Knowledge-Based Systems 209 (2020) 106435

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A reliable cross-site user generated content modeling method based on topic model

Baoxi Liu, Peng Zhang *, Tun Lu, Ning Gu

School of Computer Science, Fudan University, Shanghai, China

ARTICLE INFO

Article history:
Received 12 March 2020
Received in revised form 11 August 2020
Accepted 1 September 2020
Available online 21 September 2020

Keywords:
Multiple social network sites
User generated content modeling
Topic model
Weibo
Douban

ABSTRACT

Nowadays, social network sites (SNSs) have been significant platforms for content sharing in our daily life. With the emergence of different kinds of social network sites and users' diverse needs for content sharing, their content sharing practices are generally taken place in multiple SNSs. To construct models that can characterize users' content sharing practices in a composite context constituted by multiple social network sites (cross-site user generated content modeling) has been an emerging research topic in web data mining and human behavior research. However, previous methods such as Dirichlet Multinomial Mixture model (DMM), Biterm Topic Model (BTM), Twitter-LDA and MultiLDA have limited representation ability or are based on unreliable assumption, which cannot characterize the user generated content (UGC) accurately from the perspective of multiple SNSs. In this paper, we first conduct an empirical study to investigate the characteristics of users' content sharing practices in cross-site context, based on which we propose a more reliable cross-site UGC model named CrossSite-LDA (C-LDA). We then evaluate the performances of the C-LDA model with four state-of-the-art models based on the two data sets sampled from Weibo–Douban and Facebook–Twitter. Results show that the C-LDA has better performances in perplexity, word coherence, topic KL divergence, UCI and UMass metrics compared with existing models, which suggests its superior accuracy on modeling users' content characteristics in cross-site context.

1. Introduction

Nowadays, social network sites (SNSs) have been important mediums for content sharing in our daily life. A report [1] released in 2018 suggests that more than 73% of users utilize Facebook, Twitter and some other popular social network sites to distribute news, events, life moments, etc. These user generated content often contains valuable information such as user sentiments, user interests, demographic information, etc, which makes UGC to be significant corpus for opinion mining, user portrait construction and personalized recommendation. Therefore, many studies have focused on mining user generated content from SNSs. For example, studies [2,3] collected large-scale tweet data, and then performed user sentiment analysis and election prediction based on that; Studies [4–6] extracted topics from user generated content, and utilized that to infer users' interests for book and music recommendations; In studies [7,8], authors mined topics and demographic information from users' content in SNSs to relieve the cold start problem of product recommendations in e-commerce sites; Some studies attempted to infer user demographic information based on user generated content, such as age [9], gender [10] emotion [11] and occupation [12] with the goal of constructing a more accurate user portrait, etc.

The key to mining user generated content is to model the corpus accurately, which motivates researchers to propose many UGC modeling methods in recent years. The most commonly used methods are Term Frequency–Inverse Document Frequency (TF–IDF), Latent Dirichlet Allocation (LDA) and Twitter-LDA. The TF–IDF method models each user's post as a series of keywords generated by calculating the Term Frequency and Inverse Document Frequency [13]. This approach is easy to implement, while it has disadvantages like high dimensions of word vector and limitation to distinguish polysemy. For these problems, the LDA model performs UGC modeling by representing the post as distribution of topics, which is driven based on the co-occurrence of words [14]. However, in the SNS context, the user's post is usually short, which results that the LDA model generally suffers the problem of word sparsity. The newly proposed Dirichlet Multinomial Mixture (DMM) [15] and Twitter-LDA [16] model solve this problem by assuming that each post corresponds to one topic. Biterm topic model (BTM) [17] takes a step forward, and this method directly models the word co-occurrence patterns based on observed biterms (word–word co-occurrence patterns) in texts, which can alleviate the document-level word sparsity. These model have been proved more accurate for UGC modeling

* Corresponding author.
E-mail address: zhpll@126.com (P. Zhang).

https://doi.org/10.1016/j.knosys.2020.106435
0950-7051/

and widely utilized for content modeling in Twitter, Facebook, Weibo, etc.

Nowadays, as the emergence of different kinds of social network sites and users' diverse needs for content sharing, their content sharing practices are generally taken place in multiple social network sites [18,19]. Some previous research has begun to combine two or more SNSs to explore users' content sharing behaviors. One important finding is that there exists some relevance (i.e., similarity, association, complementation, etc.) of user generated content between different SNSs. For example, a study calculated the JSD scores of content topics between Twitter and Instagram. They found that most users corresponded to higher JSD scores, which suggests that the content users have shared on different social network sites has a higher topic similarity [20]. A similar work also found that users often shared content with similar topics to different social network sites. But considering the SNS norms and audiences, they generally adjust their content so that the content can be accepted by different audiences [21]. Another study found that considering audience differences, users usually described the same events or topics from different perspectives, which means posts on different SNSs complement each other but point to same events or topics [22]. Therefore, in nowadays SNSs context, user's content sharing is no longer a "local" practice that is limited in every single site but a composite activity conducted across multiple social network sites. Modeling the post by treating SNSs in isolation ignores the content relevance among different SNSs and thus cannot characterize the content accurately anymore. To construct a model that can characterize users' content sharing practices in a composite context constituted by multiple social network sites (cross-site UGC modeling) has been an emerging research topic in web data mining and human behavior research [23–25].

The aforementioned methods like LDA, Twitter-LDA, DMM and BTM are difficult to solve the cross-site UGC modeling problem accurately. One of the most important reasons is that these methods have limited content representation capability. They only consider document (post), topic and word while ignore factors related to platforms like content relevance between different platforms, users' platform preferences, etc. Recently, a study constructed a topic model named MultiPlatform-LDA (MultiLDA) which considered above relevance to characterize users' content sharing behaviors in multiple SNSs context [20]. In this model, the content relevance between different platforms is represented by users' platform preferences for specific topics, based on which the cross-site content characteristics can be learned from user-topic distribution and topic-platform distribution. Although this method provides a fresh and feasible approach to model UGC in the sharing context constituted by multiple social network sites, it is built based on a basic assumption - "users have social media platforms preference specific to topics". However, whether this assumption is reliable or not is not investigated in the paper, which gives future research potential to continue to focus on cross-site UGC modeling and construct a reliable model (i.e., a model that can truly characterize users' content sharing practices in the context constituted by multiple SNSs) to promote its accuracy in further.

However, to construct a reliable model is very challenging. First, abstracting users' content sharing practices from the perspective of multiple social network sites is non-trial. Cross-site content sharing is a complicated practice that is driven by both intrinsic factors like user motivations and expectations as well as extrinsic factors such as platform functionalities and audiences. The existence of so many influential factors and the difficulty to observe many of such factors bring challenges to understanding and abstracting users' cross-site content sharing practices. Second, based on a fine-grained understanding of such a practice, to construct a quantitative model to characterize the procedures and concepts of cross-site content sharing is also tricky. Third, UGC modeling in a single SNS is a challenging task because of problems like word sparsity, low corpus quality, etc. In the composited context constituted by multiple social network sites, UGC modeling becomes more complicated as more factors will be involved, such as users' platform preferences, topic relevance, etc.

For the above questions and challenges, we focus on proposing a new and reliable cross-site UGC modeling method in this paper mostly based on Twitter-LDA and MultiLDA. For this goal, we first conduct an empirical study by combining quantitative methods and qualitative methods to investigate users' content sharing practices in multiple SNSs context. The analysis results reveal that SNS users essentially have platform preferences for content sharing. It is generally independent of specific topics, which is contrary to the assumption of MultiLDA. Furthermore, we also find that users tend to utilize different language styles in different platforms, which also gives us valuable insights to construct a novel cross-site UGC model. Based on these findings, we propose a reliable cross-site UGC model named CrossSite-LDA (C-LDA). In this model, we define elements including platform, user, topic and word as well as relationships between them like user-platform, platform-word, user-topic, and topic-word distribution to represent the procedures and characteristics of cross-site content sharing practices revealed by our empirical study. For example, user's preferences for different platforms are represented by user-platform distribution in our model, and the language style differences on different platforms can be represented by platform-word distribution. To validate the performances of our proposed method, we collect real data from two popular social network sites — Weibo and Douban and conduct comparative experiments by setting DMM, BTM, Twitter-LDA, and MultiLDA as baselines. Results show that our C-LDA model has lower perplexity and better word coherence, topic KL divergence, UCI and UMass scores, which suggests its superior performances on modeling users' cross-site content characteristics. Furthermore, we also conducted experiments from Twitter and Facebook cross-site data sets to verify the robustness of the C-LDA model. To conclude, the main contributions of our work are given below.

- We investigate users' cross-site content sharing practices by a qualitative and quantitative combined empirical study, which provides valuable insights for cross-site content modeling and research on cross-site content sharing behavior.
- Based on the findings of empirical study, we propose a novel and more reliable cross-site UGC model which can more accurately represent characteristics of users' content sharing practices in multiple SNSs context.
- We collect real user generated content from Weibo and Douban and validate the C-LDA model's better performances in perplexity, word coherence, topic KL divergence, UCI and UMass scores compared with existing models. We also verify the robustness of C-LDA model by introducing Twitter–Facebook cross-site data set.

The rest of this paper is organized as follows. In Section 2, we review the related works about social network content analysis and content modeling methods. Section 3 gives our preliminary analysis which includes semi-structured interviews and quantitative data analysis. In Section 4, we propose our C-LDA model and give the model definition and inference. Then, the data sets and model evaluations are given in Section 5. Finally, the conclusion is given in Section 6.

2. Related work

In this section, we review works related to cross-site UGC analysis and modeling. Since understanding the user content sharing characteristics is significant for content modeling, we first review SNS content analysis related works. Then we discuss SNS content modeling approaches. We find that despite the increase in the user profile and social relationship studies, there are relatively few studies on cross-site UGC modeling.

Several studies have analyzed the user content sharing characteristics in social network sites in terms of content topic, quantity, category, flow, etc. One important finding is that users' content in different SNSs generally exhibits some differences as well as some relevance. Some studies revealed that users' content in different platforms exhibited some differences. For example, in [19,26,27], authors studied the category differences of user content sharing on different SNSs. They found that the content published on Twitter was more about news, interests and opinion expressions, while the content on Facebook was mostly related to daily activities. They also found Instagram engaged more of the users' hearts and Twitter captured more of their minds. Ottoni et al. [18] analyzed the differences of the content quantity, content category and content flow on Twitter and Pinterest. They found that the content sharing on Twitter was more numerous and diverse than that on Pinterest, and new content tended to germinate on Pinterest and then flowed to Twitter. Some research have revealed that user generated content on different social network sites has relevance. For example, Sleeper et al. [21] found that users often shared content with a similar topic to different social network sites. But considering the platform norms and audiences, they generally adjusted their content so that the content could be accepted by different audiences. Zhao et al. [22] found that considering audience differences, users usually described the same events or topics from different perspectives, which means posts on different platforms complement each other and point to the same events or topics. Manikonda et al. [19] and Lee et al. [20] analyzed the topic distributions on Twitter and Instagram. Their studies suggest that both of the two platforms have overlap topics and users enjoy higher topic distribution similarity. They also found users usually adjusted their content to cater to the different audiences on different social network sites.

Many works attempted to model user content sharing practices from the view of a single social network site. To the best of our knowledge, the most prevalent model is Twitter-LDA [16]. This model assumes that each tweet contains only one topic and describes the generation of tweets utilizing elements like the user, topic and word. Their experiment results show that Twitter-LDA outperforms standard LDA model [14] and the Author Topic Model [28] in topic discovery. In [29], Diao et al. proposed a topic model to detect the burst topics on Twitter by dividing the content topics into global topics and personal topics. In [30], Pal et al. aimed to identify authoritative users on Instagram. They presented a novel Authority Learning Framework (ALF) to identify authoritative users on Instagram. In [31], motivated by an empirical study that different content had different likelihood of getting propagated on different SNSs, Hoang et al. developed the V2S framework to model user content sharing and then predicted the content propagation on Twitter. In [32], Jang et al. first proposed a user content modeling method and then analyzed users' topic characteristics of different ages. Ferrara et al. [33] and Xu et al. [34] introduced the post tags and labels to improve the topic modeling results. In [17],authors proposed Biterm topic model (BTM) and this method relieve the problem of word sparsity by directly modeling the word co-occurrence patterns based on observed biterms (word–word co-occurrence patterns) in texts. In [4,7,8], authors first built a model to extract content topics and then utilized these topics to make recommendations for users. In [35], Xuan et al. presented a nonparametric relational topic model that could discover topic size and the hidden topics automatically. To achieve integrated intelligence that involves both perception and inference, some studies attempted to combine the topic model with deep learning. For example, Li et al. [36] proposed a Recurrent Attention Topic Model (RATM) for document embedding. Different from the conventional topic models, RATM considered two important factors: one is sequential orders among sentences, and the other is attention mechanism. Wang et al. [37] proposed a Bayesian deep learning (BDL) framework, which merges the perception ability of neural networks and inference ability of probabilistic graphical models.

In multiple SNSs context, there is some work focusing on modeling and analyzing user profiles and social relationships. For example, Abel et al. [38] analyzed the completeness, consistency of user profiles on different platforms, and then modeled these aggregated user profiles data. Their modeling results were used for tag recommendations and they solved the cold start problem. Cho et al. [39] attempted to infer real world events from social network content. They first designed a model to independently identify the events on both Instagram and Twitter and then integrated the results to infer hot events. In [40], Farseev et al. proposed a fusion-based model on a large-scale multi-source dataset to infer users' age and gender. Chen et al. [41] aggregated users' gender, age, and other information on different platforms to infer users' real-world information. As for social relationship modeling, Magnani et al. [42] established a social relationships model across multiple SNSs. They first established the users' social relationships in one SNS and then built a cross-site social relationship model through account mapping between different platforms. Guo et al. [43] proposed a Social-Relational Topic Model (SRTM), which considered social-relationship among users for topic modeling. SRTM jointly modeled texts and social links for learning the topic distribution and topical influence of each user. They applied the SRTM model on Weibo and Twitter data sets and found SRTM could alleviate the effect of topic-irrelevant links by analyzing relational users' topics of each link. Cho et al. [44] proposed the Constrained Latent Space Model (CLSM) that incorporated users' social interactions and attributes for topic modeling and applied their model on six social network sites.

Despite the increase in the user profile and social relationship studies, there are relatively few studies on cross-site UGC modeling. In recent, Lee et al. [20] attempted to establish associations between the user, platform and topic to model the content relevance across multiple SNSs. Their topic modeling results and platform choice prediction results are indeed improved. However, this model is built based on an unreliable assumption and users' platform preferences and platform-word relevance are not properly represented. These state-of-the-art UGC modeling methods remain flawed, and one of the reasons is the lack of understanding of the user content sharing practices on multiple SNSs. Our study attempts to bridge this gap by conducting an empirical study to understand user content sharing practices and build a reliable cross-site UGC model.

3. Preliminary analysis

As mentioned above, abstracting and understanding users' content sharing practices is the foundation of constructing a reliable cross-site UGC model, while it is a challenging task. In our research, we attempt to explore this question through empirical studies that combine both quantitative and qualitative methods. Quantitative methods such as statistical models and machine learning methods can discover users' content sharing characteristics from the large-scale corpus, but they cannot abstract UGC

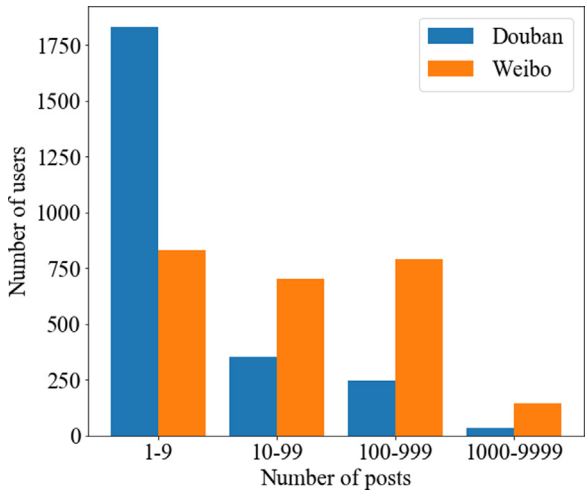

Fig. 1. A comparison of the post quantity between Weibo and Douban.

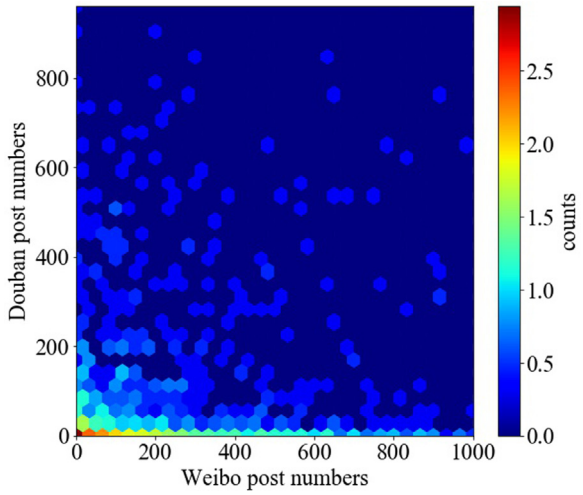

Fig. 2. A visual comparison of Weibo and Douban post numbers by heat map.

procedures on multiple SNSs because this procedure is implicit. On the contrary, qualitative methods can conduct fine-grained observations and understanding of user behaviors and abstract UGC practices, but they cannot uncover content sharing characteristics from the large-scale corpus. Therefore, we combine these two methods to investigate the cross-site UGC practice characteristics. First, we analyze the users' cross-site content sharing behaviors utilizing Weibo and Douban as research sites. Two important findings are that the content quantity of users on different platforms is significantly different, and users tend to utilize different language styles between different platforms. Then, we conduct a semi-structured interview and qualitative analysis according to grounded theory [45]. On one hand, above findings of the quantitative analysis are further understood through such a procedure, and on the other hand, we also obtain new insights that the users' content on different platforms is relevant in terms of topic but different in language style as well.

3.1. Research sites and data collection

We conducted our quantitative analysis on two popular social network sites — Weibo and Douban, which are two of the biggest SNSs in China. Weibo was launched in 2009. By the end of 2019, there are 374.1 million users on that to share content, obtain hot news and follow celebrities [46]. Douban is another popular site in China. It was launched in 2005 and now has 66 million registered users, with its monthly user-visits over 100 million [47]. One reason why we choose Weibo and Douban as our research sites is that many Douban users provide their Weibo ID or URL on their homepage, which allows us to identify cross-site users (i.e. users that have accounts in both Weibo and Douban). By utilizing crawler tools, we randomly obtained 210,000 Douban users based on the snowball sampling rules. After matching, we got 2483 cross-sites users. We then sampled these users' 664,418 posts in Weibo and 185,939 posts in Douban.

3.2. Results of quantitative analysis

We conducted quantitative analysis on content number and linguistic of Weibo and Douban, and we mainly obtained the following insights.

Post number difference. We first give a visual comparison of the users' content quantity on Weibo and Douban, and the results are shown in Figs. 1 and 2. In Fig. 1, the X-axis lists four different posts numbers range from 1 to 9999, and the Y-axis shows the number of users. Fig. 1 indicates that most Douban users share less than 10 messages within three months but they are active on Weibo platform. Actually, the total posts on Weibo reached 664,418, which are nearly three times larger than that of Douban (185,939). Fig. 2 is a heat map showing the number of Weibo posts and Douban posts per user has in our data set. As can be seen from Fig. 2, there is a "bright line" along the X-axis and Y-axis, but the brightness of the X-axis is significantly higher than that of the Y-axis. It indicates that users tend to use one of the two platforms, and in our data set, users prefer to share more content on Weibo. Similar phenomenons have also been found in Ref. [18].

The above analysis visually depicts the quantity difference of user posts between Weibo and Douban. To validate whether this difference is significant, we conducted another quantitative analysis by utilizing the Kolmogorov–Smirnov test (KS-test) method. We treated the number of users' posts on both platforms as two distributions and applied KS-test to these distributions. The result of the KS-test shows $p < 0.001$, which implies that there is indeed a significant difference in users' content quantity between two platforms.

Linguistic difference. We also studied the linguistic differences of users' content on Weibo and Douban. Users' content sharing practices in SNSs are usually driven by intrinsic factors such as motivations and expectations as well as extrinsic factors such as platform functionalities and audiences. The existence of these factors might compel users to utilize different language styles in different platforms. In this section, we conducted a fine-grained linguistic comparison of users' content on Weibo and Douban.

In our study, we investigated linguistic differences by focusing on emotion words (positive words and negative words), tense words (past tense words, present tense words and future tense words), pronouns words (first-person pronouns, second-person pronouns and third-person pronouns), etc. Similar as Ref. [48], we randomly sampled 100 most recent Weibo and Douban status updates from 100 users. For each platform, we totally collected 10,000 posts. After some basic processing such as word segmentation, removing stop words and word category classification (utilize the LIWC Chinese dictionary), we analyzed the language style differences of posts between Weibo and Douban.

Table 1
A comparison of linguistic differences between Weibo and Douban.

Linguistic features		Douban		Weibo		KS-test
Category	Example word	*Mean*	*SD*	*Mean*	*SD*	*p*-value
Word count	–	3257	1762	2549	1191	0.026
Pronouns	I, him, they	71.23	52.1	94.71	64.01	0.046
1st person	I, me, mine	36.81	24.55	57.61	39	0.001
2st person	You, your, yours	16.13	16.57	19.96	17.14	0.211
3rd person	She, him, they	18.29	19.75	17.14	17.87	0.310
Family	Mom, dad, son	7.17	9.28	12.19	10.26	0.046
Friend	Friend, buddy	6.13	5.18	14.71	14.09	0.015
Body	Hands, cheek	6.25	2.05	8.20	2.16	0.039
Health	Illness, pain	13.93	9.98	14.65	15.61	0.699
Present tense	Is, want, does	26.37	15.76	25.4	19.67	0.281
Future tense	Will, gonna	24.02	16.08	23.48	18.41	0.468
Affective	Happy, sad	163.5	95.3	112.27	102.66	0.028
Swear	Swear, fuck	19.81	12.99	9.5	10.59	0.009
Religion	Church, Jesus	9.19	9.74	12.16	11.85	0.111
Work	Job, finish	60.17	48.66	79.19	70.22	0.078
Negative	Sad, hurt, boring	83.70	48.06	31.53	44.43	<0.001
Positive	Nice, love, good	38.54	34.68	67.53	54.26	<0.001
Leisure	Movie, travel	8.37	2.13	7.34	2.50	<0.001
Past tense	Done, had, went	20.81	16.22	14.58	13.46	<0.001
Emoticon	😊😮😄	2.21	1.46	11.24	6.92	<0.001

The results are exhibited in Table 1. It indicates there exist many linguistic differences of users' content between Weibo and Douban in terms of post length, personal pronoun words, family words, friend word, etc. The personal pronoun words, family words, friend words, positive emotion words and emoticons appear more frequently in Weibo content, while the leisure words, past tense words, negative emotion words and swear words are more common-used on Douban. To evaluate the significance of such differences, we also conducted analysis from a user to user perspective by utilizing KS-test. The results show that most word categories exhibit differences between two platforms. Based on which we obtain insights that people tend to adjust their content and utilize different language styles in different SNSs context to express themselves.

3.3. Results of qualitative analysis

We also conducted a qualitative analysis based on the thoughts of grounded theory [49]. Grounded theory is an inductive and comparative methodology that provides systematic guidelines for gathering, synthesizing, analyzing and conceptualizing qualitative data for the purpose of theory construction. It is capable of inducing and extracting implicit patterns from human behavior, which makes it be effective and prevalent in human research like psychology, pedagogy, and communication. According to the procedure of grounded theory, we first conducted semi-structured interviews to sample the analysis data and then coded the corpus iteratively to obtain reliable findings.

We recruited interviewees to participate in our study and collected their experiences and opinions of multiple SNSs use. We selected these interviewees based on standard sampling [49]. According to the thoughts of grounded theory, for our research, participants need to have multiple SNSs accounts, abundant SNSs using experiences, and good presentation skills. As young people are the main participants of SNSs, we recruited 20 interviewees(10 males and 10 females) that age from 20 to 30. These 20 interviewees include students, company employees, etc., and all of them have more than 5 years experiences in SNSs use. Figs. 3 and 4 show the quantity and popularity of different SNSs among these interviewees. Fig. 3 is the cumulative proportion of the number of SNSs these participants use, which shows all of them utilize more than one platform. Fig. 4 shows the popularity of different platforms among these participants. The popularity is in accordance with the 2019 China statistical report on internet development [50], which indicates the good representation of our recruited participants.

Our interview is semi-structured so that we can collect the multifaceted viewpoints of respondents. We asked respondents about two types of questions. The former is about platform selection, and the latter is about language style. For the platform selection, we are interested in how users make choices when they face multiple SNSs to share content. Results show that although interviewees employ multiple SNSs, the common-used platforms are generally the same, which means users essentially have platform preferences. For example, P1, P2, and P6 said:

"The most important reason is that I don't have enough time to keep multiple platforms active at the same time, and it is also not necessary to do that because most of my friends are Weibo users. Moreover, I am used to using the Weibo site. Except for some professional questions, almost everything I followed, such as news, celebrities, etc. can be found on the Weibo platform".

The above illustration suggests three reasons for users' platform preferences. The first is that users are time-constrained, and keeping multiple SNSs active at the same time can be time-consuming. Respondents P4 and P9 also explained *"I won't spend all my time on social media."* Another factor is the audience. As respondent P5 said *"the reason why I share content and present myself is that I want my friends to see my status and keep in touch with friends."* Time and audiences are two external factors that restrict users' platform choices. A more critical intrinsic factor is that users have formed a habit of utilizing specific platforms. Respondents P1, P2, P3, P7, P8, P11, P13, P18 explicitly gave their viewpoints that they were used to using some specific platforms such as Weibo or WeChat.

For platform selections, we also asked respondents whether the content topic would influence their platform choices. Sixteen interviewees believe that content topic has little impacts on their platform choices, and they just choose their favorite platforms to share content. P5 explained that *"I wouldn't be so dogmatic to choose platforms based on the topic of my post. I share content to whichever platform I like, and this process doesn't even require 'thinking'"*. Our findings reveal that the assumption - "users have social media platforms preference specific to topics" [20] are less likely to be reliable. Users' platform choices are independent of the content topic, and users essentially have platform preferences for content sharing.

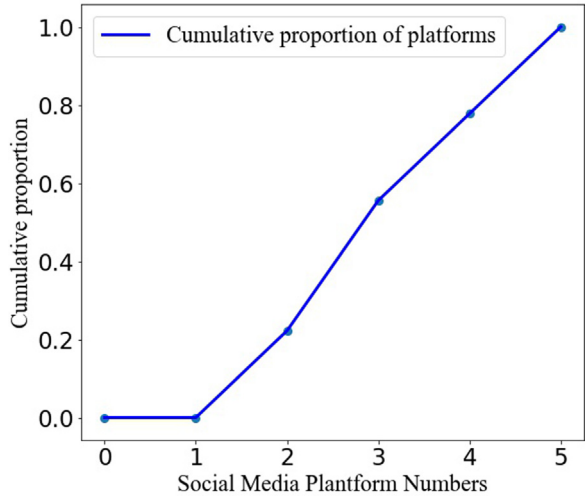

Fig. 3. Cumulative proportion of user's SNSs.

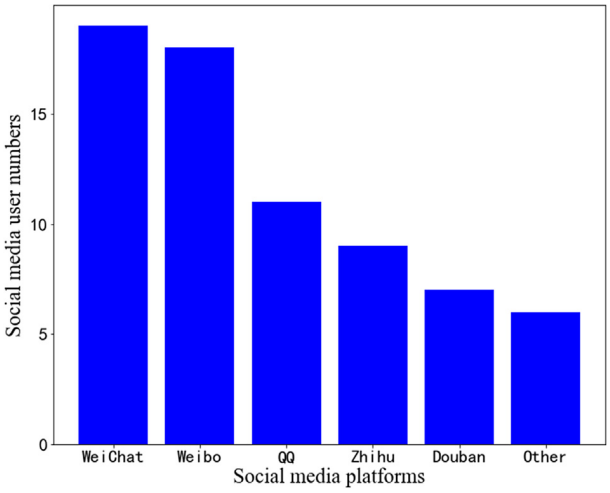

Fig. 4. Popularity of different SNSs.

In the quantitative analysis, we have found that users utilize different language styles in different platforms. We further focused on this phenomenon and investigated what factors influence users' word choices. Several respondents explicitly pointed out that even if they shared a message related to the same event on different platforms, their word choices would vary greatly. The reason can be summarized in two aspects. The former is the 'culture' of SNSs. For example, users use 'Jinli' (锦鲤) and 'C center' (C位) on Weibo to express "bring me good luck", and on Douban the 'Douyou' (豆友) means your friends. These platform cultures have influences on users' word choices and even produce new platform buzzwords over time. Another important factor is the audiences of SNSs, like the P5, P13, P17 said,

"Sometimes I update status related to same events on both WeChat and Weibo, but I will not use extreme words on WeChat because my parents and supervisors also use WeChat, and I do not want to leave a bad impression on them. But for my Weibo posts, the word choices are more open, and sometimes even include swearing, because my Weibo audiences almost are peers and strangers".

Table 2
Notation of C-LDA Model.

Symbol	Description
α, χ, β, δ, γ	Dirichlet prior distribution
U, E, S, K, V	Set of users, platforms, posts, topics and vocabularies
N_u	Posts number of user u
$N_{(u,s)}$	Words number of user u's sth post
π	Bias toward background topic distribution
ϱ_e	Word distribution of platform e
θ_u	Topic distribution of user u
σ_u	Platform distribution of user u
ϕ_k	Word distribution of topic k
ϕ_B	Word distribution of background topic
$Z_{(u,s)}$	Topic of sth post of user u
$Y_{(u,s,n)}$	Controller of user u's nth word of sth post
$P_{(u,s)}$	Platform of sth post of user u
$W_{(u,s,n)}$	user u's nth word of sth post
n_v^k	Times of word v is marked as topic k
n_v^B	Times of word v is marked as background topic B
n_k^u	Times of user u chooses the topic k
n_e^u	Times of user u chooses the platform e
n_v^e	Times of word v appears in the platform e
$Y_{\neg(u,s,n)}$	All other words except $W_{(u,s,n)}$
$Z_{\neg(u,s)}$	All other posts except sth post of user u

The above illustration gives us new insights that users sometimes post content related to the same topic on different platforms. But considering the platform audiences, users generally adjust their language styles to meet their needs for maintaining their image, which results that the content on different platforms is relevant in topic but different in language style.

Based on the quantitative analysis and qualitative analysis, we obtain the following results:

- Affected by platform audiences, personal habits and available time, users essentially have platform preferences for content sharing.
- Affected by platform audiences and platform culture, users tend to utilize different language styles on different platforms.
- Users' content on different platforms is relevant in terms of topic but different in language style.

4. Model

Based on the results of above empirical study, we propose a novel cross-site UGC modeling method named CrossSite-LDA (C-LDA) mostly based on topic model. Before introducing the C-LDA model in detail, we summarize the notations in Table 2.

In the C-LDA model, there are five key elements, namely, user, platform, post, topic and word, which are represented by U, E, S, K and V. We assume that all of the content is a bag of words model. We use $P_{(u,s)}$ to denote the platform of sth post of user u, and $N_{(u,s)}$ is the word count of sth post of user u. Lastly, the $W_{(u,s,n)}$ denotes nth word of user u's sth post.

4.1. C-LDA generative process

Compared with other user generated content modeling methods, the most significant contribution of the C-LDA model is that the C-LDA directly introduces the user-platform and the platform-word distributions to represent the users' platform preferences and the language style differences on different platforms.

Moreover, these two user content sharing characteristics have been validated by empirical study.

Fig. 5 shows the plate diagram of the C-LDA model. We suppose that there are $|U|$ users, $|E|$ platforms, $|K|$ topics, and variables χ, δ, α, β, γ all represent the Dirichlet prior distribution of user-platform, platform-word, user-topic, topic-word and background topic word. We assume that each user u has a multinomial distribution σ_u and θ_u for platform e and topic k, representing their platform and topic selection process. Similarly, for each platform and topic, there is a multinomial distribution of ϱ_e and ϕ_k over vocabulary v. We also set up a background topic word multinomial distribution ϕ_B for each platform to capture the noisy words which have high frequency in SNSs but not meaningful. We set a bias variable $Y_{(u,s,n)}$ to control whether the word belongs to the topic-word distribution ϕ_k or the background topic-word distribution ϕ_B. Similarly, bias variable $Y_{(u,s,n)}$ has a multinomial distribution π and a Dirichlet prior distribution γ.

Now we give the joint probability of C-LDA model and the detail formulation process of its each part. We assume that the C-LDA contains a total of $|K|$ topics and $|V|$ terms. We define the Dirichlet prior distribution α, χ, β, γ for the hidden variables $Z_{(u,s)} = k(k \in \{1, 2, 3, \ldots, K\})$, $P_{(u,s)} = e(e \in \{1, 2, 3, \ldots, E\})$, controller $Y_{(u,s,n)} = c(c \in \{0, 1\})$ and word $W_{(u,s,n)} = v(v \in \{1, 2, 3, \ldots, V\})$. When $Y_{(u,s,n)} = 0$ (i.e. the word is generated from the background topic), the joint probability distribution is:

$$\begin{aligned} &p(Y_{(u,s,n)} = 0, Z_{(u,s,n)} = k, P_{(u,s,n)} = e, W_{(u,s,n)} = v|\alpha, \beta, \gamma, \chi, \delta) \\ &\quad = p(Y_{(u,s,n)} = 0|\pi, \gamma) \cdot p(Z_{(u,k)} = k|\theta_u, \alpha) \cdot p(P_{(u,s)} = e|\sigma_u, \chi) \\ &\quad \cdot p(W_{(u,s,n)} = v|\varrho_u, \phi_B, \delta, \beta). \end{aligned} \tag{1}$$

When $Y_{(u,s,n)} = 1$ (i.e. the word is generated from the real topic), the joint probability distribution is:

$$\begin{aligned} &p(Y_{(u,s,n)} = 1, Z_{(u,s,n)} = k, P_{(u,s,n)} = e, W_{(u,s,n)} = v|\alpha, \beta, \gamma, \chi, \delta) \\ &\quad = p(Y_{(u,s,n)} = 1|\pi, \gamma) \cdot p(Z_{(u,k)} = k|\theta_u, \alpha) \cdot p(P_{(u,s)} = e|\sigma_u, \chi) \\ &\quad \cdot p(W_{(u,s,n)} = v|\varrho_u, \phi_k, \delta, \beta). \end{aligned} \tag{2}$$

In Eq. (1) and Eq. (2), $p(Y_{(u,s,n)} = 0|\pi, \gamma)$, $p(Y_{(u,s,n)} = 1|\pi, \gamma)$, $p(Z_{(u,s)} = k|\theta_u, \alpha)$ and $p(P_{(u,s)} = e|\sigma_u, \chi)$ are evaluated by:

$$\begin{aligned} p(Y_{(u,s,n)} = 0|\pi, \gamma) &= \frac{n_0 + \gamma}{n_0 + n_1 + 2\gamma}, \\ p(Y_{(u,s,n)} = 1|\pi, \gamma) &= \frac{n_1 + \gamma}{n_0 + n_1 + 2\gamma}, \\ p(Z_{(u,s)} = k|\theta_u, \alpha) &= \frac{n_k^u + \alpha}{\sum_{k=1}^{K}(n_k^u + \alpha)}, \\ p(P_{(u,s)} = e|\sigma_u, \chi) &= \frac{n_e^u + \chi}{\sum_{e=1}^{E}(n_e^u + \chi)}, \end{aligned} \tag{3}$$

where n_0 denotes the number of times that background words occur in corpus, n_1 represents the number of times that topic words occur in corpus, n_k^u represents the number of times that the kth topic occurs in the uth user's posts and n_e^u means the number of times that user u choose platform e.

Now we give a look at $p(W_{(u,s,n)} = v|\varrho_e, \phi_B, \delta, \beta)$, $p(W_{(u,s,n)} = v|\varrho_e, \phi_k, \delta, \beta)$ in Eq. (1) and Eq. (2). It describes the generation of a word $W_{(u,s,n)}$ base on the ϕ_k, ϕ_B and ϱ_e, which means the word generation in C-LDA model considers not only topics (including real topic and background topic) but also the influence of platforms. When $Y_{(u,s,n)}$=0, word $W_{(u,s,n)}$ can be evaluated as:

$$\begin{aligned} &p(W_{(u,s,n)} = v|\varrho_e, \phi_B, \delta, \beta) \\ &\quad = p(W_{(u,s,n)} = v|\varrho_e, \delta) \cdot p(W_{(u,s,n)} = v|\phi_B, \beta), \end{aligned} \tag{4}$$

where $p(W_{(u,s,n)}|\varrho_e, \delta)$ is described in Eq. (5) and $p(W_{(u,s,n)}|\phi_B, \beta)$ is formulated as Eq. (6).

$$p(w_{(u,s,n)} = v|\varrho_e, \delta) = \frac{n_v^e + \delta}{\sum_{v=1}^{V}(n_v^e + \delta)}, \tag{5}$$

$$p(w_{(u,s,n)} = v|\phi_B, \beta) = \frac{n_v^B + \beta}{\sum_{v=1}^{V}(n_v^B + \beta)}, \tag{6}$$

where n_v^e denotes the number of times that word v occurs in platform e and n_v^B represents the number of times that word v occurs in background topic B.

When $Y_{(u,s,n)} = 1$, we can evaluate $p(W_{(u,s,n)} = v|\varrho_e, \phi_k, \delta, \beta)$ as follows:

$$\begin{aligned} &p(W_{(u,s,n)} = v|\varrho_e, \phi_k, \delta, \beta) \\ &\quad = p(W_{(u,s,n)} = v|\varrho_e, \delta) \cdot p(W_{(u,s,n)} = v|\phi_k, \beta), \end{aligned} \tag{7}$$

where $p(W_{(u,s,n)} = v|\varrho_e, \delta)$ is also obtained by Eq. (5), and $p(W_{(u,s,n)} = v|\phi_k, \beta)$ is formulated as:

$$p(W_{(u,s,n)} = v|\phi_k, \beta) = \frac{n_v^k + \beta}{\sum_{v=1}^{V}(n_v^k + \beta)}, \tag{8}$$

where n_v^k denotes the number of times that word v occurs in topic k, and the definitions of other symbols are same as Table 2.

Based on the above joint distribution, the generation process of the C-LDA model is exhibited in Algorithm 1. Since most users' content on SNSs is short, we follow the Ref. [16]'s strategy and assume that each post contains only one topic. The sth post of user u is generated as follows. First, for each user, the corresponding topic multinomial distribution θ_u and platform multinomial distribution σ_u are sampled. After then, we sample the topic $Z_{(u,s)}$ and platform $P_{(u,s)}$ of sth post from θ_u and σ_u. Next, we sample each word $W_{(u,s,n)}$ of user's sth post respectively. But as we mentioned before, to distinguish noise words from real topic words, we set two distributions, namely, topic word ϕ_k and background topic word ϕ_B. Both of two distributions share a common Dirichlet prior distribution β. Finally, we set a bias variable $Y_{(u,s,n)}$ to control where the word $W_{(u,s,n)}$ sampled from. If it is a topic word, word $W_{(u,s,n)}$ will be sampled based on $p(W_{(u,s,n)} = v|\phi_k, \beta)$, while if the word is a background word, word $W_{(u,s,n)}$ will be sampled based on $p(W_{(u,s,n)} = v|\phi_B, \beta)$. The bias variable is sampled from a multinomial distribution π, and the π also has a Dirichlet prior distribution γ.

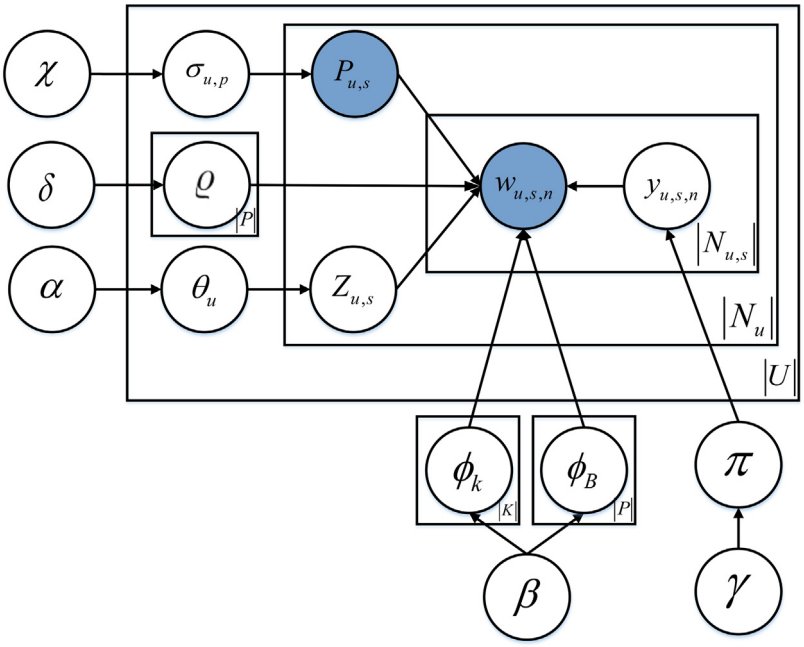

Fig. 5. Plate diagram of C-LDA model.

Algorithm 1 C-LDA model Generative Process

Input: Users' content of different platforms
Output: Distribution of user-topic,topic-word,user-platform,platform-word

Sample ϕ_B, π from $Dir(\beta)$, $Dir(\gamma)$
for each $k \in [1, ..., K]$ **do**
  Sample topic-word distribution ϕ_k from $Dir(\beta)$
end for
for each $u \in [1, ..., U]$ **do**
  Sample user-platform distribution σ_u from $Dir(\chi)$
  Sample user-topic distribution θ_u from $Dir(\alpha)$
  for each user $u's$ post $s \in [1, ..., S]$ **do**
    Sample the topic $Z_{(u,s)}$ from $Multi(\theta_u)$
    Sample the platform $P_{(u,s)}$ from $Multi(\sigma_u)$
    for each word $W_{(u,s,n)} \in [1, ..., N_{u,s}]$ **do**
      Sample word bias $Y_{(u,s,n)}$ from topic bias $multi(\pi)$
      if $Y_{(u,s,n)}$=1 **then**
        Sample word $W_{(u,s,n)}$ from $multi(\varrho_e, \theta_k)$
      end if
      if $Y_{(u,s,n)}$=0 **then**
        Sample word $W_{(u,s,n)}$ from $multi(\varrho_e, \theta_B)$
      end if
    end for
  end for
end for

4.2. Model inference

Same to parameter estimation of many other topic models, we apply Gibbs sampling to infer the parameters of C-LDA model. The procedure is listed below.

- Initialization: For each user, we first randomly assign a topic for each post, and we also randomly initialize the latent coins(topic word or background word) of all words.
- Sampling: Review the corpus and sample a coin for each word and sample a topic for each post according to the update formula (The update formulas have given in Eq. (12), Eq. (13) and Eq. (19)).
- Iterate Step 2 until reaching convergence.
- Calculate the co-occurrence matrixes of topic-word, noisy topic-word, user-topic, etc.

4.2.1. Sampling bias variable $Y_{(u,s,n)}$

The bias variable $Y_{(u,s,n)}$ in our C-LDA model is used to distinguish the noisy word and topic word. When $Y_{(u,s,n)} = 1$, the word $W_{(u,s,n)}$ corresponds to the topic word and the Gibbs sampling formula is exhibited in Eq. (9) and Eq. (13).

$$\begin{aligned}
&p(Y_{(u,s,n)} = 1|Y_{\neg(u,s,n)}, W, Z) \\
&= \frac{p(Y_{(u,s,n)} = 1, Y_{\neg(u,s,n)}, W_{(u,s,n)} = v, W_{\neg(u,s,n)}, Z)}{p(Y_{\neg(u,s,n)}, W, Z)} \\
&= p(Y_{(u,s,n)} = 1, W_{(u,s,n)} = v, Z_{(u,s)} = k|Y_{\neg(u,s,n)}, W_{\neg(u,s,n)}, Z_{\neg(u,s)}) \\
&\cdot p(Y_{\neg(u,s,n)}, W_{\neg(u,s,n)}, Z_{\neg(u,s)}) \cdot \frac{1}{p(Y_{\neg(u,s,n)}, W, Z)} \\
&\propto p(Y_{(u,s,n)} = 1, W_{(u,s,n)} = v, Z_{(u,s)} = k|Y_{\neg(u,s,n)}, W_{\neg(u,s,n)}, Z_{\neg(u,s)}) \\
&\propto p(Y_{(u,s,n)} = 1|\pi_{\neg(u,s,n)}, \gamma) \cdot p(W_{(u,s,n)} = v|\phi_{k,\neg(u,s,n)}, \beta).
\end{aligned} \tag{9}$$

In Eq. (9), we utilize $\neg(u, s, n)$ to denote all other words excluding the current word $W_{(u,s,n)}$, and $p(Y_{(u,s,n)} = 1|\pi_{\neg(u,s,n)}, \gamma)$ and $p(W_{(u,s,n)} = v|\phi_{k,\neg(u,s,n)}, \beta)$ are formulated as Eq. (10) and Eq. (11).

$$p(Y_{(u,s,n)} = 1|\pi_{\neg(u,s,n)}, \gamma) = \frac{n_{1,\neg(u,s,n)} + \gamma}{n_{1,\neg(u,s,n)+} + n_{0,\neg(u,s,n)} + 2\gamma}, \tag{10}$$

$$p(W_{(u,s,n)} = v|\phi_{k,\neg(u,s,n)}, \beta) = \frac{n^k_{v,\neg(u,s,n)} + \beta}{\sum_{v=1}^{V}(n^k_{v,\neg(u,s,n)} + \beta)}. \tag{11}$$

In Eq. (10) and Eq. (11), the $n^k_{v,\neg(u,s,n)}$ denotes the number of times that vth word occurs in topic k but ignoring current word $W_{(u,s,n)}$; $n_{0,\neg(u,s,n)}$ and $n_{1,\neg(u,s,n)}$ represent that number of times that background word and topic word occurs in corpus without considering $W_{(u,s,n)}$, respectively. Therefore, when $Y_{(u,s,n)} = 1$, the

evaluation is formulated as Eq. (12).

$$p(Y_{(u,s,n)}=1|Y_{\neg(u,s,n)},W,Z)=\frac{n_{1,\neg(u,s,n)}+\gamma}{n_{(\cdot)}+2\gamma}\cdot\frac{n^{k}_{v,\neg(u,s,n)}+\beta}{\sum_{v=1}^{V}(n^{k}_{v,\neg(u,s,n)}+\beta)}, \tag{12}$$

where $n_{(\cdot)}$ means the sum of $n_{0,\neg(u,s,n)}$ and $n_{1,\neg(u,s,n)}$.

For $Y_{(u,s,n)}=0$, it represents the word is sampled from the background word distribution. The sampling formula is similar to above equation and is shown in Eq. (13).

$$\begin{aligned}
&p(Y_{(u,s,n)}=0|Y_{\neg(u,s,n)},W,Z)\\
&=\frac{p(Y_{(u,s,n)}=0,Y_{\neg(u,s,n)},W_{(u,s,n)}=v,W_{\neg(u,s,n)},Z)}{p(Y_{\neg(u,s,n)},W,Z)}\\
&=p(Y_{(u,s,n)}=0,W_{(u,s,n)}=v,Z_{(u,s)}=k|Y_{\neg(u,s,n)},W_{\neg(u,s,n)},Z_{\neg(u,s)})\\
&\cdot p(Y_{\neg(u,s,n)},W_{\neg(u,s,n)},Z_{\neg(u,s)})\cdot\frac{1}{p(Y_{\neg(u,s,n)},W,Z)}\\
&\propto p(Y_{(u,s,n)}=0,W_{(u,s,n)}=v,Z_{(u,s)}=k|Y_{\neg(u,s,n)},W_{\neg(u,s,n)},Z_{\neg(u,s)})\\
&\propto p(Y_{(u,s,n)}=0|\pi_{\neg(u,s,n)},\gamma)\cdot p(W_{(u,s,n)}=v|\phi_{B,\neg(u,s,n)},\beta)\\
&=\frac{n_{(0,\neg(u,s,n))}+\gamma}{n_{(\cdot)}+2\gamma}\cdot\frac{n^{B}_{v,\neg(u,s,n)}+\beta}{\sum_{v=1}^{V}(n^{B}_{v,\neg(u,s,n)}+\beta)},
\end{aligned} \tag{13}$$

where $n^{B}_{v,\neg(u,s,n)}$ denotes the number of times that word v occurs in background words but ignoring word $W_{(u,s,n)}$, and other symbols are same as definitions in Eq. (10) and Eq. (11).

4.2.2. Sampling the topic $Z_{(u,s)}$ of user's posts

Based on aforementioned joint distribution and the sampling results of $Y_{(u,s,n)}$. For each user, we sampled the topic for each post and the sampling formula is shown as Eq. (14).

$$\begin{aligned}
&p(Z_{(u,s)}=k|Z_{\neg(u,s)},P,W,Y)\\
&=\frac{p(Z_{(u,s)}=k,Z_{\neg(u,s)},P_{(u,s)}=e,P_{\neg(u,s)},W_{(u,s,n)}=v,W_{\neg(u,s,n)},Y)}{p(Z_{\neg(u,s)},P,W,Y)}\\
&=p(Z_{(u,s)}=k,P_{(u,s)}=e,W_{(u,s,n)}=v|Z_{\neg(u,s)},P_{\neg(u,s)},W_{\neg(u,s,n)},Y)\\
&\cdot p(Z_{\neg(u,s)},P_{\neg(u,s)},W_{\neg(u,s,n)},Y)\cdot\frac{1}{p(Z_{\neg(u,s)},P,W,Y)}\\
&\propto p(Z_{(u,s)}=k,P_{(u,s)}=e,W_{(u,s,n)}=v|Z_{\neg(u,s)},P_{\neg(u,s)},W_{\neg(u,s,n)})\\
&=p(Z_{(u,s)}=k|\theta_{u,\neg(u,s)})\cdot p(P_{(u,s)}=e|\sigma_{u,\neg(u,s)})\cdot p(W_{(u,s,n)}=v|\phi_{k,\neg(u,s)})\\
&\cdot p(W_{(u,s,n)}=v|\varrho_{(e,\neg(u,s))}).
\end{aligned} \tag{14}$$

In Eq. (15), $\neg(u,s)$ means ignoring the current post (sth post of user u), and $p(Z_{(u,s)}=k|\theta_{u,\neg(u,s)})$, $p(P_{(u,s)}=e|\sigma_{u,\neg(u,s)})$, $p(W_{(u,s,n)}=v|\phi_{k,\neg(u,s)})$ and $p(W_{(u,s,n)}=v|\varrho_{(e,\neg(u,s))})$ are formulated as Eq. (15), Eq. (16), Eq. (17) and Eq. (18).

$$p(Z_{(u,s)}=k|\theta_{u,\neg(u,s)})=\frac{n^{u}_{k,\neg(u,s)}+\alpha}{\sum_{k=1}^{K}(n^{u}_{k,\neg(u,s)}+\alpha)}, \tag{15}$$

where $n^{u}_{k,\neg(u,s)}$ denotes the number of times that the kth topic occurs in the uth user's posts but ignoring the current post (sth post).

$$p(P_{(u,s)}=e|\sigma_{u,\neg(u,s)})=\frac{n^{u}_{e,\neg(u,s)}+\chi}{\sum_{e=1}^{E}(n^{u}_{e,\neg(u,s)}+\chi)}, \tag{16}$$

where $n^{u}_{e,\neg(u,s)}$ represents that without considering the sth post, the number of times that user u update the status on platform e.

$$p(W_{(u,s,n)}=v|\phi_{k,\neg(u,s)})=\frac{n^{k}_{v,\neg(u,s)}+\beta}{\sum_{v=1}^{V}(n^{k}_{v,\neg(u,s)}+\beta)}, \tag{17}$$

where $n^{k}_{v,\neg(u,s)}$ means the number of times that the vth word in dictionary occurs in the kth topic without considering the sth post.

$$p(W_{(u,s,n)}=v|\varrho_{k,\neg(u,s)})=\frac{n^{e}_{v,\neg(u,s)}+\delta}{\sum_{v=1}^{V}(n^{e}_{v,\neg(u,s)}+\delta)}, \tag{18}$$

where $n^{e}_{v,\neg(u,s)}$ denotes the number of times that the vth word in dictionary occurs in the eth platform without considering the sth post.

Therefore, based on above evaluation, the $p(Z_{(u,s)}=k|Z_{\neg(u,s)},P,W,Y)$ can be formulated as:

$$\begin{aligned}
&p(Z_{(u,s)}=k|Z_{\neg(u,s)},P,W,Y)\\
&=\frac{n^{u}_{k,\neg(u,s)}+\alpha}{\sum_{k=1}^{K}(n^{u}_{k,\neg(u,s)}+\alpha)}\cdot\frac{n^{u}_{e,\neg(u,s)}+\chi}{\sum_{e=1}^{E}(n^{u}_{e,\neg(u,s)}+\chi)}\cdot\frac{n^{k}_{v,\neg(u,s)}+\beta}{\sum_{v=1}^{V}(n^{k}_{v,\neg(u,s)}+\beta)}\\
&\cdot\frac{n^{e}_{v,\neg(u,s)}+\delta}{\sum_{v=1}^{V}(n^{e}_{v,\neg(u,s)}+\delta)}.
\end{aligned} \tag{19}$$

5. Experiments

In this section, we first introduce the data sets, baselines, evaluation metrics and model settings. Then we conduct an experimental evaluation of the C-LDA model. We compare the C-LDA model with four state-of-the-art methods: DMM, BTM, Twitter-LDA and MultiLDA model from perplexity, words coherence, topics KL divergence, UCI and UMass metrics. Our experiments prove that the C-LDA model outperforms DMM, BTM, Twitter-LDA and MultiLDA in above metrics, which suggests its superior performances on modeling users' content characteristics on multiple SNSs context.

5.1. Experiment setup

Data set. As we mentioned before, we have obtained 850,037 real cross-site posts, among which 664,418 are Weibo posts and 185,939 are Douban posts. Based on the above real data sets, we train our model and conduct experiments to validate the performances of our model.

Baselines. We use the DMM, BTM, Twitter-LDA model and MultiLDA model as our baselines. These four models are the state-of-the-art topic models, which have been widely used to model the user generated content in SNSs. However, it should be noted that the DMM, BTM and Twitter-LDA models do not take platform information into account. It assumes that all posts are generated from a single platform. Therefore, these three models actually cannot represent the relevance of users content on different platforms.

Metrics. We use perplexity, word coherence, topics KL divergence, UCI [51] and UMass [52] as our metrics and these five metrics are widely used to evaluate the generalization performance and topic modeling ability of the language model.

A common understanding of perplexity is the uncertainty of the sentence. This uncertainty refers to the degree of perplexity (Eq. (20) gives the mathematical formula of perplexity), and a model with lower perplexity indicates that the model has a strong generalization ability.

$$perplexity(D_{test})=exp\{-\frac{\sum_{d=1}^{M}logp(w_d)}{\sum_{d=1}^{M}N_d}\} \tag{20}$$

The final results in above topic models are distributions of user-topic, topic-word, etc. These topics we learned are relatively

independent in semantics, and there should be large semantic differences between different topics. Therefore, we apply symmetric Kullback–Leibler (KL) divergence to measure the topic differences among users' content. The greater the difference, the more obvious the discrimination between topics, indicating that the model works better. We calculate the symmetric KL divergence of the two topic-word distributions as follows. In Eq. (21), θ_a and θ_b are topic word distributions, and θ_{mid} is their average distribution.

$$sKL(\theta_a, \theta_b) = \frac{1}{2}(D_{KL}(\theta_a \| \theta_{mid}) + D_{KL}(\theta_b \| \theta_{mid})) \tag{21}$$

KL divergence describes the difference between topics, while word coherence scores measure the semantic consistency within one topic. For the topic model, words within one topic should be relevant, thus we use an improved PMI [52] (Pointwise Mutual Information) as another evaluation indicator. For each topic, words coherence scores for the top M words is calculated according to Eq. (22). Then we average them to obtain the final scores. The larger the words coherence values, the better the model models the user generated content. To evaluate the impact of M on the results, in this paper we set M to 10, 15 and 20. In Eq. (22), the $\#(w)$ represents the number of posts containing the word w. The $w^t = (w_1^t, \ldots, w_M^t)$ represents the top M words with high probability in the topic t, and $\#(w_m, w_l)$ refers to the number of times that two words appear in the same post.

$$Coherence(t; w^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} log \frac{\#(w_m^{(t)}, w_l^t) + 1}{\#(w_l^{(t)})} \tag{22}$$

The UCI metric also defines a word pair's score to be Pointwise Mutual Information between two words. But the difference is that the word probabilities in UCI metric are computed by counting word co-occurrence frequencies in a sliding window over an external corpus [51]. According to the author, this metric can be thought of as an external comparison to known semantic evaluations. The UCI metric is calculated as shown in Eq. (23), and ϵ show in Eq. (23) is set to 1. In the following experiments, we use a 20 words sliding window passed over the Sougou Chinese Internet data set[1] (as our external corpus), and same as [51], we compute the coherence with the top 10 words from each topic that have the highest weight.

$$score(w_i, w_j, \epsilon) = log \frac{p(w_i, w_j) + \epsilon}{p(w_i)p(w_j)} \tag{23}$$

The UMass metric defines the score to be based on document co-occurrence. As shown in Eq. (24), $D(w_i, w_j)$ counts the number of documents containing words w_i and w_j, and $D(w_j)$ counts the number of documents containing word w_j. The ϵ is also set to 1. But different with the UCI metric, the UMass metric utilizes original corpus (corpus used to train our C-LDA model) to compute these counts. According to the author, this metric is more intrinsic in nature. It attempts to confirm that the models learned data known to be in the corpus [52].

$$score(w_i, w_j, \epsilon) = log \frac{D(w_i, w_j) + \epsilon}{D(w_j)} \tag{24}$$

Parameter setting. Same as the model setting in Ref. [16], in Twitter-LDA and DMM model, we treat all the content of each user as a document. For BTM, we aggregate all the user's content into one document. It is important to note that the DMM, BTM and Twitter-LDA model does not consider platform information associated with the posts. Neither the user-platform distribution nor platform-words distribution can be expressed by above models. In MultiLDA and C-LDA model, we treat the user's content on different platforms as a document [20]. All of the above models use the same data set and have the same prior parameters, iterations numbers $T = 500$, $\alpha = 50/K$, $\beta = 0.01$, $\delta = 0.01$, $\chi = 0.01$, $\gamma = 0.01$.

[1] http://www.sogou.com/labs/resource/ca.php.

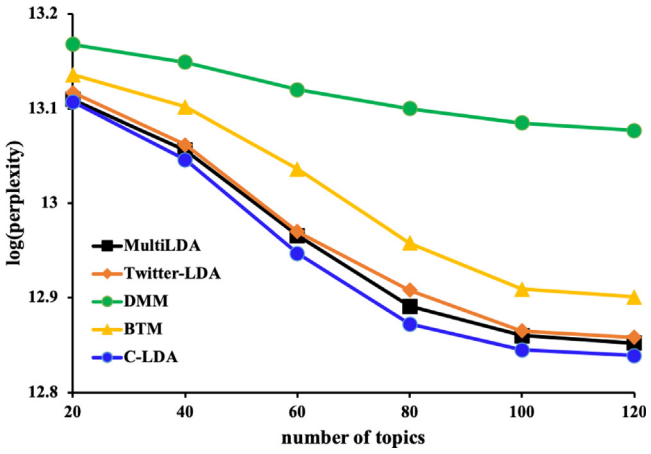

Fig. 6. Log(Perplexity) DMM, BTM, Twitter-LDA, MultiLDA and C-LDA.

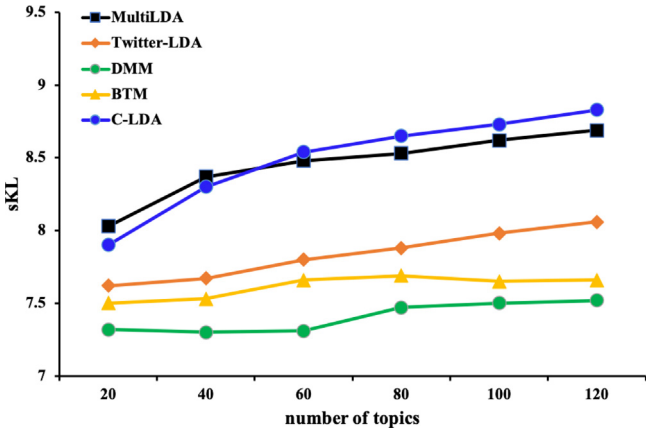

Fig. 7. Topics KL divergence of DMM, BTM, Twitter-LDA, MultiLDA and C-LDA.

5.2. Performance evaluation

5.2.1. Perplexity

The perplexity comparison results are shown in Fig. 6. We notice that the DMM model (green line) has the largest perplexity values compared to other four models. The reason may be that DMM model attempts to solve the problem of word sparsity by assuming that each post contains only one topic. However, social media content often contains noisy words (words have high frequency but not meaningful) such as "haha", "oh", "add", "call", etc, and DMM model fail to distinguish these noisy words and topic words. Fig. 6 shows that as the number of the topics k increases, the BTM, Twitter-LDA, MultiLDA, and our C-LDA model achieves smaller perplexity. When the topics number reaches 100, the perplexity values of these four models are almost unchanged, which indicates that these models begin to converge. Fig. 6 also shows that the perplexity values of BTM model are slightly larger than Twitter-LDA, MultiLDA and C-LDA models. We notice that MultiLDA and the Twitter-LDA have similar perplexity trends, but there is an improvement in MultiLDA when the number of topics reaches 80. This is consistent with the results of Ref. [20]. The results also reveal that the perplexity

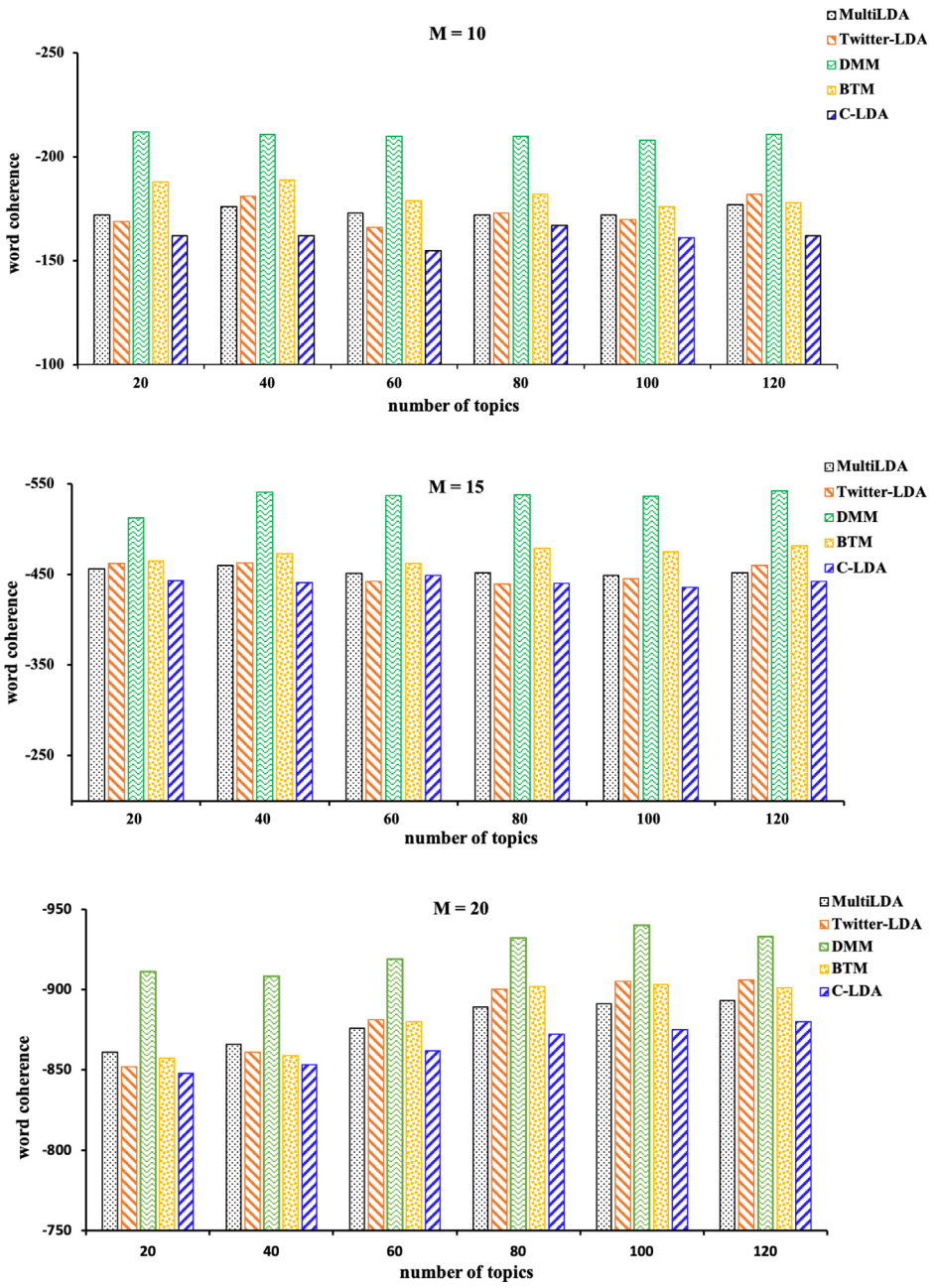

Fig. 8. Word coherence of DMM, BTM, Twitter-LDA, MultiLDA and C-LDA.

values of our C-LDA model (blue line) are generally lower than that of the MultiLDA model, and the C-LDA model shows certain improvement when the number of topics k reaches 60, which is most noticeable when the number of topics k reaches 80. The possible explanation is that users essentially have platform preferences rather than choosing a platform based on the topic of the content.

5.2.2. Topics KL divergence

The comparison results of topics KL divergence are shown in Fig. 7. It indicates that C-LDA and MultiLDA model achieve larger topic KL divergence than DMM, BTM and Twitter-LDA model. The main reason is that these three models ignore the platform elements in the process of cross-site UGC modeling. Fig. 7 also suggests that the C-LDA model achieves larger KL divergence values gradually with the number of topics increases, and when the topic number k reaches 50, the C-LDA model outperforms MultiLDA model. It means that the assumption - "users have social media platforms preference specific to topics [20]" are less likely to be reliable, and users essentially have platform preferences for content sharing. We also notice that the DMM, BTM and Twitter-LDA model generate topics with smaller KL divergence compared with C-LDA and MultiLDA. The results suggest that these models may not suitable for cross-site UGC modeling, especially because these models are incapable to represent the content relevance across different platforms.

5.2.3. Word coherence within one topic

In the experiments of word coherence, we compute the coherence scores with the top 10, 15, 20 words from each topic that have the highest weight. The comparison results are shown in Fig. 8. First, the results reveal in almost all cases, the word coherence scores generated by our C-LDA model are larger than other

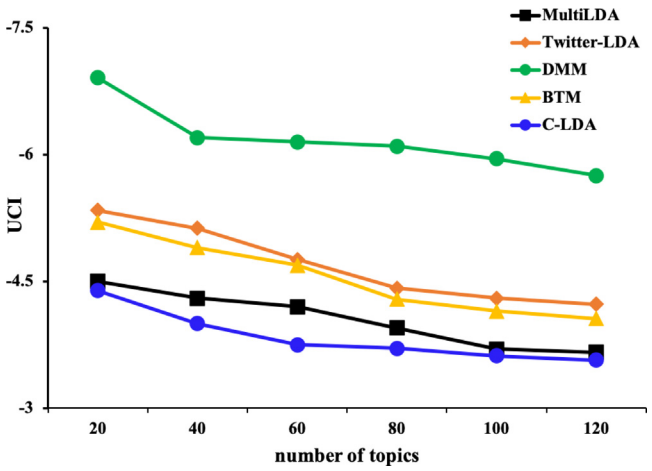

Fig. 9. The UCI Metric of DMM, BTM, MultiLDA, Twitter-LDA and C-LDA.

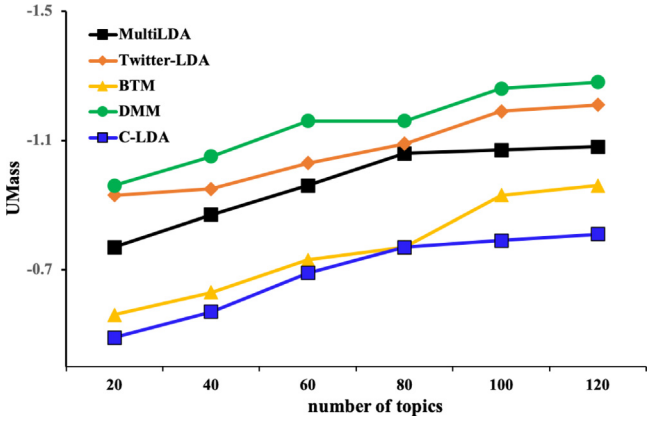

Fig. 10. The UMass Metric of MultiLDA, Twitter-LDA, DMM, BTM and C-LDA.

models, which means C-LDA model can generate high-quality topics and is highly robust. Second, Fig. 8 shows a noticeable phenomenon that the DMM model has the lowest word coherence scores in all cases, regardless of the value of words number M and topics number k. The reason is that the topic generated by the DMM model contain noisy words. This also suggests that distinguishing topic words from noisy words is very important for improving the quality of cross-site UGC modeling. Third, as we expected, we find that as the number of words M increases, word coherence scores declined in all models. Fig. 8 also shows that when $M = 10$ and $M = 15$, the word coherence scores of Twitter-LDA, MultiLDA, DMM, BTM and C-LDA show little change as the number of topics increases. But when the number of words M reaches 20, we find that the word coherence scores of all models decreases as the number of topics k increases. One possible explanation is that when the number of words M is less than 20, the topic cannot be fully represented. In short, Fig. 8 show that our C-LDA model performs better than other models in word coherence evaluation, and from the view of Pointwise Mutual Information, the continuous increase in the number of topics might have a negative impact on modeling user generated content.

5.2.4. UCI metric

The evaluation results of UCI metric are shown in Fig. 9. It suggests that the C-LDA model and the MultiLDA model are

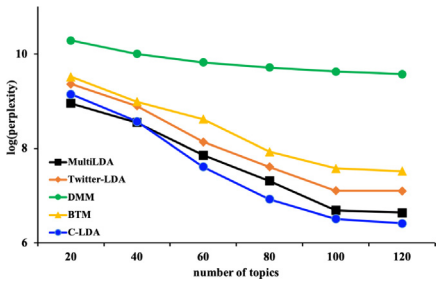

Fig. 11. Perplexity of five models.

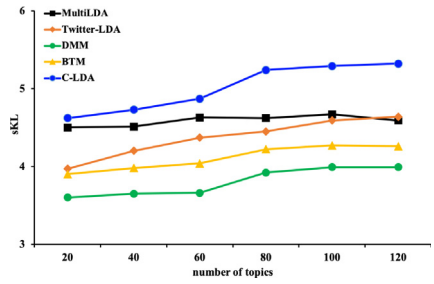

Fig. 12. KL divergence of five models.

superior to other models. That is mainly because in both of two models, the platform elements are introduced into their cross-site UGC modeling process. The results also reveal that the UCI values of our C-LDA model are generally larger than that of the MultiLDA model, and the C-LDA model shows certain improvement when the number of topics k reaches 40, which is most noticeable when the number of topics k reaches 60. The possible explanation is as we said before that users essentially have platform preferences rather than "choosing a platform based on the topic of content". We also notice that the UCI score of DMM model is generally lower than the other four models. The reason is as we mentioned before: although DMM model assumes that each post contains only one topic, it does not distinguish between topic words and noisy words. As a result, the noisy words are mixed into their topics, so its UCI score is much lower than the other models. Fig. 9 also shows that the BTM and Twitter-LDA model have similar performance in UCI metric. Finally, we notice as the number of topics k increases, the UCI scores of the five models all show an increasing trend, which is consistent with the results observed in [51].

5.2.5. UMass metric

Fig. 10 gives the comparison results of the UMass Metric. We notice that the UMass value of the C-LDA model is larger than other models, which means C-LDA model performs better than others. Fig. 10 also reveals that the C-LDA and Multi-LDA models begin to converge when the number of topics reaches 80, while the other three models converge when the number of topics reaches 100. Such results indicate that for the UGC modeling in the multiple platform context, MultiLDA and C-LDA model might converge slightly faster than other models. In addition, compared with UCI metric, we find that the BTM model outperform Multi-LDA in UMass metric. This is mainly because the UMass metric is calculated based on documents, while the MultiLDA model treats each user's content published on different platforms as a document, and is modeled based on the assumption that "users have a platform preference for topics". Therefore, its UMass value is lower than BTM and C-LDA model, which also confirms that the

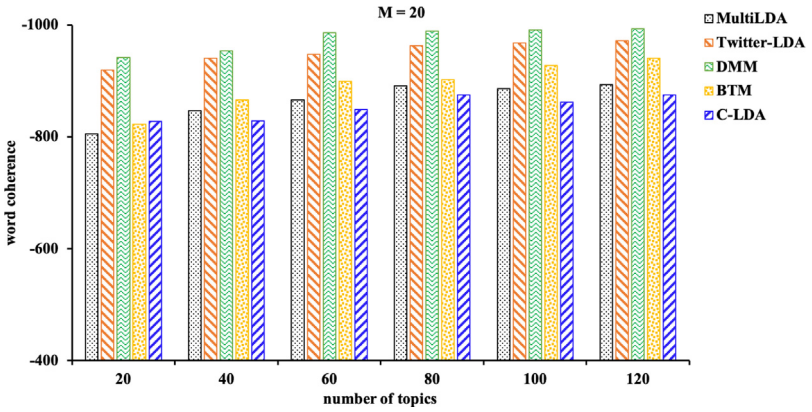

Fig. 13. Word coherence of five model when $M = 20$.

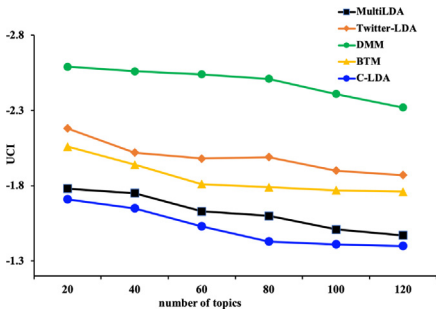

Fig. 14. The UCI score of five models.

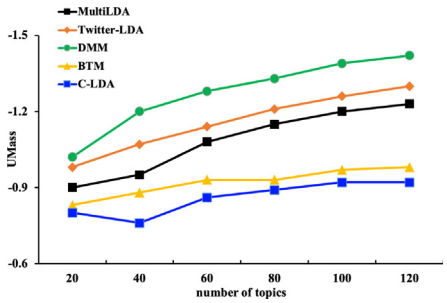

Fig. 15. The UMass score of five model.

above assumption may not be reliable. Finally, Fig. 10 also shows that the UMass scores of the above five models all decrease with the increase of the number of topics k, and eventually tend to converge.

5.3. Model robustness verification

The data set we utilized in the above experiments mainly includes two platforms — Weibo and Douban, and is limited to Chinese corpus, which may not be enough to prove that the C-LDA model has strong robustness on different platforms. Thus, to verify the robustness of our C-LDA model, we built another cross-site data set which includes two other platforms — Facebook and Twitter(Some users share their Facebook and Twitter accounts on Aboutme.com,[2] giving us the chance to identify cross-platform users). For Twitter data we collected user tweets by API,[3] and for Facebook, we collect user status by open source Data crawler.[4] In the end, we crawled a total of 32,431 user posts. After some basic processing such as lowercase the word, word segmentation and removing stop words, we conduct experiments and evaluate the performances of the C-LDA model base on above data set. In the same model setting, we make comparison with above four baselines: DMM, BTM, Twitter-LDA and MultiLDA model, and the evaluation metrics are also perplexity, KL divergence, word coherence, UCI and UMass. The only difference is that in UCI metric, we utilize Wikipedia data[5] as our external corpus. The experiment results are shown in Figs. 11–15

On the whole, the performances of C-LDA model in Facebook–Twitter data set is very similar to that in Weibo–Douban data set. First, like the results in Figs. 6–7, Figs. 11–12 show that the C-LDA model has the smallest perplexity and the largest topic KL divergence, indicating that our C-LDA model are superior to other models. We also notice that the perplexity of MultiLDA is similar to C-LDA model (outperform Twitter-LDA, DMM, BTM, but lower than C-LDA), which suggests that the platform element is critical for cross-site UGC modeling and should not be ignored. Secondly, Fig. 13 shows that the C-LDA model has the largest word coherence scores in most cases except that it is lower than Twitter-LDA and BTM model when the number of topic $k = 20$, but the model has not yet converged. We can also observe a noticeable phenomenon from Fig. 13 that DMM model has lowest word coherence scores compared with other models. The possible reason is same as we explained before: DMM model assumes that each post contains only one topic, but it does not distinguish between topic words and noisy words. Third, as for the evaluation results of UCI and UMass metrics, Figs. 14–15 also shows that the results are similar to those in the Weibo–Douban platforms. The C-LDA model has the largest UCI and UMass scores and outperform other models. In summary, we evaluate the performances of the C-LDA model on different platforms. The experimental results show that no matter on the Weibo–Douban data set or on Facebook–Twitter data set, our C-LDA model has superior performances, which indicates C-LDA model's strong robustness.

5.4. Complexity of C-LDA model

In this part, we analyze the time and space complexity of C-LDA model, and the meaning of the notations is also shown in Table 1.

First, we analyze the space complexity of the C-LDA model. As shown in Algorithm 1, we can see C-LDA model needs to

[2] http://aboutme.com/.
[3] https://developer.twitter.com/en.
[4] https://github.com/rugantio/fbcrawl.
[5] https://dumps.wikimedia.org/enwiki/latest/.

store following variables: user-platform distribution σ, platform-word distribution ϱ, user-topic distribution θ and topic word distribution ϕ. Their space sizes are $U{\cdot}E$, $E{\cdot}V$, $U{\cdot}K$, $K{\cdot}V$, where U, E, K, V represent the size of users, platforms, topics, and words, respectively. We can see none of them need much space. Actually, when dealing with huge data sets, C-LDA spends most space to store the train corpus. We need to store all content published by all users on different platforms. Assuming that there are D documents and the average length of each document is $\bar{L}$, the space complexity of the C-LDA model is $O(D\bar{L})$.

Next, we analyze the time complexity of each iteration of C-LDA model. We suppose that the Gibbs sampling algorithm runs T times. In each iteration, as shown in Algorithm 1, first, we need to initialize the model hyperparameters. Then, for each user, we sample the topic for each post according to the formula $p(Z_{(u,s)}|Z_{\neg(u,s)}, P, W, Y)$. Next, for words in each post, we assign a topic or background topic according to the formula $p(Y_{(u,s,n)}|Y_{\neg(u,s,n)}, W)$, and complete a sampling iteration. In summary, the time complexity of the C-LDA model is $O(T \cdot U \cdot S \cdot K \cdot \bar{V})$ where T is the number of iterations, U is the number of users, S is the total number of content posted by users, and K is the number of topics, $\bar{V}$ is the average length of each post of the user.

5.5. Model parameters influence

There are two types of parameters in our C-LDA model, the former is the hyperparameters of the C-LDA model such as α, β, γ, δ, and χ. The hyperparameter value is usually determined based on the experiences or actual experimental results. For α, β, and γ, existing research [16] has proved that the model has the best performances when $\alpha = 50/K$ (K is the number of topics), β=0.01 and γ=0.01, so we set $\alpha = 50/K$, $\beta = 0.01$ and $\gamma = 0.01$. For δ and χ we test three different candidate values 0.1, 0.01, 0.001, and find that when δ and χ take 0.01, the model works best, so we finally set $\alpha = 50/K$, $\beta = 0.01$, $\gamma = 0.01$, δ and χ as the training hyperparameters in this paper.

For the later parameters, topics number k and words number M, we have also given a detailed analysis in our comparative experiment (Section 5.2 Performance Evaluation). We set topics number k from 20 to 120 and set words number M to 10, 15, 20 respectively. Then, we evaluate how the topics number k and words number M influence the C-LDA performances. Overall, as the number of topics k increases, C-LDA model achieves smaller perplexity and larger topic KL divergence, indicating that as the number of topics k increases, C-LDA model achieves better modeling results. When the number of topics k reaches 100, the values of perplexity, KL divergence UCI and UMass are almost unchanged, suggesting that C-LDA model has reached convergence. The experimental results also show that as the number of words M increases, the word coherence score shows a decreasing trend, which is also in line with the expected results.

6. Conclusion

In this paper, we propose a reliable cross-site UGC model named CrossSite LDA (C-LDA). Instead of giving the model directly without reasonableness verification, we first conduct an empirical study to investigate the users' content sharing practices in multiple SNSs context. The study reveals that users essentially have platform preferences and tend to utilize different language styles on different platforms. Base on these findings, we propose the C-LDA model to represent the procedure and content characteristics of cross-site content sharing practice revealed by above empirical study. We evaluate the performances of C-LDA model based on the real SNSs data sets. Experiment results show that the C-LDA has better performances in perplexity, word coherence, topic KL divergence UCI and UMass metrics compared with existing models, which suggests its superior performances on modeling users' content characteristics in multiple SNSs context. For future works, we would like to extend the C-LDA model by introducing different types of posts like text, image, etc to enhance the representation ability of the model.

CRediT authorship contribution statement

Baoxi Liu: Data curation, Methodology, Writing - original draft. **Peng Zhang:** Conceptualization, Methodology, Writing - review & editing. **Tun Lu:** Conceptualization, Resources. **Ning Gu:** Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grants No. 61902075, No. 61932007 and National Key Research and Development Project under Grant No. 2018YFC0832303.

References

[1] M.A. Aaron Smith, Social media use in 2018, 2018, https://www.pewresearch.org/internet/2018/03/01/social-media-use-in-2018/. (Accessed 1 March 2018).
[2] A. Tumasjan, T.O. Sprenger, P.G. Sandner, I.M. Welpe, Predicting elections with twitter: What 140 characters reveal about political sentiment, in: Fourth International AAAI Conference on Weblogs and Social Media.
[3] N. Anstead, B. O'Loughlin, Social media analysis and public opinion: The 2010 UK general election, J. Comput.-Mediat. Commun. 20 (2) (2014) 204–220.
[4] P. Zhang, H. Gu, M. Gartrell, T. Lu, D. Yang, X. Ding, N. Gu, Group-based latent Dirichlet allocation (group-LDA): Effective audience detection for books in online social media, Knowl.-Based Syst. 105 (2016) 134–146.
[5] A. Paudel, B.R. Bajracharya, M. Ghimire, N. Bhattarai, D.S. Baral, Using personality traits information from social media for music recommendation, in: 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS), IEEE, 2018, pp. 116–121.
[6] K. Xu, X. Zheng, Y. Cai, H. Min, Z. Gao, B. Zhu, H. Xie, T.-L. Wong, Improving user recommendation by extracting social topics and interest topics of users in uni-directional social networks, Knowl.-Based Syst. 140 (2018) 120–133.
[7] W.X. Zhao, S. Li, Y. He, E.Y. Chang, J.-R. Wen, X. Li, Connecting social media to e-commerce: Cold-start product recommendation using microblogging information, IEEE Trans. Knowl. Data Eng. 28 (5) (2015) 1147–1159.
[8] Y. Zhang, M. Pennacchiotti, Predicting purchase behaviors from social media, in: Proceedings of the 22nd International Conference on World Wide Web, ACM, 2013, pp. 1521–1532.
[9] J. Zhang, X. Hu, Y. Zhang, H. Liu, Your age is no secret: Inferring microbloggers' ages via content and interaction analysis, in: Tenth International AAAI Conference on Web and Social Media, 2016.
[10] C. Peersman, W. Daelemans, L. Van Vaerenbergh, Predicting age and gender in online social networks, in: Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents, ACM, 2011, pp. 37–44.
[11] D. Tang, Z. Zhang, Y. He, C. Lin, D. Zhou, Hidden topic–emotion transition model for multi-level social emotion detection, Knowl.-Based Syst. 164 (2019) 426–435.
[12] Y. Liu, L. Zhang, L. Nie, Y. Yan, D.S. Rosenblum, Fortune teller: predicting your career path, in: Thirtieth AAAI Conference on Artificial Intelligence, 2016.
[13] J. Ramos, et al., Using tf-idf to determine word relevance in document queries, in: Proceedings of the First Instructional Conference on Machine Learning, vol. 242, Piscataway, NJ, 2003, pp. 133–142.
[14] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (Jan) (2003) 993–1022.

[15] L. Rigouste, O. Cappé, F. Yvon, Inference and evaluation of the multinomial mixture model for text clustering, Inf. Process. Manage. 43 (5) (2007) 1260–1280.
[16] W.X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, X. Li, Comparing twitter and traditional media using topic models, in: European Conference on Information Retrieval, Springer, 2011, pp. 338–349.
[17] X. Yan, J. Guo, Y. Lan, X. Cheng, A biterm topic model for short texts, in: Proceedings of the 22nd International Conference on World Wide Web, 2013, pp. 1445–1456.
[18] R. Ottoni, D. Las Casas, J.P. Pesce, W. Meira Jr., C. Wilson, A. Mislove, V. Almeida, Of pins and tweets: Investigating how users behave across image- and text-based social networks, in: Eighth International Aaai Conference on Weblogs and Social Media, 2014.
[19] L. Manikonda, V.V. Meduri, S. Kambhampati, Tweeting the mind and instagramming the heart: Exploring differentiated content sharing on social media, in: Tenth International AAAI Conference on Web and Social Media, 2016.
[20] R.K.-W. Lee, T.-A. Hoang, E.-P. Lim, On analyzing user topic-specific platform preferences across multiple social media sites, in: Proceedings of the 26th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, 2017, pp. 1351–1359.
[21] M. Sleeper, W. Melicher, H. Habib, L. Bauer, L.F. Cranor, M.L. Mazurek, Sharing personal content online: Exploring channel choice and multi-channel behaviors, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, ACM, 2016, pp. 101–112.
[22] X. Zhao, C. Lampe, N.B. Ellison, The social media ecology: User perceptions, strategies and challenges, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, ACM, 2016, pp. 89–100.
[23] P. Zhang, H. Zhu, T. Lu, H. Gu, W. Huang, N. Gu, Understanding relationship overlapping on social network sites: a case study of weibo and douban, Proc. ACM Hum.-Comput. Interact. 1 (CSCW) (2017) 1–18.
[24] H. Liu, B. Liu, H. Zhang, L. Li, X. Qin, G. Zhang, Crowd evacuation simulation approach based on navigation knowledge and two-layer control mechanism, Inform. Sci. 436 (2018) 247–267.
[25] B. Liu, H. Liu, H. Zhang, X. Qin, A social force evacuation model driven by video data, Simul. Model. Pract. Theory 84 (2018) 190–203.
[26] S.E. Lindley, C.C. Marshall, R. Banks, A. Sellen, T. Regan, Rethinking the web as a personal archive, in: Proceedings of the 22nd International Conference on World Wide Web, ACM, 2013, pp. 749–760.
[27] X. Zhao, S.E. Lindley, Curation through use: understanding the personal value of social media, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2014, pp. 2431–2440.
[28] M. Rosen-Zvi, T. Griffiths, M. Steyvers, P. Smyth, The author-topic model for authors and documents, in: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, AUAI Press, 2004, pp. 487–494.
[29] Q. Diao, J. Jiang, F. Zhu, E.-P. Lim, Finding bursty topics from microblogs, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, Association for Computational Linguistics, 2012, pp. 536–544.
[30] A. Pal, A. Herdagdelen, S. Chatterji, S. Taank, D. Chakrabarti, Discovery of topical authorities in instagram, in: Proceedings of the 25th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, 2016, pp. 1203–1213.
[31] T.-A. Hoang, E.-P. Lim, Microblogging content propagation modeling using topic-specific behavioral factors, IEEE Trans. Knowl. Data Eng. 28 (9) (2016) 2407–2422.
[32] J.Y. Jang, K. Han, P.C. Shih, D. Lee, Generation like: comparative characteristics in instagram, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, ACM, 2015, pp. 4039–4042.
[33] E. Ferrara, R. Interdonato, A. Tagarelli, Online popularity and topical interests through the lens of instagram, in: Proceedings of the 25th ACM Conference on Hypertext and Social Media, ACM, 2014, pp. 24–34.
[34] J. Xu, R. Compton, T.-C. Lu, D. Allen, Rolling through tumblr: characterizing behavioral patterns of the microblogging platform, in: Proceedings of the 2014 ACM Conference on Web Science, ACM, 2014, pp. 13–22.
[35] J. Xuan, J. Lu, G. Zhang, R.Y. Da Xu, X. Luo, Bayesian Nonparametric relational topic model through dependent gamma processes, IEEE Trans. Knowl. Data Eng. 29 (7) (2016) 1357–1369.
[36] S. Li, Y. Zhang, R. Pan, M. Mao, Y. Yang, Recurrent attentional topic model, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.
[37] H. Wang, D.-Y. Yeung, Towards Bayesian deep learning: A framework and some existing methods, IEEE Trans. Knowl. Data Eng. 28 (12) (2016) 3395–3408.
[38] F. Abel, E. Herder, G.-J. Houben, N. Henze, D. Krause, Cross-system user modeling and personalization on the social web, User Model. User-Adapt. Interact. 23 (2–3) (2013) 169–209.
[39] H. Cho, J. Yeo, S.-W. Hwang, Event grounding from multimodal social network fusion, in: 2016 IEEE 16th International Conference on Data Mining (ICDM), IEEE, 2016, pp. 835–840.
[40] A. Farseev, L. Nie, M. Akbari, T.-S. Chua, Harvesting multiple sources for user profile learning: a big data study, in: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, ACM, 2015, pp. 235–242.
[41] T. Chen, M.A. Kaafar, A. Friedman, R. Boreli, Is more always merrier?: a deep dive into online social footprints, in: Proceedings of the 2012 ACM Workshop on Workshop on Online Social Networks, ACM, 2012, pp. 67–72.
[42] M. Magnani, L. Rossi, The ml-model for multi-layer social networks, in: 2011 International Conference on Advances in Social Networks Analysis and Mining, IEEE, 2011, pp. 5–12.
[43] W. Guo, S. Wu, L. Wang, T. Tan, Social-relational topic model for social networks, in: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, ACM, 2015, pp. 1731–1734.
[44] Y.-S. Cho, G. Ver Steeg, E. Ferrara, A. Galstyan, Latent space model for multi-modal social data, in: Proceedings of the 25th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, 2016, pp. 447–458.
[45] K. Charmaz, L.L. Belgrave, Grounded theory, in: The Blackwell Encyclopedia of Sociology, Wiley Online Library, 2007.
[46] S.W. data center, 2018 Weibo User Development Report, 2019, https://data.weibo.com/report/index/. (Accessed 15 March 2019).
[47] Douban, About douban, 2019, https://jobs.douban.com/about//. Accessed 2019.
[48] H. Lin, L. Qiu, Two sites, two voices: Linguistic differences between facebook status updates and tweets, in: International Conference on Cross-Cultural Design, Springer, 2013, pp. 432–440.
[49] B.G. Glaser, A.L. Strauss, Discovery of Grounded Theory: Strategies for Qualitative Research, Routledge, 2017.
[50] CNNIC, The 44th China Statistical Report on Internet Development, 2019, http://www.cac.gov.cn/2019-08/30/c_1124938750.htm/. (Accessed 1 August 2019).
[51] K. Stevens, P. Kegelmeyer, D. Andrzejewski, D. Buttler, Exploring topic coherence over many models and many topics, in: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012, pp. 952–961.
[52] D. Mimno, H.M. Wallach, E. Talley, M. Leenders, A. McCallum, Optimizing semantic coherence in topic models, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2011, pp. 262–272.