User-Sentiment Topic Model: Refining User’s Topics with Sentiment Information

Tong Zhao*, Chunping Li*, Qiang Ding‡, Li Li‡
*School of Software, Tsinghua University, Beijing 100084, China
‡Shannon Lab, Huawei Technologies Co. LTD, Beijing 100095, China
zt882001@hotmail.com, cli@tsinghua.edu.cn, q.ding@huawei.com, jollylili.li@huawei.com

ABSTRACT
In large social networks, users feel free to share their feelings about anything they are interested in and many research works have focused on modeling users’ interests on social network for product recommendations or personal services. Unfortunately, there are fewer works about finding why users like or dislike something. More specifically, there are many researches about sentiment analysis of users’ opinion toward products or topics, but fewer are focused on why they hold this feeling and which aspects or factors of the product (topic) lead to users’ different opinions about it. In this paper, we present a hierarchical generative model, called user-sentiment topic model (USTM), which captures users’ topics with sentiment information. Our aim is to use USTM to refine users’ topics with different sentiment trends including positive, negative and neutral, which can be further used in social network analysis to find influential users on topic level with sentiment information. The experiment results on three datasets show that our proposed USTM can capture user’s interests with their sentiment well, making it useful for social network analysis.

Categories and Subject Descriptors
H.3.1 [Content Analysis and Indexing]: Indexing methods

General Terms
Algorithms, Experimentation

Keywords
Topic Model, Sentiment Detection, User-Sentiment Topic Model.

1. INTRODUCTION
Since the number of Web sourced data now exceeds the capability of any manual work, the automatic extraction technology of useful information from text has become an increasingly important research area. In particular, extracting user’s interests with the content of document simultaneously has received considerable attention in recent years. This has led to the proposal of various topic models that can infer latent topics with author’s interests automatically, where each topic is a latent variable that has a probability distribution over words. These topic models introduce author’s interests as topic distributions according to their document contents [2,3,4]. These proposed topic models aim at using topics or sets of topics to represent author’s interests and provide useful description for the generative process of various data, which could be applied in different aspects such as social network analysis [5,11,13], expertise recommendation [4] and collaborative filtering [12], etc.

However, most topic models combined with author’s information have not considered author’s sentiment with his interested topics. The previous topic models view an author as an observed variable and use this variable to infer topic distribution without considering their sentiment trend on the topic. While as the advent of social network and social media analysis, the attention of topic model research has focused more on what the user’s opinion holds in these interests [6,7,8,9]. Some researchers have made attempts toward efficient opinion detection, opinion extraction from web data and have proposed methods to capture the sentiment from documents [6,15,16]. However, these works could not study well on user’s interest level but just in documents level. It is essential to identify user’s sentiment to his interests. According to the previous probability generative model, topics or user’s interests are extracted only with the probability of co-occurrence, which means if a user talks about one topic frequently, the models will consider that the user is interested in it but do not care the sentiment trend on the topic. Users who talk much about a topic could have different opinions about it. It is better to distinguish these users instead of viewing them in the same community. For example some people talk a lot about the topic they like, while even if others dislike it, they could also discuss a lot about it in negative aspects. In conventional methods, these users will be clustered in the same community, but it is obvious that people with different opinions should be separated.

In this paper, we present a novel probability generative model called user-sentiment topic model (USTM), which extends Author-Topic Model [2] by adding a sentiment layer between the author and topic to form a coherent model that captures author’s interested topics with sentiment information. The key differences between USTM and others are that: (1) USTM is a fully unsupervised model to detect the sentiment trend of authors; (2) USTM can extract the topics with users’ different sentiment trends, which means USTM incorporates sentiment analysis into topic extraction for investigating the opinions users hold toward topics instead of just modeling what they talk about. Being different from other existed works for finding users’ sentiment on the topic which only show how many people like or dislike it by using classifiers, our model demonstrates why people like or dislike this topic by refining the topic extraction with different
sentiment trends. As Figure 1 shows, given a topic about iPhone, traditional works only model some people like it and some dislike it, while the result of USTM shows that some people like iPhone because of its user experience and some people are not satisfied about its battery.

One thing should be noted that our model considers sentiment and topic simultaneously, by which it makes that not all extracted topics have both positive and negative sentiment trends. Due to this setting, if there is a topic that most people like and no one shows negative feelings about it, it will not appear in any topics with high probability under negative sentiment label. Therefore, it is reasonable that topics with sentiment label modeled by USTM have better precision on users’ sentiment analysis towards it.

![Figure 1: An example of modeling users’ interests with sentiment information by USTM.](image)

The rest of the paper is organized as follows. Section 2 introduces related works. Section 3 presents User-Sentiment Topic Model (USTM) including generative process and model inference. We show our datasets, prior experiment setup in Section 4 and present the experimental results and analysis in Section 5. Finally, in section 6 we have the concluding remarks and future works.

2. RELATED WORK

Recently, the researching works about topic model have focused more on two aspects, i.e., topic models combined with author’s information and topic models combined with sentiment information. Topic model is a stream of new research in machine learning and natural language models for clustering words in order to find the underlying topics. Latent Dirichlet Allocation (LDA) [1] can robustly discover multinomial word distributions of these topics. When merging author’s information in topic model, some extended works based on LDA have been proposed. Author-Topic model [2] learns topics conditioned on the mixture of authors that composed a document which each author has a distribution over topics, Author-Persona Topic model [4] allows each author’s documents to be divided into several clusters in order to find personas of the author by the topic distribution over these documents clusters. Compared with these works that represent author’s interests or personas by topic distribution, Author Interest Topic model (AIT) [3] considers adding a document class layer between the author layer and the topic layers which represents the author’s interests as a mixture of document classes. Latent Interest Topic model (LIT) [13] is extended from AIT by adding an author class layer to model the relations between different authors, while Role-Author-Recipient model [11] considers the roles of users by assigning a role-layer to the pair of author and recipient. Moreover, researching works have tried to utilize the semantic analysis of topic model with social graph to detect communities. In [17], the authors proposed the CUT (Community-User-Topic) model which discovered communities using semantic content of the social graph. One of the first attempts to combine link based community discovery methods with content based method is the Community-Author-Recipient Topic (CART) model [18]. However, since CART, there have been few advanced attempts on combining link and content analysis to detect communities more effectively. The Topic-Link LDA model [19] and Rational Topic Models [20] draw latent topical and community distributions for each node in documents networks. Then they generate links between documents based on topical similarity and community membership similarities of their authors. All these works have shown the availability of topic models to do Social Network Analysis but they all ignore the sentiment information among the topics.

When combined sentiment detection with topic models, this probability generative model works also show a strong suitability. Several unified models of topics and sentiment have been proposed, and they extend basic topic model works to explain the sentiment trends with topic from documents such as reviews or comments [6,7,8]. Topic Sentiment Mixture (TSM) model [6] represents the sentiment as a language model separated from topics, which means TSM considers the topic and sentiment separately, the word samples from either topics or sentiments. Multi-Aspect Sentiment (MAS) model [8] aims at modeling topics to the predefined aspects that are explicitly rated by users in reviews, from which the sentiment is modeled on the aspect level according to the sentiment distribution from a weighted combination from extracted topics and words. Joint Sentiment/Topic (JST) model [7] presents a novel way to detect the sentiment of document with topic extraction and its sampling process considers that the topics are associated with sentiment and document, which can model the topic and sentiment simultaneously. JST takes much similar way as our work but it only detects the sentiment on document level.

Moreover, to limit the meaning of extracted topics, extra labels or constraints are incorporated into topic model frameworks. Labeled-LDA [32] is a supervised topic model for credit attribution in multi-labeled corpora, which constrains each topic from particular labels by assigning different prior probabilities from multi-labels in the corpora. From this definition, the topics extracted by Labeled-LDA are more meaningful and distinguishable towards labels. Our proposed model learns the idea from Labeled-LDA to distinguish the sentiment labels of topics. However, our model needs only a general sentiment paradigm word list but Labeled-LDA needs to provide labels to each document in the corpora.

There are also lots of works about sentiment analysis. As social networks become more popular, there have been some works on sentiment analysis on popular social network websites such as Twitter, but most of them [21, 22, 23, 24, 25] focus on tweet level instead of on the user level. There are also some previous works on automatically determining user level opinion or ideology [26, 27, 28, 29, 30], however, the topic used for these user level sentiment analysis is always manual to determine. Furthermore, the reason why users hold such opinions or the aspects of the topic leading to user’s different opinions are not revealed.
Different from all these works, USTM is an unsupervised generative model that captures user’s sentiment on topic level by considering topic and sentiment simultaneously. In this way, each topic extracted by USTM has a sentiment label and users have distribution over all these sentiment labeled topics. USTM aims at obtaining the sentiment-refined topics for investigating user-level sentiment analysis. Unlike previous works to do user-level sentiment analysis which either firstly extract topics from documents of users and then use classifier to do sentiment analysis or firstly fix a topic and find user’s sentiment on the topic [31], USTM can not only discover whether users like or dislike the topics but also reveal why they show these opinions by refining topics with sentiment information. Table 1 shows the notations used in this paper.

<table>
<thead>
<tr>
<th>SYMBOL</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Number of authors</td>
</tr>
<tr>
<td>K</td>
<td>Number of topics</td>
</tr>
<tr>
<td>D</td>
<td>Number of documents</td>
</tr>
<tr>
<td>V</td>
<td>Number of unique words</td>
</tr>
<tr>
<td>S</td>
<td>Number of sentiment labels</td>
</tr>
<tr>
<td>x_u</td>
<td>users associated with token i in document d</td>
</tr>
<tr>
<td>l_i</td>
<td>Sentiment label associate with token i in document d</td>
</tr>
<tr>
<td>x_d</td>
<td>topic associated with token i in document d</td>
</tr>
<tr>
<td>w_a</td>
<td>the l_i word token in document d</td>
</tr>
<tr>
<td>θ_x,i</td>
<td>the multinomial distribution of topic specific to author x and sentiment label l</td>
</tr>
<tr>
<td>φ_z,l</td>
<td>the multinomial distribution of word specific to topic z and sentiment label l</td>
</tr>
<tr>
<td>π_x</td>
<td>the multinomial distribution of sentiment label specific to author x</td>
</tr>
<tr>
<td>α, β, γ</td>
<td>the fixed parameters of Dirichlet distribution priors</td>
</tr>
</tbody>
</table>

3. Refining User’s Topic with Sentiment Information

In order to model user’s topics with sentiment labels, we extend the AT model to User-Sentiment Topic model (USTM), by adding a sentiment layer between the author layer and the topic layer. In this section, we will describe this model in details.

3.1 User-Sentiment Topic model

As we mentioned above, User-Sentiment Topic model (USTM) is an effective four-layer topic model based on the Latent Dirichlet Allocation (LDA). The LDA model is a generative model with a better statistical foundation by defining a topic-document distribution, which allows the LDA model to inference on new documents and overcome the problem of overfitting. In USTM, we keep the merits of LDA and merge the authors and sentiment information simultaneously in it. We assume that sentiment information is associated with user, topics are associated with both user and sentiment, and words are associated with sentiment label and topics. Figure 2 shows the generative process with the graphical models.

Assume that we have a corpus D with a collection of documents as \{d_1, d_2, ..., d_N\}, each document d of the corpus contains a sequence of words denoted by d=\{w_1, w_2, ..., w_{N_d}\} and each word in documents has an index from a vocabulary dictionary of V distinct terms denoted by \{1, 2, ..., V\}. S is the number of sentiment labels; K is the number of topics we want to extract and \(A_d\) are users associated with document d.

The procedure of generating the word \(w_{d,i}\) in document d can be described in four steps. Firstly, USTM samples a user x uniformly from \(A_d\) and then according to the user x, one chooses a sentiment label l from the user-sentiment multinomial distribution \(\pi_{Ad}\).

Thirdly, one chooses a topic z randomly from the topic distribution \(\theta_{x,l}\) which \(\theta_{x,l}\) is chosen conditioned on both the sentiment label l and the sampled user x with hyper parameter α. It is the key difference between our model and AT model. In AT model, for each author, there is only one topic distribution, but in USTM, each user has S topic distribution associated with different sentiment label, which provides USTM means for detecting the user’s topics on different sentiment labels. Finally, one chooses the word \(w_{d,i}\) from the distribution \(\phi_{z,l}\) conditioned with both topic z and sentiment label l with hyperparameter β. The graphical model in Figure 2(c) shows the generative process of the USTM.

The formal definition of the generative process corresponding to the hierarchical Bayesian model is as follows.

- For each document d, sample an user x uniformly from \(A_d\).
- For each user x, sample a sentiment label from the user-sentiment multinomial distribution \(\pi_{x}\) from \(\gamma\).
- For the user x and the sentiment label l, sample an user-topic-sentiment multinomial distribution \(\theta_{x,|l}\).
- For each word \(w_i\) in the document d, sample a sentiment label \(l_i\sim \pi_x\)
  - Sample a topic \(z_i\sim \theta_{x,i}\)
  - Sample a word \(w_i\) from the multinomial distribution \(\phi_{z,i}\) from \(\beta\) over words defined by the topic \(z_i\) and sentiment label \(l_i\).

The hyperparameter α and β can be seen as the prior observation counts for the number of times topic z associated with the user x & the sentiment label l sampled from the document d and the number of times words \(w_i\) associated with the topic z and the sentiment label l, before observing any words in the document.

3.2 Learning and Inference

In USTM inference process, we choose the alternative parameter estimation method, i.e., Gibbs sampling, to sample from the posterior distribution over parameters. The sampling distribution for a word given the specific topic z, sentiment label l and the uniformly sampled user x is \(P(z_i=j, S_i=m | w_i=m, z, S, a, x, \alpha, \beta, \gamma)\), where \(z_i, S_i\) represent the assignments of the \(t_{ih}\) word in a document to topic \(j\) and sentiment \(m\) respectively, \(w_i=m\) represents the observation that the \(t_{ih}\) word is the \(m_{ih}\) word in the

Figure 2 shows the generative process with the graphical models.
documents lexicon, and $z_i, S_i$ represent all topic and sentiment assignments excluding the $i_{th}$ word.

The joint probability of words $w$ assigned with sentiment label $l$ and topic $z$ with the specific user can be divided into three terms:

$$P(w, z, l, x) = P(w | z, l, x) P(z, l, x) P(l | x).$$

Each of the three terms represents a joint distribution inferred by USTM, which output as three matrix: $V \times K \times S$ matrix $\varphi$, $K \times S \times U$ matrix $\Theta$ and $S \times U$ matrix $\Pi$. In order to capture user’s positive and negative interests, we need to infer the joint user-sentiment-topic distribution $\Theta$ which models the relationships between topics, sentiments and users. We also need to infer the joint topic-sentiment-word distribution $\varphi$ and the user-sentiment distribution $\Pi$ which models relationships between words, topics and sentiment labels and the user’s sentiment trends separately at the same time to complete the sampling.

The probability of the words, sentiment and topic assignments of the entire corpus is then expanded as follows:

$$p(z, l, x, w | \alpha, \beta, \gamma) = p(w | z, l, x, \alpha) p(l | z, \beta) p(x | l, \gamma) p(x)$$

$$= \int p(w | z, l, x, \alpha) p(\theta | \alpha) d\theta \int p(z | l, x, \beta) d\beta \int p(x | l, \gamma) d\gamma$$

For the inference of these three distributions, we firstly integrate out $\theta, \varphi, \Pi$ by the Dirichlet distribution and Euler integration, and obtain the distribution $P(w, z, l, x | \alpha, \beta, \gamma)$ as follows:

$$p(z, l, x, w | \alpha, \beta, \gamma) = p(w | z, l, x, \alpha) p(l | z, \beta) p(x | l, \gamma) p(x)$$

$$= \frac{\Gamma(\sum_{u} \alpha_{u})^{\alpha_{u}} \Gamma(\sum_{r} \Gamma(\sum_{u} \alpha_{u} + \beta_{u}) \Gamma(\sum_{r} \Gamma(\sum_{u} \alpha_{u} + \beta_{u} + \gamma_{r}))^{\Gamma(\sum_{r} \Gamma(\sum_{u} \alpha_{u} + \beta_{u} + \gamma_{r})})^{\Gamma(\sum_{r} \Gamma(\sum_{u} \alpha_{u} + \beta_{u} + \gamma_{r})})}{\prod_{r} \Gamma(\sum_{u} \Gamma(\sum_{u} \alpha_{u} + \beta_{u} + \gamma_{r}))^{\Gamma(\sum_{r} \Gamma(\sum_{u} \alpha_{u} + \beta_{u} + \gamma_{r})})^{\Gamma(\sum_{r} \Gamma(\sum_{u} \alpha_{u} + \beta_{u} + \gamma_{r})})}$$

where $n_{z, l, w}$ represents the number of word $v$ assigned with topic $z$ and sentiment label $l$; $n_{x, l, w}$ represents the number of topic $z$ assigned with user $x$ and sentiment label $l$; $n_{l, w}$ represents the sentiment label $l$ associated with user $x$.

For each token in documents, the predictive topic and sentiment label assignment with the specific user is given by:

$$p(z_{dl}, l_{dl} | x_{dl} \bar{w}_{dl}, z_{dl-1}, l_{dl-1}, x_{dl}) = \frac{p(x_{dl}, w_{dl}, z_{dl}, l_{dl} | l_{dl-1}, \alpha_{l}, \beta_{l}, \gamma_{l})}{p(x_{dl}, w_{dl}, z_{dl}, l_{dl} | \alpha_{l}, \beta_{l}, \gamma_{l})}$$

During the parameter estimation, the algorithms only need to keep track of the three count matrices we mentioned, $V \times K \times S$ matrix (word by topic and sentiment) $\varphi$, $K \times S \times U$ matrix (topic by user and sentiment) $\Theta$ and $S \times U$ matrix $\Pi$. From these matrices, we can easily estimate the word-topic-sentiment distribution $\varphi$, topic-user-sentiment distribution $\Theta$ and user-sentiment distribution $\Pi$ by:

$$\Theta_{x, l, z} \propto \frac{n_{z, l, w} + \alpha}{\sum_{z=1}^{S} (n_{z, l, w} + \alpha)}$$

$$\varphi_{z, w} \propto \frac{n_{z, w} + \beta}{\sum_{w=1}^{V} (n_{z, w} + \beta)}$$

$$\Pi_{x, l} \propto \frac{n_{x, l} + \gamma}{\sum_{l=1}^{U} (n_{x, l} + \gamma)}$$

4. EXPERIMENTS

4.1 Datasets

We conduct our experiments on one Chinese dataset and two English datasets. The Chinese dataset was crawled from the comments about IT products from DOUBAN.com which consists of 7192 documents and 52167 unique words with 6129 users, we name it as douban. The two English datasets are public and available: one is a movie review dataset consisting of 1000 positive reviews and 1000 negative reviews with 312 users, we name it as movie review published by Cornell University (http://www.cs.cornell.edu/People/pabo/movie-review-data/).
And the other one is Enron email dataset published by CMU (http://www.cs.cmu.edu/~enron/). It should be noted that as USTM is a fully unsupervised model we do not use any of the sentiment label information of documents to detect the sentiment trends but just to evaluate the precision of our work. Table 2 shows the statistics of our used datasets.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Document</th>
<th>Users</th>
<th>Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>douban</td>
<td>7192</td>
<td>6129</td>
<td>52167</td>
</tr>
<tr>
<td>Movie review</td>
<td>2000</td>
<td>312</td>
<td>45011</td>
</tr>
<tr>
<td>Enron</td>
<td>39861</td>
<td>150</td>
<td>28102</td>
</tr>
</tbody>
</table>

### Table 2: Statistics of datasets.

#### 4.2 Preprocessing

#### 4.2.1 Chinese Word Segmentation

According to the particular feature of Chinese that there is no separator to distinguish each word in Chinese sentence, we firstly employ the ICTCLAS (Chinese Lexical Analysis System, http://ictclas.org/) to Chinese word segmentation job. After the processing of ICTCLAS, each document contains a serial of word tokens separated by a space. ICTCLAS can be used for Chinese word segmentation, POS tagging, named entity recognition, new word recognition and so on, here we use it just to do the Chinese word segmentation and POS tagging with its own dictionary on our dataset. The English datasets do not need this preprocessing.

#### 4.2.2 Filtering

After the process of word segmentation, the other work of preprocessing is the word filtering. In order to reduce the vocabulary size, numbers, punctuations and other non-Chinese characters were removed. Stop words and other Conjunction or Modal words were also removed. When finishing the word filtering, documents were also filtered to improve the result of topic extraction and sentiment detection, from which documents containing less than 5 words were removed. The punctuations and non-English characters were also removed from movie review dataset and Enron dataset.

#### 4.3 Model Priors

The sentiment detection is somehow more challenging than topic extraction, since authors express their opinion in a more subtle way while topic can be simply extracted according to the co-occurrence of words. One way to improve the precision of sentiment detection is to incorporate the prior information such as sentiment lexicon. In our experiments, we tried several sentiment lexicons with different size as our prior information and found that as the more prior information was incorporated into our model, the performance for modeling topics with sentiment was not better as we expected. The reason for this we considered was the problem of overfitting. Due to the usage of sentiment lexicon that we matched each word in vocabulary to the lexicon to label the prior sentiment trend of the words, it would be constrained for each word in documents to generate a proper sentiment label for specific documents if too much prior sentiment knowledge had been set. Hence, we finally choose the sentiment paradigm word list used in [7] consisting 21 positive and 21 negative paradigm words respectively as shown in Table 3.

<table>
<thead>
<tr>
<th>Positive</th>
<th>dazzling brilliant phenomenal excellent fantastic gripping mesmerizing riveting spectacular cool awesome thrilling moving exciting love wonderful best great superb still beautiful</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>sucks terrible awful unwatchable hideous bad cliched boring stupid slow waste unexciting rubbish tedious unbearable pointless cheesy frustrated awkward disappointing</td>
</tr>
</tbody>
</table>

### Table 3: Paradigm word list.

#### 4.4 Model Setup

In the experiment, the model priors were only used to initialize the sentiment label for each word in documents. We set the sentiment labels as positive, negative and neutral. The model starts by initializing the matching sentiment label of words and assigning the topic and user (from the users of document) randomly to each word. For each dataset, we randomly choose 3/4 documents as training set and others as test sets.

In our simulation, the number of topics was fixed at 100 and the smoothing parameters $\alpha, \beta, \gamma$ were set as 1/K, 0.02 and 0.01 respectively. We ran five independent Gibbs sampling chains for 1000 iterations on each dataset. For comparison, we also ran Author Topic (AT) model for five Gibbs sampling chains on each dataset with the same iterations and employ AT model + SVM as baseline method for evaluating the result of extracting users’ interested topics with sentiment information. The experiments performed on the machine with 2.13GHz Xeon 5605 processor.

#### 5. EXPERIMENTAL RESULT AND ANALYSIS

For qualitatively and quantitatively evaluating our model, we consider three performance metrics:

- **Convergence Analysis**: we use perplexity value to measure the convergence of USTM compared to LDA. Labeled LDA and Author Topic model when inferring topic and sentiment information on test datasets. We also conduct experiments with different settings such as different number of topics or iterations to evaluate the convergence of our model.

- **Accuracy Analysis**: To measure the accuracy of user’s interests with sentiment detected by USTM, we choose the related topics extracted from both USTM with sentiment information and AT model without sentiment label. In order to obtain users’ sentiment towards the topic from AT model, we employ SVM as a majority-vote mechanism for sentiment decision, which means that topic-related users’ sentiment label is simply the majority sentiment label among their on-topic documents according to SVM. We use AT model + SVM as baseline method and measure the performance via P@N, MAP and MRR on both USTM and baseline method.

- **Scalability Analysis**: To further employ our approach for social network analysis, we evaluate the running time of our model with different number of topics on datasets and propose an efficient sampling approach to enhance the efficiency of estimating.
- **Case study:** We use a case study to demonstrate how effectively our model can identify users’ interests with sentiment information.

### 5.1 Convergence Analysis

To measure the ability of a model acting as a generative model, we use perplexity to measure the convergence of our model. In order to avoid the problem of being hard to reproduce the experiments on Chinese datasets, we will detail our experiment on movie review dataset and conduct experiments with different settings.

Perplexity, which is widely used in the language modeling fields to assess the predictive power of a model, is an important indicator to demonstrate the generalization performance of a model. A lower perplexity value means that the words are less surprising to the model and their user preference is similar with model’s prediction. Given test dataset D, the perplexity value can be calculated as follows.

\[
Perplexity(D) = \exp\left\{ \frac{\sum_{d \in D} \log p(w_d | D_{max})}{\sum_{d \in D} N_d} \right\}
\]

Figure 3 shows the result of the perplexity comparison of LDA, Labeled LDA, AT model and USTM with K=30. We employ Labeled LDA as a baseline method by defining its labels according documents’ sentiment information such as positive, negative and neutral. From this figure, we find that USTM presents significantly lower perplexity value than Labeled LDA, AT model and LDA. This is due to the ability of USTM to refine the topics from different sentiment trends for users rather than no-sentiment topic extraction from AT model. This implies that USTM can well identify user’s topics from different sentiment trends by incorporating knowledge of both user and sentiment to provide a better prior for the content of new documents. We also took experiments with different size of topics to rerun the perplexity comparison to investigate the best size of topic of USTM and Author Topic model. Figure 4 shows that USTM obtains better perplexity performance with a smaller size of topics than LDA and AT model. From the curves, we can see that the perplexity values of LDA and AT model decrease rapidly with an increasing K. However, USTM has a relatively stable and lower perplexity value even when K is small, which illustrates that USTM owns a better predicting ability for unseen documents by incorporating sentiment into the process of extracting topics and this leads to its better perplexity performance when K is small than other two topic models.

![Figure 3: Perplexity values versus number of iterations by LDA, Labeled LDA, AT model and USTM. X-axis represents the number of iterations and Y-axis represents the perplexity value.](image)

### 5.2 Accuracy Analysis

To evaluate the results of modeling users’ interests with sentiment information by our model, we shall find a specific topic and topic-related users at first, then manual scoring should be used to rank the related users according to the relatedness and sentiment towards the topic. The relate-user rank is used as a criteria rank. Based on this rank, we evaluate the user-rank results of our model and the baseline method on several criterions to illustrate that our model can better obtain users’ interests with different sentiment trends. The evaluation criterions include P@N, MRR (Mean Reciprocal Rank) and MAP (Mean Average Precision).

Due to the generate process of USTM which models topic and sentiment simultaneously, not all topics with positive label have related topics with negative sentiment. Hence, the first task in our evaluation is to find a specific topic. We choose the “stock” topic and “president” topic from Enron dataset and “iPhone” topic from douban dataset. Table 4 presents the details of these topics (words in Chinese are all translated into English). To better understand what the topic is about, we just demonstrate the meaningful verb and noun of each topic and sentiment words are not shown here. All topics chosen for evaluating have their own sentiment labels modeled by USTM: “stock” topic with positive sentiment, “President” topic with negative sentiment and “iPhone” topic with both positive and negative sentiment, e.g., considering the positive topic about “iPhone” from Table 4, USTM reflects that people have nice feeling about the OS and the product of apple; however, when considering the upgrade for OS version and the payment for software, people express their negative sentiment for iPhone. When given the specific topic with sentiment, we ranked the topic-related user list modeled by USTM and compared the accuracy of this user list to the result of both baseline method. Figure 5 demonstrates the results of evaluation criterions for baseline and our USTM. We can see that our USTM outperforms baseline method on four topics in terms of P@N and MAP, the results in MRR are similar between USTM and baseline method. The reason to this fact we consider is that baseline method ranks the users with specific topic and sentiment as a linear process, which means that user lists are prepared by AT model and then filtered by sentiment label. In contrast, USTM considers sentiment and topic simultaneously making it rank users by integrated factor of both topic relatedness and sentiment information.

![Figure 4: Perplexity values versus number of topics. X-axis represents the number of topics and Y-axis represents the perplexity value.](image)
Table 4: Details of Topics related to “stock”, “president” and “iPhone”.

<table>
<thead>
<tr>
<th>“Stock”</th>
<th>“President”</th>
<th>“iPhone”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive sentiment</td>
<td>Negative sentiment</td>
<td>Positive sentiment</td>
</tr>
<tr>
<td>stocks</td>
<td>Houston</td>
<td>Apple</td>
</tr>
<tr>
<td>stock</td>
<td>business</td>
<td>Jobs</td>
</tr>
<tr>
<td>Time</td>
<td>President</td>
<td>OS</td>
</tr>
<tr>
<td>prefer</td>
<td>Vice</td>
<td>Creative</td>
</tr>
<tr>
<td>earnings</td>
<td>Office</td>
<td>Desire</td>
</tr>
<tr>
<td>contact</td>
<td>Global</td>
<td>pursue</td>
</tr>
<tr>
<td>UP</td>
<td>EES</td>
<td>notebook</td>
</tr>
<tr>
<td>offers</td>
<td>Trading</td>
<td>video</td>
</tr>
</tbody>
</table>

![Graph showing performance of users' sentiment detection on topics with USTM and Baseline method.](image)

Figure 5: Performance of users’ sentiment detection on topics with USTM and Baseline method.

We also calculate the average JS divergence to evaluate the performance of refining topic extraction with sentiment by USTM. Table 5 shows the results of the distance comparison calculated from the movie review dataset. As is shown in Table 5, the JS Divergence is calculated from LDA, AT and USTM and we observe that LDA models topics in documents level without considering the author information, so the JS Divergence is low over topics. AT model considers the topic distribution on authors, making it obtain an increase of JS Divergence over topics. We also calculate the JS Divergence of user distribution over the positive, negative and cross of both topics extracted by USTM. Since USTM infers user’s topics with sentiment information improved from AT, this leads to an increase in the JS Divergence over positive and negative topics and also leads to an increase in JS Divergence over words compared to that of AT. From the JS Divergence of the cross-positive-negative topic-word distribution, we can also see that USTM can truly distinguish the positive and negative topics over words. This result shows that the sentiment layer allows USTM to extract topics under different sentiment labels which refines the topic extraction with sentiments of users.

![Graph showing average JS divergence between user, topic and word distributions from LDA, AT model and USTM. All models are learned from the fixed number of topic, K=100.](image)

Table 5: Average JS divergence between user, topic and word distributions from LDA, AT model and USTM: All models are learned from the fixed number of topic, K=100.

<table>
<thead>
<tr>
<th></th>
<th>LDA</th>
<th>AT</th>
<th>USTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>JS(θ_s_pos / θ_s_neg)</td>
<td>0.153</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>JS(θ_s_pos / θ_s_pos)</td>
<td>-</td>
<td>0.1678</td>
<td>-</td>
</tr>
<tr>
<td>JS(θ_s_neg / θ_s_pos)</td>
<td>-</td>
<td>-</td>
<td>0.5958</td>
</tr>
<tr>
<td>JS(θ_s_pos / θ_s_neg)</td>
<td>-</td>
<td>-</td>
<td>0.5641</td>
</tr>
<tr>
<td>JS(φ_c_pos / φ_c_neg)</td>
<td>0.3987</td>
<td>0.5555</td>
<td>-</td>
</tr>
<tr>
<td>JS(φ_c_pos / φ_c_pos)</td>
<td>-</td>
<td>-</td>
<td>1.5065</td>
</tr>
</tbody>
</table>

5.3 Scalability Analysis

To further apply our model to social network analysis in our future work, we evaluate the running time of our model on datasets. Figure 6 shows the average runtime per iteration for different number of topics K. In order to improve the efficiency of our approach, we also propose an efficient sampling strategy to our model.

Convenient sampling approaches view each word in documents as the same status and assign sampling times for them only based on their occurrence. However, when a document contains words such as “algorithm” and “take”, we care more about the topic distribution on the words “algorithm” instead of “take”. Specifically, the sampling process of topic modeling should focus more on those important and meaningful words and pay less attention to other words in order to enhance the efficiency of the total running time. Based on the intuition above, we utilize Wikipedia as background knowledge to distinguish the statuses of words in documents and assign more sampling times to those meaningful words which can be matched by Wikipedia concepts. The detailed illustration of the efficient sampling approach is explained in [33].

From Figure 6, we can see that the efficient sampling approach shows a significant improvement on all datasets compared to the standard sampling approach. As the number of topics increases, the improvement of efficiency is more remarkable.

![Graph showing running time comparison between standard sampling approach and efficient sampling approach on different datasets. X-axis represents the number of topics and Y-axis represents the average runtime per iteration (in seconds), which is calculated by the total runtime/1000.](image)

Figure 6: Running time comparison between standard sampling approach and efficient sampling approach on different datasets. X-axis represents the number of topics and Y-axis represents the average runtime per iteration (in seconds), which is calculated by the total runtime/1000.

5.4 Case Study

We further present a case study to demonstrate the effectiveness of our proposed model. Table 6 shows the case, a common topic “Batman” from both AT model and USTM. We see that from the result of AT model, there is only one topic about “Batman” talking about actors or directors of the movie without any sentiment information. While in USTM, two topics about “Batman” are detected but with different sentiment labels which can be illustrated by the representative words of the two topics. In “Batman” topic with positive label, people share their admiring feelings about the character; meanwhile, we also find
that people express their opposite feelings about the story and the antagonists such as “Fangs” in negative labeled topic. We also investigate the users discovered by USTM related to the two sentiment labeled topics and make a comparison to the users ranked by AT model. As shown in Figure 7, one can observe that the users who can be found in the result from AT model are assigned to different topics by USTM according to their sentiment towards the “Batman” topic. Moreover, USTM detects some other users who are interested in this topic with sentiment but are ignored by AT model. From this case study, we see that our model can better detect users related to topics and distinguish their sentiments towards the topics. Recalling the intuition of our model, our objective is to find why people are interested in something? Is that attributed to their admiring feelings or dissatisfied feelings? We find that our proposed model can better answer this question by considering users’ interests and their sentiment simultaneously.

Table 6: Topics Extraction by AT model and USTM, all the topics are about the movie Batman.

<table>
<thead>
<tr>
<th>Topic Extraction by AT model</th>
<th>Topic Extraction by USTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without Sentiment Label</td>
<td>Positive Sentiment Label</td>
</tr>
<tr>
<td>Topic 18</td>
<td>Topic 78</td>
</tr>
<tr>
<td>Batman</td>
<td>Skillful</td>
</tr>
<tr>
<td>Robin</td>
<td>Batman</td>
</tr>
<tr>
<td>Arnold</td>
<td>Granted</td>
</tr>
<tr>
<td>Spawn</td>
<td>Meaningful</td>
</tr>
<tr>
<td>Schwarzenegger</td>
<td>twister</td>
</tr>
<tr>
<td>Schumacher</td>
<td>Peak</td>
</tr>
<tr>
<td>Comic</td>
<td>Volcano</td>
</tr>
<tr>
<td>Action</td>
<td>Reality</td>
</tr>
<tr>
<td>Devil</td>
<td>Stewart</td>
</tr>
<tr>
<td>Series</td>
<td>Tommy</td>
</tr>
</tbody>
</table>

6. CONCLUSIONS

In this paper, we present user-sentiment topic model (USTM) which can detect the user’s interests with sentiment information. In contrast to most of the works for sentiment analysis on users’ interests, our USTM analyses users’ interests and their sentiment trends simultaneously. Most of existing approaches in sentiment classification rely on supervised learning, but USTM is a fully unsupervised model which provides much flexibility for applications. Experiments have been conducted to evaluate the effect of USTM based on three different datasets and the results demonstrate that USTM performs better on modeling user’s interests with sentiment and the topics extracted by USTM are informative and clear towards the sentiment label. For future work, we plan to carry out more experiments on large scale dataset and incorporate the relationships of social network to analysis the user influence on topic level with sentiment information.

7. REFERENCES


