Terror Attack Prediction Based on Time Series Forecasting of Twitter Data

Ranjit Mishra*

The use of social media such as Twitter by terrorist organisations to spread their propaganda and to recruit new members is well-documented.[1] As per an estimate, there were about 46,000–90,000 Twitter accounts in 2014, which either advocated for Islamic State of Iraq and Syria (ISIS) or were run by supporters of the group.[2] In 2015, Twitter reported that it had banned 1,25,000 ISIS sympathetic accounts.[3] In 2016, it published an update that 3,25,000 accounts had been deleted.[4] Social media empowered ISIS recruiting, thereby helping the group draw at least 30,000 foreign fighters, from over 100 countries, to the battlefields of Syria and Iraq. The social media aided the seeding of new franchises in places ranging from Libya and Afghanistan to Nigeria and Bangladesh. It was the vehicle ISIS used to declare war on infidels worldwide; ISIS carried out deliberately choreographed executions for viral distribution through which it inspired acts of terror in five continents. While the security agencies worldwide have analysed the Twitter data for countering the propaganda and also for running de-radicalisation programmes,[5] its usage for predicting terror attacks is not publicly known. Post-event sentiment analysis is also common on all social media platforms.

A time series is a collection of observations generated sequentially through time. The special features of a time series are that the data are ordered with respect to time and that successive observations are expected

* Shri Ranjit Mishra is an officer of Indian Police Service, Bihar Cadre, 2007 batch, and is currently posted as Deputy Inspector General of Police at Patna.

ISSN 0976-1004 print

Journal of Defence Studies, Vol. 17, No. 1, January–March 2023, pp. 125–135

to be dependent. This dependence from one time period to the next is tested through various statistical measures.

Time series analysis is used for several objectives:

1. to obtain a concise description of the features of a specific time series;
2. to construct a model that explains the time series behaviour (in terms of trend, variance over time or seasonality); and
3. to use the model to forecast the behaviour of the series in the future and to test the accuracy of such predictions by comparisons to occurrence in reality.

Time series forecasting is an important area of Machine Learning (ML) and can be cast as a supervised learning problem. It involves building models inferred from historical data and using them to predict future observations. Machine learning methods such as Regression, Neural Networks, Support Vector Machines, Random Forests and XGBoost can be applied to it.

One of the earliest instances of Time Series Analysis to study terrorism was done by Gabriel Weimann and Hans-Bernd Brosius in 1988, using the Box–Jenkins approach.[6] This approach uses the Long Short-Term Memory (LSTM) network, which is a type of recurrent neural network (RNN) used in deep learning as very large architecture can be successfully trained using it.

The LSTMs are a special kind of RNN, capable of learning long-term dependencies, and are explicitly designed to overcome the long-term dependency problem. Remembering information for long periods of time is practically their default behaviour, not something they struggle to learn. All RNNs have a chain of repeating modules of neural network. In standard RNNs, this repeating module may have a very simple structure, such as a single *tanh* layer. LSTMs also have a chain-like structure, but the repeating module has a different structure. Instead of having a single neural network layer, there are four such layers, which interact in a very special way.

The main idea behind LSTM cells is to learn the important parts of the sequence seen so far and to forget the less important ones. This is achieved by the so-called gates, that is, functions that have different learning objectives such as:

- Illustrate a compact representation of the time series.

- Showcase how to combine new input with the past representation of the series.
- Identify what to forget about the series.
- Identify what to output as a prediction for the next time step.

Designing an optimal LSTM-based model can be a difficult task as it requires careful hyperparameter tuning. The most important parameters are as follows:

- Identifying how many LSTM cells are to be used to represent the sequence. Note that each LSTM cell will focus on specific aspects of the time series processed so far. Having just a few LSTM cells could make the model incapable of capturing the structure of the sequence, while too many LSTM cells might lead to overfitting.
- Designing the exact architecture might require careful fine tuning and several trials.[7]

Dataset

The Kaggle dataset[8] for the ISIS tweets was used for training and prediction purpose. It has 17,410 tweets from 100+ pro-ISIS fanboys from all over the world since the November 2015 Paris Attacks. The dataset includes the following:

Name:
Username:
Description:
Location:
Number of followers at the time the tweet was downloaded:
Number of statuses by the user when the tweet was downloaded:
Date and timestamp of the tweet:
The tweet itself (English translation):

To evaluate the proposed scheme, the dataset was converted into a time series. A list of a thousand stop words was created. Stop words are a set of commonly used words in a language, for example, "a", "the", "is", "are", "and" and "etc.", which do not carry any useful information, and are thus filtered out in Text Mining and Natural Language Processing (NLP). Based on the follower count for each user, the filtered tweets were analysed and a propaganda score was given to the users. The idea was to filter out irrelevant content, find out who is talking to whom and use Social Network Analysis to find out the important players in the

network. For this purpose, only the original tweets were required, so that it could be determined who is having an influence on whom. This, however, does not mean that all re-tweets were removed, rather those were extracted and treated like new tweets from new users. Also, for this purpose, tweets were analysed for actual vs re-tweets to figure out more important characters in the network.

Every re-tweet from a given user was extracted and then re-assigned to the first handle in that text string. For example, "RT @GIIMedia_CH004: Rules Of Imarah Part2 - Conditions of Imamah" created @GIIMedia_CH004 as a new user and assigned everything after his/her handle in the text string as a tweet for that user. A dictionary was created for every user using information they had mentioned (affiliates), every hashtag they had used and every tweet they had sent. In addition, tweets were consolidated for every user in a document and added to their dictionary.

Exploratory Data Analysis

Within the ISIS supporting community, there is a diverse range of actor types, including fighters, propagandists, recruiters, religious scholars and unaffiliated sympathisers. To make meaningful sense of the Twitter data, tweets from serious actors (important members of ISIS) must be identified. Social network analysis (SNA) is the process of investigating social structures by using the networks and graph theory. This technique is used in intelligence, counterintelligence and law enforcement activities, and allows mapping of covert organisations such as an espionage ring, an organised crime family or a street gang. Security agencies worldwide use electronic surveillance programmes to generate the data needed to perform this type of analysis on terrorist cells and other networks deemed relevant to national security. The National Security Agency (NSA) has been performing SNA on call detail records (CDRs), also known as metadata, since September 11 attacks.[9]

Generating the social network from Twitter data broadly involves the following:

Separation of Tweets and Re-tweets

Distribution of 17,000 tweets as original tweet vs re-tweet, is shown in Figure 1.

At the very first step, usernames were scraped from tweets, where the users hadn't mentioned themselves. These usernames were then

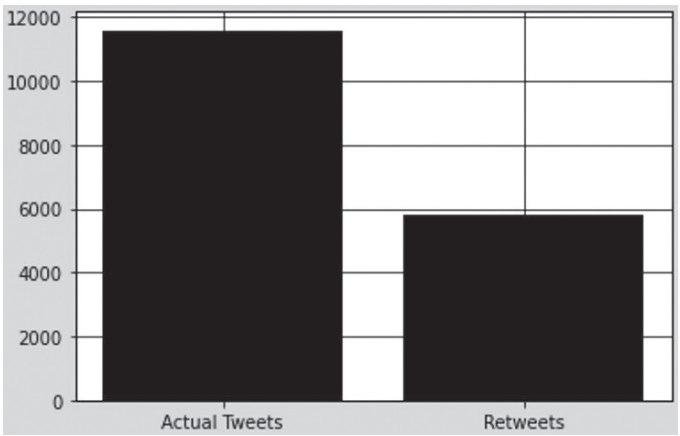

Figure 1 Distribution of tweets and re-tweets in the Kaggle dataset
Source: Author's own.

categorised as users within the dataset or out of it. To map out the most influential (most tweeted) users, it is important to count how many times they have been mentioned. This was done by counting the list. In the Kaggle dataset, the most tweeted user currently is 'Rami'. Originally, 'WarReporter1' was the most tweeted user, but removing tweets where the sender and receiver were the same user has changed the statistics drastically. The tweets where the originator of the tweet was also the recipient after few re-tweets, were filtered out.

Then the users who sent out the maximum number of tweets were separated from those who received the maximum number of tweets (see Table 1).

Table 1 Top 5 Tweet Senders and Receivers in the Kaggle Dataset

Highest Senders		*Highest Receivers*	
MaghrabiArabi	49	RamiAlLolah	53
WarReporter1	30	Nidalgazaui	34
AsimAbuMerjem	27	MilkSheikh2	26
Uncle_SamCoco	27	WarReporter1	15
Moustiklash	20	IshfaqAhmad	15

Source: Author's own.

The top 5 most-tweeted ones are mapped to a graph (see Figure 2).

- Only senders
- Only receivers
- Senders and receivers

Mapping Edges between Nodes

The next step involved mapping edges between nodes where the users had tweeted each other; the edge thickness depends on the number of communications made. The weight depends on the number of tweets sent between two connected users; ideally the increase in weight is represented by a thicker edge. Looking at the graph in Figure 2, it can be inferred that a number of interesting exchanges have happened between users, for instance, between 'AsimAbuMerjem' and 'MilkSheikh2', and 'moustiklash' and 'lshraqAhmad'. Later analysis would try to determine the nature of these exchanges. For the time being, the task of altering the node size to reflect a combination of the number of tweets and followers is being pursued. It is proving tricky to find some mathematical function that will produce large values when the number of tweets and followers grow together.

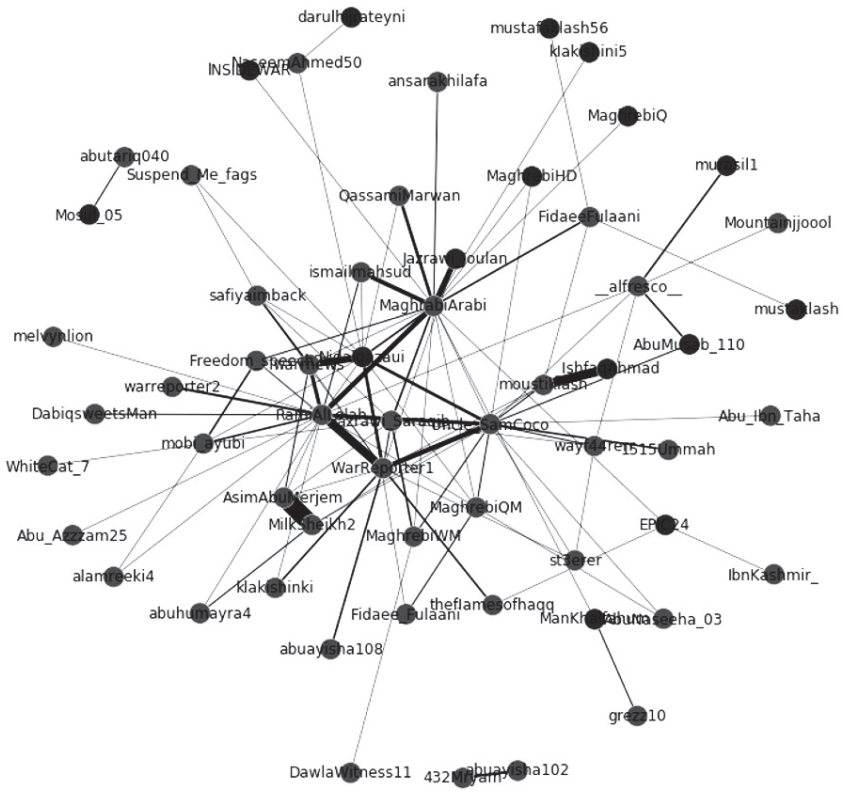

Figure 2 Social Network of the Twitter Dataset with Weighted Edge
Source: Graph generated by the author using NetworkX package in Python

Top 10 Twitter handles and their tweet frequency converted into a time series, are shown in Figures 3 and 4.

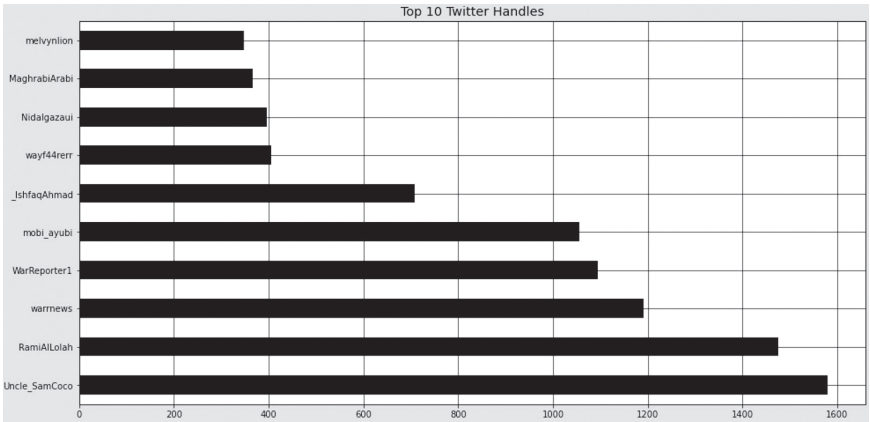

Figure 3 Top 10 Twitter Handles
Source: Author's own.

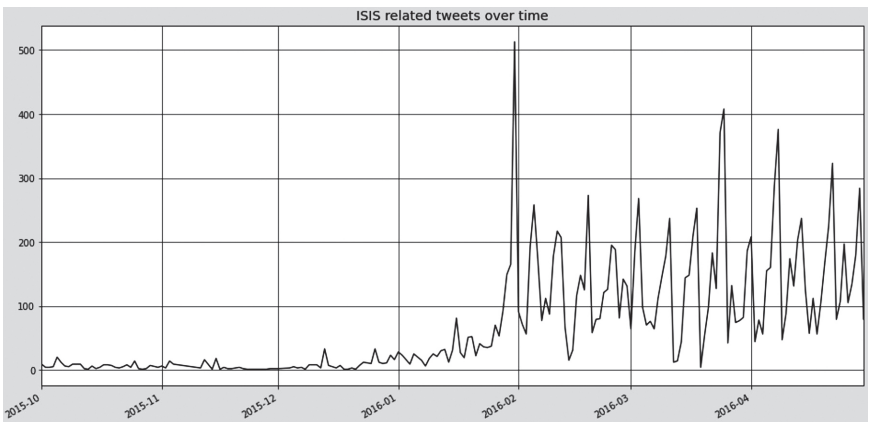

Figure 4 Time Series of tweets
Source: Author's own.

Short-Term Analysis

To show the efficacy of the technique, the Time Series data of April 2016 was taken and mapped with known terror events (Figure 5). One of the pertinent questions is, whether any event concerning terrorism or the ISIS led to a rise in the number of tweets during that time period. Below are some headlines of April 2016 that relate to the number of tweets during this time-frame.

19 April 2016: Taliban send message with deadly Kabul attack as fighting season begins

21 April 2016: Obama and King Salman of Saudi Arabia meet, but deep rifts remain

24 April 2016: U.S. Cyberattacks target ISIS in a new line of combat

27 April 2016: Suicide Bombing near historic mosque in Turkey Wounds 13

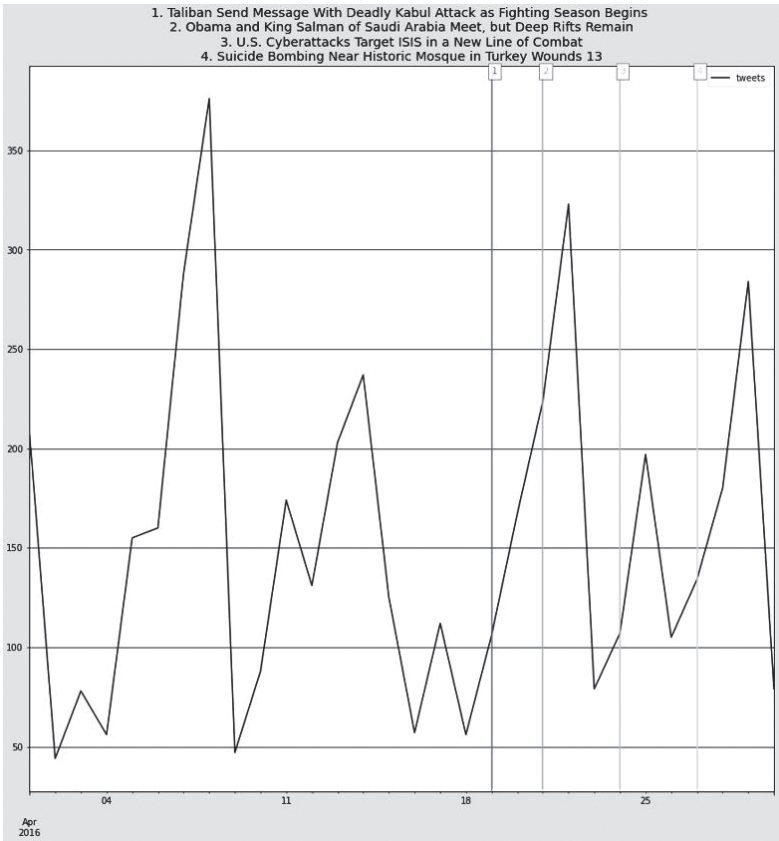

Figure 5 Time Series of April 2016 and Real Reported Terror Incidents
Source: Author's own.

Time Series Forecasting

It can be concluded from the plot shown in Figure 5 that each terror incident is reported when the Twitter plot is ascending, and that there is a trough in the plot between every two incidents. By using the LSTM

model, the plot for the next seven days (in May 2016) was predicted and the accuracy was checked with the available dataset to observe which terror incidents occurred during that period.

A two-layer LSTM network was used for forecasting with Adam optimizer algorithm (Figure 6).

```
# Generate LSTM network
model = Sequential()
model.add(LSTM(4, input_shape=(1, lookback)))
model.add(Dense(1))
model.compile(loss='mse', optimizer='adam')
history=model.fit(X_train, Y_train, validation_split=0.2, epochs=
100, batch_size=1, verbose=2)
```

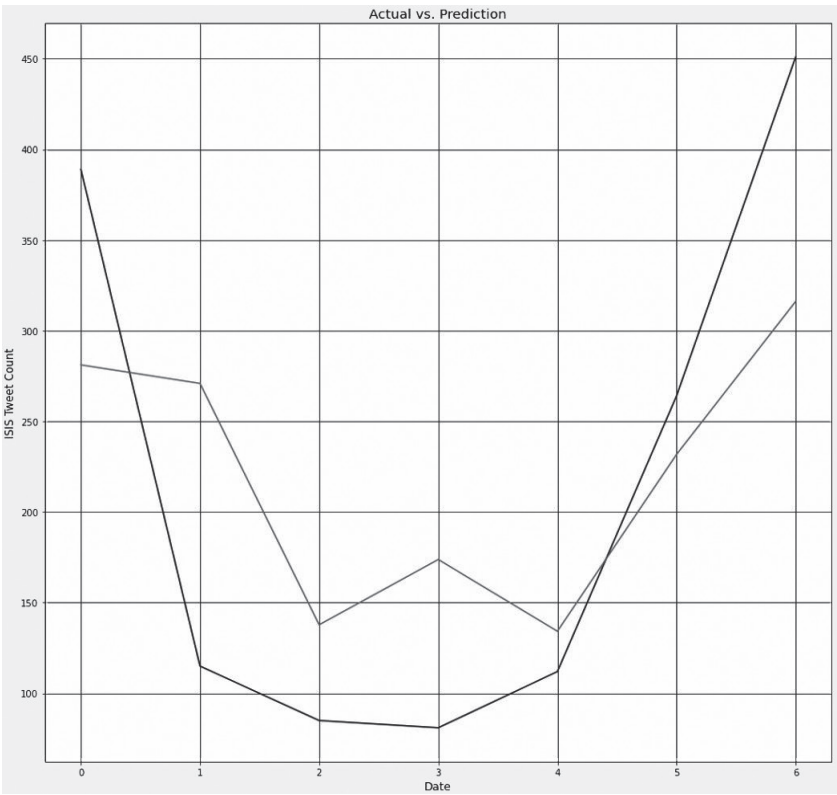

Figure 6 The Light Grey Line is the Actual Data of Tweets While the Dark Grey Line is the Predicted Data Based on the LSTM Network
Source: Author's own.

Analysis

Figure 6 shows a steep rise in the number of pro-ISIS tweets from 3 May to 6 May 2016 based on the time series forecasting. As per the analysis, there must have been at least two ISIS-related terror incidents during this period.

Cross Verification[10]

3 May 2016: A U.S. Navy SEAL was killed in Iraq after ISIS attacked a Peshmerga base. Roughly 100 ISIS fighters broke through a front-line checkpoint and drove 3 to 5 kilometres to the Peshmerga base. The U.S. responded with F-16s and drones that dropped more than 20 bombs.
5 May 2016: ISIS captures the Shaer gas field near Palmyra.

Thus, the prediction of at least two terror attacks in the seven-day period appears to be correct.

Based on this prediction, it can be said that analysis of Twitter data and time series forecasting based on this analysis may be useful in predicting terror-related incidents.

However, the limitation of this approach, as is true with all machine learning models, is that it is very data-intensive. To make a highly dependable prediction model, a huge amount of data across different scenarios will be needed.

Notes

1. Majid Alfifi, Parisa Kaghazgaran, James Caverlee and Fred Morstatter, 'Measuring the Impact of ISIS Social Media Strategy', 2018.
2. J. M Berger and Jonathon Morgan, 'Defining and Describing the Population of ISIS Supporters on Twitter', The Brookings Institution, 5 March 2015.
3. Brendan I. Koerner, 'Why ISIS is Winning the Social Media War', *Wired*, 1 May 2017, available at https://www.wired.com/2016/03/isis-winning-social-media-war-heres-beat/.
4. 'An Update on Our Efforts to Combat Violent Extremism', Twitter Blog, 18 August 2016, available at blog.twitter.com/official/en_us/a/2016/an-update-on-our-efforts-to-combat-violent-extremism.html.
5. Adam Badawy and Emilio Ferrara, 'The Rise of Jihadist Propaganda on Social Networks', *Journal of Computational Social Science*, Vol. 1, No. 8, pp. 1–18, 2018; Matthew Benigni and Kathleen M. Carley, 'From Tweets to Intelligence: Understanding the Islamic Jihad Supporting Community on Twitter', in Kevin S. Xu, David Reitter, Dongwon Lee and Nathaniel Osgood (eds), *Social, Cultural, and Behavioral Modeling*, 9th International

Conference, SBP-BRiMS 2016, Washington, DC, USA, 28 June–1 July 2016, Proceedings, Springer, pp. 346–55; Jonathon M. Berger and Jonathon Morgan, 'The ISIS Twitter Census: Defining and Describing the Population of ISIS Supporters on Twitter', *The Brookings Project on US Relations with the Islamic World 3*, Vol. 20, 2015, 4–1.

6. Gabriel Weimann and Hans-Bernd Brosius, 'The Predictability of International Terrorism: A Time-Series Analysis', *Terrorism*, Vol. 11, No. 6, 1988, pp. 491–502, DOI: 10.1080/10576108808435746.
7. More details about it may be found at https://en.wikipedia.org/wiki/Long_short-term_memory.
8. Available at https://www.kaggle.com/datasets/aliaaied/isis-twitter.
9. See 'Social Network Analysis', available at https://en.wikipedia.org/wiki/Social_network_analysis.
10. From Wikipedia, https://en.wikipedia.org/wiki/Timeline_of_ISIL-related_events_(2016)#May_2016) and https://www.wilsoncenter.org/article/timeline-the-rise-spread-and-fall-the-islamic-state.