

# Integrating Topic Related Opinions of Users on Social Media: a Separate Way

Xie Songxian, Tang Jintao, and Wang Ting

School of Computer Science, National University of Defense Technology,  
Yanwachi Street of Changsha, Hunan, China, 410073  
{xsongx, jtaotang, tingwang}@nudt.edu.cn

**Abstract.** Social media such as Twitter, has enabled more and more people to express their opinions on the web freely, making it an extremely valuable source for mining opinions of users about all kinds of topics. In this paper we study how to automatically integrate topic related opinions expressed by a user in not so well-written User-Generated Content (UGC). We propose a subjectivity model by combining topics a user talks about and his opinions towards these topics. There have been several generative topic-sentiment mixture models designed to capture sentiments and topics in text simultaneously. In contrast to generative models, we construct our model with LDA topic model and sentiment analysis techniques in a separate way. We compare our model and generative models in a series of experiments on Twitter data. Results show that the proposed model is effective and can generate useful aligned integrated opinion summaries of users. Furthermore, the proposed model is more suitable for social media context, thus can catch more fine-grain opinions than generative models and get better performance in a link prediction task.

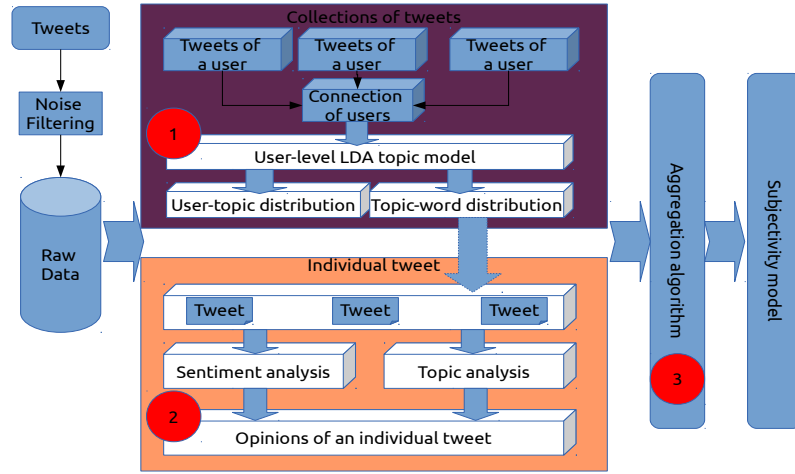
**Keywords:** generative model, social media, opinion mining, subjectivity model

## 1 Introduction

With the rise of text-based social media such as Twitter, millions of users are more and more willing to publish online short messages to express their opinions on a great variety of topics they are interested in. The wide coverage of topics, dynamics of discussion, and abundance of opinions imbedded in the social media data make them extremely valuable for mining users' opinions about all kinds of topics (e.g., products, political figures, etc.), which in turn can enable a wide range of applications, such as opinion search for ordinary users, opinion tracking for business intelligence, and user behavior prediction for targeted advertising. However, with such a large scale of information source, it is quite challenging to integrate and digest all the opinions from different users. For example, a query "iPhone" on Twitter (as of Jan. 14, 2014) returns 830,879 tweets of 231,233 users, suggesting that there are many users have expressed opinions more than once about iPhone in their tweets. To enable an application to benefit from all kinds of opinions of different users, it is thus necessary to automatically integrate and present an integrated opinion summary for each user [1]. In fact, users often publish many messages on the topics they are interested in, therefore how to find these topics

and integrate opinions towards each topic poses special challenge for opinion mining related researchers.

In this paper, we propose a two-parts model ( name as subjectivity model) by incorporating topic model and sentiment analysis into one framework, in which part 1 represents topics of interest distribution of users, while part 2 represents the distribution of opinions towards these topics. Specifically, we propose a general method to solve this integration problem in three steps illustrated as in Figure 1: (1) extract topics of interest from tweets of a user using user-level LDA; (2) extract separated sentiment and topic for each tweet with sentiment and topic analysis (3) summarize and integrate the extracted opinions towards topics to form a subjectivity model for each user.



**Fig. 1.** Framework of two-parts subjectivity model.

Several studies have tried to combine sentiment into a unified generative topic model to mine topic related sentiments in reviews or blogosphere. Mei *et al.* [2] proposed Topic-Sentiment Mixture (TSM) model that can reveal latent topics of a blog and the topics associated sentiments. Lin *et al.* [3] proposed a Joint Sentiment/Topic (JST) model based on LDA that can detect topic and sentiment simultaneously in reviews. This paper presents a comparative study of these closely related generative models for unsupervised user level opinion mining.

This paper makes the following contributions:

- We define a new problem of opinion integration at user-level. To the best of our knowledge, there are few works that meet such a problem.
- We propose a two-parts model and a separated approach for integrating opinions scattered around in tweets of a user for the topics he is interested in.
- We evaluate the proposed method both qualitatively and quantitatively comparing with state-of-the-art generative models. The results show that our method is effective for integrating opinions for users and a friend recommendation task.

Collecting and digesting opinions about a topic from different users are critical for many tasks such as shopping, medical decision making, and social interactions. Our proposed method is quite general and can be applied to integrate opinions about any topic in any domain, thus potentially has many interesting applications. The rest of the paper is organized as follows. In Section 2, related works are presented, then we formally define the novel problem of opinion integration in Section 3. After that, we present our model and analyze the difference with generative model in Section 4. We discuss our experiments and results in Section 5. Finally, we conclude in Section 6.

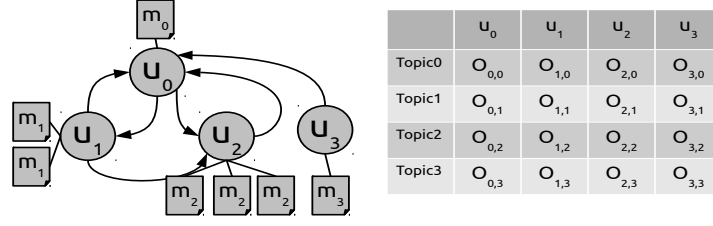
## 2 Related Works

Sentiment analysis is a popular research area and previous researches have mainly focused on reviews or news comments [4,5]. Recently, there have been some works on sentiment analysis on Twitter, mainly focusing on the tweet level [6,7,8,9,10], of which, the techniques employed are generally standard tweet-level algorithms that ignore links between users. There have been also some previous works on automatically determining user-level opinions or ideology [11,12,13,14], generally looking at information embedded in the contents that the users generate. Most of related researches mainly focused on identification of sentimental object [15], or detection of objects sentimental polarity [16] without considering the topic aspects.

Since the introduction of LDA model [17], various extended LDA models have been used for topic extraction from large-scale corpora at user level. Rosen-Zvi *et al.* [18] introduced the Author-Topic (AT) model, which includes authorship as a latent variable. Ramage *et al.* [19] introduced Labeled LDA to supervise the generation of topics via user tags. Topic models can also be utilized in sentiment analysis to correlate sentiment with topics [2,3]. Mei *et al.* [2] and Lin *et al.* [3] incorporated topic models and sentiment analysis for reviews and blogs. However, our work attempts to integrate opinions of a user by combining topic models and sentiment analysis together into a two-parts framework.

## 3 Opinion Integration Problem

As we describe in Section 1, a user usually publishes multiple messages on multiple topics when he using social media. Therefore what's the opinion of a user on a special topic can not be determined from just one tweet, but should be iterated from the topic related tweets he has published. In this paper, we put forward a new problem which is defined as *Opinion Integration Problem* (OIP). We focus on user-level rather than tweet-level opinion because the end goal of opinion mining technologies is to find out what a person thinks but not what only a piece of message states, determining the opinion expressed in individual text is usually a middle step for that ultimate objective. Additionally, it is plausible that there are cases where opinions in some of a users tweets are genuinely ambiguous because they are restricted to be very short, but his overall opinion can be determined by looking at his collection of tweets [10]. We show a typical scenario of user-level topic related opinion integration problem on Twitter in Figure 2. As denoted in the figure, a subnetwork of Twitter (users  $\{u_i\}$  and their directional rela-



**Fig. 2.** Illustration of Opinion Integration Problem.

tions), and their associated tweets (denoted as  $\{m_i\}$ ) are used to extract topics (denoted as  $\{Topic_j\}$ ) and opinions (denoted as  $O_{i,j}$ ). Opinion  $O_{i,j}$  is not the opinion imbedded in single tweet  $m_i$ , but the integrated opinion form from all tweets  $\{m_i\}$ .

There are two important factors that should be taken into considerations for the OIP problem. Firstly, topics of all users and tweets talks about and should be determined in a same topic space so as the target of opinion is normalized. Secondly but most importantly, opinions and topics are closely related, tweets of a user around some topic often cover a mixture of features related to that topic with different preferentials. Different opinions may be expressed by the user towards different features, where users may like some aspects of an topic but dislike other aspects. Therefore, how to integrate all opinions of tweets related to a topic into one opinion and represent it reasonably poses special challenge. There are different ways to meet such a challenge, generatively as TSM model or separately as we describe in this paper, and we will define our model and compare them in this paper.

## 4 Subjectivity Model

In this section, we give a formal definition of the model we work with to meet the challenges of OIP problem. Usually user-level opinion is to classify each users sentiment on a specific topic into one of two polarities: “Positive” and “Negative”. “Positive” means that the user supports or likes the target topic, whereas “Negative” stands for the opposite. However we adopt a broad “opinion” definition as sentiment coverage towards a topic over a fine-grain strength values to differentiate opinions of users. The notion of “opinion” is quite vague; we adopt this broad definition to ensure generality of the model. We frame the problem in the context of Twitter to keep things concrete, although adaptation of our model to other social-network settings is straightforward. We name our model as “**Subjectivity Model**” as it models the subjective information in the content generated by a user. Therefore, we give a formal definition of the subjectivity model under the context of Twitter as follows.

### 4.1 Definition

Let  $G = (V, E)$  denote a social network on Twitter, where  $V$  is a set of users, and  $E \subset V \times V$  is a set of follow relationships between users. For each user  $u \in V$ , there is a

tweets collection  $M_u$  denoting his message history. We assume that there is a topic space  $T$  containing all topics users in  $V$  talk about, and a sentiment valence space  $S$  to evaluate their opinions towards these topics. For the “subjectivity” of a user  $u \in V$ , we refer to both topics and opinions articulated in his tweets collection  $M_u$ .

**Definition 1 (Subjectivity Model).** *The subjectivity model  $P(u)$  of user  $u$ , is the combination of topics  $\{t\}$  the user talks about in topic space  $T$  and his opinions  $\{O_t\}$  towards each topic distributed over sentiment valence space  $S$ .*

$$P(u) = \{(t, w_u(t), \{d_{u,t}(s) | s \in S\}) | t \in T\} \quad (1)$$

where:

- with respect to user  $u$ , for each topic  $t \in T$ , its weight  $w_u(t)$  represents the distribution of the user’s interests on it, subject to  $\sum_{t=1}^{|T|} w_u(t) = 1$ .
- opinion of the user towards topic  $t$  is modelled as a topic-dependent sentiment distribution over sentiment valence space  $S$ ,  $O_t = \{d_{u,t}(s) | s \in S\}$ , subject to  $\sum_{s=1}^{|S|} d_{u,t}(s) = 1$ .

Subjectivity Model captures users subjectivity by considering both topic and opinion simultaneously. Topic weights of users are modelled by a topic distribution and users have opinion distribution on each topic over all sentiment valence space. Subjectivity Model aims at obtaining the sentiment-refined topics for investigating user-level opinion mining. Unlike previous works to do document-level sentiment analysis which either calculate the overall document sentiment or model aspect related opinions as sentiment polarity classification [10]. We now address the problems of estimating the free parameters in Subjectivity Model and inferring topic related opinions once the parameter values have been learned.

## 4.2 Establishment of Subjectivity Model

According to the definition of subjectivity model, there are two distributions to model the subjectivity: the topic distribution and the opinion distribution for each topic. Both of them need to be inferred from historic content produced by users. However, content analysis on Twitter is challenging: the volume of tweets is so huge while a single tweet is very short with limit of 140 characters, and informal languages are widely used, which make many supervised learning approaches and natural language processing techniques invalid. Hence effectively modeling content on Twitter requires techniques that can readily adapt to these challenges and require little supervision. With state-of-the-art topic model and sentiment analysis techniques, we establish subjectivity model by identifying topics and opinions in an unsupervised way simultaneously. The main advantage of our establishing framework is that it employs social-network structure to help us overcome both the paucity of textual information in short tweets and the lack of a large amount of labeled data [20].

**Topic Analysis** The topics of a tweet are latent and have to be inferred from its content. Previous studies have tried to identify topics from tweets by finding key words [21], extracting entities [22] or linking tweets to external knowledge categories [23], however, the sparsity is a main problem for these methods because even users have common local topics they still might refer to a topic with different vocabulary. Works show that topic models such as **Latent Dirichlet Allocation (LDA)** model [17] have been efficient ways to characterize latent topics of large volume corpus. Topics of LDA are broader in concept, since a single topic consists of the whole collection of related words. Therefore we adopt a user-level LDA model to find latent topics for users. To distill the topics, documents of LDA should naturally correspond to tweets content. As our goal is to understand the topics that each user is interested in rather than the topics each single tweet talks about, we aggregate the tweets published by each user into a single document, and replace documents of LDA with aggregated tweet documents. So a document stands for a user in our model, and a user can be represented as a multinomial distribution over topics, which corresponds to the topic distribution of the user’s subjectivity model.

Formally, given a set of users  $U$  and the number of topics  $K$ , a user  $u$  ( $u \in U$ ) could be represented by a multinomial distribution  $\theta_u$  over topics with a Dirichlet prior parameterized by  $\alpha$ . A topic  $k$  ( $k \in K$ ) is represented by a multinomial distribution  $\beta_k$  with another Dirichlet prior parameterized by  $\eta$ . The parameters  $\theta_u$  and each  $\beta_k$  can be estimated by Gibbs sampling or variational inference. We use variational inference-based topic model package Gensim [24].

**Opinion Analysis** Users often express opinions towards their topics of interest by publishing topic-related tweets. In order to explore the opinions of users, we need to understand sentiment embedded in each tweet. Sentiment analysis mainly depends on machine learning or rule-based approaches. Machine learning approaches often need labelled data for the training process, which is often impossible for Twitter because of the large volume of tweets and its dynamic language characteristics. Therefore we adopt rule-based approaches, which could adapt to Twitter with good flexibility by changing its particular characteristics into rules [25].

The SentiStrength package has been built especially to cope with sentiment analysis in short informal text of social media [25]. It combines lexicon-based approaches with sophisticated linguistic rules adapted to social media, which is suitable for analyzing sentiment of tweets in our research settings. SentiStrength assigns two values to each tweet standing for sentiment strengths: a positive (within  $[1, 5]$ ) and a negative (within  $[-5, -1]$ ) sentiment measurement, both ranging from 1 to 5 on absolute integer scales, with 1 denoting neutral sentiment and 5 denoting highest sentiment strength. Sentiment assigned by SentiStrength is not a simple binary value but a fine-grained strength, which can catch fine opinion distributions in a user’s subjectivity model. For the convenience calculation, we map the output of SentiStrength to a single-scaled sentiment valence space  $[0, 8]$  as follows:

$$o = \begin{cases} p+3 & \text{if } |p| > |n| \\ n+5 & \text{if } |n| > |p| \\ 4 & \text{if } |p| = |n| \end{cases} \quad (2)$$

Where  $p$  denotes the positive sentiment value and  $n$  denotes negative sentiment value. In the sentiment valence space  $[0, 8]$ , value 4 indicates neutral sentiment, while values above 4 indicate positive sentiment and values below 4 indicate negative sentiment. With the sentiments of all tweets, we can aggregate opinions towards a topic as a sentiment distribution over sentiment valence space  $[0, 8]$ .

**Concrete Subjectivity Model** With statistical topic analysis and opinion analysis described above, we can concrete Subjectivity Model now. For user set  $U$  of a social network, we denote tweet set published by a user  $u$  as  $M_u = \{m_i | i \in [1, \dots, N]\}$ . Each  $M_u$  is concatenated to a single document  $d_u$  to construct Topic Space  $T = \{t_i | i = 1, \dots, K\}$ . A topic model is built with parameter  $\theta$  representing the distribution of each user over topics in the Topic Space  $T$ , and parameter  $\beta$  represents the distribution of each topic over the vocabulary of all tweets. SentiStrength is applied to each tweet  $m$  in collection  $M_u$  and outputs sentiment strength  $s_m$  for tweet  $m$ . We build the subjectivity model  $P(u)$  for user  $u$  as Algorithm 1:

---

**Algorithm 1** Establishment of subjectivity model .

---

**Input:**

- The user set of a local network,  $U$ ;
- The tweet set published by each user  $u$ ,  $M_u$ ;

**Output:**

- The subjectivity model for each user  $u$ ,  $P(u)$ ;
  - 1: Topic analysis with a user-level LDA as Section 4.2, getting a topic model  $P(\theta, \beta | M_u, U)$ ;
  - 2: **for all** tweet  $m \in M_u$  **do**
  - 3:   Sentiment analysis as Section 4.2, outputting sentiment of  $m$ ,  $s_m$ ;
  - 4: **end for**
  - 5: **for** user  $u \in U$  **do**
  - 6:   the topic distribution is the corresponding component of parameter  $\theta$ ,  $\theta_u$ ;
  - 7:   the topics he tweets about are  $Z_u = \{t | p(t | \theta_u) > 0, t \in T\}$ ;
  - 8: **end for**
  - 9: **for** tweet  $m \in M_u$  **do**
  - 10:   topics of  $m$  can be identified by the topic model,  $Z_m = \{t | p(t | \theta, \beta, Z_u) > 0, t \in T\}$ ;
  - 11: **end for**
  - 12: **for** each topic  $t \in Z_u$  **do**
  - 13:   **for** sentiment value  $s \in S$  **do**
  - 14:     count the number of tweets which talk about topic  $t$  with sentiment value  $s$ ,  $N_s = \sum_{m \in M_u} I(s_m), \text{ if } s_m = s \& t \in Z_m$ ;
  - 15:   **end for**
  - 16:   calculating opinion towards topic  $t$ ,  $O_t = \left\{ \frac{N_s}{\sum_{s \in S} N_s} \right\}$ ;
  - 17: **end for**
  - 18: establishing subjectivity model of user  $u$ ,  $P(u) = \left\{ \left( t, p(t | \theta_u), \left\{ \frac{N_s}{\sum_{s \in S} N_s} \right\} \right) | t \in Z_u, s \in S \right\}$ ;
  - 19: **return**  $P(u)$ ;
-

In the algorithm, we assume the sentiment of tweet  $m$  is related to every topic it talks about in  $Z_m$  for simplicity. Accordingly subjectivity model  $P(m)$  for tweet  $m$  as:

$$P(m) = \{(t, p(t|\theta, \beta), d_{m,t}(s)) = 1.0 | t \in Z_m, s \in S\} \quad (3)$$

Noting that, the opinion towards each topic is a distribution of 1.0 on a single sentiment value  $s$  of tweet  $m$ .

### 4.3 Comparison with Generative Model

Several probability generative models unifying topics and sentiment have been proposed, and they extend basic topic model to explain the sentiment related to topic from documents such as reviews or comments [2,3,26]. Topic Sentiment Mixture (TSM) model [2] represents the sentiment as a language model separated from topics, which means TSM considers the topic and sentiment orthogonally, the word samples from either topics or sentiments. Multi-Aspect Sentiment (MAS) model [26] 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 [3] 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. These models are similar as our work in mining topic-related opinion at document level, all of which try to model topic and sentiment at the same time in a generative way. Specifically they learn a general word-sentiment distribution to model the sentiment of blogs or reviews, which may not work well for short and informal social media languages such as tweets. Compared to the traditional topic-based representation, sentiment representation is deemed to be more challenging as sentiment is often embodied in subtle linguistic mechanisms such as the use of sarcasm or incorporated with highly domain-specific information. Intuitively, sentiment is dependent on contextual information, such as language usage characteristics. Sentiment of tweets is determined not only by formal words but also by various special characteristics of Twitter languages such as emoticon, capitalized words, repeated letters and exclamation mark, etc. Those features can not be easily modeled by a probabilistic distribution. However, rule-based sentiment analysis methods can catch some subtle sentiment of tweets by transforming these characteristics into rules. Therefore, we construct our model in a framework analyzing topics and opinions separately but not in a unified generative way as TSM and JST. In the next section We will carry out series of experiments to compare them with our method.

## 5 Experiment

In this section, we present evaluation experiments of our method using Twitter data. We performed evaluation experiments in terms of performance of both topic opinion extraction and link prediction assessment.



### 5.1 Dataset and Settings

We use an off-the-shelf dataset [27], which is constructed by crawling Twitter. The details about the dataset can be summarized as Table 1. We use the 139,180 users with

**Table 1.** Twitter Dataset Statistics

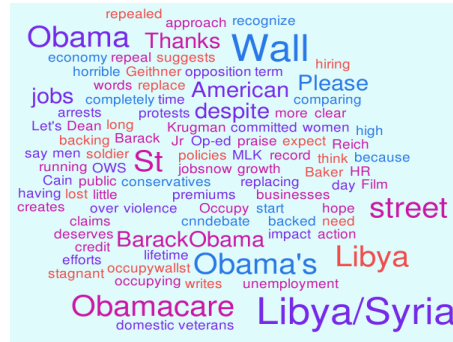
Total users	139,180
Tweets per user	549
Friends per user	14.8
Followers per user	14.9

their following relationships and tweets, as our data set to test our model and compare it with other models.

Besides our model, we also conduct a set of comparing experiments on both JST and TSM. We set topic number to 100 in the experiments. The symmetry Dirichlet prior  $\alpha$  and  $\beta$  were set to  $50/T$  and 0.01 respectively for all models. The asymmetry sentiment prior  $\gamma$  empirically was set to (0.01, 1.8) for JST. Results of JST were averaged over 5 runs with 2000 Gibbs sampling iterations.

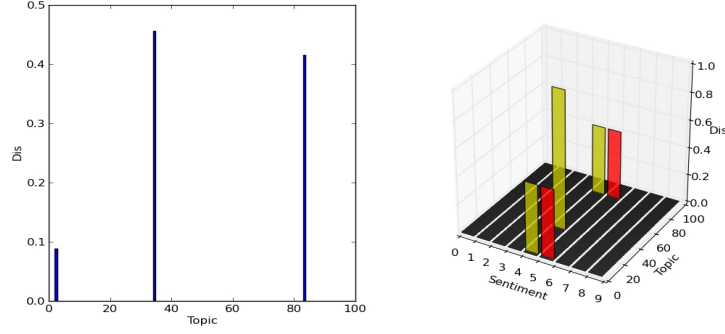
### 5.2 Case Study

In order to qualitatively evaluate the effectiveness of our unsupervised topic modeling approach, we give a vivid subjectivity model example of a user with 533 tweets, with his all tweets illustrated as Figure 3 in a word-cloud way. Figure 4 is the visualized sub-



**Fig. 3.** Word Cloud of the User.

jectivity model of the user in a  $[0, 100]$  topic space and a  $[0, 8]$  sentiment valence space, which is constructed according to our method. It is obvious that the user is interested in three topics (topic 2: “#Obamacare”, topic 32: “#libya” and topic 83: “#occupywallst”), and the left part of Figure 4 denotes the weights of his topics of interest. The right



**Fig. 4.** Subjectivity model example. The left subgraph denotes interests distribution on topic 2, 32 and 83: ( $w_u(2) = 0.08, w_u(32) = 0.48, w_u(83) = 0.44$ ). The right subgraph denotes opinions towards topics:  $O_2 = (d_{u,2}(4) = 0.5, d_{u,2}(5) = 0.5)$ ,  $O_{32} = (d_{u,32}(4) = 1.0)$ ,  $O_{83} = (d_{u,83}(4) = 0.5, d_{u,83}(5) = 0.5)$ .

part denotes the opinions towards three topics of the user, in which he is neutral to topic “#libya” with 100% distribution on sentiment strength value 4, positive to topic “#Obamacare” and “#occupywallst” with 50% on value 4 and 50% on value 5. From the example, it is demonstrated that our model can give a detail description for the subjectivity of users in that it can model not only the interest distribution but also opinion coverage over a fine-grain sentiment. To the best of our knowledge, generative models such as TSM and JST can only give relative coverage over positive, negative and neutral sentiment polarities.

### 5.3 Link Prediction Performance

It is hard for us to directly evaluate the effectiveness of our model quantitatively, therefore we adopt link prediction of social network analysis as an applicable task to test our model and compare our model with TSM and JST model. When a user forms a link in a network such as Twitter, then the principle of homophily [28], or “birds of a feather flock together” [29] suggests that users that are “connected” may tend to hold similar opinions [30]. The connection a user creates may correspond to approval or a desire to pay attention, or suggestive of the possibility of a shared opinion. Therefore we calculate the content similarity between two users as a measurement to predict whether they will connect to each other. For the content similarity we calculate three kinds of similarities including the words similarity with bag-of-words model, the topic similarity with LDA model and subjectivity similarity with our model TSM model and JST model. As for the subjectivity similarity, we mean both topic and opinion similarity between two users by adding them together. In our experiment, we adopt cosine similarity to calculating the similarities.

We randomly select 10,000 target users from our dataset with at least 80 followers. The followers of a target user is regard as positive instances, and the same number of non-followers are sampled as negative instances for that user. We combine all positive

and negative instances into a balanced test dataset, and calculate three kinds of similarities between every target user and their test users as features for a supervised classifier. We adopt logistic classifier and run Leave-One-Out Cross-Validation (LOOCV) 100 times for each user to normalize the results. The final performance is evaluated by averaging all results of the 10,000 users. Accuracy is used as our performance measurement, and the result is list in Table ?? As can be observed from the result table

**Table 2.** Accuracy performance. BOW denotes words similarity, LDA denotes topic similarity, and SUB,TSM, JST denotes subjectivity similarity with our model, TSM and JST model separately. A significant improvement over BOW with \* and LDA with ‡ ( $p < 0.05$ ).

Feature	Accuracy(%)
BOW	55.85
LDA	61.76 *
TSM	70.94 * ‡
JST	66.25 * ‡
SUB	71.53 * ‡

that: firstly, the performance of bag-of-words model is the lowest with 55.85% accuracy, and we think the reason lying in the sparsity of tweets caused by informal usage of language by different users, if topic model is applied, the sparsity is lessened by clustering different words into one topic; secondly, if topic related opinions are considered, the prediction accuracy is improved significantly from only topic similarity in that all three models (TSM 70.94%, JST 66.25% and SUB 71.53%) outperform the LDA model; finally, compared with two generative models, our model (SUB) outperforms JST significantly, and get a little better performance than TSM but the improvement is not significant, we think it is because sentiment analysis of our model is more suitable for the Twitter language, for it can extract subtle sentiment imbedded in special language characteristics such as repeated letters and emoticons. Furthermore, our model can output more fine-grain opinions than just three polarities, which can catch more fine opinion similarity between users.

## 6 Conclusion

In this paper, we study a novel opinion integration problem of social media users. We propose a Subjectivity Model to solve this problem in a separated framework. Users historical messages are leveraged to construct their topic related opinions through topic model and sentiment analysis techniques. With this model, we can automatically generate an integrated opinion summary that consists of both topic of interest distribution and topic related opinion distribution for a user. The proposed model is compared with generative models such as TSM and JST model to show that it is more suitable for the social media language. We experimented on Twitter data and the results show that the proposed Subjectivity model can effectively model topic related opinion with two probabilistic distributions and clearly outperforms generative models in a link prediction

task. In the future, we will apply our model in several social network analysis applications to testify its effectiveness.

## References

1. Lu, Y., Zhai, C.: Opinion integration through semi-supervised topic modeling. In: Proceedings of the 17th international conference on World Wide Web, ACM (2008) 121–130
2. Mei, Q., Ling, X., Wondra, M., Su, H., Zhai, C.: Topic sentiment mixture: modeling facets and opinions in weblogs. In: Proceedings of the 16th international conference on World Wide Web, ACM (2007) 171–180
3. Lin, C., He, Y.: Joint sentiment/topic model for sentiment analysis. In: Proceedings of the 18th ACM conference on Information and knowledge management, ACM (2009) 375–384
4. Pang, B., Lee, L.: Opinion mining and sentiment analysis. *Foundations and trends in information retrieval* **2**(1-2) (2008) 1–135
5. Liu, B.: Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies* **5**(1) (2012) 1–167
6. Barbosa, L., Feng, J.: Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics (2010) 36–44
7. Davidov, D., Tsur, O., Rappoport, A.: Enhanced sentiment learning using twitter hashtags and smileys. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics (2010) 241–249
8. Jiang, L., Yu, M., Zhou, M., Liu, X., Zhao, T.: Target-dependent twitter sentiment classification. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Association for Computational Linguistics (2011) 151–160
9. Li, G., Hoi, S.C., Chang, K., Jain, R.: Micro-blogging sentiment detection by collaborative online learning. In: Data Mining (ICDM), 2010 IEEE 10th International Conference on, IEEE (2010) 893–898
10. Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., Li, P.: User-level sentiment analysis incorporating social networks. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (2011) 1397–1405
11. Agrawal, R., Rajagopalan, S., Srikant, R., Xu, Y.: Mining newsgroups using networks arising from social behavior. In: Proceedings of the 12th international conference on World Wide Web, ACM (2003) 529–535
12. Mostafa, M.M.: More than words: Social networks text mining for consumer brand sentiments. *Expert Systems with Applications* **40**(10) (2013) 4241–4251
13. Malouf, R., Mullen, T.: Taking sides: User classification for informal online political discourse. *Internet Research* **18**(2) (2008) 177–190
14. Yu, B., Kaufmann, S., Diermeier, D.: Classifying party affiliation from political speech. *Journal of Information Technology & Politics* **5**(1) (2008) 33–48
15. Liu, H., Zhao, Y., Qin, B., Liu, T.: Comment target extraction and sentiment classification. *Journal of Chinese Information Processing* **24**(1) (2010) 84–89
16. Zhai, Z., Liu, B., Xu, H., Jia, P.: Constrained lda for grouping product features in opinion mining. In: Advances in knowledge discovery and data mining. Springer (2011) 448–459
17. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *the Journal of machine Learning research* **3** (2003) 993–1022
18. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proceedings of the 20th conference on Uncertainty in artificial intelligence, AUAI Press (2004) 487–494

19. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: 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: Volume 1-Volume 1, Association for Computational Linguistics (2009) 248–256
20. Lin, C., He, Y., Everson, R.: A comparative study of bayesian models for unsupervised sentiment detection. In: Proceedings of the Fourteenth Conference on Computational Natural Language Learning, Association for computational linguistics (2010) 144–152
21. Chen, J., Nairn, R., Nelson, L., Bernstein, M., Chi, E.: Short and tweet: experiments on recommending content from information streams. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM (2010) 1185–1194
22. Abel, F., Gao, Q., Houben, G.J., Tao, K.: Analyzing user modeling on twitter for personalized news recommendations. In: User Modeling, Adaption and Personalization. Springer (2011) 1–12
23. Macskassy, S.A., Michelson, M.: Why do people retweet? anti-homophily wins the day! In: ICWSM. (2011)
24. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta, ELRA (May 2010) 45–50 <http://is.muni.cz/publication/884893/en>.
25. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology* **61**(12) (2010) 2544–2558
26. Zhao, T., Li, C., Ding, Q., Li, L.: User-sentiment topic model: refining user’s topics with sentiment information. In: Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, ACM (2012) 10
27. Li, R., Wang, S., Deng, H., Wang, R., Chang, K.C.C.: Towards social user profiling: unified and discriminative influence model for inferring home locations. In: KDD. (2012) 1023–1031
28. Lazarsfeld, P.F., Merton, R.K.: Friendship as a social process: A substantive and methodological analysis. In Berger, M., Abel, T., eds.: *Freedom and control in modern society*. Van Nostrand, New York (1954)
29. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. *Annual review of sociology* (2001) 415–444
30. Thelwall, M.: Emotion homophily in social network site messages. *First Monday* **15**(4) (2010)