Organization Detection: how to find organization accounts on twitter?

1. **Introduction:**

   As the development of internet, more and more businesses turn their eyes on the social media. Researchers have done lots of work on social media user analysis (Shlomo Argamon et al.2005 and Yoram Bachrach et al. 2012) and user demography classification (David Huffaker 2004 and John D. Burger et al. 2011). There is barely research focusing on organization account detection yet, which has enormous potential in both academic and business fields. Actually in social media-based marketing, we care about personal users as latent consumers and expecting rule out the noises including organization users such as official account. Or at least, we have different strategies for different types of accounts. This is a common problem in industry, which drives me into the study of the organization detection project.

2. **Data:**

   Our data came from Haewoon Kwak et al. (2010), including 6499 "celebrity" profiles. All the accounts in this list have at least 10,000 followers so that the proportion of organization profiles would not be so unbalanced. There are 26 different features in each profile in data. All fields except gender are returned by user method (users/show) of Twitter API. The values of different features are separated by tab and every account occupies one line.

   Data format:
   numeric_id \t verified \t profile_sidebar_fill_color \t profile_text_color \t followers_count \t protected \t location \t profile_background_color \t utc_offset \t statuses_count \t description \t friends_count \t profile_link_color \t profile_image_url \t notifications \t profile_background_image_url \t screen_name \t profile_background_tile \t favourites_count \t name \t url \t created_at \t time_zone \t profile_sidebar_border_color \t following \t gender (infered by name) \n
   Because the profiles were unclassified, we needed to label all the profiles manually. We had our data classified into three different categories: personal, organizational and unknown. Organizational accounts are accounts run by personal and tweets about none-personal life and more than 1 people and tweets about none-personal life. Personal accounts are run by one person and tweets about his personal life. Unknown accounts are uncertain to tell even by

human beings due to all kinds of reasons. Here are some examples of accounts in different categories.

Figure 2.1: An account run by personal and tweets about personal life and labeled as a personal account.

Figure 2.2: An account run by personal and tweets about none-personal life and labeled as an organizational account.

Figure 2.3: An account run by more than 1 people and tweets about none-personal life and labeled as organizational accounts.

3. Methods:

We removed all unknown profiles and 4791 profiles remained, consists of 3214 personal accounts and 1577 organizational accounts.

We removed redundant low-level features in each profile and only kept followers_count, protected, statues_count, friends_count, nitifications, profile_background_title, favouritesc_count and following. We also kept the features of description, screen name and personal url for further study on high-level features.

We retrieved the descriptions of both personal profiles and organizational accounts and used them for high-level lexical feature extraction, which we will talk about later.

We also added another high-level feature: url_containing_name. We assign this feature a Boolean value indicating whether the personal url in profile contains the screen name. We came up with this idea intuitively according to the observation that most organizations have their own websites which has the address containing their name.

Linguistic feature extraction

We spent lots of time on syntactic analysis of the description. We thought there are some lexical differences between a personal description and an organizational description. So we put descriptions from personal and organizational accounts into two different files. Then we tried to use character n-grams and term (word) n-grams to analysis them.

To specify the location of the very first word in one sentence, we added one "#" before each sentence in term n-gram detection. Similarly, we added "#" at both the head and end of one word in character n-gram detection. After we calculated the cumulative occurrences of n-grams in profiles, we sorted them from the most frequent to the least. We post the results below.

Top frequent n-grams for personal	**Character bigram**	**Character trigram**	**Term unigram**	**Term bigram**
1	e r - 5777	# # a - 5435	# - 11818	# (space) - 794

2	i n - 5579	# # t - 4230	and - 1788	# i - 496
3	# a - 5435	# # m - 3639	of - 1046	social media - 177
4	# t - 4231	# # i - 3333	the - 997	i am - 160
5	a n - 4072	# # s - 3160	i - 871	i'm - 144

Table 3.1: Top frequent n-grams and occurrence for personal profiles

Top frequent n-grams for personal	**Character bigram**	**Character trigram**	**Term unigram**	**Term bigram**
1	# t - 2477	# # t - 2477	# - 4032	# - 234
2	# a - 2109	# # a - 2109	the - 852	# the - 148
3	i n - 2074	# # o - 1325	and - 754	# com - 99
4	e r - 1611	# # s - 1295	of - 339	# we - 95
5	t h - 1558	# # f - 1255	to - 336	the official - 94

Table 3.2: Top frequent n-grams and occurrence for organizational profiles

After cross checking the n-grams in both sides, we decided to discard the character n-gram features due to the high homogeneity (which can be justified in more detail in n-gram files). We applied term (word) n-grams as high-level features.

We employed two factors to select useful n-grams. Occurrence m implied the least number of weighted occurrences in either personal or organization descriptions. Difference n implied the least number of differences in weighted occurrence between in personal descriptions and organizational descriptions. The absence of an n-gram was regarded as 0 occurrences in difference calculation. We set couples of these two factors and experimented on the selected n-gram features.

Before the occurrence calculation, we had another step to balance the data proportion. In our data, the quantity of personal profiles is almost twice of the one of organizational profiles. So we normalized the occurrences by an easy equation:

$$NO = TO * TP / CO,$$

in which NO is the normalized occurrence number, TO is the true occurrence number, TP is the number of total profiles (here is 6499) and CO is the number of class occurrence in total profiles (3214 for personal and 1577 for organizational).

After pruning the n-gram features, we had relatively small feature sets as we can see in table below.

	# Unigrams	**# Bigrams**
m=30, n=30	22	10
m=30, n=50	29	13
m=50, n=30	23	11
m=50, n=50	34	17
m=50, n=100	57	26
m=100, n=50	35	25
m=100, n=100	83	49

Table 3.3: number of n-gram features in experiment with different factors

4. Experiments:

We made use of WEKA CLI to convert our csvs file into arff files. Then we used the SMO (with normalized poly kernel) and MLP algorithms in WEKA explorer. We applied 10 fold cross validation for the experiments. Here are some results of experiments with different input value of factors.

Algorithm(m, n)	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Accuracy**
SMO(30, 30)	0.8	0.78	0.784	0.726	79.8006%
MLP(30, 30)	0.765	0.772	0.765	0.814	77.1695%
SMO(30, 50)	0.804	0.803	0.789	0.731	80.2676%
MLP(30, 50)	0.757	0.761	0.758	0.799	76.1497%
SMO(50, 30)	0.802	0.801	0.786	0.727	80.0585%
MLP(50, 30)	0.757	0.764	0.758	0.804	76.4423%
SMO(50, 50)	0.806	0.805	0.793	0.738	80.552%
MLP(50, 50)	0.764	0.77	0.764	0.808	77.021%
SMO(50, 100)	0.815	0.816	0.806	0.753	81.5635%
MLP(50, 100)	0.787	0.792	0.789	0.835	79.1806%
SMO(100, 50)	0.809	0.809	0.797	0.741	80.8737%
MLP(100, 50)	0.758	0.761	0.759	0.813	76.1497%
SMO(100, 100)	**0.828**	**0.828**	**0.82**	0.771	**82.8386%**
MLP(100, 100)	0.789	0.794	0.79	**0.848**	79.4105%

Table 4.1: the weighed performance in different metrics (the best values of each metric are written in bold face)

5. Conclusions and future work:

We could conclude from the results that SMO algorithm usually gave us better performance in every metrics but the ROC Area. Generally speaking, more features generated better results.

In the experiments, we actually achieved very promising numbers of metrics for classification. Even for the organization recall rate, which was known for the most difficulty in our experiments, we achieved 0.629 in experiment MLP (50, 100) (which can be referred in the result summary text file).

For the future work, we plan to test on more data, which is uneasy due to the enormous work of manually labeling. We are also looking for some other high-level linguistic features to improve the performance. The analysis of tweets context also leads to an interesting direction which may involve more methods of user analysis. At the end, we want to improve the algorithms by implementing ourselves or try on some other algorithm to pursue better results.

6. Reference:

[1] Argamon S, Dhawle S, Koppel M, et al. Lexical predictors of personality type[J]. 2005.

[2] Bachrach Y, Kosinski M, Graepel T, et al. Personality and patterns of Facebook usage[C]//Proceedings of the 3rd Annual ACM Web Science Conference. ACM, 2012: 24-32.

[3] Huffaker D. Gender similarities and differences in online identity and language use among teenage bloggers[D]. Georgetown University, 2004.

[4] Burger J D, Henderson J, Kim G, et al. Discriminating gender on Twitter[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 1301-1309.

[5] Kwak H, Lee C, Park H, et al. What is Twitter, a social network or a news media?[C]//Proceedings of the 19th international conference on World wide web. ACM, 2010: 591-600.