On Analyzing User Topic-Specific Platform Preferences Across Multiple Social Media Sites

Roy Ka-Wei LEE
Living Analytics Research Centre
Singapore Management University
roylee.2013@smu.edu.sg

Tuan-Anh HOANG
L3S Research Center
Leibniz University of Hanover, Germany
hoang@l3s.de

Ee-Peng LIM
Living Analytics Research Centre
Singapore Management University
eplim@smu.edu.sg

ABSTRACT

Topic modeling has traditionally been studied for single text collections and applied to social media data represented in the form of text documents. With the emergence of many social media platforms, users find themselves using different social media for posting content and for social interaction. While many topics may be shared across social media platforms, users typically show preferences of certain social media platform(s) over others for certain topics. Such platform preferences may even be found at the individual level. To model social media topics as well as platform preferences of users, we propose a new topic model known as MultiPlatform-LDA (MultiLDA). Instead of just merging all posts from different social media platforms into a single text collection, MultiLDA keeps one text collection for each social media platform but allowing these platforms to share a common set of topics. MultiLDA further learns the user-specific platform preferences for each topic. We evaluate MultiLDA against TwitterLDA, the state-of-the-art method for social media content modeling, on two aspects: (i) the effectiveness in modeling topics across social media platforms, and (ii) the ability to predict platform choices for each post. We conduct experiments on three real-world datasets from Twitter, Instagram and Tumblr sharing a set of common users. Our experiments results show that the MultiLDA outperforms in both topic modeling and platform choice prediction tasks. We also show empirically that among the three social media platforms, "*Daily matters*" and "*Relationship matters*" are dominant topics in Twitter, "*Social gathering*", "*Outing*" and "*Fashion*" are dominant topics in Instagram, and "*Music*", "*Entertainment*" and "*Fashion*" are dominant topics in Tumblr.

Keywords

user preference; multiple social networks; topic modeling

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.3038914

1. INTRODUCTION

Motivation. According to a 2014 survey conducted by Pew Research Center [1], more than half of the internet users (52%) use two or more social media platforms. Users may tweet on Twitter, post pictures on Instagram, blog in Tumblr, and be engaged in other social media platforms. This surge of users using multiple social media platforms has opened up new challenges to learn users' topical interests.

Learning user topical interests in social media is a widely studied research topic [6, 3, 24, 18, 10, 29]. Most works study topics in the text content of social media. There are also studies that learn latent topics (or clusters) from user behaviors (e.g., forwarding posts, expressing "likes", etc.) and network features [21, 8]. Most of them demonstrate the applications of the learned user topical interests in e-commerce and services recommendation [28, 30]. Nevertheless, all these studies have been confined to textual content from single social media platforms.

With the same users using multiple social media platforms, the holistic approach is to learn user topical interests considering the users' combined social media data. For example, one could learn from a user's Twitter data that she is interested in IT gadgets but the same user is interested in food and fashion based on her Instagram posts. This approach, however, requires two major challenges to be tackled, namely *user linkage* and *multi-platform topic modeling*. The former refers to linking user accounts from different social media platforms belonging to the same users. The latter is topic modeling in the multi-platform context where heterogeneous media types and users' platform preferences are the additional model elements. User linkage is a highly active research topic but is not the focus of this paper [23, 25, 27, 4]. In this paper, we assume that user linkage has already been performed and focus on the second major challenge, multi-platform topic modeling.

Research Objectives and Contributions. We propose a generative model that can learn topics from users' combined social media data as well as their platform preferences. A simple way to perform multi-platform topic modeling is to apply an existing topic model such as LDA [2] on the directly combined content of the same users. Unfortunately, such an approach does not work when the content is of different media types, nor does it consider the platform preferences of the users when the latter share content of different topics.

Figure 1 shows the methodology used in our research. We first construct a topic model for multiple social media platforms. In this paper, we propose *MultiPlatform-LDA (MultiLDA)*, a topic model that jointly learns the topical interests and platform preferences of users who have accounts on multiple social media platforms. Next, we have a data processing step to gather social media data from multiple platforms, to conduct user linkage (if required) and to turn all rich media content (i.e., images and videos) to words using the state-of-the-art image captioning software. The identification and crawling of this dataset itself is a major challenge. In total, we have gathered about 5.8 million text and rich media posts from 2,785 users who have accounts on Twitter, Instagram and Tumblr.

Finally, we evaluate the multi-platform topic model(s). We perform two sets of experiments to evaluate MultiLDA: (i) we first use likelihood and perplexity to evaluate the model's ability to learn users' topical interests from observed text and rich media posts, and (ii) we also evaluate the predictive power of MultiLDA model. Lastly, we also conduct an empirical study on the real-world data using our model, where we learn and report the popular topics on different social media platforms and the individuals' platform preferences.

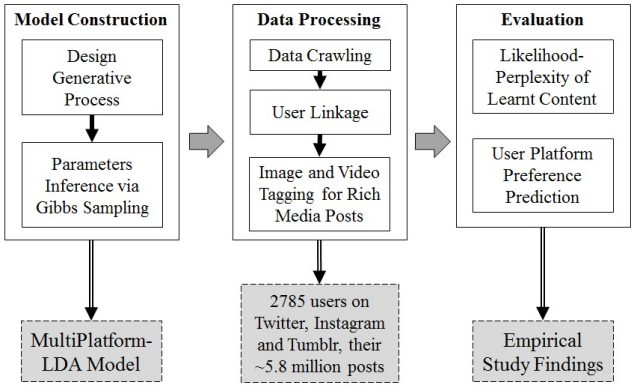


Figure 1: Research Framework

On the whole, this paper improves the state-of-the-art topic modeling research and derives several interesting findings. These include:

- In modeling text and rich media content from multiple social media platforms, MultiLDA outperforms TwitterLDA, another state-of-the-art topic model for modeling social media text.
- In the prediction of users' platform choices, MultiLDA predicted users' platform choice with an high average accuracy of 0.947, outperforming TwitterLDA's average accuracy by 30%.
- In our empirical study, we found different social media platforms having different popular topics. E.g., users prefer to post music related topics in Tumblr while sharing food related topics in Instagram. Also, while most users tend to conform the general topic distribution of social media platforms (i.e., post content with popular topics in the platform), individual user platform preference still exists. MultiLDA was able to model this individual user platform preference effectively.

Paper Outline. The rest of this paper is organized as follows: We first discuss the related works in Section 2. We then present the MultiLDA model in Section 3. Section 4 presents the real-world data and experimental evaluations for our proposed model. The empirical study on the real-world data using our model will also be discussed in this section. Finally, we conclude the paper and discuss the future works in Section 5.

2. RELATED WORKS

In this section, we review two groups of existing research works related to our research. The first group discusses the research studies on analyzing topics in social media particularly Twitter, Tumblr and Instagram, where we collected datasets to evaluate our proposed model. The second group focuses on studying user behaviors across multiple social media platforms.

2.1 Topic Analysis in Social Media

2.1.1 Single Platform

Topic analysis of social media content is a widely researched field. Pal *et. al.* proposed a framework to find topical authorities in Instagram by inferring the Instagram users' topic interests from their self-reported biographies [20]. Jang *et. al.* attempted to characterize and detect Instagram user age group by applying LDA model [2] to learn the difference in topic interests between teens and adult users [11]. Ferrara *et. al.* conducted an empirical study and analyzed the topic interests of Instagram users using hashtags in the captions of Instagram posts [6]. Similar studies were also done in Tumblr. Xu *et. al.* proposed to learn the topic interests of Tumblr users using the tags labeled on their posts [24]. In an empirical study conducted on Tumblr, Chang *et. al.* applied the LDA model to discover Tumblr users' topic interests [3].

Other than Instagram and Tumblr, Twiiter is another most widely studied social media platform in topic analysis. Michelson *et. al.* began studying the topic interests of Twitter users by examining the entities mentioned by users in their tweets [18]. Hong *et. al.* applied LDA model and author-topic model [22] to discover the topic interests of Twitter users [10]. Further research works were also done to improve the performance of LDA model by experimenting different ways of forming documents using tweets [16]. Other works also proposed to jointly model individual user and community topic interests [8].

Among the many works on Twitter topic analysis, the work by Zhao *et. al.* [29] is particularly close to ours. In this work, the researchers proposed TwitterLDA model which is a variant of LDA, in which (a) tweets of the same user are aggregated to form documents; (b) each user has a topic distribution; (c) users share a common background topic; and (d) a topic is assigned to each tweet. It is important to note that TwitterLDA was designed to learn topics from a single platform. This is different from our proposed model where we take into consideration the topic distributions for different platforms.

2.1.2 Single Platform Multiple Behaviors

Besides using social media content to model user topical interests in social media platforms, researchers also proposed modeling the topics of posts using multiple behaviors adopted by the users. For example, Qiu *et. al.* proposed to model the topics of tweets and their associating posting behaviors (e.g., retweet) in Twitter [21, 9]. Our research differs from such studies that we model user-generated content from multiple social media platforms, instead of behaviors.

2.1.3 Multiple Platforms

There are also works that apply topic models on multiple social media platforms. Guo *et. al.* proposed a model that considers social-relationship among users for topic modeling and applied their model on Sina Weibo and Twitter datasets [7]. Cho *et. al.* designed a model that incorporates users' social interactions and attributes for topic modeling and applied their model on six social media platforms [5]. However, many of these works do not link the users across platforms but perform the topic analysis on each platform independently. Our research differs from such studies by analyzing topical interests of a set of common users with accounts on multiple social media platforms.

2.2 User Behaviors Across Multiple Social Media Platforms

The vast and increasing volume of research on user linking provide the foundation for deeper user behavior studies across multiple social media platforms [23, 25, 27, 4]. For instance, Kumar *et. al.* analyzed the user migration patterns across seven social media platforms by examining the same individual user's participation in different platforms [12]. Zafarani and Liu also conducted an empirical study to investigate user-platform joining behavior and found that most users join and stay active in less than three social media platforms [26].

Despite the increase in cross social media platforms studies, there are relatively few studies on user cross-platform content publishing behaviors. Meo *et. al.* presented a macro-level analysis of users sharing behaviors on Flickr, Delicious and StumpbleUpon [17]. Ottoni *et. al.* studied the users' activities across Twitter and Pinterest and found that users tend to post items to Pinterest before posting them on Twitter [19]. Similar findings were made by Lim *et. al.* who also found that users exhibited varied information sharing behaviors on different social media platforms [14]. While both [19] and [14] investigate the posting of same content across multiple platforms, i.e., the duplication of posts across different platforms, the topic interests and the diverse types of content are however neglected. For instance, a user may not simply duplicate and share a post across platforms. Instead, she may share different types of content that share the same topic interests across different platforms. For example, a user may share a text post in Twitter and a photo in Instagram. Although the types of content shared on the two platforms are different, both the text and photo may share the same topic (e.g., Food).

Our study attempts to bridge this gap in the state-of-the-art works by examining the topic interest of the diverse types of content published by users on multiple platforms. Furthermore, we also attempt to study how the topic interests of a post could influence the user's platform choice to publish the post. For example, a user who is interested in architecture design and fashion may choose to share his architecture design posts in Tumblr while sharing the fashion posts in Instagram.

3. MULTIPLATFORM-LDA MODEL

3.1 Notations

Before we present our proposed model, we first summarize the notations in Table 1. Given a set of users and their posts on some social media platforms, we use $\mathcal{U}$, $\mathcal{S}$, and $\mathcal{P}$ to denote the sets of users, posts, and platforms respectively. We use S_u to denote the number of user u's posts across all the platforms. The s-th post of user u is then denoted by the pair $(p_{u,s}, N_{u,s})$ where $p_{u,s}$ is the platform of the post, and $N_{u,s}$ is the content of the post. In this work, we focus on text content and assume that $N_{u,s}$ is a bag of words. The n-th word of the post $(p_{u,s}, N_{u,s})$ is then denoted by $N_{u,s,n}$. Lastly, we use $\mathcal{V}$ to denote the vocabulary of all the words found in the dataset.

Table 1: Notations

Symbol	Description
$\mathcal{V}$	Vocabulary of words in users' content
$\mathcal{U}/\mathcal{S}/\mathcal{P}$	Sets of users, posts and platforms
K	Number of topics
$\mathcal{S}_u$	Set of posts of user u
$N_{u,s}$	Set of words of s-th post of user u
$p_{u,s}$	Platform of s-th post of user u
$w_{u,s,n}$	n-th word of s-th post of user u
$z_{u,s}$	Topic of s-th post of user u
$y_{u,s,n}$	Coin of n-th word of s-th post of user u
ϕ_k	Word distribution of topic k
ϕ^B	Word distribution of background topic
π	Bias toward background topic
θ_u	Topic distribution of user u
$\sigma_{u,k}$	Platform distribution of user u for topic k
$\mathcal{P}$	Bag of platforms of all posts
$\mathcal{C}$	Bag of coins of all words
$\mathcal{C}_{-u,s,n}$	Bag of coins of all words except $\omega_{u,s,n}$
$\mathcal{Z}$	Bag of topics of all posts
$\mathcal{Z}_{-u,s}$	Bag of topics of all posts except the s-th post of user u
$\mathbf{D}^c_{-u,s,n}$	Tuple $(\mathcal{C}_{-u,s,n}, \mathcal{Z}, \mathcal{S}, \mathcal{P}, \alpha, \beta, \mu, \gamma)$
$\mathbf{n_y}(c, \mathbf{D}^c_{-u,s,n})$	#times in $\mathbf{D}^c_{-u,s,n}$ that words are associated with the coin c
$\mathbf{n_b}(\omega, \mathbf{D}^c_{-u,s,n})$	#times in $\mathbf{D}^c_{-u,s,n}$ that the word ω is associated with the background topic
$\mathbf{n_w}(\omega, z, \mathbf{D}^c_{-u,s,n})$	#times in $\mathbf{D}^c_{-u,s,n}$ that the word ω is associated with topic z
$\mathbf{D}^z_{-u,s}$	Tuple $(\mathcal{Z}_{-u,s}, \mathcal{C}, \mathcal{S}, \mathcal{P}, \alpha, \beta, \mu, \gamma)$
$\mathbf{n_{wz}}(\omega, z, \mathbf{D}^z_{-u,s,n})$	#times in $\mathbf{D}^z_{-u,s}$ that word ω is associated with topic z
$\mathbf{n_p}(p, z, \mathbf{D}^z_{-u,s})$	#times in $\mathbf{D}^z_{-u,s}$ that posts about topic z are associated with platform p
$\mathbf{n_z}(k, u, \mathbf{D}^z_{-u,s})$	#times in $\mathbf{D}^z_{-u,s}$ that posts of user u are associated with topic k

3.2 Generative Process

Our model is designed based on the assumption that users have social media platforms preference specific to topics. That is, given a topic, users may prefer to generate content about the topic more on a specific social media platform than other platforms. For example, a user may post more gourmet related photos in Instagram but post more tweets

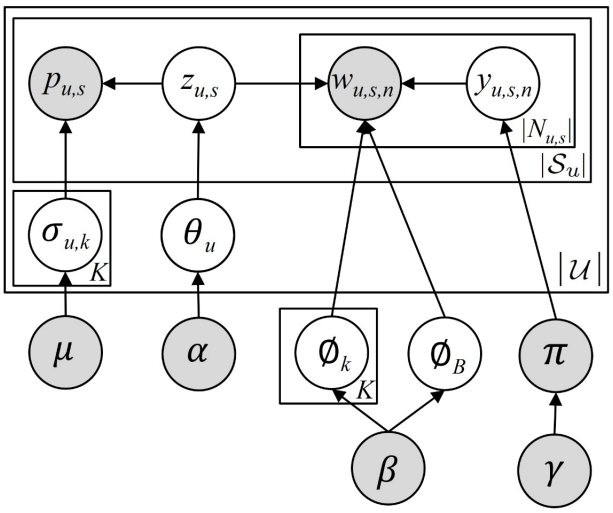


Figure 2: Plate diagram of MultiLDA model

about sports and entertainment on Twitter. Thus, to model the users' interests accurately, it is important to learn both the topics of the user generated content and topic-specific platform preference.

Based on the above assumption, we design MultiLDA model with plate diagram shown in Figure 2, to simulate the generation of observed users' content from their hidden topical interests and topic-specific platform preference. We assume that there are K topics across all the social media platforms. Each topic k has a multinomial distribution ϕ_k over the vocabulary $\mathcal{V}$. We also assume that there is a background topic that captures the background words used across the platforms. Similarly, this background topic also has a multinomial distribution ϕ_B over the vocabulary. The bias toward the background topic is characterized by a binomial distribution π. To capture users' topical interests, we assume that each user u has a multinomial distribution θ_u over K topics. Lastly, to capture the u's topic-specific platform preference, we assume that, for each topic k, u has a multinomial distribution $\sigma_{u,k}$ over the set of platforms P. The bias toward the background topic π has Beta prior γ, and the topics' word distributions ϕ_k and ϕ_B have common symmetric Dirichlet prior β. Similarly, users' topic distributions θ_u's and users' topic-specific platform distributions $\sigma_{u,k}$'s have symmetric Dirichlet priors α and μ respectively.

In MultiLDA model, the s-th post of user u is generated as follows. The post's topic $z_{u,s}$ is first chosen by sampling from u's topic distribution θ_u. As posts are short, we assume that each post has only one topic. The post's content is then generated by sampling its words where each word is sampled independently from the others. For each word $w_{u,s,n}$, a biased coin $y_{u,s,n}$ is flipped to decide where the word is sampled from. The bias of the coin is set to the bias toward the background topic π. The word is sampled from the word distribution of the chosen topic (i.e., $\phi_{z_{u,s}}$) if the coin is *head*, i.e., $y_{u,s,n} = 1$, or that of background topic (i.e., ϕ^B) otherwise. Lastly, the post's platform is chosen by sampling from u's platform distribution specific to the chosen topic, i.e., $\sigma_{u,k}$. The whole generative process of the MultiLDA model is summarized in Algorithm 1.

Algorithm 1 Generative Process for MultiLDA

sample $\phi_B \sim Dir(\beta)$
sample $\pi \sim Beta(\gamma)$
⊡ "Topic Plate"
for topic $k \in \{1, \cdots, K\}$ **do**
  sample the topic's word distribution $\phi_k \sim Dir(\beta)$
end for
⊡ "User Plate"
for user $u \in \mathcal{U}$ **do**
  sample u' topic distribution $\theta_u \sim Dir(\alpha)$
  for topic $k \in \{1, \cdots, K\}$ **do**
    sample u's platform distribution for the topic $\sigma_{u,k} \sim Dir(\mu)$
  end for
  ⊡ "Post Plate"
  for post $s \in \mathcal{S}_u$ **do**
    sample the post' topic $z_{u,s} \sim Multi(\theta_u)$
    ⊡ "Word Plate"
    for word $w_{u,s,n}$ of the post **do**
      sample the word's coin $y_{u,s,n} \sim Bernoulli(\pi)$
      if $y_{u,s,n} = 0$ **then**
        sample the word from background topic $w_{u,s,n} \sim Multi(\phi_B)$
      else
        sample the word from the post's topic $w_{u,s,n} \sim Multi(\phi_{z_{u,s}})$
      end if
    end for
    sample the post's platform $p_{u,s} \sim Multi(\sigma_{u,z_{u,s}})$
  end for
end for

3.3 Inference Via Gibbs Sampling

Like in other LDA-based models, the inference problem in the MultiLDA model is intractable [2]. We therefore adopt sampling-based approach to estimate the model's parameters from a given dataset. Specifically, we first randomly initialize the latent topics of posts and latent coins of all words in the dataset. We then use a collapsed Gibbs sampler [15] to iteratively sample the coin for every word and topic for every post. These iterations result in a sample set which allows us to estimate the model's parameters.

Sampling coin for a word. Consider the word $\omega_{u,s,n}$, we denote the bag of coins of all other words by $\mathcal{C}_{-u,s,n}$. Also, we denote the bag of topics of all the posts by $\mathcal{Z}$, and denote the bag of platforms of all posts by $\mathcal{P}$. The coin $y_{u,s,n}$ is then sampled according to the following equations.

$$p(y_{u,s,n} = 0|\mathbf{D}^c_{-u,s,n}) \propto \frac{\mathbf{n_b}(\omega_{u,s,n}, \mathbf{D}^c_{-u,s,n}) + \beta}{\sum_{\omega \in \mathcal{V}}[\mathbf{n_b}(\omega, \mathbf{D}^c_{-u,s,n}) + \beta]} \cdot \frac{\mathbf{n_y}(0, \mathbf{D}^c_{-u,s,n}) + \gamma_0}{\mathbf{n_y}(0, \mathbf{D}^c_{-u,s,n}) + \mathbf{n_y}(1, \mathbf{D}^c_{-u,s,n}) + \gamma_0 + \gamma_1} \quad (1)$$

$$p(y_{u,s,n} = 1|\mathbf{D}^c_{-u,s,n}) \propto \frac{\mathbf{n_w}(\omega_{u,s,n}, z_{u,s}, \mathbf{D}^c_{-u,s,n}) + \beta}{\sum_{\omega \in \mathcal{V}}[\mathbf{n_w}(\omega, z_{u,s}, \mathbf{D}^c_{-u,s,n}) + \beta]} \cdot \frac{\mathbf{n_y}(1, \mathbf{D}^c_{-u,s,n}) + \gamma_1}{(\mathbf{n_y}(0, \mathbf{D}^c_{-u,s,n}) + (\mathbf{n_y}(1, \mathbf{D}^c_{-u,s,n}) + \gamma_0 + \gamma_1)} \quad (2)$$

In Equations 1 and 2, $\mathbf{D}^c_{-u,s,n}$ denotes the tuple ($\mathcal{C}_{-u,s,n}$, $\mathcal{Z}, \mathcal{S}, \mathcal{P}$,

$\alpha, \beta, \mu, \gamma)$, and $\mathbf{n_y}(c, \mathbf{D}^c_{-u,s,n})$ ($c = 0$ or 1) is the number of times in $\mathbf{D}^c_{-u,s,n}$ that words are associated with the coin c. In Equation 1, $\mathbf{n_b}(\omega, \mathbf{D}^c_{-u,s,n})$ is the number of times in $\mathbf{D}^c_{-u,s,n}$ that the word ω is associated with the background topic. Similarly, in Equation 2, $\mathbf{n_w}(\omega, z, \mathbf{D}^c_{-u,s,n})$ is the number of times in $\mathbf{D}^c_{-u,s,n}$ that the word ω is associated with topic z. In these equations, the first terms on the right hand side are the posterior information of $y_{u,s,n}$, i.e., the likelihoods that the word $\omega_{u,s,n}$ is generated by the background topic (Equation 1) or by topic $z_{u,s}$ (Equation 2). The second terms are the prior information of $y_{u,s,n}$, i.e., the likelihood of $y_{u,s,n} = c$ given coins of all other words.

Sampling topic for a post. Now consider the s-th post of user u, we denote the bag of topics of all other posts by $\mathcal{Z}_{-u,s}$. Also, we denote the bag of coins of all the words by $\mathcal{C}$. The topic $z_{u,s}$ is then sampled according to the following equation.

$$p(z_{u,s} = z|\mathbf{D}^z_{-u,s}) \propto \propto \prod_{y_{u,s,n}=1} \frac{\mathbf{n_{wz}}(\omega_{u,s,n}, z, \mathbf{D}^z_{-u,s}) + \beta}{\sum_{w \in \mathcal{V}}[\mathbf{n_{wz}}(\omega, z, \mathbf{D}^z_{-u,s}) + \beta]} \cdot \frac{\mathbf{n_p}(p_{u,s}, z, \mathbf{D}^z_{-u,s}) + \mu}{\sum_{p \in \mathcal{P}}[\mathbf{n_p}(p, z, \mathbf{D}^z_{-u,s}) + \mu]} \cdot \frac{\mathbf{n_z}(z, u, \mathbf{D}^z_{-u,s}) + \alpha}{\sum_{k=1}^{K} \mathbf{n_z}(k, u, \mathbf{D}^z_{-u,s}) + \alpha} \quad (3)$$

In Equation 3, $\mathbf{D}^z_{-u,s}$ denotes the tuple $(\mathcal{Z}_{-u,s}, \mathcal{C}, \mathcal{S}, \mathcal{P}, \alpha, \beta, \mu, \gamma)$. $\mathbf{n_{wz}}(\omega, z, \mathbf{D}^z_{-u,s,n})$ is the number of times in $\mathbf{D}^z_{-u,s}$ that word ω is associated with topic z. $\mathbf{n_p}(p, z, \mathbf{D}^z_{-u,s})$ is the number of times in $\mathbf{D}^z_{-u,s}$ that posts about topic z are associated with platform p. Lastly, $\mathbf{n_z}(k, u, \mathbf{D}^z_{-u,s})$ is the number of times in $\mathbf{D}^z_{-u,s}$ that posts of user u are associated with topic k. In the equation, the first and second terms on the right hand side are the posterior information of $z_{u,s}$, i.e., the likelihoods that the post's words and platform are generated by the topic z respectively. The third term is the prior information of $z_{u,s}$, i.e., the likelihood of $z_{u,s} = z$ given topics of all other posts.

In our experiments, we used symmetric priors with $\alpha = 50/K$, $\beta = 0.01$, $\mu = 0.01$, and $\gamma_0 = \gamma_1 = 0.01$. Each time, we run the model for 600 iterations of Gibbs sampling. The first 100 iterations were ignored to remove the effect of the random initialization. We take 25 samples with a gap of 20 iterations in the last 500 iterations to estimate the model's parameters.

4. EXPERIMENTS

In this section, we perform some experiments to evaluate MultiLDA and to compare with TwitterLDA, the state-of-the-art topic model for short social media posts. We first elaborate how we obtain the multi-platform social media dataset required for the experiments. Next, we describe the experiment setup and evaluation criteria. The platform choice prediction task is then introduced as part of our evaluation experiments. Finally, we present several empirical findings on user topics and platform choices learned by MultiLDA.

4.1 User Linked Social Media Dataset

Our model evaluation requires a dataset combining social media data from multiple platforms and we want these platforms to share some common users. We selected three popular social media platforms, namely (a) Twitter, a short-text microblogging site; (b) Instagram, a photo-sharing social media site; and (c) Tumblr, a social networking and blogging site that supports a wide range of rich media such as pictures, videos, etc.

We began by gathering a set of 234,289 Singapore-based Twitter users who declared Singapore location in their user profiles. These users were identified by an iterative snowball sampling process starting from a small seed set of well known Singapore Twitter users followed by traversing the follow links to other Singapore Twitter users until the sampling iteration did not get many more new users. From these Twitter users, we obtained a subset of them having user account(s) on Instagram, Tumblr, or both. Among the above Twitter users, We selected users who also mentioned their Instagram and/or Tumblr accounts (in the form of username or hyperlink) in their Twitter bio descriptions. As some users chose to mention their other social media accounts on Instagram or Tumblr, we also gathered the linked user accounts of other social media platforms by scanning the bio descriptions of Instagram and Tumblr users. As some of these linked user accounts may no longer exist, we performed checking of account existence using the respective social media APIs. Those user accounts which no longer exist were removed from our dataset. We further filtered away inactive users who did not make at least five posts in year 2015 on any platform which the users have accounts with.

Table 2: Number of users in each particular social media platform who use another platform

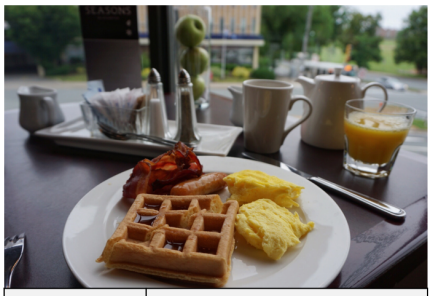
	Twitter	Instagram	Tumblr
Twitter	**2696**	2446	272
Instagram	-	**2537**	111
Tumblr	-	-	**362**

In total, we have gathered 2,785 users who form the *base user set*. Table 2 shows the breakdown of overlapping users between the three social media platforms. Twitter users form the largest group with 2,696 of them (see the first diagonal entry) in the base user set. Instagram is slightly smaller with 2,537 users. Tumblr users form the smallest user group with 362 users. There are 2,446 overlapping users between in our Twitter and Instagram data. The common users between Tumblr and the other platforms are much fewer. Not shown in Table 2, our dataset also has 22 users active on all the three platforms. Note that this dataset construction is biased towards Twitter which was conveniently used as the first social media platform to find the other linked accounts from Instagram and Tumblr. This bias should not affect our findings if the Instagram and Tumblr users without Twitter accounts have topical interests similar to those with Twitter accounts.

Table 3: Number and types of base users'posts in each social media platform

	Twitter	**Instagram**	**Tumblr**
Text	4,923,083	-	135,853
Photo	-	223,325	515,530
Video	-	-	27,015

To learn the users' topics and platform preferences, we gathered all posts generated by each user of our base user



User Caption:	*Tastes so good!*
Clarifai Tags:	*<breakfast>, <food>, <coffee>, <indoors>, <dawn> <tea>, <bread>, <delicious>*

Figure 3: Example of photo posted with caption and Clarifai generated tags

set in year 2015 using the platform-specific APIs. Table 3 shows the number and types of posts published by the base users in the three social media platforms. From Twitter, we collected nearly 5 million tweets. From Instagram, we gathered 223,325 photo images. From Tumblr, we obtained 135,853 text messages, 515,530 photo images and 27,015 videos. In total, we have 5.8 million posts from all these base set users to be used in our multi-platform topic modeling experiments.

4.1.1 Tagging Rich Media Posts

Other than tweets from Twitter and text posts from Tumblr, the photos and videos from Instagram and Tumblr rich media objects that have to be converted to text content before we can apply topic modeling on them. One possible way to extract the user annotated text associated with these photos and videos. Unfortunately, we found that about 23% of our Tumblr posts do not have user annotated text. We also found that the user-provided annotations may not accurately describe the content. In this work, we therefore relied on *Clarifai* [1], a third-party visual recognition API that is well known to accurately recognize objects and scenes in rich media, to generate word tags for the photos and videos. The generated tags will then replace the photos and videos in topic modeling. In the case of Tumblr, we thus have posts that are originally text messages as well as posts that are a bag of tags returned by Clarifai.

For example, Figure 3 shows a photo posted in Instagram with caption and the *Clarifai* generated tags. While the caption expresses the user opinion about the food in the scene, the visual recognition tool is able to better describe most if not all objects in the photo. This makes the generated tags suitable for modeling topics relevant to the photo.

4.2 Performance Evaluation

We evaluate MultiLDA model in two aspects, namely (i) the effectiveness in modeling topics in content from multiple social media platforms, and (ii) the accuracy of predicting users' platform choices as they generate posts.

[1]https://clarifai.com/

4.2.1 Experiment Setup

Baseline. We use the TwitterLDA as our baseline. While TwitterLDA is the state-of-the-art topic model for tweet posts, they can be easily adapted to "tag" posts. It is important to note that TwitterLDA model does not consider platform information associated with the posts. It assumes that all posts are from a single platform.

Training and Test Datasets. For each base user, we randomly selected 80% to 90% of posts of the user to form the training set, and use the remaining posts as the test set. We then learn the MultiLDA and TwitterLDA models using the training set, and apply the learned models on the test set.

4.2.2 Post Content Modeling

To evaluate the effectiveness of MultiLDA and TwitterLDA in modeling posts across platforms, we compute the likelihood of the training set and perplexity of the test set. The model with the higher likelihood or the lower perplexity is considered superior in the task.

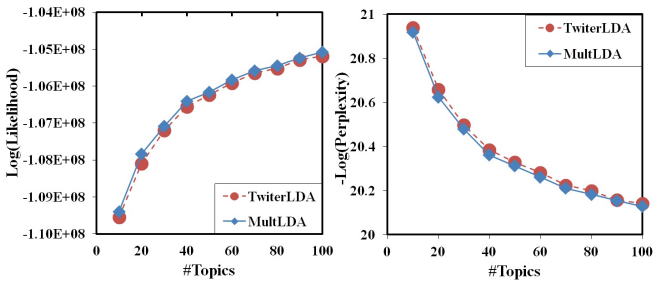


Figure 4: Log(Likelihood) and -Log(Perplexity) of MultiLDA and TwitterLDA

Figure 4 shows the likelihood and perplexity achieved by MultiLDA and TwitterLDA as we vary the number of topics K. As expected, as we use a larger number of topics, both models achieve higher likelihood and smaller perplexity. The quantum of improvement, however, reduces as K increases. We notice that the improvement reaches a plateau when K is 80 or above.

The figure also shows that MultiLDA outperforms TwitterLDA in likelihood and perplexity by a very small margin. A possible explanation is our choice of the multi-platform dataset which has relatively sufficient data generated by each user. When a user has sufficient training data from multiple platforms, TwitterLDA is able to learn the user topics quite well compared with MultiLDA. It suggests that there are not many users with strong topic-specific platform preferences for MultiLDA to yield much higher likelihood or lower perplexity than TwitterLDA.

4.2.3 Platform Choice Prediction

To evaluate the predictive power of MultiLDA and TwitterLDA, we get them to predict users' platform choices given the content of the test posts. The platform choice of a test post is predicted by MultiLDA by (i) assigning the post's topic using the trained MultiLDA, and then (ii) selecting the most probable platform for the assigned post topic where the most probable platform is determined by the user's topic-specific platform distribution.

For TwitterLDA which does not model platform choices, we generate the predicted platform choice of a given test post by (i) assigning the particular post's topic using the trained TwitterLDA, and then (ii) returning the most popular platform choice for the assigned topic according to the training set.

We use *Average F1* to measure the accuracy of platform choice prediction results. For each platform p (i.e., Twitter, Instagram, or Tumblr), we first define its precision, recall and *F1* as follows.

$$Prec_p = \frac{\#\text{ posts with } p \text{ as the correctly predicted platform}}{\#\text{ posts with } p \text{ as the predicted platform}}$$

$$Recall_p = \frac{\#\text{ posts with } p \text{ as the correctly predicted platform}}{\#\text{ posts with } p \text{ as the platform}}$$

$$F1_p = \frac{2 \cdot Prec_p \cdot Recall_p}{Prec_p + Recall_p}$$

We measure $Prec_p$, $Recall_p$ and $F1_p$ by taking average of their values over three runs of prediction each using a different randomly selected training and test sets. By taking the average over three platforms, we obtain the *Average F1* as $\frac{1}{3}\sum_p F1_p$

Experiment Results. Figure 5 shows the F1 scores of both MultiLDA and TwitterLDA for each platform and the average F1 with number of topics varying from 20 to 100. We also include a baseline which always predicts Twitter (the platform with most posts) as the platform choice. We observe that MultiLDA outperforms TwitterLDA model in every platform although the margin is small on the Twitter platform. On Instagram and Tumblr, MultiLDA significantly performs better than TwitterLDA by more than 50% and 30% respectively. The figure also shows that the prediction results do not change significantly for different number of topics. Considering all three platforms, MultiLDA improves the Avg F1 by 30% compared with TwitterLDA.

This good prediction accuracy of MultiLDA suggests that individual-level platform preferences still matter. We will further examine and discuss this in the empirical study subsession.

4.3 Platform Topics Analysis

In this section, we want to study how different the same user shares topics at different platforms. We then analyze the differences (and some similarities) of popular topics among the three social media platforms. We will also present two prediction case studies to validate the different approaches of platform choice prediction by MultiLDA and TwitterLDA. The number of topics in the MultiLDA model is set to 100 for this empirical analysis.

4.3.1 User Topic Similarity Between Platforms

For any pair of platforms p_i and p_j, we compute for each user u the Jensen-Shannon Divergence (JSD) between the u's topic distributions on p_i and p_j as follows.

$$JSD(p_i\|p_j|u) = \frac{1}{2}D(p_i\|p_j|u) + \frac{1}{2}D(p_j\|p_i|u)$$

where $D(p_i\|p_j|u)$ is the Kullback-Leibler divergence defined by:

$$D(p_i\|p_j|u) = \sum_k P(k|p_i,u) log \frac{P(k|p_i,u)}{P(k|p_j,u)}$$

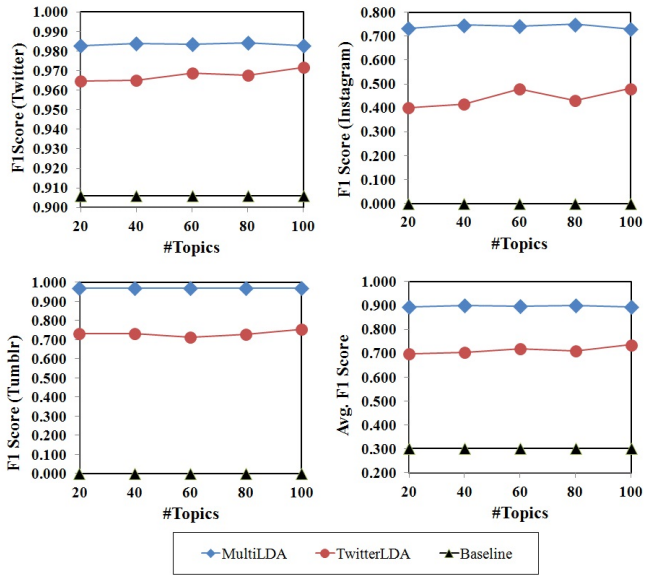


Figure 5: F1 scores for Twitter (top left), Instagram (top right), and Tumblr (bottom left) platforms, and the average F1 score of the three platform (bottom right)

where $P(k|p_i,u)$ denotes probability of a topic k when user u posts on platform p_i.

JSD measures how similar a user shares topics at two different platforms. It returns a value between 0 and 1. A JSD score of 1 means that the user has identical topic distribution on both platforms. A zero JSD score means completely different topic distributions are shared on the two platforms.

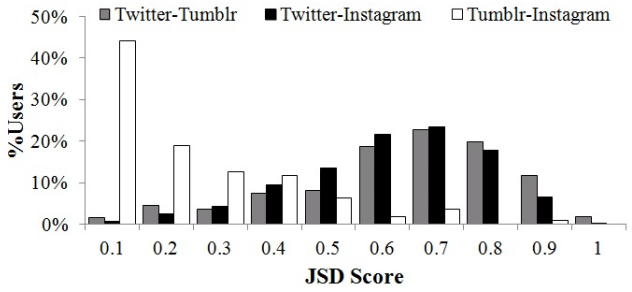


Figure 6: JSD score distributions of users for (Twitter, Instagram), (Twitter, Tumblr) and (Instagram,Tumblr)

Figure 6 depicts the JSD score distribution of users having accounts on different platform pairs. The figure shows that most users enjoy higher JSD (or higher topic distribution similarities) between Twitter and Instagram, and between Twitter and Tumblr. Even so, there are very few users with JSD more than 0.8. Among users with Instagram and Tumblr posts, most of them see much smaller topic distribution similarity. In fact, there are many of them having JSD $\leq$ 0.1.

4.3.2 Platform Specific Popular Topics

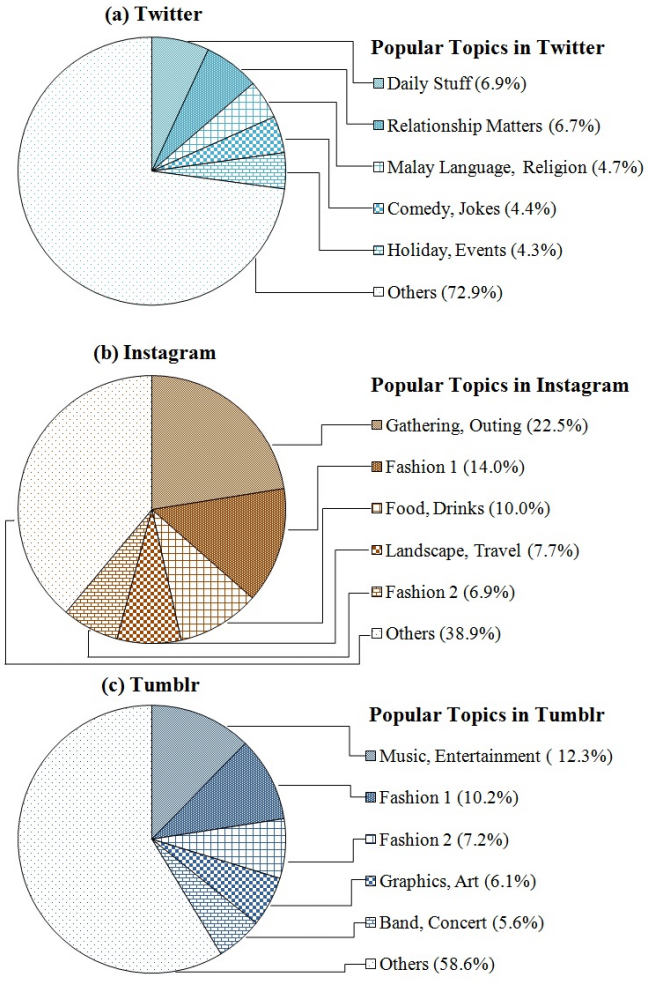


Figure 7: The proportion of top topics in (a) Twitter, (b) Instagram, and (c) Tumblr

Figure 7 shows the top five popular topics among the base users' posts in (a) Twitter, (b) Instagram and (c) Tumblr. The labels of the topics are manually assigned after examining the topics' top words. When two topics are very similar, we add numbers behind the topic labels (e.g., "*Fashion 1*", "*Fashion 2*", etc.) to distinguish them. The number in parentheses represents the topic likelihood value. For each topic, the top words are those having the highest likelihoods given the topic, and the top posts are those having the lowest perplexities given the topic.

From the charts, we notice some differences among the popular topics of the three social media platforms. In particular, the popular topics on Twitter are very different from those in Instagram and Tumblr. The popular topics in Twitter are about daily chatters while the popular topics on Instagram and Tumblr tend to be more visual (e.g., *Fashion* and *Landscape*). Instagram and Tumblr are observed to share some common popular topics (e.g., Fashion) but there are also some notable differences. For example, topics such as *Music and Entertainment* are popular on Tumblr but not in Instagram. On the other hand, topics such as *Gatherings, Food and Drinks* are popular on Instagram but not on Tumblr.

The differences in popular topics of the three social media platforms suggest that the users could be using each platform for different purposes (e.g., a user uses Twitter for news sharing but sharing posts about their pop idols in Tumblr). Another explanation could be due to the difference in the networks of friends in different platforms. Lee and Lim [13] found that most users prefer to maintain different friendships in different social media platforms while keeping only a small clique of common friends across platforms. Thus, the content shared might cater for the different audience from different social media platforms.

4.3.3 Case Study 1: Individual User Preferences

As discussed in the earlier section, the presence of individual user's platform preferences enables MultiLDA model to outperform TwitterLDA model. Among the users in our dataset, we found *User1659* who made 95 and 20 posts in Twitter and Instagram respectively. The prediction accuracies for *User1659*'s posts are 0.916 and 0.083 for MultiLDA and TwitterLDA respectively. The accuracy difference is significantly large. As we examine into the posts of *User1659*, we found that many of the user's Twitter posts fall into the *Music and Entertainment* topic which is popular on Tumblr. Hence, TwitterLDA model wrongly predicted most of *User1659*'s posts to be on Tumblr.

However, there are only a few of such cases in our dataset. The majority (87%) of the base users in our dataset have their posts predicted with more than 0.7 prediction accuracy using the TwitterLDA model.

4.3.4 Case Study 2: Advantage of Popular Topics in Platforms

Although the MultiLDA model was able to outperform the TwitterLDA model on most users' platform choice prediction, there are a few instances where TwitterLDA outperforms MultiLDA by a small margin. For example, in *User2709*'s platform choice predictions, TwitterLDA achieved a prediction accuracy of 1.0 while MultiLDA achieved a prediction accuracy of 0.875. We examine the two wrong predictions made by MultiLDA and found that the two posts are published in Tumblr and they fall into the "*Music and Entertainment*" topic. As *User2709* had not published posts of this topic on Tumblr in the training set, MultiLDA was not able to learn and predict the platform choice correctly. Conversely, TwitterLDA had predicted the platform choice correctly as "music and entertainment" is a popular topic on Tumblr.

There are very few (< 5 instances) of such exceptions in our dataset. However, this points to an interesting future work of extending MultiLDA to use a combination of global and user preferences.

5. CONCLUSION

In this paper, we proposed a novel topic model known as MultiPlatform-LDA (MultiLDA), which jointly models social media topics as well as platform preference of users. We evaluated MultiLDA using real-world datasets from three social media platforms and benchmarked against the state-of-the-art topic model. Our experiment results have shown that MultiLDA outperforms TwitterLDA in both topic modeling and platform choice prediction tasks. We have also

empirically shown that users exhibited different topical interests across platforms and the different social media platforms have different popular topics. For future works, we would like to take into consideration the type of post (i.e. text or image) in MultiLDA. We would also like to extend MultiLDA to use a combination of global and user platform preferences.

6. ACKNOWLEDGMENTS

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centres in Singapore Funding Initiative and partially supported by the ERC Grant (339233) ALEXANDRIA.

7. REFERENCES

[1] Social media site usage 2014. Tech. rep., Pew Research Center, Jan 2015.

[2] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of machine Learning research* (2003).

[3] Chang, Y., Tang, L., Inagaki, Y., and Liu, Y. What is tumblr: A statistical overview and comparison. *ACM SIGKDD Explorations Newsletter 16*, 1 (2014).

[4] Chen, Y., Zhuang, C., Cao, Q., and Hui, P. Understanding cross-site linking in online social networks. In *Proceedings of the 8th Workshop on Social Network Mining and Analysis* (2014).

[5] Cho, Y.-S., Ver Steeg, G., Ferrara, E., and Galstyan, A. Latent space model for multi-modal social data. In *WWW*.

[6] Ferrara, E., Interdonato, R., and Tagarelli, A. Online popularity and topical interests through the lens of instagram. In *HYPERTEXT* (2014).

[7] Guo, W., Wu, S., Wang, L., and Tan, T. Social-relational topic model for social networks. In *CIKM*.

[8] Hoang, T.-A., and Lim, E.-P. On joint modeling of topical communities and personal interest in microblogs. In *SOCINFO* (2014).

[9] Hoang, T.-A., and Lim, E.-P. Microblogging content propagation modeling using topic-specific behavioral factors. *IEEE Transactions on Knowledge and Data Engineering 28*, 9 (2016).

[10] Hong, L., and Davison, B. D. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics* (2010).

[11] Jang, J. Y., Han, K., Shih, P. C., and Lee, D. Generation like: Comparative characteristics in instagram. In *CHI* (2015).

[12] Kumar, S., Zafarani, R., and Liu, H. Understanding user migration patterns in social media. In *AAAI* (2011).

[13] Lee, K.-W. R., and Lim, E.-P. Friendship maintenance and prediction in multiple social networks. In *HYPERTEXT* (2016).

[14] Lim, B. H., Lu, D., Chen, T., and Kan, M.-Y. #mytweet via instagram: Exploring user behaviour across multiple social networks. In *ASONAM* (2015).

[15] Liu, J. S. The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *J. Amer. Stat. Assoc* (1994).

[16] Mehrotra, R., Sanner, S., Buntine, W., and Xie, L. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *SIGIR* (2013).

[17] Meo, P. d., Ferrara, E., Abel, F., Aroyo, L., and Houben, G.-J. Analyzing user behavior across social sharing environments. *ACM Trans. Intell. Syst. Technol. 5*, 1 (2014).

[18] Michelson, M., and Macskassy, S. A. Discovering users' topics of interest on twitter: a first look. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data* (2010).

[19] Ottoni, R., Las Casas, D. B., Pesce, J. P., Meira Jr, W., Wilson, C., Mislove, A., and Almeida, V. Of pins and tweets: Investigating how users behave across image-and text-based social networks. In *ICWSM* (2014).

[20] Pal, A., Herdagdelen, A., Chatterji, S., Taank, S., and Chakrabarti, D. Discovery of topical authorities in instagram. In *WWW* (2016).

[21] Qiu, M., Zhu, F., and Jiang, J. It is not just what we say, but how we say them: Lda-based behavior-topic model. In *SIAM SDM* (2013).

[22] Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. The author-topic model for authors and documents. In *UAI* (2004).

[23] Vosecky, J., Hong, D., and Shen, V. Y. User identification across multiple social networks. In *2009 First International Conference on Networked Digital Technologies* (2009).

[24] Xu, J., Compton, R., Lu, T.-C., and Allen, D. Rolling through tumblr: characterizing behavioral patterns of the microblogging platform. In *WebSci* (2014).

[25] Zafarani, R., and Liu, H. Connecting users across social media sites: a behavioral-modeling approach. In *SIGKDD* (2013).

[26] Zafarani, R., and Liu, H. Users joining multiple sites: Distributions and patterns. In *ICWSM* (2014).

[27] Zhang, H., Kan, M.-Y., Liu, Y., and Ma, S. Online social network profile linkage. In *Asia Information Retrieval Symposium* (2014).

[28] Zhang, Y., and Pennacchiotti, M. Predicting purchase behaviors from social media. In *WWW* (2013).

[29] Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. Comparing twitter and traditional media using topic models. In *ECIR* (2011).

[30] Zhao, W. X., Li, S., He, Y., Chang, E. Y., Wen, J.-R., and Li, X. Connecting social media to e-commerce: Cold-start product recommendation using microblogging information. *IEEE Transactions on Knowledge and Data Engineering 28*, 5 (2016).