4-2015

Chalk and cheese in Twitter: Discriminating personal and organization accounts

Richard Jayadi OENTARYO
Singapore Management University, roentaryo@smu.edu.sg

Jia-Wei LOW

Ee Peng LIM
Singapore Management University, eplim@smu.edu.sg

DOI: https://doi.org/10.1007/978-3-319-16354-3_51

Chalk and Cheese in Twitter: Discriminating Personal and Organization Accounts

Richard J. Oentaryo, Jia-Wei Low, and Ee-Peng Lim

Living Analytics Research Centre, Singapore Management University
{roentaryo, jwlow, eplim}@smu.edu.sg

Abstract. Social media have been popular not only for individuals to share contents, but also for organizations to engage users and spread information. Given the trait differences between personal and organization accounts, the ability to distinguish between the two account types is important for developing better search/recommendation engines, marketing strategies, and information dissemination platforms. However, such task is non-trivial and has not been well studied thus far. In this paper, we present a new generic framework for classifying personal and organization accounts, based upon which comprehensive and systematic investigation on a rich variety of content, social, and temporal features can be carried out. In addition to generic feature transformation pipelines, the framework features a gradient boosting classifier that is accurate/robust and facilitates good data understanding such as the importance of different features. We demonstrate the efficacy of our approach through extensive experiments on Twitter data from Singapore, by which we discover several discriminative content, social, and temporal features.

Keywords: account type classification, gradient boosting, social media

1 Introduction

Social media provide a platform not only for social interaction among users, but also for businesses, government agencies, and other interest groups to engage users with news and campaign events. As such, social media see the strong presence of both ordinary users and organizations. Unfortunately, these two kinds of social media accounts are not clearly differentiated, as the account types are not specified when accounts are created. In some cases, one could manually judge the account type by examining the account name, description, and content postings. However, such kind of intelligent judgment is a non-trivial task for machines.

We define *organization account* as a social media account that represents an institution, corporation, agencies, news media, or common interest group, whereas *personal account* is of non-organizational nature and usually managed by an individual. An accurate labeling of these account types will bring about many benefits. Firstly, organization and oersonal accounts exist for different purposes and thus demand for different types of support and services. For example, organization accounts may require templates to standardize the format of their

content postings, and dashboard to track their social media performance, say the amount of positive and negative sentiments on their product brands. Personal accounts, in contrast, would likely benefit from friend and content recommendation. Such differentiation of services is presently not possible until the account type can be made known or accurately predicted, which is the focus of this paper.

From the information retrieval perspective, the ability to distinguish personal and organization accounts is useful for enriching and providing context to search or recommendation engines. For example, when searching for a certain trending topic, one may be interested to separate/categorize between official information coming from a credible institution or news source and subjective opinions/views from individuals. From the social science standpoint, much work on social media, such as friend recommendation, community discovery, topic modeling, etc., has been done often assuming that social media accounts are owned by ordinary users. The presence of organization accounts clearly introduces biases to the results analysis and should be treated differently from personal accounts.

In this work, we attempt to address the account type classification problem, which involves assigning social media accounts into the *personal* and *organization* categories. This problem has not been well studied in the past. Nevertheless, recently there is a surge of interest in developing methods for differentiating the two account types, such as the works in [7, 12, 14, 15]. However, most of these works focused on a limited set of social, temporal, or content features [7, 12], or relied on assumptions that may impose significant biases in their evaluation (e.g., using only geotagged tweets [15] or small data samples [12, 14]).

Contributions. Deviating from the previous works, we approach the account classification task by developing a generic framework that facilitates systematic studies on a rich set of content, temporal, and social features, and that offers accurate/robust prediction method. Specifically, our key contributions include:

- We develop a generic framework for account type classification that can cater for various features using generic set of feature transformation pipelines. At its core is the *gradient boosting* classification method [8], which provides not only accurate and robust prediction but also facilitates data understanding.
- We present a new empirical study on Twitter data involving a large (unbalanced) pool of personal and organization accounts. We conduct exploratory analyses on a variety of content, social, and temporal factors associated with personal and organization accounts, based on which we systematically devise a comprehensive set of predictive features for account type classifcation.
- Extensive experiments have also been carried out to evaluate the impacts of different features, and to compare the performance of our gradient boosting approach with other classification methods. We also identify several key features important for the distinction of personal and organization accounts.

2 Related Work

The abundance of user-generated data in social media has recently attracted great interest in inferring the latent attributes of users (e.g., gender [3], political

stand [6], ethnicity [5]). Most of these works, however, have treated organization and personal accounts equally. Yet, the ability to distinguish the two is practically important for marketing and information dissemination. Nonetheless, several efforts have been recently made to this end. Tavares and Faisal [12] distinguished between personal, managed, and bot accounts in Twitter, using only the temporal features of the tweets. De Choudhury *et al.* [7] classified Twitter accounts as organization, journalist/blogger, or individual. They utilized structural features, textual features, and binary features indicating the presence of named entitites and associations with news topics. Yan *et al.* [14] called the personal and organization accounts closed and open accounts respectively, and used the diversity of the follower distribution as features. Recently, Yin *et al.* [15] devised a probabilistic method that utilizes temporal, spatial and textual features to classify personal communication and public dissemination accounts.

Proposed approach. Our work differs from the abovementioned approaches in several ways. For instance, Tavares and Faisal [12] focused only on temporal features without considering other feature types. Meanwhile, Yin *et al.* [15] used only geotagged tweets, which may yield significant bias against non-mobile (e.g., desktop) users who do not share their location. In contrast, we use a comprehensive set of content, social, and temporal features, and we consider all tweets with or without geotag information. In [7], De Choudhury *et al.* utilized only simple social features based on in-degree and out-degree centrality metrics. By comparison, our work involves more sophisticated social features that go beyond simple degree centrality. Moreover, we utilize temporal features (e.g., tweet distribution per hour or weekday) in our classification model. Compared to [14], our approach takes into account a more comprehensive set of temporal and social features (encompassing many node centrality and diversity measures). We further elaborate our classification method and feature set in Sections 4 and 5.

3 Data Exploration

In this study, we use the Twitter data of users from Singapore collected from March to May 2014. Starting from a set of popular seed users (having many followers) based in Singapore, we crawled their network based on the follow, retweet, and user mention links. In turn, we added into our user base those followers/followees, retweet sources, and mentioned users who declare Singapore in their profile location. This led to a total of 160,143 public Twitter accounts whose profiles can be accessed. To establish the ground truth, we took accounts whose "urlEntities" field ends with ".com.sg", ".gov.sg", or ".edu.sg". This choice allows us to clearly identify organization accounts for deriving high-quality ground truths, though it may impose labeling bias and miss other, less common types of organization. Nevertheless, we show later that our prediction method can work well for organization accounts that are not from these domains (cf. Section 5.4).

Using this procedure, we were able to identify 885 organization accounts. Through random sampling of the Twitter data, we also obtained 1,135 personal accounts. All labels have been manually inspected by humans. In total, we have

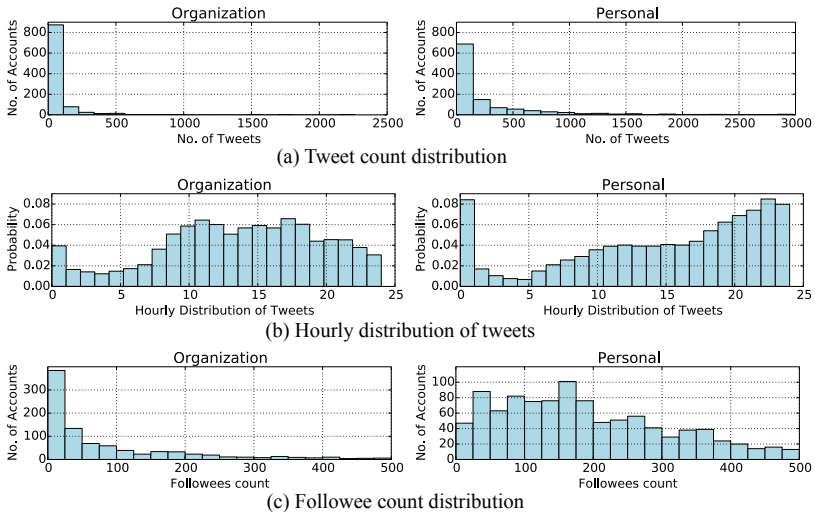

Fig. 1. Data distributions for personal and organization accounts

2,020 labeled accounts, involving 1.18 million tweets. One may argue that it is better to balance the label distribution, e.g., by using the same number (885) of personal accounts. However, we can expect that the full Twitter population would naturally have more personal accounts than organization accounts [7, 14]. Hence, we maintain the current label distribution (i.e., 885 vs. 1,135), and let our classification algorithm internally take care of the skewed distribution.

Content analysis. We first conducted analysis on the number of tweets for personal and organization accounts. Fig. 1(a) shows the distribution of tweet counts for the two accounts. From the figure, we can see that the tweet counts generally follow a long-tail distribution. It is also shown that personal accounts tend to tweet more than organization accounts. We then conducted *Kolmogorov-Smirnov* (K-S) test [11] to check whether the two distributions are significantly different[1]. In this case, we obtained a p-value of 2.8779×10^{-36}, which is smaller than typical significance level (e.g., 0.01 or 0.05). Hence, we can reject the null hypothesis that the two distributions are identical. That is, the distributions of personal and organization accounts are significantly different.

Temporal analysis. Next, we conducted a temporal data analysis to check whether the tweet dynamics of the personal and organization accounts are different. Fig. 1(b) shows the hourly distribution of the tweet counts. As the purpose of setting organization accounts is chiefly about information dissemination, we can see that their tweet activities tend to be more aligned with business operation/working hours. On the other hand, we observe that personal accounts tend to tweet more towards the end of the day, peaking around midnight. Using the

[1] We use two-sample K-S test, which is a nonparametric statistical test to quantify the distance between the empirical distribution functions of two samples.

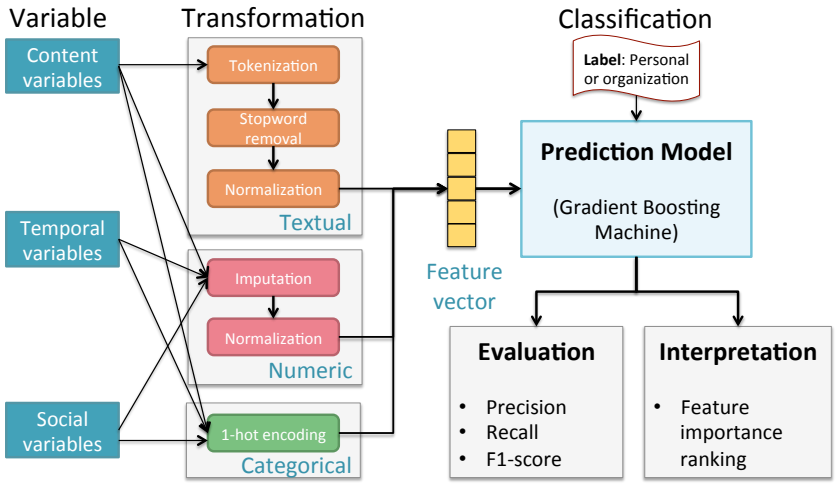

Fig. 2. Proposed framework for account type classification

K-S test, we again obtained p-value $< 10^{-100}$ and concluded a significant difference between the two. This suggests that the temporal distribution of tweets could be a useful feature for distinguishing the two account types.

Social analysis. We also analyzed the interaction patterns among accounts. Fig. 1(c) shows the distributions of the followee counts for personal and organization accounts. We can see that in general personal accounts have more followees than organization accounts. Again, this may be attributed to the fact that organization accounts are set up mainly for dissemination purposes, and so unlikely to be interested in other accounts. The significant difference between the two distributions is evident in our K-S test, with p-value of 8.07×10^{-61}.

4 Proposed Framework

Our proposed account classification framework is outlined in Fig. 2. It takes three types of (raw) input variables: *content*, *temporal*, and *social*. Each variable type goes through a specific transformation pipeline (cf. "Transformation" in Fig. 2) to derive feature vector representation suitable for our classification model. The choice of pipeline for a given variable depends on the semantics of the variable. Using the feature vector and the class label, we build the model (cf. "Predictive Model" in Fig. 2). We then evaluate the model performance based on several metrics (cf. "Evaluation" in Fig. 2). The framework also has a specialized module to extract knowledge structure from the model (cf. "Interpretation" in Fig. 2).

4.1 Feature Transformation Module

Our framework has three types of transformation pipeline, which can be generically used to transform any content, temporal, and social variables into feature

vector representation for our classification model. For convenience, we refer the collection of tweets belonging to a user as the user's tweet *document*.

Textual pipeline. For text variables such as tweet documents, we convert them into *bag-of-words* vector representation [10]. This involves several steps:

- *Tokenization*: We break a tweet document into its constituent word tokens. Delimiters, such as punctuation marks and white spaces, were used as word boundaries. At the end of this process, we obtain bags of word frequencies.
- *Stop-word removal*: We then discard words that appear very frequently and contribute little to discriminating the tweets of a user from those of other users. In this work, we use the list of English stop-words in [9].
- *Normalization*: We then applied the *term frequency–inverse document frequency* (TF-IDF) scheme [10] to obtain normalized word frequencies. The scheme puts greater importance on words that appear frequently in a document, and deems words that occur in many documents as less important. Our TF-IDF vectors span unigram, bigram, and trigram representations. More advanced methods such as BM25 and part-of-speech tagging [10] can be included, but for simplicity we use only the TF-IDF method in this work.

Numerical pipeline. The transformation steps of numerical variables (such as count or ratio variables; cf. Table 1) include:

- *Imputation*. We first impute the missing feature values by replacing them with some constant value, or else the average of the other, existing feature values. In this work, we impute missing values with a constant value of zero.
- *Normalization*: This step performs feature normalization by (re)scaling each feature to a unit range [0, 1]. This normalization serves to address the feature scaling issues in classification methods that rely on some distance metric.

Categorical pipeline: In our framework, all categorical variables are binary-encoded. For example, a categorical variable with four possible values: "A", "B", "C", and "D" is encoded using four binary features: "1 0 0 0", "0 1 0 0", "0 0 1 0", and "0 0 0 1", respectively. This is also known as *one-hot encoding* scheme.

4.2 Prediction Model

For our classification task, we employ an ensemble model called *gradient boosting machine* (GBM) [8]. The learning procedure in GBM involves consecutively fitting new models to provide more accurate estimate of the response variable (i.e., class label). The centerpiece of GBM is to construct new base-learners so that they are maximally correlated with the negative gradient of the specified loss function, associated with the entire ensemble [8].

It is worth noting that the loss function used in GBM can be arbitrary, thus providing practitioners with the flexibility to select the most appropriate loss function to the task requirements. GBM is also relatively easy to implement, allowing practitioners to experiment with different model designs. In this work, we focus on using the *binomial loss* function in GBM, which is suitable for our (binary) classification task [8]. As the base learners in GBM, we choose decision tree [2] for both computational efficiency and interpretability reasons.

4.3 Evaluation Module

To evaluate our approach, we use a *stratified* 10-fold cross-validation (CV) procedure, whereby we split the Twitter data into 10 folds of training and testing data, each retaining the class label proportion as per the original data. We then report the average performance as well as its variation (i.e., standard deviation). The stratification is needed to ensure that each fold is a good representative of the whole, i.e., retains the (unbalanced) label distribution in the original data.

In this work, we consider several evaluation metrics popularly used in information retrieval, namely *Precision*, *Recall*, and *F1-score* [10]:

$$Precision = \frac{TP}{TP + FP}; Recall = \frac{TP}{TP + FN}; F1 = \frac{2PrecisionRecall}{Precision + Recall} \quad (1)$$

where TP, FP and FN are the true positives, false positives, and false negatives respectively. Here we treat the organization account as the positive class.

4.4 Interpretation Module

The ability to describe and interpret the derived predictive model is important for many applications. A useful interpretation of our GBM classifier involves understanding those particular features that are most influential in contributing to the classification performance as well as its variance. To this end, we utilize the feature importance metric derived based on the decision tree influences [2]. Specifically, the feature importance corresponds to the expected fraction of the samples that each decision tree contributes within the ensemble models [8].

5 Experiment

This section presents our empirical studies on the Twitter data we have collected. All evaluations were based on the stratified 10-fold CV method (cf. Section 4.3).

5.1 Features Extracted

Based on findings in Section 3, we devised numerous content, social and temporal features for our account classification task. Table 1 lists all features used in our work, as well as their corresponding types and feature transformation pipelines. For convenience, we shall use the term "user" and "account" interchangably. We do not use categorical features in this work for now, although the implementation of the categorical pipeline is readily available in our framework.

For the textual contents of tweet documents, we use the TF-IDF representation for tweet documents, as described in Section 4.1. We also construct a number of numerical features from the content and social variables. These include count and ratio features, such as the total counts of entities (e.g., "MentionCount"), the counts of unique entities (e.g., "MentionUnique"), and the ratio of unique over total counts (e.g., "MentionUniqueRatio").

Table 1. List of features used

Feature	Type	Pipeline	Description
TweetContent	C	X	TF-IDF vector of a user's tweet document
TweetCount	C	N	No. of tweets of a user (March–May 2014)
SourceUnique	C	N	No. of unique applications a user tweets from
SourceUniqueRatio	C	N	No. of unique applications / total no. of tweets
HashtagUnique	C	N	No. of unique hashtags
HashtagCount	C	N	Total no. of hashtags
HashtagUniqueRatio	C	N	No. of unique hashtags / total no. of hashtags
HashtagCountRatio	C	N	No. of hashtags / total no. of tweets
ListedCount	S	N	No. of Twitter lists at which a user appears
FavouritesCount	S	N	No. of tweets a user has marked as favourite
MentionUnique	S	N	No. of unique (user) mentions
MentionCount	S	N	Total no. of (user) mentions
MentionUniqueRatio	S	N	No. of unique mentions / total no. of mentions
MentionCountRatio	S	N	No. of mentions / total no. of tweets
MentionClusterCoeff	S	N	Clustering coefficient for mention graph
MentionMentionedRatio	S	N	No. of mentions / no. of mentioneds
FollowersCount	S	N	No. of followers of a user
FolloweesCount	S	N	No. of followees of a user
FolloweeClusterCoeff	S	N	Clustering coefficient for followee graph
FollowerFolloweeRatio	S	N	No. of followers / no. of followees
FolloweeFollowerMean	S	N	Mean of the no. of followers of a user's followees
FolloweeFollowerMedian	S	N	Median of the no. of followers of a user's followees
FolloweeFollowerStdDev	S	N	Deviation of the no. of followers of a user's followees
FolloweeFollowerEntropy	S	N	Entropy of the no. of followers of a user's followees
FolloweeFolloweeMean	S	N	Mean of the no. of followees of a user's followees
FolloweeFolloweeMedian	S	N	Median of the no. of followees of a user's followees
FolloweeFolloweeStdDev	S	N	Deviation of the no. of followees of a user's followees
FolloweeFolloweeEntropy	S	N	Entropy of the no. of followees of a user's followees
FolloweeTraceMean	S	N	Mean of the trace of no. of followees over time
FolloweeTraceMedian	S	N	Median of the trace of no. of followees over time
FolloweeTraceStdDev	S	N	Deviation of the trace of no. of followees over time
FolloweeTraceEntropy	S	N	Entropy of the trace of no. of followees over time
FollowerTraceMean	S	N	Mean of the trace of no. of followers over time
FollowerTraceMedian	S	N	Median of the trace of no. of followers over time
FollowerTraceStdDev	S	N	Deviation of the trace of no. of followers over time
FollowerTraceEntropy	S	N	Entropy of the trace of no. of followers over time
AccountAge	T	N	Total duration from since account created till now
AverageTweetCount	T	N	No. of tweets / account age
ProbWeekend	T	N	Probability of a user tweeting on the weekend
ProbMorning	T	N	Probability of a user tweeting in the morning
ProbAfternoon	T	N	Probability of a user tweeting in the afternoon
ProbEvening	T	N	Probability of a user tweeting in the evening
ProbNight	T	N	Probability of a user tweeting at night
Hour-x	T	N	Probability of a user tweeting at hour x
Weekday-x	T	N	Probability of a user tweeting at day x

Type – C: content, S: social, T: temporal; Pipeline – N: numeric, X: textual

For social features, we also consider two-hop centrality features such as clustering coefficient (CC) and some first- and second-order statistics of the followees' followees (or followees' followers) of a user. The purpose of including two-hop features is to allow us to account for a sufficiently large community of users. The CC metric measures the extent to which a user's neighborhood form a clique. For a user i, CC is the number of edges between the user's neighbors N_i divided by the total number of possible edges between them, i.e., $|N_i| \times (|N_i| - 1)$.

As for the statistics of the followees/followers, we use first-order statistics such as mean and median, as well as second-order statistics such as standard deviation and entropy. The second-order metrics are used to quantify the *diversity* of the entities associated with a user's neighborhood. To obtain the entropy, we

Table 2. Impacts of different features for account type classification (10-fold CV)

(a) Classification results

Features	Precision	Recall	F1-Score
C	0.841 ± 0.030	0.782 ± 0.046	0.809 ± 0.025
S	0.865 ± 0.031	0.858 ± 0.040	0.860 ± 0.025
T	0.758 ± 0.029	0.777 ± 0.054	0.766 ± 0.028
C, S	0.879 ± 0.036	0.886 ± 0.034	0.882 ± 0.024
C, T	0.854 ± 0.019	0.861 ± 0.058	0.856 ± 0.033
S, T	0.890 ± 0.032	0.889 ± 0.047	0.889 ± 0.024
C, S, T	$\mathbf{0.909 \pm 0.023}$	$\mathbf{0.904 \pm 0.041}$	$\mathbf{0.906 \pm 0.016}$

C: content, S: social, T: temporal

(b) Statistical significance (p-value)

	C, S	C, T	S, T	C, S, T
C	0.005^{**}	0.005^{**}	0.005^{**}	0.005^{**}
S	0.007^{**}	0.333	0.013^{*}	0.005^{**}
T	0.005^{**}	0.005^{**}	0.005^{**}	0.005^{**}
C, S	-	0.021^{*}	0.241	0.009^{**}
C, T	-	-	0.009^{**}	0.007^{**}
S, T	-	-	-	0.008^{**}

C: content, S: social, T: temporal

* / **: significant at 95% / 99%

Table 3. Benchmarking results of different algorithms (10-fold CV)

Algorithm	Precision	Recall	F1-Score	p-value
Support vector machine	0.859 ± 0.045	0.843 ± 0.033	0.850 ± 0.029	0.005^{**}
Logistic regression	0.863 ± 0.037	0.842 ± 0.039	0.852 ± 0.030	0.005^{**}
Decision tree	0.808 ± 0.023	0.827 ± 0.043	0.817 ± 0.023	0.005^{**}
Random forest	0.878 ± 0.028	0.899 ± 0.032	0.888 ± 0.029	0.008^{**}
Gradient boosting	$\mathbf{0.909 \pm 0.023}$	$\mathbf{0.904 \pm 0.041}$	$\mathbf{0.906 \pm 0.016}$	-

**: significant at 99%

first take the normalized count (i.e., probability density) $p_{i,j}$ for each neighbor $j \in N_i$ of user i, and then compute the entropy $-\sum_{j=1}^{|N_i|} p_{i,j} \log p_{i,j}$.

We also devise more advanced social features dubbed *trace*, describing the dynamics of social entities over time. For instance, the "FolloweeTraceMean" feature in Table 1 means the average of the trace vector of followee counts over time. Here each element in the trace vector is the followee count observed for time period t . In this work, we set the observation period as $t = 3$ days.

Finally, we devise a number of temporal features based on the periodicity of the tweet counts observed at different time spans. In particular, we bin the tweets by time and compute the probability of tweeting in the morning (4:00-11:59am), afternoon (12:00pm-4:59pm), evening (5:00-7:59pm), and night (8:00pm-3:59am). We also compute the probability of the tweets occurring in the weekend. To capture daily and hourly distribution of tweets (cf. Fig. 1(b)), we also compute the probability of tweeting at Weekday-x (where $x \in \{0, 1, \ldots, 6\}$ for Monday to Sunday), and Hour-x (where $x \in \{0, 1, \ldots, 23\}$ for 24 hours).

5.2 Performance Assessment

We first evaluated the impact of different features to the overall classification performance of GBM, and then compared the GBM results using all features to the results of several other popular classification algorithms. Table 2 illustrates the impact of different features. Looking at the results of individual content (C), social (S), and temporal (T) features, we can see that the social features alone gave the highest F1-score, followed by the content features and temporal features. The performance of combination of content and social features is higher than either of the individual baseline. The same conclusion applies for the combination of social and temporal features. Lastly, the GBM model that uses all content, social, and temporal features was able to achieve the highest F1 score.

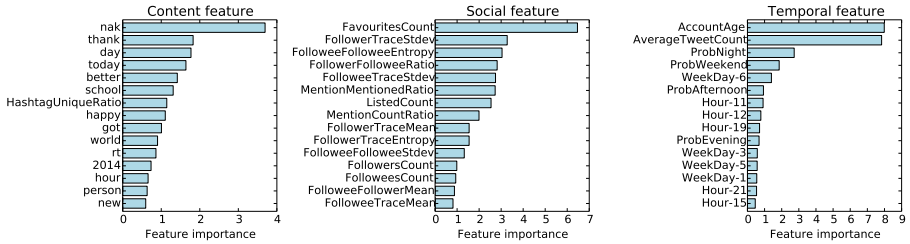

Fig. 3. Top 15 features for Twitter account type classification

Table 4. Prediction results on unseen data

(a) Confusion matrix

		Predicted									
		Top 20		Top 40		Top 60		Top 80		Top 100	
		Per	*Org*	*Per*	*Org*	*Per*	*Org*	*Per*	*Org*	*Per*	*Org*
Actual	*Per*	20	0	40	0	60	0	79	1	99	1
	Org	1	19	3	37	5	55	7	73	12	88

Per: personal, *Org*: organization

(b) Organization accounts

Domain	**No. of accounts**
.com	66
.sg	14
.org	5
None	3

To evaluate the contributions of different feature combinations, we conducted the *Wilcoxon signed-rank* statistical test [13][2]. From the p-values in Table 2(b), we can see that the overall pairwise differences of $F1$ are statistically significant, except for two cases. Nevertheless, it is clear that combining all feature types (content, social, and temporal) gave substantially better results than using the constituent features (cf. Table 2(b), last column), which is our primary interest.

We further benchmarked the results of our approach against those of other classification algorithms. These include *support vector machine* and *logistic regression* [4], which are linear models widely used in information retrieval. We also used decision tree baseline [2], as well as *random forest* [1]—a popular bootstrap aggregating method to create an ensemble of decision trees. For fairness, we used all three feature types in this benchmark. As evident from Table 3, our GBM method consistently outperforms the other algorithms across all evaluation metrics. We also found that the improvements are statistically significant according to the Wilcoxon signed-rank test, as per the last column of Table 3. This in turn justifies the accuracy and robustness traits of our approach.

5.3 Feature Importance

Using the trained GBM model, we can now evaluate the importance of different features, as described in Section 4.4. Fig. 3 shows the top 15 most important features produced by GBM for each feature type. Several interesting insights are observed. For example, the top textual feature "nak" is the short form of

[2] The Wilcoxon test provides a non-parametric alternative to the t-test for matched pairs, when the pairs cannot be assumed to be normally distributed

"want" in Malay language, which is often used for informal communication. From Fig. 3 and our manual inspections, we also found that special word such as "rt" (which stands for retweet) is indicative of the account type (in this case, personal accounts tend to retweet more). Among the non-textual features, "HashtagUniqueRatio" is ranked among the top. A closer look at the data shows that organization accounts often have more unique hashtags than personal accounts, suggesting that the former have a more focused topic of interest.

As for social features, it is shown that "FavouritesCount" emerges as the top feature. Indeed, our internal inspections reveals that personal accounts tend to have larger favourite counts. Despite this observation, using "FavouritesCount" alone is not sufficient to obtain good classification results, and the collective contribution of the other social features remains important. We also found the diversities of the no. of followees/followers over time (e.g., "FollowerTraceStdDev", "FolloweeTraceStdDev", "FollowerFolloweeEntropy") to be discriminative of the account types. From our inspections, we found that the deviations of followee trace for organization accounts are moderate in general. This is likely due to the fact that most organizations utilize Twitter as a dissemination platform.

With regard to temporal features, we discovered that organization accounts have gained traction on Twitter only in the recent 2-3 years, whereas many personal accounts were created 4-5 years ago. This explains why the account age is one of the top features. We also noticed that personal accounts have higher "AverageTweetCount" than organization accounts. In addition, we conclude from "ProbWeekend" and "Weekday-6" that personal accounts tend to tweet more than organization accounts during the weekend. The results also suggest that the probability of tweeting in the afternoon ("ProbAfternoon") or evening ("ProbEvening") is discriminative. Lastly, there are several critical hours (e.g., "Hour-11", "Hour-12", "Hour-19"–possibly related to lunch/dinner time) as well as critical days (e.g., "Weekday-1" (Tuesday), "Weekday-3" (Thursday), "Weekday-5" (Saturday)) that are useful for the account type classification.

5.4 Out-of-Sample Generalization

To assess the ability of our model to generalize, we used our trained GBM model to predict for all unlabeled data. We then picked the top K organization accounts and top K personal accounts based on the prediction scores. We varied K from 20 to 100 and examined the prediction results for all the top accounts, so as to see how the GBM predictions match with our manually-examined labels. Table 4(a) summarizes the results. It is shown that, under varied K, our approach produced good performance on unseen data, achieving robust accuracies of $98.75 - 100\%$ for personal accounts and $88 - 95\%$ for organization accounts.

Table 4(b) shows the domain type breakdown of the 88 correct predictions for the top 100 organization accounts. We can see that our approach can correctly predict for organization accounts with domain types other than those of the labeled (training) data. Note here that the domain extensions ".com" and ".sg" in the unlabeled data are different from the ".com.sg" extension in the labeled data. In sum, these results justify the generalization ability of our approach.

6 Conclusion

We put forward a generic framework for discriminating personal and organizational accounts in social media. Our framework provides a generic set of feature transformation pipelines that supports integration of rich content, social, and temporal features. With gradient boosting as its core, our approach achieves accurate/robust performance and provides useful insights on the data. We have empirically demonstrated the effectiveness and interpretability of our approach using Singapore Twitter data. Moving forward, we wish to apply our method to Twitter data from a larger region. We also plan to build a multi-attribute prediction method that can integrate information from heterogeneous social networks.

Acknowledgments. This research is supported by the Singapore National Research Foundation under its International Research Centre@Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority (MDA).

References

1. L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
2. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. 1984.
3. J. D. Burger, J. Henderson, G. Kim, and G. Zarrella. Discriminating gender on Twitter. In *EMNLP*, pages 1301–1309, 2011.
4. C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM-TIST*, 2:27:1–27, 2011.
5. J. Chang, I. Rosenn, L. Backstrom, and C. Marlow. ePluribus: Ethnicity on social networks. In *ICWSM*, pages 18–25, 2010.
6. R. Cohen and D. Ruths. Classifying political orientation on Twiter: It's not easy! In *ICWSM*, pages 91–99, 2013.
7. M. De Choudhury, N. Diakopoulos, and M. Naaman. Unfolding the event landscape on Twitter: Classification and exploration of user categories. In *CSCW*, 2012.
8. J. H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.
9. D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *JMLR*, 5:361–397, 2004.
10. C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
11. N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, 19(2):279–281, 1948.
12. G. Tavares and A. A. Faisal. Scaling-laws of human broadcast communication enable distinction between human, corporate and robot Twitter users. *PloS One*, 8(7):1–11, 2013.
13. F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–88, 1945.
14. L. Yan, Q. Ma, and M. Yoshikawa. Classifying Twitter users based on user profile and followers distribution. In *DEXA*, pages 396–403, 2013.
15. P. Yin, N. Ram, W.-C. Lee, C. Tucker, S. Khandelwal, and M. Salathé. Two sides of a coin: Separating personal communication and public dissemination accounts in Twitter. In *PAKDD*, pages 163–175, 2014.