# A Non Parametric Theme Event Topic Model for Characterizing Microblogs

**Himabindu Lakkaraju**
IBM Research - India
Manyata Embassy Business Park
Bangalore, Karnataka - 560045
klakkara@in.ibm.com

**Hyung-Il Ahn**
IBM Research - Almaden
650 Harry Road
San Jose, California 95120-6099
hiahn@us.ibm.com

## Abstract

In recent times, *microblogging* sites like Facebook and Twitter have gained a lot of popularity. Millions of users world wide have been using these sites to post about topics that interest them and also to voice their opinions on several current events. In this paper, we present a novel non-parametric probabilistic model called Theme-Event Model (TEM) for analyzing the content on microblogs. We also describe an online gibbs-sampling based inference algorithm for this model that enables its usage on large scale data. Experimentation carried out on real world data extracted from Facebook and Twitter demonstrates the efficacy of the proposed approach for characterizing content on microblogs.

## 1 Introduction

Microblogging sites like Facebook, Twitter are housing huge volumes of user generated content. The content being posted on these sites ranges from *themes* that interest individual users (for ex. football) to responses to *events* that occur (for ex. demise of michael jackson). Analyzing the content on these sites poses new challenges and opportunities because of the scale of the data and also the nature of the content. The content on these microblogging sites is a product of an intricate interplay of themes of individual interests and events occuring across the world. To illustrate, let us consider a microblogger who is an avid football fan, it is most likely that this user will author several posts related to football. Thus, one of his themes of interest is 'football'. Further, when a particular event relevant to this theme occurs for ex. FIFA worldcup, it is likely that this user is going to post about this event. Thus, the postings on microblogs can be seen as a combination of theme related content and user responses to events occuring at a particular instant in time. Characterizing content on these microblogs can benefit from capturing the interplay of these two aspects. Further, the nature of the content on these sites is different from traditional text in the sense that each post typically has relatively few words and the syntax of the language used by the microbloggers is certainly not as refined as that which can be found in traditional blogs or news articles.

Most of the work involving analysis of the content on microblogs [4] relies on traditional text analysis techniques like Latent Dirichlet Allocation [2]. A few variants of traditional algorithms have been proposed to tackle microblogs [6]. However, the methodology proposed makes use of meta-information like hashtags which are specific to certain microblogging sites. In this work, we attempt to overcome these limitations by proposing a rich non-parametric probabilistic model TEM that effectively captures the generation of the content on microblogs. Further, we describe an online gibbs sampling based inference algorithm for the proposed model and experiment with real world datasets in order to study the effectiveness of the proposed methodology.

## 2 Our Approach

In this section, we discuss the approach that we propose for modeling the generation of the content in microblogs. We assume that each post is generated by a single underlying theme (for ex. sports) and a single underlying event (for ex. FIFA world cup). Note that this is different from the traditional

admixture modeling [2], which associates a distribution over multiple topics for each document. However, considering the terseness of posts on microblogs, we conjecture that it is unlikely for such posts to correspond to multiple themes and events. We further employ Dirichlet Process [7] for modeling the generation of the themes and events. The model that we describe can be essentially broken down into three components - Identifying underlying theme of a post; Identifying the event associated with the post; Modeling each post as a theme/event mixture. We describe each of these components in detail below.

**Identifying underlying theme :** One of the most significant factors influencing the generation of a post is the *theme*. Each post has some underlying theme that is manifested into words. For ex. posts like 'watching a football match today', 'just finished playing football match', 'excited to watch NFL finals today' carry an underlying theme which is 'football'. When we see posts corresponding to such underlying theme from a user, it is easy to conclude that the user is interested in football. Thus, themes are governed by individual user interests. These user interests in turn can be inferred from the user's previous postings. In order to assimilate this aspect into our approach, we model the probability that the post $p$ authored by user $u$ is generated by the theme $k$ - $P_Z(z_{u,p} = k|z_{u,1:(p-1)}, \alpha_Z)$ as a dirichlet process [7] -

$$P_Z(z_{u,p} = k|z_{u,1:(p-1)}, \alpha_Z) \propto \begin{cases} n_{u,k} & \text{if k}\leq\text{K} \\ \alpha_Z & \text{if k=K+1} \end{cases} \tag{1}$$

$n_{u,k}$ is the number of posts authored by user $u$ which have been assigned theme $k$. $\alpha_Z$ is a positive real value which ensures that a new theme can be sampled if necessary. $K$ denotes the number of themes for which $\sum_{u \in \mathcal{U}} n_{u,k} \neq 0$. As new themes are sampled, the value of $K$ changes. This essentially gives the model flexibility to dynamically increase the number of themes whenever need arises, thus relieving us from the burden of parametrizing the number of themes. Note that while dealing with data sources like Facebook and Twitter, it is extremely tough to assign a value to the parameters like the number of themes and events, hence we resort to a non-parametric prior as discussed above.

**Identifying associated event :** In addition to a broad underlying theme, microtext based postings are usually associated with an event. For instance, though posts like 'following FIFA worldcup', 'excited to watch NFL finals today' relate to the same underlying theme, the events that they refer to are different. In the first case, it is football worldcup, in the second case it is NFL. As discussed above, the theme of a post is governed by user interest, on the other hand, the event which a post refers to is governed by other posts from across the globe during a particular period of time. To illustrate, during the FIFA world cup time frame, a lot of posts mentioned words like FIFA, world cup etc., the volume of these posts indicate that this is a significant event. Thus, an event is governed by world-wide trends, precisely, how many more posts across the globe are referring to the same event. Analogous to the theme identification above, we model the probability that the post $p$ authored by user $u$ is generated by the event $q$ - $P_E(e_{u,p} = q|e_{u\in\text{U},1:(p-1)}, \alpha_E)$ where $\mathcal{U}$ is the set of all the users as a dirichlet process -

$$P_E(e_{u,p} = q|e_{u\in\text{U},1:(p-1)}, \alpha_E) \propto \begin{cases} n_{t,q} & \text{if q}\leq\text{Q} \\ \alpha_E & \text{if q=Q+1} \end{cases} \tag{2}$$

$n_{t,q}$ is the number of posts authored at time instant $t$ that have been assigned the event $q$. Note that this count is different from the count used for theme identification. As discussed, themes are user-centric and hence counts in equation 1 depend upon how many relevant posts $u$ has authored. On the other hand, the counts in equation 2 are not user specific. These counts are taken w.r.t entire set of posts at time instant t. Though $P_Z$ and $P_E$ are both distributions drawn from a dirichlet process, the counts that direct them are different. Further, the time instant t can be chosen to span a convenient period of time like a day, aweek or a month.

**Theme Event Model - Modeling Content Generation on Microblogs :** In this subsection, we put all the pieces described above together and describe the content generation process on microblogs. The generative process is highlighted in Table 1. Following the methodology of topic models [2], we associate a multinomial distribution with each theme $k$ and each event $q$. Each theme $k$ is modeled as multinomial distribution of a finite vocabulary of words, where $\phi^Z_{k,j}$ denotes the probability of the $j^{th}$ word in the vocabulary appearing under the $k^{th}$ theme. Each event $q$ is modeled in an analogous manner. For each post $p$ authored by user $u$, a theme $z_{u,p}$ and an event $e_{u,p}$ are chosen as described in equations 1 and 2. Each post is further associated with a distribution $\pi_{u,p}$ that governs the proportion of theme related words and event related words in it. Each word in the post is either

Table 1: Generative Process of Theme Event Model

1. For each theme $k$,
   Choose $\phi_k^Z \sim Dir(\beta^Z)$
2. For each event $q$,
   Choose $\phi_q^E \sim Dir(\beta^E)$
3. For each post $p$ authored by user $u$,
   a. Choose theme $z_{u,p} \sim P_Z(z_{u,p}|z_{u,1:(p-1)}, \alpha_Z)$
   b. Choose event $e_{u,p} \sim P_E(e_{u,p}|e_{u\in \mathrm{U},1:(p-1)}, \alpha_E)$
   c. Choose $\pi_{u,p} \sim Beta(\alpha_{mix})$
   d. For each word index $n$ of post $p$
      i. Choose $r_{u,p,n} \sim Mult(\pi_{u,p})$
      ii. If $r_{u,p,n} = 1$, choose $w_{u,p,n} \sim \phi^Z_{z_{u,p}}$
          else if $r_{u,p,n} = 2$, choose $w_{u,p,n} \sim \phi^E_{e_{u,p}}$

generated from the theme of the post or the event of the post, the random variable $r_{u,p,n}$ determines if the $n^{th}$ word in a post $p$ authored by a user $u$ is generated by the theme or the event.

## 3 Inference

In this section, we present the inference algorithm for the proposed approach. The inference task here is to compute the conditional distribution over the set of hidden variables for all the word occurrences in the collection of posts. Exactly computing this distribution is intractable. Approximation tecnhniques like collapsed gibbs sampling [3] and variational inference [2] can be applied, however, these techniques involve making multiple passes over the entire corpus and hence are not very suitable for large scale data. Here, we employ forward sampling [1] which bases the estimates of the hidden variables on the data encountered so far. This provides us with a suitable process for carrying out the inference at the expense of suboptimal estimates at the beginning of the markov chain. This forward sampling based approach concurs with the online streaming of the data that is encountered on social media where data keeps pouring in and predictions should be made based on the data seen so far. This renders the proposed approach truly online and scalable for streaming data.

**Update Equations :** At the heart of our inference procedure lies collapsed gibbs sampling technique [3]. For each post $p$ authored by user $u$, there are two hidden variables (after integrating out intermediate parameters) that need to be estimated - $z_{u,p}$, the underlying theme of the post; $e_{u,p}$, the event associated with the post. Further, for each word $w_{u,p,n}$ in the post, the value of $r_{u,p,n}$ needs to be estimated. This parameter determines if the word is generated by the theme of the post or the event of the post. $z_{u,p}$ can be sampled as in equation 1 and $e_{u,p}$ can be sampled as in equation 2. The update equation for $r_{u,p,n}$ is given below :

$$P(r_{u,p,n} = c|\cdot) \propto \begin{cases} (nr_c^{u,p} + \alpha^c_{mix}) \frac{\left(n_v^{z_{u,p}} + \beta^Z\right)}{\left(\sum_{r=1}^{V}(n_r^{z_{u,p}} + \beta^Z)\right)} & \text{if c = 1} \\ (nr_c^{u,p} + \alpha^c_{mix}) \frac{\left(n_v^{e_{u,p}} + \beta^E\right)}{\left(\sum_{r=1}^{V}(n_r^{e_{u,p}} + \beta^E)\right)} & \text{if c = 2} \end{cases} \quad (3)$$

Here, $nr_c^{u,p}$ is the number of words in the post $p$ authored by user $u$ that have been associated with $c$ (note that c=1 corresponds to theme and c=2 corresponds to event). The word $w_{u,p,n}$ corresponds to the $v^{th}$ word of the vocabulary. $n_v^{z_{u,p}}$ corresponds to the number of times word $v$ has been assigned theme $z_{u,p}$. Similarly, $n_v^{e_{u,p}}$ denotes the number of times word $v$ has been assigned event $e_{u,p}$.

**Implementation :** We have experimented with Twitter data of about 20 million tweets. Handling this kind of data automatically calls for distributed inference. We ran all our experiments on 32 GM RAM, 8-core machines. All the operations are split across 7 threads with a master thread synchronizing all the operations. Data is read by the master thread 35K posts at a time and split equally across all the threads with 5K posts being handled by a single thread. Each thread maintains its own copy of the state that it modifies during the course of execution. 100 gibbs iterations are run over each batch of 5K posts in every thread. Thread states are synchronized upon completion of each such batch iteration.

## 4 Experimental Results

In this section we discuss in detail the experiments that we carried out using the proposed model on real world social media datasets extracted from Facebook (300K posts obtained by extracting feeds from publicly available profiles over a span of three months) and Twitter (a subset of 20 million tweets crawled over a time span of about 2 months, [5]). We present here two different kinds of

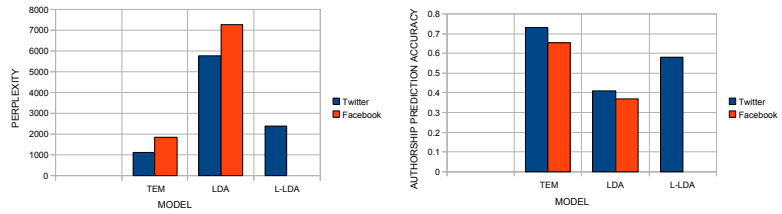


Figure 1: a. Perplexity of various approaches b. User Authorship Prediction Accuracy

experimental results - Perplexity evaluation and User authorship prediction and a brief description of some additional experimentation.

**Perplexity Evaluation :** Perplexity is one of the most widely employed empirical measure to detect how well a given model will be able to generalize to the test data. Though in our case, the inference as such is online and does not distinguish data as training set and test set, this metric is still valid when we are trying to argue how well the proposed approach would incorporate incoming data. The perplexity, used by convention in language modeling, is monotonically decreasing in the likelihood of the test data, and is algebraically equivalent to the inverse of the geometric mean per-word likelihood. In order to compute model perplexity, we consider a sample of 3 million tweets spread out across last 15 day period and a sample of 40K posts from the last months activity on Facebook. Figure 1a. depicts these measurements on Twitter and Facebook data for our model TEM and baselines Labeled-LDA [6] and LDA [2]. It can be seen that TEM gives the least perplexity in both the cases. Note that Labeled-LDA [6] cannot be applied for Facebook data as meta information like hashtags are not available on Facebook.

**Predicting User Authorship :** Our approach TEM mainly aims at characterizing content on microblogs. However, this need not be an end result in itself. We briefly describe here the quantitative evaluation we carried out on a prediction task. The question that we aim at answering is *Given a post p and user u, is user u likely to be the author of post p ?* A question like this would actually help us evaluate the proposed approach quantitatively on real world data. We perform this task both on Facebook and Twitter datasets. In case of Twitter, a sample of about 3M tweets from the last 15 days of the crawled data is considered as the test set. For Facebook, we consider a sample of about 40K posts spanning the last one month as the test set.

*Evaluation* In order to predict whether a user is the author of a particular post, we make use of k-nearest neighbor classifier (k=5). To create a training set for this classifier, for each user, we consider all the posts that the user has authored as positives, and further create a negative set for training the classifier by sampling an equal number of posts from themes that the user has not authored in the recent past ( last 1 month ). Now, we consider each post that the user has authored from the test sample and predict the algorithms response. In order to perform the classification, we consider the following features of the posts - Theme of the post ($z_{u,p}$), Event associated with the post ($e_{u,p}$), mixture distribution of the post $\pi_{u,p}$ and Time of authorship. KL-divergence is used to measure the distance between the corresponding theme distributions, event distributions, mixture distributions $\pi$ and the difference in the number of days is used as the distance metric for time. In case of the baselines LDA and Labeled LDA, we consider the post level topic distribution $\theta_p$ as a feature along with the time of authorship.

*Discussion* We compute the accuracy metric for the classification task. The results are presented in Figure 1b. It can be seen that our approach TEM performs significantly better on both Twitter and Facebook datasets. As can be seen, standard topic modeling baselines do not perform very well on this task. Labeled-LDA performs better than LDA, but in spite of using hash-tags, is significantly outperformed by our approach.

**Additional Experimentation :** TEM can be used for a wide variety of purposes like analyzing themes of interest for individual users and to capture events occuring at a particular point in time. We performed experimentation related to both these aspects. Due to space constraints, a detailed account of this experimentation is not provided here. For evaluating user interests, tweets from a set of 12 Twitter users was collected for a period of 15 days and then based upon their activity, the algorithm put forth a few themes that the users may be interested in. These users evaluated the themes output by the algorithm and about 76.32% of the themes suggested by the algorithm were interesting to the users. Further, while analyzing the events that our approach discovered, several significant events like demise of michael jackson, release of certain new movies etc. were well captured.

## References

[1] A. Ahmed, Y. Low, M. Aly, and V. Josifovski. Scalable distributed inference of dynamic user interests for behavioral targeting. In *ACM SIGKDD Conference on Knowledeg Discovery and Data Mining*, pages 373–382, 2011.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 2004.

[4] L. Hong and B. Davidson. Empirical study of topic modeling in twitter. In *KDD Workshop on Social Media Analytics*, 2010.

[5] J.Yang and J.Leskovec. Patterns of temporal variation in online media. In *ACM International Conference on Web Search and Data Mining*, pages 373–382, 2011.

[6] D. Ramage, S. T. Dumais, and D. J. Liebling. Characterizing microblogs with topic models. In *International AAAI Conference on Weblogs and Social Media*, 2010.

[7] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 2006.