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

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

Abstract*— These are supplemental materials to the paper *Simulating User-Level Twitter Activity with XGBoost and Probabilistic Hybrid Models

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.

*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