Report

Dangerous events forecasting

1. Description of the problem

Forecasting events like terrorist attacks, conflicts, and any other mass violence can help to assess a risk and take a measure to prevent them. Nowadays we have datasets with description of events from around the world, automatically collecting information from the Internet mass media sources. Can we predict the future events with this given information? Let's consider a certain country – France and time period 2013-04-01 – 2016-09-03.

Note: the problem is not new, so you can check:

- http://data.gdeltproject.org/documentation/ISA.2013.GDELT.pdf
- http://foreignpolicy.com/2014/01/03/half-a-billion-clicks-cant-be-wrong/

2. Approach to the problem

The main idea to forecast dangerous events is to build time series from their characteristic – AvgTone of event. In this context, AvgTone is the score ranges from -100 (extremely negative) to +100 (extremely positive). Common values range between -10 and +10, with 0 indicating neutral. This score, calculated automatically, can be used for measuring of the importance of an event. For example, an event like terrorist attack has a AvgTone less than -15. So, let's define a dangerous event as event having:

- AvgTone < -15
- GoldsteinScale = -10

GoldsteinScale is also a numeric score from -10 to +10, capturing the theoretical potential impact that type of event will have on the stability of a country. For more information see:

- http://gdeltproject.org/data/lookups/CAMEO.goldsteinscale.txt
- http://gdeltproject.org/data/lookups/CAMEO.eventcodes.txt

When we have a time sequence of AvgTone, we can apply appropriate methods for forecasting this set of data: ARMA/ARIMA model, neural network or genetic algorithm.

Time sequence has a day as a time unit. However, there are many events during a certain day, so before applying method, AvgTone of events, happened in the same day, should be transformed according the following rules:

- If there are no a dangerous events during a day, tone is average of AvgTones these events.
- If there are dangerous events during a day, tone is equal tone of dangerous event with minimum of tone.

The whole algorithm includes the following steps:

- Loading and merging datasets
- Cleaning and transforming data
- Visualizing data
- Applying neural network

Let's consider these steps with more details.

2.1. Loading and merging datasets

Datasets source is GDELT event files. In this webpage, we can see zipped csv files representing a certain dates. There are about 1200 files for period 2013-04-01 – 2016-09-03. Each file consists of about 160 000 – 170 000 records. Therefore, the total amount of records is about 200 000 000.

The first step is to load datasets on a local machine and during the loading to merge datasets representing the same month. After this step there are 42 files with the total size 68 GB.

Each record in these files describes one event with the following format:

- event represented as «Actor1 performed an action upon Actor2», where Actor can be a certain person, organization, group, country and so on. Note: everything described by codes;
- location of an event;
- tone and other characteristic;
- hyperlink;
- a date an event was added to a database.

For more details about structure of datasets see http://data.gdeltproject.org/documentation/GDELT-Data_Format_Codebook.pdf.

2.2. Cleaning and transforming data

First of all, we need to extract information concerning events in France. After this step we delete duplicates applying the rules:

- delete duplicates with the same URL
- delete duplicates with the same characteristic:
    - FractionDateN'
    - 'QuadClass'
    - 'GoldsteinScale'
    - 'AvgToneN'
    - 'Actor1CountryCode'
    - 'Actor2CountryCode' .

Note: Description of events in given datasets are not always correct because of machine processing.

Finally, we have the following time series of AvgTone of events:

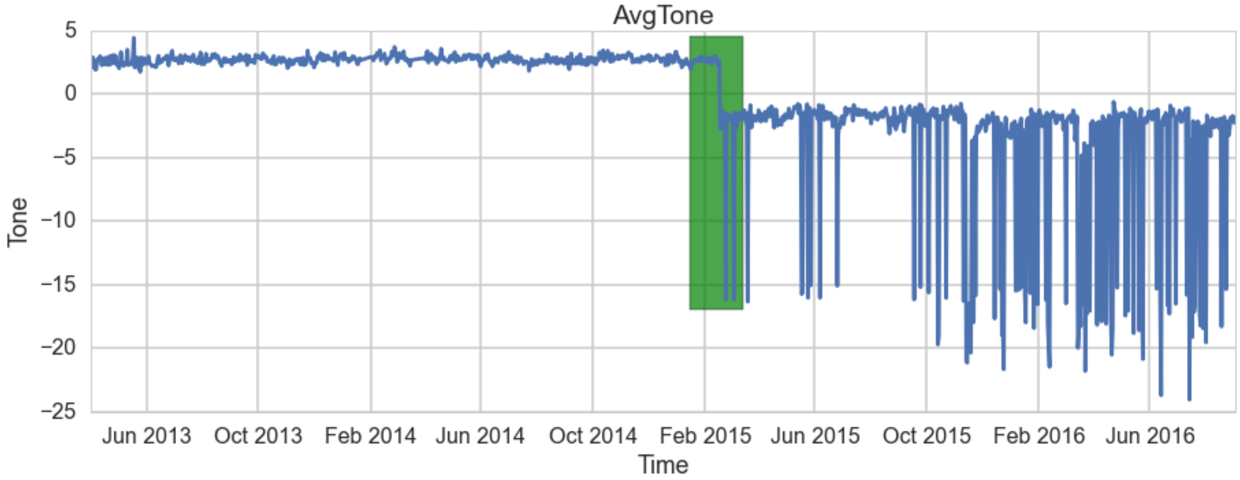

Figure 1

There was a significant change in behavior of AvgTone ('green' area in the picture above), but it's difficult to find reasons for that. Let's have a close look at events in 2016.

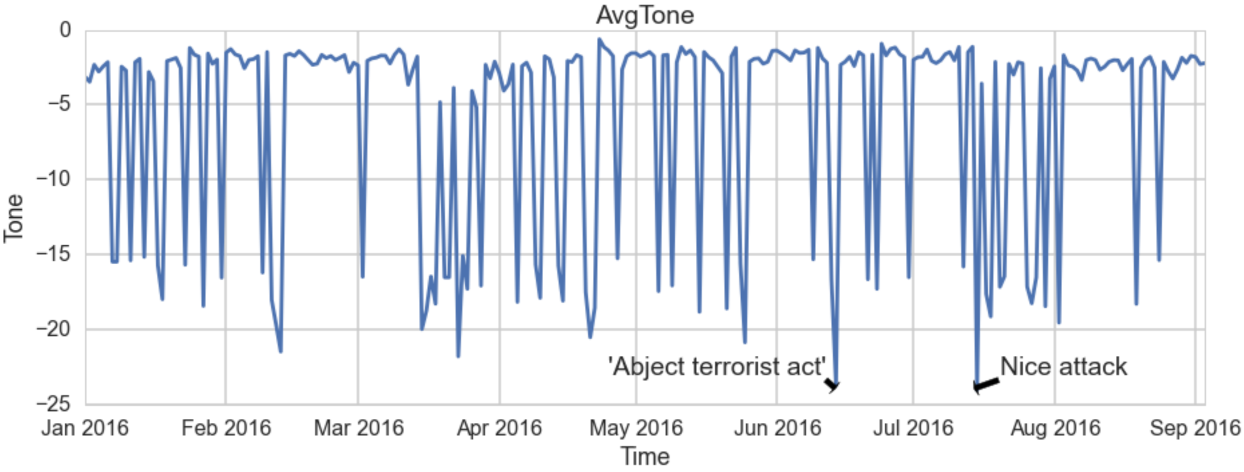

Figure 2

2.3. Applying neural network

Before applying neural network (NN), I tested stationarity of the time series:

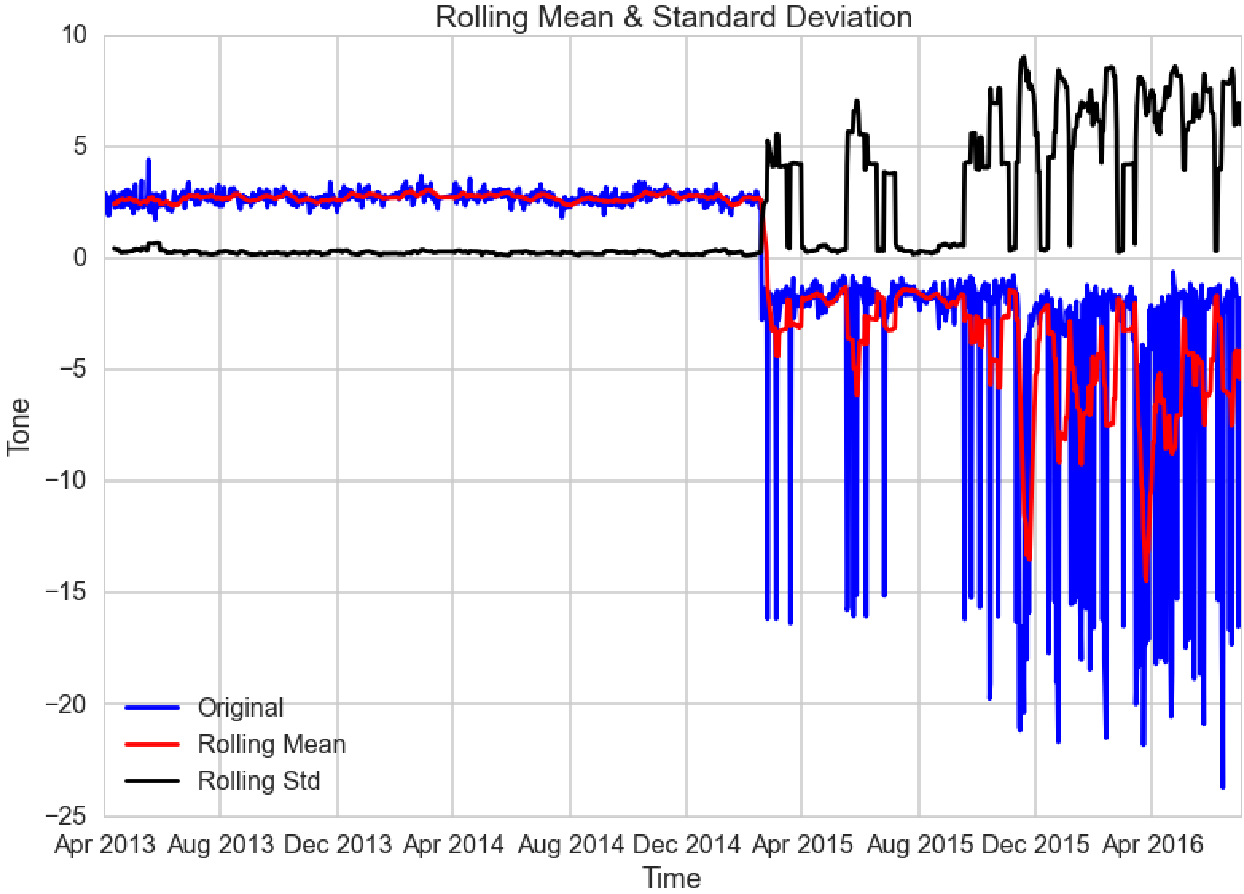

Figure 3

Standard deviation and mean are both changing with time so it's not the stationary time series. In this case, it is better to implement NN instead of ARIMA model.

In this project, I used so-called LSTM network - the Long Short-Term Memory Networks. I used window method. It means that for prediction t+1 value NN used t-lag-1, t-lag-2, …, t values, where lag – Window size.

3. Results

There are results below after implementation of NN (lag = 30).

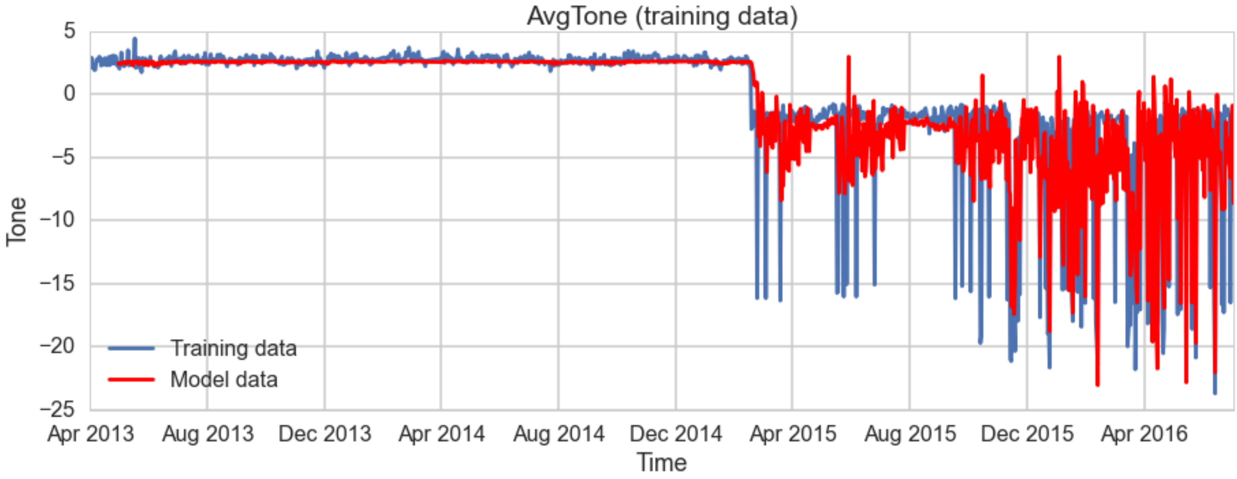

Figure 4

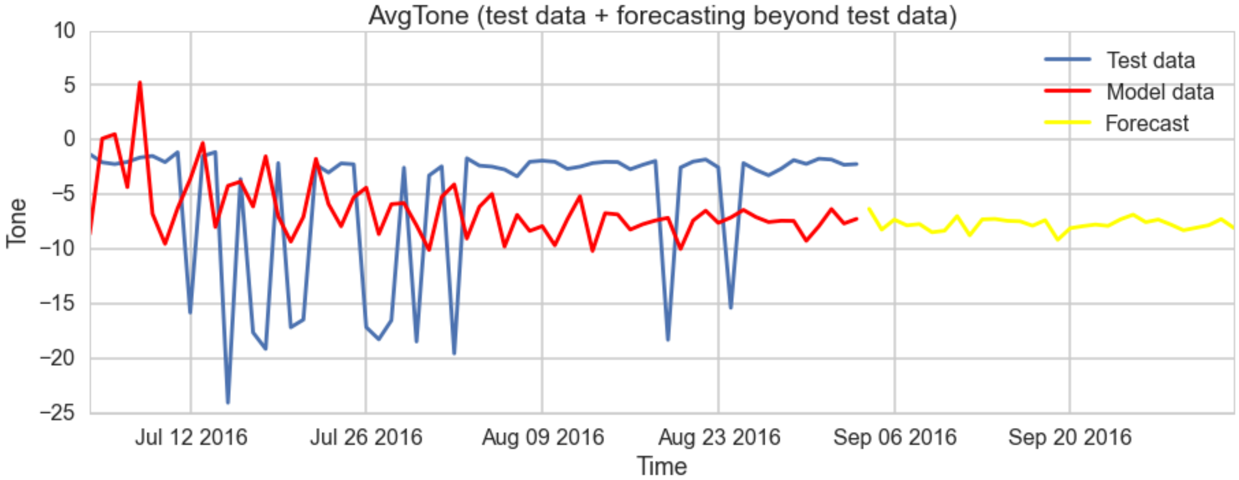

Figure 5

We can see significant differences between test and model data in the last picture. Moreover, forecast line (yellow color) collapses, so approach needs some improvements.

The main reasons for this kind of results are the following:

- Initial data: As I mentioned earlier description of events in given datasets are not always correct. I suppose there is no automatic method to exclude incorrect records.
- Model: current model includes only one variable – a time. It’s reasonable to add more variables, but before dependencies should be calculated.

4. Conclusion

This project represents one of the biggest task in Data Science – prediction continues variable.

Approach with representing events as a time series of AvgTone and implementing NN requires some improvements:

- find appropriate transformation for input data;
- play with parameters of NN;
- use genetic algorithm instead of NN;
- build a model for forecasting based on not only time, but also some other independent variables.

I think the most important step is the last improvement, but finding a more appropriate model is the individual task with its own research. Here, in this project, we can see a common approach and model for forecasting.