Exploring Time-Sensitive Variational Bayesian Inference LDA for Social Media Data

Anjie Fang[1], Craig Macdonald[2], Iadh Ounis[2], Philip Habel[2], Xiao Yang[2]

[1]a.fang.1@research.gla.ac.uk,[2]{firstname.secondname}@glasgow.ac.uk
University of Glasgow, UK

Abstract. There is considerable interest among both researchers and the mass public in understanding the topics of discussion on social media as they occur over time. Scholars have thoroughly analysed sampling-based topic modelling approaches for various text corpora including social media; however, another LDA topic modelling implementation—Variational Bayesian (VB)—has not been well studied, despite its known efficiency and its adaptability to the volume and dynamics of social media data. In this paper, we examine the performance of the VB-based topic modelling approach for producing coherent topics, and further, we extend the VB approach by proposing a novel time-sensitive Variational Bayesian implementation, denoted as TVB. Our newly proposed TVB approach incorporates time so as to increase the quality of the generated topics. Using a Twitter dataset covering 8 events, our empirical results show that the coherence of the topics in our TVB model is improved by the integration of time. In particular, through a user study, we find that our TVB approach generates less mixed topics than state-of-the-art topic modelling approaches. Moreover, our proposed TVB approach can more accurately estimate topical trends, making it particularly suitable to assist end-users in tracking emerging topics on social media.

1 Introduction

Perhaps the greatest technological change over the past decade has been the advent and growth of social media. Yet despite social media's ubiquity, scholars still wrestle with the appropriate tools for best capturing the topics of discussion conveyed over these platforms [1–3]. To this end, researchers have employed various topic modelling approaches [1, 4–8], e.g. Latent Dirichlet Allocation (LDA), but these efforts have proved challenging, as models applied to social media data can produce topics that are mixed and lack coherence, and are generally difficult to interpret [1]. To deal with the short nature of social media posts, LDA enhancement methods such as single topic assignment [1, 9, 10] and document pooling [2, 11] have been proposed to improve the coherence of the generated topics within the sampling-based topic modelling approaches. However, another LDA implementation, the Variational Bayesian (VB)-based topic modelling approach, has not been well studied on social media posts. As the VB approach has been shown to be more efficient for large datasets [6, 12], it can be argued that VB can better handle the increasing volume and dynamicity of social media data.

It has been previously shown that the time dimension of documents (e.g. news articles) can help a topic modelling approach to provide more valuable information [7, 8, 13], for example, capturing the topic changes or topical trends over time. Apart from these additional benefits, we argue that distinguishing topical word usage over time can also help to generate more coherent and less mixed topics, thereby assisting the end-users in interpreting discussions on social media. We propose a time-sensitive VB (TVB) approach for social media data that embraces the time dimension of social media data. We extend the traditional VB approach by incorporating a Beta distribution, which is reported to fit various patterns [14]. The employed Beta continuous distribution is used to represent each topic's volume over time, i.e. the topical trend, similar to what has been used in [7]. However, we notice that time could have a negative bias on the topic inference when a Beta distribution does not fit the topics' trends. To solve this problem, we introduce a balance parameter to alleviate the bias of time.

To evaluate the performance of the proposed TVB approach, we create a ground truth Twitter dataset covering 8 large events. We evaluate our TVB approach together with several baselines from the literature (e.g. Twitter LDA (TLDA) [1], the Topic Over Time approach (TOT) [7]) in terms of topical coherence, the extent to which the generated topics are mixed, or the estimation errors of the topical trends. Our empirical results suggest that incorporating the time dimension does indeed help to enhance the coherence of the topics generated by the TVB approach compared with the traditional VB and sampling approaches. Moreover, we show that our TVB model can outperform the state-of-the-art LDA enhancement approaches (i.e. TLDA and TOT) in generating less mixed topics. This conclusion is further supported by conducting a user study to validate the results of the quantitative evaluation. Finally, we compare our TVB approach with the TOT approach when estimating the topical trends. We find that our proposed TVB model better estimates the topical trends.

The contributions of this paper are three-fold: 1) we study the VB approach and develop its enhancement for social media, 2) we propose a time-sensitive TVB approach by integrating the time dimension in the modelling process and 3) we show the advantages of the TVB approach in generating better quality topics and estimating more accurate topical trends.

The rest of this paper is organised as follow: Section 2 provides basic background on two LDA implementations, i.e. sampling & VB approaches, followed by a description of related work in Section 3. We describe our TVB approach in Section 4. Following that, we describe our dataset in Section 5 and the experimental setup in Section 6. The results are shown and discussed in Section 7. Finally, we provide concluding remarks in Section 8.

2 Two LDA implementations: Sampling & VB approaches

In topic modelling approaches, a topic k is represented by a distribution β_k (k is the topic index and K is the number of topics) over N terms drawn from a Dirichlet prior η, where N is the size of the vocabulary. A document in a corpus $\boldsymbol{w}$ is represented by $w_d = \{w_{d,1}, ..., w_{d,i}, ..., w_{d,N}\}$ (d is the document index and D is the number of documents in $\boldsymbol{w}$) and has a topic belief distribution θ_d drawn from the Dirichlet prior α. A document w_d is associated with topic assignment $z_d = \{z_{d,1}, ..., z_{d,i}, ..., z_{d,N}\}$. The sampling approach [4, 5], which is based on a Markov Chain Monte Carlo sampling, estimates the real posterior distributions (e.g. β_k & θ_d). In a typical sampling approach, such as the collapsed Gibbs sampling, each word is assigned a topic according to Equation (1) in order to construct a Markov Chain on latent topics, where $n^{w_{d,i}}_{-(d,i),k}$ is the number of $w_{d,i}$ occurring in topic k and $n^{d}_{-(d,i),j}$ is the number of words from document w_d occurring in topic k not including the current one. After a number of iterations, $\boldsymbol{\beta}$ ($\{\beta_1, .., \beta_K\}$) and $\boldsymbol{\theta}$ ($\{\theta_1, .., \theta_D\}$) can be estimated from the converged Markov Chain.

$$p(z_{d,i} = k | z_{-(d,i)}, \boldsymbol{w}) = \frac{n^{w_{d,i}}_{-(d,i),k} + \eta}{n_{-(d,i),k} + N\eta} \times (n^{d}_{-(d,i),j} + \alpha) \qquad (1)$$

The VB approach [6, 12] approximates the variational distribution by minimising the distance from the true distribution. Specifically, an expectation maximization (EM) algorithm is used to maximise the lower bound of the log-likelihood of all documents, which equivalently minimises the distance between the variational distribution and the true posterior distribution. In the E step of EM, the variational Dirichlet prior γ_d of all documents are optimised together with $\phi_{D \times N \times K}$, which represents the words' topic belief within documents. In the M step of EM, $\phi_{D \times N \times K}$ is used to update the variational Dirichlet prior $\lambda_{K \times N}$ of $\boldsymbol{\beta}$. The parameters' optimisation formulas in the EM algorithm are shown in Equation (2). Finally, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ can be obtained when the lower bound converges. Importantly, since the VB approach does not have the topic assignment step, the single topic assignment strategy (mentioned in the introduction and discussed further in Section 3) cannot be applied. The main advantage of the VB approach is that the lower bound converges much more quickly than the sampling approach especially on large datasets [6, 12]. Moreover, the VB approach can be intuitively implemented in parallel since the updates of γ_d & ϕ_d among documents do not impact each other, while the sampling approach cannot be easily parallelised as it is intrinsically sequential [15]. Because of the increasing volume of social media data and its dynamicity, it could be argued that the VB approach offers various advantages for those interested in interpreting discussions on social media as events transpire. In the next section, we review a number of existing methods, which aim to improve the quality of topic models and/or integrate the time dimension in the topic model.

$$\phi_{d,i,k} \propto exp\{E[log\beta_{k,i}] + E[log\theta_{d,k}]\},\ \gamma_{d,k} = \alpha + \textstyle\sum_{i,k} \phi_{d,i,k},\ \lambda_{k,i} = \eta + \textstyle\sum_{d,i,k} \phi_{d,i,k} \qquad (2)$$

3 Related Work

Three methods are mainly used in the literature to adapt topic modelling approach for short social media posts: 1) A post is assigned to a single topic under the assumption that a post represents a single topic. This method was used in Twitter LDA as proposed by Zhao et al. [1] and later applied in [9, 10]. Indeed, this method brings more coherent words for a topic. However, we argue that this method can generate multi-theme topics[1] since the underlying assumption cannot be upheld in all situations. For example, the same words can be used across

[1] A mixed topic contains keywords pertaining to multiple different topic themes.

multiple topics. Assigning all words in a tweet to a single topic could increase word overlaps and thus result in mixed topics. 2) Multiple posts are combined into a virtual document [2, 11], also known as the pooling strategy (e.g. tweets from a single user are combined into a single document [2]). The pooling method can increase the number and occurrence of words, which makes it easier to apply a topic modelling approach. 3) Topical words are connected using word representations (e.g. word embedding). Sridhar et al. [16] improved the topical coherence by applying soft clustering over word representations in a topic model. Nguyen et al. [17] introduced an additional word topic belief distribution calculated using word representations in the sampling approaches. Li et al. [18] assigned the semantically similar words under the same topic. All of these approaches improved the topical coherence by connecting similar words in order to overcome the shortness of posts on social media. We do not deploy the single topic assignment method in our approach since it cannot be applied in the VB approach, as mentioned in Section 2. Given that the central aim of this paper is to integrate time to the VB approach, we do not adopt the pooling method in our modelling process.

Early work on time-sensitive topic modelling by Blei et al. [13] was based on a Markov assumption that the topic parameters are in a sequential structure over time. Later on, Blei et al. [19] used Brownian motion to estimate the topical evolutions over time. The proposed model was claimed to have a better predictive perplexity. However, these state-space models did not integrate the timestamps of documents in the generative process. Assuming that the topic proportion changes over time, Wang et al. [7] proposed a non-Markov topic model (TOT) using a Beta continuous distribution, which was reported to generate more interpretable topics and trends. Their work is based on a sampling approach, in which the timestamps of documents are incorporated in the generative process without considering time dependency for topics or words. Another recent work from [8] leveraged a time-dependent function to capture topical dynamics.

Although the sampling approach is still the preferred choice in analysing social media data, the advantages of the VB approach for a large corpus should not be ignored. For example, Hoffman et al. [20] and Braun et al. [12] recently proposed a VB-based solution to quickly inference a large number of documents. In this paper, we offer a solution to apply an enhanced VB approach (TVB) for social media data, which incorporates time in the topic modelling process. Our TVB approach is based on the same assumption as the TOT approach but is implemented using VB. In the next section, we introduce our TVB approach and elaborate further the differences between the TOT and TVB approaches.

4 Integrating the time dimension in the VB Approach

Our proposed TVB approach extends the traditional VB approach by integrating the time dimension of social media data. In this section, we explain how we implement the EM algorithm in our proposed TVB approach and compare it with the traditional VB and TOT approaches. To integrate the timestamps of social media posts, we deploy a continuous probability distribution τ for each topic. This time distribution τ_k represents the proportion of topic k over time. Theoretically, any continuous distribution can be used to simulate the topic proportion over time. However, to better estimate topical trends, the continuous distribution has to approximate the real topical trends. Indeed, recently, the Beta distribution has drawn a lot of attentions for accommodating a variety of shapes given an x-axis interval [14]. Therefore, we choose to use a Beta distribution since it can more accurately fit the various shapes of topical trends. Next, we describe the generative process and the EM implementation of our TVB approach.

Generative process. Similar to the traditional LDA generative process, each word $w_{d,i}$ in a document d is assigned a topic assignment $z_{d,i}$ according to θ_d, where i is the word index. Since words ($\boldsymbol{w}$) in social media posts are associated with timestamps ($\boldsymbol{t}$), in the TVB approach, a pair $(w_{d,i}, t_{d,i})$ is drawn from $\beta_{z_{d,i}}$ and $\tau_{z_{d,i}}$, respectively, where a Beta distribution τ_k is parametrised by two shape parameters, ρ_k^1 and ρ_k^2. A similar strategy was previously applied in a time-sensitive sampling approach [7]. The process is defined as follows:

$$z_{d,i}|\theta_d \sim Dirichlet(\alpha), \quad w_{d,i}|z_{d,i}, \beta_{z_{d,i}} \sim Dirichlet(\eta), \quad t_{d,i}|z_{d,i}, \tau_{z_{d,i}} \sim Beta(\rho^1_{z_{d,i}}, \rho^2_{z_{d,i}})$$

EM Implementation of the TVB approach. The core part of a variational inference is to minimise the distance between the variational distributions $q(\theta_d|\gamma_d)$ & $q(\beta_k|\lambda_k)$ and the two true posterior distributions $p(\theta_d|\alpha)$ & $q(\beta_k|\eta)$, i.e. maximising the lower bound of a document log-likelihood $p(\boldsymbol{w}, \boldsymbol{t}|\alpha, \eta)$ shown in Equation (3). The right part of the equation is the lower bound of all documents, L. Commonly, the derivative of L is taken over parameters (γ, ϕ, λ) and thus the parameter optimisation formulas can be obtained by maximising the lower bound L. To achieve this, we first decompose all the items in L.

$$\begin{gathered} log\ p(\boldsymbol{w}, \boldsymbol{t}|\alpha, \eta) \geq L(\boldsymbol{w}, \boldsymbol{t}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) = \textstyle\sum_d E_q[log\ p(w_d, t_d, z_d, \theta_d, \boldsymbol{\beta}, \boldsymbol{\tau}|\alpha, \eta)] \\ = \textstyle\sum_d (E_q[log\ p(w_d|z_d, \boldsymbol{\beta})] + E_q[log\ p(z_d|\theta_d)] + E_q[log\ p(\theta_d|\alpha)] + E_q[log\ p(\boldsymbol{\beta}|\eta)] \\ +E_q[log\ p(t_d|z_d, \boldsymbol{\tau})] - E_q[q(\theta_d, z_d, \boldsymbol{\beta})]) \end{gathered} \tag{3}$$

The sixth item of the lower bound L, the log-expectation of joint variational probability, is decomposed as shown in Equation (4). These decomposed items together with the first five items in L can be expanded by leveraging the properties of the Dirichlet and Beta distributions. Finally, we have the expanded L shown in Equation (5), where B is the Beta function and $\mathit{\Gamma}$ is the Gamma function.

$$E_q[q(\theta_d, z_d, \boldsymbol{\beta})] = \textstyle\sum_k E_q[log\ q(\theta_{d,k}|\gamma_{d,k})] + \sum_i E_q[log\ q(z_{d,i}|\phi_{i,k})] + \sum_{i,k} E_q[log\ q(\beta_{k,i}|\lambda_{k,i})] \tag{4}$$

$$\begin{gathered} L(\boldsymbol{w}, \boldsymbol{t}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) = \textstyle\sum_d (\sum_{i,k} \phi_{d,i,k}\ E_q[log\ \beta_{k,i}] + \sum_{i,k} \phi_{d,i,k} E_q[log\ \theta_{d,k}] \\ +log\mathit{\Gamma}(K\alpha) - Klog\mathit{\Gamma}(\alpha) + \textstyle\sum_k (\alpha - 1) E_q[log\theta_{d,k}] \\ +log\mathit{\Gamma}(\textstyle\sum_{i,k} \eta) - \sum_{i,k} log\mathit{\Gamma}(\eta) + \sum_{i,k} (\eta - 1) E_q[log\ \beta_{k,i}] \\ + \textstyle\sum_{i,k} \phi_{d,i,k}((\rho_k^1 - 1) log\ t_{d,i} + (\rho_k^2 - 1) log\ (1 - t_{d,i})) - \sum_k (\sum_i \phi_{d,i,k}\ log\ B(\rho_k^1, \rho_k^2)) \\ -log\mathit{\Gamma}(\textstyle\sum_k \gamma_k) + \sum_k log\mathit{\Gamma}(\gamma_k) - \sum_k (\gamma_k - 1) E_q[log\theta_{d,k}] \\ - \textstyle\sum_{i,k} \phi_{d,i,k} log\phi_{d,i,k} - log\mathit{\Gamma}(\sum_{i,k} \lambda_{k,i}) + \sum_{i,k} log\mathit{\Gamma}(\lambda_{k,i}) - \sum_{i,k} (\lambda_{k,i} - 1) E_q[log\ \beta_{k,i}]) \end{gathered} \tag{5}$$

To maximise L, we first optimise $\phi_{d,i,k}$ by setting $\frac{\partial L_{\phi_{d,i,k}}}{\partial \phi_{d,i,k}} = 0$ and obtain the $\phi_{d,i,k}$ optimisation formula shown in Equation (6). Compared with the traditional VB approach, the third item in Equation (6), the time statistics, is the additional feature we add to incorporate timestamps. Intuitively, the time statistics can have a direct impact on the term topic belief $\phi_{d,i,k}$. If a word $w_{d,i}$ is

Algorithm 1: Our TVB approach for Latent Dirichlet Allocation.

```
Initialize $\lambda_{N \times K}$, $\gamma_{D \times K}$
while $L$ not converges do
    E step:
    for $d < D$ do
        repeat
            for $i < N^d$ & $k < K$ do
                $\phi_{d,i,k} \propto exp(E_q[log\beta_{k,i}] + E_q[log\theta_{d,k}]$
                $+\delta((\rho_k^1 - 1)log\ t_{d,i} + (\rho_k^2 - 1)log\ (1 - t_{d,i}) - log\ B(\rho_k^1, \rho_k^2)))$
                $\gamma_{d,k} = \alpha + \sum_{i,k} \phi_{d,i,k}$
        until $\gamma_d$ converges;
    M step:
    $\psi(\rho_k^1) - \psi(\rho_k^1 - \rho_k^2) = \frac{\sum_{d,i,k} \phi_{d,i,k} log\ t_{d,i}}{\sum_{d,i,k} \phi_{d,i,k}}, \psi(\rho_k^2) - \psi(\rho_k^1 - \rho_k^2) = \frac{\sum_{d,i,k} \phi_{d,i,k} log\ (1 - t_{d,i})}{\sum_{d,i,k} \phi_{d,i,k}}$
    $\lambda_{k,i} = \eta + \sum_{d,i,k} \phi_{d,i,k}$, $\forall i \in N$
```

highly used in topic k at a time point t, $\phi_{d,i,k}$ is likely to be promoted if a post has the word $w_{d,i}$ with a timestamp t. However, the estimated time distribution may not always fit a topic's trend well. A drifted time distribution could give a negative bias on $\phi_{d,i,k}$. To solve this problem, we introduce a balance parameter δ, to control the impact of the time statistics on $\phi_{d,i,k}$ and alleviate such bias. Note that the influence of time in the TOT approach cannot be adjusted, e.g. through a δ parameter. Similar to $\phi_{d,i,k}$, we obtain the optimisation formula of γ and λ (shown in Equation (7)) by setting their derivative of L to zero.

$$\phi_{d,i,k} \propto exp(E_q[log\beta_{k,i}] + E_q[log\theta_{d,k}] + \delta((\rho_k^1 - 1)log\ t_{d,i} + (\rho_k^2 - 1)log\ (1 - t_{d,i}) - log\ B(\rho_k^1, \rho_k^2))) \quad (6)$$

$$\gamma_{d,i} = \alpha + \textstyle\sum_{i,k} \phi_{d,i,k}, \quad \lambda_{k,i} = \eta + \textstyle\sum_{d,i,k} \phi_{d,i,k} \quad (7)$$

Meanwhile, to maximise L, we can also take the partial derivative with respect to the parameters of Beta distribution, ρ_k^1/ρ_k^2. Actually, this step is equivalent to maximising the likelihood of the timestamps in topics. By optimising ρ_k^1/ρ_k^2, we also obtain the estimated topical trends. Taking the derivative to zero, we obtain the optimisation formula of ρ_k^1/ρ_k^2 shown in Equation (8). Since the Digamma function (ψ, log-derivative of Γ) is involved in the optimisation equation, it is impossible to calculate ρ_k^1/ρ_k^2 directly. In our TVB approach, we estimate ρ_k^1/ρ_k^2 using a parameter optimisation algorithm and we set their initial values following [21]. Note that, while we use EM to estimate ρ_k^1/ρ_k^2, the method of moment [7] is used in the TOT approach. In summary, in the iterative EM algorithm, we update ϕ and γ for each document (social media post) in the E step. In the M step, λ and ρ_k^1/ρ_k^2 are updated using the statistics information (ϕ) from all posts. At the same time, all the timestamps are taken into account to estimate ρ_k^1/ρ_k^2. Algorithm 1 shows the EM algorithm in our TVB approach.

$$\psi(\rho_k^1) - \psi(\rho_k^1 - \rho_k^2) = \frac{\sum_{d,i,k} \phi_{d,i,k} log\ t_{d,i}}{\sum_{d,i,k} \phi_{d,i,k}}, \ \psi(\rho_k^2) - \psi(\rho_k^1 - \rho_k^2) = \frac{\sum_{d,i,k} \phi_{d,i,k} log\ (1 - t_{d,i})}{\sum_{d,i,k} \phi_{d,i,k}} \quad (8)$$

5 Ground Truth Datasets

To evaluate our proposed TVB approach together with the existing topic modelling approaches, we create a Twitter dataset containing 8 selected popular hashtag-events that occurred in July and August 2016. This dataset was collected using Twitter API by searching for 8 hashtags: `#gopconvention`, `#teamgb`,

#badminton, #gameofthrone, #juno, #nba, #pokemongo and #theresamay. For each hashtag-event, we randomly sample 2,000 tweets, hence we obtain a Twitter dataset containing 16,000 tweets. Such a balanced dataset has several advantages: 1) The reasonable size (16K) of the Twitter corpus allows for the efficient conduct of our experiments, i.e. all approaches can quickly converge; 2) We avoid generating dominant and duplicated topics, thereby focusing the evaluation on the quality and coherence of the topics; 3) These predefined hashtags provide readily usable ground-truth labels, i.e. each hashtag-event is associated with the top 10 used words in its corresponding tweets. These labels of the 8 hashtag-events are used to match a generated topic with a hashtag-event. This enables us to evaluate how close the estimated topical trend to its real trend (further details are given in Section 6); 4) This ground truth dataset allows humans to more effectively examine the generated topics and to conduct a user study described in Section 7. Indeed, since this dataset contains a limited number of topics, it is more feasible for human interpreters to evaluate all the generated topics of a given topic model in the conducted user study. In the next section, we explain how we apply various topic modelling approaches on this dataset and the used metrics.

6 Experimental Setup

We compare our new proposed TVB approach to 4 baselines from the literature, namely TOT [7], TLDA [1], and the traditional sampling (Gibbs) [4] and VB [6] approaches. In particular, the TOT approach is included since it is the most closely related work that integrates the time dimension into the topic modelling process. We use 3 different metrics to evaluate the quality of the generated topic models: 1) the topical coherence, 2) the degree to which the topics are mixed and 3) the topical trends estimation error. In the following, we explain the experimental setup used for the topic modelling approaches and each of the metrics.

Topic Modelling Setup. For all approaches (Gibbs, TLDA, TOT, VB & TVB), η is set to 0.01 according to [4, 5]. We do not follow the traditional setting for α ($\alpha = 50/K$), and set instead α to 0.4 for all approaches in our experiments, since in other separate preliminary experiments we noticed that a smaller α helps to generate topics with higher coherence for short texts. The number of topics is set to 10, which is slightly higher than the real number of topic (8 in our dataset corresponding to 8 hashtags) because a slightly higher number of topics assures that all hashtag-events can be extracted. As our Twitter dataset is not very big and contains distinguishable topics, all approaches can converge fast. Hence, for the sampling approaches (Gibbs, TLDA and TOT), we set the maximum number of iterations to 50. For the traditional VB and our proposed TVB approaches, we set the number of iterations to 10 as the VB approaches converge more quickly. Each experiment for each approach is repeated 10 times in order to conduct statistical significance. In TLDA, a document contains several tweets from a single Twitter user. However, most of users in our Twitter dataset have only one tweet. Hence, we create a virtual Twitter user by assigning 5 random tweets to this user. For all the other approaches, a document represents a single tweet.

Metric 1: Coherence Metric. A coherence metric is used to evaluate whether a generated topic is interpretable by humans. A higher score indicates that the topic is easier to understand. Following [22, 23], we use a word embedding (WE) representations-based coherence metric to evaluate the coherence of the generated topics, which has been reported to have a high agreement with human judgments. In order to capture the semantic similarity of the latest hashtags and Twitter handle names, we train our WE model using 200 million English tweets posted from 08/2015 to 08/2016. The obtained WE model has 5 million tokens. We use the average coherence (Aver) to evaluate all topics in a topic model. Meanwhile, we also examine the top 2/7[2] most coherent topics in a model for more effective coherence evaluation, i.e. C@2 & C@7 metrics, following to [24].

Metric 2: Topics Mixing Degree. A generated topic can be a mixture of several topic themes (multi-theme topics). The coherence score is calculated by averaging the similarity of each two words in the top 10 words of a generated topic. Consider that if a topic contains two topic themes, as long as the coherence of words under a theme in this topic is high, the coherence metric can still yield a higher coherence. Although this multi-theme topic is interpretable by humans, a user expects to see the generated topics only containing a single topic theme. Therefore, it is necessary to identify the multi-theme topics. Since the generated multi-theme topics often contain the same topic theme, these multi-theme topics can be similar to each other. Thus, to quantify the extent to which a given topic in a model is mixed (MD), we use Equation (9) to calculate the topic similarity in the entire topic model (containing K topics). The higher the similarity, the more likely that the model has more multi-theme topics. A similar methodology is used in [25] to identify the background topics.

$$MD(\boldsymbol{\beta}) = \sum_k \sum_{k'} cosine(\beta_k, \beta_{k'})/|\boldsymbol{\beta}|^2 \tag{9}$$

Metric 3: Topical Trends Estimate Error. Both the TOT and our TVB approaches estimate the topical trends. To evaluate the topical trends over time, we calculate the distance/error between the real topic trends and the estimated topical trends (using the Beta distributions in the TOT/TVB models). The error is calculated using the method shown in Equation (10), where $PDF_k(t)$ is the probability density of the real timestamps of topics, which is obtained through the ground truth Twitter dataset. The ERR score ranges from 0 to 2. The generated topics are matched to the ground-truth topics if the top 10 words of a generated topic have at least 3[3] same words in the top 10 words of a hashtag event.

$$ERR(\tau) = \frac{\sum_k \int_0^1 |\tau_k(t) - PDF_k(t)| dt}{K} \tag{10}$$

7 Results

In this section, we analyse our experimental results shown in Table 1. The listed scores are the average scores of 10 models generated by each approach with respect to the 3 types of metrics (described in Section 6). For the coherence metrics (Aver, C@2 & C@7), a higher score means more coherent topics, whereas lower scores for the MD and ERR metrics indicate higher quality models. The subscript indicates whether a given approach is significantly[4] better than the other one.

[2] Considering that the number of topics is 10, the top 2 and 7 most coherent topics are reasonable choices for all-side coherence evaluation.

[3] 3 mutual words in the top 10 words is a reasonable minimum number to indicate a similar topic.

[4] p-values ($p<0.05$) are calculated by the t-test using 10 models of each two approaches.

Table 1. The topic coherence, mixing degree and topic trends estimation error.

Models	Coherence			MD	ERR
	Aver	C@2	C@7		
Gibbs (G)	0.154	0.204	0.168	$0.051_{W,T}$	×
TLDA (W)	$0.177_{G,V,T,T'}$	$0.248_{G,V,T,T'}$	$0.198_{G,V,T,T'}$	0.102_{T}	×
VB (V)	0.151	0.201	0.165	$0.049_{W,T}$	×
TOT (T)	$0.160_{G,V}$	0.205	0.175_{V}	0.149	1.358
TVB(T'), $\delta = 0.4$	0.152	0.202	0.165	$0.043_{W,T}$	1.211_{T}
TVB(T'), $\delta = 0.6$	0.153	0.204	0.166	$0.042_{W,T}$	1.256_{T}
TVB(T'), $\delta = 0.8$	0.158_{V}	$0.221_{G,V,T}$	0.174_{V}	$0.047_{W,T}$	1.206_{T}
TVB(T'), $\delta = 1.0$	0.156_{V}	0.209	0.170	$0.055_{W,T}$	1.168_{T}

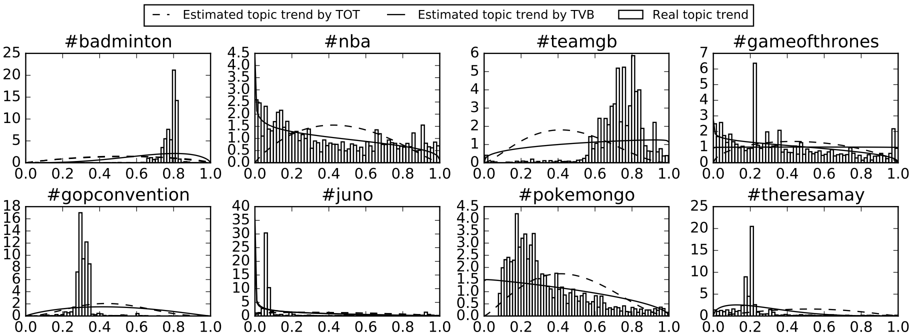

Fig. 1. The real and estimated topical trends, where x-axis and y-axis represent the timeline and density probability, respectively.

For example, the average coherence score (`Aver`) of TVB$_{\delta=0.8}$ (δ is the balance parameter) is significantly better than that of the VB approach, indicating that TVB generates topics with higher coherence than VB. To help understand the topical trends, we randomly choose one TOT and one TVB models and list their estimated topical trends together with the real trends in Figure 1. Next, we will first analyse the results in terms of topical coherence and topical mixing degree. Then, we discuss the performances of the approaches in estimating topical trends.

Topical Coherence and Topical Mixing Degree. Our experiments involve two types of topic modelling approaches: Sampling & VB approaches (shown as G, W, T & V, T' in Table 1) and two topic enhanced methods: single topic assignment (W) and incorporation of the time dimension (T, T'). First, for the topical coherence, it is clear that the single topic assignment, TLDA (W), significantly outperforms all of the other approaches. However, we can still see the positive impact of the time dimension in improving the coherence of models in both the TVB and TOT approaches. For example, the `Aver` coherence score of TOT is better than that of the Gibbs and VB approaches. In particular, for `C@2`, the TVB models outperform all of the other approaches, except the TLDA models with $\delta = 0.8$, while the TVB models with a lower/higher δ (T' with $\delta = 0.4, 0.6/1.0$) do not. This indicates that alleviating the bias of the time statistics (described in Section 4) helps to generate topics with a higher coherence. In terms of the `MD` metric, the TLDA models have higher mixing scores indicating they have more multi-theme topics. As argued in Section 3, aggressively assigning all words in a tweet under the same topic theme can result

Table 2. Topic samples from TLDA and TVB$_{\delta=0.8}$ models, where the underlined words have a different topic theme from the others in a topic. Note that, we present a human with the top 10 words of a topic in our user study. We list the top 5 words for each topic in this table due to space limitations.

Topic	TLDA	TVB$_{\delta=0.8}$
1	#rio #badminton #olympics #iamteamgb wei	#badminton #rio #mas #olympics wei chong
2	#jupiter #juno @nasa orbit @nasajuno	#juno burn engine complete unlock #jupiter
3	#nbasummer nba #basketball @nba basketball	nba #basketball sign wire basketball
4	@gameofthrones #emmys season outstanding	thanks @gameofthrones #iamteamgb #emmys
5	#rncincle trump speech melania donald	#rncincle trump @realdonaldtrump speech
6	#rio #badminton #iamteamgb team gold	#iamteamgb win medal #rio @teamgb
7	#iamteamgb #theresamaypm thanks #jupiter	#theresamaypm watch #brexit minister prime
8	thrones game pokemon season like #pokemon	pokemon basketball team usa #pokemon news

in multi-theme topics. The MD results confirm that this single topic assignment indeed causes multi-theme topics, which is the reason why the TLDA models exhibit a very high mixing degree. Besides, we notice that the TOT models have the highest topical mixing degree. This might be caused by the strong time bias in the sampling approach. Consider that if two topics have similar trends (topical proportions over time, e.g. #nba and #pokemongo in Figure 1), it is likely that these two topics would mix, and thus it causes the generation of multi-theme topics in the TOT models. In this situation, reducing the importance of the time statistics by the balance parameter δ is equally increasing the importance of the words statistics (the first two items in Equation 6), hence avoiding the negative bias of time statistics.

To verify that our generated TVB models have less multi-theme topics than TOT & TLDA, we also conduct a user study to compare the mixing degree of their generated topic models. Since the MD scores of the TOT models are significantly higher than those of the TLDA models, we choose to compare the mixing degree between the TVB$_{\delta=0.8}$ and TLDA models using human judgements. If the users confirm that the TVB approach generates less multi-theme topics than TLDA, it is reasonable to conclude that the TVB approach generates less multi-theme topics than TOT. In our user study, we ask 8 expert end-users whether a given topic contains multiple themes. Specifically, both the TVB and TLDA approaches generate 10 models. We pair these 20 models randomly and generate 10 pairs, where each pair has one model from TVB and another one from TLDA. For each pair, we present a human with all the generated topics of the 2 models. The human is asked to identify all of the multi-theme topics from 2 given models (10 topics per model). A model in a pair is preferred (i.e. obtains a vote), if a human finds less multi-theme topics in this model pair. Each pair gets 3 judgements from 3 different humans. An approach gets a credit if its model in a pair obtains the majority votes out of 3. In the end, among the 10 pairs, our TVB approach gets 7 credits while the TLDA approach gets 2 credits, expect that 3 humans do not have agreement on one pair out of the 10. This user study confirms the results we obtained from MD that our TVB approach generates less multi-theme topics. We list two topic examples of our TLDA and TVB$_{\delta=0.8}$ models in Table 2. Both models generate human interpretable topics. However, we can see more multi-theme topics in the TLDA models, such as "badminton"(topic 1), "teamgb" (topic 6), "theresamaypm" (topic 7) and "pokemon"(topic 8), while the TVB model has less multi-theme topics: "gameofthrone" (topic 4) and "pokemon"(topic 8). In fact, it is easy to mix the topics “theresamaypm” and “teamgb” since they are all popular topics in the UK, and it is possible that the word usage in these two topics is similar. However, the topical trends of these two topics are not similar: “theresamaypm” was popular around 11/07/2016 when Theresa May became the new UK Prime Minister, while “teamgb” was highly discussed during the Olympic Games (from 05/08/2016 to 21/08/2016) (See the topical trends in Figure 1). Our TVB approach can identify these different topical trends by integrating time.

Topical Trends Estimation Error. Both the TOT and TVB approaches estimate topical trends. The `ERR` metric indicates the distance between the real topical trends and the estimated ones (smaller distances are better). The `ERR` scores in Table 1 suggest that our TVB approach generates significantly more accurate topical trends than the TOT approach. The main reason is that the TOT approach has a very high mixing degree (see Table 1), which shows that it has more multi-theme topics similar to the TLDA approach. It could be difficult to match the real topics with the generated topics (explained in Section 6), and thus the multi-theme topics in the TOT model result in less accurate topical trends. Unlike the TOT/TLDA approaches, our TVB model has less multi-theme topics, which results in a more accurate estimation of the topical trends. In Figure 1, both chosen models have duplicated topics, which are `#badminton` & `#juno` and `#gameofthrone` & `#juno` in the TOT and TVB models, respectively. Since the TOT models have more multi-theme topics, it is difficult to match the generated topics with the real ones. For example, the topic theme `#nba` is mixed with `#pokemongo` in the TOT model. As a result, the estimated trend of TOT for `#nba` is not accurate. Although both the TOT and TVB models do not exactly fit the real topical trends using Beta distributions, it is still clear that the estimated trends from the TVB model are closer to the real trends than those of the TOT model as illustrated in Figure 1.

Apart from the used three metrics, it is worth recalling that all the VB and TVB models in our experiments are obtained by setting the maximum iteration to 10, while it is set to 50 for the TLDA, TOT & Gibbs models (see Section 6). Using less iterations, our TVB approach can still provide very competitive results, which indicates its advantage in terms of convergence speed.

8 Conclusions

In this paper, we proposed a time-sensitive Variational Bayesian (TVB) topic modelling approach to improve the quality of generated topics and to estimate topical trends by leveraging the time dimension of social media posts. Our proposed TVB approach, extending the traditional Variational Bayesian approach, employed a Beta distribution to integrate time, where the time statistics were controlled by a balance parameter to alleviate bias. Through experimentation over a ground truth Twitter dataset covering 8 hashtag events, we showed that the time dimension helps to generate more coherent topics in our models with the set balance parameter. Backed by a user study, we find that our TVB approach generated less mixed topics compared with two state-of-the-art baselines. Moreover, our TVB approach can more accurately estimate the topical trends of social media posts.

References

1. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.P., Yan, H., Li, X.: Comparing Twitter and traditional media using topic models. In: Proc. of ECIR. (2011)
2. Mehrotra, R., Sanner, S., Buntine, W., Xie, L.: Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In: Proc. of SIGIR. (2013)
3. Fang, A., Ounis, I., Habel, P., Macdonald, C., Limsopatham, N.: Topic-centric classification of Twitter user's political orientation. In: Proc. of SIGIR. (2015)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research **3** (2003) 993–1022
5. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences **101** (2004) 5228–5235
6. Blei, D.M., Jordan, M.I.: Variational methods for the dirichlet process. In: Proc. of ICML. (2004)
7. Wang, X., McCallum, A.: Topics over time: a non-markov continuous-time model of topical trends. In: Proc. of SIGKDD. (2006)
8. Hong, L., Dom, B., Gurumurthy, S., Tsioutsiouliklis, K.: A time-dependent topic model for multiple text streams. In: Proc. of SIGKDD. (2011)
9. Cheng, X., Yan, X., Lan, Y., Guo, J.: Btm: Topic modeling over short texts. In: Proc. of TKDE. (2014)
10. Yan, X., Guo, J., Lan, Y., Xu, J., Cheng, X.: A probabilistic model for bursty topic discovery in microblogs. In: Proc. of AAAI. (2015)
11. Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential twitterers. In: Proc. of ICWSM. (2010)
12. Braun, M., McAuliffe, J.: Variational inference for large-scale models of discrete choice. Journal of the American Statistical Association **105** (2010) 324–335
13. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proc. of ICML. (2006)
14. Guolo, A., Varin, C., et al.: Beta regression for time series analysis of bounded data. The Annals of Applied Statistics **8** (2014) 74–88
15. Asuncion, A., Welling, M., Smyth, P., Teh, Y.W.: On smoothing and inference for topic models. In: Proc. of CUAI. (2009) 27–34
16. Sridhar, V.K.R.: Unsupervised topic modeling for short texts using distributed representations of words. In: Proc. of NAACL-HLT. (2015)
17. Nguyen, D.Q., Billingsley, R., Du, L., Johnson, M.: Improving topic models with latent feature word representations. In: Proc. of TACL. (2015)
18. Li, C., Wang, H., Zhang, Z., Sun, A., Ma, Z.: Topic modeling for short texts with auxiliary word embeddings. In: Proc. of SIGIR. (2016)
19. Wang, C., Blei, D., Heckerman, D.: Continuous time dynamic topic models. In: Proc. of CUAI. (2008)
20. Hoffman, M., Bach, F.R., Blei, D.M.: Online learning for latent dirichlet allocation. In: Proc. of NIPS. (2010)
21. Johnson, N.L., Kotz, S., Balakrishnan, N.: Chapter 21: beta distributions. Continuous Univariate Distributions Vol. 2 (1995)
22. Fang, A., Macdonald, C., Ounis, I., Habel, P.: Topics in tweets: A user study of topic coherence metrics for Twitter data. In: Proc. of ECIR. (2016)
23. Fang, A., Macdonald, C., Ounis, I., Habel, P.: Using word embedding to evaluate the coherence of topics from twitter data. In: Proc. of SIGIR. (2016)
24. Fang, A., Macdonald, C., Ounis, I., Habel, P.: Examining the coherence of the top ranked tweet topics. In: Proc. of SIGIR. (2016)
25. AlSumait, L., Barbará, D., Gentle, J., Domeniconi, C.: Topic significance ranking of LDA generative models. In: Proc. of ECMLPKDD. (2009)