Simulating User-Level Twitter Activity with XGBoost and Probabilistic Hybrid Models - Supplemental Materials

Fred Mubang [1] and Lawrence O. Hall[1]

I. Supplemental Material Information

This document contains supplemental information to the main VAM paper.

II. Annotation Set

Table I contains the 21 topics from the annotation set. The bolded topics are the final 10 topics chosen for training and testing VAM.

Twitter Topic Annotation Set Information			
Topic	**Weighted Average IAA**	**Label Count in Annotation Set**	**F1**
controversies/pakistan/students	**0.9308**	220	**0.97**
controversies/china/border	**0.9126**	309	**0.77**
leadership/sharif	**0.8980**	236	**0.86**
controversies/pakistan/baloch	**0.8589**	276	**0.71**
controversies/china/uighur	**0.8567**	25	**0.86**
leadership/bajwa	**0.8464**	722	**0.88**
benefits/development/roads	**0.8326**	571	**0.83**
benefits/covid	0.8276	242	0.67
benefits/development/energy	**0.8171**	335	**0.73**
benefits/jobs	**0.8124**	216	**0.75**
opposition/propaganda	**0.8046**	439	**0.75**
benefits/connections/afghanistan	0.7599	64	0.29
opposition/kashmir	0.7550	99	0.55
controversies/pakistan/bajwa	0.7533	165	0.73
controversies/china/exploitation	0.7379	210	0.57
leadership/khan	0.7376	246	0.63
controversies/pakistan/army	0.7269	129	0.19
controversies/china/naval	0.7261	24	0
controversies/china/funding	0.6225	46	0.4
benefits/development/maritime	0.6215	324	0.65
controversies/china/debt	0.6053	79	0.57

TABLE I: Twitter Topic Annotation Set Information. IAA stands for Inner Annatator Agreement. Topics were chosen for the Twitter dataset if the Inner-Annatator Agreement was at least 0.8 and if the F1 score of the BERT classifier on the test set was at least 0.7. The final chosen topics are in bold.

III. ARIMA Models

As previously mentioned in the main paper, ARIMA, ARMA, AR, and MA models were used as baselines against VAM. The models were trained in the following way. The ARIMA model has $p > 0$, $d > 0$, and $q > 0$. The AR model has $p > 0$, $d = 0$, and $q = 0$. The ARMA model has $p > 0$, $d = 0$, and $q > 0$. Lastly, the Moving Average (MA) model has $p = 0$, $d = 0$, and $q > 0$.

In order train each of these ARIMA-based models, a grid search was performed with p and q's possible values being 0, 24, 48, and 72, and d's possible values being 0, 1, and 2. This is the same grid search approach used in [1]. A different model was trained per topic/output-type pair. So, for example, the *(Benefits/Jobs, # of new users)* pair had its own ARIMA, ARMA, AR, and MA models. The validation data was used to select the best model parameters for the test period and the *RMSE* metric was used to select the best model parameters.

*This work is partially supported by DARPA and Air Force Research Laboratory via contract FA8650-18-C-7825.

[1]Department of Computer Science , University of South Florida, 4202 E Fowler Ave, Tampa, FL 33620, USA fmubang@usf.edu lohall@usf.edu

Twitter Hourly Active New/Old Frequencies		
Topic	**Avg. New User Freq (%)**	**Avg. Old User Freq (%)**
controversies/china/uighur	78.72	21.28
controversies/pakistan/students	75.0	25.0
benefits/jobs	66.67	33.33
opposition/propaganda	59.74	40.26
controversies/pakistan/baloch	50.0	50.0
leadership/bajwa	47.62	52.38
benefits/development/energy	47.5	52.5
benefits/development/roads	42.55	57.45
controversies/china/border	34.94	65.06
leadership/sharif	28.26	71.74

TABLE II: This table shows the average hourly proportion of new to old users per topic.

IV. New and Old User Information

Table II shows the average hourly proportion of new to old users per topic in the Twitter dataset.

V. Twitter Network Counts

Table III contains the node and edge counts of each of the 10 Twitter networks. The largest network in terms of nodes is the *controversies/china/border* network with 443,666 nodes. The smallest network in terms of nodes is the *controversies/pakistan/students* network, with 10,650 nodes.

Furthermore note that Table III also contains columns for *Edges* and *Temporal Edges*. An edge is defined as a user-user interaction (u, v), while a temporal edge is defined as a user-user interaction at some timestep t, or (u, v, t).

VI. User Assignment Diagram

Figure 1 is a pictoral representation of the User Assignment Module for easier understanding.

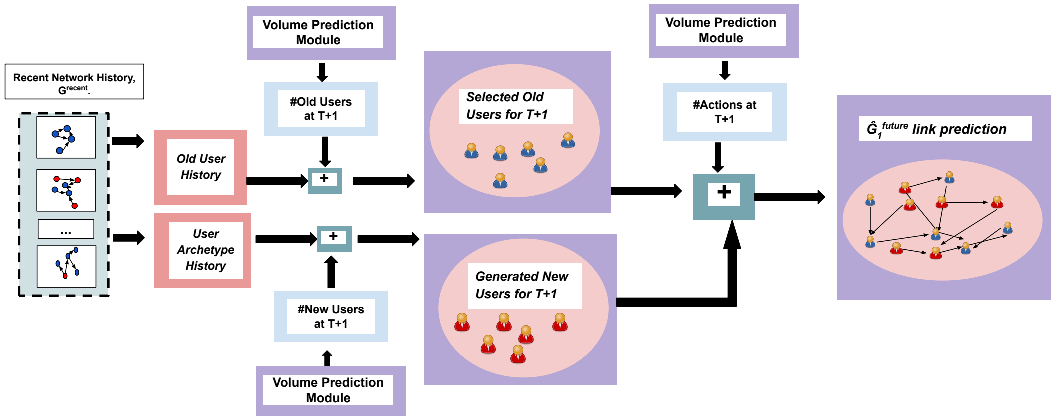

Fig. 1: This is an overview of the user-assignment module for 1 future timestep prediction at $T+1$. The recent network history (G^{recent}) is used to obtain *Old User History* and *User Archetype History*. This information, along with the counts from the *Volume Prediction* module, is used to predict the active old and new users at time $T+1$. These user sets, and the action volume counts are used to predict the links in the G_1^{future} set of edges for $T+1$.

Twitter Topic Network Counts			
Topic	**Nodes**	**Edges**	**Temporal Edges**
controversies/china/border	443,666	1,170,374	1,438,123
opposition/propaganda	170,942	281,023	296,690
controversies/china/uighur	133,542	164,484	171,590
controversies/pakistan/baloch	133,343	253,247	294,114
benefits/development/roads	74,042	148,345	179,432
benefits/jobs	71,914	98,038	110,304
benefits/development/energy	69,836	128,115	153,246
leadership/sharif	47,775	130,333	169,864
leadership/bajwa	35,320	87,836	99,783
controversies/pakistan/students	10,650	20,456	27,182

TABLE III: Twitter network information by topic.

VAM Jaccard Similarity Variation Information Per Topic			
Topic	**Mean**	**Standard Deviation**	**Variation Coefficient**
leadership/sharif	0.1352	0.0023	0.0172
opposition/propaganda	0.0958	0.002	0.0205
controversies/china/uighur	0.1621	0.0036	0.0222
benefits/development/roads	0.1192	0.0029	0.0247
controversies/pakistan/baloch	0.0567	0.0023	0.0408
controversies/china/border	0.0851	0.005	0.0589
benefits/development/energy	0.0744	0.0071	0.096
controversies/pakistan/students	0.062	0.0089	0.144
benefits/jobs	0.068	0.0123	0.1806
leadership/bajwa	0.1008	0.0185	0.1838

TABLE IV: VAM Jaccard Similarity Variation Information Per Topic

VII. Weighted Jaccard Similarity

As mentioned in the main paper, in order to measure the accuracy of the old user prediction task, the Weighted Jaccard Similarity metric was used, which is also known as the Ruzicka Similarity [2]. It was used to measure how well VAM predicted the old users in each hour, as well as how "influential" they were. In this case, influence is defined quantitatively as the number of retweets, replies, and quotes a user's tweets received.

We used the Weighted Jaccard Similarity in a similar fashion to the work of [1]. Let A represent the actual old user set within a particular hour, and let P represent the predicted set of old users within a particular hour. Furthermore, let $\mathbf{a}$ and $\mathbf{p}$ represent vectors that contain the weights of each user in the A and P sets, respectively. For example, $\mathbf{a}_k$ represents the weight of user A_k from the A set. With this in mind, the weighted Jaccard Similarity is defined as follows:

$$J(\mathbf{a},\mathbf{p}) = \frac{\sum_k min(\mathbf{a}_k,\mathbf{p}_k)}{\sum_k max(\mathbf{a}_k,\mathbf{p}_k)}$$

.

VIII. User Assignment Variation Information

As previously discussed in the main document, there were 5 trials run for the user-assignment algorithm (because it's probabilistic). The metric results for the Earth Mover's Distance, Relative Hausdorff Distance, and Jaccard Similarity were averaged for each of the 5 trials.

In Tables IV, V, and VI, we show the mean, standard deviation, and variation coefficient across all 5 trials for each topic and each of the 3 metrics. The values in the "Mean" column are the same values seen in the metric result tables in the main document. The "Standard Deviation" columns show the standard deviations across each of the 5 trials for each topic and metric. The "Variation Coefficient" columns show the variation coefficients, which are calculated by dividing the standard deviation by the mean. The variation coefficient gives a more clear view of how much each trial's metric result varied from the mean because it is a ratio of the standard deviation to mean. As one can see, in general, the variation coefficients are quite low, indicating that in general the different trial metric results did not vary by much.

REFERENCES

[1] F. Mubang and L. O. Hall, "VAM: An End-to-End Simulator for Time Series Regression and Temporal Link Prediction in Social Media Networks," *(In Review) IEEE Transactions on Computational Social Systems*, 2021.

[2] S.-H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *Int. J. Math. Model. Meth. Appl. Sci.*, vol. 1, 01 2007.

VAM Relative Hausdorff Distance Variation Information Per Topic			
Topic	**Mean**	**Standard Deviation**	**Variation Coefficient**
controversies/pakistan/baloch	0.9015	0.0061	0.0068
benefits/development/roads	0.7651	0.0057	0.0075
controversies/pakistan/students	0.6138	0.0069	0.0112
controversies/china/uighur	0.7696	0.0089	0.0115
leadership/sharif	0.7985	0.0115	0.0144
leadership/bajwa	1.0904	0.0175	0.0161
benefits/development/energy	0.688	0.0117	0.017
controversies/china/border	0.9512	0.0176	0.0185
opposition/propaganda	1.2339	0.0239	0.0194
benefits/jobs	0.6669	0.0165	0.0247

TABLE V: VAM Relative Hausdorff Distance Variation Information Per Topic

VAM Earth Mover's Distance Variation Information per Topic			
Topic	**Mean**	**Standard Deviation**	**Variation Coefficient**
controversies/pakistan/baloch	0.0358	0.0003	0.0085
benefits/development/energy	0.1896	0.0029	0.015
benefits/development/roads	0.1076	0.002	0.0184
controversies/pakistan/students	0.1945	0.0038	0.0195
benefits/jobs	0.2137	0.0044	0.0207
controversies/china/border	0.1144	0.0026	0.0228
leadership/sharif	0.082	0.0021	0.0258
leadership/bajwa	0.1971	0.0069	0.0351
controversies/china/uighur	0.1137	0.0043	0.038
opposition/propaganda	0.0963	0.0045	0.0472

TABLE VI: VAM Earth Mover's Distance Variation Information per Topic