Spatio-Temporal Event Detection from Multiple Data Sources

Aman Ahuja[1], Ashish Baghudana[1], Wei Lu[2], Edward A. Fox[1], and Chandan K. Reddy[1]

[1] Virginia Tech, VA, USA {aahuja,ashishb,fox}@vt.edu,reddy@cs.vt.edu
[2] Singapore University of Technology & Design, Singapore luwei@sutd.edu.sg

Abstract. The proliferation of Internet-enabled smartphones has ushered in an era where events are reported on social media websites such as Twitter and Facebook. However, the short text nature of social media posts, combined with a large volume of noise present in such datasets makes event detection challenging. This problem can be alleviated by using other sources of information, such as news articles, that employ a precise and factual vocabulary, and are more descriptive in nature. In this paper, we propose Spatio-Temporal Event Detection (STED), a probabilistic model to discover events, their associated topics, time of occurrence, and the geospatial distribution from multiple data sources, such as news and Twitter. The joint modeling of news and Twitter enables our model to distinguish events from other noisy topics present in Twitter data. Furthermore, the presence of geocoordinates and timestamps in tweets helps find the spatio-temporal distribution of the events. We evaluate our model on a large corpus of Twitter and news data, and our experimental results show that STED can effectively discover events, and outperforms state-of-the-art techniques.

Keywords: Topic modeling · Probabilistic models · Event detection.

1 Introduction

Social media platforms such as Twitter and Facebook have become a central mode of communication in people's lives. They are regularly used to discuss and debate current events, ranging from natural calamities such as *Hurricane Harvey*, to political incidents like the *U.S. Elections*. These events span different locations and time periods. With strong Internet penetration and the ubiquity of location-enabled smartphones, a large number of social media posts also have the geographical coordinates of the users. These are rich sources of information, aiding location-specific event detection and analysis.

Event detection aims to discover content describing an important occurrence. Applications of event detection include the modeling of a disease outbreak, such as an epidemic of influenza, based on Twitter data [4], and reactions to sporting events [20]. Hence, significant research has been conducted on mining topics from Twitter data [1, 19]. However events are not merely topics, and have aspects of time and location. Researchers have previously defined events with an

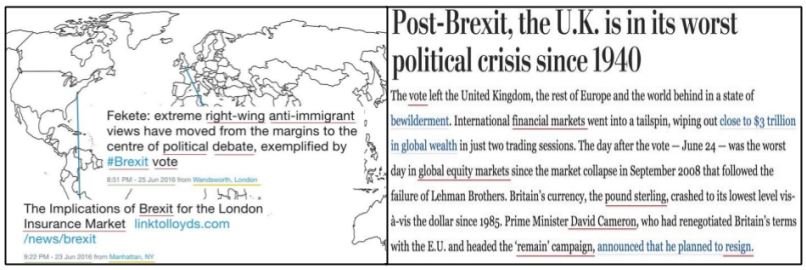


Fig. 1: Tweets and news related to *Brexit*, originating from different parts of the world, and containing several aspects of the event.

approximate geolocation, temporal range and a set of words [18]. However, these definitions do not account for events that span across multiple regions, such as *Hurricane Harvey*, nor do they identify sub-themes of an event, such as *destruction and damage*, and *help and relief*.

In this paper, we propose **S**patio **T**emporal **E**vent **D**etection (STED), a probabilistic model that discovers events using information from various data sources, such as news and Twitter. It combines the location, time, and the text, from tweets, aided with textual information from news articles, to discover the various parameters associated with an event. An event is characterized by the following three attributes:

- The *time* of occurrence. For instance, most tweets about *Rio Olympics* occur in August 2016. Each event, therefore, has a temporal mean and variance.
- A *regional distribution* describing where the event occurs. Global events such as *Brexit* have tweets from several countries, whereas tweets about the *Burning Man Festival* are concentrated in Nevada, US. Hence, an event can occur in one or more regions. Regions are defined by their geographical center and the corresponding covariance.
- A *set of topics* describing the event, where each topic is a facet of the event. *U.S. Elections 2016* contain several topics such as the Republican and Democratic campaigns, as well as the FBI investigation into Russian meddling.

We use timestamp and geolocation information from tweets to estimate the temporal and regional distributions of events, respectively. We supplement the vocabulary in tweets with news text to provide larger context about the facts surrounding the event. This is summarized in Figure 1. This ensures that noisy tweets are eliminated and do not contribute to aspects of an event, while news articles provide more factual information about the event.

2 Related Work

2.1 Topic Modeling

Topic modeling has been widely studied in the domain of text mining to discover latent topics. One of the earliest methods to discover topics in text documents

was probabilistic Latent Semantic Indexing (pLSI) [9]. However, since pLSI was based on the likelihood principle and did not have a generative process, it cannot assign probabilities to new documents. This was alleviated by Latent Dirichlet Allocation (LDA) [6], which models each document as a mixture over topics, and topics as a mixture over words. Inspired by its success, LDA has been extended and applied to various corpora including microblogs such as tweets [19], as well as news documents [14].

2.2 Event Extraction from Text

The most common data-driven approach to event extraction uses text clustering. Within text clustering, the two major paradigms are similarity-based methods and statistical techniques. Similarity-based efforts generally use cosine similarity [11]. These techniques are fast and efficient, however they ignore all the statistical dependencies between variables. Graphical models bring more insight to event detection by modeling dependencies and hierarchies [5]. Another class of event detection models uses spikes in activity as an indication of an event. These bursts change the distribution of the existing data and are detected as new events [12, 13]. These models rely on detecting words that have a sudden increase in activity, while trying to penalize terms that occur consistently in the data. Thus, events are defined only by a subset of terms that have increased co-occurrence.

2.3 Geospatial and Temporal Models

With social media platforms like Twitter and Facebook allowing users to embed their locations in posts, there has been an increase in the availability of data with geospatial and temporal information. As a result, several researchers have incorporated this information in event detection systems. [16] built an earthquake detection system by correlating Twitter messages during a disaster event in Japan. A sudden increase in the volume of tweets in a specific region within a timeframe indicated an event. [7] introduced the Geographical Topic Model where they aimed to discover variation of different topics in latent regions. However, it does not assume a dependency between the latent topics and regions. [2, 10] proposed probabilistic models that address the problem of modeling geographical topical patterns on Twitter. This improved upon prior models that used predefined region labels instead of actual latitude and longitude. However, their focus was more on geographical topics, rather than events. The model proposed in [18] explicitly uses geospatial and temporal information to discover events. It assumes that every event has a single temporal and regional distribution. Furthermore, the authors use data only from Twitter. We improve upon this approach by allowing an event to be spatially distributed by creating a joint model for news and social media. While social media provides quick and short details about an event, the text often contains personal opinion. When combined with news data, event summaries are both factual, and provide views of the people about an event.

3 The Proposed STED Model

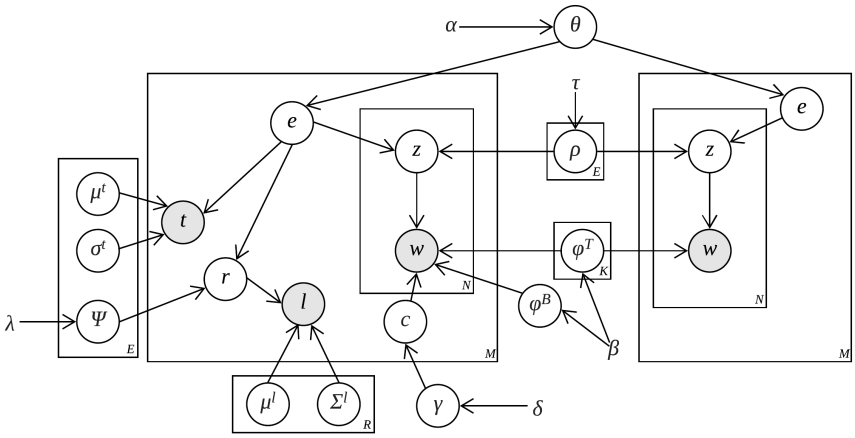


Fig. 2: Plate notation for the model.

In this section, we introduce STED, a probabilistic graphical model that discovers events, and their aspects, across different geographical regions and temporal ranges, from a multimodal corpus of geo-tagged microblogs, such as tweets, and news. Our model is built on the following observations:

- An event refers to an incident that is discussed widely in news and social media, such as "U.S. Elections 2016" and "Rio Olympics 2016". Events have a definite geospatial and temporal distribution. Thus, a particular event is more likely to be discussed within a specific period of time.
- An event can be discussed in multiple geographical regions. Each region can be represented using a bivariate Gaussian distribution, with a geographical center μ_r^l, and variance determined by a diagonal covariance matrix Σ_r^l. For example, "U.S. Elections 2016" is discussed in New York, California, and Texas – each of which belongs to a different region – but not as much in Asia.
- Similarly, an event has a temporal distribution given by its mean time of occurrence, μ_e^t, and variance, σ_e^t. "Brexit Vote" and "Rio Olympics" may have similar regional distributions but occurred during different months – June 2016 and August 2016, respectively.
- News articles and tweets (now with an increased character limit of 280) cover several topical aspects within an event. "Trump Campaign" and "Clinton Campaign" form two topics in the event "U.S. Elections 2016".
- Finally, different events can have recurring subthemes – both "Hurricane Irma" and "Hurricane Harvey" speak about wind speeds, damage, and loss of life, despite having different geospatial and temporal distributions.

3.1 Problem Statement

Given a set of news articles $D^n = \{d_1^n, \ldots, d_{|D^n|}^n\}$, a set of tweets $D^m = \{d_1^m, \ldots, d_{|D^m|}^m\}$, their geolocations $L^m = \{l_1^m, \ldots, l_{|L^m|}^m\}$, their timestamps $T^m =$

Table 1: Notations used in this paper.

Symbol	Description	Symbol	Description	Symbol	Description
M	number of documents	c_m	category of tweet m	ψ	event-region distribution
N_m	number of words in document m	α	Dirichlet prior for θ	γ	tweet-category distribution
E	number of events	β	Dirichlet prior for ϕ^B, ϕ^T	ϕ^T	topic-word distribution
Z	number of topics	δ	Beta prior for γ	ϕ^B	background word distribution
R	number of regions	λ	Dirichlet prior for ψ	$n_{e,g}$	number of documents with event e
V	vocabulary	τ	Dirichlet prior for ρ		and region g
e	event	μ_r^l	geographical center of region r	$p_{e,k}$	number of words with event e
z	topic	Σ_r^l	regional variance of region r		and topic k
r	geographical region	μ_e^t	temporal mean of event e	$s_{k,v}$	number of times term v is used
l_m	latitude, longitude of tweet m	σ_e^t	temporal variance of event e		with topic k
t_m	time of tweet m	θ	corpus-event distribution	$q_{c,v}$	number of times term v is used
w	word	ρ	event-topic distribution		in category c

$\{t_1^m, \ldots, t_{|T^m|}^m\}$, the goal of STED is to find for each event e, a ranked list of topics and regions, temporal mean μ_e^t and variance σ_e^t, as well as a ranked list of words for each topic k. For each geographical location $r \in R$, STED also finds it's geographical mean μ_r^l and variance Σ_r^l.

3.2 Model Definition

STED is a generative model that incorporates the key characteristics described above. It discovers latent events and their corresponding latent topics from a corpus of long documents, such as news, and short geotagged documents, such as social media posts. Figure 2 illustrates the plate notation of our model.

- The model assumes there are E events, K topics, and R regions, the values of which are fixed.
- It models each event e as a mixture of topics and regions, along with a definite temporal distribution.
- For each news article, an event e is drawn from the corpus event distribution θ. Subsequently, for each word in the document, a topic z is drawn from the event topic distribution ρ_e.

Algorithm 1 Generative Process of STED.

Draw event distribution $\theta \sim Dir(\alpha)$
for each event e **do**
  Draw region distribution $\psi_e \sim Dir(\lambda)$
  Draw topic distribution $\rho_e \sim Dir(\tau)$
end for
for each topic z **do**
  Draw word distribution $\phi_z^T \sim Dir(\beta)$
end for
Draw background word distribution $\phi^B \sim Dir(\beta)$
for each news article m **do**
  Draw event $e_m \sim Mult(\theta)$
  for each word n **do**
    Draw topic $z_{m,n} \sim Mult(\rho_{e_m})$
    Draw word $w_{m,n} \sim Mult(\phi_{z_{m,n}}^T)$
  end for
end for
For tweets, draw category distribution $\gamma \sim Beta(\delta)$
for each tweet m **do**
  Draw category $c_m \sim Bin(\gamma)$
  if $c_m = 1$ **then**
    Draw event $e_m \sim Mult(\theta)$
    Draw timestamp $t_m \sim N(\mu_{e_m}^t, \sigma_{e_m}^t)$
    Draw region $r_m \sim Mult(\psi_{e_m})$
    Draw geolocation $l_m \sim N(\mu_{r_m}^l, \Sigma_{r_m}^l)$
    for each word n **do**
      Draw topic $z_{m,n} \sim Mult(\rho_{e_m})$
      Draw word $w_{m,n} \sim Mult(\phi_{z_{m,n}}^T)$
    end for
  else if $c_m = 0$ **then**
    for each word n **do**
      draw word $w_{m,n} \sim Mult(\phi^B)$
    end for
  end if
end for

– Since tweets are often noisy, they may or may not be related to an event. Hence, for every tweet, a category c is sampled from the category distribution.
  - If $c = 1$, an event e is drawn from the corpus event distribution θ, and the region r, geolocation (latitude, longitude) l, and time t of the tweet are drawn. Subsequently, for each word, a topic is sampled from the event topic distribution ρ_e.
  - If $c = 0$, the tweet is regarded as a noisy tweet and each word in the document is sampled from the background word distribution ϕ^B.

The detailed generative process of STED is described in Algorithm 1.

3.3 Model Inference

We use the Gibbs-EM algorithm [3, 8] for inference in the STED model. We first integrate out the model parameters θ, ρ, ψ, γ, ϕ^T, and ϕ^B using Dirichlet-Multinomial conjugacy. After this, the latent variables left in the model are e, r, c, z, μ^l, Σ^l, μ^t, and σ^t. We sample the latent variables e, z, r, and c in the E-step of the algorithm using the following equations:

Sampling Event For a news article m, the event e_m can be sampled by:

$$P(e_m = x|*) \propto (n_{x,*}^{-m} + \alpha_x) \times \frac{\prod_{j \in Z_m} \prod_{y=0}^{p_{x,j}^m - 1} (p_{x,j}^{-m} + \tau_{x,j} + y)}{\prod_{y=0}^{N_m - 1} (\sum_{j=1}^{K} (p_{x,j}^{-m} + \tau_{x,j}) + y)} \tag{1}$$

For tweets, given the region is g and the timestamp is t,

$$P(e_m = x|*) \propto (n_{x,*}^{-m} + \alpha_x) \times \frac{n_{x,g}^{-m} + \lambda_g}{\sum_{i=1}^{R} n_{x,i}^{-m} + \lambda_i} \times \frac{\prod_{j \in Z_m} \prod_{y=0}^{p_{x,j}^m - 1} (p_{x,j}^{-m} + \tau_{x,j} + y)}{\prod_{y=0}^{N_m - 1} (\sum_{j=1}^{K} (p_{x,j}^{-m} + \tau_{x,j}) + y)} \times \mathcal{N}(t_m | \mu_x^t, \sigma_x^t) \tag{2}$$

Sampling Topic For a news article or tweet m, the topic z_{mn} for the word n with vocabulary index t can be sampled by:

$$P(z_{mn} = k|e = x) \propto \frac{s_{k,t}^{-mn} + \beta_t}{\sum_{r=1}^{V} s_{k,r}^{-mn} + \beta_r} \times \frac{p_{x,k}^{-mn} + \tau_{x,k}}{\sum_{j=1}^{K} p_{x,j}^{-mn} + \tau_{x,j}} \tag{3}$$

Sampling Region Geographical region is sampled only for tweets with category $c = 1$, given their corresponding event is e and geo-coordinates is l_m, as follows:

$$P(r_m = g|*) \propto \frac{n_{e,g}^{-m} + \lambda_g}{\sum_{j=1}^{R} n_{e,j}^{-m} + \lambda_j} \times \mathcal{N}(l_m | \mu_g^l, \Sigma_g^l) \tag{4}$$

Sampling Category The category (background or event-related) is sampled only for tweets, using the following equation:

$$P(c_m = d) \propto \frac{q_{d,*}^{-m} + \delta}{\sum_{i=0}^{1} q_{i,*}^{-m} + \delta} \times \frac{\prod_{r=1}^{V} \Gamma(q_{d,r}^{-m} + \beta_r)}{\Gamma(\sum_{r=1}^{V} q_{d,r}^{-m} + \beta_r)} \times \frac{\prod_{r \in V_m} \prod_{y=0}^{N_m - 1} (q_{d,r}^{-m} + \beta_r + y)}{\prod_{y=0}^{N_m - 1} (\sum_{r=1}^{V} (q_{d,r}^{-m} + \beta_r) + y)} \tag{5}$$

After sampling the latent variables e, z, c, and r, the geographical center μ_r^l, and covariance matrix Σ_r^l, is updated for each region r. The temporal mean μ_e^t and variance σ_e^t, is also updated for each event e.

3.4 Priors for Model Initialization

The STED model uses a bivariate Normal distribution on the location variable l. The mean μ_r^l and covariance Σ_r^l for the regions in R serve as the prior for this Normal distribution. To initialize these parameters, we run K-means clustering on the tweet geo-coordinates. The values of the mean and average co-variance obtained for the clusters are used as the prior μ_r^l and Σ_r^l for latent regions. The latent variables e, z, and c for all the tweets and news articles are randomly initialized, and all the distribution parameters are set using the initialized values of variables they use.

4 Experiments

4.1 Dataset Description and Preprocessing

For empirical evaluation of STED, we estimate its performance on a large real-world data, composed of tweets and news articles from the year 2016.

1. *Tweet Data:* This dataset consists of tweets with geolocations collected through 2016 using the Twitter Streaming API for a period of 7 months from June 2016 to December 2016. The Twitter Streaming API is believed to give a 1% random sample of tweets streaming on Twitter. We further performed a random sampling and obtained 1 million tweets from the collected data. Subsequently, we filtered out all the tweets that had less than 90% English characters and encoded the remaining tweets with an ASCII codec. This final dataset contained 715,262 tweets.
2. *News Data:* We collected news data from the articles published in Washington Post for the time period mentioned above. This dataset contained 148,769 news articles.

All the documents in both the datasets were lowercased and preprocessed to remove common stop words and punctuation marks. Tweets were further processed to remove all usernames and URLs. However, we retained all hashtags as they contain valuable information about events.

4.2 Performance Evaluation

For quantitative comparison of STED against baseline techniques, we use the following two metrics:

- *Perplexity:* This is a measure of the degree of uncertainty in fitting test documents to a language model. It is defined as the negative log-likelihood of test documents using the trained model.

$$Perp(D) = exp\left\{ \frac{-\sum_{d \in D} \log p(w_d)}{\sum_{d \in D} N_d} \right\} \tag{6}$$

  A lower perplexity indicates better predictive performance. $p(w_d)$ is the joint probability of the word w_d occurring in an event-related and non-event related document d, and N_d is the number of words in document d.

– *Topic Coherence:* It is measured using Pointwise Mutual Information (PMI), which is a measure of information overlap between two variables. Prior research indicates that PMI is well correlated with human judgment of topic coherence [15].

$$\text{PMI-Score} = \frac{1}{EZ} \sum_{e=1}^{E} \sum_{z=1}^{Z} \sum_{i<j} \log \left\{ \frac{P(w_i, w_j)}{P(w_i)P(w_j)} \right\} \tag{7}$$

where E = number of events, and Z = number of topics. $P(w_i)$ indicates the proportion of documents containing word w_i. Consequently, $P(w_i, w_j)$ indicates the proportion of documents containing words w_i and w_j. A higher PMI score shows better topic coherence.

4.3 Baseline Methods

We compare the aforementioned metrics on the following models:

1. **LDA [6]:** An implementation of LDA using collapsed Gibbs sampling.
2. **GeoFolk [17]:** A spatial topic model that aims to discover topics and their geographical centers.
3. **BGM [18]:** A Bayesian Graphical Model to discover latent events from Twitter, that models events with geographical and temporal centers, and their associated variances. We refer to this model as **BGM**.
4. **STED-T:** A variant of our model which uses only tweets to discover events.

4.4 Parameter Setting

To initialize STED, the following hyperparameters are required: α, β, τ, λ, and δ. These hyperparameters serve as priors for each of the distributions. We used symmetric values for these hyperparameters, all of which were derived empirically. Specifically, we set $\alpha = 1$, $\beta = 0.01$, $\tau = 0.1$, $\delta = 0.1$, and $\lambda = 10$. The priors μ_e^t and σ_e^t for temporal mean and variance, as well as μ_r^l, and Σ_r^l were set as specified in Section 3.4.

We ran our model, its variant, and BGM, for 50 EM iterations, with 10 Gibbs sampling steps in each E-step of the iteration. We varied both the number of events and the number of topics. The other baseline models were run for 500 Gibbs sampling iterations.

4.5 Experimental Results

Quantitative Results In this section, we discuss the quantitative metrics of the STED model. We compare the perplexity and topic coherence of our model against baselines, and also show how the addition of news articles improves the performance of the model. Since our model is hierarchical, we measure these metrics by first varying the number of events, fixing the number of topics to 50, and then varying the number of topics, fixing the number of events to 10.

a. Perplexity: We observe that the perplexity of STED is consistently better than that of all baseline models (Figure 3(a)). This shows that event detection

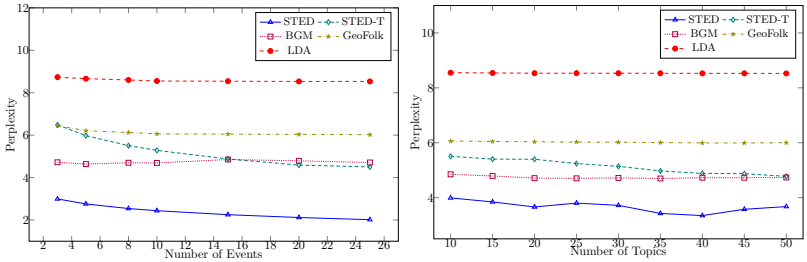
is not merely dependent on words in each document, but also on the spatial and temporal distribution of the documents. We further observe that STED outperforms STED-T, indicating that the addition of news data improves the predictive power of the model.

We also notice that even though BGM performs worse than STED, it's performance is at-par with STED-T. Therefore, for a dataset such as tweets, that contains geolocation information, it is better to consider latent regions as a mixture of Gaussian distributions, rather than using predefined regions based on the coordinates. The rise in perplexity beyond $e = 10$ can be explained by overfitting of the BGM model. This trend remains the same even when we vary the number of topics while keeping the number of events constant (Figure 3(b)).

It is also interesting to see that the perplexity plots are uniformly flat for most of the baselines, indicating that the dataset was noisy. Despite the noise, the qualitative results show that STED correctly identified events of world importance, that occurred during the timeframe that the dataset was collected.

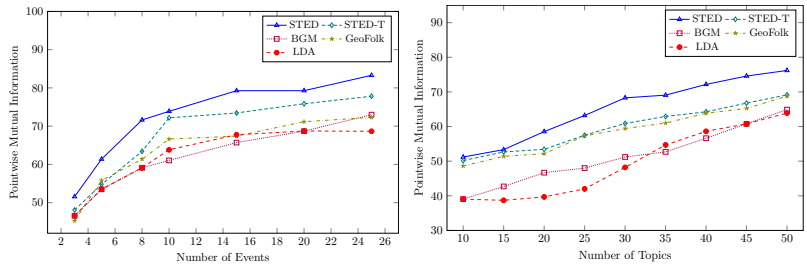
b. Topic Coherence: As described in Section 4.2, we use PMI as an indicator of topic coherence. We compare the normalized PMI score of our model to those of LDA, GeoFolk, and BGM, for the top twenty words in each event (or topic).

Figures 3(c) and 3(d) show that STED has the highest PMI score. GeoFolk and LDA perform comparably in topic coherence, i.e., the topics are equally interpretable in both of these models. This is expected since GeoFolk only accounts



(a) Perplexity (varying number of events)

(b) Perplexity (varying number of topics)



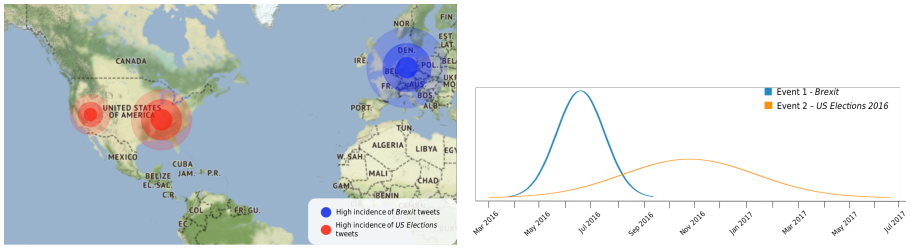
(c) Topic Coherence (varying number of events)

(d) Topic Coherence (varying number of topics)

Fig. 3: Performance comparison of STED with other baseline methods.

for geographical topics and does not consider temporal information. Moreover, it is trained only on tweets, and not news data. For the same reason, STED outperforms STED-T. This demonstrates that vocabulary from news articles improves the readability of topics generated from the model. BGM makes the implicit assumption that events are concentrated in a specific region, which fails for events with more distributed geolocations, such as U.S. Elections or Brexit. The joint modeling makes STED's PMI score higher than BGM and GeoFolk.

Qualitative Results For qualitative evaluation of our model, we identify two events, *U.S. Elections 2016* and *Brexit*. We characterize these events across three features – latent regions in which these events were prevalent, their temporal distributions, and their topics.



(a) Geographic regions (b) Temporal variation

Fig. 4: Geographic regions and temporal variation for events *Brexit* and *U.S. Elections 2016*.

Tweets about *U.S. Elections 2016* were largely localized to North America, while those about *Brexit* were concentrated in Europe (Figure 4a). Since we model temporal distributions as Gaussian (Figure 4b), we observe that the event *U.S. Elections 2016* was centered at Oct. 31, 2016 (elections were held on Nov. 8, 2016), and *Brexit* was centered at June 30, 2016 (actual vote happened on June 23, 2016).

(a) U.S. Elections 2016 (b) Brexit

Fig. 5: Word clouds representing the top two topics generated from STED.

In Figure 5, we show the top two topics for each of these events, generated from STED. Each topic is described by its corresponding top-ranking words. The first topic in Figure 5(a) describes the Republican campaign with references to Donald Trump, Jeb Bush, and Ted Cruz, while the second topic details the Democratic campaign focusing on Hillary Clinton, and Bernie Sanders. The first topic in Figure 5(b) mentions the Prime Minister of Britain, David Cameron, and the second topic illustrates the anti-immigration sentiment prevalent at the time of the vote.

5 Conclusion

In this paper, we presented STED, a novel probabilistic topic model to extract latent events, from a heterogenous corpus of documents from multiple data sources, such as long and short documents. Because of the growing importance of social media, which also has location and time information, but limited textual information, we used Twitter data as one of the data sources in STED. To overcome the sparsity of textual information available in social media data, we use a much more elaborate form of data, such as news articles, as other source. This improves the predictive power of the model, by providing relevant vocabulary, along with spatial and temporal information. Furthermore, the use of latent regions helps define events more naturally – geospatially distributed, but temporally centered. The results obtained on Twitter and news data from 2016 show that the model obtains meaningful results and outperforms state-of-the-art techniques on several quantitative metrics. We hope that STED can be an important tool in detecting the different aspects of events, such as disasters, and help government agencies better plan and mitigate such events.

Acknowledgments

This work was supported in part by the US National Science Foundation grants IIS-1619028, IIS-1707498, and IIS-1838730.

References

1. Ahuja, A., Wei, W., Carley, K.M.: Microblog Sentiment Topic Model. In: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW). pp. 1031–1038. IEEE (2016)
2. Ahuja, A., Wei, W., Lu, W., Carley, K.M., Reddy, C.K.: A probabilistic geographical aspect-opinion model for geo-tagged microblogs. In: Data Mining (ICDM), 2017 IEEE International Conference on. pp. 721–726. IEEE (2017)
3. Andrieu, C., De Freitas, N., Doucet, A., Jordan, M.I.: An introduction to MCMC for Machine Learning. Machine Learning **50**(1-2), 5–43 (2003)
4. Aramaki, E., Maskawa, S., Morita, M.: Twitter catches the flu: detecting influenza epidemics using Twitter. In: Proceedings of the Conference on Empirical methods in Natural Language Processing. pp. 1568–1576. Association for Computational Linguistics (2011)

5. Benson, E., Haghighi, A., Barzilay, R.: Event discovery in social media feeds. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. pp. 389–398. Association for Computational Linguistics (2011)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of machine learning research **3**(Jan), 993–1022 (2003)
7. Eisenstein, J., O'Connor, B., Smith, N.A., Xing, E.P.: A latent variable model for geographic lexical variation. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. pp. 1277–1287. Association for Computational Linguistics (2010)
8. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences **101**(suppl 1), 5228–5235 (2004)
9. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval. pp. 50–57. ACM (1999)
10. Hong, L., Ahmed, A., Gurumurthy, S., Smola, A.J., Tsioutsiouliklis, K.: Discovering geographical topics in the Twitter stream. In: Proceedings of the 21st international conference on World Wide Web. pp. 769–778. ACM (2012)
11. Kumaran, G., Allan, J.: Text classification and named entities for new event detection. In: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval. pp. 297–304. ACM (2004)
12. Mathioudakis, M., Koudas, N.: TwitterMonitor: trend detection over the Twitter stream. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data. pp. 1155–1158. ACM (2010)
13. Matuszka, T., Vinceller, Z., Laki, S.: On a keyword-lifecycle model for real-time event detection in social network data. In: Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on. pp. 453–458. IEEE (2013)
14. Newman, D., Chemudugunta, C., Smyth, P., Steyvers, M.: Analyzing entities and topics in news articles using statistical topic models. In: ISI. pp. 93–104. Springer (2006)
15. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 100–108. Association for Computational Linguistics (2010)
16. Sakaki, T., Okazaki, M., Matsuo, Y.: 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)
17. Sizov, S.: Geofolk: latent spatial semantics in web 2.0 social media. In: Proceedings of the third ACM international conference on Web search and data mining. pp. 281–290. ACM (2010)
18. Wei, W., Joseph, K., Lo, W., Carley, K.M.: A bayesian graphical model to discover latent events from Twitter. In: ICWSM. pp. 503–512 (2015)
19. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.P., Yan, H., Li, X.: Comparing Twitter and traditional media using topic models. In: European Conference on Information Retrieval. pp. 338–349. Springer (2011)
20. Zubiaga, A., Spina, D., Amigó, E., Gonzalo, J.: Towards real-time summarization of scheduled events from Twitter streams. In: Proceedings of the 23rd ACM Conference on Hypertext and Social Media. pp. 319–320. ACM (2012)