5-2013

It Is Not Just What We Say, But How We Say Them: LDA-based Behavior-Topic Model

Minghui QIU
Singapore Management University, minghui.qiu.2010@smu.edu.sg

Feida ZHU
Singapore Management University, fdzhu@smu.edu.sg

Jing JIANG
Singapore Management University, jingjiang@smu.edu.sg

It Is Not Just What We Say, But How We Say Them: LDA-based Behavior-Topic Model

Minghui Qiu * Feida Zhu * Jing Jiang *

Abstract

Textual information exchanged among users on online social network platforms provides deep understanding into users' interest and behavioral patterns. However, unlike traditional text-dominant settings such as offline publishing, one distinct feature for online social network is users' rich interactions with the textual content, which, unfortunately, has not yet been well incorporated in the existing topic modeling frameworks.

In this paper, we propose an LDA-based behavior-topic model (B-LDA) which jointly models user topic interests and behavioral patterns. We focus the study of the model on online social network settings such as microblogs like Twitter where the textual content is relatively short but user interactions on them are rich. We conduct experiments on real Twitter data to demonstrate that the topics obtained by our model are both informative and insightful. As an application of our B-LDA model, we also propose a Twitter followee recommendation algorithm combining B-LDA and LDA, which we show in a quantitative experiment outperforms LDA with a significant margin.

1 Introduction

Since its advent in [1], LDA (Latent Dirichlet Allocation) has been widely used for topic modeling in various domains. Variants of LDA have been proposed to enhance the model to address different challenges [2; 3; 4]. Existing variants have mostly focused on the textual content as the subject of the topic modeling, be it a set of news articles, web-blogs, micro-blogs, etc., which is perfectly fine if only the topics of the text body are of interest. However, in many applications, it is also interesting to study the context in which the text is generated, consumed and interacted with, especially in online settings where text is an integral part of the social interactions. The way people interact with the text is critical in understanding user behavior patterns and modeling user interest in social network analysis. In a word, what is important is not just what we say, but how we say them as well.

*School of Information System, Singapore Management University

Take Twitter, the most popular micro-blogging service, for example. There are essentially four ways Twitter users interact with tweets, which are "post," "retweet," "reply" and "mention." We call these interactions the behavioral information associated with the textual content of tweets. While the textual content indicates the topics of interest to users, these rich behavior information could provide insight into user's online social personalities and behavioral profiles. The benefits of integrating behavioral information into topic modeling can be summarized from the following three aspects.

Firstly, user groups with similar topics of interest but different behavioral patterns can be identified, which is important for building more accurate user models for online social profiling. By applying LDA on a Twitter data set based in Singapore, we selected a set of Twitter users related to Singapore politics and having a similar number of tweets. Figure 1 shows four users selected from this set who, by just examining their topic distributions computed by traditional LDA, are hardly distinguishable because of their almost identical topics of interest. However, if we look into their behavioral patterns, drastic differences can be easily observed. For example, *Fake_PMLee* is an active user who mostly publishes his own original tweets, oftentimes jokes, about Singapore news, while *YamKeng* is a Singapore Member of Parliament who often engages in conversations with others on Twitter to directly reach out to individuals. Clearly these different behaviors they exhibit on Twitter suggest their different motivations in using the platform.

Secondly, the user clusters with distinct behaviorial patterns usually represent different user profiles that are easily identifiable. For example, Table 1 lists the top 5 users for the dimension of "post" (PO) in the same Twitter data on Singapore politics. It is easy to note that these users form a group of coherent behavioral patterns: they post a lot of original tweets and seldom engage in interactions like reply, retweet and mention with other users. Closer examination reveals that they are all official news media accounts.

Thirdly, as shown by our experiments, more ac-

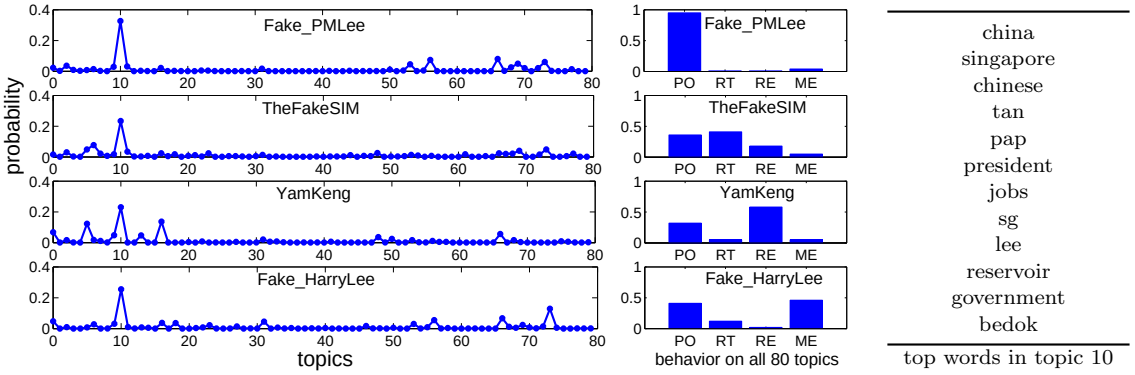


Figure 1: Users with similar topic interests but different behavioral patterns and top words in topic 10. PO: post. RT: retweet. RE: reply. ME: mention.

curate recommendation of users to follow can also be achieved by identifying the cluster of users who are more behavior-driven in following others.

name	PO	RT	RE	ME	Descriptions
sg_story	0.9988	0.0004	0.0004	0.0004	SG News
SGnews	0.9978	0.0005	0.0003	0.0014	SG News
sgdrivers	0.9969	0.0011	0.0010	0.0010	News on traffic etc.
singaporesurf	0.9954	0.0016	0.0015	0.0015	News media
tocsg	0.9952	0.0014	0.0016	0.0018	TheOnlineCitizen

Table 1: Top-5 users in "PO = post" dimension. The 4 columns in the middle show the probabilities of each posting behavior for each user.

Micro-blogging service is just one example of the many scenarios where behavioral and textual information are integrated. Wikipedia articles are collaboratively edited in a number of ways; News articles online can be thumbed up or thumbed down, and, with a simple click, shared to various other social platforms; Reviews on products and services are rated, bookmarked and recommended to friends.

Summing up, in traditional text-dominant settings, users' rich interaction with the textual information, unfortunately, has not yet been well incorporated in the existing topic modeling techniques proposed for these challenges. In this paper, we propose an LDA-based behavior-integrated topic model, called B-LDA, which jointly models the topic interests and interactions of a user with the topics. To the best of our knowledge, this is the first topic model to incorporate user interaction into the modeling of topics on a text corpus. We demonstrate the usefulness of our B-LDA model by evaluating the model in both topic analysis and followee recommendation. Our experiments on real Twitter data show that B-LDA can not only qualitatively uncover more informative topics, but also quantitatively provides better followee recommendation for behavior-driven users.

2 Model

In this section, we present our joint behavior-topic model. We first give a brief review of the LDA model and its variant T-LDA for micro-blogging settings, then present our proposed model B-LDA.

2.1 LDA and its variants

LDA [1] has been widely used in textual analysis [2; 3; 4]. The original LDA is used to find hidden 'topics' in the documents, where a topic is a subject like 'arts' or 'education' that is discussed in the documents. After applying the model, each document can be represented in the semantic topic space which is a lower dimensional space. In this case, documents are featured by their semantic meaning, which can help many tasks including text classification, document clustering, information retrieval, etc.

While the literature has witnessed the successful application of LDA on traditional documents like news articles, it is still an open and yet popular research question on whether LDA and its variants will work on micro-blogs like Twitter [5; 6; 7]. The original setting in LDA, where each word has a topic label, may not work well with Twitter as tweets are short and a single tweet is more likely to talk about one topic. Hence, Twitter-LDA (T-LDA) [7] has been proposed to address this issue. T-LDA also addresses the noisy nature of tweets, where it captures background words in tweets. As experiments in [7] have shown that T-LDA could capture more meaningful topics than LDA, we extend it to jointly model the topic interests and behaviors of a user in micro-blogs like Twitter.

2.2 LDA-based Behavior-Topic Model

Table 2 summarizes the set of notations and descriptions of our model parameters.

Notations	Descriptions
U	the total number of users
N_u	the total number of tweets by user u
$L_{u,n}$	the total number of words in u's n-th tweet
T	the total number of topics
V	the vocabulary size
b	a behavior in $\mathcal{B} = \{post, retweet, reply, mention\}$
y	a switch
z	a topic label
ϕ_t	topic-specific word distribution
ψ_t	topic-specific behavior distribution
ϕ'	background word distribution
θ_u	user-specific topic distribution
φ	Bernoulli distribution
α, η, β', β, γ	Dirichlet priors

Table 2: Notations and descriptions.

We now present our B-LDA model. First, we assume that there are T hidden topics, where each topic has a multinomial word distribution ϕ_t and a multinomial behavior distribution ψ_t. Each tweet has a single hidden topic which is sampled from the corresponding user's topic distribution θ_u ($1 \le u \le U$). We further assume that given a tweet with hidden topic t ($1 \le t \le T$), the words in this tweet are generated from two multinomial distributions, namely, a background model and a topic specific model. The background model ϕ' generates words commonly used in many tweets; they are similar to stop words. The topic specific model ϕ_t generates words related to topic t. When we sample a word w ($1 \le w \le V$), we use a switch $y \in \{0, 1\}$, which is sampled from a Bernoulli distribution φ, to decide which word distribution the word comes from. Specifically, if $y = 0$, the word w is sampled from ϕ'; otherwise, it is sampled from ϕ_t. We also assume the behavior pattern b ($b \in \mathcal{B}$) is sampled from the behavior distribution ψ_t. Lastly, we assume θ_u, ψ_t, ϕ', ϕ_t and φ have Dirichlet priors α, η, β', β and γ respectively. Figure 2 shows the plate notation of the model. The generative process for all posts is in Figure 3.

In our model, we assume a universal behavior distribution instead of a personalized behavior distribution for each topic, as the former ensures the behavior information is a property of "topic." In this case, both a user's personal behaviors and topic interests can be reflected by looking at her "topic" distribution. In other words, a "topic" in our model is with behavior pattern, which is studied in Section 4.1. Note that our model is designed mainly for the settings where the text is mostly short and about a single topic. It is not hard to modify our model to remove this special assumption, and we leave it to our future work to design a more general topic-behavior model.

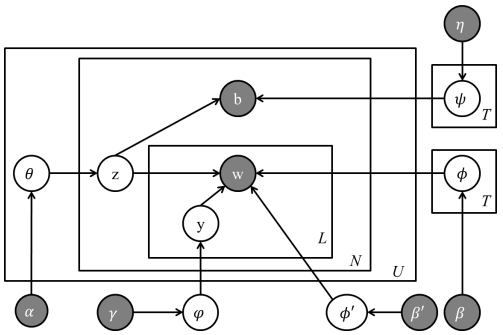


Figure 2: LDA-based behavior-topic model (B-LDA)

- For each topic $t = 1, \cdots, T$
  - Draw $\psi_t \sim \text{Dir}(\eta)$, $\phi_t \sim \text{Dir}(\beta)$
- Draw $\phi' \sim \text{Dir}(\beta')$, $\varphi \sim \text{Dir}(\gamma)$
- For each user $u = 1, \cdots, U$
  - Draw topic distribution $\theta_u \sim \text{Dir}(\alpha)$
  - For u's n-th tweet, $n = 1, \cdots, N_u$
    - Draw a topic $z_{u,n}$ from θ_u
    - For each word $l = 1, \cdots, L_{u,n}$
      - Draw $y_{u,n,l}$ from Bernoulli(φ)
      - Draw $w_{u,n,l} \sim \phi'$ if $y_{u,n,l} = 0$, otherwise draw $w_{u,n,l} \sim \phi_{z_{u,n}}$
    - Draw a posting behavior $b_{u,n} \sim \psi_{z_{u,n}}$

Figure 3: The generative process for all posts in B-LDA.

Learning and Parameter Estimation

We use collapsed Gibbs sampling to obtain samples of the hidden variable assignment and to estimate the model parameters from these samples. Due to space limit, we leave all the detailed derivation and running time analysis to the supplementary pages[1].

With Gibbs sampling, we can make the following estimation of the model parameters:

$$\theta_{u,t} = \frac{n_u^t + \alpha}{\sum_{t=1}^{T} n_u^t + T\alpha}, \quad \text{user-topic distribution} \tag{2.1}$$

$$\psi_{t,b} = \frac{n_t^b + \eta}{\sum_{b=1}^{B} n_t^b + B\eta}, \quad \text{topic-behavior distribution} \tag{2.2}$$

$$\phi_{t,w} = \frac{n_{t,y=1}^w + \beta}{\sum_{w=1}^{V} n_{t,y=1}^w + V\beta}, \quad \text{topic-word distribution} \tag{2.3}$$

where n_u^t is, when given the user u, the number of times t is sampled, n_t^b is the number of times behavior b co-occurs with topic t, and $n_{t,y=1}^w$ is, given the topic t, the number of times w is sampled as topical word.

[1] http://www.mysmu.edu/phdis2010/minghui.qiu.2010/papers/BLDA_supp.pdf

3 Followee Recommendation with B-LDA

In this section, we look at how B-LDA can be applied for an important task on Twitter — followee recommendation, i.e., recommending who to follow on Twitter, by making use of the model parameters. Note that our aim is not to propose a new recommendation model, but rather to study how the behavior aspect of users, and accordingly our proposed B-LDA model, can be applied in recommendation to make a difference.

Existing studies for followee recommendation essentially focus on the textual content of either the target user herself or her followees [8; 9]. This works well when users follow others only based on whether they share similar topic of interests, regardless of how they interact with the topics. However, our observation is that it is not true for all the users, which seems to echo the findings in [10] that Twitter functions both as a news media and a social network. For example, some users prefer to follow users who always generate original tweets. These users tend to use Twitter more like an information source and news media. In contrast, some other users prefer to follow and interact with users who are also heavily engaged in retweeting and replying others. To them, Twitter's social network aspect is more valued. These observations drive home an important message: behavior information is also a factor when users decide who to follow. We refer to the kind of users who care about the behavioral patterns of their followees, explicitly or implicitly, as *"behavior-driven"*.

3.1 Behavior-driven index β_K

To capture the behavior factor in users' following style, we propose a new index to measure the extent to which a user is behavior-driven follower. The index is based on the following intuition: if a user tends to follow users with certain behavioral patterns, the set of all her followees will naturally form a small number of clusters within each of which the followees would share similar behavioral patterns. This gives us the idea of using k-nearest-neighbor to measure the modularity of a user's followee set in the joint behavior-topic distribution space to gauge the user's behavior-driven index.

Given a user space S defined by user topic distribution and a user $v \in S$, let δ_v^K be the set of the k-nearest-neighbors of v. The k-nearest-neighbor distance (D_{knn}) for a single user v is defined as follows:

$$D_{knn}(\mathcal{S}, K, v) = \frac{1}{K} \sum_{f \in \delta_v^K} (1 - sim(\theta_v^{\mathcal{S}}, \theta_f^{\mathcal{S}})), \tag{3.4}$$

where K is the neighborhood size, $\theta_v^{\mathcal{S}}$ is v's topic distribution given $\mathcal{S}$, and $sim()$ is cosine similarity.

Equation (3.4) can be used to measure how close any member in a given user u's followee set is to other members in the set. We also need to define the k-nearest-neighbor distance (D_{knn}) for a set of users to measure the modularity of u's followee set as a whole. Given a set of users $\mathcal{U}$, the k-nearest-neighbor distance (D_{knn}) for $\mathcal{U}$ is defined as follows:

$$D_{knn}(\mathcal{S}, K, \mathcal{U}) = \sum_{w \in \mathcal{U}} D_{knn}(\mathcal{S}, K, w). \tag{3.5}$$

Now we are ready to define our behavior-driven index β_K for a given user u, which is the ratio of the k-nearest-neighbor distances of u's followee set in two spaces — $\mathcal{S}_T$ the pure topic space and $\mathcal{S}_B$ the joint behavior-topic space. The idea is that the more behavior-driven a user u is, the closer u's followees will be in the joint behavior-topic space $\mathcal{S}_B$ than in the pure topic space $\mathcal{S}_T$. Given a user u, let $\mathcal{F}_u$ denote u's followee set. β_K is defined as follows:

$$\beta_K = \frac{D_{knn}(\mathcal{S}_T, K, \mathcal{F}_u)}{D_{knn}(\mathcal{S}_B, K, \mathcal{F}_u)}. \tag{3.6}$$

In the experiment section, we find about half of the users are behavior-driven to at least some degree. We then use a threshold τ to draw the definition — if $\beta_K \geq \tau$, we define user u as a behavior-driven follower, and if $\beta_K < \tau$, u is a topic-driven follower. In our experiments, we use our proposed B-LDA to form the joint behavior-topic space $\mathcal{S}_B$. As for $\mathcal{S}_T$, we would use either LDA or T-LDA, whichever gives the better performance. The detailed results are in Section 4.2.

3.2 Followee Recommendation Algorithm

We present a followee recommendation algorithm in Algorithm 1 for a user u and a set of non-followees $\mathcal{T}_u$ to recommend.

Algorithm 1 Followee Recommendation

```
1: Input: user u, followees F_u, neighborhood size K, model M,
   non-followees T_u
2: Output: the ranked users in T_u
3: procedure FeeRec(u, F_u, K, M, T_u)
4:    for each user w in T_u do
5:       Find its K closest followees δ_w^K from F_u
6:       Set distance d(w) as the average of its distances to δ_w^K
7:    end for
8:    Rank users in T_u according to their distance d(·)
9:    return ranked T_u
10: end procedure
```

Based on the β_K index in Equation 3.6, we propose a combined recommendation method by first examining whether a user is a topic-driven follower or a behavior-driven follower, then using the corresponding model to

perform followee recommendation. In Algorithm 1, $\mathcal{M}$ is given, while in the combined approach, $\mathcal{M}$ is obtained by evaluating the given user's β_K index. Specifically, if $\beta_K \geq \tau$, the follower is defined as behavior-driven, and we use $\mathcal{M}_{\text{B-LDA}}$; Otherwise, the follower is topic-driven, and we use $\mathcal{M}_{\text{LDA}}$ or $\mathcal{M}_{\text{T-LDA}}$.

4 Empirical Evaluations

In this section we present our empirical study of B-LDA for two application domains: (I) Topic Analysis and (II) Followee Recommendation.

Data setup

We use real Twitter data to evaluate our proposed model. Our base data set contains 151,055 Singapore-based Twitter users and their tweets, which are collected by starting from a seed set of active Singapore users and tracing their follower and followee links up to two hops. From this base set, 5000 users are randomly selected, among whom 1000 are further randomly selected to obtain all their followees, which makes a total of 9688 users. A total of 11,882,441 tweets of these 9688 users published between September 1 and November 30, 2011 are used in our experiments. For the application of followee recommendation, we provide recommendations for the randomly selected 1000 users.

We compare our B-LDA model with LDA and T-LDA in our study. For all the models, the number of topics T is set as 80, α is $50/T$ and β is 0.01. For B-LDA, γ is set as 10, β' is 0.1, η is 0.01. Each model runs for 400 iterations of Gibbs sampling. We take 40 samples with a gap of 5 iterations in the last 200 iterations to assign values to all the hidden variables. Below we first present topic analysis on behavior dimension, followed by a quantitative evaluation on fee recommendation.

4.1 Topic Analysis

To see how integrating behavioral information into topic modeling could make a difference, we show some empirical studies on topics obtained by B-LDA.

Topics grouped by dominant behavior

In contrast to LDA, B-LDA generates topics each enhanced by a behavior distribution, which is denoted as $\psi_{t,b}$ in the output. Just like LDA is expected to generate topics each containing words most relevant to a coherent topic, we would like B-LDA to generate topics which are identified with some dominant behavior. To measure in general whether a topic contains some dominant behavior, we use the idea of entropy and identify for each topic its associated behavior distribution. The lower the entropy score, the more dominant the behavior the topic is identified with. For a given topic t, we first give the definition of its entropy $e(t)$ on behavior distribution as: $e(t) = \sum_{b \in \mathcal{B}} -p(b|t) \times \log p(b|t)$, where $\mathcal{B}$ is the set of all possible types of behavior.

How do we identify the associated behavior distribution of a given topic? In B-LDA, $p(b|t)$ is equal to $\psi_{t,b}$ in Equation (2.2). For T-LDA and LDA, we compute $p(b|t)$ as: $p(b|t) = \frac{C(t,b)+\delta}{\sum_{b \in \mathcal{B}} C(t,b)+|\mathcal{B}|\delta}$. Note that, the ways of computing $C(t,b)$ are different in T-LDA and LDA. For T-LDA, $C(t,b)$ is computed by counting all tweets with topic t and behavior b. For LDA, $C(t,b)$ is the number of times a word is labeled as topic t and its corresponding tweet has behavior b. A normalization factor δ is introduced, which is similar to the hyperparameter η in Equation (2.2). To make fair comparison, we set $\delta = \eta = 0.01$.

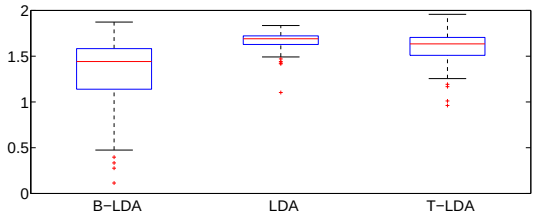
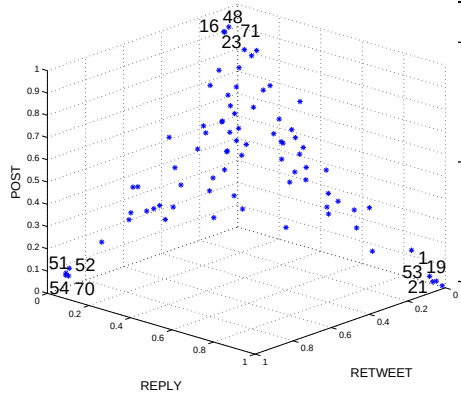


Figure 4: Comparison of topics from B-LDA, LDA and T-LDA in terms of entropy on behavior distribution.

We find B-LDA has a lower entropy score than both LDA and T-LDA as shown in Figure 4, which means topics generated by B-LDA tend to be characterized by a few dominant types of behavior. Note that in LDA, one tweet is associated with multiple topics but with one behavior. In this case, the chance of many topics sharing the same behavior is higher comparing to the setting of one tweet sharing one topic in T-LDA and B-LDA. That is the reason why entropy of topics in T-LDA and B-LDA are with a higher variance than in LDA.

Topics of distinct behavioral pattern

Now we show the topics with distinct behavioral patterns associated, i.e., those ranked top for one type of behavior. Figure 5 shows the distribution of all 80 topics on the behavior dimensions of "PO," "RT" and "RE" together with the top topical words of those ranked top along each behavior dimension. For the "PO" dimension, topic 16 is related to daily news which is mainly contributed by news media accounts who mostly post original tweets. Topic 23 is mostly users' daily personal updates which seldom interest others to retweet or reply. Topic 71 is also related to personal updates, but more on things related to cell phones, laptops, etc. Top-4 topics in the "RT" dimension are topics related to jokes like topic 51 which is a mixture of jokes and funny things



	ID	Top Topical Words	Label
PO	48	today, scorpio, aquarius, aries, pisces, leo, libra	horoscope
	16	#singapore, #news, #business, lee, minister, #local	SG news
	71	phone, omg, time, goona, internet, laptop, home	personal updates
	23	home, time, gonna, back, work, dinner, house	personal updates
RT	51	#sosingaporean, sg, #bvssg, siri, leh, friend, money	jokes on siri etc.
	70	love, people, type, life, person, make, things, smile	popular quotes
	54	people, love, #pisces, #aquarius, #taurus, scorpio	horoscope
	52	super, junior, music, video, snsd, mama, shinee	MAMA concert
RE	21	lol, yeah, :p, good, man, time, nice, bad, thought	-
	19	time, tmr, meet, eh, la, work, school, yeah, free	-
	53	:p, dont, omg, im, ah, lah, sleep, reply, wait, text	-
	1	ur, dun, time, wad, ppl, wan, nt, de, ya, abt, tt	-

Figure 5: Topic distribution on "PO = post", "RE = retweet", and "RE = reply" dimension, and top ranked topics in each behavior and related topical words. 'ID' in the table corresponds to topic id in the left figure. Labels are manually assigned.

shared by user @SoSingaporean and @BvsSG, popular quotes like topic 70, daily horoscope topic 54, and topic 52 on a music event - MAMA concert. We can also tell that topical words used in reply are more informal than the other behavior types which are hard to be labeled.

One interesting observation is that both topic 48 and topic 54 are related to horoscope, but associated with different behavior. One group is mostly original tweets while the other is getting retweeted all the time. Both focused on the topic of horoscope, they would be hardly distinguishable by examining their topical words by LDA or T-LDA. This means this topic of horoscope is split into two topics with different associated behavior in B-LDA. Next, we show more such cases.

Topics split by different behavioral patterns
In order to find topics in T-LDA or LDA that would be split into multiple topics in B-LDA, we first identify relationships among the topics in different models by measuring their topic similarity. In particular, we use KL-divergence on the word distribution of two topics to measure their distance. Specifically, for topic t' in model $\mathcal{M}_1$ and t in $\mathcal{M}_2$, the distance between them is:

$$D(\phi_t^{\mathcal{M}_2}||\phi_{t'}^{\mathcal{M}_1}) = \sum_{w=1}^{V} p(w|\phi_t^{\mathcal{M}_2}) \times \log \frac{p(w|\phi_t^{\mathcal{M}_2})}{p(w|\phi_{t'}^{\mathcal{M}_1})}.$$

We focus on T-LDA as B-LDA is an extension of T-LDA, which suggests a high correlation between the topics in the two models. As KL-divergence is asymmetric, to measure the distance from topic t' in T-LDA to topic t in B-LDA, it is better to use $D(\phi_t^{\text{B-LDA}}||\phi_{t'}^{\text{T-LDA}})$, where $\phi_t^{\text{B-LDA}}$ is computed in Equation (2.3).

For each topic t, we find its precedent topic t^* in T-LDA by finding the topic with the minimum distance. We define a valid 'topic group' in B-LDA as a topic group that contains at least two topics and all the topics in the group share the same precedent topic in T-LDA. As a result, among the 80 topics of B-LDA, we find 16 topic groups. In particular, topic 48 and topic 54 in Figure 5, which are both related to horoscope, are in the same topic group, which means they are indeed merged into one topic of horoscope in T-LDA.

Table 3 presents more such sample topic groups and their topical words. The precedent topic in T-LDA is shown in the first row of each case and the group of topics in B-LDA into which it is split are shown in the second and third row. The table shows topics within the same topic group share similar topical words but are associated with different behavior patterns. For example, topic 13 and topic 16 share common top topical words like 'news' and 'police', but the latter is essentially a topic of original tweets while the former get retweeted almost as much.

Another observation is that the retweet topics tend to contain more hashtags in top words than the topics of other behavior types in the same topic group. Such examples include topic 3 and topic 61 in Table 3, and topic 48 and topic 54 in Figure 5. Note that this observation is not true for topic 13 and 16, where the latter is mainly from original posts but contains more hashtags than topic 13. Close examination shows that topic 16 is mainly contributed by news media accounts like 'YahooSG', while topic 13 is from non-media account. As Twitter hashtags can serve to classify and promote tweets [11], our observation shows that news media accounts tend to use more hashtags in their tweets to propagate and promote them. In general, topics in B-LDA tend to feature more distinct

Model	ID	Top Topical Words	PO	RT	RE	ME
T-LDA	65	#singapore, #news, news, #business, china, #int'l, cna, minister, police, stocks	0.67	0.25	0.05	0.03
B-LDA	13	police, obama, news, occupy, people, street, president, wall, man, video, cna	0.51	0.44	0.03	0.02
	16	#singapore, #news, #business, lee, minister, #local, news, pm, police, s'pore	0.90	0.08	0.01	0.01
T-LDA	62	eat, food, dinner, hungry, chicken, #sosingaporean, curry, lunch, sauce, rice	0.51	0.11	0.33	0.05
B-LDA	3	eat, food, hungry, dinner, chicken, lunch, rice, ice, cream, ate, nice, love, meal	0.59	0.08	0.28	0.05
	61	curry, sauce, indian, #replacesongnameswithcurrysauce, #sosingaporen, #sgedu	0.28	0.59	0.10	0.03

Table 3: Sample topic groups and their topical words and behavioral patterns. 'ID' is topic id.

behavioral pattern than those in T-LDA, which can help identify users groups with distinct behavioral pattern, for example news media accounts.

4.2 Followee Recommendation

We show how our proposed B-LDA model can improve followee recommendation results in this section. The task is to recommend users to follow for a target user u. We randomly pick one followee from u's current followee set, and then combine her with another m ($m = 1000$) randomly-selected users who are not in her followee list. Any recommendation algorithm would generate a ranking of these $m+1$ users according to Algorithm 1, where the higher the real followee is ranked, the better the performance of the recommendation algorithm has. We repeat this process for R ($R = 10$) runs where each run we pick a different followee and obtain an average rank of the real followee. Note that in the algorithm, the distance measure between two users w and f is $1 - sim(\theta_w^{\mathcal{M}}, \theta_f^{\mathcal{M}})$, where $\mathcal{M}$ is a given model and $sim()$ is computed by cosine similarity. Our task is general recommendation evaluation task, if we set R as 1, it is similar to the one studied in [12].

4.2.1 Comparison of Models

We consider LDA and T-LDA as our baselines. The model setting is the same as discussed at the beginning of Section 4. We compare B-LDA with LDA in terms of two criteria: the average rank of the real followee, which is defined as: $\bar{r} = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} \bar{r}(u)$, where $\bar{r}(u)$ is the average rank of the randomly-picked real followees from u's followee list in R runs; and mean reciprocal rank: MRR $= \frac{1}{|\mathcal{U}|}\sum_{u\in\{\mathcal{U}\}} \frac{1}{\bar{r}(u)}$.

The comparison among these models in Table 4 suggests these findings: 1). In general, all these models perform better by setting a smaller neighborhood size K. 2). B-LDA is a direct extension of T-LDA. The fact that it significantly outperforms T-LDA in terms of both real followee's rank and MRR demonstrates the benefit of adding behavior information into topic model for the task. 3). In terms of MRR, B-LDA and LDA report similar performance, which suggests there exist both behavior-driven and topic-driven followers.

K	$\bar{r}$			MRR		
	B-LDA	LDA	T-LDA	B-LDA	LDA	T-LDA
1	**294***	301	302	0.022	0.024	0.016
2	**295***	302	300	0.022	0.024	0.016
3	**298***	305	301	0.021	0.023	0.016
4	**300***	307	303	0.021	0.022	0.015
5	**303***	309	306	0.021	0.022	0.015

Table 4: Comparisons of models by average rank of real followee $\bar{r}$ and MRR. * indicates the result is significantly better than T-LDA at 5% significance level by Wilcoxon signed-rank test.

4.2.2 Evaluation on behavior-driven followers

We show in this part an empirical study on how to choose a suitable K value for β_K index and use it to identify behavior-driven followers.

We use Cohen's Kappa coefficient to measure the correlation between the ranking results and the D_{knn} metric in Equation (3.5). We find that, by setting $K = 1$, D_{knn} has the highest correlation score with the ranking results in both B-LDA and LDA, 0.7 for B-LDA and 0.8 for LDA, and both B-LDA and LDA yield better recommendation results. This shows that the proposed D_{knn} metric provides a good characterization of the modularity of a user's followee set in both topic and joint behavior-topic space by setting $K = 1$. We then use β_1 to judge whether a follower is topic-driven or behavior-driven. Specifically, if user u's corresponding $\beta_1 \geq \tau$ ($\tau = 1$), then u is a behavior-driven follower; otherwise, u is topic-driven follower. Figure 6 shows the histogram of 1,000 target users' β_1 values, binned into intervals of 0.01, where we find 53% of all the target users are behavior-driven followers.

Model	B-LDA	LDA	p-value
$\bar{r}$	**322**	370	4E-033
MRR	**0.0124**	0.0076	N/A

Table 5: Comparisons of B-LDA and LDA by $\bar{r}$ and MRR on behavior-driven followers.

We report the performance of two models on behavior-driven followers in Table 5. On this set of

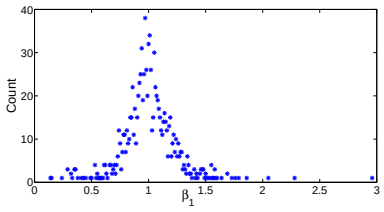


Figure 6: Histogram of β_1 values on target users.

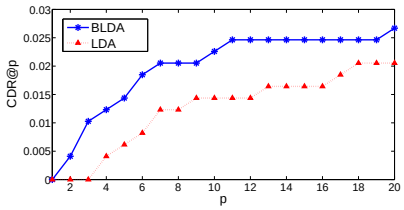


Figure 7: Comparison of B-LDA and LDA by CDR on behavior-driven follwers.

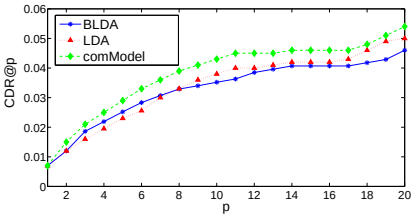


Figure 8: Comparison of comModel, B-LDA and LDA by CDR.

users, B-LDA outperforms LDA in terms of both real followee' rank and MRR score. As MRR score is a single numeric value, we cannot perform significant test on it. For real followee' rank, significant test shows a very low p-value of 4E-033, which means B-LDA significantly outperforms LDA.

We also evaluate the two models in terms of cumulative distribution of ranks (CDR) for real followees. CDR@p is the percentage of users whose real followee is ranked at least at rank p, defined as CDR@p $= \frac{|\{u \in \mathcal{U} | \bar{r}(u) \leq p\}|}{|\mathcal{U}|}$. Figure 7 shows that B-LDA could give a better recommendation for behavior-driven users.

4.2.3 Evaluation on the Combined Approach

We compare the combined model (comModel) proposed in Section 3.2 with B-LDA and LDA on target user set $\mathcal{U}$ in Table 6. The combined model significantly outperforms B-LDA and LDA in terms of both real followee' rank and MRR. We also report cumulative distribution of ranks (CDR) for real followees in Figure 8, from which comModel is observed to provide better recommendations than B-LDA and LDA. In all, the combined model shows a promising followee recommendation results.

Model	B-LDA	LDA	comModel
$\bar{r}$	294	301	**277**†
MRR	0.022	0.024	**0.030**

Table 6: Comparisons of comModel, B-LDA and LDA by average rank of real followee $\bar{r}$ and MRR. † indicates the result is significantly better than all other results at 5% significance level by Wilcoxon signed-rank test.

5 Related Work

As arguably the most popular and representative microblogging service, Twitter has attracted an ever-growing amount of attention from the research community [5; 6; 7; 13]. An important observation in Twitter is that people use the platform for different purposes. For example, as studied in [14], user activities in Twitter can be thought of as information seeking, information sharing or social activity. Similarly, content analysis in [13] reveals that tweets can be categorized from "information sharing" to "self promotion," and two kinds of users are identified: users who pass on non-personal information and users who tweet about themselves. Our study differs from these studies in that we study the textual content at semantic topic level and at the same time examine how users interact with these topics, combining topic discovery with user behavior modeling.

Another observation is that users' behavioral patterns are associated with semantic meanings. In [15], it is found that a user's retweeting behavior is a strong indicator of the user's topical interest. And [16] studies the sources of retweets and proposes a factor graph model to predict user retweeting behavior. In our work, we are looking at how the behavior associated with the textual content could help in a range of applications including topic analysis and followee recommendation. In a recent work in [17], content and user interactions are studied to discover communities in social networks, where they assume a community specific interaction proportion. However, this is mainly designed for finding communities, while in our work, we would like to characterize topics with behaviors, and further characterize users by looking at their topic distribution.

From the topic modeling perspective, LDA (Latent Dirichlet Allocation) has been widely used for topic modeling in various domains. Variants of LDA have been proposed to enhance the model to tackle different tasks including mining online reviews [18; 19; 20], sentiment analysis [21; 22] and community discovery [23; 24]. These existing variants normally look at how texts or links are generated and how to extract opinion words from textual contents. Besides these, it is also important to study the context in which the text is consumed and interacted with as it can help to understand user behavior patterns and model user interest. Another dif-

ference is that, we focus the study of the model on microblogs like Twitter where the textual content is relatively short but user interactions on them are rich.

The work in [7] compares Twitter with traditional news media, where they find Twitter has made itself an important and unique information source on a diverse range of topics which are different from all traditional news media [7]. From the topic perspective, [5] studied the characteristics of tweets and applied labeled-LDA to Twitter, but the model relies on labeled topic types and other information like emoticons, social signals and hash-tags. Our model, on the other hand, is an unsupervised one and studies how users interact with textual content. [6] is an interesting piece of work on finding topic-sensitive influential twitterers based on LDA [1]. The recent work [25] compares LDA and Author-Topic Model [4] on Twitter. It shows the effectiveness of topic modeling especially on real-world classification tasks. The major difference between our model and these topic models in Twitter [1; 4; 7] is that topics in our model are enhanced by behavioral patterns. Experiments show our model can also help to perform better followee recommendation on behavior driven followers. In summary, the novelty of our model lies in the integration of users' topic of interests and their associated behavioral patterns, which, as supported by our experiments, better characterizes users and topics.

6 Conclusion

In this paper, we propose a behavior-integrated topic model based on LDA, called B-LDA, which jointly models user topic interests and behavioral patterns on microblogging services like Twitter. We compare our model with standard LDA as well as Twitter-LDA on real Twitter data. Firstly, experiment results show our model can find topics with dominant behaviors; Secondly, we propose an index β_K to characterize users who are behavior-driven followers, Thirdly, experiment results demonstrate that B-LDA significantly outperforms other models on followee recommendation for these behavior-driven followers; Finally, based on the β_K index, we propose a new recommendation framework combining B-LDA and LDA which gives promising recommendations.

Acknowledgments

This research/project is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office. We thank the reviewers for their valuable comments.

References

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

[2] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 2006, pp. 113–120.

[3] I. Titov and R. McDonald, "Modeling online reviews with multi-grain topic models," in *WWW*, 2008, pp. 111–120.

[4] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths, "Probabilistic author-topic models for information discovery," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2004, pp. 306–315.

[5] D. Ramage, S. T. Dumais, and D. J. Liebling, "Characterizing microblogs with topic models," in *ICWSM*, 2010.

[6] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *WSDM*, 2010, pp. 261–270.

[7] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *Proceedings of the 33rd European conference on Advances in information retrieval (ECIR)*, 2011, pp. 338–349.

[8] J. Hannon, M. Bennett, and B. Smyth, "Recommending twitter users to follow using content and collaborative filtering approaches," in *RecSys*, 2010, pp. 199–206.

[9] S. M. Kywe, E.-P. Lim, and F. Zhu, "A survey of recommender systems in twitter," in *SocInfo*, 2012, pp. 420–433.

[10] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *WWW*, 2010, pp. 591–600.

[11] E. Cunha, G. Magno, G. Comarela, V. Almeida, M. A. Gonçalves, and F. Benevenuto, "Analyzing the dynamic evolution of hashtags on twitter: a language-based approach," in *Proceedings of the Workshop on Languages in Social Media*, 2011, pp. 58–65.

[12] W.-Y. Chen, J.-C. Chu, J. Luan, H. Bai, Y. Wang, and E. Y. Chang, "Collaborative filtering for orkut communities: discovery of user latent behavior," in *WWW*, 2009, pp. 681–690.

[13] M. Naaman, J. Boase, and C.-H. Lai, "Is it really about me?: message content in social awareness streams," in *CSCW*, 2010, pp. 189–192.

[14] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities," in *WebKDD/SNA-KDD*, 2007, pp. 56–65.

[15] M. J. Welch, U. Schonfeld, D. He, and J. Cho, "Topical semantics of twitter links," in *WSDM*, 2011, pp. 327–336.

[16] Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, and Z. Su, "Understanding retweeting behaviors in social networks," in *CIKM*, 2010, pp. 1633–1636.

[17] M. Sachan, D. Contractor, T. A. Faruquie, and L. V. Subramaniam, "Using content and interactions for discovering communities in social networks," in *WWW '12*, 2012, pp. 331–340.

[18] N. Burns, Y. Bi, H. Wang, and T. Anderson, "A twofold-lda model for customer review analysis," *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, pp. 253–256, 2011.

[19] S. Moghaddam and M. Ester, "Ilda: interdependent lda model for learning latent aspects and their ratings from online product reviews," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 2011, pp. 665–674.

[20] W. X. Zhao, J. Jiang, H. Yan, and X. Li, "Jointly modeling aspects and opinions with a maxent-lda hybrid," in *EMNLP*, 2010, pp. 56–65.

[21] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, "Topic sentiment mixture: modeling facets and opinions in weblogs," in *WWW*, 2007.

[22] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," in *CIKM*, 2009, pp. 375–384.

[23] Y. Liu, A. Niculescu-Mizil, and W. Gryc, "Topic-link lda: joint models of topic and author community," in *ICML*, 2009, pp. 665–672.

[24] D. Zhou, E. Manavoglu, J. Li, C. L. Giles, and H. Zha, "Probabilistic models for discovering e-communities," in *WWW*, 2006, pp. 173–182.

[25] L. Hong and B. D. Davison, "Empirical study of topic modeling in Twitter," in *Proceedings of the First Workshop on Social Media Analytics*, 2010, pp. 80–88.