Proceedings of the 2021 Winter Simulation Conference
S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo, and M. Loper, eds.

SIMULATING ONLINE SOCIAL RESPONSE: A STIMULUS/RESPONSE PERSPECTIVE

Huajie Shao[*], Tarek Abdelzaher[*], Sam Cohen[†], James Flamino[††], Jiawei Han[*], Minhao Jiang[*],
Gyorgy Korniss[††], Omar Malik[††], Aamir Mandviwalla[†††], Yuning Mao[*], Yu Meng[*], Wenda Qiu[*], Dachun Sun[*],
Boleslaw Szymanski[†††], Ruijie Wang[*], Chaoqi Yang[*],
Zhenzhou Yang[*], Shuochao Yao[*], Lake Yin[†††], Xinyang Zhang[*], Yu Zhang[*]

[*]Department of Computer Science, University of Illinois at Urbana Champaign, Urbana, IL 61801, USA
[†]Department of Information Technology and Web Science,
[††]Department of Physics, Applied Physics, and Astronomy,
[†††]Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180, USA

ABSTRACT

The paper describes a methodology for simulating online social media activities that occur in response to external events. A large number of social media simulators model information diffusion on online social networks. However, information cascades do not originate in vacuum. Rather, they often originate as a *reaction to events* external to the online medium. Thus, to predict activity on the social medium, one must investigate the relation between external stimuli and online social responses. The paper presents a simulation pipeline that features stimulus/response models describing how social systems react to external events of relevance to them. Two case studies are presented to test the fidelity of different models. One investigates online responses to events in the Venezuela election crisis. The other investigates online responses to developments of the China Pakistan Economic Corridor (CPEC). These case studies indicate that simple macroscopic stimulus/response models can accurately predict aggregate online trends.

1 INTRODUCTION

Physical events and their depictions in mainstream news media engender reactions in the online world. An interesting question (addressed in the simulator described in this paper) is therefore to understand how various *topic-specific discussions* on an online social medium would react to a particular external event. For example, how might online discussions of military aid get affected by election news? What groups will be activated online by the signing of a deal to develop a hydroelectric power plant with foreign aid in a given location? What repercussions might a particular border skirmish have on online discussions pertaining to the underlying geopolitical context? To answer such questions, we develop stimulus/response models of online social systems, where the stimulus comprises reported world events and the response comprises the corresponding online social media footprints separated by online discussion topic. We train our models by observing a series of key events in two different scenarios and build models of online social response using a set of estimation techniques. These models are then tested on other events withheld from the training data. Different modeling solutions are compared to understand their strengths and limitations.

A key challenge in accurately reconstructing topic-specific online social responses is to understand the correlations between event types (for events transpiring in the *physical* world) and the different discussion topics (for discussions occurring *online* on a particular social medium). This question is somewhat separable from the more commonly asked question of how a particular online information cascade (e.g., series of retweets) might evolve (Arnaboldi et al. 2017), or how a group of online agents might react to each other or react to specific online nodes (Gyarmati and Trinh 2010). The difference lies in that we explicitly consider interactions between the online medium and external stimuli, as opposed to interactions that occur entirely within the online medium. The main contribution of this work lies in developing and understanding the fidelity of stimulus/response models that correlate online activities with external events.

The rest of this paper is organized as follows. Section 2 briefly outlines related work. Section 3 summarizes the simulation pipeline. Sections 4, 5, and Section 6 describe the three main stages of the pipeline; online stimulus estimation from descriptions of external events, assessment of online influence, and prediction of online response. Section 7 overviews two case-studies used for evaluation. Section 8 presents evaluation results. We discuss the findings in Section 9 and conclude the paper in Section 10.

2 RELATED WORK

The simulator described in this paper complements and evolves a series of tools reported in previous work (Yao et al. 2018; Abdelzaher et al. 2020) that focus on *agent-based* online social media simulation approaches. Contrary to the aforementioned agent-based solutions, we take a *macroscopic view* of the online social medium, modeling it collectively as a dynamic system that responds to external stimuli by emitting online microblogs.

A complete review of *online social media* simulation is beyond the scope of this section. Several survey papers covered aspects of social media modeling, analysis, and simulation in recent years (Ghani et al. 2019; Li et al. 2017). Most simulators extend agent-based social network simulation models (Madey et al. 2003; Railsback et al. 2006), often targeting specific applications. For example, Lanham et al. (Lanham et al. 2013) proposed an agent-based simulation model in support of crisis de-escalation; Zhang et al. (Zhang et al. 2015) built an agent-based model to explore peer influence;In contrast to agent-based solutions, we propose a macroscopic stimulus/response simulation technique.

Social media simulation is tightly linked to modeling different aspects of individual or collective user behavior and propagation patterns (Althoff et al. 2017; Lei et al. 2018) in social networks. Many aspects of online behavior were studied in recent years. For example, Shao et al. (Shao et al. 2018; Shao et al. 2020; Shao et al. 2019) proposed some unsupervised learning models to evaluate the reliability of users and the veracity of their reported claims. Liu et al. (Liu et al. 2019) explored content propagation among users with the same interest in location-based social networks. Arnaboldi et al. (Arnaboldi et al. 2017) studied the impact of ego network structures on information diffusion in social networks.

Models were also developed for predicting the number of events or user activities over time on social media. Bidoki et al. (Hajiakhoond Bidoki et al. 2019) adopted LSTM in Recurrent Neural Networks (RNN) to predict bursts based on cross-platform social media data. Xu et al. (Xu et al. 2018) proposed a novel model to predict the peak time and peak volume of bursting hashtags. In order to predict future event arrivals, Liu et al. (Liu et al. 2019) captured latent interactions among different social communities with a latent Hawkes process model. While the above listing is anecdotal and incomplete, this paper differs from such examples in its focus on generating holistic (and realistic) simulation traces of online discussions in the context of some background physical events. Specifically, we develop a novel simulator that predicts topic-based response to *external stimulae*, when events occur in the real world that impact online discussion of different topics on social media.

3 THE PIPELINE: STIMULUS, PERCEPTION, AND RESPONSE

Our simulation pipeline is composed of three stages: (i) *computing the external stimulus*, (ii) *assessing its perception* (or influence) within the social medium, and (iii) *estimating the social response* (in a topic-specific fashion). We do not use the concepts of *stimulus* and *perception* interchangeably. Rather, by *stimulus* we refer to an external quantity or signal that changes in the environment. In contrast, by *perception* we refer to the signal as perceived by the online social medium. Eventually, the *response* of the online social medium (e.g., the emitted posts) is computed as a function of signal perception. The distinction between stimulus, perception, and response is motivated by other biological systems with imperfect sensors. For example, in human auditory systems, the original stimulus might be the external absolute sound intensity as measured by an accurate sensor. The perception is intensity as *perceived* by the human. This perceived intensity is different from the absolute external ground-truth in that it is typically scaled logarithmically due to nonlinearities in human auditory perception, and it depends on the frequency (leading to the notion of Mel-frequency cepstrum (On et al. 2006)). The response follows as a function of the perceived signal.

In social systems, several phenomena impact perception. For example, hearing the same news repeatedly is not as exciting as hearing the news the first time. Thus, social systems get gradually desensitized to persistent stimuli. The perception of news importance may also exhibit a nonlinearity, where it may grow with a (non-linear) function of popularity.

Accordingly, our simulator first computes an *external topic-specific stimulus* time-series (where by *external* we mean external to the online medium). In the current implementation, this time-series simply represents the volume of topic-specific news coverage of the underlying events. Second, we perform signal conditioning on the external topic-specific stimulus time-series to better approximate its *perceived* influence on the social medium. We call the result of that conditioning, the *perception* time-series. Finally, a response is computed to the perceived signal. In the following three sections, we discuss stimulus, perception, and response, respectively.

4 COMPUTING THE STIMULUS

4.1 The DMG Stimulus Family

This stimulus family uses a data mining and NLP (natural language processing) approach to transform real-world events into a stimulus time-series by analyzing the news articles text. As a proxy for world events, we use the GDELT event database (https://www.gdeltproject.org/). GDELT monitors news media from around the world continuously and generates corresponding event records, containing event types, parties involved, geological locations, timestamps, media sources (article URLs), and other details. More specifically, to compute the stimulus, news article URLs are collected from the database that are linked to the GDELT events. Those articles are then analyzed for relevance to a predefined set of user-supplied topics, using a text classifier, C, that computes a measure of similarity d_{ij} between each article, a_j, and each topic, c_i. The stimulus, $U_i(k)$, for topic, c_i, within a given time interval, k (set in our simulator to be one day), is set proportional to the amount of news on the topic within that time interval, as determined by classifier C. Let $\mathscr{P}_k$ be the set of articles published in the kth time interval. Thus, the stimulus time-series is given by:

$$U_i(k) = \sum_{a_j \in \mathscr{P}_k} d_{ij} \tag{1}$$

In practice, for efficiency of implementation, the set $\mathscr{P}_k$ is pruned ahead of time by extracting from GDELT only those events that are relevant to the underlying general geopolitical context, as well as their related articles. In our current implementation, we classify text based on the pretrained language model, BERT (Devlin et al. 2019), that has been widely used in the NLP domain for various downstream tasks. The BERT model leverages a large pretraining corpus on Wikipedia, then fine-tunes its model parameters on specific tasks (in our case, the topic classification task). The BERT classifier was fine-tuned by LEIDOS under the DARPA SocialSim program (Brian Kettler 2020) using tweet text with topic labels. Although the news text and tweet text may differ in expressions, empirical results show that it has the ability to transfer from tweet text classification to news text classification. We call the stimulus time-series, $U_i(k)$, described above, DMG-BERT.

This method requires topic-labeled data. Recently, unsupervised text classification methods were developed as well, such as WeSTClass (Meng et al. 2018) and LOTClass (Meng et al. 2020). We chose BERT-based methods instead because BERT models were shown to have a more robust performance. We leave unsupervised extensions to future work.

4.2 The CORR Stimulus Family

This family is a slight modification to the above, where instead of classifying articles by tweet topics, we break this operation into two. First, we leverage the internal GDELT event type field to categorize articles (linked to different GDELT events) by the corresponding GDELT event type. Second, we use a BERT-based classifier, similarly to the previous section, to compute a correspondence between GDELT event categories and topics of interest. The BERT classifier, in this case, is tuned based on labeled articles not labeled tweets. We use it to compute a semantic correlation between GDELT event types and topics of interest. In this paper, we use about 100 different GDELT event types. During the learning phase, we collect all the news articles corresponding to each event type, e_j, into a set, and compute the average similarity score, D_{ij}, between articles in each set and each social media topic c_i, forming a correlation-like matrix, D. At inference time, the *stimulus* for topic c_i, within time interval, k, is given by:

$$U_i^{CORR}(k) = \sum_j D_{ij}\ GDELT_j(k) \tag{2}$$

where $GDELT_j(k)$ is the number of articles published on GDELT in the time interval k that are of GDELT event type, j. We call this alternative stimulus time-series, CORR-TUNED-BERT. An advantage of this version is that it leverages the categorization of articles by GDELT event types. For topics aligned with GDELT event types, this solution tends to do well. Otherwise, the DMG-BERT stimulus is preferred.

5 COMPUTING THE PERCEPTION

The algorithms described above produce a topic-specific stimulus time-series for each topic, c_i. In general, the mapping between this stimulus and its perception by the social system is an open question that calls for more collaborative research between computer scientists, cognitive scientists, and social psychologists. There are several factors at play. For example, humans tend to pay more attention to more outlying news (Fiske 1980; Lamberson and Soroka 2018). They are drawn to statistical surprise (Varshney 2019) and tend to get desensitized to repetitive and aging topics in favor of new changes (Shoemaker 1996). They also tend to be more influenced by bursty coverage (e.g., when coverage of the same new event is carried by more sources within a shorter period of time), as opposed to when coverage is more scattered or the covered events are more loosely related (Doyle, Szymanski, and Korniss 2017). Our simulator approximates the effects of the above factors in order to compute how a stimulus is *perceived* by the online social medium. This function is described next.

5.1 Rewarding More Bursty and Coherent Signals

The first step in the mapping between *external stimulus* and its *perception* is to capture the human preference for bursts of more focused coverage (Doyle, Szymanski, and Korniss 2017). To do so, we need to analyze the statistical features of news articles. In particular, we focus on the normalized word use frequency distribution (with stop-words filtered) in each time interval, k, for each topic, c_i. It is easy to observe that a more topically-focused coverage results in a more skewed word use frequency distribution as the word usage among many articles converges to a focused set, boosting a few common keywords at the expense of others, producing the skew. In contrast, coverage of multiple scattered events that are loosely connected by a topic tends to result in a more uniform word use distribution.

We fit the empirically observed frequency distribution of words used in the set of articles on each topic, c_i, in each time interval, k, to a Zipf distribution (Tullo and Hurford 2003). Let us denote it coefficient by $s_i(k)$. Recall that a Zipf distribution states that when $N >> 1$ words are ranked by frequency of use, the probability of use of the word at rank m is approximately:

$$\text{Zipf}(m; s, N) = \frac{1/m^s}{\sum_{n=1}^{N}(1/n^s)}, \tag{3}$$

where s is the coefficient characterizing the distribution. The higher the s, the more skewed the distribution. A higher skew implies a more focused coverage (and thus one that is more likely to garner attention). We therefore use the Zipf coefficient, $s_i(k)$, as a weight (i.e., a multiplicative factor) multiplied by the corresponding stimulus time-series value ($U_i(k)$ or $U_i^{CORR}(k)$) computed in Section 4, essentially associating more importance to more focused coverage of a topic.

5.2 Desensitization and Attenuation of Stable Stimuli

Information consumers are "wired" to favor new content, while getting desensitized to old news (Fiske 1980; Varshney 2019; Shoemaker 1996). To simulate this effect, we apply (i) a *high-pass filter*, and (ii) a non-linearity to the topic-specific stimulus time-series. The high-pass filter attenuates signal components that change slowly or are sustained over long periods of time. A logarithmic non-linearity models saturation of human attention (much like with auditory perception). Combining this computation with that of the previous section, the entire conditioning algorithm becomes as follows:

$$Y_i(k) = \begin{cases} s_i(0)\ U_i(0) & \text{if } k = 0 \\ \max\{0, pY_i(k-1) + s_i(k)(U_i(k) - U_i(k-1))\} & \text{if } k > 0 \end{cases} \tag{4}$$

$$Z_i(k) = q_a Y_i(k) + q_b \log\left(1 + Y_i(k)\right) \tag{5}$$

where $U_i(k)$ is the stimulus computed in Section 4, $p \in (0,1)$ is a parameter of the high-pass filter (for attenuating sustained signals), and q_a and q_b are parameters of the non-linearity. The output, $Z_i(k)$, is then used as the perception time-series. In subsequent sections, it is called DMG-ENT-BERT.

6 SIMULATED RESPONSE

Given the perception time-series described above, the next step is compute aspects of online response. For each time interval (e.g., day), k, of the simulation, the simulator predicts (i) the *total number of posts* that day on the online social medium, (ii) the *total number of users* activated on the online social medium that day, and (iii) the *number of new or reactivated users* involved that day. We define a *new or reactivated user* (new user for short) as someone who has not posted online on the topic for at least some predetermined minimum window of time (e.g., two months). One can think of these time-series as envelopes of daily activity. Using those envelopes as aggregate constraints, it becomes possible to use a more conventional agent-based simulator to fill-in the actual sequence of individual posts and their sources. Below, we first describe models for *envelope prediction*. For completeness, we also describe a trivial algorithm for the generation of the sequence of individual posts, which we call *envelope filling*.

Given the computed perception time-series, the simplest assumption is that the online social medium will respond to the perception time-series *linearly*. We can use regression to compute a scaling factor and time-shift between the perception time-series and each of the three outputs using training data. Those values can then be applied to subsequent perception inputs to predict the outputs.

A slightly more detailed model may be to use classification and regression tree (CART) algorithms (Lewis 2000) to predict the social response given the perception time-series. These algorithms build a tree-structure by recursively splitting the training data via feature threshold cuts to minimize the overall weighted loss that the split data achieves. Within each split, a different regression model is applied. Thus, by increasing the depth of the model tree, nonlinear time-series data can be fitted arbitrarily more closely to a collection of linear models. To cope with data with specific characteristics, such as sudden large spikes, we construct the model tree using linear regression and LASSO regression as the underlying regression techniques.

Finally, long-range dependencies may exist in a wide range of naturally occurring phenomena. Employing detrended cross-correlation analysis (Podobnik and Stanley 2008; Zhou 2008) we observed long-range dependencies between the external stimulus and the response time-series. To capture these long-range dependencies, another model is thus to use Long Short-Term Memory (LSTM) Neural Networks (Hochreiter and Schmidhuber 1997). LSTM networks are well suited to time-series prediction with long-range dependencies and time-lags between stimulus and response, such as the prediction of traffic flow (Azzouni and Pujolle 2018). LSTMs are a specialized form of Recurrent Neural Networks (RNNs). RNNs have hidden layers that consist of recurrent cells and the state of the cell is determined both by states in the past and the current input. LSTMs differ from standard RNNs by the introduction of input gates that prevent the state and output of the cell from being affected by irrelevant information. Since LSTMs, CART algorithms, and regression solutions are standard modeling tools, we skip their mathematical formulation above in the interest of brevity.

With envelopes of online activity predicted by (one of) the algorithms discussed above, it remains to populate the envelopes with actual posts. Solutions for doing so have been described in past work (Yao et al. 2018; Abdelzaher et al. 2020) and are not the focus of this paper. Instead, we use a simple approach that selects posting times, probability of new user activation, and probability of current user activation to match those measured in training data. We repeat the process until enough (event and user) activations are generated to reach the respective envelopes. This process produces a realistic looking set of posts that add up to the right total volume as predicted by response estimation algorithms covered above.

7 TWO CASE STUDIES

The algorithms described above comprise the components of a *simulator* concerned with *realistic reproduction of topic-specific activity levels on the social medium*, given a set of concurrent physical events that transpired and their representation in the news. To train and test this simulator, we consider real news on past events. We use some period of time for training and use a subsequent period for testing. Testing exploits the observation that our simulator can be used for (limited) prediction. Specifically, given today's events on GDELT, one should be able to invoke the simulator to predict the volume of online discussion on different topics. It is then possible to compare those predictions to data directly collected from the online social medium and compute a notion of error. In this evaluation, we use two data sets collected by LEIDOS from the complete Twitter firehose as part of the DARPA SocialSim program, originally created by Jonathan Pfautz (Davis, O'Mahony, and Pfautz 2019) then managed by Brian Kettler (Brian Kettler 2020). The two case studies are described below.

7.1 Venezuela Elections

In January 2019, the presidential elections in Venezuela resulted in naming Nicolás Maduro the president. His opposition, Juan Guaidó, declared the elections unconstitutional and swore himself as acting president. Venezuela (and the international community) would remain divided with some countries recognizing Maduro and others recognizing Guaidó. It is in this context that the simulator was used to predict online activity on multiple topics, including protests, violence, military operations, international aid, and support/opposition (to the two parties in question), among several other topics. In our experiment, we use GDELT events and Twitter posts about Venezuela from December 24th, 2018 to February 14th, 2019 as training data (to estimate parameters of perception and response models, given a stimulus), and use GDELT events from February 15th to March 14th, 2019 for testing the fidelity of online response simulation, given GDELT events in the testing window. (All data was collected by LEIDOS as part of the SocialSim program.)

7.2 The New Silk Road

The China–Pakistan Economic Corridor (CPEC) is a collection of infrastructure projects in Pakistan financed by China as part of its global economic development program known informally as the New Silk Road (or the Belt and Road initiative). It refers to the development of land and maritime routes reminiscent of the ancient trade routes that once connected China to the Roman Empire. An initiative of this scale is surrounded with much debate on economic and political implications. In this context, we used the simulator to predict several aspects of that debate in Pakistan, focusing on online response to specific physical activities such as road development, energy projects, jobs, local political figures, and border tensions, among other topics.

In our experiments, we use the GDELT and Twitter data from March 30th to July 13th, 2020 as training data (to estimate parameters of perception and response models, given a stimulus), and use data from July 14th to August 17th, 2020 to test fildelity of social response estimation given GDELT events in the testing window. (All data was collected by LEIDOS as part of the SocialSim program.)

8 EVALUATION RESULTS

To evaluate the simulator, we considered four metrics. First, the *Absolute Percentage Error (APE)* on the number of total posts (and users): This metric evaluates the accuracy of models at predicting the overall number of posts (and activated users) that would be present in a testing window. Second, *Root Mean Square Error (RMSE)* over normalized cumulative time-series: This metric evaluates the RMSE between the real and simulated cumulative response time-series, normalized by the final total. Third, *Relative Hausdorff Distance (RH)* over the in-degree distribution of the user interaction graph: This metric evaluates the user interaction structure of the simulated responses. RH distance (Aksoy et al. 2019) is used because it requires the distribution to match at every point, so any outlier behavior in the ground truth must be approximated correctly (Simpson et al. 2015). Finally, *Earth Movers Distance (EM)* applied on graph's PageRank distributions: This metric evaluates the difference of two user interaction graph's PageRank distributions (Rubner, Tomasi, and Guibas 1998). We then compare the following models:

- Replay: This is the simplest baseline. It replays the most recent past. This baseline should accurately reproduce statistical properties of online response (since it, in fact, replays real data seen recently on the medium), but it offers no correspondence with current external stimuli. Improvements over this baseline can thus be attributed to stimulus/response models.
- Linear-DMG-BERT: This approach classifies GDELT articles using DMG-BERT as described in Section 4.1, then utilizes linear regression to predict the response time-series. This algorithm assumes that perception is equal to stimulus, skipping the adjustment described in Section 5.
- Linear-DMG-ENT-BERT: This approach is similar to Linear-DMG-BERT, except that a perception model (DMG-ENT-BERT) is computed, as described in Section 5, before computing response.
- Piecewise-Linear: This approach enhances Linear-DMG-BERT by using a piecewise-linear model (based on decision trees) instead of a linear model to predict response.
- Linear-CORR-TUNED-BERT: In contrast to the above two, this approach modifies Linear-DMG-BERT by using the CORR stimulus family.
- LSTM: Finally, this approach adopts LSTM, a variant of Recurrent Neural Networks (RNN), to predict the desired time series data.

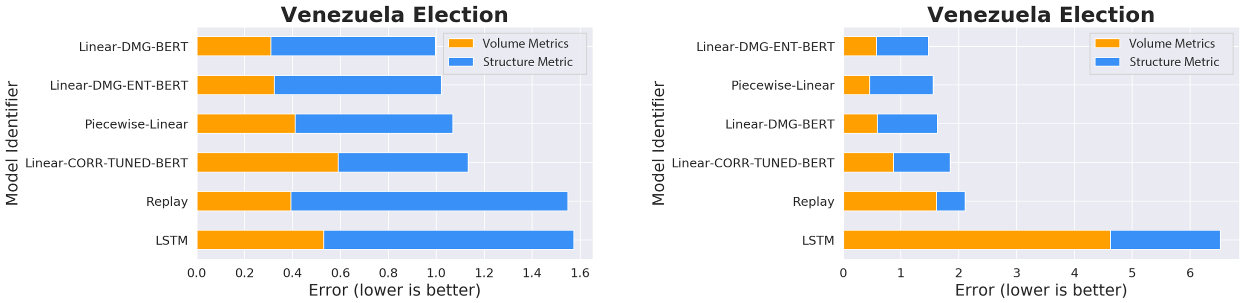
8.1 Simulation Results: Venezuela

We first evaluate the performance of the proposed models on the Venezuela data set. Table 1 shows the evaluation results based on the four metrics: APE, RMSE, RH distance, and EM Distance, explained above. Note that, Linear-DMG-ENT-BERT is better than the baseline, Replay, in almost all metrics, suggesting benefits to stimulus/response models.

Table 1: Error metrics for the Venezuela dataset, from February 15th to March 14th, 2019.

Models & Metrics	Volume				Structure	
	APE		RMSE		RH Distance	EM Distance
	#Events	#Activated Users	#Events	#Activated Users	Graph Degree Dist.	PageRank
Linear-DMG-ENT-BERT	33.0%	90.0%	0.14	0.35	1.60	0.001
Linear-DMG-BERT	32.5%	89.0%	0.15	0.33	1.73	0.000
Linear-CORR-TUNED-BERT	51.5%	175.0%	0.14	0.71	1.48	0.003
Piecewise-Linear	38.0%	75.5%	0.19	0.44	1.75	0.002
LSTM	40.0%	615.5%	0.17	0.58	2.95	0.001
Replay	29.5%	170.0%	0.15	0.41	1.65	0.002

To facilitate rough overall comparisons, we combine the metrics and separate two different evaluation periods. Fig. 1 shows the results. The four APE and RMSE metrics are averaged and shown in orange. The RH and EM distances are averaged and shown in blue. We find that simpler models beat Replay, in aggregate, whereas the neural network model does not. This effect is most likely attributed to overfitting.



(a) Time window: Feb 15th to 28th, 2019

(b) Time window: Mar 1st to 14th, 2019

Figure 1: Combined error for the Venezuela dataset. *Volume metrics* display the average of the four volume metrics in Table 1, whereas *structure metrics* display the average of the two structure metrics in Table 1, disaggregated by simulation time window.

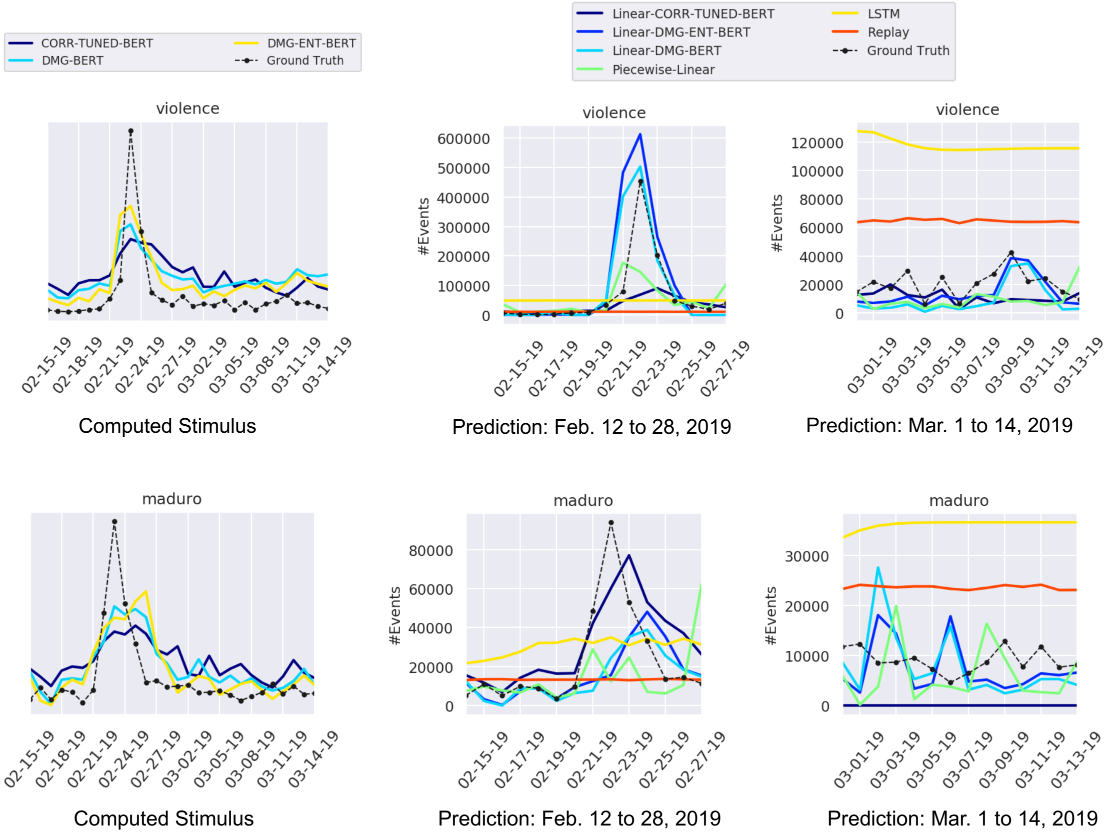
Fig. 2 shows example predictions of online activity compared to ground truth activity for two specific topics; *violence* and *Maduro* (the president). The figures in the left column show the stimulus and perception time series for each topic. Specifically, DMG-BERT is the stimulus family computed in Section 4.1. CORR-TUNED-BERT is the stimulus family computed in Section 4.2. DMG-ENT-BERT is the perception time-series computed in Section 5. These are the inputs to the response models. Since different inputs have different scales, we discard the scale of the y-axis in the leftmost figures altogether to focus on shape similarity (and thus information overlap with ground truth, shown in black dashed lines). Note that, input scale is not important since the response models will re-scale the input anyway. The other two columns show predictions of daily posts in two different simulation intervals compared to ground truth. The figures reveal a correlation between ground-truth and predicted online response (especially in February data).

8.2 Simulation Results: Silk Road

Next, we show results for the Silk Road dataset. Table 2 shows the performance comparison for different methods using the same same four metrics: APE, RMSE, RH distance, and EM distance. It can be observed that Linear-DMG-ENT-BERT performs the *best* under different metrics. Next, we break the test period into two overlapped one month periods and show the error in each period in Fig. 3. Overlap illustrates the effect of a sliding window, where one month simulations are made based on all training data collected up to the beginning of that window. As before, the four APE and RMSE metrics are averaged and shown in orange. The RH and EM distances are averaged and shown in blue. Observe that Linear-DMG-ENT-BERT has the best performance in both testing windows. Besides, it can be seen that all models beat the baseline, Replay, except for LSTM. We believe this is because LSTMs have too many parameters for the small data set, leading to overfitting. We present some examples to show the predicted time series for different models in Figure 4.

9 DISCUSSION

The results shown in this paper demonstrate that stimulus/response models generally improve prediction over mere (statistically-accurate) replay. They also confirm the intuition that simpler models work better than more complex (e.g., neural network) models, when the amount of training data is limited. The results also suggest that separating stimulus and perception generally improves response prediction accuracy.



Computed Stimulus

Prediction: Feb. 12 to 28, 2019

Prediction: Mar. 1 to 14, 2019

Computed Stimulus

Prediction: Feb. 12 to 28, 2019

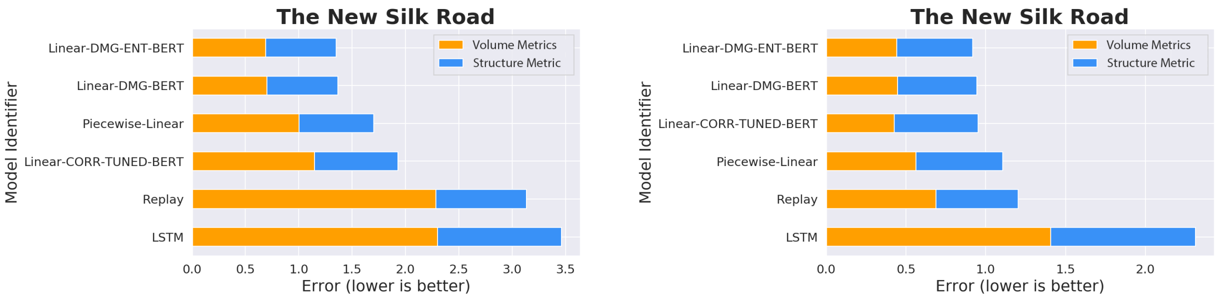
Prediction: Mar. 1 to 14, 2019

Figure 2: Stimulus and response for the Venezuela case study.

Specifically, the algorithm that explicitly models perception (Linear-DMG-ENT-BERT) generally performs best. Finally, the results show that while in some cases significant correlations exist between GDELT events (classified by BERT) and related topic-specific discussions on online social media (Twitter in this case), more work is needed to fully predict online response. There are examples where significant spikes occur in online social media discussions that do not correspond to spikes in the related stimulus time-series. For instance, as exemplified by Figure 4 (bottom row), there are spikes in discussions of leadership/Bajwa (a local leader), that do not correspond to spikes in GDELT events of that class, according to BERT. More sources of external influence might therefore need to be uncovered for an accurate response simulation. For instance, online activity spikes not driven by external news might be an indication of organized online information campaigns. We leave further investigation of this hypothesis to future work.

Table 2: Error metrics for the Silk Road dataset, from July 14th to August 17th, 2020.

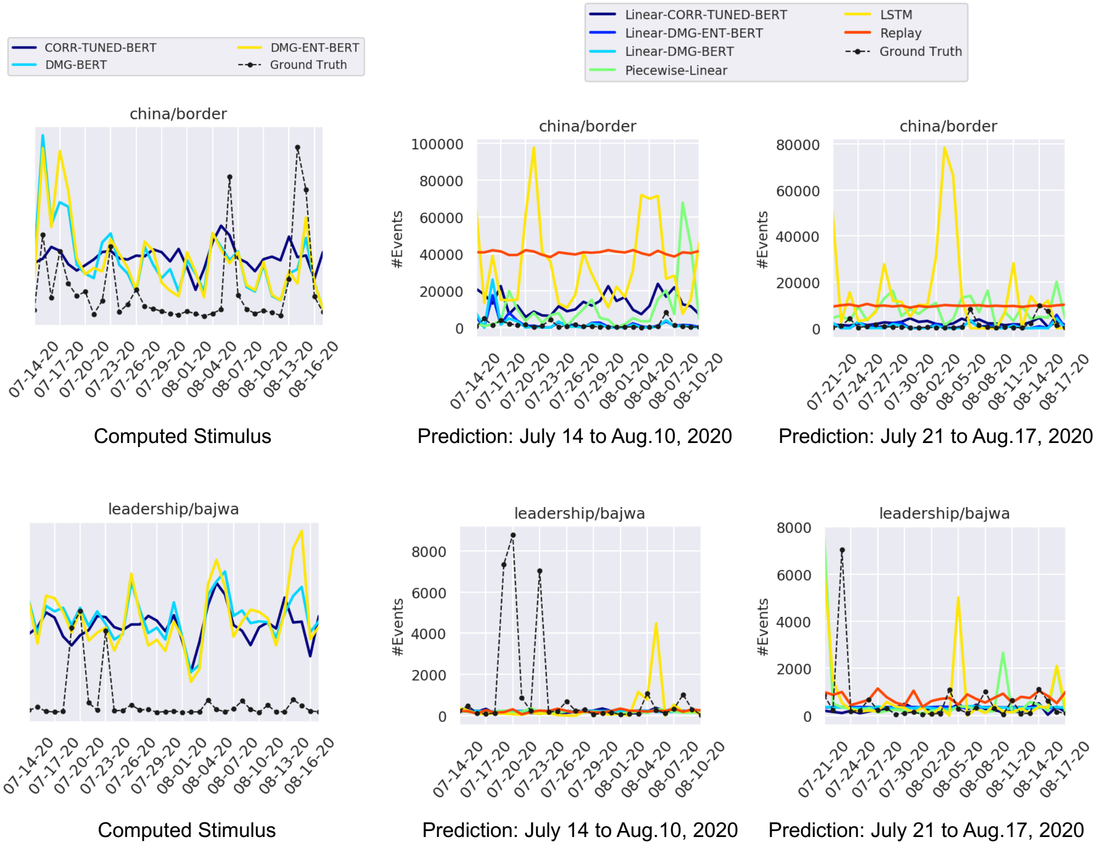
Models & Metrics	Volume				Structure	
	APE		RMSE		RH Distance	EM Distance
	#Events	#Activated Users	#Events	#Activated Users	Graph Degree Dist.	PageRank
Linear-DMG-ENT-BERT	70.0%	120.5%	0.18	0.19	1.14	0.001
Linear-DMG-BERT	71.5%	121.0%	0.18	0.19	1.16	0.001
Linear-CORR-TUNED-BERT	111.5%	165.5%	0.19	0.19	1.31	0.002
Piecewise-Linear	109.5%	168.5%	0.17	0.18	1.25	0.002
LSTM	260.0%	432.5%	0.24	0.25	2.07	0.005
Replay	267.5%	289.0%	0.18	0.20	1.36	0.003



(a) Time window: July 14th to Aug 10th, 2020

(b) Time window: July 21th to Aug 17th, 2020

Figure 3: Combined error for the Silk Road dataset. *Volume metrics* display the average of the four volume metrics in Table 2, whereas *Structure metrics* display the average of the two structure metrics in Table 2, disaggregated by simulation time window.



Computed Stimulus

Prediction: July 14 to Aug.10, 2020

Prediction: July 21 to Aug.17, 2020

Computed Stimulus

Prediction: July 14 to Aug.10, 2020

Prediction: July 21 to Aug.17, 2020

Figure 4: Stimulus and response for the Silk Road case-study.

10 CONCLUSIONS

In this paper, we presented a simulation pipeline that features stimulus/response models describing how social systems react to external events of relevance to them. Different from information diffusion mechanisms, we explored the relation between external stimuli and online social responses to predict activity bursts on the social medium. Two case-studies were conducted to demonstrate the fidelity of different models using external stimulus.

REFERENCES

Abdelzaher, T., J. Han, Y. Hao, A. Jing, D. Liu, S. Liu, H. H. Nguyen, D. M. Nicol, H. Shao, T. Wang et al. 2020. "Multiscale online media simulation with SocialCube". *Computational and Mathematical Organization Theory*:1–30.

Aksoy, S. G., K. E. Nowak, E. Purvine, and S. J. Young. 2019. "Relative Hausdorff distance for network analysis". *Applied Network Science* 4(1):1–25.

Althoff, T., P. Jindal, and J. Leskovec. 2017. "Online actions with offline impact: How online social networks influence online and offline user behavior". In *Proceedings of the tenth ACM international conference on web search and data mining*, 537–546.

Arnaboldi, V., M. Conti, A. Passarella, and R. I. Dunbar. 2017. "Online social networks and information diffusion: The role of ego networks". *Online Social Networks and Media* 1:44–55.

Azzouni, A., and G. Pujolle. 2018. "NeuTM: A neural network-based framework for traffic matrix prediction in SDN". In *NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium*, 1–5.

Brian Kettler (URL accessed date) 2020. "Computational Simulation of Online Social Behavior (SocialSim)". https://www.darpa.mil/program/computational-simulation-of-online-social-behavior.

Davis, P. K., A. O'Mahony, and J. Pfautz. 2019. *Social-Behavioral Modeling for Complex Systems*. John Wiley & Sons.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019, June. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Doyle, C., B. K. Szymanski, and G. Korniss. 2017. "Effects of communication burstiness on consensus formation and tipping points in social dynamics". *Physical Review E* 95(6):062303.

Fiske, S. T. 1980. "Attention and weight in person perception: The impact of negative and extreme behavior.". *Journal of personality and Social Psychology* 38(6):889.

Ghani, N. A., S. Hamid, I. A. T. Hashem, and E. Ahmed. 2019. "Social media big data analytics: A survey". *Computers in Human Behavior* 101:417–428.

Gyarmati, L., and T. A. Trinh. 2010. "Measuring user behavior in online social networks". *IEEE network* 24(5):26–31.

Hajiakhoond Bidoki, N., A. V. Mantzaris, and G. Sukthankar. 2019. "An LSTM model for predicting cross-platform bursts of social media activity". *Information* 10(12):394.

Hochreiter, S., and J. Schmidhuber. 1997, Nov. "Long short-term memory". *Neural Comput* 9(8):1735–1780.

Lamberson, P., and S. Soroka. 2018. "A model of attentiveness to outlying news". *Journal of Communication* 68(5):942–964.

Lanham, M. J., G. P. Morgan, and K. M. Carley. 2013. "Social network modeling and agent-based simulation in support of crisis de-escalation". *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 44(1):103–110.

Lei, K., Y. Liu, S. Zhong, Y. Liu, K. Xu, Y. Shen, and M. Yang. 2018. "Understanding user behavior in Sina Weibo online social network: A community approach". *IEEE Access* 6:13302–13316.

Lewis, R. J. 2000. "An introduction to classification and regression tree (CART) analysis". In *Annual meeting of the society for academic emergency medicine in San Francisco, California*, Volume 14.

Li, M., X. Wang, K. Gao, and S. Zhang. 2017. "A survey on information diffusion in online social networks: Models and methods". *Information* 8(4):118.

Liu, S., S. Yao, D. Liu, H. Shao, Y. Zhao, X. Fu, and T. Abdelzaher. 2019. "A latent hawkes process model for event clustering and temporal dynamics learning with applications in GitHub". In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, 1275–1285. IEEE.

Liu, Y., A. Liu, N. N. Xiong, T. Wang, and W. Gui. 2019. "Content propagation for content-centric networking systems from location-based social networks". *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 49(10):1946–1960.

Madey, G., Y. Gao, V. Freeh, R. Tynan, and C. Hoffman. 2003. “Agent-based modeling and simulation of collaborative social networks”. *AMCIS 2003 Proceedings*:237.

Meng, Y., J. Shen, C. Zhang, and J. Han. 2018. “Weakly-Supervised Neural Text Classification”. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 983–992. ACM.

Meng, Y., Y. Zhang, J. Huang, C. Xiong, H. Ji, C. Zhang, and J. Han. 2020. “Text Classification Using Label Names Only: A Language Model Self-Training Approach”. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

On, C. K., P. M. Pandiyan, S. Yaacob, and A. Saudi. 2006. “Mel-frequency cepstral coefficient analysis in speech recognition”. In *2006 International Conference on Computing & Informatics*, 1–5. IEEE.

Podobnik, B., and H. E. Stanley. 2008, February. “Detrended cross-correlation analysis: a new method for analyzing two nonstationary time series”. *Physical review letters* 100(8):084102.

Railsback, S. F., S. L. Lytinen, and S. K. Jackson. 2006. “Agent-based simulation platforms: Review and development recommendations”. *Simulation* 82(9):609–623.

Rubner, Y., C. Tomasi, and L. J. Guibas. 1998. “A metric for distributions with applications to image databases”. In *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, 59–66. IEEE.

Shao, H., D. Sun, S. Yao, L. Su, Z. Wang, D. Liu, S. Liu, L. Kaplan, and T. Abdelzaher. 2020. “Truth discovery with multi-modal data in social sensing”. *IEEE Transactions on Computers*.

Shao, H., S. Yao, Y. Zhao, L. Su, Z. Wang, D. Liu, S. Liu, L. Kaplan, and T. Abdelzaher. 2019. “Unsupervised fact-finding with multi-modal data in social sensing”. In *2019 22th International Conference on Information Fusion (FUSION)*, 1–8. IEEE.

Shao, H., S. Yao, Y. Zhao, C. Zhang, J. Han, L. Kaplan, L. Su, and T. Abdelzaher. 2018. “A constrained maximum likelihood estimator for unguided social sensing”. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, 2429–2437. IEEE.

Shoemaker, P. J. 1996. “Hardwired for news: Using biological and cultural evolution to explain the surveillance function.”. *Journal of communication*.

Simpson, O., C. Seshadhri, and A. McGregor. 2015. “Catching the head, tail, and everything in between: A streaming algorithm for the degree distribution”. In *2015 IEEE International Conference on Data Mining*, 979–984. IEEE.

Tullo, C., and J. Hurford. 2003. “Modelling Zipfian distributions in language”. In *Proceedings of language evolution and computation workshop/course at ESSLLI*, 62–75.

Varshney, L. R. 2019. “Must surprise trump information?”. *IEEE Technology and Society Magazine* 38(1):81–87.

Xu, W., P. Shi, J. Huang, and F. Liu. 2018. “Understanding and predicting the peak popularity of bursting hashtags”. *Journal of computational science* 28:328–335.

Yao, S., Y. Hao, D. Liu, S. Liu, H. Shao, J. Wu, M. Bamba, T. Abdelzaher, J. Flamino, and B. Szymanski. 2018. “A predictive self-configuring simulator for online media”. In *2018 Winter Simulation Conference (WSC)*, 1262–1273. IEEE.

Zhang, J., L. Tong, P. Lamberson, R. Durazo-Arvizu, A. Luke, and D. Shoham. 2015. “Leveraging social influence to address overweight and obesity using agent-based models: the role of adolescent social networks”. *Social science & medicine* 125:203–213.

Zhou, W. X. 2008, Jun. “Multifractal detrended cross-correlation analysis for two nonstationary signals”. *Phys Rev E Stat Nonlin Soft Matter Phys* 77(6 Pt 2):066211.