7-2014

Lifetime lexical variation in social media

Lizi LIAO
Singapore Management University, lzliao@smu.edu.sg

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

Ying DING
Singapore Management University, ying.ding.2011@smu.edu.sg

Heyan HUANG
Beijing Institute of Technology

Ee Peng LIM
Singapore Management University, eplim@smu.edu.sg

Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence

Lifetime Lexical Variation in Social Media

Lizi Liao
School of Computer Science
Beijing Institute of Technology
liaolizi.llz@gmail.com

Jing Jiang
School of Information System
Singapore Management University
jingjiang@smu.edu.sg

Ying Ding
School of Information System
Singapore Management University
ying.ding.2011@phdis.smu.edu.sg

Heyan Huang*
School of Computer Science
Beijing Institute of Technology
hhy63@bit.edu.cn

Ee-Peng Lim
School of Information System
Singapore Management University
eplim@smu.edu.sg

Abstract

As the rapid growth of online social media attracts a large number of Internet users, the large volume of content generated by these users also provides us with an opportunity to study the lexical variation of people of different ages. In this paper, we present a latent variable model that jointly models the lexical content of tweets and Twitter users' ages. Our model inherently assumes that a topic has not only a word distribution but also an age distribution. We propose a Gibbs-EM algorithm to perform inference on our model. Empirical evaluation shows that our model can learn meaningful age-specific topics such as "school" for teenagers and "health" for older people. Our model can also be used for age prediction and performs better than a number of baseline methods.

Introduction

With the rapid growth of user-generated content in social media, there has been a tremendous amount of work on content analysis on social media. In particular, Twitter, arguably the most popular microblog site, has attracted much attention in the research community. Content analysis on Twitter ranges from search (Liang, Qiang, and Yang 2012), recommendation (Phelan, McCarthy, and Smyth 2009) to topic discovery (Ramage, Dumais, and Liebling 2010) and event detection (Sakaki, Okazaki, and Matsuo 2010). In particular, the goal of topic analysis is to discover the major topics exhibited in a large corpus, which can be used to better understand and summarize the corpus. The discovered topics can also be used to assist other tasks such as classification and recommendation. To discover topics from a Twitter corpus, models such as standard LDA, Author-Topic Model, Labeled LDA and Twitter-LDA have been used (Hong and Davison 2010; Ramage, Dumais, and Liebling 2010; Zhao et al. 2011). In addition to only the words contained in tweets, researchers have also exploited other types of data to study special kinds of topics. For example, Eisenstein et al. (2010) combined Twitter content with users' location information to infer regional variants of base topics. Diao et al. (2012) designed an LDA model that takes timestamps of tweets into consideration and discovers "bursty" topics. These models assume that the content published by a user may be influenced by factors such as location and time.

We observe that another important factor that heavily influences a social media user's published content is her age. For example, using standard LDA, we find that a popular topic in our Twitter data set is related to "homework," "school," "exam," etc. Obviously one may suspect that this topic is frequently discussed by school children. However, to the best of our knowledge, currently there is no principled model that can help verify this hypothesis and formally characterize this kind of topics that have a strong age association. Assuming that we are able to obtain the age information of a large set of users, a naive way to solve the problem above is to first divide the users into age groups and then perform standard topic modeling within each age group to learn age-specific topics. However, this naive solution suffers from several shortcomings: (1) The choice of the boundaries between age groups are arbitrary. (2) The topics learned within different age groups are independent and cannot be easily linked together. (3) The naive solution cannot differentiate between topics that are age-insensitive, i.e. topics that are general to all age groups, versus topics that are more age-specific.

In this paper, we design a novel, principled latent variable model that naturally links content words with users' ages. Our model is based on standard LDA and Gaussian mixture models. We assume that each topic has not only a word distribution but also an age distribution. To perform inference on our model, we design a Gibbs-EM algorithm. Using the tweets of a set of users whose ages are known, we are able to infer a wide range of topics over different age groups. We show both qualitatively and quantitatively that our model learns meaningful age-specific topics. A direct application of our model is age prediction. We show that our model can effectively perform age prediction better than some baseline

*The corresponding author.

methods.

Our contributions can be summarized as follows: (1) We propose a principled probabilistic model to link content with users' ages. (2) We design an efficient Gibbs-EM algorithm to perform inference on our model. (3) Using a large set of Twitter users and their tweets, we show empirically that our model is able to discover meaningful topics with clear age associations.

Related Work

While designing latent variable models to jointly model content words together with other types of data is not new, to the best of our knowledge, jointly modeling text and age information is new, partly because traditional documents do not contain authors' age. We review a few lines of related work below.

Basic Topic Models on Twitter

There have been many studies which model the topics on Twitter using latent variable models. Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) is a general model for finding latent topics from any document collection. Hong and Davison (2010) empirically evaluated the performance of LDA on Twitter. They found that by aggregating all tweets published by the same user into a single document the learned topics had higher quality. They also applied an extended version of the Author-Topic Model (Steyvers et al. 2004) to Twitter and found that it did not perform well. Labeled LDA is another extension of LDA which assumes that each document has a set of known labels (Ramage et al. 2009). By treating metadata such as hashtags on Twitter as labels, Ramage, Dumais, and Liebling (2010) characterized the topics on Twitter into four categories: substance, status, style and social. Zhao et al. (2011) proposed a Twitter-LDA model, which assigns a single topic to an entire tweet, and found the model to produce more meaningful topics than standard LDA. All these studies discover the traditional kind of topics, which are simply word distributions. In comparison, our learned topics also have age distributions.

Augmented Topic Models

There have also been a number of topic models on Twitter that jointly model text and other metadata. Besides word distributions, these topics are often augmented with some other characteristics. Eisenstein et al. (2010) proposed a geographic topic model that incorporates the locations of users. Each base topic has regional variants, whose word distributions deviate from that of the base topic. With the observed geotags of tweets, the model is able to learn these regional variants of base topics and draw insights into geographic lexical variations. Diao et al. (2012) proposed a latent variable model on Twitter which considers timestamps of tweets. Tweets published on the same day are assumed to be more likely to be about the same topic. The learned topics thus have a "bursty" pattern over time. Qiu, Zhu, and Jiang (2013) proposed a behavior-topic model that considers both the content and the type of a tweet. Replies, at-mentions and retweets are the special types of tweets considered. A topic has not only a distribution over words but also a distribution over tweet types. With this model, they can characterize different users who have similar topic interests but tend to publish different types of tweets.

While our model is also an augmented topic model on Twitter, we model users' ages, which is another type of metadata that has not been studied in existing topic models for Twitter. We use Gaussian mixture models to generate users' ages, which is very different from the aforementioned models.

There have also been many general topic models that combine text with other types of data but are not specifically designed for Twitter. Examples include correspondence-LDA (Blei and Jordan 2003) for annotated data, combining LDA with probabilistic matrix factorization for recommendation (Wang and Blei 2011), etc. In particular, our model bears similarity to the supervised LDA model proposed by Blei and McAuliffe (2007). In supervised LDA, each document has associated with it a numerical response variable, which is stochastically generated from the topic assignment of the words in the document. In our model, the age of a user is also stochastically generated from the topic assignment of the words. However, the supervised LDA model uses a set of coefficients to transform the topic distribution of a document into a single numerical value, which serves as the Gaussian mean to generate the response variable, whereas in our model each topic has its own Gaussian mean and variance.

Age Prediction Using User-Generated Content

There have long been interests in how various morphological, phonological and stylistic aspects of language may change over time as a person grows old (Fischer 1958; Labov 1972). With the availability of the large amount of user-generated content, recently there has been much work trying to associate the content and writing styles with user's ages (Schler et al. 2006; Argamon et al. 2007; Nguyen et al. 2013). The goal is often to predict an user's age based on her published content. Typically a supervised classification approach is taken and various useful linguistic features are identified and evaluated. Nguyen et al. (2013) used only unigram features to perform age prediction on Twitter and achieved good performance on their data set. Rao et al. (2010) used stacked-SVM and more complex features such as socio-linguistic features and n-gram features to predict a number of user attributes including gender, age and regional origin.

While in our experiments, we also use age prediction as a task to evaluate the effectiveness of our model, our model is not designed solely for age prediction. The topics learned by our model offer insights into the lifestyles of people of different age groups and shed light on how we can better understand and explore Twitter data.

Method

Model

We design the following model to combine topics with users' ages. We assume that there is a set of U users whose

ages are known. For each user there is a set of tweets. Let a_u, a positive integer, denote the age of the u-th user. Let $w_{u,n}$, an index between 1 and V, denote the n-th word in the tweets published by the u-th user, where V is the vocabulary size. Note that we essentially treat all the tweets of the same user as a bag of words. Table 1 shows the notation and descriptions of our model parameters.

Notation	Description
U	the total number of users
N_u	the total number of words of user u
T	the total number of topics
V	the total number of unique words
a_u	the age of user u
$\boldsymbol{\theta}_u$	user specific topic distribution
$\boldsymbol{\phi}_t$	topic specific word distribution
μ_t	mean of age for topic t
σ_t	standard deviation of age for topic t
α, β	Dirichlet priors

Table 1: Notation and descriptions.

Our model makes the following assumptions. There exist T topics that explain all the tweets published by all the users. Like in standard LDA, each topic has a word distribution $\boldsymbol{\phi}_t$. In addition, each topic has an age mean μ_t and an age variance σ_t^2. We will explain later how μ_t and σ_t^2 are related to the users' ages. We also assume that each user has a topic distribution $\boldsymbol{\theta}_u$. We assume that $\boldsymbol{\theta}_u$ has a uniform Dirichlet prior parameterized by α and $\boldsymbol{\phi}_t$ has a uniform Dirichlet prior parameterized by β.

We introduce two kinds of hidden variables in our model. The first kind of hidden variables is similar to the one in standard LDA: for the n-th word published by the u-th user, hidden variable $z_{u,n}$ is the topic for that word. The second kind of hidden variables helps us associate a user's age with topics: for the u-th user, hidden variable y_u is a topic chosen uniformly from $z_{u,1}, z_{u,2}, \ldots, z_{u,N_u}$. Formally, let $\boldsymbol{\pi}_u$ denote a multinomial distribution determined by the hidden variables $z_{u,1}, z_{u,2}, \ldots, z_{u,N_u}$ in the following way:

$$\pi_{u,t} \quad = \quad \frac{\sum_{n=1}^{N_u} \delta(z_{u,n}, t)}{N_u},$$

where $\delta(t, t')$ is 1 if t and t' are the same and 0 otherwise. In other words, $\pi_{u,t}$ is the fraction of user u's words that are assigned to topic t. Then y_u is randomly sampled from $\boldsymbol{\pi}_u$. The topic y_u will be used to generate the user's age.

We assume the following generative process of the data:

- For each user, draw $\boldsymbol{\theta}_u \sim \text{Dirichlet}(\alpha)$.
- For each topic, draw $\boldsymbol{\phi}_t \sim \text{Dirichlet}(\beta)$.
- For each word, draw $z_{u,n} \sim \text{Discrete}(\boldsymbol{\theta}_u)$, and then draw $w_{u,n} \sim \text{Discrete}(\boldsymbol{\phi}_{z_{u,n}})$.
- For each user, set $\boldsymbol{\pi}_u$ as defined above. Draw $y_u \sim \text{Discrete}(\boldsymbol{\pi}_u)$, and then draw $a_u \sim \text{Gaussian}(\mu_{y_u}, \sigma_{y_u}^2)$.

The model is also depicted in Figure 1. Basically we assume that each topic is associated with a Gaussian distribution parameterized by μ_t and σ_t^2. The age of an user is generated by a mixture of these Gaussian distributions weighted by the hidden topic assignment of the user's words.

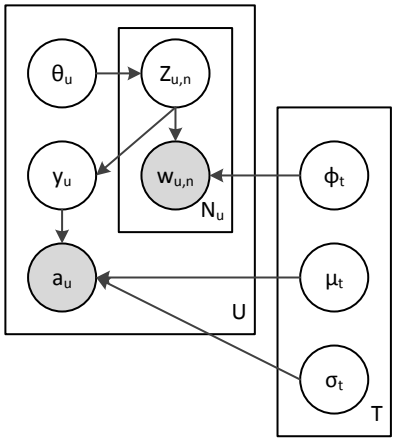

Figure 1: The plate notation of our model. The hyperparameters are omitted. Note that the variables $\boldsymbol{\pi}_u$ are not shown because they are deterministically set from $z_{u,n}$.

Inference

Given a set of observed data, our goal is to find the best parameters $\boldsymbol{\theta}_u$ for each user and $\boldsymbol{\phi}_t$, μ_t and σ_t for each topic. Given that our model has hidden variables, it is easier to use the EM algorithm to solve the optimization problem. Here we develop a Gibbs-EM algorithm to perform the inference.

First of all, let us use $\boldsymbol{w}$ to denote all the observed words $w_{u,n}$ and $\boldsymbol{a}$ to denote all the observed ages a_u. Similarly, $\boldsymbol{z}$ denotes all $z_{u,n}$ and $\boldsymbol{y}$ denotes all y_u. $\boldsymbol{\theta}$ denotes all $\boldsymbol{\theta}_u$, $\boldsymbol{\phi}$ denotes all $\boldsymbol{\phi}_t$, $\boldsymbol{\mu}$ denotes all μ_t and $\boldsymbol{\sigma}$ denotes all σ_t.

Specifically, we use Gibbs-EM to solve the following optimization problem:

$$\boldsymbol{\mu}^*, \boldsymbol{\sigma}^* \quad = \quad \arg\max_{\boldsymbol{\mu}, \boldsymbol{\sigma}} p(\boldsymbol{w}, \boldsymbol{a} | \boldsymbol{\mu}, \boldsymbol{\sigma}, \alpha, \beta).$$

Essentially this is to maximize the likelihood of the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ given the observed data and hyperparameters. Here we treat $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ as hidden variables that have been integrated out. After we find $\boldsymbol{\mu}^*$ and $\boldsymbol{\sigma}^*$, we collect samples of $(\boldsymbol{y}, \boldsymbol{z})$ from $p(\boldsymbol{y}, \boldsymbol{z} | \boldsymbol{w}, \boldsymbol{a}, \boldsymbol{\mu}^*, \boldsymbol{\sigma}^*, \alpha, \beta)$ and use these samples to estimate $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, just like in collapsed Gibbs sampling for standard LDA.

We now describe our Gibbs-EM algorithm in more detail. The algorithm runs iteratively. In the $(k+1)$-th iteration, during the E-step, we use collapsed Gibbs sampling to collect samples of $(\boldsymbol{y}, \boldsymbol{z})$ from the distribution $p(\boldsymbol{y}, \boldsymbol{z} | \boldsymbol{w}, \boldsymbol{a}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\sigma}^{(k)}, \alpha, \beta)$, where $(\boldsymbol{\mu}^{(k)}, \boldsymbol{\sigma}^{(k)})$ are parameters estimated from the M-step in the k-th iteration. Let us use $\mathcal{S}^{(k+1)}$ to denote these samples. Then during the M-step, using $\mathcal{S}^{(k+1)}$, we find $(\boldsymbol{\mu}^{(k+1)}, \boldsymbol{\sigma}^{(k+1)})$ that maximize the following objective function:

$$\begin{aligned} & \boldsymbol{\mu}^{(k+1)}, \boldsymbol{\sigma}^{(k+1)} \\ = \quad & \arg\max_{\boldsymbol{\mu}, \boldsymbol{\sigma}} \sum_{(\boldsymbol{y}, \boldsymbol{z}) \in \mathcal{S}^{(k+1)}} \ln p(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{a} | \boldsymbol{\mu}, \boldsymbol{\sigma}, \alpha, \beta). \end{aligned}$$

Due to the space limit, we leave out the derivation details and show the formulas we use in the E-step and the M-step.

E-step To re-sample the value of $z_{u,n}$ given $\boldsymbol{z}_{\neg(u,n)}$ and $\boldsymbol{y}$, the formula is as follows:

$$p(z_{u,n} = t|\boldsymbol{y}, \boldsymbol{z}_{\neg(u,n)}, \boldsymbol{w}, \boldsymbol{a}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\sigma}^{(k)}, \alpha, \beta) \propto \frac{C_{u,t} + \alpha}{C_{u,\cdot} + T\alpha} \cdot \frac{C_{u,t} + \delta(y_u, t)}{C_{u,\cdot}} \cdot \frac{C_{t,w_{u,n}} + \beta}{C_{t,\cdot} + V\beta}. \quad (1)$$

To re-sample the value of y_u given $\boldsymbol{y}_{\neg u}$ and $\boldsymbol{z}$, the formula is as follows:

$$p(y_u = t|\boldsymbol{y}_{\neg u}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{a}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\sigma}^{(k)}, \alpha, \beta) \propto \frac{C_{u,t}}{C_{u,\cdot}} \cdot \frac{1}{\sigma_t^{(k)}\sqrt{2\pi}} \exp\Big(-\frac{(a_u - \mu_t^{(k)})^2}{2(\sigma_t^{(k)})^2}\Big). \quad (2)$$

Here all the C variables denote the various counters, where the current variable that is being sampled is excluded. For example, in Eqn. (1), $C_{u,t}$ is the number of words published by user u that have been assigned to topic t, excluding the word $w_{u,n}$.

M-step The formulas to estimate $\boldsymbol{\mu}^{(k+1)}$ and $\boldsymbol{\sigma}^{(k+1)}$ are as follows:

$$\mu_t^{(k+1)} = \frac{\sum_{u=1}^{U} x_{u,t} a_u}{\sum_{u=1}^{U} x_{u,t}},$$

$$\sigma_t^{(k+1)} = \sqrt{\frac{\sum_{u=1}^{U} x_{u,t}(a_u - \mu_t^{(k+1)})^2}{\sum_{u=1}^{U} x_{u,t}}},$$

where

$$x_{u,t} = \frac{\sum_{(\boldsymbol{y},\boldsymbol{z}) \in \mathcal{S}^{(k+1)}} \delta(y_u, t)}{|\mathcal{S}^{(k+1)}|}.$$

Experiments

This section presents the empirical evaluation of our model. We first describe our data set, including how we obtain the ground truth of users' ages. We show the effectiveness of our model by demonstrating that the learned topics are meaningful. We then use the task of age prediction to quantitatively evaluate our model, comparing it with a few baseline methods that represent the state-of-the-art for age prediction using only unigram features.

Data Set and Ground Truth

Our experiments are based on Twitter. We used the following strategy to crawl Twitter users. Starting from a set of 59 popular seed users in Singapore, we first crawled these users' direct followers and followees and then crawled their followers/followees' followers and followees, i.e. we crawled all users who are either one or two hop(s) away from the seed users. In this way, we obtained 2,891,761 users. We then deleted users with more than 2000 followees as these are unlikely to be real people. Among the remaining users, we were able to obtain the age information from user profiles for 21,831 users, accounting for 0.75% of the total users. This shows that age information in Twitter is extremely sparse. After we got these users' IDs and age information, we used Twitter's public timeline API to crawl these users' latest 200 tweets. Those users who have less than 200 tweets were deleted. Finally, we got 16,017 users' tweets and age information. To give an overview of our data, we plot the age distribution of our data in Figure 2.

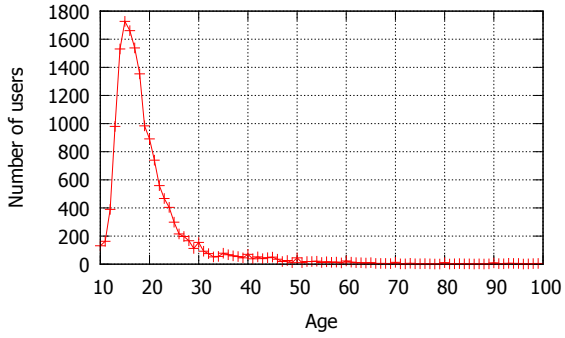

Figure 2: Age distribution of our data set.

As we can see from the figure, this data set is very unbalanced, with most users between 10 and 30. While this shows the real distribution of users, our preliminary experiments show that it is not easy to discover topics more specific to people above 30 using this unbalanced data set. Thus, we randomly selected up to 30 users of each age to form our data set, which finally consists of 1564 users. In our age prediction experiments, we randomly selected 150 users from the 1564 users as our test data.

In our experiments, we set α to 0.25 and β to 0.2. We empirically choose 200 topics. We run 32 iterations of Gibbs-EM, where during each iteration in the E-step we run 400 iterations of Gibbs sampling.

Qualitative Evaluation

To give a clear view of the topics learned by our model, we first divide the learned topics into four groups based on the learned age mean values. The four age groups are teenagers ("10s"), people in their 20s, in their 30s, and people above 40 ("40s+"). Then within each age group, we rank topics in increasing order of the age variance. The top-6 topics for each group are shown in Table 2.

We can see from the table that within each age group the topics generally make sense. For example, for both teenagers and people in their 20s, we see many Internet slang words, but these words are less used by the older age groups. For teenagers, we see a topic on school/class/exam and another topic on the pop musician Justin Bieber, which are clearly relevant to this age group. For people in their 20s, we also observe interesting topics like the fourth one and the sixth one. The fourth one is related to acne treatment and skin care, while the sixth one is related to nba games. Those are also relevant to this age group. In comparison, for people in their 40s and above, there is a topic related to arthritis/pain/disease, showing that people around this age start to have health problems.

We also show the topics with the largest age variance at the bottom of Table 2. These supposedly are topics generally popular among all age groups. The first topic is about

Age Group	Mean	Standard Deviation	Top Words
10s	15.82	3.86	fuck, fucking, shit, bitch, hate, ass, gonna, hell, damn, man
	16.47	5.39	na, sa, ko, ang, mo, ng, ako, ka, lang, pa
	17.14	6.65	sexy, sex, hot, ass, mb, big, pics, cam, teen, pussy
	17.33	7.33	fan, day, idol, biggest, vip, line, number, greyson, cross, thousand
	17.72	7.36	school, class, tomorrow, exam, college, study, homework, year, test, studying
	19.02	7.62	justin, bieber, beliebers, love, belieber, fans, back, selena, world, justin's
20s	27.76	8.72	ng, jackie, cover, nh, ch, band, tr, click, version, kh
	27.36	9.07	ho, facebook, foto, su, una, ja, ei, guam, today, se
	21.83	9.48	nigga, beats, musik, da, bruh, yo, niggas, dis, tht, aint
	29.95	9.71	acne, treatment, popular, products, oz, natural, pimples, reviews, remedies, skin
	28.36	9.99	quotes, theme, update, images, item, themes, english, fix, issue, core
	23.79	10.09	video, favorited, heat, nba, game, bulls, vs, lebron, nbadraft, playoffs
30s	31.62	8.91	tv, watch, online, august, time, face, mo, ordinary, millionaire, extraordinary
	30.48	9.12	une, vid, jai, playlist, aim, km, nike, viens, gps, mv
	30.47	9.90	website, contact, business, designs, domains, week, replies, sale, official, sorted
	31.99	10.03	travel, tours, hotels, india, tour, star, china, asia, hotel, group
	38.04	10.33	keyword, complete, tools, program, software, download, fitness, install, cookies, package
	34.69	10.78	goverment, healthcare, chinese, ge, employee, china, charged, world, justice, employ
40s+	40.22	8.01	surf, traffic, cash, sites, exchange, join, surfing, jackpot, promoting, amp
	42.02	11.36	internet, business, start, starting, year, biz, average, real, wanted, users
	40.07	12.62	miles, walking, km, fine, steps, traveled, began, workout, doggy, lost
	47.05	12.76	book, abuse, abused, lives, power, poetry, women, battered, pointers, love
	40.10	13.26	arthritis, walk, pain, fiona, florida, disease, intended, forum, honored, juvenile
	46.01	13.64	rest, reps, rounds, box, squats, minutes, lb, jumps, ball, plank
general topics	26.78	24.92	photo, posted, facebook, photos, album, photoset, holder, training, enter, park
	28.32	24.41	ni, nak, tak, aku, dah, je, la, tu, nk, lah

Table 2: Top-topics with the lowest/highest age variances.

$\mu = 25.60$ $\sigma = 15.68$	$\mu = 39.08$ $\sigma = 11.12$	$\mu = 40.94$ $\sigma = 16.86$	$\mu = 48.90$ $\sigma = 18.22$
twilight	lottery	weight	insurance
girl	winning	health	meds
kristen	law	loss	health
dawn	attraction	fat	supplies
robert	orlando	diet	medicine
breaking	win	healthy	medications
edward	draw	excercise	plan
movie	tonights	body	pharmacy
pattinson	add	fitnes	allergiess
moon	tonight	tips	breakdown

Table 3: Some other representative sample topics.

posting photos online. The second topic contains some commonly used Malay words, because Singapore has a large Malay population.

Besides the topics shown in Table 2, there are also some other interesting topics with meaningful age association. Due to the limit of space, we only show four representative ones in Table 3. Recall that μ represents the age mean of a topic and σ is the standard deviation of age. For example, the third topic is about weight control, which has a mean age of around 41.

Quantitative Evaluation

In this section, we quantitatively evaluate our model. We use the task of age prediction to perform the evaluation. The task is defined as predicting the age of a test user based on the test user's tweets. For training, a set of users with their tweets and age information are used. As mentioned earlier, around 10% of the users are used for testing and the rest are used for training.

Baselines We consider the following representative baseline methods, which all use only unigram features from the tweets. This ensures fair comparison with our model as our model also relies on unigrams.

- **Age Topic Model (ATM)** This is the model we proposed. During the training stage, we obtain the age distributions as well as the word distributions per topic. During testing, for a given test user, we first assign topic labels to each word in her tweet and then apply the learned topic specific age distributions to estimate her age.
- **Supervised Latent Dirichlet Allocation (sLDA)** As we have discussed in related work, our model bears some similarity to the supervised LDA model. Here we combine all tweets of a user into a single document and treat the age of the user as the observed response variable. We can then directly apply sLDA to the data to perform age prediction.
- **Support Vector Regression (SVR)** Another baseline method is to simply treat each unigram word as a fea-

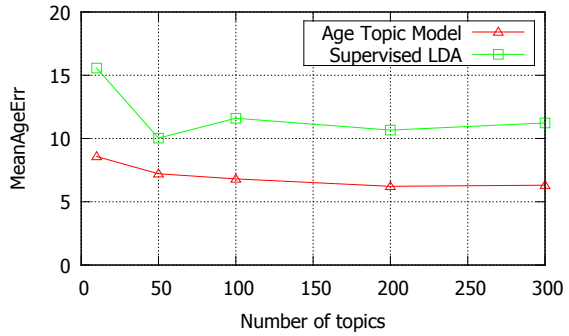

Figure 3: The effect of various topic numbers on mean age error.

ture and directly learn a model from the feature vectors of users. As shown by Nguyen et al. (2013), unigram features can achieve decent performance on their data set. Specifically, we first construct a feature vector for each user by combining all her tweets and using TF-IDF term weighting to weigh the features. We then apply support vector regression (Fan et al. 2008) to learn a linear model from the training data. This model is then used for age prediction.

Evaluation Metrics We introduce the following metrics to help us evaluate the performance of age prediction task. We compare the predicted age of a user versus her actual age. The first metric we consider is **AgeErr**, which quantifies the error gap in years between the actual age of the user a_u and the predicted age $\hat{a}_u$. The **AgeErr** for user u is defined as:

$$\text{AgeErr}(u) = |a_u - \hat{a}_u|. \quad (3)$$

In order to give a clear insight into the distribution of age prediction errors, the next metric **MeanAgeErr** used is the mean absolute age errors over a set of U_{test} test users, and the metric **Accuracy** considers the percentage of users with their **AgeErr** capped within d years:

$$\text{MeanAgeErr} = \frac{\sum_u \text{AgeErr}(u)}{U_{\text{test}}}, \quad (4)$$

$$\text{Accuracy}(d) = \frac{|\{u : \text{AgeErr}(u) \leqslant d\}|}{U_{\text{test}}}. \quad (5)$$

Performance Analysis To show how performance changes as the number of topics varies, we plot the results of the ATM model and the sLDA model in Figure 3. As the SVR method does not have topics, it is not shown in the figure. We can see that as the topic number increases from 10 to 50, the performance of our ATM model and the sLDA model both increases. When the topic number further increases, the ATM model achieves better results until the topic number reaches 200. The ATM model gets the best results under the setting of $T = 200$, while the sLDA model achieves the best results under the setting of $T = 50$. In the rest of this section, we use the results of the ATM and the sLDA models under these optimal settings.

d	0	2	5	10	20
ATM	0.107	**0.273**	**0.607**	**0.807**	**0.973**
sLDA	**0.140**	0.207	0.507	0.607	0.873
SVR	0.007	0.127	0.193	0.271	0.680
Mean	0.007	0.067	0.233	0.367	0.867

Table 4: Accuracy for all methods.

We then present the results in terms of accuracy of our model and the baselines in Table 4. The method Mean refers to predicting users' age as the mean age of all training users. As shown in Table 4, our ATM model achieves the best performance. The performance of SVR is very poor. In our preliminary experiments, we found that SVR could perform much better when more training data was used. However, when less training data is available, its performance drops substantially, leading to non-competitive results.

Comparing ATM with sLDA, we can see that although when the threshold d is zero, the sLDA model performs slightly better than our ATM model, the ATM model performs much better than the sLDA model when we loosen the threshold. In the sLDA model, each word is first associated with a latent topic and then multiplied by a learned regression coefficient. They accumulate to influence the mean of the Gaussian distribution. With learned variance, age is drawn from this Gaussian distribution. In this way, supervised LDA suits the task of predicting exact age well. However, our age topic model focuses on discovering the age-specific topics. In age topic model, each topic is assumed to have a Gaussian distribution over age. The user in this model is treated as a mixture of various topics, thus her age is reflected in a Gaussian mixture distribution. As mentioned above, this model is capable of uncovering meaningful and coherent topics and revealing the association between topics and age. Thus, it achieves much better performance when we loosen the age threshold.

Conclusion

The access to huge amount of user generated data enables us to investigate lifetime linguistic variation of people. This paper presents an age topic model that jointly models latent topics and user's ages. The core premise of the model is that age influences the topic composition of a user, and each topic has a unique age distribution. Content and age is thus combined to shape the observed lexical variations. Our model uncovers coherent topics and their age distributions, offering insights into the lifestyles of people of different age groups. Experiments show that the model outperforms strong alternatives in an age prediction task.

We see this work as a step towards leveraging generative models to jointly model user contents and their personal attributes. In a general sense, our model may be used to model other user attributes with numerical values. We hope to explore this possibility in our future work.

Acknowledgments

This research is supported in part by the PhD Joint Education grants from Beijing Institute of Technology, Chinese

National Program on Key Basic Research Project (Grant No. 2013CB329605). This research is also partially supported by the Singapore National Research Foundation under its International Research Centre@Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority (MDA).

References

Argamon, S.; Koppel, M.; Pennebaker, J. W.; and Schler, J. 2007. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday* 12(9).

Blei, D. M., and Jordan, M. I. 2003. Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, 127–134.

Blei, D. M., and McAuliffe, J. D. 2007. Supervised topic models. In *NIPS*, volume 7, 121–128.

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

Diao, Q.; Jiang, J.; Zhu, F.; and Lim, E.-P. 2012. Finding bursty topics from microblogs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 536–544.

Eisenstein, J.; O'Connor, B.; Smith, N. A.; and Xing, E. P. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1277–1287.

Fan, R.-E.; Chang, K.-W.; Hsieh, C.-J.; Wang, X.-R.; and Lin, C.-J. 2008. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.* 9:1871–1874.

Fischer, J. L. 1958. Social influences on the choice of a linguistic variant. *Word* 14:47–56.

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

Labov, W. 1972. *Sociolinguistic patterns*. Number 4. University of Pennsylvania Press.

Liang, F.; Qiang, R.; and Yang, J. 2012. Exploiting real-time information retrieval in the microblogosphere. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, 267–276.

Nguyen, D.; Gravel, R.; Trieschnigg, D.; and Meder, T. 2013. "how old do you think i am?": A study of language and age in Twitter. In *Seventh International AAAI Conference on Weblogs and Social Media*.

Phelan, O.; McCarthy, K.; and Smyth, B. 2009. Using Twitter to recommend real-time topical news. In *Proceedings of the Third ACM Conference on Recommender Systems*, 385–388.

Qiu, M.; Zhu, F.; and Jiang, J. 2013. It is not just what we say, but how we say them: LDA-based behavior-topic model. In *Proceedings of the 13th SIAM International Conference on Data Mining*, 794–802.

Ramage, D.; Hall, D.; Nallapati, R.; and Manning, C. D. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 248–256.

Ramage, D.; Dumais, S.; and Liebling, D. 2010. Characterizing microblogs with topic models. In *Proceedings of the Fourth International Conference on Weblogs and Social Media*.

Rao, D.; Yarowsky, D.; Shreevats, A.; and Gupta, M. 2010. Classifying latent user attributes in Twitter. In *Proceedings of the Second International Workshop on Search and Mining User-generated Contents*, 37–44.

Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2010. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, 851–860.

Schler, J.; Koppel, M.; Argamon, S.; and Pennebaker, J. W. 2006. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 199–205.

Steyvers, M.; Smyth, P.; Rosen-Zvi, M.; and Griffiths, T. 2004. Probabilistic author-topic models for information discovery. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 306–315.

Wang, C., and Blei, D. M. 2011. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 448–456.

Zhao, W. X.; Jiang, J.; Weng, J.; He, J.; Lim, E.-P.; Yan, H.; and Li, X. 2011. Comparing Twitter and traditional media using topic models. In *Proceedings of the 33rd European Conference on Information Retrieval*, 338–349.