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

Anonymous

I. Supplemental Material Information

This document shows supplemental information to the main VAM paper.

A. 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.

B. New and Old User Table

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

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.

C. 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).

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.

D. User Assignment Diagram

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

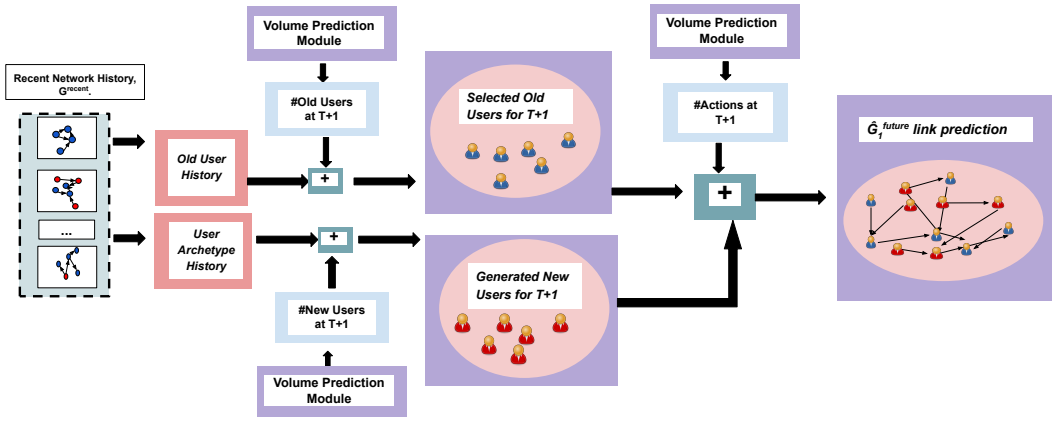

Fig. 1: This is an overview of the user-assignment module for 1 future timestep prediction at $T+1$. The recent network history (G^{recent}) is used to obtain *Old User History* and *User Archetype History*. This information, along with the counts from the *Volume Prediction* module, is used to predict 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 $T+1$.