Modeling of Social Media Behaviors Using Only Account Metadata

Fernanda Carapinha[1], John Khoury[2], Shai Neumann[2], Monte Hancock[1(✉)], Federico Calderon[1], Mendi Drayton[1], Arvil Easter[1], Edward Stapleton[1], Alexander Vazquez[1], and David Woolfolk[1]

[1] 4Digital, Los Angeles, USA
practicaldatamining@gmail.com
[2] Eastern Florida State College, Cocoa, USA

Abstract. Applications in Augmented Cognition can be hampered by obstacles to the effective instrumentation of the data space, making the collection of informative feature data difficult. These obstacles usually arise from technical limitations, but can also be present due to methodological and legal considerations. We address a specific instance of the difficulty of characterizing a complex behavior space under legally constrained data collection: the instrumentation of social media platforms, where privacy, policy, and marketing considerations can severely hamper 3rd-party data collection activities. This paper documents our constrained empirical analysis and characterization of the behaviors of Twitter account-holders from their account metadata alone. The characterization is performed by coding user account data as feature vectors in a low-dimensional Euclidean space, then applying parametric and non-parametric methods to the resulting empirical distribution. Suggestions for future work are offered.

Keywords: Twitter · Social media · Behavior modeling · Metadata

1 Background

Social Media is entrenched in the human psyche across the globe. Individuals feel its impact in relationships, knowledge, while businesses cite its effectiveness in brand management, public relations, and product promotion. Even government sees the value in utilizing social media to reach its citizens and promote policy as well as provide alerts and notifications in case of emergencies. As a result, social media has become a lucrative marketing and sales channel for businesses. It offers sellers immediate access to a demographically diverse, socially active, and relatively affluent international market. They interconnect users in social cliques, where buying and product experiences are shared, occasionally resulting in geometric growth of product awareness ("going viral") [1].

As technology progresses, how can users of this medium ensure their ROI? Twitter has been transformed from a personal microblogging site to a robust information portal. It has become a defacto media channel, and as such is concerned that its audience is engaged and authentic. These concerns are, however, not as easily determined as in

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC 2016, Part II, LNAI 9744, pp. 393–401, 2016.
DOI: 10.1007/978-3-319-39952-2_38

years past. Analyzing social media accounts' metadata can unlock clues to an its user authenticity. Does the account belong to active, passive, or anomalous users? These statuses are important to entities wishing to fully leverage the capabilities promised by Twitter's technology. After all, the Twittersphere is populated by countless users and the accuracy of their information, and identification provide real incentives for personal, economic and even political ends. Conversely, inaccurate identification and targeting can place the user at risk or at the very least exposed to unintended consequences.

The design of a reliable method of identification will ensure conditions are optimal to offer interested parties greater opportunity for the monetization of social media data, having overcome the present technical decision-theoretic challenges:

1. There are billions of individual social media accounts
2. Privacy controls limit access to the most informative data elements
3. Technology has reached a level where automated account-holders can impersonate human users, and these constitute a growing proportion of account-holders

2 Data

A large number of fields are available for collection via the Twitter API (Application Program Interface) and they are divided into five object classes. Those object classes are; Users, Tweets, Entities, Entities in Objects, and Places [2]. Due to the focus of this research on user metadata the APIs reviewed were almost exclusively from the "User" object class. Information from the Twitter Developer API guide was used to identify a number of metadata categories for the research team to collect and use in the Twitter-space behavioral data analysis. A bot program was written to collect the data for later inspection and calculations. The sample size for this research was 100,001 users and spanned 13 direct and 3 calculated data fields.

Each Twitter user has a unique 64 bit integer allocated to them for a user identifier, the field in the API is designated "id" and would be analogous to a Primary or Unique Key from a relational database. The id field is essential in separating information about one user from that of another. All the other user object data fields are related back to this unique identifier. In order to better understand these relationships and the possibilities therein, the following definitions provide the field identifiers.

Twitter API Fields Utilized.

1. id – The unique/primary identifier for the user (64 bit integer).
2. created_at – UTC date and time of the initial user creation.
3. default_profile – true/false identifier for whether the user has altered the default theme or background.
4. default_profile_image – true/false identifier for whether the user has replaced the default user avatar.
5. favorites_count – total lifetime number of tweets the user has favorited.
6. followers_count – total number of other users following that unique user.
7. friends_count – total number of users the user is following.

8. location – user provided location, not always accurate (e.g. Mars, etc.).
9. protected – indicates if the user's tweets are protected (only viewable by their followers).
10. statuses_count – total number of tweets sent by the user.
11. time_zone – user defined time zone, not always accurate (e.g. "Eastern Time (US & Canada)").
12. utc_offset – the offset from GMT/UTC described in seconds.
13. verified – user identity has been verified by Twitter (e.g. Joe Actor might be verified as being the real celebrity and not an impersonator).

3 Methodology

The purpose of this paper is to characterize categories of user behavior in communication via tweeting as these are reflected in the metadata only: no tweet content. Data mining technology is used with the aid of statistical methods: exploratory data analysis, cluster analysis, factor analysis, multiple regression, and multinomial logistic regression.

Given 19 variables and 100001 cases, various approaches were used to study the data. The data included both categorical and quantitative variables.

Data cleaning was performed prior to analysis. Evidence of corrupt data was found in the case of one of the features that involved ratio calculations. It appeared that the formula was not executed appropriately to all cases. After corrections were applied, the data set was reexamined and judged to be ready for analysis. There were no missing values in the data set. In addition, issues of division by zero or undefined non-linear transformations of some features were addressed prior to analysis.

In order to visualize the data, a number of graphs were investigated, such as histograms, scatter plots, box plots, and stem and leaf displays.

Descriptive statistics were obtained for each feature, including 5 number summary, outliers were identified.

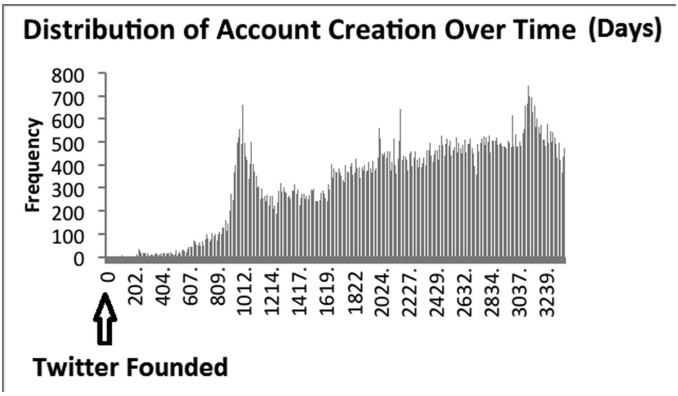
The correlation among variables were calculated and analyzed.

Using SPSS, variable reduction techniques were applied to the original data, showing variable importance, leading to a smaller number of variables to be used in the two-step clustering on SPSS.

In total, two clustering techniques were applied to the data. A two-step clustering run on SPSS revealed primary clusters, including an order of importance of the variables. Consistent with this clustering technique, a multinomial logistic regression was run on SPSS to check accuracy of cluster assignment. Subsequent to the identification of primary clusters, two additional iterations for each cluster were performed using the same two-step clustering method. A second clustering technique created 200 clusters using a dedicated clustering program. Those 200 clusters were examined for distinctive characteristics with particular attention paid to small clusters involving high z scores in a number of features.

4 Analysis

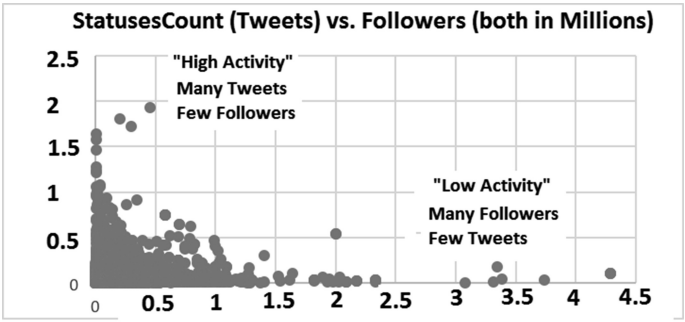
The data set was examined at different levels. One dimensional review included a time series of account creation. It appears to reveal the period when Twitter actually took off and grew exponentially as a platform. That date and time corresponds to the South by South West Interactive conference in March 2007, some 234 days after the first account of the data set was created in July 2006. The distribution suggests that once the product reached a certain level of traction, rates of account creation may have been associated with world events [3].



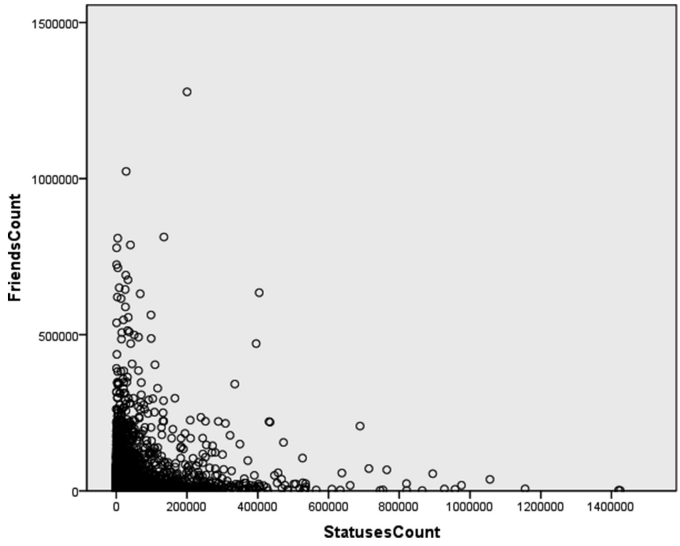
Two dimensional review of the data included creation of the correlation matrix (immediately below) and examination of scatter plots. Features that exhibited high correlation were later examined in the context of variable reduction.

	1-UP IDX	TwitterUserId	Day_Num	DefaultProfile	DefaultProfileImage	FavoritesCount	FollowersCount	FriendsCount	ProtectedUser	StatusesCount	TimeZone	UtcOffset	VerifiedUser
1-UP IDX	1.00												
TwitterUserId	0.99	1.00											
Day_Num	0.97	0.94	1.00										
DefaultProfile	0.15	0.14	0.15	1.00									
DefaultProfileImage	0.04	0.03	0.04	0.25	1.00								
FavoritesCount	-0.01	-0.01	-0.01	-0.03	-0.03	1.00							
FollowersCount	-0.01	-0.01	-0.01	-0.01	-0.01	0.02	1.00						
FriendsCount	0.00	0.00	0.00	-0.05	-0.03	0.07	0.16	1.00					
ProtectedUser	0.03	0.03	0.04	0.00	0.01	-0.02	-0.01	-0.04	1.00				
StatusesCount	-0.03	-0.03	-0.03	-0.08	-0.05	0.18	0.03	0.17	-0.04	1.00			
TimeZone	0.10	0.09	0.09	-0.04	-0.06	-0.02	0.01	0.08	0.00	0.02	1.00		
UtcOffset	0.16	0.15	0.15	0.13	0.05	-0.03	0.00	0.04	-0.01	-0.02	0.84	1.00	
VerifiedUser	-0.04	-0.04	-0.04	-0.03	-0.01	0.03	0.08	0.08	-0.02	0.02	-0.01	-0.03	1.00

The scatter plot for StatusesCount vs followers indicates that in a small number of cases accounts produce a lot of tweets and yet have few followers, while a small number of twitter accounts have many followers but they produce a small number of tweets.



Similarly, the scatter plot for StatusesCount vs friends indicates the same phenomenon.



Another scatter plot that exhibited a noticeable characteristic was the "Followers-Favorites" ratio vs "Friends-Favorites" ratio. Multidimensional analysis was performed using SPSS and a dedicated clustering program. The Two Step Cluster Analysis procedure was used as an exploratory tool in order to reveal grouping (or clusters) within a large data set containing 23 variables and 100001 cases. The clusters were based on both categorical and continuous variables.

The following Five variables were selected as *Categorical:*

Default Profile; Default Profile image; Protected User; Verified User; Duplicates

The following seven variables were selected as *Continuous:*

Favorites Count; Epoch Time; Followers Friends Ratio KGH; Status Count; Friends Favorites Ratio KHF; Location_ Symbol; Friends Count

Factor Analysis and Principal Cluster Analysis, PCA, were used in order to reduce the Dimension of the space from 23 to 12. The number of clusters was selected automatically by the procedure, based on Schwartz's Bayesian Criterion. Then, each of the clusters were sub-clustered by repeating the two step clustering procedure.

The Likelihood Distance measure, which assumes that variables in the cluster are independent, was used. Continuous variables are assumed to have a Gaussian distribution and each categorical variable was assumed to have a multinomial distribution. However, the procedure is fairly robust to departures from both assumptions of independence and normality.

A model summary was created along with a fair cluster quality. The model showed 12 variables and 100001 cases were valid and no missing data. A Pie Chart revealed the percentages of cases in each cluster, along with a table showing the size of the smallest and largest cluster. Clusters profiles, centroids of each cluster, show clusters are well separated.

Multinomial logistic regression was performed with an independent variable being the cluster identified in a two-step SPSS clustering that resulted in three primary clusters. The table below summarizes the results

5 Results

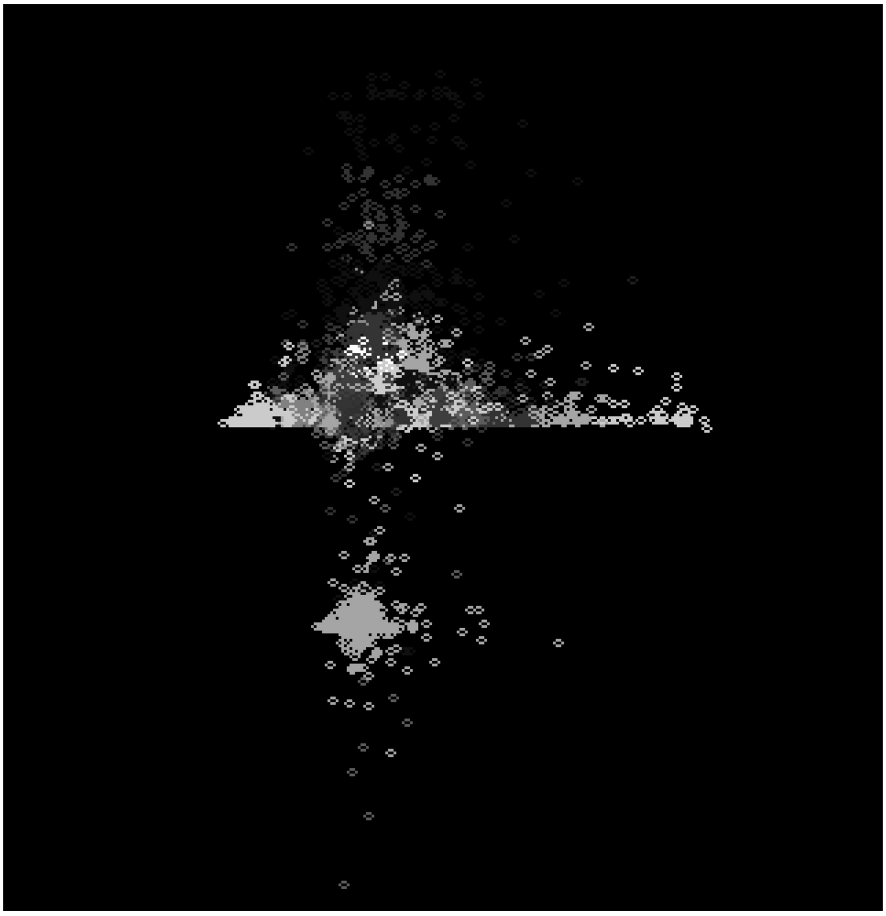
Classification

Actual cluster membership	Predicted			
	Cluster 1	Cluster 2	Cluster 3	Percent correct
Cluster 1	1127	796	1333	34.6 %
Cluster 2	174	42873	47	99.5 %
Cluster 3	206	228	53217	99.2 %
Overall percentage	1.5 %	43.9 %	54.6 %	97.2 %

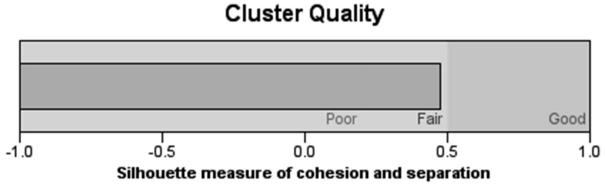
The results indicate that if a data point belongs to clusters 2 or 3, there is better than 99 % probability that the multinomial logistic regression procedure would correctly identify the data point as belonging to the correct cluster. If the data point belonged to cluster 1, which was the smallest cluster, there was only slightly more than one in three chance the multinomial logistic regression would identify the data point correctly. The overall percent accuracy from this perspective is 97.2 %. From another perspective, numbers in this table may be used to derive estimates of conditional probabilities of belonging to a cluster given a multinomial logistic regression prediction of belonging to a specific cluster. The results are as follows: P (belonging to cluster 1 given multinomial logistic regression predicts data point belongs to cluster 1) = 74.8 %, P (belonging to cluster 2 given multinomial logistic regression predicts data point belongs to cluster 2) = 97.7 %, P (belonging to cluster 3 given multinomial logistic regression predicts data point belongs to cluster 3) = 97.5 %. This seems to indicate

that the multinomial logistic regression performs well in this context and can be used to validate cluster assignments. Numbers appear to confirm that clusters might be meaningful in this feature space.

Below is an oblique ortho-projection of the clustered account metadata. Notice the presence of a variety of coherent subpopulations. These correspond to the levels of activity described in the semantic:



The charts below illustrate the sub-clustering of the primary cluster that contained 55566 cases leading to two sub-clusters, one that contains 45413 cases, the other 10153 cases.

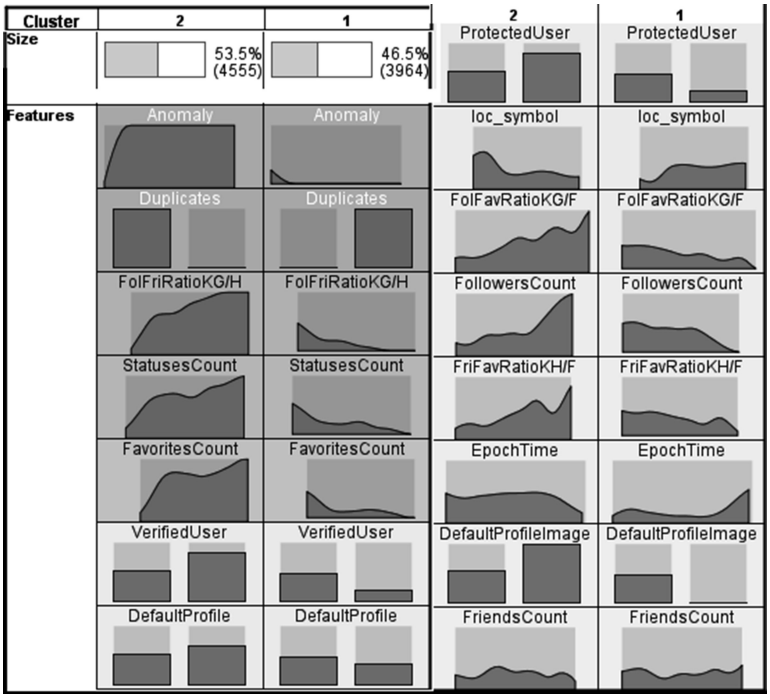


The Figure below shows Cluster profiles: the cells show different distributions of features within a single cluster for each varia

Additional Comments: In the future, we should collect more refined information on other properties of the variables with additional attributes, if available.

Other approaches could be used such as Simulation, Monte Carlo method, in order to construct some predictive regression model and use it on another more refined data.

200 Clusters: A dedicated clustering program was used to create 200 clusters for the original data set. Clusters varied in terms of number of cases, but of particular interest were small clusters that were characterized by high z-scores in one or more of the features. A number of clusters contained as few as 4 cases out of the 100001 and showed high z scores in a number of specific features. These small clusters represent cases that are far out in multidimensional feature space.



6 Future Work

Similarly, sub clustering of the smallest primary cluster produced two sub clusters, one that contains only 5 cases, consistent with small clusters obtained through separate 200 cluster procedure.

In this paper we presented an initial quantitative analysis based on the behavior of participating social media account users. During the analysis, a bot detection model was incorporated to collect and analyze data in an effort to investigate human and non-human behavior based on a structured framework. In the future, we aim to develop and incorporate prediction algorithms to assess the accuracy, scalability and resiliency of data based on selected features which would measure user's behavior.

References

1. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, pp. 56–65. ACM, August 2007
2. Twitter. API Overview, Object: Users. Twitter Developers Field Guide, 20 February 2016. https://dev.twitter.com/overview/api/users
3. Mischaud, E.: Twitter: expressions of the whole self. In: An Investigation into User Appropriation of a Web-Based Communications Platform. Media@lse, London (2007). Accessed 20 Oct 2011
4. Aladwani, A.M.: Facilitators, characteristics, and impacts of Twitter use: theoretical analysis and empirical illustration. Int. J. Inf. Manag. 15–25 (2015)
5. Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S.: Detecting automation of Twitter accounts. IEEE Trans. Dependable Secure Comput. 1–14 (2012)
6. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., Tesconi, M.: Fame for sale: efficient detection of fake Twitter followers. Decis. Support Syst. 56–71 (2015)
7. Ferrarra, E., Varol, O., Davis, C., Menczer, F., Flammini, A.: The Rise of Social Bots. Cornell University, New York (2015)
8. Kang, H., Wang, K., Soukal, D., Behr, F., Zheng, Z.: Large scale bot detection for search engine. In: WWW 2010 Proceedings of the 19th International Conference on World Wide Web, pp. 501–510. ACM, New York (2010)
9. Twitter. API Overview, Object: Users. Twitter Developers Field Guide, 20 February 2016. https://dev.twitter.com/overview/api/users
10. Xiao, C., Freeman, D.M., Hwa, T.: Detecting clusters of fake accounts in online social. In: AISec 2015 Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security, pp. 91–101. ACM, New York (2015)