A. Training Details

In this section we provide details about training.

To facilitate comparison with previous works, all the datasets are used by Du et al. (2016) and Mei & Eisner (2017), except for 911-Calls and Earthquake. Details about data pre-processing and train-dev-test split, as well as downloadable links, can be found in the aforementioned papers. For the 911-Calls dataset, we exclude zipcodes (and the associated events) whose occurrences are scarce, i.e., we only keep zipcodes that have the top 75 frequent occurrences. The dataset contains 141 types of events, and we cluster them into three categories, namely EMS, fire, and traffic. We do not exclude any events in the Earthquake dataset. Earthquakes are partitioned into two categories, "small" and "large", where small earthquakes are the ones whose Richter scale is equal to or lower than 1.0. We perform this partition because of the imbalance in data, i.e., most of the recorded earthquakes are on small magnitude. Models are trained on 911-Calls and Earthquake with different number of training events. In each experiment, we equally divide the events that are not in the training set in half to construct the development set and the test set.

There are three sets of hyper-parameters that we use throughout the experiments, and they are summarized in Table 7. Besides layer normalization and residual connection, we also employ the dropout technique to avoid overfitting. Table 8 contains the specific parameters that are applied for the training of each dataset. In the table, from left to right columns specify: name of the dataset, the set of applied hyper-parameters, batch size, learning rate, and solver for the integral approximation (MC stands for Monte Carlo integration, and NU stands for numerical integration with the trapezoidal rule), respectively. In the 911-Calls and the Earthquakes datasets, we also employ the graph regularization method, and the corresponding regularization parameter is set to be 0.01 in all the experiments. We use a single NVIDIA RTX graphics card to run all the experiments.

Table 7. Sets of hyper-parameters used in training.

Parameters	# head	# layer	M
Set 1	3	3	64
Set 2	6	6	128
Set 3	4	4	512
Parameters	$M_K = M_V$	M_H	**dropout**
Set 1	16	256	0.1
Set 2	64	2048	0.1
Set 3	512	1024	0.1

Table 8. Hyper-parameters used for training each dataset.

Dataset	set	batch	lr	solver
Retweets	1	16	5×10^{-3}	MC
MemeTrack	1	128	1×10^{-3}	MC
Financial	2	1	1×10^{-4}	NU
MIMIC-II	1	1	1×10^{-4}	NU
StackOverflow	3	4	1×10^{-4}	NU
911-Calls	2	1	1×10^{-5}	MC
Earthquake	3	1	1×10^{-5}	MC