



# Experimental evaluation of baselines for forecasting social media timeseries

Kin Wai Ng<sup>1</sup>, Frederick Mubang<sup>1</sup>, Lawrence O. Hall<sup>1</sup>, John Skvoretz<sup>1</sup> and Adriana Iamnitchi<sup>2\*</sup>

\*Correspondence:

[a.iamnitchi@maastrichtuniversity.nl](mailto:a.iamnitchi@maastrichtuniversity.nl)

<sup>2</sup>Maastricht University,  
Minderbroedersberg 4-6, 6211 LK  
Maastricht, The Netherlands  
Full list of author information is  
available at the end of the article

## Abstract

Forecasting social media activity can be of practical use in many scenarios, from understanding trends, such as which topics are likely to engage more users in the coming week, to identifying unusual behavior, such as coordinated information operations or currency manipulation efforts. To evaluate a new approach to forecasting, it is important to have baselines against which to assess performance gains. We experimentally evaluate the performance of four baselines for forecasting activity in several social media datasets that record discussions related to three different geo-political contexts synchronously taking place on two different platforms, Twitter and YouTube. Experiments are done over hourly time periods. Our evaluation identifies the baselines which are most accurate for particular metrics and thus provides guidance for future work in social media modeling.

**Keywords:** Forecasting; Social media time series; Baselines

## 1 Introduction

The importance of baselines for scientific progress cannot be overestimated. In their absence, the evaluation of substantive theoretical models is compromised: exact and accurate fit is rare, if not impossible, and so whether the fit is “good enough” is left to the eye of the beholder. It is much better to have an understanding of how much can be accomplished by relatively simple and widely utilized models of the process of interest and then compare the value added by one’s substantive theoretical model. These simple and widely used models are “baselines.” In certain circumstances, they can themselves be successful at capturing key regularities at relatively low cost in terms of parameters estimated [1].

In a new domain of phenomena, like that of social media timeseries, baselines are even more important. And in such a new domain, it is worth exploring multiple baseline candidates since it is not *a priori* known which baseline does the best job at capturing key regularities. Indeed, it is possible, as we show, that which baseline does best depends on which key regularities one believes it is essential to capture. In the next section, we introduce three different baselines for social media time series: ARIMA [2–6], Hawkes process [7–14], and the Shifted Baseline [6, 15–19]. They have each been widely used in the field of time series analysis.

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Our objective is to use data from multiple social media platforms streams to assess which of these baselines has better predictive power against the empirical data, so called ground truth (GT) data streams for capturing key properties (e.g., volume of activities, temporal patterns, distribution characteristics). Social media platforms function as an *attention economy* [20, 21] in which providers of content (both individuals and organizations) in effect compete for the market share of attention. In general, there is reason to believe that activities in attention economies are amenable to analysis and prediction [22–24].

Via experimental evaluation from six independent datasets (three geo-political contexts over two different social media platforms) and a total of 30 time series of topics in social media, we show that each baseline has advantages and limitations based on the end-task and particular time series characteristics that we aim to capture. Two prediction time periods, daily and weekly, are examined. We observed that, for example, ARIMA better estimates the overall volume of activities over a long forecasting horizon, but it fails to produce accurate temporal patterns/signals present in the GT. When the detection of anomalous or irregular periods of activity is of interest, comparing our predictions against ARIMA could be problematic and might lead to misleading conclusions. For these scenarios, Hawkes processes can be a more appropriate benchmark. Simple baseline approaches that consider shifting past observations as the forecast proved to be highly competitive and a strong benchmark for many scenarios, especially, for next day forecasts. However, when topics are heavily influenced by external variables (e.g., real-world events, economic fluctuations, emergency responses, etc.), the shifted baseline is likely to be less useful over longer prediction windows. Finally, we show that ensemble approaches could provide guidance for better benchmarks.

## 2 Traditional baselines for time series

*What is a competitive baseline against which to compare approaches for predicting time series for social media activity?* This question was motivated by our work on the DARPA SocialSim project with a focus on forecasting social media activity in selected contexts at micro-level granularity [6, 15, 16, 23–27]. In the DARPA challenges, our forecasts were evaluated against what we call below the Shifted Baseline, while reviewers of our work submitted for publication consistently suggested a comparison with ARIMA models. This paper addresses this disjuncture systematically. We selected three well-known time-series generators used in different applications, and applied them to three datasets from two social media platforms each. The three baselines typically used for time series forecasting are briefly presented below. In addition, we combine two of these baselines, ARIMA and Hawkes, into an ensemble that simply averages their outputs in an attempt to obtain a better baseline.

### 2.1 ARIMA

ARIMA is a traditional approach to modeling dynamical systems. This model is derived from a composition of autoregressive and moving average modeling processes, which are commonly represented by  $p$  and  $q$ , with the addition of a differencing term (I) represented by  $d$ . In short, ARIMA assumes that data are linear and follow a particular probability distribution. That is, for any observed variable  $X$ , it can be decomposed into fixed trend, seasonal, and irregular components over time. The output of the ARIMA model can be

expressed as a linear combination of past values of  $X_t$  and lagged forecast errors  $\epsilon_t$ :

$$X_t = \mu + \phi_1 X_{t-1} + \dots + \phi_p X_{t-p} - \theta_1 \epsilon_{t-1} - \dots - \theta_q \epsilon_{t-q}, \quad (1)$$

where  $\phi_1, \dots, \phi_p$  and  $\theta_1, \dots, \theta_q$  are the regression weights to be estimated,  $\mu$  represents the constant trend component, and  $\epsilon_{t-1}, \dots, \epsilon_{t-q}$  are the random errors.

Despite suffering from limitations when presented with non-linear tasks due to its linearity assumption [2], ARIMA is widely accepted and used as a base model in most time series research [3–6].

## 2.2 Hawkes processes

Hawkes processes have been applied in a variety of settings in order to describe or predict univariate or multivariate data. Unlike discrete-time models (in which temporal data is aggregated into uniform time intervals), Hawkes processes rely on discrete events in continuous time (that is, with exact timestamps) to model the likelihood of an event occurring in the future. They are based on the idea that events that are observed in time frequently cluster together. For example, a post published by an influential social media user is likely to be followed by a significant number of interactions from other social media users. Hawkes processes model the inter-arrival timing of these events/interactions by using a conditional intensity function, which can be influenced by past/historical events. Hawkes processes can be defined as follows:

$$\lambda(t) = \mu + \sum_{t_i < t} \varphi(t - t_i), \quad (2)$$

where  $\mu$  is the conditional intensity rate, and  $\varphi$  is a triggering kernel function that takes as input the delay between the current and previous timestamps of events. In this paper, we used the exponential decay kernel function defined by  $\varphi(x) = \alpha\beta \exp(-\beta x)$ , where  $\alpha$  and  $\beta$  regulate the growth and decay rate of events based on observed data.

Hawkes processes were originally defined to describe earthquake dynamics [7], but previous work has used them for forecasting tasks such as estimating price fluctuations in stocks [8] or reproducing conversation dynamics as event sequences [9]. In social media, Hawkes processes have been adopted to model information diffusion and to predict the popularity of content [10–13]. Despite some success, they often struggle to capture characteristic patterns usually observed on social media discussions such as topics with low volume of activities [14].

## 2.3 Shifted baseline

The Shifted (or Persistence) Baseline model is the third common baseline approach we investigate. It is defined as follows. Let  $T$  be the current timestep of interest, and let  $S$  be the length of the desired predicted time series. The Shifted Model predicts the time series for period  $T + 1$  to  $T + S$  by moving forward, or “shifting forward” the time series that spans period  $T - S + 1$  to  $T$ . The underlying assumption of this model is that the immediate future of the time series will approximately replicate its immediate past.

The Shifted baseline has been used in many domains: in social media time series prediction [6, 15, 16], in healthcare time series prediction (patients’ weekly average expenditures on certain pain medications) [17], and in weather time series forecasting [18, 19].

## 2.4 Ensemble Hawkes and ARIMA

The quantitative evaluations presented later suggested a combination of the Hawkes processes approach and the ARIMA approach in an ensemble could be useful. As shown in the next section, Hawkes processes offer better approximations of fluctuations in activity over time, while ARIMA offers better approximations of the amplitude of activity. Therefore, these results suggest an ensemble of the two approaches might prove valuable. Our ensemble is simple: it averages the two source predictions over time, assuming that each approach contributes equally to the final predicted value.

## 3 Datasets

We use data from three different geo-political contexts, each with their own different reactions on social media. The datasets contain a record of user activity on two social media platforms, Twitter and YouTube. The focus on multiple datasets from different contexts and platforms attempts to capture regularities attributable to a baseline rather than features related to data or a social media platform.

### 3.1 Three geo-political contexts

Our datasets include a period of political unrest in Venezuela in early 2019, discussions around the Chinese-Pakistan Economic Corridor, and discussions around the Chinese-East Africa Development Project.

*Venezuelan Political Crisis of 2019 (Vz19)* In 2019, a humanitarian, economic and political crisis engulfed Venezuela as the presidency was disputed between Juan Guaidó and Nicolás Maduro, each claiming to be the country's rightful president. The political turmoil had its roots in the controversial re-election of Nicolás Maduro as the head of the state on January 10th, which was boycotted by opposition politicians and condemned by the international community as fraudulent. Consequently, on January 23rd, the opposition-led National Assembly ratified Juan Guaidó as the country's interim president in an effort to restore democracy and constitutional rights. This event contributed to unprecedented conflicts and high political tension which resulted in nationwide protests, militarized responses, international humanitarian aid intervention attempts, and incidents of mass violence and arrests. Our dataset covers two major events unfolding in the country. First, the declaration of Juan Guaidó as Venezuela's interim president which triggered massive anti-government protests across the country. Second, the arrival of humanitarian aid for the country which was met with violent standoffs between protesters and military forces. Our models are trained on data from the first event and evaluated on data from the second.

*Chinese-Pakistan Economic Corridor (CPEC)* CPEC is a strategic economic project under the Belt and Road Initiative (BRI) launched by China aimed at strengthening and modernizing road, port and energy transportation systems in Pakistan. While China's investment in Pakistan has largely taken the form of infrastructure development, the project has been heavily criticized with claims of lack of transparency, self-benefit and imposing unsustainable debt on the Southeast Asian country [28]. For example, the Indian government firmly stands against CPEC as the project involves traversing parts of the disputed Jammu and Kashmir regions [29]. Conflict around this project plays out in discussions on social media, where it has been reported that both Pakistani and Chinese state actors

strategically promote Chinese interests to facilitate building projects and garner further investment. On the other hand, factions opposed to CPEC perpetuate unfavorable narratives such as the project only being beneficial for China to boost its influence and trading power [30, 31].

*Belt and Road Initiative in East Africa (BRIA)* As part of the same BRI flagship project, China has committed a substantial amount of resources and investment across Africa, especially in the east African nations. Similar to the CPEC scenario, several concerns have arisen over China’s strategic intentions in Africa. Chinese officials claim BRIA is a purely economic construct to enhance international cooperation. However, other parties suggest that the project is a geopolitical tool to boost China’s global strategic influence [32]. Due to the heavy influence of Chinese media in Africa [33], BRIA activities have been promoted extensively on the continent through public pronouncements from ambassadors and government officials. At the same time, many citizens have openly expressed resistance to China’s BRIA as they believe it creates an era of external dependency, internal conflicts, and a rising debt trap [32].

### 3.2 Twitter and YouTube data collection

The datasets used in this work were collected by Leidos and are part of the DARPA-sponsored SocialSim program [34]. Twitter data was collected using the GNIP API and YouTube data was collected through its public API query tool. The collection was based on a list of keywords relevant to each context and was compiled by our SocialSim collaborators. Table 1 presents the search terms for data collection and the 66-day time periods considered in this work, chosen at the beginning of the intervals of data provided. For Twitter, activities consist of tweets, retweets, replies and quotes. For YouTube, activities consist of videos and comments. The resultant social media datasets consist of 13,984,719 Twitter activities done by 1,052,526 users and 116,725 YouTube activities by 70,651 users in Vz19; 1,892,182 Twitter activities by 631,920 users and 133,931 YouTube activities by

**Table 1** Search terms used for data collection and time periods considered

Domain	Period	Search Terms
Vz19	December 24, 2018-February 28, 2019	#23Ene, #23Feb, Aid Venezuela, #BravoPueblo, Caracas, Maracaibo, Chacao, Maturín, #Chavismo, FANB, #FreeVenezuela, #FueraDictadura, Fuerza Venezuela, GNB, #GuaidóPresidente, Juan Guaidó, Nicolás Maduro, #LasCallesSonDelChavismo, Libertad para Venezuela, Freedom for Venezuela, #VamosBien, #MaduroDictator, #SOSVenezuela, #VenezolanosEnElMundo, Venezuela Aid Live, #WeAreMaduro, Yankee go Home, #HandsOffVenezuela, #FebreroRebelde, #NoMasDictatura, #AbajoCadenas, Venezuela Crisis Humanitaria, Maduro Illegítimo
CPEC	March 30, 2020-June 4, 2020	gwadar, china-pak, china pakistan economic corridor, cpec, cathaypak, gwadar port, cphcg, pakchina_friendship, baltistan, baluchistan, balochistan, kashmir, Imran Khan, Belt and Road Initiative, beltandroadpod, OborChina, #BRI, #obor, one belt and road
BRIA	February 1, 2020-April 7, 2020	ctgnafrica, cdafricanews, china belt and road, ukanda mmoja barabara moja, bring kenyans in china home, evacuate ugandans in wuhan, kenya under construction, keep kenya moving, china confucius institute, uhuru kenyatta, Yoweri Museveni, John Magufuli, Paul Kagame, Emmerson Mnangagwa, Edgar Lungu, Karuma Hydropower, Isimba Hydroelectric, Entebbe-Kampala Expressway, Mombasa-Nairobi Standard Gauge Railway, Standard Gauge Railway

73,268 users in CPEC; and lastly 1,046,827 Twitter activities by 472,663 users and 5703 YouTube activities by 4168 users in BRIA.

### 3.3 Topic assignment

As an effort to investigate how activity/information diffuses for different topics of discussion in the social media ecosystem, Leidos employed annotators and subject matter experts (SME) to identify a set of relevant topics relating to each of the three contexts. After a thorough exploration of the datasets corpus, SME crafted a total of 18 topics for Vz19, 12 topics for CPEC and 12 topics for BRIA. For each of the three contexts, the annotators created initial annotation sets comprised of a small subset of tweets and YouTube posts (i.e., videos and comments), which were then labeled with their corresponding topics. The annotations were conducted manually and over a subsample of 11,218 distinct text documents for each of the three contexts. The annotation process consisted of an 8 to 1 ratio of single annotator annotations to all annotator annotations. The 18 topics identified in Vz19 had a weighted average Cohen's Kappa score of 0.64 for inter-annotator agreement. The 12 topics in CPEC had inter-annotator agreement of 0.74 for Cohen's Kappa. Lastly, the 12 topics in BRIA had an average Cohen's Kappa score of 0.65.

Due to the substantial variation in agreement scores among individual topics, we decided to focus this work on the 5 topics with the highest inter-annotator agreement scores from each context. Table 2 presents the topics chosen alongside their agreement scores and a brief description. Because it is not feasible to manually label millions of messages, BERT models [35] were trained and tested for topic annotation. The BERT models for each context were trained on 10,097 distinct text documents and evaluated on a 10% test set (1121 texts). Stratified sampling was used to ensure that the train and test sets had approximately the same percentage of samples of each topic class as in the original manually annotated corpus. The F1 score of the BERT models on Vz19 and CPEC annotation test sets were 0.66 and 0.73, respectively. For BRIA, a XLM-Roberta-Large pre-trained model [36] was used instead to label the texts. Its F1 score was 0.83 on the annotation test set.

After topic assignment and only considering the top 5 topics in inter-annotator agreement (see Table 2), we ended up with 7,986,918 activities by 757,496 users in Twitter and 66,788 YouTube activities by 43,342 for Vz19; 407,621 Twitter activities by 184,961 users and 1072 YouTube activities by 889 users for CPEC; and 905,018 Twitter activities by 437,941 users and 3365 YouTube activities by 2616 users for BRIA. Table 2 also shows the number of activities broken down per topic in both Twitter and YouTube. We observe that both platforms exhibit clear differences in terms of the magnitude of activities. Topic activity on Twitter is orders of magnitude larger than YouTube activity in every context. This is expected as YouTube is mainly used for audiovisual media consumption rather than micro-blogging, where large-scale user interactions are more common. We also find differences across contexts, for instance, the Venezuelan political crisis in 2019 (Vz19) spawned much larger discussions, compared to CPEC and BRIA, on both social media platforms as different external events unfolded in the country.

## 4 Evaluation

Our objective is to assess how well different baselines can capture key properties of time series in the context of per-topic social media activity. To this end, we evaluate the baselines described above against real data and across a set of relevant performance metrics.

**Table 2** Topic Inter-Annotator Agreement scores (IAA) and corresponding descriptions in each context. We include the aggregated volume of activities per topic in both Twitter and YouTube over the periods we consider

Domain	Topic	IAA	Volume		Description
			Twitter	YouTube	
Vz19	maduro/dictator	0.78	1,281,021	25,776	Posts that describe Maduro as a dictator, or when its administration is represented as a dictatorship.
	arrests	0.70	842,546	1944	Describes arrests or people who have been taken prisoners during on-the-ground events.
	other/chávez	0.70	958,450	18,372	Posts with terms related to the late president of Venezuela, Hugo Chávez.
	military	0.69	3,502,074	12,895	Posts about the Venezuelan military, security forces, or other armed militarized organization.
	international/aid	0.68	2,336,919	12,168	Descriptions of foreign humanitarian aid sent or requested to be sent to Venezuela.
CPEC	controversies/china/border	0.91	180,078	260	Discussions related to the disputed Chinese-Indian border.
	leadership/sharif	0.89	82,327	132	Positively discusses leadership by Nawaz Sharif, former prime minister of Pakistan.
	controversies/china/uighur	0.85	25,854	52	Posts that bring attention to the Uighur controversies in Xinjiang China.
	controversies/pakistan/baloch	0.85	84,621	410	Statements that advocate for Baloch independence, or emphasize exploitation of Baloch by Pakistan or China.
	benefits/development/roads	0.83	43,831	222	Discussions on road or railway projects in Pakistan related to CPEC.
BRIA	covid	0.82	774,284	2018	Coronavirus and related topics such as vaccines in Africa.
	debt	0.82	23,803	205	Debts or loans, where the lender is China and the lendee is an African nation
	infrastructure	0.78	28,397	205	Roads, railways, ports or other infrastructure project as part of the BRI in Africa.
	travel	0.78	224,582	815	Travel between countries, usually by flights. Often in the context of COVID and related to Africa and China.
	mistreatment	0.68	58,306	670	Description of mistreatment of particular groups of people such as Kenyans, Africans, or Chinese people.

**Table 3** Time periods chosen for data split into training, validation and testing sets

Domain	Training (52 days)	Validation (1 week)	Testing (1 week)
Vz19	2018-12-24 to 2019-02-14	2019-02-15 to 2019-02-21	2019-02-22 and 2019-02-28
CPEC	2020-03-30 to 2020-05-21	2020-05-22 to 2020-05-28	2020-05-29 and 2020-06-04
BRIA	2020-02-01 to 2020-03-24	2020-03-25 to 2020-03-31	2020-04-01 and 2020-04-07

### 4.1 Experimental setup

We evaluate performance in forecasting social media activity on two platforms, Twitter and YouTube, on a number of topics from three different geo-political contexts. We split the time series samples into training, validation and testing sets for each context as shown in Table 3. The baselines forecast one week of social media activity at hour granularity, thus, 168 datapoints corresponding to the 168 hours in a week.

For Hawkes, we use the sequence of previous events timestamps in order to estimate an intensity function that approximates the likelihood of events happening in the future. The model parameters are: (1) the intensity rate,  $\mu$ , (2) the infectivity factor which represents

**Table 4** ARIMA best parameter combinations over the validation set for each topic on two different platforms

Domain	Topic	(p, q, d)	
		Twitter	YouTube
Vz19	arrests	(72, 1, 48)	(48, 2, 72)
	international/aid	(24, 2, 48)	(72, 1, 96)
	maduro/dictator	(96, 2, 72)	(96, 1, 72)
	military	(72, 1, 72)	(72, 1, 96)
	other/chavez	(48, 2, 72)	(48, 1, 24)
CPEC	benefits/development/roads	(24, 1, 72)	(72, 2, 72)
	controversies/china/border	(24, 2, 72)	(48, 2, 72)
	controversies/china/uighur	(24, 1, 48)	(24, 2, 48)
	controversies/pakistan/baloch	(48, 2, 96)	(48, 2, 72)
	leadership/sharif	(24, 2, 48)	(24, 1, 24)
BRIA	covid	(96, 1, 48)	(72, 2, 96)
	debt	(96, 1, 96)	(48, 1, 96)
	infrastructure	(24, 1, 24)	(72, 1, 96)
	mistreatment	(96, 1, 96)	(48, 1, 48)
	travel	(24, 1, 48)	(24, 2, 72)

how many expected events are caused by a previous event,  $\alpha$ , and (3) the exponential delay which is the probability distribution of time between events,  $\beta$ . These parameters are fit from the training data using an expectation-maximization algorithm that seeks to maximize the log-likelihood function. The output of the fit model results in timestamps of future events, which we aggregate into discrete-time intervals for evaluation (in our case, hour granularity).

For ARIMA, we use the validation set to select a combination of optimized parameters  $(p, q, d)$  for each topic. That is, each hour of predicted activity per topic is dependent on the previous  $p$  hours of activities and the previous  $q$  hours of estimation errors. The differencing factor  $d$  indicates the order of transforming a non-stationary time series into a stationary one.

We ran a grid search over different combinations of  $(p, q, d)$ . At hourly granularity, we considered 24, 48, 72, and 96 previous hours for both  $p$  and  $q$  values. Values of 0, 1 and 2 were considered for  $d$ . Table 4 shows the parameter combinations that obtained the lowest RMSE score over the validation period for each topic.

We report six performance metrics. Here,  $y_t$  indicates the true value of the time series and  $\hat{y}_t$  the forecast series.  $Y_t$  and  $\hat{Y}_t$  represent the aggregated volume of activities for the ground truth and the forecast, respectively:

1. Absolute percentage error (APE), which evaluates the error over the aggregated volume of activities within the forecasting period,

$$APE = \frac{|Y_t - \hat{Y}_t|}{Y_t} \times 100. \tag{3}$$

2. Root mean squared error (RMSE), which measures the differences between predicted and actual values over time,

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2}. \tag{4}$$

- Symmetric mean absolute percentage error (sMAPE), which measures relative errors over time,

$$sMAPE = \frac{1}{n} \sum_{t=1}^n \left( \frac{200 \times |y_t - \hat{y}_t|}{|y_t| + |\hat{y}_t|} \right). \tag{5}$$

- Dynamic time warping (DTW), for measuring the similarity between two time series that may vary in timing. DTW takes as input two time sequences,  $X$  and  $Y$ , of length  $n$ , where  $X = \{x_1, \dots, x_n\}$  and  $Y = \{y_1, \dots, y_n\}$  and computes the matching cost based on the following formulation,

$$D(i, j) = d(x_i, y_j) + \min\{D(i, j - 1), D(i - 1, j), D(i - 1, j - 1)\}, \tag{6}$$

where  $d$  is a distance function (e.g., euclidean distance), and consequently,

$$DTW(X, Y) = \sum_{i=1}^n \sum_{j=1}^n D(i, j). \tag{7}$$

- Volatility error (VE), for measuring the absolute difference between the standard deviation ( $\sigma$ ) of the ground truth and forecasted time series,

$$VE = |\sigma(y_t) - \sigma(\hat{y}_t)|. \tag{8}$$

- Skewness error (SkE), for measuring asymmetry in the volume of activities over time. The SkE metric is measured by computing the adjusted Fisher–Pearson standardized moment coefficient (Eq. (9)) of both the ground truth and forecast time series,

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} \times \frac{\sum_{i=1}^n (x_i - \bar{x})/n}{s^3}, \tag{9}$$

where  $n$  is the length of the time series,  $\bar{x}$  is the sample mean and  $s$  is the sample standard deviation, and then calculating their absolute difference,

$$SkE = |G_1 - \hat{G}_1|, \tag{10}$$

where  $G_1$  and  $\hat{G}_1$  are the standardized moment coefficients for the ground truth and forecast, respectively,

We selected these metrics for evaluation as they measure specific characteristics of time series that we believe are of relevance to social media forecasting solutions. First, we use APE to assess how well a model can capture the total volume of activities per-topic. This property is important for scenarios where the identification of viral/popular topics across social media platforms is of interest. For temporal patterns and volume of activity over time, we use DTW, RMSE and sMAPE measures. The temporal characteristics of time series are valuable for identifying anomalous periods of activity and making informed decisions based on predictions. Finally, we measure skewness and volatility errors, which are important characteristics that can describe social media time series distributions in terms

of their shape and variability. The authors of [37] noted the importance of these 2 particular metrics because of instances in which a model reported strong performance using more typical metrics such as RMSE, but were weak at capturing variability or “spikiness” upon visual representation.

## 4.2 Hardware and runtime costs

All experiments were run on a computer with an Intel Xeon E5-2630 v4 CPU. Each CPU had two sockets, 10 cores, 2 threads per core, and 512 GB of memory. The ARIMA-based models took roughly 12 hours per topic. We ran ARIMA experiments in parallel across 5 computers, and ended up with a total training time duration of 36 hours for all topics. Hawkes processes took on average 2 minutes for each topic and a total run time of 1.5 hours for all topics. The shifted baseline is trivially created by moving historical predictions forward; thus, there is no training phase involved.

## 4.3 Empirical results

In comparing the performance of the baselines against the ground truth, we answer the following questions:

1. Which baselines better capture the volume of activity?
2. Which baselines better capture the temporal patterns?
3. Which baselines better capture the skewness and volatility?
4. How does the duration of the forecast window affect the performance of the baselines?

A high level overview of results is found in Table 5 which presents the performance of baselines for predicting time series of activities across social media topics over different metrics, contexts and platforms at hour granularity. Highlighted in bold are the scores for the baseline that performed best for a metric and topic for a platform. Before we discuss the results in detail, we want to emphasize that they vary significantly by topic and by platform. Thus, we try to identify consistent patterns, but stress the patterns hold in most but not all cases.

We then utilized an aggregate metric, called the *Overall Normalized Metric Error* (ONME) [37] that better illustrates relative performance among models. Tables 6 and 7 contain the results of this metric. Utilizing this new metric, we find that the Shifted Baseline seems to be the strongest performing baseline.

### 4.3.1 Volume of activity

APE and sMAPE are measures for the volume of activity. Figure 1 presents the APE and Fig. 2 the sMAPE per topic for Twitter and YouTube. We make the following observations.

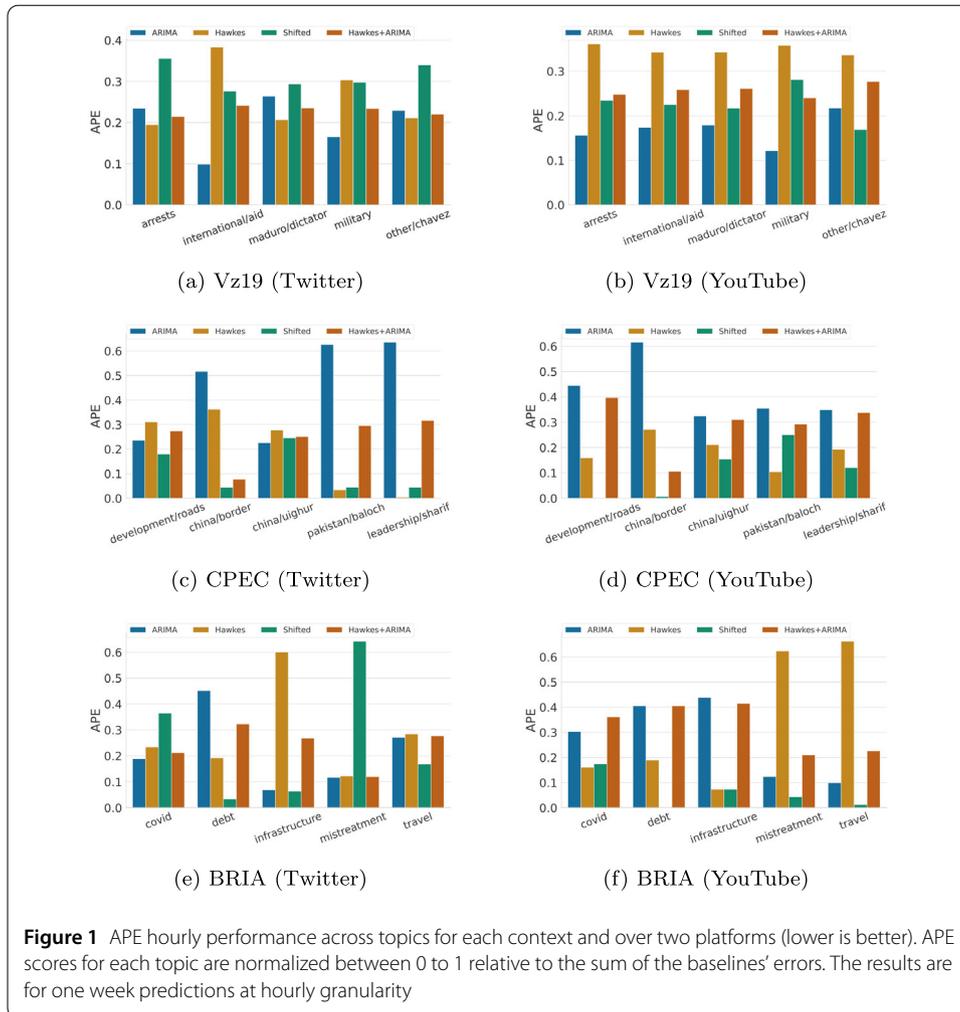
First, there is consistency across topics within the same context, but not across contexts with regards to predicting the total volume of activities. For example, in CPEC, ARIMA performs poorly across most topics on the two platforms (Figs. 1c and 1d), yet in Vz19 it is best at predicting the total volume of activity overall (2 out of 5 topics in Twitter and 4 out of 5 topics in YouTube). Another example is that Hawkes processes perform the worst on all Vz19 topics in YouTube (Fig. 1b). A third example is the shifted baseline which performs the best overall in CPEC (6 out of 10 topics) and BRIA (7 out of 10 topics). One reason for the good performance of the shifted baseline in CPEC/BRIA (or, conversely, its poor performance for Vz19 topics) is the context of our datasets: social media activity

**Table 5** Forecasting performance for total activity predicted across topics in each domain and platform. Baselines were set to forecast one week of activity at hour granularity (168 hours). To compute each metric, we aggregated the predicted activity across all topics and compare with the total number of activities in ground truth. Thus, the performance results are implicitly weighted by the number of activities in each topic. Bold entries are for the best aggregate metric

Domain	Platform	Metric	ARIMA	Hawkes	Shifted	Hawkes + ARIMA	
Vz19	Twitter	APE	<b>35.54</b>	74.61	67.88	55.07	
		RMSE	<b>20,216.11</b>	24,015.25	24,312.27	21,898.09	
		SMAPE	81.25	93.59	98.17	<b>76.38</b>	
		DTW	<b>0.53</b>	0.67	0.83	0.57	
		Skewness	1.37	<b>0.05</b>	1.13	0.90	
		Volatility	17,562.68	17,836.05	<b>15,222.80</b>	18,309.18	
	YouTube	APE	<b>40.70</b>	81.07	52.39	60.88	
		RMSE	<b>128.06</b>	173.08	148.72	148.04	
		SMAPE	<b>45.13</b>	119.75	73.88	66.81	
		DTW	0.55	<b>0.40</b>	0.68	0.49	
		Skewness	2.71	1.65	<b>1.45</b>	2.52	
		Volatility	92.85	104.28	<b>77.50</b>	106.69	
	CPEC	Twitter	APE	270.56	75.97	<b>0.64</b>	97.29
			RMSE	2082.84	<b>767.49</b>	1055.06	1029.92
SMAPE			123.11	108.25	<b>84.60</b>	98.15	
DTW			0.92	<b>0.45</b>	0.74	0.91	
Skewness			3.04	<b>0.29</b>	0.75	3.11	
Volatility			133.45	447.03	266.56	<b>208.28</b>	
YouTube		APE	46.11	61.14	<b>5.70</b>	29.53	
		RMSE	1.54	1.57	2.13	<b>1.52</b>	
		SMAPE	<b>103.58</b>	113.69	116.64	114.56	
		DTW	0.73	<b>0.61</b>	0.63	0.74	
		Skewness	2.75	<b>0.83</b>	1.50	0.93	
		Volatility	0.74	0.47	<b>0.33</b>	0.62	
BRIA		Twitter	APE	<b>33.30</b>	39.34	43.84	36.32
			RMSE	<b>430.54</b>	733.79	731.99	533.80
	SMAPE		60.34	69.60	<b>54.31</b>	61.53	
	DTW		0.48	<b>0.45</b>	0.44	<b>0.45</b>	
	Skewness		3.09	0.13	<b>0.08</b>	0.40	
	Volatility		294.21	142.47	213.30	<b>120.38</b>	
	YouTube	APE	53.99	78.71	<b>10.27</b>	11.03	
		RMSE	2.80	3.42	3.55	<b>2.64</b>	
		SMAPE	127.34	116.49	117.78	<b>113.58</b>	
		DTW	0.87	<b>0.44</b>	0.66	0.49	
		Skewness	3.05	<b>1.43</b>	1.76	1.44	
		Volatility	2.04	0.17	<b>0.14</b>	0.97	

in Vz19 (Venezuela political unrest) is strongly correlated to exogenous activity (such as street protests, power outages, political declarations), thus simply replaying the past does not capture potential reactions to current events. Moreover, the shifted baseline replays the actions from seven days before, while Hawkes processes and ARIMA at least take into account recent history.

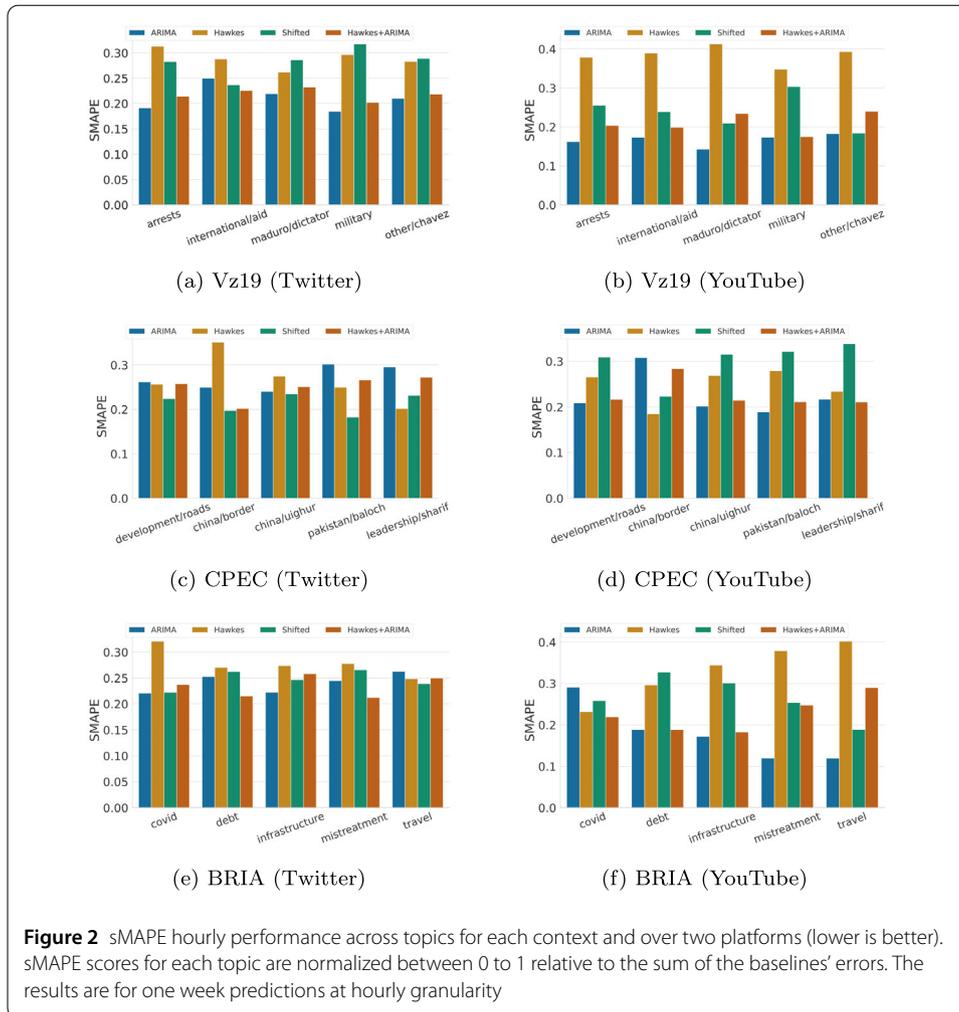
Second, we note that ARIMA performs consistently better than other baselines in sMAPE for 12 out of 15 topics on YouTube (as shown in Figs. 2b, 2d, 2f). For low-activity topics, as is the case for most topics in YouTube, ARIMA can more accurately predict the volume of activities per hour while baselines such as Hawkes or the Shifted baseline often over predict. On the other hand, for Twitter, there is not a persistent winner across most topics using sMAPE. Yet ARIMA is rarely ranked as the worst performing baseline in this metric (only 3 out of 15 topics in total).



### 4.3.2 Temporal patterns

Visual inspection of the time series in Figs. 3 and 4 show that in many instances ARIMA produces a highly unrealistic pattern of variation that diminishes the seeming advantage in performance discussed above. While in some cases ARIMA captures well the periodicity in the ground truth data at hour granularity (as in, for example, Figs. 3a, 3b and 3g for Twitter or Figs. 4a, 4b and 4c for YouTube), it either misses the amplitude of the signal by an order of magnitude (as shown in Fig. 3e), or creates a highly regular (and unrealistic) pattern (Fig. 3a). In other cases, however, it predicts a linear time series that is highly unrealistic, as in Fig. 3d.

Yet despite these visible limitations, traditional time series error metrics such as RMSE show a more favorable picture of ARIMA when compared to both Hawkes and the shifted baselines. For example, in Table 5, which shows the weighted performance over all topics in a context/platform, ARIMA appears to be the best performing baseline on RMSE (5 out of 6 possible instances). This observation also holds when evaluated per topic, where ARIMA outperforms other baselines in 23 out of 30 instances (as shown in Fig. 5). This contradiction emphasizes the importance of model performance evaluation across additional dimensions. For instance, if we desire to accurately capture patterns over time, then metrics that measure similarity between temporal sequences (e.g., DTW) might be

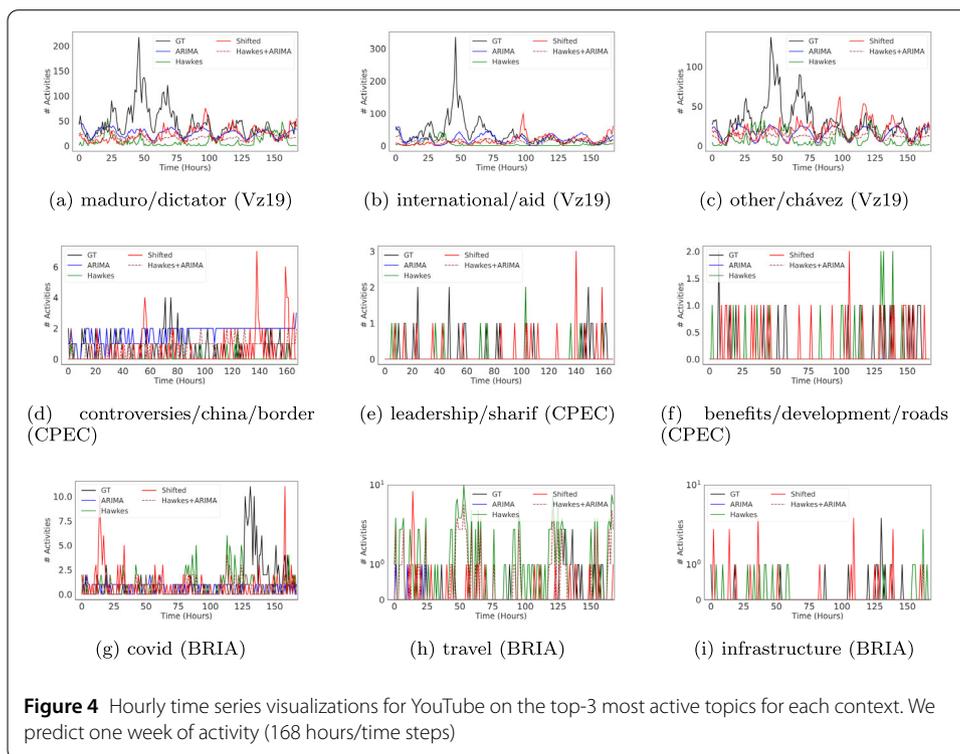
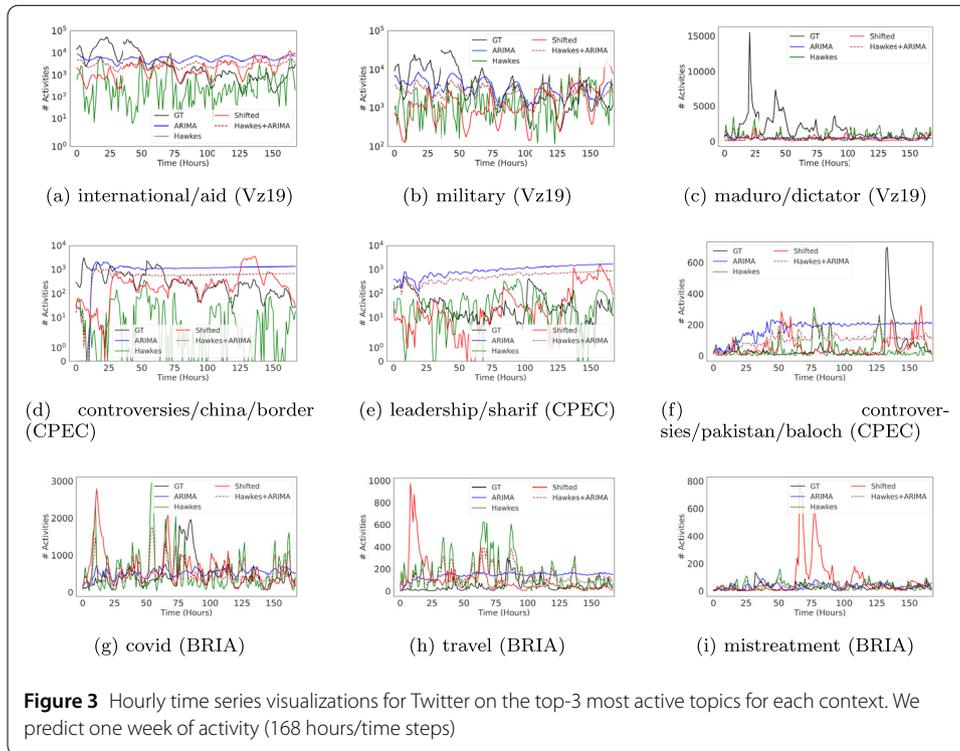


a more appropriate alternative. From Fig. 6, we observe that ARIMA often fails to outperform other baselines, especially Hawkes, with regards to DTW as in many cases it produces unrealistic temporal sequences. This observation excludes ARIMA as a meaningful methodology for creating representative social media time series patterns for scientific purposes.

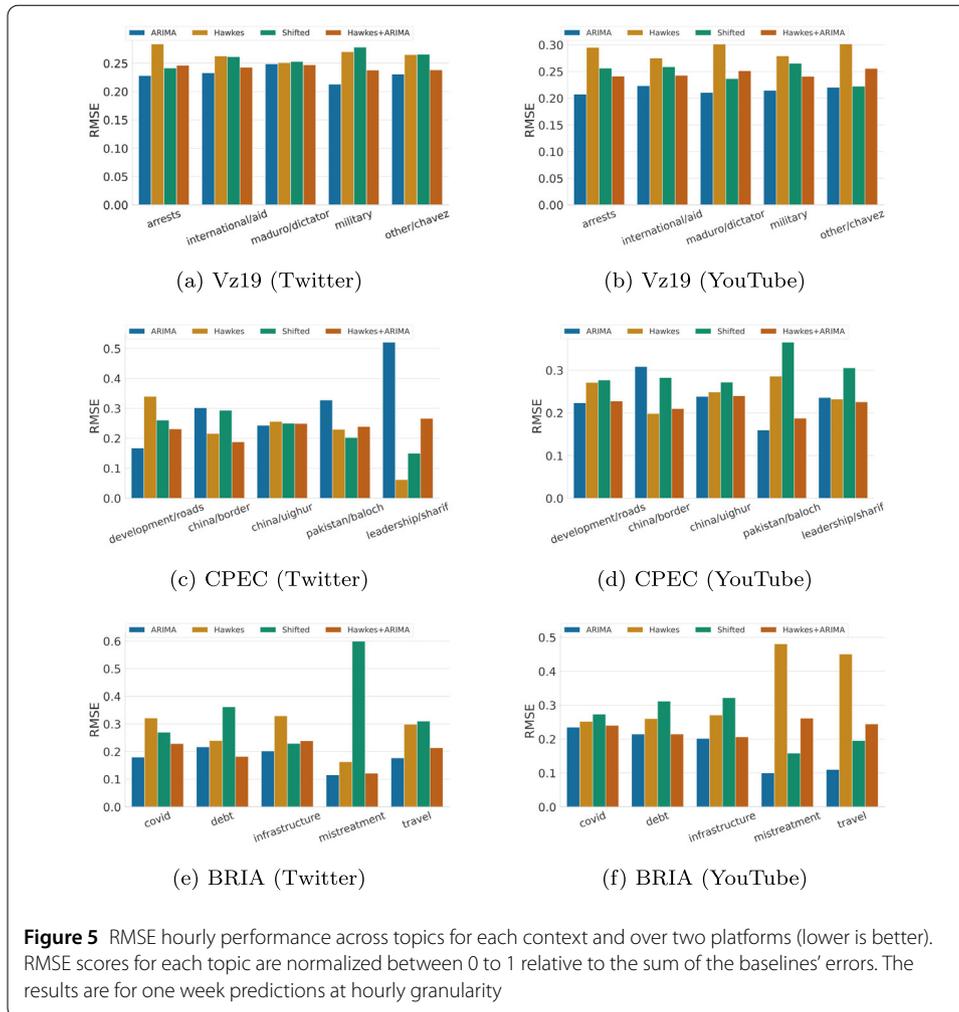
#### 4.3.3 Skewness and volatility

Despite ARIMA's unrealistic predictions, we noticed that traditional time series forecasting metrics fail to penalize it enough. Relying only on these particular metrics to make decisions on which baseline to select can be misleading. In fact, a representative baseline for social media time series is one that not only produces realistic activity levels over time but also captures well key properties related to the distribution of data. Figures 7 and 8 show baselines performance on capturing two basic characteristics of a distribution: spread (in terms of volatility error) and shape (in terms of skewness).

While in some topics ARIMA captures volatility well (e.g., "china/border" in CPEC or "travel" in BRIA as shown in Figs. 7c and 7e), on the majority of topics it is the worst across different contexts and platforms. From time series visualizations, it is evident that ARIMA completely misses the variations observed in the ground truth time series data.



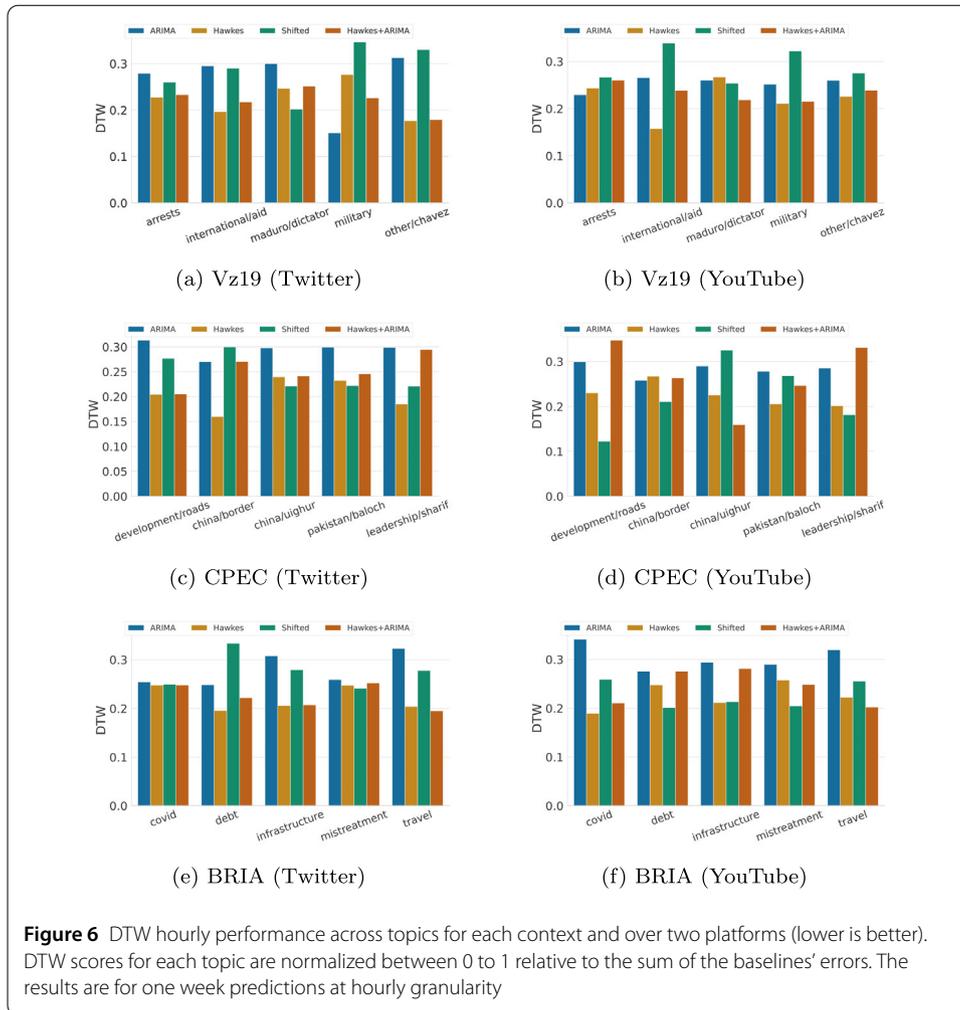
For example, in highly volatile topics such as “controversies/china/border” (Fig. 3d) or “leadership/sharif” (Fig. 3e), ARIMA predicts smooth time series with no signs of volatility. On the other hand, Hawkes and Shifted baselines do a better job at capturing rapid



changes despite sometimes exhibiting irregular up-and-down trends. Similarly, ARIMA does poorly on capturing the shape of the data distribution based on its symmetry (Fig. 8). These patterns become clearer when looking at the aggregated results on volatility and skewness across topics in Table 5, where ARIMA's limitations on distribution-based metrics are more evident. Finally, we might want to select relevant baselines as informed by the utility of time series predictions: should we compare a particular time series against different baselines for different performance metrics? For example, consider ARIMA as a relevant baseline for volume, but another baseline for capturing representative patterns over time. To explore this question, we experimented with an ensemble approach (Hawkes + ARIMA), which takes advantage of both ARIMA's volume predictions and Hawkes' temporal patterns to possibly produce more accurate predictions. Overall the ensemble approach exhibits well-rounded performance across different metrics (Table 5), and proves promising for improving in some topics.

#### 4.3.4 Overall normalized metric error analysis

We sought to have a more clear understanding of which models perform the best, so to that end, we utilized an aggregate metric called the *Overall Normalized Metric Error*, pre-



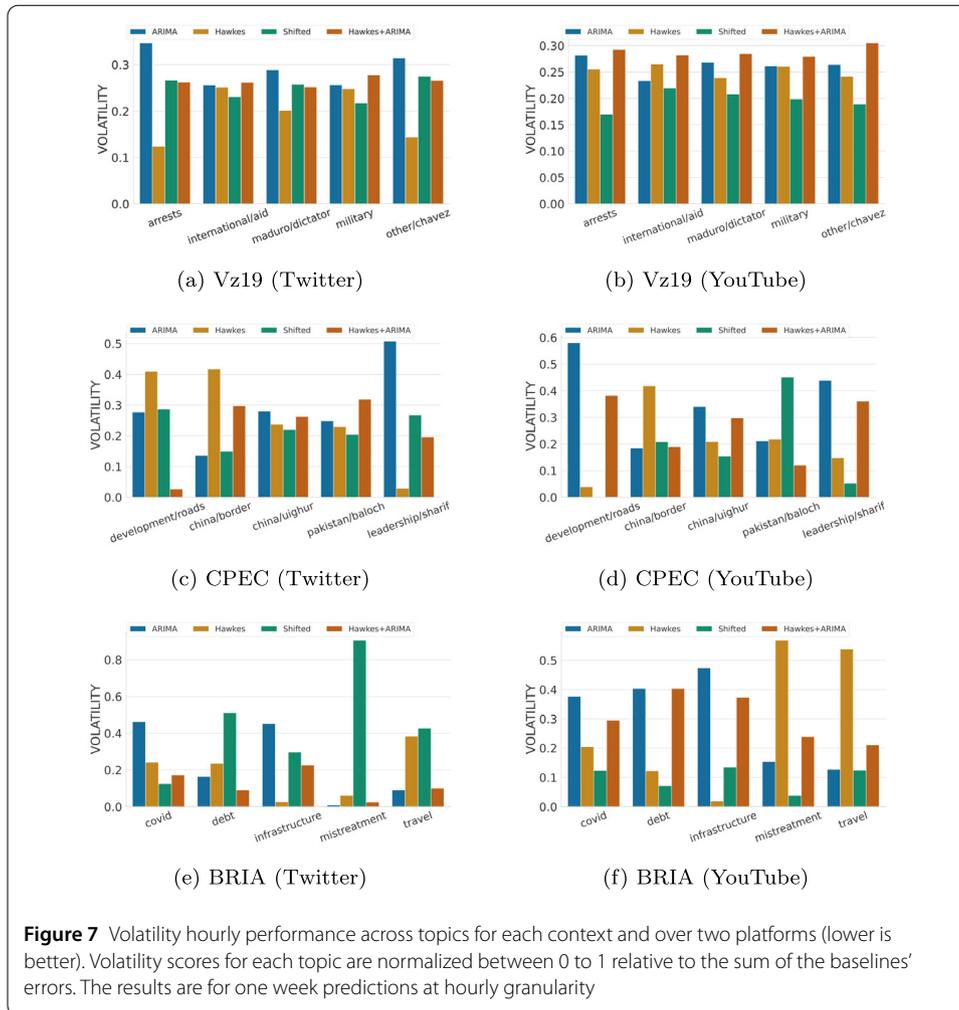
viously used in [37]. It combines the 6 metrics shown in Table 5 into 1 metric per model. Intuitively, it shows how each model performed relative to the other models.

Tables 6 and 7 contain the ONME results for each (*domain, platform*) pairing for each model. Table 6 contains both the ONME results as well as the values (called “normalized errors”) used to calculate them. Table 7 contains only the ONME values from 6 by themselves, for easier viewing and interpretation of results.

For a given (*domain, platform*) pairing, the ONME was calculated in the following way.

1. For each model, we calculated a *normalized error* score. This was done by taking the metric result for a particular model and metric, and then dividing that metric result by the sum of all the other models' metric results for the current (*domain, platform*) pair of interest. We indicate the normalized error for each metric using the prefix “n” (e.g. *nAPE*). These normalized errors are shown in Table 6.

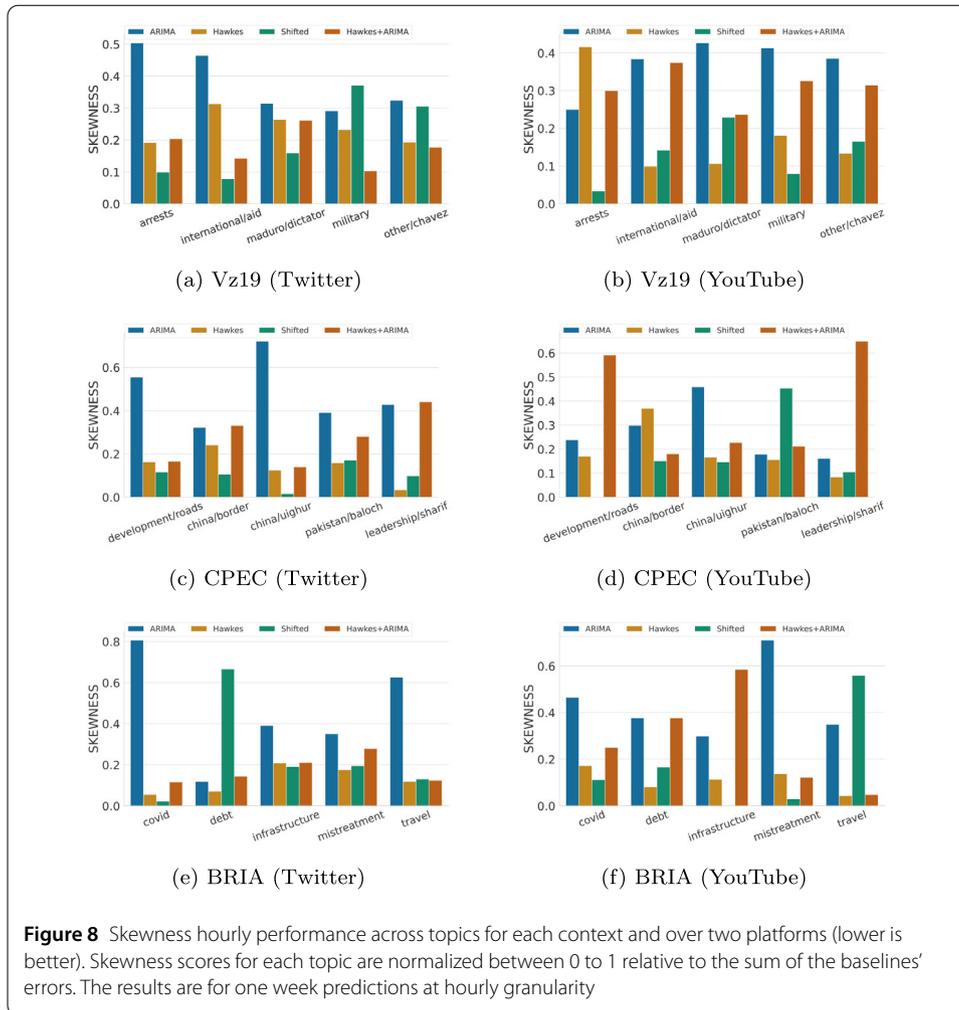
For example, for (*Vz19, Twitter*), we calculated ARIMA's normalized APE score by taking its raw APE from Table 5 (35.54), and then dividing it by the sum of all the models' APE scores for (*Vz19, Twitter*) from Table 5. So, the sum of all APEs would be  $35.54 + 74.61 + 67.88 + 55.07$ , which is 233.1 We then divided 35.54 by this number, which yielded ARIMA's normalized APE score (*nAPE*) for (*Vz19, Twitter*),



which was about 0.1525. We repeated this process for each model and metric. The normalized error results for each model can be seen in Table 6.

2. By repeating this process for each model and metric, this resulted in 6 normalized metric values per model, each between 0 and 1, for a given  $(domain, platform)$  pair. Intuitively, each normalized metric score indicates the relative performance of a model on a particular metric, in comparison to the other models. Note that since each error value is in the range of 0 to 1, the sum of each row in Table 6 adds up to 1.
3. Finally, for each model, we calculated the ONME score by taking the average of the 6 normalized metric error values for that particular model and  $(domain, topic)$  pairing. For example, for the ARIMA model in the  $(Vz19, Twitter)$  category, its nAPE, nRMSE, nSMAPE, nDTW, nSkewness, and nVolatility results were 0.1525, 0.2235, 0.2325, 0.2038, 0.3971, and 0.2548; respectively. The average of these 6 values is 0.244, which is the ONME value shown in Table 6.

As previously mentioned, Table 7 contains only the ONME results from Table 6 by themselves, for easier viewing and interpretation. There is also a new row, that shows the average ONME for each model across all domains and platforms.



As one can see in the table, the Shifted Baseline has the best overall performance. Out of the 6 (*domain, platform*) pairings, it had the lowest ONME three times. In contrast, the Hawkes, ARIMA, and Hawkes + ARIMA models won in this regard only once each.

Furthermore, the Shifted model had the lowest average ONME across all domains and platforms. Its score was 0.2189. In the second, third, and fourth places were the Hawkes + ARIMA, Hawkes, and ARIMA models; respectively. Their ONMEs were 0.2334, 0.2368, and 0.3109; respectively. It is notable that ARIMA performed much worse than the other models in this regard, as it is such a commonly used time series baseline.

#### 4.3.5 Forecast duration vs. baseline performance

One of our objectives is to understand how the baselines perform under different prediction window durations. Towards this, we developed a new set of models that predict just one day of activity at the hourly granularity (thus, 24 values, each corresponding to one hour of the day) to contrast with the 7 day predictions just discussed. We refer to these new models as *Short-Term*, while the previous models are called *Long-Term*.

For the Short-Term configuration, the Shifted model predictions were created by shifting the previous 24 hours of ground truth into the next day. This was done for all seven days in the test period.

**Table 6** Normalized error results for each metric result along with the Overall Normalized Metric Error (ONME)

Domain	Platform	Metric	ARIMA	Hawkes	Shifted	Hawkes + ARIMA
Vz19	Twitter	nAPE	<b>0.1525</b>	0.3201	0.2912	0.2363
		nRMSE	<b>0.2235</b>	0.2655	0.2688	0.2421
		nSMAPE	0.2325	0.2679	0.281	<b>0.2186</b>
		nDTW	<b>0.2038</b>	0.2577	0.3192	0.2192
		nSkewness	0.3971	<b>0.0145</b>	0.3275	0.2609
		nVolatility	0.2548	0.2588	<b>0.2208</b>	0.2656
		ONME	0.244	<b>0.2307</b>	0.2848	0.2405
	YouTube	nAPE	<b>0.1732</b>	0.3449	0.2229	0.259
		nRMSE	<b>0.2142</b>	0.2895	0.2487	0.2476
		nSMAPE	<b>0.1477</b>	0.3919	0.2418	0.2186
		nDTW	0.2594	<b>0.1887</b>	0.3208	0.2311
		nSkewness	0.3253	0.1981	<b>0.1741</b>	0.3025
		nVolatility	0.2435	0.2735	<b>0.2032</b>	0.2798
		ONME	<b>0.2272</b>	0.2811	0.2352	0.2565
CPEC	Twitter	nAPE	0.6087	0.1709	<b>0.0014</b>	0.2189
		nRMSE	0.422	<b>0.1555</b>	0.2138	0.2087
		nSMAPE	0.2973	0.2614	<b>0.2043</b>	0.237
		nDTW	0.3046	<b>0.149</b>	0.245	0.3013
		nSkewness	0.4228	<b>0.0403</b>	0.1043	0.4325
		nVolatility	<b>0.1265</b>	0.4236	0.2526	0.1974
		ONME	0.3637	0.2001	<b>0.1702</b>	0.266
	YouTube	nAPE	0.3236	0.4291	<b>0.04</b>	0.2073
		nRMSE	0.2278	0.2322	0.3151	<b>0.2249</b>
		nSMAPE	<b>0.231</b>	0.2535	0.2601	0.2554
		nDTW	0.2694	<b>0.2251</b>	0.2325	0.2731
		nSkewness	0.4576	<b>0.1381</b>	0.2496	0.1547
		nVolatility	0.3426	0.2176	<b>0.1528</b>	0.287
		ONME	0.3087	0.2493	<b>0.2083</b>	0.2337
BRIA	Twitter	nAPE	<b>0.2179</b>	0.2575	0.2869	0.2377
		nRMSE	<b>0.1772</b>	0.302	0.3012	0.2197
		nSMAPE	0.2455	0.2832	<b>0.221</b>	0.2503
		nDTW	0.2637	<b>0.2473</b>	0.2418	<b>0.2473</b>
		nSkewness	0.8351	0.0351	<b>0.0216</b>	0.1081
		nVolatility	0.3819	0.1849	0.2769	<b>0.1563</b>
		ONME	0.3536	0.2183	0.2249	<b>0.2032</b>
	YouTube	nAPE	0.3506	0.5111	<b>0.0667</b>	0.0716
		nRMSE	0.2256	0.2756	0.2861	<b>0.2127</b>
		nSMAPE	0.268	0.2451	0.2479	<b>0.239</b>
		nDTW	0.3537	<b>0.1789</b>	0.2683	0.1992
		nSkewness	0.3971	<b>0.1862</b>	0.2292	0.1875
		nVolatility	0.6145	0.0512	<b>0.0422</b>	0.2922
		ONME	0.3682	0.2413	<b>0.19</b>	0.2004

**Table 7** Overall Normalized Metric Error results for each domain and platform per model

Domain	Platform	ARIMA	Hawkes	Shifted	Hawkes + ARIMA	Winner
Vz19	Twitter	0.244	<b>0.2307</b>	0.2848	0.2405	Hawkes
	YouTube	<b>0.2272</b>	0.2811	0.2352	0.2565	ARIMA
CPEC	Twitter	0.3637	0.2001	<b>0.1702</b>	0.266	Shifted
	YouTube	0.3087	0.2493	<b>0.2083</b>	0.2337	Shifted
BRIA	Twitter	0.3536	0.2183	0.2249	<b>0.2032</b>	Hawkes + ARIMA
	YouTube	0.3682	0.2413	<b>0.19</b>	0.2004	Shifted
All-Domains	All-Platforms	0.3109	0.2368	<b>0.2189</b>	0.2334	Shifted

**Table 8** The Short-Term ARIMA model best parameter combinations. Similar to the Long-Term configurations, a validation set was used to determine the combinations for each topic and platform

Domain	Topic	(p, d, q)	
		Twitter	YouTube
Vz19	arrests	(24, 1, 24)	(48, 2, 72)
	international/aid	(24, 1, 48)	(72, 1, 24)
	maduro/dictator	(24, 2, 72)	(72, 2, 72)
	military	(24, 1, 24)	(24, 2, 72)
	other/chavez	(48, 2, 72)	(24, 1, 96)
CPEC	benefits/development/roads	(96, 1, 96)	(48, 2, 24)
	controversies/china/border	(96, 1, 96)	(24, 1, 24)
	controversies/china/uighur	(96, 2, 48)	(48, 2, 96)
	controversies/pakistan/baloch	(24, 2, 48)	(24, 1, 24)
	leadership/sharif	(24, 1, 72)	(96, 1, 96)
BRIA	covid	(24, 2, 24)	(48, 1, 72)
	debt	(48, 2, 24)	(72, 1, 48)
	infrastructure	(48, 2, 24)	(48, 2, 96)
	mistreatment	(24, 2, 96)	(24, 2, 24)
	travel	(24, 2, 24)	(48, 2, 24)

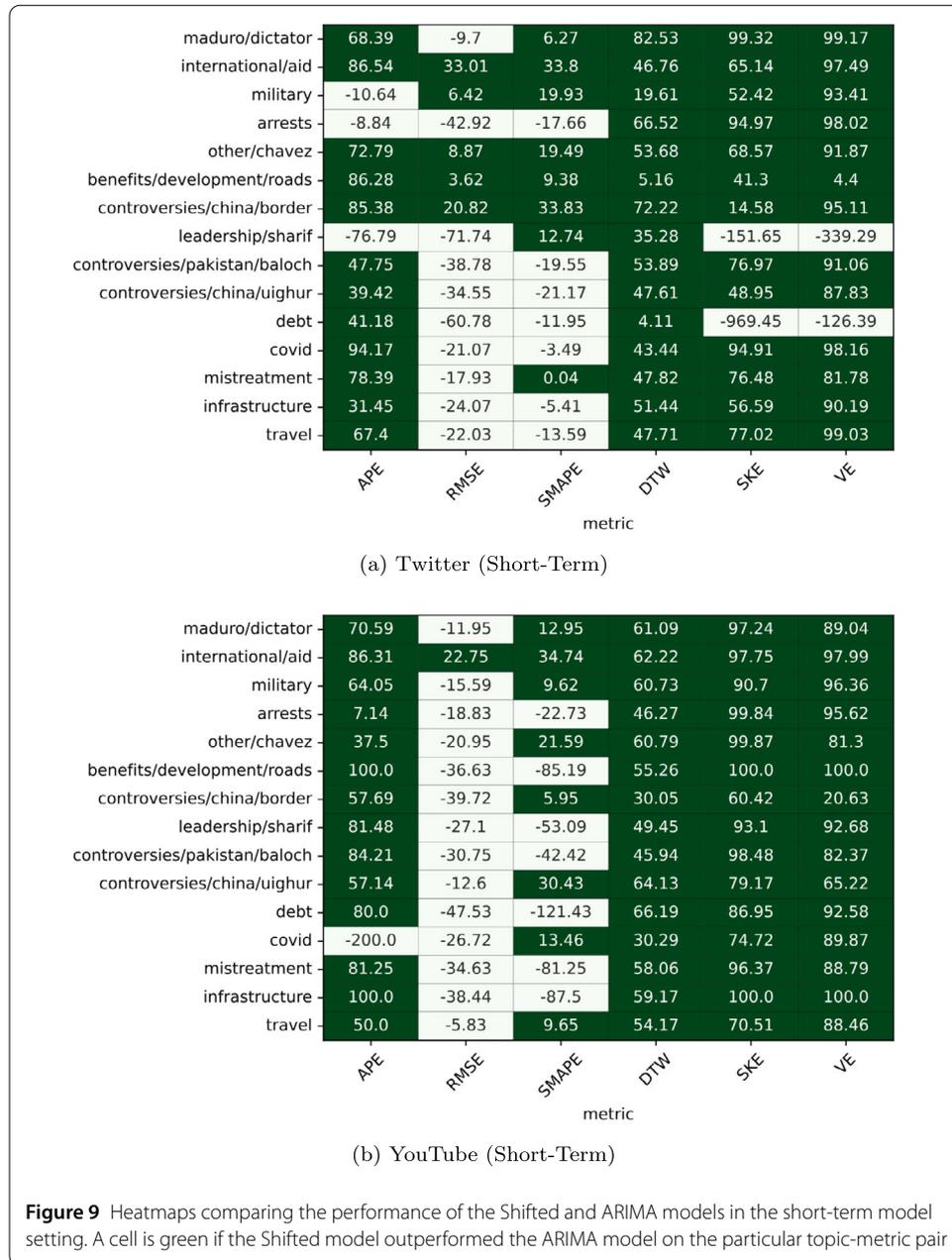
All of the Short-Term ARIMA models predicted 24 hours out into the future at a time, for all 7 test day periods. Their  $p$ ,  $d$ , and  $q$  hyperparameters were determined using a validation set, in a similar fashion to the Long-Term ARIMA model hyperparameters in Table 4. Table 8 contains the Short-Term ARIMA hyperparameter combinations.

As a reminder, the Long-Term models predict the seven days of activity following the training period at hour granularity. In the Long-Term configuration, the Shifted model predictions were created by shifting a full 168-hour block of past ground truth into the 168-hour test period and the ARIMA models predicted 168 hours of values without any ground truth after the start time as input to the model. The predictions of the previous 24-hour blocks were fed into the model as inputs for the next 24-hour block's prediction. We focus on the Shifted model because it had the strongest Overall Normalized Metric Error, and on the ARIMA model because of its high popularity in the literature [2–6].

Note that we use the terms *Short-Term* and *Long-Term* in a relative sense, and not an absolute one. The purpose of the following analysis is to better understand how extending the prediction windows of the Shifted and ARIMA models would affect their relative performance in comparison to one another. The reader may have their own preference for prediction window size, however we just want to demonstrate how extending or shortening it may alter relative model performance.

The Fig. 9 heatmap shows the Short-Term prediction results and the Fig. 10 heatmap shows the Long-Term prediction results of the Shifted and ARIMA models. In these heatmaps, if the Shifted Model outperforms the ARIMA model, a cell is green; otherwise, the cell is white. The values in each cell are the percent improvement scores of the Shifted models over the ARIMA models (thus, the larger the value, the better Shifted performs compared to ARIMA in the metric shown on the x axis and on the topic shown on the y axis).

For short-term (one day) predictions (Fig. 9), the Shifted model outperforms ARIMA on most topics and most metrics on both platforms. For the Short-Term Twitter predictions (Fig. 9a), the Shifted model wins 66 out of 90 trials, or 73.33% of the time. Similarly, for YouTube, the Shifted model wins 68 out of 90, or 75.56% of the time (Fig. 9b). APE is the

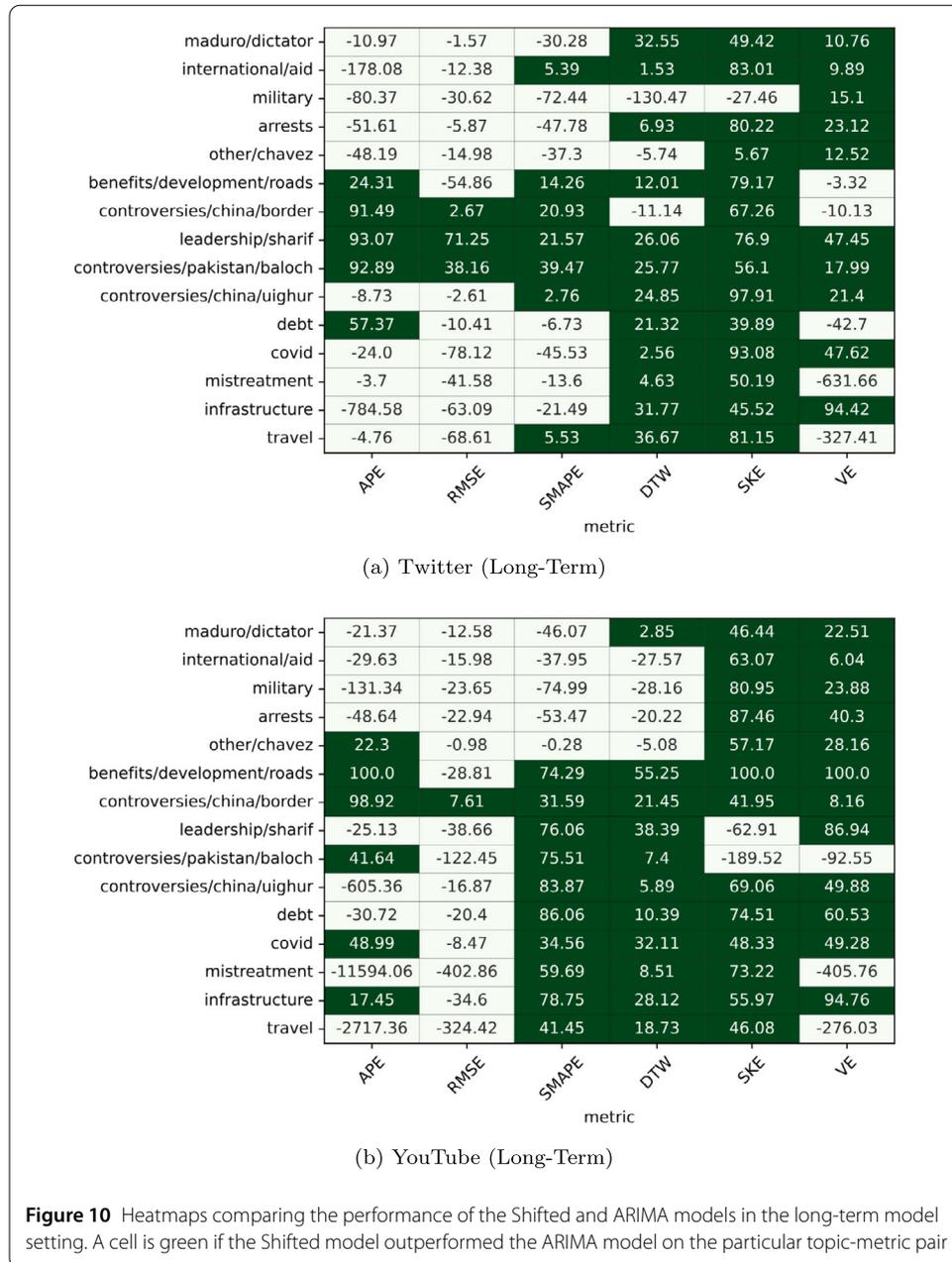


**Figure 9** Heatmaps comparing the performance of the Shifted and ARIMA models in the short-term model setting. A cell is green if the Shifted model outperformed the ARIMA model on the particular topic-metric pair

challenge for ARIMA in this configuration, a result we have not seen in the long-term prediction results presented before.

However, the Shifted model loses its relative advantages over ARIMA in the Long-Term configuration, as shown in Fig. 10. In the Long-Term setup on Twitter data (10a), the Shifted model wins only slightly more than 50%. For YouTube (Fig. 10b), the Shifted model outperforms ARIMA 59% of the time.

Moreover, the advantage that ARIMA gained in the long term predictions are mainly in the volume-related metric (APE), and the exacting-timing-volume metrics (RMSE and sMAPE); Shifted maintains advantages in the approximate-timing-volume metric (dynamic time warping), as well as in the volatility-metrics (Skewness Error and Volatility Error).



**Figure 10** Heatmaps comparing the performance of the Shifted and ARIMA models in the long-term model setting. A cell is green if the Shifted model outperformed the ARIMA model on the particular topic-metric pair

To conclude, for short-term predictions, the Shifted baseline is reliable (and in our experience, very challenging to outperform [38]) while ARIMA misses on most metrics in all contexts and both platforms. For long-term predictions, however, the two models perform similarly when aggregated over all metrics we analyzed, but with clear advantages in the volume-related and exact-timing-volume-related metrics for ARIMA and the approximate-timing-volume and volatility-related metrics for Shifted.

### 5 Summary and recommendations

This study was motivated by the desire to understand and document the relative performance benefits of various time series forecasting methodologies for social media activity. For this, we defined the objective of forecasting the number of social media activities per

hour within a discussion context for the following 168 hours without using the ground truth information during the forecasting interval. This objective is from a typical forecasting problem, and was the task during the scientific challenges which were a part of the DARPA SocialSim project output evaluation. Forecasting social media activity has numerous practical applications, from generating realistic simulations and/or synthetic datasets for scientific experiments to testing intervention techniques meant to control processes on social media.

We chose three representative techniques for forecasting time series: Shifted, that simply replays the past, Hawkes processes, and ARIMA. We used three different conversations from different geo-political contexts, each manifested simultaneously on two social media platforms (Twitter and YouTube), and each comprised of five topics. Based on a combination of manual and automated annotation, each message was annotated with/assigned to one of the topics.

Using these rich datasets, we focused on answering the following question: What is a competitive baseline for forecasting social media activity? We discovered that the answer is nuanced: the Shifted Baseline performed the best in terms of Overall Normalized Metric Error, however it was still not “universally the best”. Rather, a representative baseline should be chosen based on what the predicted time series will be used for. How should a competitive baseline be chosen for a particular scenario? It varies based on 1) the duration of the forecast window; and 2) the metrics that are most important for the forecast task (operational scenario, perhaps).

Specifically, we discovered that ARIMA best estimates the volume of activity when measured in RMSE, but it fails to accurately model temporal patterns. The Shifted baseline works well for short prediction windows in most measurements, but loses power for longer prediction windows. We speculate that Shifted does not match GT when some real-world events drive discussions online in a different manner, or maybe different processes (such as influence campaigns) are taking place on platforms. Hawkes processes create very agitated time series and never dominate in all measurements, possibly because social media activity is not only self-exciting, but external events are causing new excitement on the platform. There can also be non-transparent platform interventions (such as algorithmic promotion of particular messages on user feeds) that may disrupt the learned processes.

Overall, it appears that the Shifted baseline is the winner: replaying the recent past is a cheap and highly reliable estimation of the near future in many scenarios and performance metrics. Where the Shifted baseline will likely fail is when a topic on social media is heavily influenced by real life events, such as major news, political unrest, protests and armed conflicts. In such cases, ARIMA provides a better estimation of the overall volume of social media posts, although it fails at capturing the variation over time.

The choice of baselines depends on the particular characteristics that the problem attempts to capture/model. For example, if modeling the final volume of information cascades, or predicting next day activity (assuming previous day ground truth is always available), then a baseline like ARIMA could be a good choice. However, if the focus is to capture spiky behavior to identify anomalous periods in the future, for instance, then comparing against baselines such as Shifted or Hawkes might be more meaningful.

An ensemble of baselines could result in more competitive models to compare against. For example, when combining Hawkes and ARIMA by averaging, we observed an improvement for some topics in both volume of activity and temporal patterns.

We note that experiments were done with different granularity: in addition to predicting hourly activity, we also predicted daily activity for the same seven-day interval (not included here). Given how the Shifted and Hawkes processes work (that is, they predict the exact time when an event will occur, totally independent of the time granularity chosen), the only difference between day vs hour granularities are in the aggregation of the performance results over time. Thus, comparing these two methods at different time granularities tells more about the performance metrics chosen to evaluate their quality rather than about the methods themselves. ARIMA did not show significant differences in terms of volume prediction. Because many of the other metrics are not normalized, a direct comparison is not meaningful.

A final observation from this study is that different performance metrics tell incomplete and different stories—what is the minimum optimal set of performance metrics for estimating time series forecasting especially in the context of social media activity remains an open question.

Future work would include analysis of burstiness as described in [39]. Also of interest would be an analysis of the features of the time series across domains and platforms. This may yield some better understanding as to why certain models perform better or worse on certain time series. Lastly, we would try incorporating the Shifted model into the ARIMA + Hawkes ensemble model to see if better performance could be obtained.

#### Acknowledgements

The authors would like to thank Leidos for providing data.

#### Funding

This work was supported by the DARPA SocialSim Program and the Air Force Research Laboratory under contract FA8650-18-C-7825.

#### Availability of data and materials

The data that support the findings of this study are available from Leidos but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Leidos.

## Declarations

#### Competing interests

The authors declare that they have no competing interests.

#### Author contributions

All authors contributed equally to the final manuscript. All authors read and approved the final manuscript.

#### Author details

<sup>1</sup>University of South Florida, 4202 E. Fowler Ave, 33620 Tampa, FL, USA. <sup>2</sup>Maastricht University, Minderbroedersberg 4-6, 6211 LK Maastricht, The Netherlands.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 11 May 2022 Accepted: 28 February 2023 Published online: 27 March 2023

#### References

1. Mayhew BH (1984) Baseline models of sociological phenomena. *J Math Soc* 9:259–281
2. Zhang G, Patuwo BE, Hu MY (1998) Forecasting with artificial neural networks: the state of the art. *Int J Forecast* 14(1):35–62
3. Box GE, Jenkins GM, Reinsel GC, Ljung GM (2015). *Time series analysis: forecasting and control*
4. Hipel KW, McLeod AI (1994) Time series modelling of water resources and environmental systems
5. Siami-Namini S, Tavakoli N, Namin AS (2018) A comparison of arima and lstm in forecasting time series. In: 2018 17th IEEE international conference on machine learning and applications (ICMLA). IEEE, pp 1394–1401

6. Shrestha P, Maharjan S, Arendt D, Volkova S (2019) Learning from dynamic user interaction graphs to forecast diverse social behavior. In: Proceedings of the 28th ACM international conference on information and knowledge management. CIKM'19. Association for Computing Machinery, New York, pp 2033–2042. <https://doi.org/10.1145/3357384.3358043>
7. Hawkes AG (1971) Spectra of some self-exciting and mutually exciting point processes. *Biometrika* 58(1):83–90
8. Bacry E, Mastromatteo I, Muzy J-F (2015) Hawkes processes in finance. *Market Microstructure and Liquidity* 1(01)
9. Masuda N, Takaguchi T, Sato N, Yano K (2013) Self-exciting point process modeling of conversation event sequences pp 245–264
10. Rizoju M-A, Xie L, Sanner S, Cebrian M, Yu H, Van Hentenryck P (2017) Expecting to be hip: Hawkes intensity processes for social media popularity. In: Proceedings of the 26th international conference on world wide web, pp 735–744
11. Zhao Q, Erdogdu MA, He HY, Rajaraman A, Leskovec J (2015) Seismic: a self-exciting point process model for predicting tweet popularity. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 1513–1522
12. Lukasiak M, Srijith P, Vu D, Bontcheva K, Zubiaga A, Cohn T (2016) Hawkes processes for continuous time sequence classification: an application to rumour stance classification in Twitter. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: short papers), pp 393–398
13. Pinto JCL, Chahed T, Altman E (2015) Trend detection in social networks using Hawkes processes. In: Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining 2015, pp 1441–1448
14. Bacry E, Bompierre M, Gaïffas S, Muzy J-F (2020) Sparse and low-rank multivariate Hawkes processes. *J Mach Learn Res* 21(50):1–32
15. Hernandez A, Ng KW, Iamnitchi A (2020) Using deep learning for temporal forecasting of user activity on social media: challenges and limitations. In: Proceedings of temporal web analytics workshop, companion proceedings of the 2020 world wide web conference (TempWeb'20), Taipei, Taipei
16. Abdelzaher T, Han J, Hao Y, Jing A, Liu D, Liu S, Nguyen H, Nicol D, Shao H, Wang T, Yao S, Zhang Y, Malik O, Dipple S, Flaminio J, Buchanan F, Cohen S, Korniss G, Szymanski B (2020) Multiscale online media simulation with socialcube. *Computational and Mathematical Organization Theory* 26. <https://doi.org/10.1007/s10588-019-09303-7>
17. Kaushik S, Choudhury A, Sheron PK, Dasgupta N, Natarajan S, Pickett LA, Dutt V (2020) Ai in healthcare: time-series forecasting using statistical, neural, and ensemble architectures. *Front Big Data* 3:4. <https://doi.org/10.3389/fdata.2020.00004>
18. Owens M, Challen R, Methven J, Henley E, Jackson D (2013) A 27 day persistence model of near-earth solar wind conditions: A long lead-time forecast and a benchmark for dynamical models. *Space Weather* 11. <https://doi.org/10.1002/swe.20040>
19. Pérez-Ortiz M, Gutiérrez PA, Tino P, Casanova C, Salcedo-Sanz S (2018) A mixture of experts model for predicting persistent weather patterns. <https://doi.org/10.1109/IJCNN.2018.8489179>
20. Helmer S (2018) May I have your attention, please: -building a dystopian attention economy. In: Companion proceedings of the web conference 2018, pp 1529–1533
21. Huberman BA (2017) Big data and the attention economy: big data (ubiquity symposium). *Ubiquity* 2017:1–7
22. Lakkaraju H, Ajmera J (2011) Attention prediction on social media brand pages. In: Proceedings of the 20th ACM international conference on information and knowledge management, pp 2157–2160
23. Liu R, Mubang F, Hall LO, Horawalavithana S, Iamnitchi A, Skvoretz J (2019) Predicting longitudinal user activity at fine time granularity in online collaborative platforms. In: 2019 IEEE international conference on systems, man and cybernetics (SMC), pp 2535–2542. <https://doi.org/10.1109/SMC.2019.8914586>
24. Ng KW, Horawalavithana S, Iamnitchi A (2022) Social media activity forecasting with exogenous and endogenous signals. *Soc Netw Anal Min* 12(1):1–16
25. Liu R, Mubang F, Hall LO (2020) Simulating temporal user activity on social networks with sequence to sequence neural models. In: 2020 IEEE international conference on systems, man, and cybernetics (SMC), pp 1677–1684. <https://doi.org/10.1109/SMC42975.2020.9283257>
26. Bollenbacher J, Pacheco D, Hui P-M, Ahn Y-Y, Flammini A, Menczer F (2021) On the challenges of predicting microscopic dynamics of online conversations. *Appl Netw Sci* 6(1):1–21
27. Blythe J, Ferrara E, Huang D, Lerman K, Muric G, Sapienza A, Tregubov A, Pacheco D, Bollenbacher J, Flammini A et al (2019) The darpa socialsim challenge: massive multi-agent simulations of the github ecosystem. In: Proceedings of the 18th international conference on autonomous agents and MultiAgent systems, pp 1835–1837
28. Sacks D (2021) The China-Pakistan economic corridor—hard reality greets BRI's signature initiative. <https://www.cfr.org/blog/china-pakistan-economic-corridor-hard-reality-greets-bris-signature-initiative>
29. Aamir A (2020) India's opposition to CPEC on shaky ground. <https://asiatimes.com/2020/04/indias-opposition-to-cpec-on-shaky-ground/>
30. Ahmed ZS, Rojratanakiat S, Taekasem S (2021) The China-Pakistan economic corridor in social media: a critical discourse analysis. [https://uscpublicdiplomacy.org/sites/uscpublicdiplomacy.org/files/CPEC\\_Twitter\\_Discourse\\_Analysis\\_China\\_Pakistan\\_Ahmed\\_Rojratanakiat\\_Taekasem.pdf](https://uscpublicdiplomacy.org/sites/uscpublicdiplomacy.org/files/CPEC_Twitter_Discourse_Analysis_China_Pakistan_Ahmed_Rojratanakiat_Taekasem.pdf)
31. Afzal M (2020) At all costs: How pakistan and china control the narrative on the china-pakistan economic corridor. The Brookings Institution report, 1–10
32. Lokanathan V (2020) China's belt and road initiative: Implications in africa. *Observer Research Foundation Issue* 395
33. Thussu DK, De Burgh H, Shi A (2018) China's media go global
34. DARPA: Computational Simulation of Online Social Behavior (SocialSim). <https://www.darpa.mil/program/computational-simulation-of-online-social-behavior> (2021)
35. Devlin J, Chang M-W, Lee K, Toutanova K (2019) Bert: Pre-training of deep bidirectional transformers for language understanding
36. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V (2019) Unsupervised cross-lingual representation learning at scale. *arXiv preprint*. [arXiv:1911.02116](https://arxiv.org/abs/1911.02116)
37. Mubang F, Hall LO (2022) Vam: an end-to-end simulator for time series regression and temporal link prediction in social media networks. In: *IEEE transactions on computational social systems*, pp 1–12. <https://doi.org/10.1109/TCSS.2022.3180586>

38. Ng KW, Horawalavithana S, Iamnitich A (2021) Forecasting topic activity with exogenous and endogenous information signals in Twitter. In: Proceedings of the 2021 IEEE/ACM international conference on advances in social networks analysis and mining, pp 95–98
39. Karsai M, Jo H-H, Kaski K (2018) Bursty Human Dynamics. <https://doi.org/10.1007/978-3-319-68540-3>

**Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:**

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

---

Submit your next manuscript at ▶ [springeropen.com](https://www.springeropen.com)

---