A Unified Topic-Style Model for Online Discussions

Ying Ding, Jing Jiang, Qiming Diao
School of Information Systems
Singapore Management University
{ying.ding.2011, jingjiang, qiming.diao.2010}@smu.edu.sg

Abstract

Forums have become major places for online communications for many years, where people often share and express opinions. We observe that, when editing posts, while some people seriously state their opinions, there are also many people playing jokes and writing meaningless posts on the discussed topics. We design a unified probabilistic graphical model to capture both topic-driven words and style-driven words. The model can help us separate serious and unserious posts/users and identify slang words. An extensive set of experiments demonstrates the effectiveness of our model.

1 Introduction

With the fast growth of the popularity of online social media, people nowadays are very used to sharing their thoughts and interacting with their friends on the Internet. Large online social network sites such as Facebook, Twitter and Flickr have attracted hundreds of millions of users. Among these online social media platforms, forums have always played an important role with its special characteristics. Unlike personal blogs, forums allow many users to engage in online conversations with a topic focus. Unlike Facebook, forums are usually open to public and users who post in forums do not need to reveal too much personal information. Unlike Wikipedia or Freebase, forums encourage users to exchange not only factual information but more importantly subjective opinions. All these characteristics make online forums a valuable source from which we can retrieve and summarize the general public's opinions about a given topic. This is especially important for businesses who want to find out how their products and services have been received and policy makers who are concerned about people's opinions on social issues.

While the freedom with which users can post in online forums has promoted the popularity of online forums, it has also led to the diversity in post quality. There are posts which contribute positively to a discussion by offering relevant, serious and meaningful opinions, but there are also many posts which appear irrelevant, disrespectful or meaningless. These posts are uninformative, hard to consume and sometimes even destructive. Let us look at some examples. Table 1 shows two forum posts in response to a piece of news about GDP bonuses for senior civil servants in Singapore. We can see that User A's post is clearly written. User B's post, on the other hand, is hard to comprehend. We see broken sentences, many punctuation marks such as "?" and colloquial expressions such as "ha." User B is not seriously contributing to the online discussion but rather trying to make a joke of the issue. Generally speaking, User B's post is less useful than User A's post in helping us understand the public's response to the news.

	Senior civil servants to get bumper GDP bonuses
User A	let us ensure this will be the LAST time they accord themselves ceiling salary scales and bonuses. i suspect MANY citizens are eagerly looking forward to the GE.
User B	Fever night, fever night, fe..ver.. Fever like to do it Got it?????? Ha..ha..ha...

Table 1: Two example online posts.

In this work, we opt for a fully unsupervised approach to modeling this phenomenon in online discussions. Our solution is based on the observation that the writing styles of serious posts and unserious posts are different, and the writing styles are often characterized by the words used in the posts. Moreover, the same user usually exhibits

Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media, pages 33–41,
Baltimore, Maryland USA, 27 June 2014. ©2014 Association for Computational Linguistics

User	Post
User A	*Re: Creativity, Art in the eyes of beholder. your take?* The difference is, the human can get tired or sick, and then it will affect his work, but the robot can work 24 hours a day 365 days a year and yet produce the same every time.
	Re: Diesel oil spill turns Manila Bay red, poses risk to health - ST The question is, will this environmental hazard turn up on the shores of it neighbors? And maybe even affect Singapore waters?
User B	*Re: Will PAP know who i vote in GE?* Hey! Who are you??? You make. ha..ha..ha.. he..he..he.. very angry lah
	Re: Gender discrimination must end for Singapore to flourish, says AWARE Hao nan bu gen nu dou Let you win lah ha..ha..ha..

Table 2: Sample posts of two example users.

the same writing style in most of his posts. For example, Table 2 shows two example users, each with two sample posts. We can see that their writing styles are consistent in the two posts. If we treat each writing style as a latent factor associated with a word distribution, we can associate observed words with the underlying writing styles. However, not all words in a post are style-driven. Many words in forum posts are chosen based on the topic of the corresponding thread. Our model therefore jointly considers both topics and writing styles.

We apply our topic-style model to a real online forum dataset from Singapore. By setting the number of styles to two, we clearly find that one writing style corresponds to the more serious posts while the other corresponds to posts that are not so serious. This topic-style model also automatically learns a meaningful slang lexicon. Moreover, we find that topics discovered by our topic-style model are more distinctive from each other than topics produced by standard LDA.

Our contributions in this paper can be summarized as follows: 1) We propose a principled topic-style model to jointly model topics and writing styles at the same time in online forums. 2) An extensive set of experiments shows that our model is effective in separating the more serious posts and unserious posts and identifying slang words.

2 Related Work

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) has been shown to be useful for many applications. Many extensions of LDA have been designed for different tasks, which are not detailed here. Our model is also an extension of LDA. We introduce two types of word distributions, one representing topics and the other representing writing styles. We use switch variables to alternate between these two types of word distributions. We also assume an author-level distribution over writing styles. It is worth pointing out that although our model bears similarity to a number of other LDA extensions, our objectives are different from existing work. E.g., the author topic model (Rosen-Zvi et al., 2004) also assumes an author-level distribution over topics, but the author-level distribution is meant to capture an author's topical interests. In contrast, our user-level distribution is over writing styles and is meant to identify serious versus unserious users. Similar to the models by Mei et al. (2007) and Paul et al. (2010) , we also use switch variables to alternate between different types of word distributions, but our goal is to identify words associated with writing styles instead of sentiment words or perspective words.

Another body of related research is around studying text quality, formality and sarcasm. Pitler and Nenkova (2008) investigated different features for text readability judgement and empirically demonstrated that discourse relations are highly correlated with perceived readability. Brooke et al. (2010) applied Latent Semantic Analysis to determine the formality level of lexical items. Agichtein et al. (2008) presented a general classification framework incorporating community feedback to identify high quality content in social media. Davidov et al. (2010) proposed the first robust algorithm for recognition of sarcasm. González-Ibáñez et al. (2011) took a closer look at sarcasm in Twitter messages and found that automatic classification can be as good as human classification. All these studies mainly rely on supervised techniques and human annotation needs to be done, which is very time consuming. Our method is fully unsupervised, which can automatically uncover different styles and separate serious posts from unserious posts.

Our work is also related to spam/spammer detection in social media, which has been studied over different platforms for a few years. Jindal and Liu (2008) first studied opinion spam in online reviews and proposed a classification method for opinion spam detection. Bhattarai et al. (2009)

investigated different content attributes of comment spam in the Blogsphere and built a detection system with good performance based on these attributes. Ding et al. (2013) proposed to utilize both content and social features to detect spams in online question answer website. Existing work on spam detection need annotated data to learn the spam features but our model does not as it is fully unsupervised.

3 A Topic-Style Model

Writing styles can be reflected in many different ways. Besides choices of words or expressions, many other linguistic features such as sentence length, sentence complexity and use of punctuation marks may all be associated with one's writing style. In this work, however, we try to take an approach that does not rely on heavy linguistic analysis or feature engineering. Part of the reason is that we want our approach to be independent of language, culture or social norms so that it is robust and can be easily applied to any online forum.

To this end, we represent a writing style simply as a distribution over words, much like a topic in LDA. We assume that there are S latent writing styles shared by all users contributing to a forum. Meanwhile, we also assume a different set of T latent topics. We mix writing styles and topics to explain the generation of words in forum posts.

A key assumption we have is that the same user tends to maintain a consistent writing style, and therefore we associate each user with a multinomial distribution over our latent writing styles. This is similar to associating a document with a distribution over topics in LDA, where the assumption is that a single document tends to have focused topics. Another assumption of our model is that each word in a post is generated from either the background or a topic or a writing style, as determined by a binary switch variable.

3.1 Model Description

We now formally describe the topic-style model we propose. The model is depicted in Figure 1. We assume that there are T latent topics, where ϕ_t is the word distribution for topic t. There are S latent writing styles, where ψ_s is the word distribution for writing style s. There are E threads, where each thread e has a topic distribution θ_e, and there are U users, where each user u has a writing style distribution π_u.

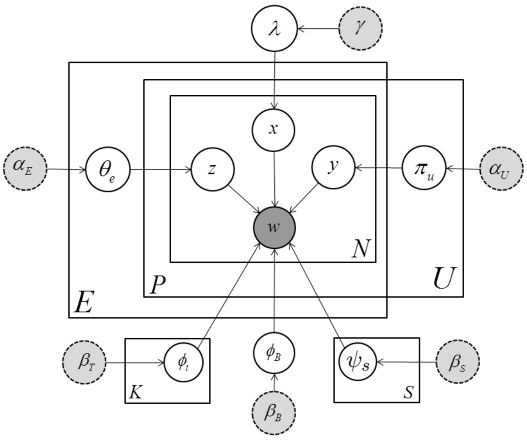

Figure 1: Topic-Style Model

Notation	Description
$\gamma, \alpha_E, \alpha_U,$ $\beta_B, \beta_T, \beta_S$	Hyper-parameters of Dirichlet distributions
λ	A global multinomial distribution over switching variables x
θ_e, π_u	Thread-specific topic distributions and user-specific style distributions
ϕ_B, ϕ_t, ψ_s	Word distributions of background, topics and styles
$x_{e,p,n}$, $y_{e,p,n}$, $z_{e,p,n}$	Hidden variables: $x_{e,p,n}$ for switching, $y_{e,p,n}$ for style of style words, $z_{e,p,n}$ for topic of topic words
e, p, n	Indices: e for threads, p for posts, n for words
$E, P_e, U,$ $N_{e,p}$	Number of threads, numbers of posts in threads, number of users and numbers of words in posts
S, K, V	Numbers of styles, topics and word types

Table 3: Notation used in our model.

For each word in a post, first a binary switch variable x is sampled from a global Bernoulli distribution parameterized by λ. If $x = 0$, we draw a word from the background word distribution. Otherwise, if $x = 1$, we draw a topic from the corresponding thread's topic distribution; if $x = 2$, we draw a writing style from the corresponding user's writing style distribution. We then draw the word from the corresponding word distribution.

The generative process of our model is described as follows. The notation we use in the model is also summarized in Table 3.

- Draw a global multinomial switching variable distribution $\lambda \sim$ Dirichlet(γ).
- Draw a multinomial background word distribution $\phi_B \sim$ Dirichlet(β_B).
- For each topic $t = 1, 2, \ldots, T$, draw a multinomial topic-word distribution $\phi_t \sim$ Dirichlet(β_T).
- For each writing style $s = 1, 2, \ldots, S$, draw a multinomial style-word distribution $\psi_s \sim$ Dirichlet(β_S).
- For each user $u = 1, 2, \ldots, U$, draw a multinomial style distribution $\pi_u \sim$ Dirichlet(α_u).
- For each thread $e = 1, 2, \ldots, E$

- draw a multinomial topic distribution $\theta_e \sim \mathrm{Dir}(\alpha_E)$.
- for each post $p = 1, 2, \ldots, P_e$ in the thread, where $u_{e,p} \in \{1, 2, \ldots, U\}$ is the user who has written the post
    - for each word $n = 1, 2, \ldots, N_{e,p}$ in the thread, where $w_{e,p,n} \in \{1, 2, \ldots, V\}$ is the word type
        - draw $x_{e,p,n} \sim$ Multinomial(λ).
        - If $x = 0$, draw $w_{e,p,n} \sim$ Multinomial(ϕ_B)
        - If $x = 1$, draw $y_{e,p,n} \sim$ Multinomial($\pi_{u_{e,p}}$), and then draw $w_{e,p,n} \sim$ Multinomial($\psi_{y_{e,p,n}}$).
        - If $x = 2$, draw $z_{e,p,n} \sim$ Multinomial(θ_e), and then draw $w_{e,p,n} \sim$ Multinomial($\phi_{z_{e,p,n}}$).

3.2 Parameters Estimation

We use Gibbs sampling to estimate the parameters. The sampling probability that assign the nth word in post p of thread e to the background topic is as follows:

$$\begin{aligned} &P(x_{e,p,n} = 0 | \boldsymbol{W}, \boldsymbol{U}, \boldsymbol{X}^{-i}, \boldsymbol{Y}^{-i}, \boldsymbol{Z}^{-i}) \\ \propto &(\gamma + n_0) \times \frac{\beta_B + n_B^{w_{e,p,n}}}{V\beta_B + n_0} \end{aligned}$$

where n_0 is the number of words assigned as background words and $n_B^{w_{e,p,n}}$ is the number of times word type of $w_{e,p,n}$ assigned to background. The probability to assign this word to style s is as follows:

$$\begin{aligned} &P(x_{e,p,n} = 1, y_{e,p,n} = s | \boldsymbol{W}, \boldsymbol{U}, \boldsymbol{X}^{-i}, \boldsymbol{Y}^{-i}, \boldsymbol{Z}^{-i}) \\ \propto &(\gamma + n_1) \times \frac{\alpha_U + n^s_{u_{e,p}}}{S\alpha_U + n^*_{u_{e,p}}} \times \frac{\beta_S + n_s^{w_{e,p,n}}}{V\beta_S + n_s^*} \end{aligned}$$

where n_1 is the number of words assigned as style words, $n^*_{u_{e,p}}$ and $n^s_{u_{e,p}}$ are the number of words written by user $u_{e,p}$ and assigned as style words, and the number of these words assigned to style s, respectively. n_s^* and $n_s^{w_{e,p,n}}$ are the number of words assigned to style s and the number of times word type of term $w_{e,p,n}$ assigned to style s. The probability to assign this word topic t is as follows:

$$\begin{aligned} &P(x_{e,p,n} = 2, z_{e,p,n} = t | \boldsymbol{W}, \boldsymbol{U}, \boldsymbol{X}^{-i}, \boldsymbol{Y}^{-i}, \boldsymbol{Z}^{-i}) \\ \propto &(\gamma + n_2) \times \frac{\alpha_E + n_e^t}{K\alpha_E + n_e^*} \times \frac{\beta_T + n_t^{w_{e,p,n}}}{V\beta_T + n_t^*} \end{aligned}$$

where n_2 is the number of words assigned as topic words, n_e^* is the number of words in thread e assigned as topic words, n_e^t is the number of words in thread e assigned to topic t, n_t^* is the number of words assigned to topic t, and $n_t^{w_{e,p,n}}$ is the number of times word type of $w_{e,p,n}$ is assigned to topic t.

After running Gibbs sampling for a number of iterations, we can estimate the parameters based on the sampled topic assignments. They can be calculated by the equations below:

$$\phi_t^w = \frac{\beta_T + n_t^w}{V\beta_T + n_t^*} \qquad \phi_s^w = \frac{\beta_S + n_s^w}{V\beta_S + n_s^*}$$

$$\theta_e^t = \frac{\alpha_E + n_e^t}{K\alpha_E + n_e^*} \qquad \theta_u^s = \frac{\alpha_U + n_u^s}{S\alpha_U + n_u^*}$$

4 Experiment

4.1 Data Set and Experiment Setup

To evaluate our model, we use forum threads from AsiaOne[1], a popular online forum site in Singapore. We crawled all the threads between January 2011 and June 2013 under a category called "Singapore," which is the largest category on AsiaOne. In the preprocessing stage, we removed the URLs, HTML tags and tokenized the text. Emoticons are kept in our data set as they frequently occur and indicate users' emotions. All stop words and words occurring less than 4 times are deleted. We also removed users who have fewer than 8 posts and threads attracting fewer than 21 posts. The detailed statistics of the processed dataset are given in Table 4.

#Users	#Words	#Tokens	#Posts/User	#Posts/Thread
580	29,619	2,940,886	205.3	69.5

Table 4: Detailed statistics of the dataset.

We fix the hyper-parameters $\gamma, \alpha_E, \alpha_U, \beta_T$ and β_S to be $10, 1, 1, 0.01$ and 0.01 respectively. we set $\beta_{B,v}$ to be $H \cdot p_B(v)$, where H is set to be 20 and $p_B(v)$ is the probability of word v as estimated from the entire corpus. The number of topics K is set to be 40 empirically.

4.2 Model Development

Before we evaluate the effectiveness of our model, we first show how we choose the number of styles to use. Note that although we are interested in separating serious and unserious posts, our model can generally handle any arbitrary number of writing styles. We therefore vary the number of writing styles to see which number empirically gives the most meaningful results.

Assuming that different styles are characterized by words, we expect to see that the discovered

[1] http://www.asiaone.com

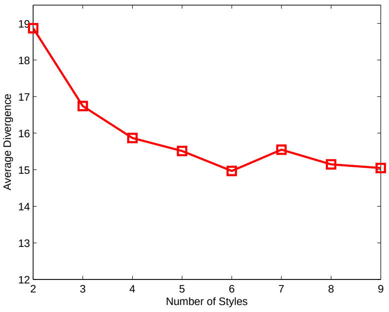

Figure 2: Average Divergence over different numbers of styles.

	Style No.	Top Words
2 styles	Style 1	singapore, people, years, government
	Style 2	BIGGRIN, TONGUE, lah, ha
3 styles	Style 1	people, make, WINK, good
	Style 2	singapore, years, government, mr
	Style 3	BIGGRIN, TONGUE, lah, ha
4 styles	Style 1	ha, lah, WINK, dont
	Style 2	singapore, year, mr, years
	Style 3	people, good, make, singapore
	Style 4	BIGGRIN, TONGUE, EEK, MAD

Table 5: Sample style words

word distributions for different styles are very different from each other. To measure the distinction among a set of styles, we define a metric called Average Divergence (AD) based on KL-divergence. Average Divergence can be calculated as follows.

$$AD(\boldsymbol{S}) = \frac{2}{N(N-1)} \sum_{i \neq j} S_{\mathrm{KL}}(s_i || s_j),$$

where $\boldsymbol{S}$ is a set of style-word distributions, N is the size of $\boldsymbol{S}$ and s_i is the i-th distribution in $\boldsymbol{S}$. $S_{\mathrm{KL}}(s_i||s_j)$ is the symmetric KL divergence between s_i and s_j (i.e., $D_{\mathrm{KL}}(s_i||s_j) + D_{\mathrm{KL}}(s_j||s_i)$). The higher Average Divergence is, the more distinctive distributions in $\boldsymbol{S}$ are.

Figure 2 shows the Average Divergence over different numbers of styles. We can clearly see that the Average Divergence reaches the highest value when there are only two styles and decreases with the increase of style number. This means the styles are mostly distinct from each other when the number is 2 and their difference decreases when there are more styles.

To get a better understanding of the differences of using different numbers of styles, we compare the top words in each style when the number of styles is set to be 2, 3 and 4. The results are shown in Table 5 where all uppercase words represent emoticons. From the top words of the first row, we can see that Style 1 is dominated by formal words while Style 2 is dominated by emoticons like BIGGRIN and slang words like “lah” and “ha.” These two styles are well distinguished from each other and humans can easily tell the difference between them. Also, Style 2 is an unserious style characterized by emoticons, slang and urban words. Table 6 shows the top words of these 2 styles excluding emoticons. From this table, we can observe that Style 2 has many slang words with high probability while top words in Style 1 are all very formal. However, styles in the second and third rows of Table 5 are not easily distinguishable from each other. In these results, there often exist two styles very similar to the styles in row 1 while the other styles look like the combination of these two styles and humans cannot tell their meanings very clearly. Based on these observations, we fix the number of styles to 2 in the following experiments.

Serious Style	Unserious Style
singapore	lah
people	ha
years	dont
government	stupid
time	leh
made	ah
year	lor
public	liao

Table 6: Top words of different styles

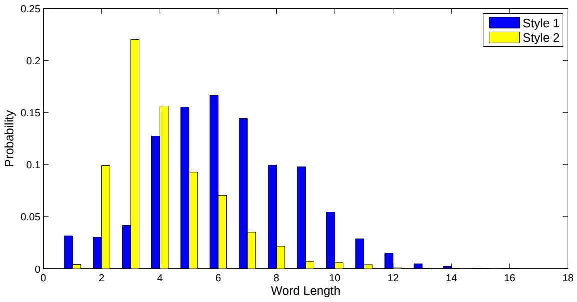

Figure 3: Word length distribution

One previous work uses word length as an indicator of formality (Karlgren and Cutting, 1994). Here, we borrow this idea and compare the word length of Style 1 and Style 2. We calculate the distributions of word length and show the results in Figure 3. It shows that the majority of words in Style 1 are longer compared with those in Style 2. To have a quantitative view of the difference between the word lengths of these two styles, we heuristically extract words labeled with Style 1 and Style 2 in our dataset in the final iteration of Gibbs sampling and apply Mann-Whitney U test on these two word length populations. The null hypothesis that the two input populations are the same is rejected at the 1% significance level. This verifies the intuition that serious posts tend to use longer words than unserious posts.

4.3 Post Identification

Our model can also be used to separate serious posts and unserious posts. We treat this as a retrieval problem and use precision/recall for evaluation.

We use a simple scoring function, which is the proportion of words assigned to the unserious style when we terminate the Gibbs sampling at the 800-th iteration, to score each post. When applying this method to our data, emoticons are all removed. For comparison, we rank post according to the number of emoticons inside a post as the baseline. After getting the result of each method, we ask two annotators to label the first and last 50 posts in the ranking list. The first 50 posts are used for evaluation of unserious post retrieval and the last 50 post are used for evaluation of serious post retrieval. This evaluation is based on the assumption that if a method can separate serious and unserious posts very well, posts ranked at the top position should be unserious ones and those ranked near to the bottom should be serious ones. The results are shown in Table 7 where our method is denoted as TSM and the baseline method is denoted as EMO. In serious post retrieval, the baseline have a perfect performance and our method is competitive. We can see that EMO has a perfect performance in identifying serious posts. When posts are ranked in reverse order according to the number of emoticons they contain, the last 50 ones do not contain any emoticons. They can be regarded as a random sample of posts without emoticons. Compared with identifying serious posts, identifying unserious posts looks much more difficult. EMO's poor performance on this task tells us that emoticon is not a promising sign to detect unserious posts. However, the word style a post uses matters more, which also proves the value of our proposed model.

		P@5	P@15	P@25	P@35
Serious	EMO	1.0	1.0	1.0	**1.0**
	TSM	1.0	1.0	1.0	0.97
Unserious	EMO	0.4	0.67	0.64	0.6
	TSM	**1.0**	**0.93**	**0.96**	**0.97**

Table 7: Precision for Serious and Unserious Post Retrieval. P@N stands for the precision of the first N results in ranking list.

		P@5	P@15	P@25	P@35
Serious	Baseline	0.6	0.8	0.8	0.83
	TSM	**1.0**	**1.0**	**1.0**	**0.94**
Unserious	Baseline	1.0	0.87	0.92	0.91
	TSM	1.0	**1.0**	**1.0**	**1.0**

Table 8: Precision for serious and unserious user retrieval.

4.4 User Identification

In this section, we evaluate the performance of TSM on identifying serious and unserious users. This identification task is very important as many research tasks such as opinion mining and expert finding are more interested in the serious users. We treat this task as a retrieval problem as well, which means we will rank users by a scoring function and do evaluation on this ranking result.

We rank user according to their style distribution π_u and pick the first 50 and last 50 users for evaluation. For each user, 10 posts are sampled to be shown to the annotators. We mix these 100 users and ask two graduate students to do the annotations. The evaluation strategy is the same as that in Section 4.3. We choose a simple baseline which ranks users by the number of emoticons they use per post. The evaluation result is shown in Table 8 for serious and unserious user retrieval respectively.

In both serious and unserious user retrieval tasks, our method gets almost perfect performance, which is better than the baseline. This means the user style distributions learned by our model can help separate serious and unserious users.

4.5 Perplexity

Perplexity is a widely used criterion in statistical natural language processing. It measures the predictive power of a model on unseen data, which is algebraically equivalent to the inverse of the geometric mean per-word likelihood. A lower perplexity means the test data, which is unseen in the training phase, can be generated by the model with a higher probability. So it also indicates that the model has a better generalization performance.

In this experiment, we leave 10% data for testing and use the remaining 90% data for training. We choose LDA as a baseline for comparison and treat each thread as a document. The perplexity for both models is calculated over different numbers of topics, which ranges from 10 to 100. The result is show in Figure 4. We can clearly see that our proposed model has a substantially lower perplexity than LDA over different numbers of topics. This proves that our model fits the forum discussion data better and has a stronger generalization power. It also indicates that separating topic-driven words and style-driven words can better fit the generation of user generated content in forum discussions.

4.6 Topic Distinction

In traditional topic modeling, like LDA, all words are regarded as topic-driven words, which are generated by mixture of topics. However, this may not be true to user-generated content in online forums as not all words are driven by discussed topics. Take the following post for example:

- Okay lah. Let them be. I mean its their KKB right? Let it rot lor.

In this post, the words "lah" and "lor" are not related to the topics under discussion. They appear in the post because the authors are used to using these words, which means these words are style driven. Style-driven words are related to a user's characteristics and should not be clustered into any topic. Without separating these two types of words, style-driven words may appear in different topics and make topics less distinct to each other.

Figure 5 compares the Average Divergence among discovered topics between TSM (Topic Style model) and LDA over different numbers of topics. We can clearly see that the Average Divergence of TSM is substantially larger than that of LDA over different numbers of topics. This proves that in TSM, the learned topics are more distinct from each other. This is because LDA mixes these two kinds of words, which introduces noise into the learned topics and decreases their distinction between each other. But topic driven words and style driven words are well separated in TSM. Figure 5 also plots the Average Divergence between the learned two styles, which is the curve denoted by DIFF. We can see the AD between different styles is even larger than that among topics in TSM. Different topics may still have some overlap in frequently used words but styles may share few words with each other. So AD of styles can get higher value. This also proves the effectiveness of our model in identifying writing styles and uncovering more distinct topics.

	P@5	P@10	P@20	P@30	P@40	P@50
E	0	0.2	0.25	0.23	0.225	0.2
T	**0.8**	**0.9**	**0.8**	**0.8**	**0.675**	**0.62**

Table 9: Slang identification precision. E: Emoticon; T:TSM.

	#Word/Post	#Post
Formal User	34.9	158.3
Informal User	14.5	381

Table 10: Mean Value of average post length and number of post for different type of users

4.7 Discovering Slang

By looking at Table 5, we notice that the unserious style contains many slang words with high probability. This indicates that the unserious style in the dataset we use is also characterized by slang words. In this section, we will show the usefulness of our model in slang discovery. The baseline method is denoted as Emoticon as it ranks words according to their probability of occurring in a post containing emoticons. We ask two Singaporean annotators to help us identify Singaporean slang in the top 50 words. The result is shown in Table 9. It tells us the unserious style learned in our model has very good performance in identifying local slang words. For people preferring unserious writing style, they would write posts in a very flexible way and use many informal words, abbreviations and slang expressions. So our unserious style will be characterized by these slang words and performs very well in identifying these slang words.

4.8 Analysis of Users

In this subsection, we analyze users in our dataset based on the result learned by TSM. Figure 8 shows the distribution of the histogram of serious style probability. The majority of users have a high serious style probability, which means most users in our dataset are more eager to give serious comments and express their opinions. This satisfies our observation that most people use forums mainly to discuss and seek knowledge on different topics and they are very eager to express their thoughts in a serious way.

We heuristically split all users into two sets according to user-style probability by setting 0.5 as threshold. Users with probability of serious style

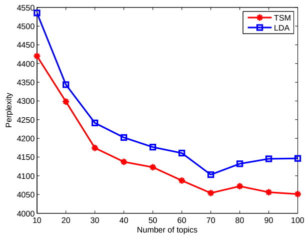

Figure 4: Perplexity

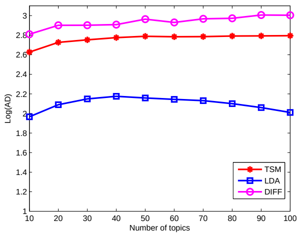

Figure 5: Average Divergence

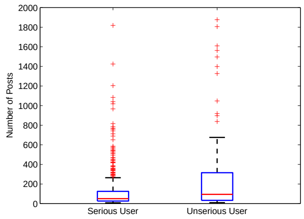

Figure 6: Box plot of post number for serious and unserious users

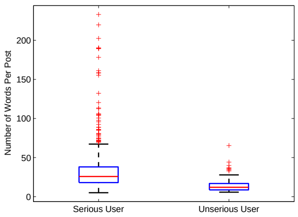

Figure 7: Box plot of average post length for serious and unserious users

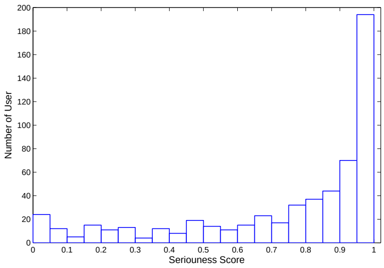

Figure 8: Seriousness Score of Users

larger than 0.5 are regarded as serious users and the remaining are unserious users. Next, we extract the number of posts each user edit and the average number of words per post for each user and compare the difference between these two user sets. Figure 6 and Figure 7 show the box plots of post number and average post length respectively. We can see that serious users edit fewer posts but use more words in each post. To see the difference between serious and unserious users more clearly, we apply Mann-Whitney U test on the post number populations and average post length populations. The Mann-Whitney U test on both data set reject the null hypothesis that two input populations are the same at the 1% significance level. The mean value for post number and average post length are also computed and shown in Table 10. We can find that serious users tend to publish fewer but longer posts than unserious users. This result is intuitive as serious users often spend more effort editing their posts to express their opinions more clearly. However, for unserious users, they may just use a few words to play a joke or show some emotions and they can post many posts without spending too much time.

5 Conclusions

In this paper, we propose a unified probabilistic graphical model, called Topic-Style Model, which models topics and styles at the same time. Traditional topic modeling methods treat a corpus as a mixture of topics. But user-generated content in forum discussions contains not only words related to topics but also words related to different writing styles. The proposed Topic-Style Model can perform well in separating topic-driven words and style-driven words. In this model, we assume that writing style is a consistent writing pattern a user will express in her posts across different threads and use a latent variable at user level to capture the user specific preference of writing styles. Our model can successfully discover writing styles which are different from each other both in word distribution and formality. Words belonging to different writing styles and user specific style distribution are captured by our model at the same time. An extensive set of experiments shows that our method has good performances in separating serious and unserious posts and users. At the same time, the model can identify slang words with promising accuracy, which is proven by our experiments. An analysis based on the learned parameters in our model reveal the difference between serious and unserious users in average post length and post number.

References

Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. 2008. Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 183–194, New York, NY, USA. ACM.

Archana Bhattarai, Vasile Rus, and Dipankar Dasgupta. 2009. Characterizing comment spam in the blogosphere through content analysis. In *Computational Intelligence in Cyber Security, 2009. CICS'09. IEEE Symposium on*, pages 37–44. IEEE.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

Julian Brooke, Tong Wang, and Graeme Hirst. 2010. Automatic acquisition of lexical formality. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 90–98, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 107–116, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhuoye Ding, Yeyun Gong, Yaqian Zhou, Qi Zhang, and Xuanjing Huang. 2013. Detecting spammers in community question answering. In *Proceeding of International Joint Conference on Natural Language Processing*, pages 118–126, Nagoya, Japan. Association for Computational Linguistics.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, pages 581–586, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 219–230, New York, NY, USA. ACM.

Jussi Karlgren and Douglass Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 2*, pages 1071–1075, Stroudsburg, PA, USA. Association for Computational Linguistics.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 171–180, New York, NY, USA. ACM.

Michael J. Paul, ChengXiang Zhai, and Roxana Girju. 2010. Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 66–76, Stroudsburg, PA, USA. Association for Computational Linguistics.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 487–494, Arlington, Virginia, United States. AUAI Press.