# Context over Time: Modeling Context Evolution in Social Media

Md. Hijbul Alam
Korea University
Seoul, Korea
hijbul@korea.ac.kr

Woo-Jong Ryu
Korea University
Seoul, Korea
skdirwj@korea.ac.kr

SangKeun Lee
Korea University
Seoul, Korea
yalphy@korea.ac.kr

## ABSTRACT

The rise of online social media has led to an explosion in user-generated content. However, user-generated content is difficult to analyze in isolation from its context. Accordingly, context detection and tracking its evolution is essential to understanding social media. This paper presents a statistical model that can detect interpretable topics along with their contexts. A topic is represented by a cluster of words that frequently occur together, and a context is represented by a cluster of hashtags that frequently occur with a topic. The model combines a context with a related topic by jointly modeling words with hashtags and time. Experiments on real datasets demonstrate that the proposed model successfully discovers both meaningful topics and contexts, and tracks their evolution.

## Categories and Subject Descriptors

H.2 [**ARTIFICIAL INTELLIGENCE**]: Learning; H.2.8 [**DATABASE MANAGEMENT**]: Database applications—*Data mining*

## Keywords

Topic model; Context and topic evolution; Social media;

## 1. INTRODUCTION

Trends in social media are very important components of exploratory data analysis and prediction. They help assess the level of public interest in a given topic. Moreover, knowledge of trends improves the user experience of IR systems, benefits time-sensitive online advertisement, and aids in learning social behavioral patterns, e.g., by predicting polls or box office revenues. Typically, a trending keyword is the result of the evolution of many topics that contain the keyword. Therefore, successfully identifying and tracking topic evolution is essential in order to meaningfully present trends. It is challenging to identify the topics relevant to a given entity since the volume of associated tweets is large.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
*DUBMOD'14,* November 03, 2014, Shanghai, China.
Copyright 2014 ACM 978-1-4503-1303-2/14/11 ...$15.00.
http://dx.doi.org/10.1145/2665994.2665996 .

To help identify tweets, metadata may be used. There are several types of metadata available in social media, such as the associated shorturl, picture, @mention, and #hashtag. Perhaps the most important type of metadata, a hashtag, plays several key roles. A hashtag facilitates navigation and defines the semantics of a tweet by linking relevant topics and events together. It also defines a virtual community. Thus, users demonstrate their interest in a topic and simultaneously align themselves with a similar community by using a hashtag, which provides valuable insight into the context of a tweet.

Formally, a topic is represented by a cluster of words that frequently occur together, and a topic evolution is a topic equipped with a time distribution. We define a context as a cluster of hashtags that frequently occur with a topic. Thus, a context is a distribution of hashtags. Similarly, we define a context evolution as a context equipped with time distribution. In social media, contexts change rapidly over time and affect the composition of topics. Therefore, it is important to extract the context of a topic and track its evolution.

To the best of our knowledge, there is no existing work that considers context discovery related to topics with time distribution. One of the pioneering works in the literature of detecting topic evolution is Topics over Time (TOT) [6]. TOT does not discover contexts. Kawamae [2] extended TOT by introducing a trend class in the Trend Analysis Model (TAM). TAM extracts word distribution and topic distribution over time simultaneously. As further extensions, TUT [5] was proposed by considering users interests. However, users and hashtags are not exchangeable, since a hashtag plays dual role. Recently, Tang et al. [4] proposed Multi-contextual LDA (mLDA) that discovers consensus topics based on multiple contexts such as hashtag, user, time, and Mehrotra et al. [3] proposed different pooling scheme such as hashtag, burst score, and time for improving LDA for microblogs. In addition, Alam et al. [1] extracts sentiment topics from online reviews. However, the above models focus on increasing topics quality and ignore context evolution.

In this paper, we propose a new probabilistic generative model, Context over Time (COT), that models in a unified framework the generation process of text and hashtags and their dependency on time. Each word is jointly modeled with time, topic, and/or hashtag. COT creates both topics and contexts with their time distributions. The contributions of this paper are as follows:

- We propose COT, which models both topics and context evolution automatically.

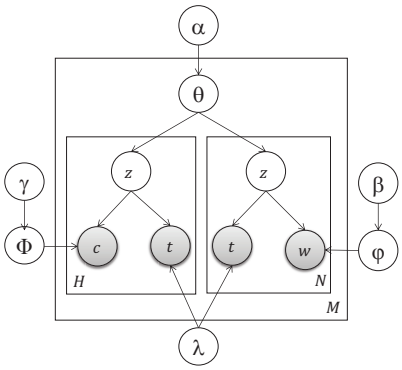



Figure 1: COT model

- We demonstrate the efficacy of COT qualitatively by showing that they increase the interpretability of topics in social media.
- We show that COT is comparable to the baseline approach TOT [6] in terms of time prediction.

# 2. CONTEXT OVER TIME

In this section, we introduce COT to detect both topic and context evolution. Using this model, we can identify, visualize, and understand the topic evolution that is discussed in different contexts.

## 2.1 Generative Model

In online social media, typically each 'document' or tweet is very short. Many tweets contain hashtags to overcome length restrictions or to provide additional perspective. For instance, tweets related to Nelson Mandela contained hashtags such as #qunu, #humanrightsday, and #worldcupdraw. These tags help to explore topics evolving from Nelson Mandela. Based on this motivation, we build the COT.

We assume that topic discovery is influenced by temporal information, word co-occurrence, and the presence of hashtags. Therefore, COT models a word jointly with a timestamp and/or a hashtag. The graphical model of COT is shown in Figure 1. It shows the relationship among a topic, a word, a timestamp, and a hashtag, and topics are responsible for generating timestamps, words, and hashtags. Table 1 describes the parameters of COT, and the generative model of COT is as follows:

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

Table 1: The parameters of COT

| | |
|---|---|
| $M, K$ | # of documents and topics, respectively |
| $N, H$ | # of words and hashtags in a document |
| $w, z, c$ | word, topic and hashtag, respectively |
| $t, d$ | timestamp and document, respectively |
| $\varphi$ | multinomial distribution over words |
| $\phi$ | multinomial distribution over hashtags |
| $\theta$ | multinomial distribution over topics |
| $\lambda_z$ | beta distribution over timestamp for a topic |
| $\alpha, \beta, \gamma$ | Dirichlet prior for $\theta$, $\varphi$ and $\phi$ respectively |

Table 2: Datasets properties

| Dataset | Total tweets | Unique words | Unique hashtags | Total words |
|---|---|---|---|---|
| Sewol Ferry | 239,117 | 8,992 | 723 | 1,790,226 |
| Nelson Mandela | 2,813,461 | 255,298 | 50,425 | 18,776,257 |

## 2.2 Inference

We apply collapsed Gibbs sampling using the following update rules in order to estimate the hidden parameters of the COT. A hashtag is sampled by Eq. (1), and a word is sampled by Eq. (2).

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

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

where $n^z_w$ is the number of times word $w$ 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 times hashtags are assigned to $z$ in the collection. $n^d_z$ is the number of words in document $d$ assigned to topic $z$, and $n^d$ is the number of words in document $d$. We choose to use symmetric Dirichlet distributions ($\alpha$=1.0, $\beta$=0.01, $\gamma$=0.01) like TOT in all of our experiments. By using a sample produced by a Gibbs sampling procedure, we compute $\varphi = \frac{n^z_w + \beta}{n^z + V\beta}$, $\phi = \frac{n^{z_h}_c + \gamma}{n^{z_h} + C\gamma}$.

# 3. EXPERIMENTAL RESULTS

We describe the datasets in Section 3.1, and list contexts discovered by COT in Section 3.2. We make quantitative comparisons in time prediction and KL divergence in Section 3.3. In all cases, for simplicity, we fix the number topics $K$=30. The topics are extracted from a single sample at the 500th iteration of the Gibbs sampler.

## 3.1 Dataset

We perform experiments on two datasets: sinking of the Sewol Ferry, and the death of Nelson Mandela. The properties of the datasets are given in Table 2.

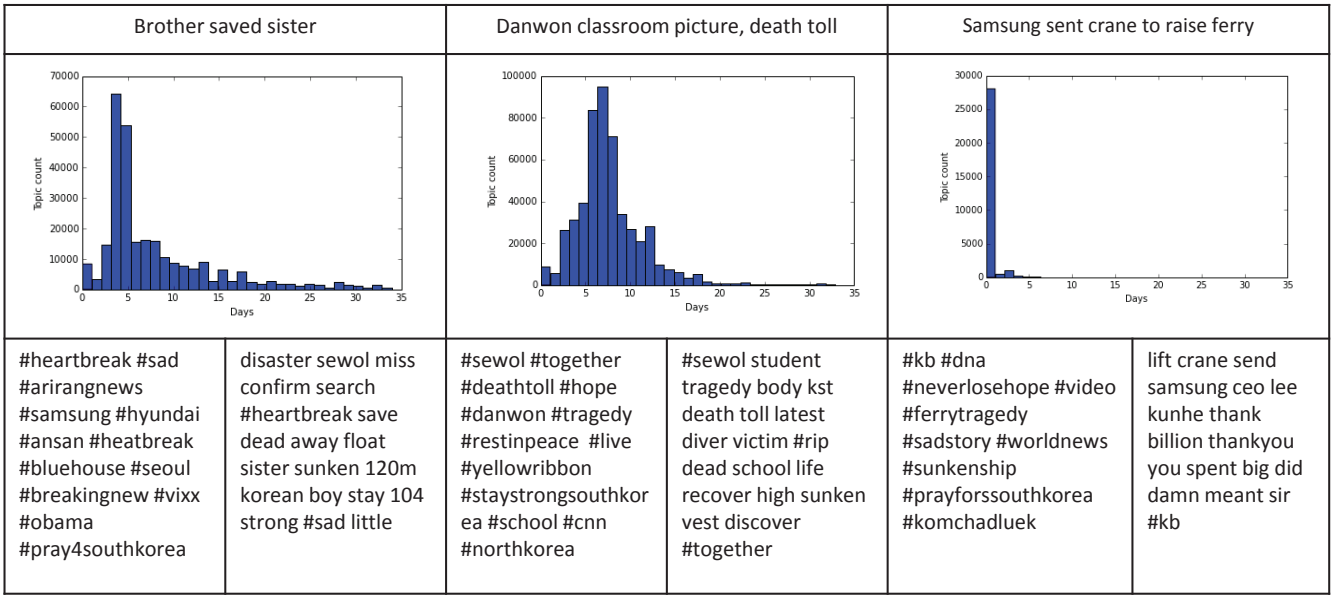



Figure 2: COT contexts, topics and their histogram over time on Sewol Ferry dataset

### 3.1.1 *Sinking of the Sewol Ferry*

The Sewol Ferry capsized while carrying 476 people on 16 April, 2014 en route from Incheon towards Jeju[1]. Using the Twitter streaming API, we collected tweets posted between 17 April, 2014 and 20 May, 2014 using the keywords "ferry" and "#prayforsouthkorea".

### 3.1.2 *Death of Nelson Mandela*

Nelson Mandela was an anti-apartheid revolutionary who served 27 years in prison and dismantled the legacy of the apartheid while he was president of South Africa. He died on 5 December, 2013 at the age of 95. Approximately 90 representatives of foreign states travelled to South Africa to attend memorial events[2]. Using the Twitter streaming API, we collected tweets posted between 6 December, 2013 and 5 January, 2014 for the keyword "Nelson Mandela".

## 3.2 Topic Visualization

Typically, people create hashtags as the details of an event develop, which drives topic evolution. Figure 2 shows the context and topic evolution discovered by COT. Due to the space limitations, we only discuss 3 of the 30 topics. Each topic is illustrated with (a) a histogram of topic distribution over time (b) the top 20 words in each topic, and (c) the top most likely hashtags to be generated concerning the topic.

COT captures both topics and contexts focused in time, which are computed from $\varphi$ and $\phi$, respectively. We manually label three topics in Figure 2. We observe that in some cases, the top 20 words of a topic reflect the messages of one or two tweets. For example, the first topic shows information related to a boy who saved his 5-year-old sister. The second topic describes the death toll and a related picture of a Danwon high school room. The third topic shows information about Samsung CEO who sent a crane to lift the ferry.

However, contexts are created by several messages or emotions in social media and the surrounding communities. For example, #vixx, and #pray4southkorea are some communities that impacted the evolution of the first topic. The other hashtags, #heartbreak, and #sad, reflect the emotions of the users in social media. In the second topic, #yellowribbon describes a movement to pay tribute to the victims by displaying yellow ribbons all over Korea. In the third topic, #video describes the conversation and surroundings recorded by a victim while ferry was sinking. Overall, we see many interesting hashtags in contexts, which were not found in the top words of topics.

Figure 3 shows the context and topic evolution discovered through the Nelson Mandela dataset. Here, we also observe that contexts discovered interesting hashtags that were not in top words of topics. We mentioned a few examples here, such as #mandelamemory, #humansrightday and #worldcupdraw. These hashtags reflect what happened during a period and reveal users' motivations to generate that topic. After the death of Nelson Mandela, people around the world share their memories of Nelson Mandela in social media through the hashtag #mandelamemory, which appears in the first topic. The final draw of the World Cup was held on 6 December, 2013, and people revived interest in appropriate Mandela's quotes about sports, reflected through the #worldcupdraw and #quoteoftheday hashtags in the second topic. The third topic shows that people also deeply mourned Mandela four days after his death, 10 December, 2013, which is Human Rights Day. COT contexts discovered all of these hashtags.

Table 3 shows that COT outperforms TOT in discovering meaningful hashtags i.e., the semantics of topics for a common topic. For example, TOT discovered one hashtag #prayforsouthkorea, whereas COT discovered many meaningful hashtags, such as #heartbreak, #sosad, and so on. Therefore, COT is an effective way to model social media and extracts contexts.

## 3.3 Quantitative Comparisons

An interesting feature common to the COT and TOT is the ability to predict timestamp [6]. Thus, COT and TOT

[1]http://en.wikipedia.org/wiki/Sinking_of_the_MV_Sewol
[2]http://en.wikipedia.org/wiki/Nelson_Mandela

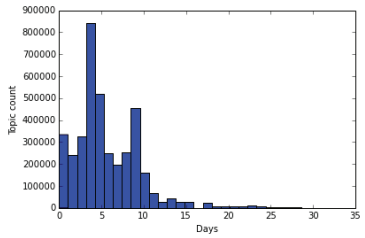
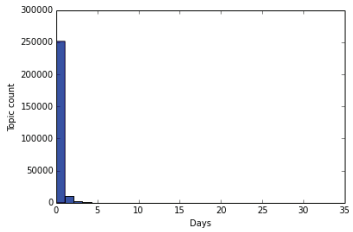
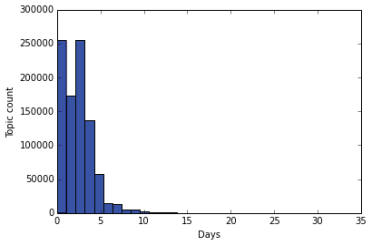
| Mandela memory | | Quote | | Kanye West controversy | |
|---|---|---|---|---|---|
|  | |  | |  | |
| #mandelamemory #nelson #obama #rip #p2 #mufc #tfb #youngturk #peopleschampion #paulasli #madibamemory #art #ripnelsonma | memory south tribute africa obama presid service funeral african photo sign 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 #cdnpoli #sirkenayo #ucl #cancer #photography #humanrightsday #atheism #ssrugbi #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 3: COT contexts, topics and their histogram over time on Nelson Mandela dataset

Table 3: Topic distribution for a common topic of different models on Sewol Ferry dataset

| | |
|---|---|
| COT context | #heartbreak #so #bangyongguk #prayfo #prayforsout #mh370 #southkorea #korea #sosad #prayforsouthkor #sewol #prayforsouth |
| COT topic | sent text student #heartbreak prayforsouthkorea helpless start hisher moms dad still alive |
| TOT | student sent good family mom start text sink #prayforsout hisher trap video |

Table 4: Accuracy and L1 Error comparison

| | Sewol Ferry | | Nelson Mandela | |
|---|---|---|---|---|
| | Accuracy | L1 Error | Accuracy | L1 Error |
| COT | 0.44 | 1.59 | 0.468 | 2.35 |
| TOT | 0.44 | 1.57 | 0.466 | 2.35 |

may be quantitatively compared by measuring how well they predict timestamps. We randomly select 1000 tweets and predict the timestamps. We also compute L1 Error which is the difference between predicted and true day. As shown in Table 4, COT is comparable to TOT in timestamp prediction accuracy and L1 Error. Similarly, we observe that average KL divergences between topics for COT vs. TOT are comparable on Sewol Ferry dataset for different number topics, which is shown in Figure 4.

## 4. CONCLUSION

In this paper, we addressed the problem of contexts extraction and evaluating the role of contexts in shaping topics across social media. We proposed a novel Context over Time (COT) model. We showed that COT discovered important hashtags in contexts effectively, which were not discovered by topics alone. We plan to work on relevant tweet timeline generation and sentiment analysis based on contexts.

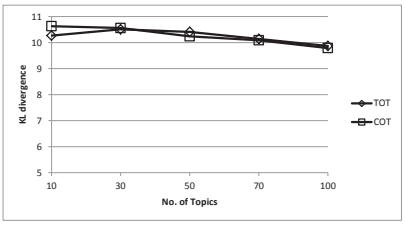



Figure 4: Average KL divergence between topics for COT vs. TOT

## Acknowledgment

This work was supported in part by the Brain Korea 21 Plus Project and the Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education, Science and Technology (No. 2012M3C4A7033344).

## 5. REFERENCES

[1] M.H. Alam and S. Lee, "Semantic Aspect Discovery for Online Reviews," in *Proc. of ICDM*, 2012, pp. 816-821.

[2] N. Kawamae, "Trend analysis model: Trend consists of temporal words, topics, and timestamps," in *Proc. of WSDM*, 2011, pp. 317–326.

[3] R. Mehrotra, S. Sanner, W. Buntine, Wray and L. Xie, "Improving LDA topic models for microblogs via tweet pooling and automatic labeling," in *Proc. of SIGIR*, 2013, pp. 889–892.

[4] J. Tang, M. Zhang, and Q. Mei, "One theme in all views: Modeling consensus topics in multiple contexts," in *Proc. of SIGKDD*, 2013, pp. 5–13.

[5] X. Tang and C. C. Yang, "TUT: A statistical model for detecting trends, topics and user interests in social media," in *Proc. of CIKM*, 2012, pp. 972–981.

[6] X. Wang and A. McCallum, "Topics over time: A non-markov continuous-time model of topical trends," in *Proc. of SIGKDD*, 2006, pp. 424–433.