# Detecting and Characterizing Events

Anonymous EMNLP submission

## Abstract

Significant events are characterized by interactions between entities (e.g., countries, organizations, individuals) that deviate from typical interaction patterns. Investigators, such as historians, commonly read large quantities of text to construct an accurate picture of who, what, when, and where an event happened. In this work, we present the Capsule model for analyzing documents to identify and characterize events of potential significance. Specifically, we develop a model based on topic modeling to distinguish between topics that describe "business-as-usual" and topics that deviate from these patterns. To demonstrate this model, we analyze a corpus of over 2 million US State Department cables from the 1970s; we provide open-source implementations of an inference algorithm for the Capsule model and a pipeline to explore its results.

## 1 Introduction

Foreign embassies of the United States government communicate with each other and with the U.S. State Department through cabled message. The National Archive collects these documents in a running corpus, which traces the (unclassified) diplomatic history of the United States. It has collected, for example, about two million cables sent between 1973 and 1978.

Typically, a cable from this collection describes diplomatic "business as usual," such as arrangements for visiting officials, recovery of lost or stolen passports, or obtaining lists of names for meetings and conferences. For example, the embassies sent 8,635 cables during the week of April 21, 1975. Here is one, selected at random,

> Hoffman, UNESCO Secretariat, requested info from PermDel concerning an official invitation from the USG RE subject meeting scheduled 10-13 JUNE 1975, Madison, Wisconsin. Would appreciate info RE status of action to be taken in order to inform Secretariat. Hoffman communicating with Dr. John P. Klus RE list of persons to be invited.

But hidden in the corpus are also cables about important diplomatic events, the cables and events that are of primary interest to historians. During that same week, the United States was in the last moments of the Vietnam war and, on April 30, 1975, lost its hold on Saigon. This resulted in the end of the Vietnam War and a max exodus of refugees from the country. One of the cables around this event is

> GOA program to move Vietnamese Refugees to Australia is making little progress and probably will not cover more than 100-200 persons. Press comment on smallness of program has recognized difficulty of getting Vietnamese out of Saigon, but "Canberra Times" Apr 25 sharply critical of government's performance. [...] Labor government clearly hopes whole matter will somehow disappear.

Our goal in this paper is to develop a method to help historians and political scientists wade through their collections, such as the 1970s cables, to find potentially important events, such as the fall of Saigon,

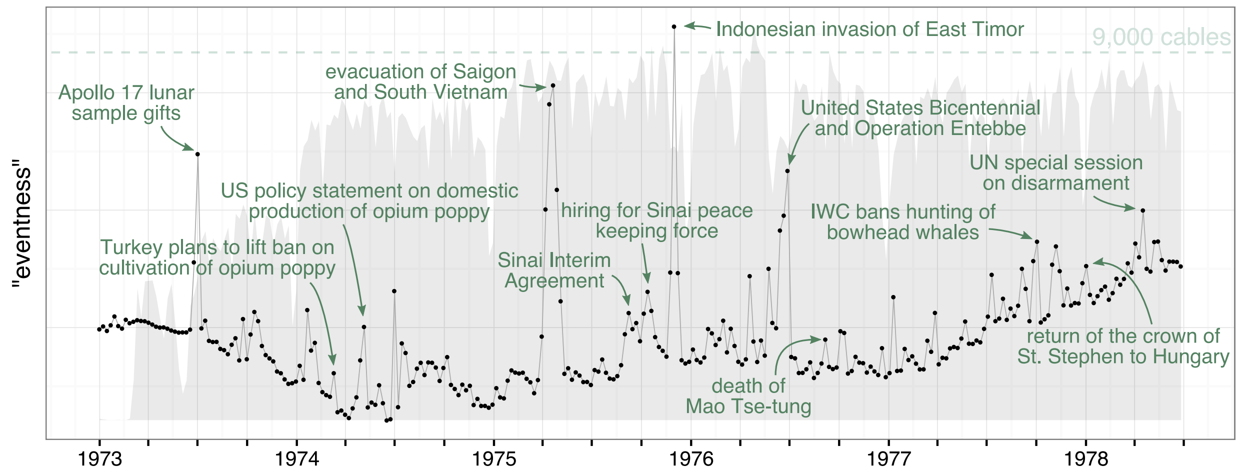


**Figure 1:** Measure of "eventness," or time interval impact on cable content (Eq. 2). Grey background indicates the number of cables sent over time. This comes from the model fit we discuss in Section 3. Capsule successful detects real-world events from National Archive diplomatic cables.

and the primary sources around them. We develop *Capsule*, a probabilistic model for detecting and characterizing important events in large collections of historical communication.

Figure 1 illustrates Capsule's analysis of the two million cables from the National Archives. The $y$-axis is "eventness", a loose measure how strongly a week's cables deviate from the usual diplomatic chatter to discuss a matter that is common to many embassies. (This is described in detail in Section 2.)

The figure shows that Capsule detects many of the important moments during this five-year span, including Indonesia's invasion of East Timor (Dec. 7, 1975), the Air France hijacking and Israeli rescue operation (June 27–July 4, 1976), and the fall of Saigon (April 30, 1975). It also identifies other moments, such as the U.S. sharing lunar rocks with other countries (March 21, 1973) and the death of Mao Tse-tung (Sept. 9, 1976). Broadly speaking, Capsule gives a picture of the diplomatic history of these five years; it identifies and characterizes moments and source material that might be of interest to a historian.

The intuition behind Capsule is this: embassies write cables throughout the year, usually describing typical business such as the visiting of a government official. Sometimes, however, there is an important event, e.g., the fall of Saigon. When an event occurs, it pulls embassies away from their typical business to write cables that discuss what happened and its consequences. Thus Capsule effectively defines an "event" to be a moment in history when embassies deviate from what each usually discusses, and when each embassy deviates in the same way.

Capsule embeds this intuition into a Bayesian model. It uses hidden variables to encode what "typical business" means for each embassy, how to characterize the events of each week, and which cables discuss those events. Given a corpus, the corresponding posterior distribution provides a filter on the cables that isolates important moments in the diplomatic history. Figure 1 illustrates the mean of this posterior.

Capsule can be used to explore any corpora with the same underlying structure: text (or other discrete multivariate data) generated over time by known entities. This includes email, consumer behavior, social media posts, and opinion articles.

We present the model in Section 2, providing both a formal model specification and guidance on how to use its posterior to detect and characterize real-worlds events. In Section 3, we evaluate Capsule and explore its results on a collection of U.S. State Department cables and on simulated data.

**Related work.** We first review previous work on automatic event detection and other related concepts.

In both univariate and multivariate settings, the goal is often that analysts want to predict whether or not rare events will occur (Weiss and Hirsh, 1998; Das et al., 2008). Capsule, in contrast, is designed to help analysts explore and understand the original data: our goal is interpretability, not prediction.

Events can also be construed as "change points" to mark when typical observations shift semi-permanently from one value to another (Guralnik and Srivastava, 1999; Adams and MacKay, 2007). Both varieties of events are important, but we focus on temporary shifts away from normal.

A common goal is to identify clusters of documents; these approaches are used on news articles (Zhao et al., 2012; Zhao et al., 2007; Zhang et al., 2002; Li et al., 2005; Wang et al., 2007; Allan et al., 1998) and social media posts (VanDam, 2012; Lau et al., 2012; Jackoway et al., 2011; Sakaki et al., 2010; Reuter and Cimiano, 2012; Becker et al., 2010; Sayyadi et al., 2009). In the case of news articles, the task is to create new clusters as novel news stories appear—this does not help disentangle typical content from rare events of interest. Social media approaches identify rare events, but the methods are designed for short, noisy documents; they are not appropriate for larger documents that contain information about a variety of subjects.

Many existing methods use document terms as features, usually weighted by tf-idf value (Fung et al., 2005; Kumaran and Allan, 2004; Brants et al., 2003; Das Sarma et al., 2011; Zhao et al., 2007; Zhao et al., 2012); here, events are bursts in groups of terms.

Topic models (Blei, 2012) reduce the dimensionality of text data; they have been used to help detect events mentioned in social media posts (Lau et al., 2012; Dou et al., 2012) and posts relevant to monitored events (VanDam, 2012). We rely on topic models to characterize both typical content and events, but grouped observations can also be summarized directly (Peng et al., 2007; Chakrabarti and Punera, 2011; Gao et al., 2012).

In addition to text data over time, author (Zhao et al., 2007), news outlet (Wang et al., 2007), and spatial information (Neill et al., 2005; Mathioudakis et al., 2010; Liu et al., 2011) can be used to augment event detection. Capsule uses author information in order to characterize the typical concerns of authors.

Detecting and characterizing relationships (Schein et al., 2015; Linderman and Adams, 2014; Das Sarma et al., 2011) is related to event detection. When a message recipient is known, Capsule can use a sender-receiver pair in place of an author, but the model could be further tailored for network interactions.

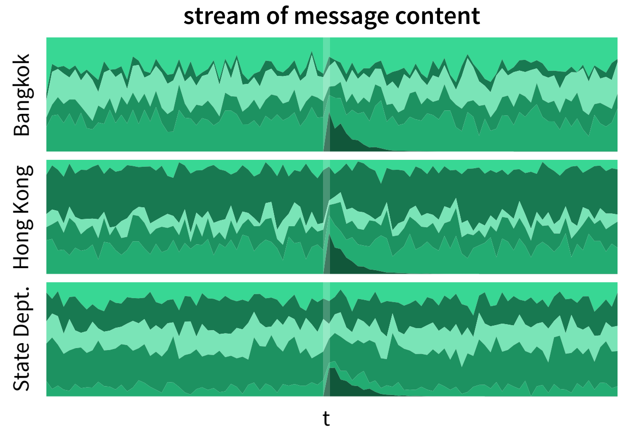


**Figure 2:** Cartoon intuition of Capsule; the $y$ axis is the stacked proportion of messages about various subjects during a given time interval. The Bangkok embassy, Honk Kong embassy, and State Department all have typical concerns about which they usually send messages. When an events occurs at time $t$, the stream of message content alters to include the event, then fades back to "business as usual." Capsule discovers entities' typical concerns as well as the event occurrence and content.

## 2 The Capsule Model

In this section we develop the Capsule model for detecting and characterizing events. Capsule relies on text data sent between entities over time, and builds on topics models. We first give the intuition on Capsule, then formally specify the model. We also describe how to explore a corpus using Capsule and how we learn its hidden variables.

Consider an entity like the Bangkok American embassy, shown in Figure 2. We can imagine that there is a stream of messages (or *diplomatic cables*) being sent by this embassy—some might be sent to the US State Department, others to another American embassy like Hong Kong. An entity will usually talk about certain topics; the Bangkok embassy, for instance, is concerned with topics regarding southeast Asia more generally.

Now imagine that at a particular time $t$, an event occurs, such as the capture of Saigon during the Vietnam war. We do not directly observe that events occur, but we do observe the message stream. Using this stream, each event will be described as a distribution over the vocabulary, similar to how topics are distributions over these same terms. When an event occurs, the message content changes for multiple entities—significant events impact multiple parties

| topic type | top terms |
|---|---|
| general | plan, visit, arrival, itinerary, visitor |
| entity | soviet, moscow, embasy, ussr, meet |
| event | vietnam, evacuation, evacuate, missionary |

**Table 1:** Top vocabulary terms for examples of each of the three topic varieties; these three types of topics blend to form the distribution of each message. They come from the model fit we discuss in Section 3.

simultaneously. The day following the capture of Saigon, for instance, the majority of the diplomatic cables sent by the Bangkok embassy and several other entities were about Vietnam war refugees. Thus we imagine that an entity's stream of messages is controlled by what it usually talks about as well as the higher level stream of unobserved events.

**Model Specification.** We now define Capsule in detail. Our data are *entities* sending *messages* over *time*. The observed variables are $w_{d,v}$, the number of times term $v$ occurs in message $d$. The message is associated with an entity (or author) $a_d$ and a time (or date) interval $i_d$.

We model each message with a bank of Poisson distributions, one per term in the vocabulary, $w_{d,v} \sim \text{Poisson}(\lambda_{d,v})$. The rate $\lambda_{d,v}$ blends the different influences on the content of the message, which are defined in terms of different types of *topics*. A topic, as in typical topic modeling (Blei et al., 2003; Canny, 2004; Gopalan et al., 2014), is a distribution over terms.

Specifically, the message blends general topics about diplomacy (e.g., diplomats, communication) $\beta_k$, an entity topic that is specific to the author of the message (e.g., terms about France) $\eta_{a_d}$,[1] and an event topic that is specific to the events of relevant recent weeks (e.g., terms about an international crisis) $\gamma_t$. Notice how messages share these topics in different configurations: all messages share the general topics; messages from the same entity share the entity topics; and messages from the same interval share the event topics.

Examples of these three types of topics are in Table 1—the general topic relates to planning travel, the entity topic captures words related to the U.S.S.R., and the event topic captures words related to the evacuation of Saigon toward the end of the Vietnam War.

Each message blends its corresponding topics with different strengths, which are drawn per message. Each message represents a different mix of the events of recent weeks, entity-specific items, and general diplomacy.

Putting this together, the Poisson rate for term $v$ in document $d$ is

$$\lambda_{d,v} = \theta_d^\top \beta_v + \zeta_d \eta_{a_d,v} + \sum_{t=1}^{T} f(i_d, t)\epsilon_{d,t}\gamma_{t,v}, \quad (1)$$

where $\theta_d$ corresponds to strength of general diplomacy, $\zeta_d$ corresponds to strength of entity-specific concerns, and $\epsilon_d$ corresponds to strength of events; $f$ is some function of decay. This function is important because events should not remain at their full strength indefinitely, but should decay over time. In our experiments, we find that exponential decay, as in Equation (3), performs well.

We place gamma priors on the topic strengths and Dirichlet priors on the topics. The distributions of general and entity topic strengths are defined hierarchically by entity, capturing the different topics that each entity tends to discuss. The prior on the entity strength is also defined hierarchically; different weeks are more or less "eventful." The full generative process is in Figure 3.

Given a collection of messages, posterior inference uncovers the different types of topics and how each message exhibits them. We will see below, how inferences about the event strengths enable us to filter the corpus to find important messages.

**Detecting and characterizing events.** Once we estimate the posterior distribution of the Capsule parameters, we can use the expectations of the latent parameters to explore the original data. To detect events, we average the per-document event relevancy parameters $\epsilon$ for each document in the interval and multiply it by the interval strength $\psi$:

$$m_t = \mathbb{E}[\psi_t]\frac{1}{|D_t|}\sum_{d \in D_t} \mathbb{E}[\epsilon_{d,t},] \quad (2)$$

where $D_t$ is the set of all cables sent in interval $t$. This measure of "eventness" provides a scaled estimate of the number of words that are related to

[1] These entity-specific topics are similar to background topics (Paul and Dredze, 2012).

- for each time step $t = 1{:}T$,
  - draw interval description over vocabulary (event topic) $\gamma_t \sim \text{Dirichlet}_V(\alpha)$
  - draw interval strength $\psi_t \sim \text{Gamma}(s_\psi, r_\psi)$
- for each entity $n = 1{:}N$,
  - draw entity-specific topics over vocabulary $\eta_n \sim \text{Dirichlet}_V(\alpha)$
  - draw entity-specific topic strength $\xi_n \sim \text{Gamma}(s_\xi, r_\xi)$
- for each topic $k = 1{:}K$,
  - draw general topic distribution over vocabulary $\beta_k \sim \text{Dirichlet}_V(\alpha)$
  - for each entity $n = 1{:}N$,
    - draw general entity concern $\phi_{n,k} \sim \text{Gamma}(s_\phi, r_\phi)$
- for each document $d = 1{:}D$ sent at time $i_d$ by author $a_d$,
  - draw local entity concern $\zeta_d \sim \text{Gamma}(s_\zeta, \xi_{a_d})$
  - for each topic $k = 1{:}K$,
    - draw local entity concern $\theta_{d,k} \sim \text{Gamma}(s_\theta, \phi_{a_d,k})$
  - for each time $t = 1{:}T$,
    - draw local interval relevancy $\epsilon_{d,t} \sim \text{Gamma}(s_\epsilon, \psi_t)$
  - for each vocabulary term $v = 1{:}V$,
    - set $\lambda_{d,v} = \theta_d^\top \beta_v + \zeta_d \eta_{a_d} + \sum_{t=1}^{T} f(i_d, t)\epsilon_{d,t}\gamma_{t,v}$
    - draw word counts $w_{d,v} \sim \text{Poisson}\left(\lambda_{d,v}\right)$

**Figure 3:** The generative process for Capsule.

an real-world event in that interval. Figure 1 shows events detected with this metric.

Given an identified event, we can characterize it in terms of its top terms under $\gamma$, but we can also use event relevancy parameters $\epsilon$ to sort documents; Section 3 explores relevant documents for events found in the National Archive diplomatic cables data. In addition to detecting and characterizing events, Capsule can be used to explore entity concerns and the general themes in a given collection.[2]

There are connections between Capsule and recent work on Poisson processes. In particular, we can interpret Capsule as a collection of related discrete time Poisson processes with random intensity measures. Further, marginalizing out the event strength prior reveals that word use from one entity can "excite" word use in another, which suggests a close relationship to Hawkes processes (Hawkes, 1971).

**Learning the hidden variables.** In order to use the Capsule model to explore the observed documents, we must compute the posterior distribution. Conditional on the observed word counts $w$, our goal is to compute the posterior values of the hidden parameters—general topics $\beta$, entity topics $\eta$, event topics $\gamma$, entity concerns $\phi$ (for general topics) and $\xi$ (for their own topic), overall event strengths $\psi$, and document-specific strengths for general topics $\theta$, entity topics $\zeta$, and event topics $\epsilon$.

As for many Bayesian models, the exact posterior for Capsule is not tractable to compute; approximating it is our central statistical and computational problem. We develop an approximate inference algorithm for Capsule based on variational methods (Jordan et al., 1999),[3] which is detailed in Appendix A. This algorithm produces a fitted variational distribution which can then be used as a proxy for the true posterior, allowing us to explore a collection of documents with Capsule.

## 3 Evaluation

In this section we explore the performance of Capsule on a collection of U.S. State Department diplomatics cables and on simulated data.

**Data.** The National Archive collects communications between the U.S. Sate Department and its embassies. We obtained a collection of these diplomatic messages from the History Lab at Columbia,[4] which received them from the Central Foreign Policy Files at the National Archives. The communications in this data set were sent between 1973 and 1978.

In addition to the text of the cables themselves, each document is supplemented with information about who sent the cable (e.g., the State Department, the U.S. Embassy in Saigon, or an individual by name), who received the cable (often multiple entities), and the date the cable was sent. We used

[2] Upon publication, we will release code for a pipeline to visualize and explore a corpus, given a Capsule fit.

[3] Source code is available at https://github.com/?????/capsule.

[4] http://history-lab.org

a vocabulary of size 6,293 and omitted cables with fewer than three terms, resulting in a collection of 2,139,324 messages sent between 27,134 entities. We selected a weekly duration for the time intervals, as few cables were sent on the weekends.

**Model settings.** We fit Capsule with $K = 100$ general topics and using an exponential decay $f$,

$$f(i_d, t) = \begin{cases} 0, & \text{if } t > i_d \\ \exp\{-(i_d - t)/\tau\}, & \text{otherwise,} \end{cases} \quad (3)$$

with mean lifetime $\tau = 3$. This mean lifetime indicates that most intervals would no longer be relevant after about three weeks. With these settings on the cables data, fitting the model takes 2.8 hours per iteration;[5] results are shown on 15 iterations.

**Results.** We begin our exploration by detecting events using Capsule. With Equation (2) as our metric of "eventness," we consider this metric over time, which is shown in Figure 1. Here, peaks correspond to real-worlds events, several of which are labeled.[6]

The tallest peak occurs the week of December 1, 1975, just prior to the Indonesian invasion of East Timor, which began December 7, 1975. As discussed in Section 2, we sort documents by their event relevancy parameters $\epsilon$ to find cables that reflect an event. Table 2 shows the top cables for the East Timor invasion. Capsule accurately identifies this real-world event and recovers relevant cables.

The second tallest peak occurs the week of April 21, 1975, just prior to the fall of Saigon on April 30, 1975; Table 3 shows the top cables for this event, which reflect the evacuation efforts that occurred during that week. Unlike the East Timor event, where the most relevant communication exists at an administrative level, the evacuation of Saigon is best captured by individuals seeking help for family and friends.

Another event peaks occurs the week of July 2, 1973; the top three words under event its description $\pi$ are *bicentennial*, *hijack*, and *mercenary*. Top cables under event relevancy $\epsilon$ surround the bicentennial celebration of United States (July 4, 1973) and the Air France hijacking incident that began on June 27; Israeli operatives rescued hostages from this incident on July 4th.

[5]Our algorithm is batch–we consider each data point for every iteration. Modifying the algorithm to stochastically sample the data would reduce the time required to achieve an equivalent model fit.

[6]Appendix B contains an analogous figure on arXiv data, which shows that Capsule does not capture weekly events on data that does not contain real-world events at that resolution.

Capsule also identifies events with smaller peaks, such as the death of Mao Tse-tung. One of the top cables for this event is sent by Kissinger to all post with public affairs guidance:

> 1. Missions should avoid all speculation about the possible effects of the death of Chairman Mao on US-PRC relationships as well as impact on internal Chinese developments.
> 2. Official comments should be limited to the statements of top level administration officials, texts of which will follow by SEPTEL and wireless file.

In the other top cables for this event, embassies generally reported on press reactions and condolence or memorial ceremonies at their various locations.

Capsule helps discovers events which follow a chain of related incidents, though connecting these events is left to the investigator. For example, Capsule discovers an event the week the Sinai Interim Agreement was signed (September 4, 1975), but it also detects an event in mid-October 1975 about the hiring of observers and technicians for the Sinai peace keeping force. Associated with this second event is a cable from London to the State department entitled *FCO views on Syrian stance*:

> Since conclusion of Sinai II negotiations, FCO officials have expressed considerable interest in prospects for next US effort to promote Syrian negotiations with Israelis. ...

This cable and the sequence of events discovered by Capsule indicate that there is a longer lasting underlying situation. Capsule cannot capture every aspect of these larger sequences of events, but it can provide insight into key moments so that investigators can explore both short-lived events and long-lasting political situations.

In addition to events, Capsule can be used to explore the general themes of a corpus and entities' typical concerns. Examples of general topics of conversation are shown in Table 4 and entity-exclusive topics are shown in Table 5; these show us how en-

| $\epsilon$ | date | entity | subject |
|---|---|---|---|
| 0.124 | 1975-12-03 | State | President's talking point on Portuguese Timor |
| 0.115 | 1975-12-04 | State | Timor we are repeating FYI a DAO message |
| 0.112 | 1975-12-04 | State | Legal problems relating to Portuguese Timor |
| 0.105 | 1975-12-04 | Secretary Peking | US Support for Timor resolution |
| 0.102 | 1975-12-07 | State | Invasion of Portuguese Timor |

**Table 2:** Top documents for the time interval of week December 1, 1975, just prior to the Indonesian invasion of East Timor, which began December 7, 1975; Capsule recovers relevant documents related to this real-world event.

| $\epsilon$ | date | entity | subject |
|---|---|---|---|
| 0.090 | 1975-04-24 | Mansfield, Mike | Assistance in evacuating family from South Vietnam |
| 0.089 | 1975-04-24 | Railsback, Tom | Assistance in evacuating friend from South Vietnam |
| 0.086 | 1975-04-24 | Koch, Edward | Assistance in evacuating family from South Vietnam |
| 0.086 | 1975-04-21 | Schweiker, Richard | Support in evacuating family from Vietnam |
| 0.081 | 1975-04-25 | Ketchum, William | Movement of South Vietnamese refugees to Guam |
| 0.080 | 1975-04-21 | Scott, Hugh | Whereabouts of missionaries in Vietnam |

**Table 3:** Top documents for the time interval of week April 21, 1975, just prior to the fall of Saigon on April 30, 1975; Capsule recovers relevant documents related to this real-world event.

| top terms |
|---|
| outlook, review, hire, personnel, invite, prepare |
| arrest, incident, security, family, guard, death, jail |
| locate, home, son, death, please, contact, father |
| request, refugee, response, service, sale, asylum |
| market, report, commercial, food, import, commerce |
| fear, leadership, back, arm, role, threaten |
| hotel, travel, reservation, visit, arrange, schedule |

**Table 4:** Top vocabulary terms for a selection of general topics, one per row, according to topic distributions $\beta_k$. Capsule identifies general diplomatic themes that can be relevant to any entity.

| entity | top terms |
|---|---|
| State | request, follow, embassy, meet, make |
| Bangkok | bangkok, thailand, thai, refugee, follow |
| Jerusalem | jerusalem, israeli, bank, report, say |
| Stockholm | swedish, sweden, trade, meet, embassy |
| Kampala | ugandan, nairobi, african, imperialist |
| Ndjamean | chadian, chad, lagos, drought, austerity |

**Table 5:** Top vocabulary terms for a selection of entities according to entity-exclusive topics $\eta_n$. Capsule identifies entity-specific themes and interests.

tity topics absorb location-specific words, preventing these terms from overwhelming the general topics.

Appendix B contains additional examples of events discovered by Capsule, and more examples of general and entity-specific topics.

These exploratory results show that our model is successfully capturing when multiple entities are discussing the same subjects and that our model can be used to explore the underlying data by providing a structured scaffold from which to view the data.

**Simulations.** We simulated data to provide a quantitative assessment of Capsule. We generated twenty data sets, each with 100 time steps, 10 general topics, and 100 entities. Each simulation contained about 55,000 documents and followed the generative process assumed by Capsule, as shown in Figure 3.

To evaluate event detection, we created a ranked list of all time intervals and computed the overlap between a method and the simulated ground at every threshold; this generates an curve under which we can compute the area and normalized based on ideal performance—we refer to this metric as event detection AUC.

The most successful of the baseline methods for event detection was average absolute error in word

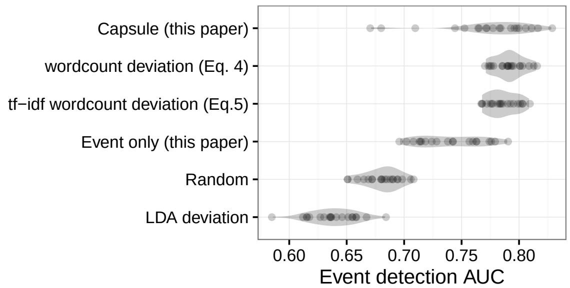


**Figure 4:** Event detection performance on twenty simulated datasets. Capsule is able to detect events as well as comparison methods, but its performance has higher variance.

count relative to the mean, or

$$\sum_{v=1}^{V}\left[\sum_{d=1}^{D}\text{abs}\left(w_{d,v}-\frac{1}{|D|}\sum_{d=1}^{D}w_{d,v}\right)\right], \qquad (4)$$

and its tf-idf variant,

$$\sum_{v=1}^{V}\text{tf-idf}(v)\left[\sum_{d=1}^{D}\text{abs}\left(w_{d,v}-\frac{1}{|D|}\sum_{d=1}^{D}w_{d,v}\right)\right]. \qquad (5)$$

Figure 4 shows that Capsule can outperform these approaches for event detection, but that it has higher variance in performance. We also consider an "event only" model—this is a model that only uses the interval-related subset of Capsule's parameters; comparing to this shows that is it important to model "business as usual" for improved event detection. LDA based approaches like average deviation from mean in topic space (Dou et al., 2012) do not perform well for event detection as deviations in topic space are too coarse to provide a meaningful signal.

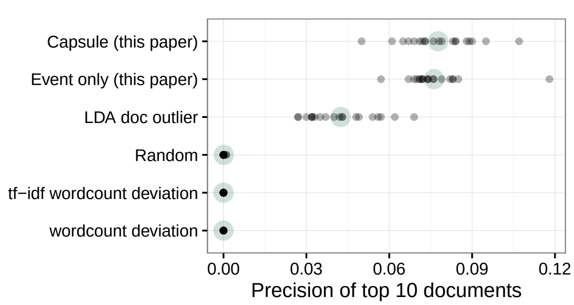


**Figure 5:** Precision of recovering the top ten most relevant documents, averaged over all time intervals. Capsule performs best, averaged over twenty simulations.

Once events have been identified, our next task is to identify relevant documents; to evaluate this, we calculate precision of recovering the top ten documents. LDA is useful in finding relevant documents by selecting documents that deviate from the mean in topic space. Word count deviations for each document (similar to Equations 4 and 5) perform close to random for document recovery.

Finding documents based on absolute deviation from the mean works better in LDA topic space, but /home/statler/achaney/cables/src/event_detect/scripts/sweep5linearv2/fitnot over the full vocabulary. Word count deviations, which performed well for event detection, performed worse than random for document recovery. Both Capsule and its event-only partial model outperform all comparison methods in terms of document recovery. Figure 5 shows precision of recovering the top ten documents.

We assessed the sensitivity of our model to three different decay functions $f$: exponential, linear, and step functions. We simulated data for each function and then fit Capsule using every permutation of $f$ and multiple settings for event decay duration. In all cases, we found that the model is not sensitive to decay shape or duration; details are in Appendix B.

**Comparisons on cables.** We ran the word count based approaches on the cables data and found that they were difficult to interpret and did not recover important historical events. For instance, none detected the evacuation of Saigon, a major historical event in the corpus. The LDA-based approaches do not yield large gains in run time over Capsule and do not provide the granularity needed to capture substantial events.

## 4 Conclusion

We have presented Capsule, a Bayesian model that identifies when events occur, characterizes these events, and discovers the typical concerns of author entities. We have shown that Capsule outperforms comparison methods and explored its results on a real-world datasets. We anticipate that Capsule can be used by historians, political scientists, and others who wish to investigate events in large text corpora.

## References


Ryan Prescott Adams and David JC MacKay. 2007. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*.

James Allan, Ron Papka, and Victor Lavrenko. 1998. On-line new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 37–45. ACM.

Hila Becker, Mor Naaman, and Luis Gravano. 2010. Learning similarity metrics for event identification in social media. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 291–300. ACM.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *JMLR*, 3:993–1022, March.

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

Thorsten Brants, Francine Chen, and Ayman Farahat. 2003. A system for new event detection. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 330–337. ACM.

John Canny. 2004. Gap: a factor model for discrete data. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 122–129. ACM.

Deepayan Chakrabarti and Kunal Punera. 2011. Event summarization using tweets. *ICWSM*, 11:66–73.

Kaustav Das, Jeff Schneider, and Daniel B Neill. 2008. Anomaly pattern detection in categorical datasets. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 169–176. ACM.

Anish Das Sarma, Alpa Jain, and Cong Yu. 2011. Dynamic relationship and event discovery. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 207–216. ACM.

Wenwen Dou, Xiaoyu Wang, Drew Skau, William Ribarsky, and Michelle X Zhou. 2012. Leadline: Interactive visual analysis of text data through event identification and exploration. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pages 93–102. IEEE.

Gabriel Pui Cheong Fung, Jeffrey Xu Yu, Philip S Yu, and Hongjun Lu. 2005. Parameter free bursty events detection in text streams. In *Proceedings of the 31st international conference on Very large data bases*, pages 181–192. VLDB Endowment.

Wei Gao, Peng Li, and Kareem Darwish. 2012. Joint topic modeling for event summarization across news and social media streams. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1173–1182. ACM.

Prem K Gopalan, Laurent Charlin, and David Blei. 2014. Content-based recommendations with poisson factorization. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3176–3184. Curran Associates, Inc.

Valery Guralnik and Jaideep Srivastava. 1999. Event detection from time series data. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 33–42. ACM.

Alan G Hawkes. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90.

Alan Jackoway, Hanan Samet, and Jagan Sankaranarayanan. 2011. Identification of live news events using twitter. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, pages 25–32. ACM.

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

Giridhar Kumaran and James Allan. 2004. 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*, pages 297–304. ACM.

Jey Han Lau, Nigel Collier, and Timothy Baldwin. 2012. On-line trend analysis with topic models:\# twitter trends detection topic model online. In *COLING*, pages 1519–1534.

Zhiwei Li, Bin Wang, Mingjing Li, and Wei-Ying Ma. 2005. A probabilistic model for retrospective news event detection. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 106–113. ACM.

Scott W Linderman and Ryan P Adams. 2014. Discovering latent network structure in point process data. *arXiv preprint arXiv:1402.0914*.

Xueliang Liu, Raphaël Troncy, and Benoit Huet. 2011. Using social media to identify events. In *Proceedings of the 3rd ACM SIGMM international workshop on Social media*, pages 3–8. ACM.

Michael Mathioudakis, Nilesh Bansal, and Nick Koudas. 2010. Identifying, attributing and describing spatial bursts. *Proceedings of the VLDB Endowment*, 3(1-2):1091–1102.

Daniel B Neill, Andrew W Moore, Maheshkumar Sabhnani, and Kenny Daniel. 2005. Detection of emerging space-time clusters. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 218–227. ACM.

Michael J Paul and Mark Dredze. 2012. A model for mining public health topics from twitter. *Health*, 11:16–6.

Wei Peng, Charles Perng, Tao Li, and Haixun Wang. 2007. Event summarization for system management. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1028–1032. ACM.

Timo Reuter and Philipp Cimiano. 2012. Event-based classification of social media streams. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, page 22. ACM.

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

Hassan Sayyadi, Matthew Hurst, and Alexey Maykov. 2009. Event detection and tracking in social streams. In *ICWSM*.

Aaron Schein, John Paisley, David M Blei, and Hanna Wallach. 2015. Bayesian poisson tensor factorization for inferring multilateral relations from sparse dyadic event counts. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1045–1054. ACM.

Courtland VanDam. 2012. A probabilistic topic modeling approach for event detection in social media. Master's thesis, Michigan State University.

Xuanhui Wang, ChengXiang Zhai, Xiao Hu, and Richard Sproat. 2007. Mining correlated bursty topic patterns from coordinated text streams. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 784–793. ACM.

Gary M Weiss and Haym Hirsh. 1998. Learning to predict rare events in event sequences. In *KDD*, pages 359–363.

Yi Zhang, Jamie Callan, and Thomas Minka. 2002. Novelty and redundancy detection in adaptive filtering. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 81–88. ACM.

Qiankun Zhao, Prasenjit Mitra, and Bi Chen. 2007. Temporal and information flow based event detection from social text streams. In *AAAI*, volume 7, pages 1501–1506.

Wayne Xin Zhao, Rishan Chen, Kai Fan, Hongfei Yan, and Xiaoming Li. 2012. A novel burst-based text representation model for scalable event detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 43–47. Association for Computational Linguistics.