# Recognizing warblers: a probabilistic model for event detection in Twitter

**Joan Capdevila** JC@AC.UPC.EDU
Universitat Politècnica de Catalunya (UPC), Barcelona Supercomputing Center (BSC)

**Jesús Cerquides** CERQUIDE@IIIA.CSIC.ES
Artificial Intelligence Research Institute - Spanish National Research Council (IIIA-CSIC)

**Jordi Torres** TORRES@AC.UPC.EDU
Universitat Politècnica de Catalunya (UPC), Barcelona Supercomputing Center (BSC)

## Abstract

Event detection in Twitter poses a set of new challenges since social networks were not specifically designed for this task. The event identification capabilities of existing probabilistic models are far from state-of-the-art. In this paper we identify three key factors which, when combined, boost the accuracy of such models. Firstly, we notice that the large amount of meaningless social data requires modeling non-event observations. Secondly, we note that the tweeting activity varies in space and time. Thirdly, we observe that the shortness of tweets hampers the application of traditional topic models. Consequently, we propose WARBLE, a new probabilistic model and variational learning algorithm for retrospective event detection that explicitly considers all three factors. The preliminary results show that the proposed model outperforms other state-of-the-art techniques in detecting various types of events while relying on a principled probabilistic framework to reason under uncertainty.

## 1. Introduction

The rise of social networking in mobile devices has turned Twitter into the *de facto* platform to perform event detection from crowdsourced data (Atefeh & Khreich, 2015). In contrast to traditional sensor networks (Wong & Neill, 2009), social networks are not application specific. This enables the identification of various types of events ranging from natural disasters (Sakaki et al., 2010) to geo-social events (Lee & Sumiya, 2010).

*Presented at ICML2016 Anomaly Detection Workshop*, New York, NY, USA, 2016. 

Twitter users, acting as sensors, can ubiquitously describe events through text messages, pictures or even video snippets. This data is usually accompanied with contextual metadata, such as posting time or geographical location, enabling the association of discovered events to real-world occurrences (Zheng, 2012). In this work, we focus on retrospective event detection from geo-located and time-stamped short text messages, such as tweets.

Twitter was not designed with event detection functionalities in mind. Thus, performing this task on the tweet stream poses a set of challenges among which we highlight three, namely rarity, variability and text-shortness. First, we notice that **rare** or anomalous patterns like events are masked by tones of non-event data such as *memes*, user conversations or *retweet* activities (Becker et al., 2011). Second, the tweeting activity **varies** in time and space since distribution of posting times is not uniform along a day (it peaks during late evening) and that of geographical locations is not uniform along space (concentrating mostly in big cities) (Li et al., 2013). Finally, the **shortness** of 140-character-long tweets hampers the application of standard topic models which rely on co-occurrence of words within the same document (Hong & Davison, 2010).

Recently, Capdevila et al. (2015) presented a technique called Tweet-SCAN to retrospectively identify events in Twitter. Tweet-SCAN performs remarkably well in terms of detection accuracy despite not addressing variability. Nonetheless, this technique does not have a proper probabilistic foundation hampering to reason about unseen observations or partially observed data in a principled way. A probabilistic approach to event detection was presented by McInerney & Blei (2014). However, this model fails to explicitly address rarity because it assumes that every tweet is generated by a latent event without distinguishing between event and non-event tweets. Thus, the detection performance of this model is far from that of Tweet-SCAN.

Against this background, we propose WARBLE, a new probabilistic model and a variational inference algorithm that explicitly addresses all three challenges. To show that these three factors, when combined, boost the detection accuracy, we evaluate WARBLE in a real dataset made of tweets in the city of Barcelona during its local festivities.

The rest of the paper is structured as follows. In section 2, we present the related work. In section 3, we introduce the detailed WARBLE model. The scheme to learn the model from tweets is described in section 4. In section 5, we show the detection performance of WARBLE and compare against other state-of-the-art techniques. Finally, we conclude this work in section 6.

## 2. Related Work

Event detection in Twitter has been deeply influenced by Topic Detection and Tracking (TDT) project, which aimed to identify stories in traditional news streams that pertain to new or previously unidentified events (Yang et al., 1998). However, the TDT project assumed that all documents were event-related, an assumption which breaks in the case of Twitter (Atefeh & Khreich, 2015).

DBSCAN-like techniques have been proposed for event detection in Twitter due to their spatial formulation and their noise resilience capabilities (Ester et al., 1996). Tamura & Ichimura (2013) proposed to use the spatio-temporal extension called ST-DBSCAN (Birant & Kut, 2007) to detect specified local events from text filtered tweets. Later, Singh (2015) extended DBSCAN to incorporate search within the textual dimension through cosine similarity over term vectors, enabling to discover various types of unspecified events. Capdevila et al. (2015) presented Tweet-SCAN which relies instead on Jensen-Shannon distance over probabilistic topic distributions (Blei, 2012). This approach enables to mitigate the high-dimensional effects of term vector models, but it conveys new challenges due to the lack of word co-occurrence in short text such as tweets (Hong & Davison, 2010). Because of this, Tweet-SCAN aggregates tweets by *hashtag* prior to event identification to learn topics from these aggregated documents.

A probabilistic model for event detection also using spatio-temporal and textual features was presented by McInerney & Blei (2014). The model was proposed for uncovering newsworthy events from tweets by using of an external news data set, from which topics were also learned prior to event detection. In contrast to DBSCAN-like algorithms, this model does not distinguish between event and non-event tweets making it hard to unequivocally associate the discovered latent events to real-world occurrences. Similarly to DBSCAN-like techniques, this model does not explicitly consider variability along time or space.

## 3. The WARBLE Model

The model proposed by McInerney and Blei presented in (McInerney & Blei, 2014) is a mixture model in which every mixture component shares the same distributional form. Formally, they assume that the $n$-th tweet $\mathbb{T}_n$ is generated according to,

$$\mathbb{T}_n \sim f(\beta_{e_n}) \tag{1}$$

where $f$ is the probability distribution function (pdf), common for all mixture components and $\beta_k$ are the distribution parameters corresponding to the $k$-th mixture component.

As argued in the introduction, a vast majority of tweets is not event-related. Therefore, we would like to address rarity of event data by introducing a new mixture component, to which we will refer as *background*, which contains those tweets which are not part of any event. In probabilistic terms, it seems clear that the distribution of tweets inside the background component should be widely different from that inside events.

Accordingly, the WARBLE model generalizes McInerney and Blei's model to handle heterogeneous components as proposed by Banfield & Raftery (1993). To do that, for each component $k$, we enable a different base function $f_k$ as

$$\mathbb{T}_n \sim f_{c_n}(\beta_{c_n}) \tag{2}$$

where the latent variables are now symbolized as $c_n$ to denote that a tweet might be generated by event component ($c_n < K$) or by background ($c_n = K$).

Moreover, geo-located tweets tends to be unevenly distributed through space and time. For example, it is known that users are more likely to tweet during late evening and from highly populated regions (Li et al., 2013). Consequently, the background component ($c_n = K$) needs to cope with *density varying spatio-temporal distributions*.

In particular, we propose to model the background component through two independent histogram distributions for time and space with parameters $T_B$ and $L_B$, respectively. The temporal histogram distribution is represented through a piecewise-continuous function which takes constant values $(T_{B_1}, T_{B_2}, ...T_{B_{I_T}})$ over the $I_T$ contiguous intervals of length $b$. Similarly, the spatial background is modeled through a 2d-histogram distribution over the geographical space, which is represented in a Cartesian coordinate system. The 2d-piecewise-continuous function is expressed through $I_L$ constant values $(L_{B_1}, L_{B_2}, ...L_{B_{I_L}})$ in a grid of squares with size $b$ x $b$ each.

Fig. 1 shows the complete probabilistic graphical model for the WARBLE model, where tweets $\mathbb{T}_n$ are represented by their temporal $t_n$, spatial $l_n$ and textual $w_{n,\cdot}$ features.

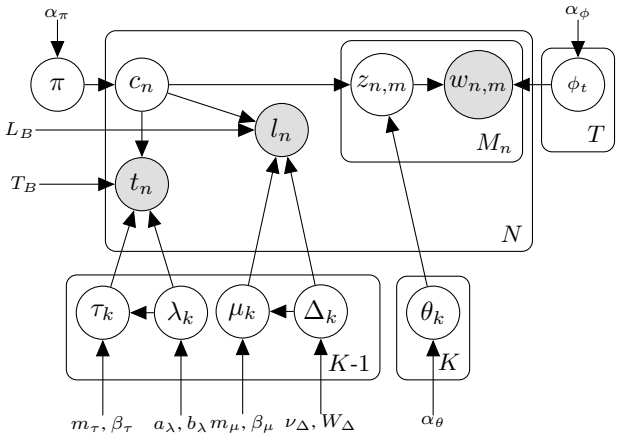


*Figure 1.* The detailed WARBLE model

The event-related components ($k < K$) generate the temporal, spatial and textual features from a Normal distribution with mean $\tau_k$ and precision $\lambda_k$, a Normal distribution with mean $\mu_k$ and precision matrix $\Delta_k$ and a Categorical distribution with proportions $\theta_k$, respectively. Moreover, priors over these distributions are assumed with hyperparameters $m_\tau$, $\beta_\tau$, $a_\lambda$, $b_\lambda$, $m_\mu$, $\beta_\mu$, $\nu_\Delta$, $W_\Delta$ and $\alpha_\theta$.

The background mixture component ($k = K$) accounts for the spatio-temporal features of a tweet, which are drawn from the histogram distributions with parameters ($L_B$) and ($T_B$) introduced earlier. However, the textual feature of the $K$-th component is not ruled by any textual background, but drawn from a Categorical distribution with proportions $\theta_K$ and hyperparameter $\alpha_\theta$.

Finally, we consider $T$ topic distributions over words $\phi = \{\phi_1, \ldots, \phi_T\}$ generated from a Dirichlet distribution with hyperparameter $\alpha_\phi$. The topic distributions $\phi$ are learned simultaneously with component assignments $c_n$ which has lately been found very promising in modeling short and sparse text (Quan et al., 2015) and we refer here as *simultaneous topic-event learning*. In contrast to traditional topic modeling, where distributions over topics are document-specific (Blei et al., 2003), the WARBLE model assumes that topics $z_{n,m}$ are drawn from component-specific distributions $\theta_k$. This enables to directly obtain topics that are event-related or background-related, providing also an interesting approach for automatic event summarization.

## 4. Learning from Tweets

In this section we describe how we can learn the WARBLE model from tweets to identify a set of events in a region during a period of interest.

### 4.1. Learning the background model

To learn the spatio-temporal background, we propose to collect geo-located tweets previous to the period of interest in order to add a sense of typicality to the model.

From the collected tweets, the temporal background is built by first computing the daily histogram with $I_T$ bins. Then, the daily histogram is smoothed by means of a low pass Fourier filter in order to remove high frequency components. The cut-off frequency $f_c$ determines the smoothness of the resulting signal. The normalized and smoothed histogram provides the parameters for the temporal background $T_{B_1}, T_{B_2}, ...T_{B_{I_T}}$.

The spatial background is built following the same procedure. However, geographical location has to be first projected into a Cartesian coordinate system in order to consider locations in a 2-d Euclidean space. The spatial range limits can be determined from the most southwestern and northeastern points. We consider now a two dimensional Gaussian filter with a given variance $\sigma$. The resulting 2d-histogram provides the parameter for the spatial background $L_{B_1}, L_{B_2}, ...L_{B_{I_L}}$.

We suggest to set the number of bins for the temporal and spatial histograms as well as the cut-off frequency and variance empirically. Future work will examine how to automatically adjust these parameters.

### 4.2. Assigning tweets to mixture components

To assign tweets to mixture components, we need to find the the most probable assignment of tweets to mixture components, given the data at hand. That is finding $c^*$,

$$c^* = \underset{c}{\operatorname{argmax}}\, p(c|l, t, w; \Gamma) \tag{3}$$

where $\Gamma$ stands for the model hyperparameters $L_B$, $T_B$, $\alpha_\pi$, $\alpha_\theta$, $\alpha_\phi$, $m_\tau$, $\beta_\tau$, $a_\lambda$, $b_\lambda$, $m_\mu$, $\beta_\mu$, $\nu_\Delta$ and $W_\Delta$. Exactly assessing $c^*$ is computationally intractable for the WARBLE model.

Therefore, we propose to first use mean-field variational Bayesian inference (Fox & Roberts, 2012; Jordan et al., 1999) to approximate $p(X|D;\Gamma)$ (where $X$ stands for the set of random variables containing $c$, $z$, $\pi$, $\tau$, $\lambda$, $\mu$, $\Delta$, $\theta$ and $\phi$, and $D$ stands for our data, namely $l$, $t$, and $w$) by a distribution $q(X;\eta)$ (where $\eta$ stands for the variational parameters). Then, assess $c^*$ from the approximation, that is

$$c^* = \underset{c}{\operatorname{argmax}}\, q(c;\eta) = \underset{c}{\operatorname{argmax}} \int_{X-c} q(X;\eta). \tag{4}$$

The functional forms for the mean-field approximation $q(X;\eta)$ and the updates for the variational parameters can be found in a technical report (Capdevila et al., 2016).

## 5. Experiments

In this section, we present the experimental setup and the comparative evaluation of WARBLE against other state-of-the-art techniques.

### 5.1. Experimental setup

To evaluate our method we rely on a set of 2173 geo-located tweets that were collected from the Twitter streaming API in the city of Barcelona during its local festivities on the 24th of September 2014, referred as "La Mercè 2014". Within this dataset, 202 tweets were associated with 7 real-world events by local experts helped with the official calendar of the festivities.

In addition to "La Mercè 2014" dataset, we consider tweets previous to the period of interest in order to learn the spatio-temporal backgrounds $L_B$ and $T_B$ as explained in Section 4.1. In particular, we collected tweets from the 20th to the 23th of September 2014 to learn the histogram distributions for the spatio-temporal features.

The WARBLE model contains several parameters and hyperparameters. Although their optimization is out of the scope for this paper, we have not experimented substantial differences in the results when varying them. The number of components $K$ is set to 8 so that the model is able to capture the 7 events occurring. Following (Capdevila et al., 2015), we set the number of topics $T$ to 30.

### 5.2. Evaluation against state-of-the-art

In what follows, we compare WARBLE against other event detection techniques in terms of F-measure. F-measure, which is defined as the harmonic mean of purity and inverse purity, has been traditionally used in event detection to compare the resulting clusters against a baseline (Yang et al., 1998).

In particular, we compare the performance of (A) McInerney & Blei model, (B) the WARBLE model without simultaneous topic-event learning, (C) the WARBLE model without modeling background, (D) the complete WARBLE model and (E) Tweet-SCAN.

For models (A), (B) and (E), which do not perform simultaneous topic-event learning, the Latent Dirichlet Allocation model (Blei et al., 2003) is first separately trained with tweets aggregated by key terms as proposed in (Hong & Davison, 2010) and then topics are transferred to the model.

Fig. 2 shows the results for each event detection model introduced earlier in terms of set matching metrics. Results show that WARBLE outperforms the existing state-of-the-art models (A) and (E) in terms of F-measure and purity. Moreover, by analyzing the results of models (B) and (C) we see a clear synergy between background modeling and simultaneous topic-event learning. Neither of them separately achieves a large increase of the F-measure, but when combined they do.

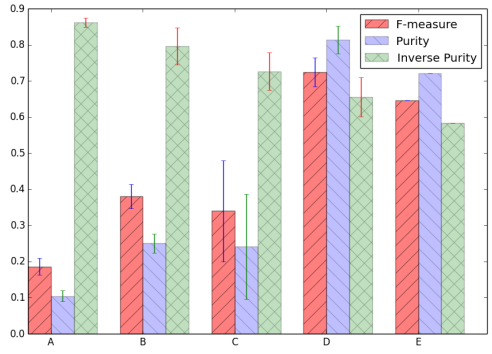


*Figure 2.* Set matching metrics. A) McInerney & Blei model (B) WARBLE w/o simultaneous topic-event learning (C) WARBLE w/o background model (D) WARBLE model (E) Tweet-SCAN

## 6. Conclusions

In this paper we identified three main challenges in event detection from Twitter data, namely rarity, variability and text-shortness. In order to address them, we proposed WARBLE, a new probabilistic model and variational learning algorithm that uncovers real-world events from tweets in an unsupervised manner. The WARBLE model explicitly tackles rarity and variability through a background component, which captures varying tweet densities in time and space. To mitigate text-shortness, our proposal simultaneously learn topics and events making it easier to find word co-occurrences among tweets within the same event.

The preliminary experimental results point out that WARBLE outperforms other state-of-the-art techniques in detecting local events from "La Mercè 2014" dataset. Moreover, the evaluation highlights the need to combine the spatio-temporal background and the simultaneous topic-event learning. Finally, this probabilistic approach to event detection paves the way to reason about unseen observations or partially observed data in a probabilistically well principled way.

## Acknowledgement

This work is partially supported by Obra Social "la Caixa", by the Spanish Ministry of Economy and Competitivity under contract TIN2015-65316-P, by the BSC-CNS Severo Ochoa program (SEV-2015-0493), by the SGR programme (2014-SGR-118) of the Catalan Government and by Collectiveware (TIN2015-66863-C2-1-R) project.

## References


Atefeh, Farzindar and Khreich, Wael. A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1):132–164, 2015.

Banfield, Jeffrey D and Raftery, Adrian E. Model-based gaussian and non-gaussian clustering. *Biometrics*, pp. 803–821, 1993.

Becker, Hila, Naaman, Mor, and Gravano, Luis. Beyond trending topics: Real-world event identification on twitter. In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*, 2011.

Birant, Derya and Kut, Alp. St-dbscan: An algorithm for clustering spatial–temporal data. *Data and Knowledge Engineering*, 60(1):208 – 221, 2007. Intelligent Data Mining.

Blei, David M. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

Blei, David M, Ng, Andrew Y, and Jordan, Michael I. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

Capdevila, Joan, Cerquides, Jesús, Nin, Jordi, and Torres, Jordi. Tweet-scan: An event discovery technique for geo-located tweets. In *Artificial Intelligence Research and Development - Proceedings of the 18th International Conference of the Catalan Association for Artificial Intelligence, Valencia, Catalonia, Spain, October 21-23, 2015.*, pp. 110–119, 2015.

Capdevila, Joan, Cerquides, Jesús, and Torres, Jordi. Variational forms and updates for the WARBLE model. Technical report, `https://www.dropbox.com/s/0qyrkivpsxxv55v/report.pdf?dl=0`, 2016.

Ester, Martin, Kriegel, Hans-Peter, Sander, Jörg, and Xu, Xiaowei. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pp. 226–231, 1996.

Fox, Charles W. and Roberts, Stephen J. A tutorial on variational Bayesian inference. *Artificial Intelligence Review*, 38(2):85–95, 2012.

Hong, Liangjie and Davison, Brian D. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pp. 80–88. ACM, 2010.

Jordan, Michael I, Ghahramani, Zoubin, Jaakkola, Tommi S, and Saul, Lawrence K. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Lee, Ryong and Sumiya, Kazutoshi. Measuring geographical regularities of crowd behaviors for twitter-based geosocial event detection. In *Proceedings of the 2Nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, LBSN '10, pp. 1–10, New York, NY, USA, 2010. ACM.

Li, Linna, Goodchild, Michael F, and Xu, Bo. Spatial, temporal, and socioeconomic patterns in the use of twitter and flickr. *Cartography and Geographic Information Science*, 40(2):61–77, 2013.

McInerney, James and Blei, David M. Discovering newsworthy tweets with a geographical topic model. *NewsKDD: Data Science for News Publishing workshop. Workshop in conjunction with KDD2014 the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2014.

Quan, Xiaojun, Kit, Chunyu, Ge, Yong, and Pan, Sinno Jialin. Short and sparse text topic modeling via self-aggregation. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pp. 2270–2276. AAAI Press, 2015.

Sakaki, Takeshi, Okazaki, Makoto, and Matsuo, Yutaka. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pp. 851–860. ACM, 2010.

Singh, Smita. Spatial temporal analysis of social media data. Master's thesis, Technische Universität München, 2015.

Tamura, K. and Ichimura, T. Density-based spatiotemporal clustering algorithm for extracting bursty areas from georeferenced documents. In *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*, pp. 2079–2084, Oct 2013.

Wong, Weng-Keen and Neill, Daniel B. Tutorial on event detection. In *KDD*, 2009.

Yang, Yiming, Pierce, Tom, and Carbonell, Jaime. A study of retrospective and on-line event detection. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 28–36. ACM, 1998.

Zheng, Yu. Tutorial on location-based social networks. In *Proceedings of the 21st international conference on World wide web, WWW*. ACM, May 2012.