World Wide Web (2017) 20:1527–1549
DOI 10.1007/s11280-017-0451-3

Hashtag-based topic evolution in social media

Md. Hijbul Alam[1] · Woo-Jong Ryu[1] · SangKeun Lee[1]

Received: 15 February 2016 / Revised: 28 February 2017 /
Accepted: 2 March 2017 / Published online: 13 March 2017

Abstract The rise of online social media has led to an explosion of metadata-containing user generated content. The tracking of metadata distribution is essential to understand social media. This paper presents two statistical models that detect interpretable topics over time along with their hashtags distribution. A topic is represented by a cluster of words that frequently occur together, and a context is represented by a cluster of hashtags, i.e., the hashtag distribution. The models combine a context with a related topic by jointly modeling words with hashtags and time. Experiments with real-world datasets demonstrate that the proposed models discover topics over time with related contexts effectively.

Keywords Topic evolution · Hashtag distribution · Topic model · Social media

1 Introduction

Over the past few years, social media platforms have undergone tremendous growth. In particular, the microblogging subgenre, the most notable example of which is Twitter, has become one of the most widely used communication tools and platforms for the broadcasting of opinions and breaking news, and for the discussion of issues. Microblogging therefore constitutes a valuable and dynamic source of insight regarding the public's views

✉ SangKeun Lee
yalphy@korea.ac.kr

Md. Hijbul Alam
hijbul@gmail.com

Woo-Jong Ryu
skdirwj@korea.ac.kr

[1] Department of Computer Science, Korea University, Seoul, Republic of Korea

and moods, providing politicians, business providers, social observers, and advertisers with significant opportunities. Examples of this prominent status include the heavy usage of Twitter as an open forum during the following events: the U.S. presidential elections of 2008 and 2012, the Tohoku tsunami in Japan in 2011, the political upheavals in the Middle East, and the Sydney cafe siege in 2015.

Social media trends form a very important component of exploratory data analysis and prediction, as they help in the assessment of the level of public interest regarding a given topic. Moreover, the knowledge regarding these trends improves the user experience with respect to information retrieval systems. In addition, it benefits time-sensitive online advertising, and aids the learning regarding social behavioral patterns, e.g., by predicting polls or box-office revenues. A key indicator of this significance is the practice of a large number of media sites such as Google,[1] Twitter,[2] and Foursquare,[3] whereby trends are highlighted through a "trending now" module on the home page.

Typically, a trending keyword is the result of the evolution of numerous topics that contain the keyword. The ability to successfully identify topics and track their evolution is therefore essential for a meaningful presentation of trends. The identification of the topics that are relevant to a given keyword is challenging, however, because of the large volume of associated tweets. For instance, from December 5, 2013, to December 20, 2013, thousands of tweets followed the death of Nelson Mandela.

To effectively identify relevant tweets, metadata can be used. Several types of metadata are available in social media, such as the associated short url, picture, @mention, and #hashtag. The hashtag, perhaps the most important type of metadata, plays several key roles. A hashtag is used to highlight themes, events or explicit topics within a tweet. The hashtag provides a simple method for people to categorize, find, and join conversations on a particular topic. Typically, users introduce many hashtags to a trending topic to provide additional perspectives and recent updates. For instance, tweets related to Nelson Mandela contained hashtags such as #qunu, #humanrightsday, #morganfreeman, and #worldcupdraw, as Twitter users linked other concurrent events and perspectives to Nelson Mandela. These tags were then helpful in an exploration of the topics that evolved from Nelson Mandela. Motivated by scenarios such as this, we have hypothesized that metadata, specifically hashtags, captures the reason why topics evolve.

In social media, metadata changes over time and affects the composition of topics. It is therefore important to extract the metadata distribution and track its evolution. In this regard, contexts for the solving of a problem are defined in a number of interesting works [13, 31], and a context can therefore be defined in many ways. In this paper, we define the distribution of metadata, including hashtags, users, hyperlinks, images, and entities, but not time, as a context. The proposed models show evolutions over time in terms of word distributions (i.e., topics) over time and relate a metadata distribution with a word distribution. We observed that hashtags deliver more semantics than other types of metadata, e.g., user names or hyperlinks. We have therefore explicitly examined hashtag distributions with topics over time, whereby Context over Time (COT) and COT+, which model the generation processes of text, metadata and time, are proposed. COT and COT+ use continuous time distributions and discrete time distributions, respectively.

[1] http://www.google.com/trends/

[2] https://twitter.com/

[3] https://foursquare.com/

In this paper, we extend our preliminary work [3] in three ways. First, we propose COT+, which adopts a discrete distribution over time, and it outperforms the other models in terms of running time and effectiveness. Second, we propose two extended models (i.e., C^2OT and C^2OT+) and show the efficacy of the modeling of hashtag distribution with a topic over time. Third, we use two applications, timeline generation and important-event discovery [16], to demonstrate the potentials and efficacy of the proposed models. The contributions of this paper are as follows:

- We propose COT and COT+, both of which effectively model hashtag based topic distribution over time.
- We demonstrate the efficacy of COT and COT+ qualitatively by showing that hashtag-based topic distributions over time increase the interpretability of social media trends.
- We show that COT+ and the extended models (i.e., C^2OT and C^2OT+) outperform the baseline approach in the discovery of timeline tweets and important events.

The remainder of this paper is organized as follows. Section 2 describes related works. Sections 3 and 4 present our schemes and experimental results, respectively. Section 5 concludes this work.

2 Related works

In the past decade, topic modeling has attracted numerous research efforts from both academia and industry. Latent Dirichlet Allocation (LDA) [6], and the probabilistic Latent Semantic Analysis (pLSA) [12] are the foundations of topic models. In these schemes, a topic consists of a cluster of words that frequently co-occur and is represented as a multinomial distribution over words. Given a corpus, a topic model can distinguish words with different semantic meanings and extract hidden topics. Rosen-Zvi et al. [27] extended the LDA model by adding the author's relationship to articles and proposed the author-topic model. As further extensions, McCallum et al. [19] proposed the author-recipient-topic-model. Lin et al. [17], Alam et al. [2], and Rao et al. [25] extended the LDA model by considering the sentiment embedded in online reviews and news. All of these models are static, however, and ignore the dynamic nature of topics.

COT departs from Labeled LDA [24] since COT is an unsupervised model, whereas Labeled LDA is a supervised one. In the Labeled LDA, it is assumed that a topic is manually associated with a label or a hashtag, and a document has a series of labels associated with it. On the other hand, COT discovers the related hashtags of a topic automatically.

Beyond the static topic modeling, Wang et al. [34] designed TOT model that associates topics with a continuous distribution over time. Similarly, Blei et al. [5] proposed Dynamic Topic Model, and Chua et al. [8] proposed Decay Topic Model by using a Gaussian time series or distribution. AlSumait et al. [4] and Lau et al. [15] presented online LDA models for text streams. In Mei et al.'s work [21], topic evolutions are computed by using the Kullback-Leibler (KL)-divergence of topics that are discovered within each interval. Ahmed et al. [1] proposed "Storyline" by incorporating the Recurrent Chinese Restaurant Process (RCRP) into LDA model. Storyline can simultaneously group articles into storylines and identify prevalent topics. Notably, however, metadata is not exploited for any of these models.

Multi-contextual LDA [31] incorporates several types of metadata in a unified framework. Mehrotra et al. [20] proposed different pooling schemes such as hashtag, burst score, and time to improve LDA for microblogs. He et al. [11] made use of citations. Qian et al.

[23] explored topic formation and evolving process. Zhou et al. [35] focused on hot topics extraction from blog news, and Si et al. [29] focused on grouping users' interest from online reviews. Notably, however, an explicit consideration of the relations of the metadata distribution of evolving topics, which is critical for modeling social media, has not been taken for any of these models.

Our goal is to extract important events, explicit topics, or themes that are related to latent topics defined by hashtags. In contrast, Trend Analysis Model (TAM) [14] extracts a trend class which can be interpreted as summarization or grouping of several latent topics. Topic User Time [32] is an extension of TAM that considers user information in generating documents and produces the ranking of users.

TOT as well as the proposed models are not dynamic topics model (DTM) as it is defined by Blei et al. [5]. In DTM, distribution of words of a topic changes over time which can be treated as a new topic. Therefore, DTM aligns the topics over time. In TOT and the proposed models, however, the distribution of words of a topic do not change over time. Therefore, new topics are not generated in TOT as well as in COT and COT+. Only time varies over the distribution of words.

3 Generative models

Our motivation is the provision of the hashtag distribution along with the topic over time, and a demonstration of an improvement of the interpretability of social media by considering the hashtag based topic evolution. We describe two probabilistic graphical models for the generation of context-aware, dynamic document collections utilizing hashtags. We observe that hashtags play dual roles in a tweet. A hashtag appears in a tweet as a word (i.e., as a part of a sentence like an anchor text), while describing an explicit topic as well. The number of hashtags appearing in the social media is huge, and exploring a large number of hashtags is challenging. Therefore, we extract the cluster of hashtags as context. In addition, the top words of a topic can be described clearly with a hashtag if they come from the same word distribution. Therefore, we include hashtags in the definition of both topic and context, and formally define the topic and context as follows:

Definition 1 (Topic) A topic φ is defined as a multinomial distribution of the words and hashtags including all types of metadata in the vocabulary V (i.e., $\{p(w|\varphi)\}_{w \in V}$).

LDA models each document as a mixture of the underlying topics and discovers the topics. TOT [34] associates the topics with continuous distributions.

Definition 2 (Context) A context ϕ is defined as a multinomial distribution of metadata only, such as hashtags, in the metadata vocabulary U (i.e., $\{p(c|\phi)\}_{c \in U}$).

Given a text collection D from social media and a predefined number K, COT discovers K hashtag distributions (i.e., contexts) and K topics that are equipped with time-continuous (λ) distributions. Similarly, COT+ discovers K hashtag distributions and K topics that are equipped with time-discrete (ψ) distributions. In this research, therefore, we attempt to extract a context (ϕ_z) (i.e., themes, events, or explicit topics defined with hashtags) that can be related to a latent topic (φ_z). Top words of a latent topic may reveal few hashtags that only point to few themes. However, through the proposed model, we can extract many themes that are related to a latent topic. Therefore, a context can be considered as a collection

of themes, events, or explicit topics defined by users using the cluster of hashtags that are related to a latent topic.

3.1 Context over Time

In online social media, each document or tweet is typically very short. Many tweets contain hashtags to overcome the length restriction or to provide additional perspectives. For instance, tweets related to Nelson Mandela contain hashtags such as #qunu, #humanrightsday, and #worldcupdraw. We have assumed that the topics were influenced by temporal information, word co-occurrence, and the presence of hashtags. The timestamping of the COT model involves the parameterization of the continuous distribution over time that is associated with each topic.

TOT treats hashtags as words and draw them from one distribution. In contrast with TOT, COT utilizes the probability of a hashtag from both distributions (i.e., φ_z and ϕ_z) with index z to generate hashtags since a hashtag is in V and U. As a result, a context is aligned and related with a topic.

Regarding topics for each document, COT draws a multinomial distribution using a Dirichlet distribution with a prior α. Topic assignment is selected from this topic distribution. Next, a hashtag is generated by randomly sampling from ϕ_z. However, a word is generated from a multinomial distribution over words φ_z. Finally, a timestamp is drawn for each word or hashtag from the beta distribution. A beta distribution over the timestamps, a hashtag distribution over the hashtags, and a term distribution over the vocabulary for each topic are each sampled once for the entire corpus. Figure 1 shows the graphical COT model, and the generative process of the model is given below (refer to Table 1 for parameters):

1. (a) Draw K word distributions $\varphi_z \sim Dir(\beta)$.
   (b) Draw K hashtag distributions $\phi_z \sim Dir(\gamma)$.
2. For each document d,
   (a) Draw a distribution of topics $\theta_d \sim Dir(\alpha)$.
   (b) For each hashtag c,
      i. Draw a topic $z_{di} \sim \theta_d$.
      ii. Draw a hashtag $c_{di} \sim \phi_z$.
      iii. Draw a timestamp $t_{di} \sim Beta(\lambda_z)$.

Figure 1 COT

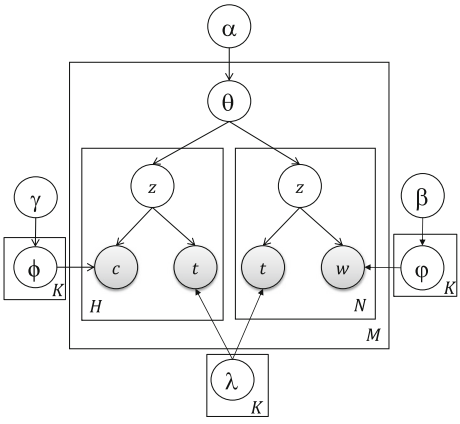


Table 1 Parameters of COT and COT+

M	number of documents
N	number of words in a document
H	number of hashtags in a document
K	number of topics
T	number of timestamps
V	number of unique words
C	number of unique hashtags
w, z, d	word, topic, and document, respectively
c, t	hashtag and time, respectively
φ	multinomial distribution over words
ϕ	multinomial distribution over hashtags
θ_d	multinomial distribution over topics
ψ	categorical distribution over timestamps
λ_z	Beta distribution over timestamp for topic z
α	Dirichlet prior for θ
β	Dirichlet prior for φ
γ	Dirichlet prior for ϕ
δ	Dirichlet prior for ψ

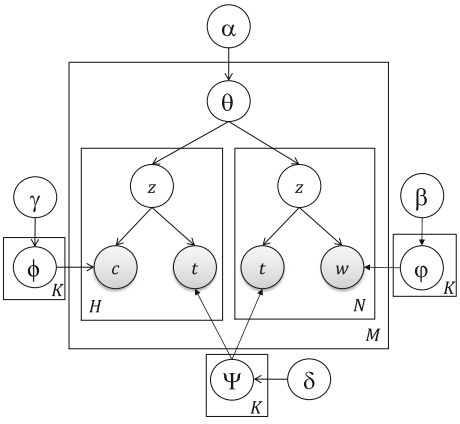
(c) For each word w,

i. Draw a topic $z_{di} \sim \theta_d$.
ii. Draw a word $w_{di} \sim \varphi_z$.
iii. Draw a timestamp $t_{di} \sim Beta(\lambda_z)$.

3.2 Context over Time+

We parameterized the timestamps with the discrete distribution that is associated with each topic, since sampling from a beta distribution is computationally expensive. We called the developed model "COT+". In contrast with COT, COT+ draws a timestamp from a

Figure 2 COT+



multinomial distribution over timestamps that is topic-specific. The timestamp multinomial distribution for each topic is sampled once for the entire corpus. The graphical COT+ model is shown in Figure 2, and the generative process of the model is given below (refer to Table 1 for parameters):

1. (a) Draw K word distributions $\varphi_z \sim Dir(\beta)$.
   (b) Draw K hashtag distributions $\phi_z \sim Dir(\gamma)$.
   (c) Draw K timestamp distributions $\psi_z \sim Dir(\delta)$.
2. For each document d,
   (a) Draw a distribution of topics $\theta_d \sim Dir(\alpha)$.
   (b) For each hashtag c,
      i. Draw a topic $z_{di} \sim \theta_d$.
      ii. Draw a hashtag $c_{di} \sim \phi_z$.
      iii. Draw a timestamp $t_{di} \sim \psi_z$.
   (c) For each word w,
      i. Draw a topic $z_{di} \sim \theta_d$.
      ii. Draw a word $w_{di} \sim \varphi_z$.
      iii. Draw a timestamp $t_{di} \sim \psi_z$.

3.3 Inference

To estimate the hidden parameters of the COT model in (1) and (2), we apply the collapsed Gibbs sampling in accordance with updated rules. Here, a hashtag is sampled by (1) and a word is sampled by (2).

$$P(z_{d,i} = z|\mathbf{z}', \mathbf{w}, \mathbf{t}, \mathbf{c}) \propto \frac{n^z_{w_{d,i}} + \beta}{n^z + V\beta} \times \frac{n^{z_h}_c + \gamma}{n^{z_h} + C\gamma} \times \frac{n^d_z + \alpha}{n^d + M\alpha} \times \frac{(1-t_d)^{\lambda_{z1}-1} t_d^{\lambda_{z2}-1}}{B(\lambda_{z1}, \lambda_{z2})} \tag{1}$$

$$P(z_{d,i} = z|\mathbf{z}', \mathbf{w}, \mathbf{t}) \propto \frac{n^z_{w_{d,i}} + \beta}{n^z + V\beta} \times \frac{n^d_z + \alpha}{n^d + M\alpha} \times \frac{(1-t_d)^{\lambda_{z1}-1} t_d^{\lambda_{z2}-1}}{B(\lambda_{z1}, \lambda_{z2})} \tag{2}$$

For COT+, the updated rules for the collapsed Gibbs sampling are given in (3) and (4), where a hashtag is sampled by (3) while a word is sampled by (4).

$$P(z_{d,i} = z|\mathbf{z}', \mathbf{w}, \mathbf{t}, \mathbf{c}) \propto \frac{n^z_{w_{d,i}} + \beta}{n^z + V\beta} \times \frac{n^{z_h}_c + \gamma}{n^{z_h} + C\gamma} \times \frac{n^d_z + \alpha}{n^d + M\alpha} \times \frac{n^{t_d}_z + \delta}{n^z + T\delta} \tag{3}$$

$$P(z_{d,i} = z|\mathbf{z}', \mathbf{w}, \mathbf{t}) \propto \frac{n^z_{w_{d,i}} + \beta}{n^z + V\beta} \times \frac{n^d_z + \alpha}{n^d + M\alpha} \times \frac{n^{t_d}_z + \delta}{n^z + T\delta} \tag{4}$$

Here, n^z_w is the number of times word w is assigned to topic z, and n^z is the number of words assigned to topic z. $n^{z_h}_c$ is the number of times hashtag c is assigned to topic z, and

n^{z_h} is the number of hashtags assigned to z in the collection. n_z^d is the number of words assigned to topic z in document d, $n_z^{t_d}$ is the number of times timestamp t_d is assigned to topic z, and n^d is the number of words in document d.

A hashtag is in vocabulary V as well as metadata vocabulary U. φ_z is a distribution over V, while ϕ_z is a distribution over U. Both distributions play roles in generating a hashtag. In other words, the probability of a hashtag in both distributions is used for the same topic index z. In (1) and (3), the first term corresponds to φ_z and the second term corresponds to ϕ_z, where w_{di} and c refer to the same hashtag. It ensures that a hashtag as both a word and a metadata, and each hashtag distribution (i.e., context) is aligned and related to a word distribution (i.e., topic).

Timestamp for each hashtag comes from the distribution over time of each topic since a hashtag is in V. Moreover, the distribution over time associated with each topic is also aligned with the hashtag distribution, since a hashtag distribution is aligned with a topic. Therefore, we did not model or derive the distribution over time separately for hashtag distributions.

Algorithm 1 and Algorithm 2 describe the collapsed Gibbs sampling procedures of COT and COT+, respectively. By using a sample produced by a Gibbs sampling procedure, the parameters φ, ϕ, θ, and ψ can be estimated as follows:

$$\varphi = \frac{n_w^z + \beta}{n^z + V\beta}, \tag{5}$$

$$\phi = \frac{n_c^{z_h} + \gamma}{n^{z_h} + C\gamma}, \tag{6}$$

$$\theta = \frac{n_z^d + \alpha}{n^d + M\alpha}, \tag{7}$$

$$\psi = \frac{n_z^{t_d} + \delta}{n^z + T\delta}. \tag{8}$$

Algorithm 1 Gibbs sampling procedure of COT

begin
01: Initialize Gibbs sampler;
02: **for** (i = 1 to max Gibbs sampling iterations) **do**
03: **for** (each document $d \in [1, M]$) **do**
04: **for** (each word $w \in [1, N_d]$) **do**
05: Exclude w, c and t associated with z from variables
05: $n^z, n_w^z, n_c^{z_h}, n^{z_h}, n_z^d, n^d, n_z^{t_d}$;
06: If a word w is a hashtag:
07: sample a new topic z using Eq. (1);
08: Else: sample a new topic z using Eq. (2);
09: Update variables $n^z, n_w^z, n_c^{z_h}, n^{z_h}, n_z^d, n^d, n_z^{t_d}$
09: using the new z;
10: **end for**
11: **end for**
12: **end for**
end

Algorithm 2 Gibbs sampling procedure of COT+

begin
Initialize Gibbs sampler;
for (i = 1 to max Gibbs sampling iterations) **do**
  for (each document $d \in [1, M]$) **do**
    for (each word $w \in [1, N_d]$) **do**
      Exclude w, c and t associated with z from variables
      $n^z, n_w^z, n_c^{z_h}, n^{z_h}, n_z^d, n^d, n_z^{t_d}$;
      If a word w is a hashtag:
        sample a new topic z using Eq. (3);
      Else: sample a new topic z using Eq. (4);
      Update variables $n^z, n_w^z, n_c^{z_h}, n^{z_h}, n_z^d, n^d, n_z^{t_d}$
      using the new z;
    end for
  end for
end for
end

4 Experimental results

Twitter contains a large number of hashtags, as described in Section 4.1. Therefore, analyzing hashtags is a challenging task. The proposed models cluster related hashtags using the latent topics extracted from data to comprehend and visualize the themes or events in a meaningful way.

We demonstrate two application scenarios, including context visualization and diverse timeline tweets generation, to demonstrate the efficacy of the proposed models. Social media analytics will benefit from the applications to systematically explore and analyze the large volume of hashtags. Context visualization task in Section 4.2 shows that we can effectively explore important events through the cluster of hashtags. A popular approach to visualize tweets is a timeline. The efficacy of a timeline depends on the ability to show diverse important events. Timeline generation task in Section 4.3 shows the efficacy of selecting diverse tweets to extract important events from the top ranked tweets, where different models are evaluated quantitatively. For simplicity, we fixed the number of topics, K=30, in all of the cases. We used symmetric Dirichlet distributions (α=1.0, β=0.01, γ=0.01, δ=0.1) that are similar to TOT in all of our experiments. The topics were extracted from a single sample at the 500th iteration of the Gibbs sampler.

4.1 Dataset

We used the following three datasets for our experiments: the sinking of the Sewol ferry, the death of Nelson Mandela, and football. The properties of the datasets are given in Table 2. From the dataset, we observed that there are a large number of unique hashtags, and therefore, the related hashtags could be clustered to comprehend and visualize the themes or events expressed by hashtags. We used an open-source toolkit [28] to detect the languages of the collected tweets, but we considered only the English-language tweets.

Table 2 Dataset properties

Dataset	Total tweets	Unique words	Unique hashtags	Unique users	Total words
Sewol ferry	239,117	8,992	723	95,610	1,790,226
Nelson Mandela	2,813,461	255,298	50,425	936,022	18,776,257
Football	3,000,000	675,144	90,660	1,287,918	24,680,824

4.1.1 Sinking of the Sewol ferry

On April 16, 2014, the Sewol ferry capsized en route from Incheon, Korea, to Jeju Island[4] while it was carrying 476 people. Using the Twitter streaming API, we collected tweets posted between April 17, 2014, and May 20, 2014, regarding the keywords "ferry" and "#prayforsouthkorea".

4.1.2 Death of Nelson Mandela

Nelson Mandela was an anti-apartheid revolutionary who served 27 years in prison and dismantled the legacy of apartheid while he was president of South Africa. He died on December 5, 2013, at the age of 95 years. Approximately 90 representatives of foreign states traveled to South Africa to attend the memorial events.[5] Using the Twitter streaming API, we collected tweets posted between December 6, 2013, and January 5, 2014, regarding the keyword "Nelson Mandela".

4.1.3 Football

We also used a recurrent event to experiment with the proposed models. Using the Twitter streaming API, we collected tweets posted between May 19, 2015, and June 17, 2015, regarding the keyword "football". Only the first three million tweets were included for our analysis.

4.1.4 Number of unique users over time

We plotted the number of unique users for different days for the Sewol-ferry dataset in Figure 3. We observed that the number of unique users is greater than 10,000 for the first six days, and for another 10 days until day 15, the number of unique users is greater than 1,000. That is, the number of unique users decreases with the passing of time. We also plotted the number of unique users for different days for the Nelson Mandela dataset in Figure 4, wherein a large number of unique users is apparent. Figure 5 shows the number of users over time for the football dataset as an example of a recurrent event. Here, we observed that the number of unique users does not decrease over time.

[4] http://en.wikipedia.org/wiki/Sinking_of_the_MV_Sewol

[5] http://en.wikipedia.org/wiki/Nelson_Mandela

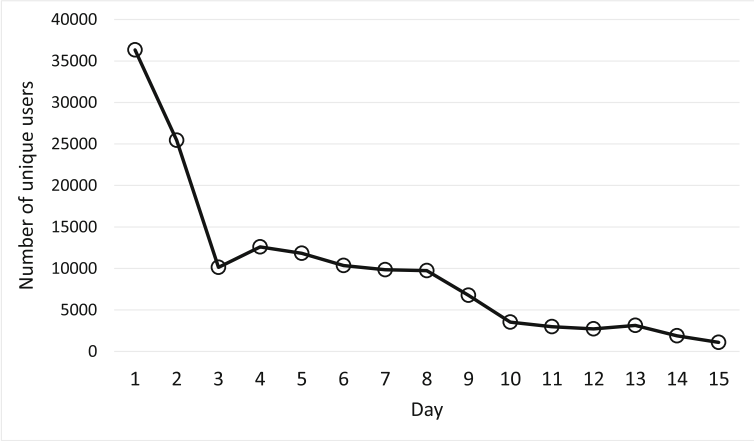


Figure 3 Number of unique users per day that results the generation of real trends in the Sewol ferry dataset

4.2 Context visualization

Context visualization experiment shows that the top words of topics failed to visualize many important events and explicit topics, whereas contexts are effective to extract and explore them.

We now discuss contexts and topics discovered by COT. As shown in Figure 6, we discuss three of the 30 topics. Each topic is illustrated with (a) a histogram of the topic distribution over time, (b) the top 20 words in each topic, and (c) the top 10 hashtags concerning the topic that are the most likely to be generated.

COT captures the topic and context, which are computed using (5) and (6), respectively. We manually labeled three topics in Figure 6. We observed that, in some cases, the top 20

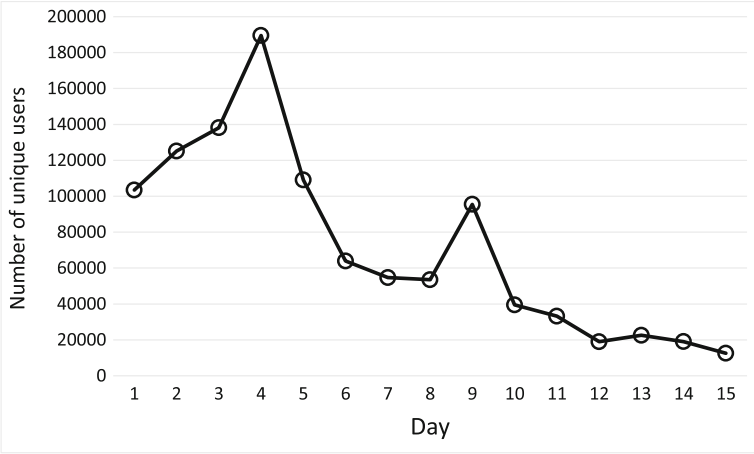


Figure 4 Number of unique users per day that results in the generation of real trends in the Nelson Mandela dataset

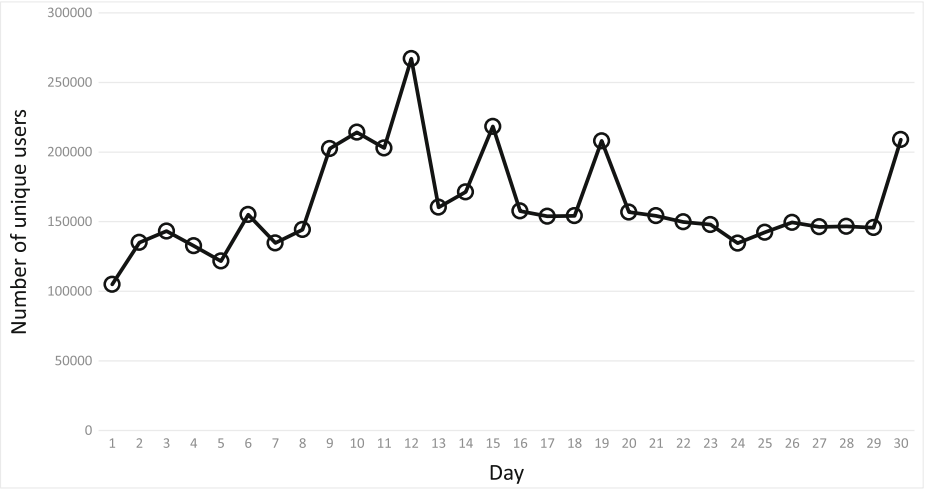


Figure 5 Number of unique users per day in the football dataset

words of a topic reflect the messages of one or two tweets. For example, the first topic shows information regarding a boy who saved his 5-year-old sister. The second topic describes the victims as mostly being from Danwon High School, and the third topic shows the analogy between the Titanic and the Sewol.

The contexts of social media are, however, created by multiple messages, emotions, or surrounding communities. For example, #heartbreaking and #sad reflect the emotions of the users of social media, and #yellowribbon describes a movement whereby tributes were paid to the victims through the display of yellow ribbons throughout Korea. For the third topic,

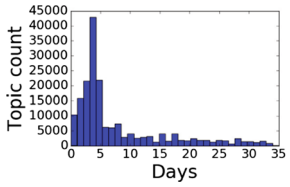
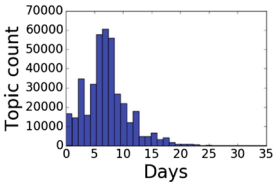
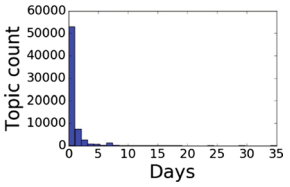
Brother saved sister		Danwon high school students		Titanic vs. Sewol	
					
Context	Topic	Context	Topic	Context	Topic
#hero #sad #news #love #life #chosen #samsung #hyunday #heartbreaking #bluehouse #seoul #yellowribbon #staystrongsouthkorea	disaster boy away save float sister sewol #hero 120m sunken victim help #sad little family #news sink out search	#sewol #rip #hope #tragedy #cnn #northkorea #fear #yellow #joongang #maywish #coastguard #tvxq #3rdwaveinfo #mayday #thailand	#sewol student body diver #rip school sunken high life dead vest discover say #hope kim sea passenger ship danwon hold	#ukiss #koreanferry #halamadrid #sink #video #pleasecomeback #np #sew #ktr #elfmemory #koreanferrydisaster #photo #breakingvideo	april titanic date year 2014 102 1912 month 15th tragedy difference but subhanallah tilt precious fishery ferryyonhap

Figure 6 COT topics, hashtag distributions (i.e., contexts), and topics' histograms over time for the Sewol ferry dataset. The histograms show the topic distributions over time, while the top hashtags along with the top words in each topic are shown below the histograms. Hashtags such as #BlueHouse, #YellowRibbon, and #Video reveal numerous important events, users' emotions in response to events, and the communities that support the events

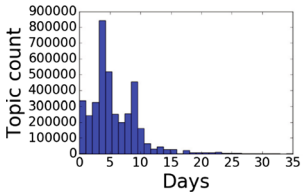
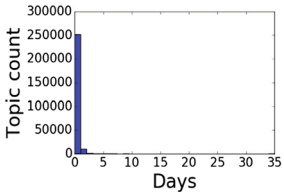
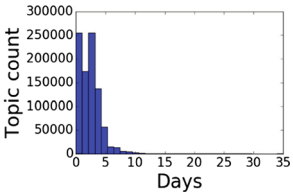
Mandela memory		Quote		Kanye West controversy	
					
Context	Topic	Context	Topic	Context	Topic
#mandelamemory #nelson #obama #rip #p2 #mufc #tfb #youngturk #peopleschampion #paulasly #madibamemory #art #ripnelsonma	memory south tribute africa obama president service funeral african photo sign pay state interpret new video watch body incredible language	#legend #morganfreeman #neverforget #36ng #worldcupdraw #allofu #quoteoftheday #brixton #reconcile #fifa	good impossible done terrorist heart watch list 2008 head formidable didnt rip did combination terror earth hill dick consider govern	#tyt #cdnpoly #sirkenayo #ucl #cancer #photography #humanrightsday #atheism #ssrugby #mandel #president #south #unionbuild	kany amp man said daughter compass wife hell west star porn say pornstar nigga drogba person fiance year support spend

Figure 7 COT topics, contexts, and topics' histograms over time for the Nelson Mandela dataset. Examples of the hashtags that refer to specific news or events (i.e., meaningful hashtags) are #WorldCupDraw, #HumanRightsDay

#video describes the conversation and surroundings recorded by a victim while the ferry was sinking. Overall, we can see many interesting hashtags in contexts that were not found in the top words of the topics.

Figure 7 shows the context and topic evolutions discovered through the Nelson Mandela dataset. Here, we also observed that the contexts uncovered interesting hashtags, and some of the hashtags are not in the top topic words. Hashtags such as #mandelamemory, #humanrightsday, and #worldcupdraw, which are mentioned here, reflect what happened during the period and reveal the users' motivations for generating the topics. After the death of Nelson

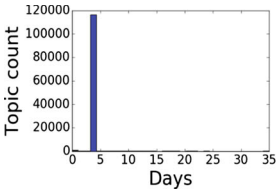
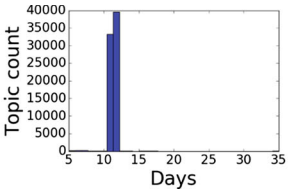
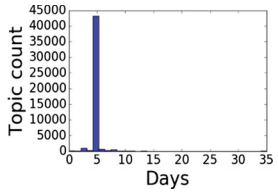
Brother saved sister		Danwon classroom picture, death toll		Rescue efforts	
					
Context	Topic	Context	Topic	Context	Topic
#hyunday #vixx #philippines #infographic #hopeformiracle #pleasecomeback #prayforsouthkorea #koreanferrydisaster #heatbreaking #dna	tragedy deaths toll sewol latest sister away saved 120m float dead body kst little boy update heavy industries	#sewol #danwon #koreaherald #sewolho #mustread #ferry #grief #empathy #Pray #info #yunakim #tears #koreaferrydisaster #koreatoday	#sewol tragedy kst school high deaths toll classroom victim #danwon latest students pictures heartbreaking 205 victims donated won	#news #live #obama #jtbc #3rdwaveinfo #yellowribbon #sadnews #pray #snsd #donating	#news missing victim sewol passengers koreans search diver #Live rescued continues sunken disaster family relative students news

Figure 8 COT+ topics, contexts, and topics' histograms over time for the Sewol ferry dataset. A discrete distribution over time has been used for the COT+ models. COT+ identified a number of day-specific topics, along with some meaningful hashtags such as #Obama and #YellowRibbon

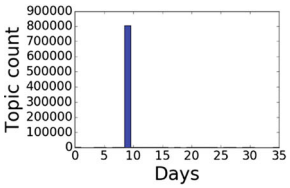
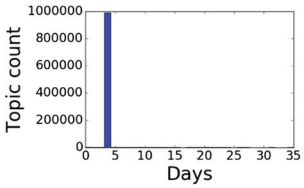
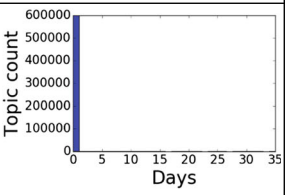
Funeral		Mandela Memory		Quote	
					
Context	Topic	Context	Topic	Context	Topic
#mandelafuneral #qunu #farewellmadiba #madibafuneral #bigroarformadiba #funeral #respect #mandelafinalfare well #goodbyemadiba #mandelasendoff	funeral south tribute africa rest qunu incredible laid bury home president #mandelafuneral live today ancestor rip watch african monuments world	#mandelamemory #madibamemory #humanrightsday #nelsonmandelame mory #championsleague #uniteblue #kingsmen #mandelamemory #barackobama #fnbstadium	memory obama service south president africa world #mandelamemory kany tribute man said live leader compass daughter african	#quote #ripnelsonmand #mtvstar #break #toronto #justice #lifechange #hivaid #wisdom #worldcupdraw	live time good amp remember not greatest celebrity thank let pass fall great legacy obama rip moment beauty heart generous

Figure 9 COT+ topics, contexts, and topics' histograms over time for the Nelson Mandela dataset. Some meaningful hashtags are #MandelaFinalFuneral, #FNBStadium, and #WorldCupDraw

Mandela, people around the world shared their memories of Mandela on social media with the hashtag #mandelamemory, which appears in the first topic. The final draw of the World Cup was held on December 6, 2013, and the revival of people's interest in Mandela was shown by sports-related Mandela quotes that were reflected through the #worldcupdraw and #quoteoftheday hashtags in the second topic. The third topic shows that people continued to deeply mourn Mandela four days after his death, on December 10, 2013, which is Human Rights Day.

TOT	#southkorea #ytn #korea #southkoreaferry #hero #hope #love #life #chosen #heartbreak #arirangnews #mh370 #rip #dearjesus #together
COT topics	#jindo #korea #hero #sad #news #obama #heartbreak #together #sewol #rip #hope #accident #sewol #sewolupdate #arirangnews #korean #southkorea #ytn #mh370 #dearjesus #ferrysinking #southkoreaferry
COT context	#yellowribbon #northkorea #yellow #expertsveiw #ripkarpalsingh #danwon #live #ansan #poem #heartbreaking #video #ripparkjiyoung #expert #ryu
COT+ topics	#hero #sewol #tragedy #heartbreak #restinpeace #southkorea #news #southkoreaferry #together #deathtoll #arirangnews #mh370 #rip #sad #danwon #hope #love #life #jindo #live #chosen
COT+ context	#dna #poem #yellow #coastguard #yellowribbon #video #northkorea #ripkarpalsingh #expertsveiw #ansan #obama #expert #ripparkjiyoung #ryu #sos #diver #bluehouse #15april #infographic

Figure 10 Meaningful hashtags, i.e., hashtags that refer to specific events or news in the Sewol ferry dataset

Table 3 Number of hashtags discovered in the top 20 words of each topic for different datasets, as computed by (5)

	Sewol ferry	Nelson Mendala	Football
TOT	15	17	52
COT	22	19	61
COT+	21	21	93

The COT+ model captures both the topics and contexts focused in time, shown in Figure 8, for the Sewol ferry dataset. The first topic is similar to COT's first topic. The second topic describes the death toll and a related picture of a room at Danwon High School. The third topic reports the rescue efforts.

Similarly to COT, COT+ contexts capture key messages, emotions, and surrounding communities. For example, #vixx and #snsd are some influential communities that are captured in the first topic and the third topic, respectively. The other tags, #pleasecomeback and #hopeformiracle in the first topic, reflect the emotions of the users conveyed on social media four days after the accident, whereby people hoped for miracles and prayed for the further rescuing of survivors. In the second topic, #KoreaHerald and #KoreaToday are the popular English news dailies that provided ferry-disaster updates. The third topic captures #yellowribbon and #Obama, the latter of which acknowledges the condolences expressed by U.S. President Barack Obama, as well as his visit to South Korea eight days after the tragedy.

Figure 9 shows the COT+ context and topic evolutions in the Nelson Mandela dataset. For example, Nelson Mandela's final funeral proceedings, which occurred nine days after his death, can be understood through the hashtag #mandelafuneral, or #mandelafinalfuneral, which appears in the first topic. The second topic is regarding the sharing of the memories of Nelson Mandela, and the third topic is about the revival of the public interest in Mandela quotes.

The difference between the topics of COT and COT+ is observable. COT models the rise, peak, and fall of topics with continuity, whereas a continuous trend does not apply for COT+. COT+ trends are instead discrete, and specific days where topics are discussed extensively are highlighted since COT+ models time based on a discrete distribution. For example, in Figure 9, the second topic is highlighted on Day 4, Human Rights Day, since people heavily revived Nelson Mandela's contributions to human rights, in addition to sharing memories of him.

Figure 10 and Table 3 show that COT+ and COT outperform TOT in terms of the discovery of hashtags that refer to a specific event or news, i.e., meaningful hashtags. We consider

Table 4 Diverse timeline tweets extracted from the top-ranked tweets

	Total top ranked tweets	Sewol Ferry		Nelson Mandela	
		No. of diverse timeline tweets	Percentage	No. of diversetweets timeline	Percentage
TOT	18,000	109	0.60	218	1.21
COT	18,000	120	0.66	217	1.20
COT+	18,000	227	1.26	846	4.70
C^2OT	6,000	242	4.03	326	5.43
C^2OT+	6,000	231	3.85	358	5.96

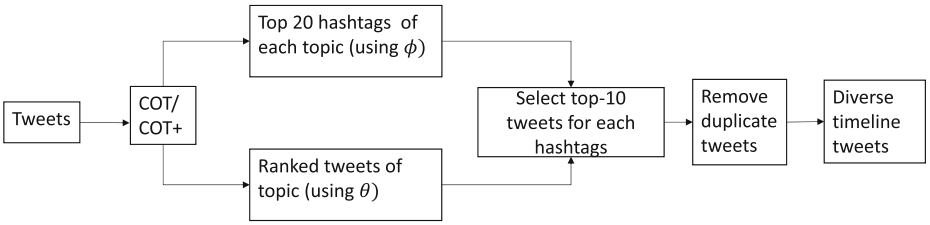


Figure 11 Diverse timeline tweets generated by the extended models, C^2OT/C^2OT+

a TOT that includes hashtag information and treats hashtags and words as a whole (i.e., with one Dirichlet prior) in all experiments. In other words, we have considered hashtags as words and draw them from one distribution with less parameter in the TOT. We curated meaningful hashtags in two ways. First, we considered all of the hashtags in the top 20 words of each topic that were computed by (5) as meaningful, since the hashtags directly appear in the topics. For example, TOT, COT, and COT+ discovered 15, 22, and 21 hashtags in the Sewol ferry dataset, respectively (as shown in Table 3). Second, we manually examined the top hashtags of each context that were computed by (6) and considered the hashtags that refer to a specific event or news as meaningful. The COT and COT+ contexts discovered at least 14 and 19 hashtags, respectively, in the Sewol ferry dataset (as shown in Figure 10).

We observed that our methodologies, COT and COT+, discover more meaningful hashtags than those discovered by TOT. It also validates our hypothesis that considering hashtags explicitly with a distribution (i.e., context ϕ), in addition to word distribution (i.e., topic φ), is very effective. It is worth noting that the automatic clustering of meaningful hashtags in the COT and COT+ contexts is challenging. We observed that the hashtags could be broadly categorized into meaningful, sentiment-oriented, and community hashtags. We leave the automatic clustering of hashtags into different categories as a future work.

4.2.1 Discussion

We have not handled the misspellings of hashtags here, as it is beyond the scope of the research. We plan to apply state-the-of-art, automatic spelling-correction techniques (e.g., [10]) or open-source spell checkers[6] to correct hashtag misspellings in the future. In addition, we have not performed a sentiment analysis in this research, and instead focused on the topics of each hashtag. For example, #heartbreaking refers to the news regarding a student's text message that was sent to his parents while the ferry was sinking, or the discovery of the dead body of the boy who saved his 5-year-old sister. We will investigate sentiment-oriented hashtags to analyze sentiment on Twitter in the future [7, 22, 26].

In topic modeling [6], the number of topics, K=30, is given as a parameter. Standard practices such as perplexity or log likelihood can be used to choose the number of topics. There is a nonparametric approach called "Hierarchical Dirichlet Process" [30] that can be used to determine the number of topics automatically. Similarly, a non-parametric approach to TOT has been proposed in [9]. A nonparametric approach to COT, i.e., a variable number of topics depending on the moment, is therefore an interesting research direction.

[6] http://hunspell.sourceforge.net/

Table 5 Number of ground-truth events extracted by different models in Sewol ferry dataset

TOT	31
COT	26
COT+	45
C^2OT	37
C^2OT+	50

4.3 Timeline generation

A growing interest in social media timeline generation has been identified. Recently, Li et al. [16] proposed a timeline-generation algorithm for individuals that discovers personalized important events, but it does not generate timelines for trending events. Motivated by this research, we generate a timeline tweets with important and diverse events, which are extracted from the top-ranked tweets for each topic. Selecting a large number of diverse tweets from the top-ranked tweets to generate a timeline provides a wide coverage of important events since the top-ranked tweets are the widely discussed tweets. We also evaluate the models quantitatively based on the diverse timeline tweets.

We generated the diverse timeline tweets by following three steps. First, we ranked the tweets for each topic based on the topic distribution, i.e., (7). Second, we selected the top 600 tweets for each topic. Third, we detected and removed any duplicate tweets from the top-ranked tweets [33]. Due to frequent re-tweets, a topic's top-ranked tweets tends to contain duplicates. In this experiment, we used the cosine similarity with a vector-space model to measure the similarity between two tweets. A tweet was selected if the cosine similarity between the tweet and the set of already selected tweets is less than a threshold of 0.1. Table 4 shows the numbers of diverse timeline tweets that were selected for both datasets. We observed that COT is comparable to TOT, while COT+ selected a greater number of diverse timeline tweets for a timeline than TOT.

The main advantage of the COT and COT+ models is the provision of contexts, i.e., a distribution of hashtags through which social emotions, active communities, key messages, events, and so on. We developed two new extended models, C^2OT and C^2OT+, based on the hashtag distribution, which are shown in Figure 11. We used the top-ranked hashtags for the diverse timeline tweets extraction by following four steps. First, we selected the top 20 hashtags computed for each topic by (6). Second, we ranked the tweets for each topic based on the topic distribution, i.e., (7). Third, for each hashtag we selected the top 10 tweets that contain the hashtag from the topic, meaning that we selected the top 200 tweets for each topic based on the context. Lastly, we removed the duplicate tweets from the top-ranked tweets according to the previously described procedure. We observed that both C^2OT and C^2OT+ outperform COT and TOT, as shown in Table 4.

COT- and COT+-based approaches are more effective than TOT regarding the discovery of important events and stories, since they select a greater number of diverse timeline tweets

Table 6 Relevant timeline tweets by different models in Sewol ferry dataset

TOT	43
COT	52
COT+	82
C^2OT	83
C^2OT+	78

Table 7 Precision, recall and F1-measure of timeline tweets relevance in Sewol ferry dataset

	Precision	Recall	F1-measure
TOT	0.39	0.15	0.21
COT	0.43	0.18	0.25
COT+	0.36	0.29	0.32
C^2OT	0.34	0.29	0.31
C^2OT+	0.37	0.27	0.29

than TOT. To verify this, we compiled the ground truth of the important events from the Sewol ferry dataset by manually examining the diverse timeline tweets extracted by all of the models. Therefore, we examined a total of 929 diverse timeline tweets, as shown in Table 4. We conducted this experiment on the Sewol ferry dataset, since the examiners closely followed the updates of the Sewol ferry tragedy over different media, meaning that it is possible to expertly judge whether a diverse timeline tweet refers to an important event.

For each approach, we manually examined whether a diverse timeline tweet corresponds to an important event. Because the evaluation of important events is a subjective task, four experts were asked independently to extract important events from the diverse timeline tweets. We applied pooling [18] for the selection of the ground-truth events from the diverse timeline tweets of each approach. Pooling is widely used for the evaluation of the relative performances of the different IR systems of a very large collection. The four experts extracted 77 events. Among them, 55 events were extracted by at least two experts, 32 events were extracted by at least three experts, and all of the experts extracted 11 events. The agreements regarding the ground-truth events by at least two experts, three experts, and four experts are 0.71, 0.41, and 0.14, respectively.

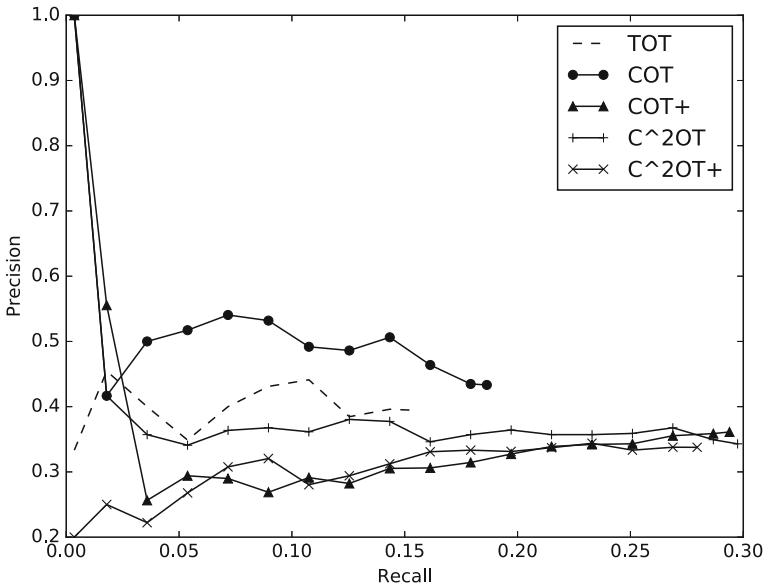


Figure 12 Precision-recall curve of timeline tweets relevance in the Sewol ferry dataset

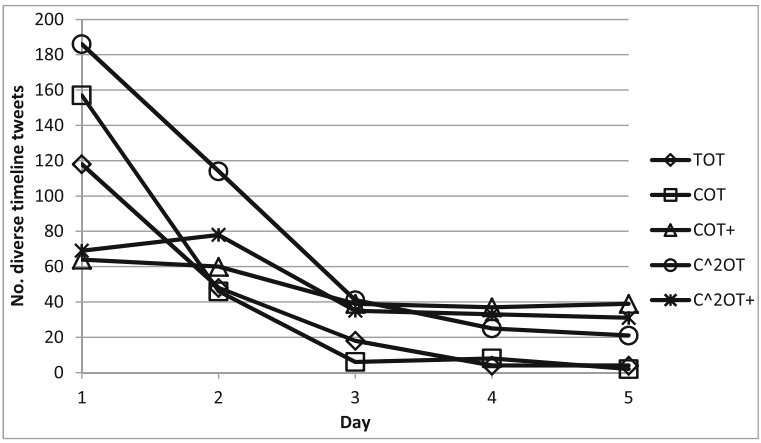


Figure 13 Number of diverse timeline tweets from day 1 to day 5 in the Sewol ferry dataset

We considered 55 events that were extracted by at least two experts as ground-truth. Table 5 shows the ground-truth events discovered by different models. We observed that COT+, C^2OT and C^2OT+ outperform TOT model.

We also evaluated the relevance of timeline tweets. We computed the precision and recall by manually checking whether a diverse timeline tweet is relevant to any of the ground-truth events. However, tweets that are related to emotions, donations, live updates, and show cancellation are not considered as relevant tweets for this experiment as there are many tweets that are related to this category. We measured the relevance of diverse timeline tweets retrieved by each model in the Sewol ferry dataset as shown in the third column of Table 4. Among 929 diverse timeline tweets, 279 tweets are labeled as relevant tweets. Table 6 shows relevant timeline tweets found in the experiments, wherein proposed models outperform TOT. Table 7 reports the precision, recall and F1-measures of timeline tweets relevance for different models, wherein proposed models outperform TOT in F1-measure.

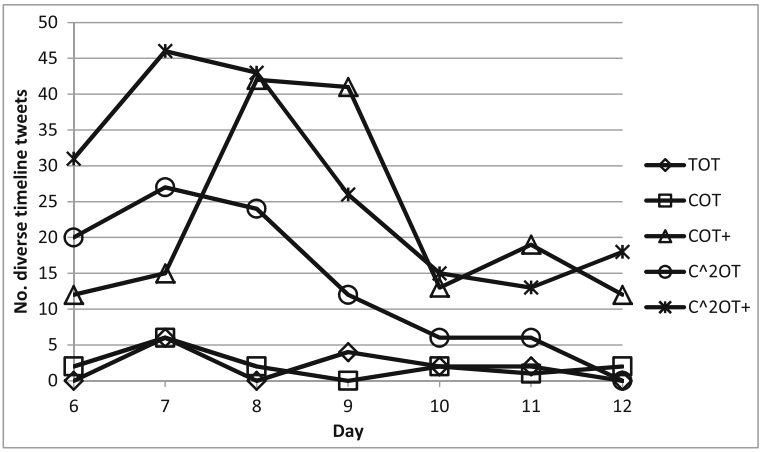


Figure 14 Number of diverse timeline tweets from day 6 to day 12 in the Sewol ferry dataset

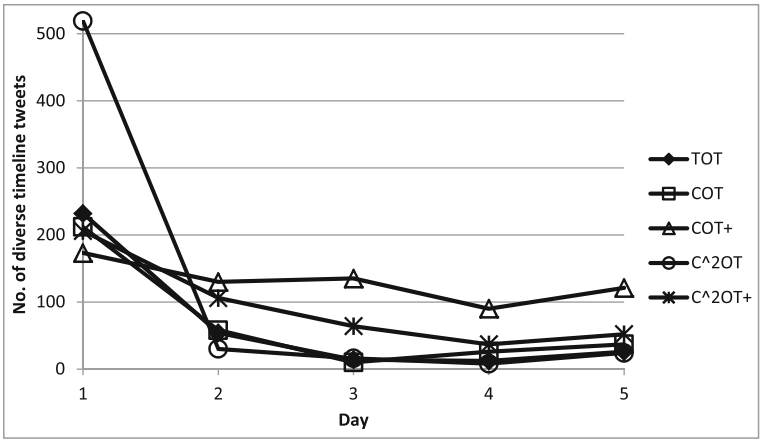


Figure 15 Number of diverse timeline tweets from day 1 to day 5 in the Nelson Mandela dataset

Figure 12 shows the precision-recall curve for timeline tweets relevance. We observed that COT outperforms TOT in the precision-recall curve.

We also experimented with diverse timeline tweets that are generated on distinct days. Typically, tweets are generated every day for a trending topic before the topic fades away. It is therefore highly probable that updated information will be incorporated for the diverse timeline tweets of distinct days. Figures 13, 14, 15, 16 show the number of diverse timeline tweets selected by the different approaches on different days. The behaviors of TOT and COT are similar in both datasets. C^2OT outperforms TOT in the Sewol ferry dataset but fluctuates in the Nelson Mandela dataset. COT+ and C^2OT+, however, outperform TOT for the selection of diverse timeline tweets on distinct days in both datasets. We therefore quantitatively verified that the modeling of both topic and context improves the interpretability of social-media trends.

We conducted experiments on the handling of recurrent events such as football. Interestingly, we observed that timelines of the extended models were more effective than the

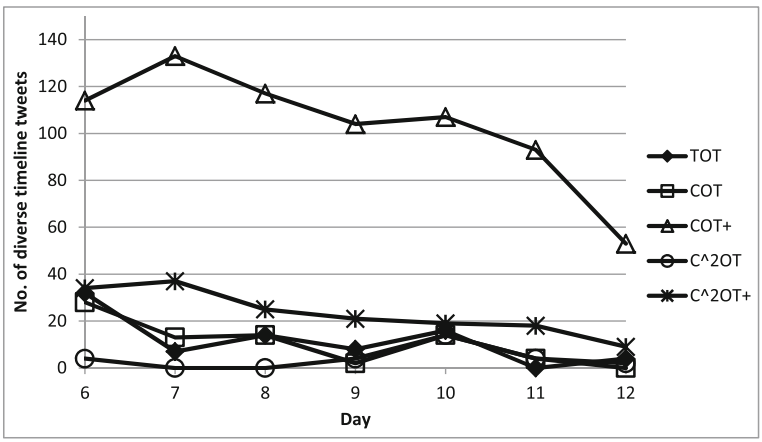


Figure 16 Number of diverse timeline tweets from day 6 to day 12 in the Nelson Mandela dataset

Table 8 Xavi-related stories discovered by different models

Stories	Models				
	TOT	COT	COT+	C^2OT	C^2OT+
Xavi won of all title in Football world				✓	✓
Xavi heading to Qatar					✓
Xavi's decisive roles in all victories				✓	✓
PSG to loan FCB legend Xavi from Al-Sadd				✓	✓
Xavi's departure from Barca		✓	✓	✓	✓
Xavi's quotes	✓	✓	✓	✓	✓

timelines of TOT, COT, and COT+ for handling of recurrent events. For example, C^2OT and C^2OT+ discovered more stories than TOT. For our analysis, we filtered diverse timeline tweets that contain "Xavi", a Spanish professional footballer, and found that TOT discovered two Xavi's quotes. COT and COT+ discovered that Xavi leaving Barca along with some Xavi's quotes. However, C^2OT and C^2OT+ discovered several stories in addition to the Xavi leaving Barca, as shown in Table 8. C^2OT and C^2OT+ outperform TOT and COT, since C^2OT and C^2OT+ extracted diverse timeline tweets using hashtags, whereas TOT and COT extracted diverse timeline tweets based on hidden topics. C^2OT and C^2OT+ discovered many Xavi related hashtags such as #GraciasXavi #6raciesXavi #XaviHernandez #XaviLegend, while COT and COT+ have discovered only #ThanksXavi hashtag. Although TOT discovered some hashtags, unfortunately, there was no Xavi related hashtag in the TOT diverse timeline tweets.

The running time of the different models is presented in Table 9. We observed that COT+ is very efficient and requires much less running time compared to the requirements of COT and TOT, since COT+ models timestamps discretely. In addition, it is worth noting that we do not model time with a hashtag distribution directly (e.g., λ_z for a topic). We model time with topics and align topics with hashtag distributions using a joint distribution. Therefore, modeling time with a discrete distribution better relates to the joint distribution than that of modeling time with a continuous distribution. As a result, we empirically observe that COT+ shows better performance than COT.

4.3.1 Discussion

Although we explicitly examined the hashtag distributions with topics over time in this work, we are not limited to hashtags in the modeling context. A high potential regarding another type of metadata or indicator, (e.g., user names, hyperlinks, user-based graphs, or other types of metadata) exists in terms of personalized timeline generation, clustering, and summarizing. The utilization of other indicators is left for a future work.

Table 9 Running time of inference algorithms in minutes

	Sewol ferry	Nelson Mandela
TOT	80	1061
COT	80	1074
COT+	11	158

5 Conclusion

In this paper, we have addressed the problem of extracting the hashtag distribution related to a topic over time in social media. We have proposed the hashtag-based topic-evolution models, Context over Time (COT) and COT+. Furthermore, we have built two extended models, C^2OT and C^2OT+. We have significantly improved the interpretability of social-media trends through the use of the proposed models. Interestingly, the proposed models have discovered numerous meaningful hashtags referring to specific events or news in hashtag distributions, which were originally not identified by latent topics alone. We have also shown that the extended models retrieve important events with high F1-measure, and effectively extract diverse timeline tweets. The conducted experiments with real-world datasets demonstrate that the joint modeling of hashtag distribution and topic over time is an effective approach. We plan to develop the sentiment-oriented COT in a distributed system, while an online and non-parametric version of COT is another future consideration. We will also investigate the automatic clustering of meaningful hashtags, sentiment-oriented hashtags, and community hashtags.

Acknowledgment This research was supported by the Basic Science Research Program and the Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (numbers 2015R1A2A1A10052665, 2015R1A2A1A15052701 and 2012M3C4A7033344).

References

1. Ahmed, A., Ho, Q., Eisenstein, J., Xing, E., Smola, A.J., Teo, C.H.: Unified analysis of streaming news. In: Proceedings of the 20th International Conference on World Wide Web (WWW), pp. 267–276 (2011)
2. Alam, M.H., Lee, S.: Semantic aspect discovery for online reviews. In: Proceedings of the 12th IEEE International Conference on Data Mining (ICDM), pp. 816-821 (2012)
3. Alam, M.H., Ryu, W.J., Lee, S.: Context over time: Modeling context evolution in social media. In: Proceedings of the 3rd Workshop on Data-Driven User Behavioral Modeling and Mining from Social Media (DUBMOD), pp. 15–18 (2014)
4. AlSumait, L., Barbara, D., Domeniconi, C.: On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In: Proceedings of the 8th IEEE International Conference on Data Mining (ICDM), pp. 3–12 (2008)
5. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of the 23rd International Conference on Machine Learning (ICML), pp. 113–120 (2006)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
7. Bravo-Marquez, F., Mendoza, M., Poblete, B.: Meta-level sentiment models for big social data analysis. Knowl.-Based Syst. **69**, 86–99 (2014)
8. Chua, F., Asur, S.: Automatic summarization of events from social media. In: Proceedings of the 7th International Conference on Weblogs and Social Media (ICWSM), pp. 81–90 (2013)
9. Dubey, A., Hefny, A., Williamson, S., Xing, E.P.: A nonparametric mixture model for topic modeling over time. In: Proceedings of the 13th SIAM International Conference on Data Mining, pp. 530–538 (2013)
10. Flor, M.: Four types of context for automatic spelling correction. Traitement Automatique Langues (TAL) **53**(3), 61–99 (2012)
11. He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., Giles, L.: Detecting topic evolution in scientific literature: How can citations help? In: Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM), pp. 957–966 (2009)
12. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. **42**, 177–196 (2001)
13. Katz, G., Ofek, N., Shapira, B.: ConSent: Context-based sentiment analysis. Knowl.-Based Syst. **84**, 162–178 (2015)

14. Kawamae, N.: Trend analysis model: Trend consists of temporal words, topics, and timestamps. In: Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM), pp. 317–326 (2011)
15. Lau, J., Collier, N., Baldwin, T.: On-line trend analysis with topic models: #twitter trends detection topic model. In: Proceedings of the 24th International Conference on Computational Linguistics (COLING), pp. 1–16 (2012)
16. Li, J., Cardie, C.: Timeline generation: Tracking individuals on twitter. In: Proceedings of the 23rd International Conference on World Wide Web (WWW), pp. 643–652 (2014)
17. Lin, C., He, Y.: Joint sentiment/topic model for sentiment analysis. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM), pp. 375–384 (2009)
18. Manning, C.D., Raghavan, P., Schutze, H.: Introduction to information retrieval. Cambridge University Press (2008)
19. McCallum, A., Wang, X., Corrada-Emmanuel, A.: Topic and role discovery in social networks with experiments on enron and academic email. J. Artif. Intell. Res. **30**(1), 249–272 (2007)
20. Mehrotra, R., Sanner, S., Buntine, W., Xie, L.: Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp. 889–892 (2013)
21. Mei, Q., Zhai, C.: Discovering evolutionary theme patterns from text: An exploration of temporal text mining. In: Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (SIGKDD), pp. 198–207 (2005)
22. Montejo-Rez, A., Daz-Galiano, M.C., Martnez-Santiago, F., Urea-Lpez, L.A.: Crowd explicit sentiment analysis. Knowl.-Based Syst. **69**, 134–139 (2014)
23. Qian, T., Li, Q., Liu, B., Xiong, H., Srivastava, J., Sheu, P.C.: Topic formation and development: A core-group evolving process. World Wide Web **17**(6), 1343–1373 (2014)
24. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 248–256 (2009)
25. Rao, Y., Lei, J., Wenyin, L., Li, Q., Chen, M.: Building emotional dictionary for sentiment analysis of online news. World Wide Web **17**(4), 723–742 (2014)
26. Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S.M., Ritter, A., Stoyanov, V.: SemEval-2015 task 10: Sentiment analysis in twitter. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval), pp. 451–463 (2015)
27. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI), pp. 487–494 (2004)
28. Shuyo, N.: Language detection library for java. http://code.google.com/p/language-detection/ (2010)
29. Si, J., Li, Q., Qian, T., Deng, X.: Users' interest grouping from online reviews based on topic frequency and order. World Wide Web **17**(6), 1321–1342 (2014)
30. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical dirichlet processes. J. Amer. Stat. Assoc. **101**(476), 1566–1581 (2006)
31. Tang, J., Zhang, M., Mei, Q.: One theme in all views: Modeling consensus topics in multiple contexts. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD), pp. 5–13 (2013)
32. Tang, X., Yang, C.C.: TUT: A statistical model for detecting trends, topics and user interests in social media. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM), pp. 972–981 (2012)
33. Tao, K., Abel, F., Hauff, C., Houben, G.-J., Gadiraju, U.: Groundhog day: Near-duplicate detection on twitter. In: Proceedings of the 22nd International Conference on World Wide Web (WWW), pp. 1273–1284 (2013)
34. Wang, X., McCallum, A.: Topics over time: A non-markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD), pp. 424–433 (2006)
35. Zhou, E., Zhong, N., Li, Y.: Extracting news blog hot topics based on the W2T methodology. World Wide Web **17**(3), 377–404 (2014)