine learning approach. In: IFIP Annual Conference on Data and Applications Security and Privacy. pp. 335–342. Springer (2010)
27. Wang, G., Song, Q., Sun, H., Zhang, X., Xu, B., Zhou, Y.: A feature subset selection algorithm automatic recommendation method. Journal of Artificial Intelligence Research (2013)
28. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.P., Yan, H., Li, X.: Comparing twitter and traditional media using topic models. In: European Conference on Information Retrieval. pp. 338–349. Springer (2011)