Mining Topic-level Opinion Influence in Microblog

Daifeng Li
Dept. of Computer Science and Technology
Tsinghua University
Beijing, China
ldf3824@yahoo.com.cn

Xin Shuai
School of Informatics and Computing
Indiana University
Bloomington
IN, USA
xshuai@indiana.edu

Guozheng Sun
Dept. of Research Center
Tencent Company
Beijing, China
gordon.gzsun@gmail.com

Jie Tang
Dept. of Computer Science and Technology
Tsinghua University
Beijing, China
jery.tang@gmail.com

Ying Ding
School of Library and Information Science
Indiana University
Bloomington
IN, USA
dingying@indiana.edu

Zhipeng Luo
Beijing University of Aeronautics and Astronautics
Beijing, China
patrick.luo2009@gmail.com

ABSTRACT

This paper proposes a *Topic-Level Opinion Influence Model (TOIM)* that simultaneously incorporates topic factor and social influence in a two-stage probabilistic framework. Users' historical messages and social interaction records are leveraged by TOIM to construct their historical opinions and neighbors' opinion influence through a statistical learning process, which can be further utilized to predict users' future opinions on some specific topic. We evaluate our TOIM on a large-scaled dataset from *Tencent Weibo*, one of the largest microblogging website in China. The experimental results show that TIOM can better predict users' opinion than other baseline methods.

Categories and Subject Descriptors

H.2.8 [**Database and Management**]: Data Mining; J.4 [**Computer Applications**]: Social and Behavioral Science

General Terms

Algorithms, Experimentation

Keywords

Opinion Mining, Sentiment Analysis, Social Influence, Topic Modeling

1. INTRODUCTION

Opinion mining, or sentiment analysis, aims to classify polarity of a document into positive or negative. There're two important factors that should be taken into considerations. One, *opinions and topics are closely related.* The online discussions around some entity, or object, often cover a mixture of features/topics related to that entity with different preferentials. Different opinions may be expressed by users towards different topics, where users may like some aspects of an entity but dislike other aspects. Two, *users' opinions are subject to social influence.* The rise of social media puts the sentiment analysis in the context of social network. Users not only express their individual opinions, but also exchange opinions with others. In the context of opinion mining, social influence refers to the phenomenon that one is inclined to agree (positive influence) or disagree (negative influence) with his/her neighbors' opinions with different degrees, depending on the influence strengths.

Several opinion mining related studies are in line with our work. Mei et. al [8] proposed Topic-Sentiment Mixture (TSM) model that can reveal latent topical facets in a Weblog collections, and their associated sentiments. Lin et. al [5] proposed a joint sentiment/topic (JST) model based on LDA that can detect topic and sentiment simultaneously. Both TSM and JST tried to model topic and sentiment at the same time but social influence is not considered. Our paper *tries to incorporate topic modeling, social influence and sentiment analysis into a two-stages model to classify users' polarities.*

We show a typical scenario of topic level opinion influence analysis on Tencent Microblog (a chinese microblogging website) in Figure 1. Like Twitter, Weibo users can post messages of up to 140 chinese characters and *follow* other users to read their messages. Followers can reply to other users' message by leaving their own comment, whose opinions can be mined from the comments. Two types of entities (user and message) and multiple types of relations (user posts/comments on message, user replies to another user) constitute a heterogenous network built on Tencent Weibo. Specifically, Lucy comments on both Lee and Peggy's messages and replies to both of them on the *visual effect* aspect of the movie *Titanic 3D*. Given the topological and text information, we aim to generate a topic opinion influence ego-network, with Lucy at the center influenced by Lee and Peggy. Their historical opinion distributions over positive/negative opinion, as well as type (*agree* means positive influence while *disagree* means negative influence) and strength (agree/disagree probability) of the opinion influence between Lucy and Lee/Peggy are calculated. Finally, How can we predict Lucy's further opinion by jointly consid-

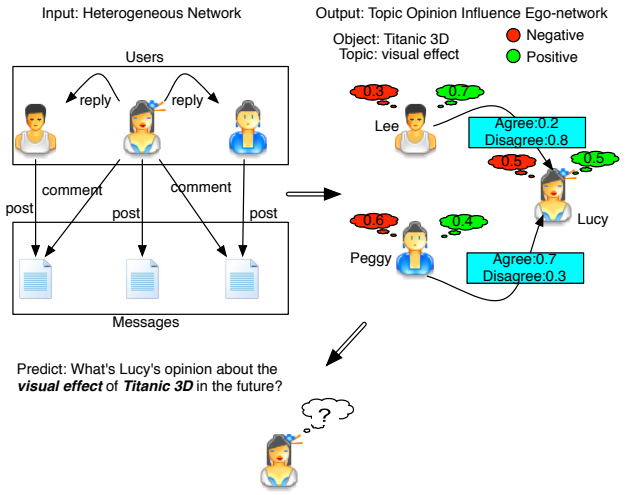


Figure 1: Motivating example

ering her own opinion preference and opinion influence from Lee and Peggy?

To solve the problem in Figure 1, we propose a *Topic-level Opinion Influence Model (TOIM)* that simultaneously incorporates topic factor and social influence in a unified probabilistic framework; users' historical messages and social interaction records are leveraged by TOIM to construct their historical opinions and neighbors' opinion influence through a statistical learning process, which can be further utilized to predict users' future opinions towards some specific topic. Our main conclusions include:

- we propose a probabilistic model to capture the dual effect of topic preference and social influence on opinion prediction problem.
- we testify our model on a new and large-scaled chinese microblog dataset from Tencent Weibo.

This paper is organized as follows: Section 2 defines the problem. Section 3 explains the mechanism of TOIM and illustrates the process of model learning process. Section 4 shows the experimental results and case studies. Section 5 lists related literature and Section 6 concludes the study.

2. PROBLEM DEFINITION

In this section, we formally define our problem of predicting a user's opinion regarding certain topic by jointly considering the user's historical opinion and opinion influence from the neighbors. Our problem starts with a *Heterogeneous Social Network* on Tencent Weibo, where nodes include all users (i.e. followers and followees) and all messages (i.e. posts and comments), and edges include all actions from users to messages (i.e. post and comment) and all connections from users to users (i.e. reply). Specifically, given a query object , a sub-graph $G = (U, M, A, E)$ can be extracted from the Heterogeneous Social Network where $U = \{u_i\}_{i=1}^{V}$ is a set of users once posted or commented on messages about the object, $M = \{m_i\}_{i=1}^{D}$ is a set of messages posted or commented from $u_i \in U$, $A = \{(u_i, m_i)|u_i \in U, m_i \in M\}$ is a set of edges indicating u_i posted or commented on m_j, and $E = \{(u_i, u_j)|u_i, u_j \in U\}$ is a set of edges indicating u_j replied to u_i. Based on G, we list several formal definitions as follows:

*DEFINITION 1. [**Vocabulary**] All distinct words from M constitute a vocabulary $W = \{w_i\}_{i=1}^{X}$. According to the word property, we further define noun vocabulary $W_N = \{n_i\}_{i=1}^{N}$ where n_i is a noun and opinion vocabulary $W_O = \{ow_i\}_{i=1}^{Q}$ where ow_j is an adjective or verb. The intuition is that a noun represents a topic while an adjective or verb indicates an opinion of the noun*

*DEFINITION 2. [**Opinion Words**] In a sentence, the opinion of a noun is often expressed by verbs or adjective. E.g. I **like** iphone4, Adele is a **marvelous** singer. Such words are called **opinion words**. We use $O(n_i)$ to denote the opinion word of a noun n_i and $O(n_i) \in W_O$.*

*DEFINITION 3. [**Topic-Noun Distribution**] An object contains several conceptually related topics $T = \{t_i\}_{i=1}^{K}$ and each topic is defined as a multinomial distribution over W_N. We define a topic-noun distribution $\mathbf{\Phi} = \{\phi_{ij}\}_{K \times N}$ where ϕ_{ij} denotes the probability that noun n_j is selected given topic t_i.*

*DEFINITION 4. [**User-Topic Distribution**] Different users have different preference over a set of topics T. We define a user-topic distribution $\mathbf{\Theta} = \{\theta_{ij}\}_{V \times K}$ where θ_{ij} denotes the probability that topic t_i is selected given user u_i.*

*DEFINITION 5. [**Topic-User-Opinion Distribution**] Different users show different opinions towards the same topic. We define a topic-user-opinion distribution $\mathbf{\Psi} = \{\psi_{i,j}^{k}\}_{K \times V \times 2}$ where $\psi_{i,j}^{k}$ denotes the probability that user u_i prefers opinion o_j given topic t_k and $o_j \in \{-1, +1\}$.*

*DEFINITION 6. [**Topic Opinion Neighbors**] For user u_i, all users that u_i replied to regarding to topic t_k constitute a set $ON(u_i, t_k)$ which is called u_i's topic opinion neighbors around t_k. Each user $u_j \in ON(u_i, t_k)$ can influence u_i's opinion of t_k.*

*DEFINITION 7. [**Topic-Opinion Influence**] For any $u_j \in ON(u_i, t_k)$, the influence of u_j on u_i can be measured by $\mathbf{\Omega} = \{\omega_{i,j,agree}^{k}\}_{K \times V \times V \times 2} \cup \{\omega_{i,j,disagree}^{k}\}_{K \times V \times V \times 2}$ where $\omega_{i,j,agree}^{k}$ denotes the probability that u_i agrees with u_j's opinion and $\omega_{i,j,disagree}^{k}$ denotes the probability that u_i disagrees with u_j's opinion on topic t_k.*

The most important four parameters are Θ, Φ, Ψ and Ω, which bind the user, topic, opinion and influence in a unified framework. Our task can be reduced to the following two steps:

- First, given G, how to estimate Θ, Φ, Ψ and Ω?
- Second, given user u_i and topic t_j, if Θ, Φ, Ψ and Ω are known, how to predict u_i's opinion of t_j if u_i post or comment on a new message?

3. MODEL DESCRIPTION

3.1 Sentiment Analysis

3.1.1 Message-level sentiment

Message level sentiment analysis is to capture the opinion word $O(n_i)$ for a noun n_i and judge the polarity of $O(n_i)$ in the context of a message. First, a parse tree is constructed to exhibit the syntactic structure of a sentence and dependency relations between words. Consequently, $O(n_i)$ can be spotted by analyzing the structure of parse tree. Second,

the polarity of $O(n_i)$ is judged by searching a corpus of Chinese sentimental words lexicon provided by Tsinghua NLP group, which consists of 5,567 positive and 4,479 negative words. Besides, two additional rules are applied to capture the real sentimental relation: One, whether there exists negation word, like *not, don't*, etc.; Two, whether there exists *adversative relation* between n_i and $O(n_i)$, like *but, however*, etc.

Based on our experiment, the number of $n_i - O(n_i)$ pairs are usually small, due to the short and noisy feature of microblog messages. In order to overcome the limitation of data sparsity, we consider the statistical co-occrrence relations from all messages we collected. For each distinct noun $n_i \in W_N$ we find out all adjectives/verbs $ow_i \in W_O$ that co-occur with n_i in all messages and pick out the top 20 most frequent co-occurrent $ow_1, ..., ow_{20}$, which constitutes a set $OS(n_j)$. For each $ow_j \in OS(n_j)$, we define a *statistical dependence(SD)*:

$$SD(n_i, ow_j) = \frac{CO(n_i, ow_j)}{AVEDIS(n_i, ow_j)}, j = 1, ..., 20 \quad (1)$$

where $CO(n_i, o_j)$ denotes the total number of co-occurent frequency of n_i and ow_j, and $AVEDIS(n_i, ow_j)$ denotes the average distance of n_i and ow_j in all their co-occurrent messages. Then, given a message, if $O(n_i)$ is not found for n_i through parse tree, we can calculate $SD(n_i, ow_j)$ as is shown in Equation 1 and finally obtain a $O(n_i)$:

$$O(n_i) = \underset{ow_j \in OS(n_j)}{Argmax}\ SD(n_i, ow_j) \quad (2)$$

3.1.2 User-level sentiment

User-level opinion regarding a topic can be easily obtained through aggregation of all message-level opinion records. We define two counters $C^k_{i,+1}$ and $C^k_{i,-1}, i = 1, ..., V, k = 1, ..., K$ to record the number of times that user u_i express positive or negative opinions towards topic t_k by scanning all u_i's message. Then Ψ can be estimated as:

$$\psi^k_{i,+1} = \frac{C^k_{i,+1}}{C^k_{i,+1} + C^k_{i,-1}}, \psi^k_{i,-1} = \frac{C^k_{i,-1}}{C^k_{i,+1} + C^k_{i,-1}} \quad (3)$$

In addition, we define another two counters $C^k_{i,j,agree}$ and $C^k_{i,j,disagree}$ to record the number of times u_i and u_j agree or disagree on topic k by scanning all their "post-reply" messages. Then Ω can be estimated as:

$$\begin{aligned} \omega^k_{i,j,agree} &= \frac{C^k_{i,j,agree}}{C^k_{i,j,agree} + C^k_{i,j,disagree}}, \\ \omega^k_{i,j,disagree} &= \frac{C^k_{i,j,disagree}}{C^k_{i,j,agree} + C^k_{i,j,disagree}} \end{aligned} \quad (4)$$

The strength of tie is also important to determine the opinion influence from neighbors, regardless of positive or negative influence. Especially, for $u_i \in ON(u_j, t_k)$, we calculate the strength of relation by:

$$\begin{aligned} s^k_{i,j,agree} &= \frac{C^k_{i,j,agree}}{\sum_{u_i \in ON(u_j,t_k)} C^k_{i,j,agree}}, \\ s^k_{i,j,disagree} &= \frac{C^k_{i,j,disagree}}{\sum_{u_i \in ON(u_j,t_k)} C^k_{i,j,disagree}} \end{aligned} \quad (5)$$

In many cases, given a pair u_i and u_j, though both of their opinions can be detected, their agreement could not be judged, for example, A supports object X while B supports Y on the same topic Z, if X and Y are opposite, then A disagrees with B, else, A agrees with B. To solve the problem, a simple competitive vocabulary corpus is generated by applying Topic Models and manual annotation, at last, 2,104 entity pairs are found, we name the data set as CoE; if object X and Y are consistent, then $CoE(X, Y) = 1$, if A and B are opposite, $CoE(X, Y) = 0$.

Besides, for many cases that objects X and Y can not be detected from CoE, we need to utilize other information in addition to the content of their messages. According to previous studies [3] [11], A metric *Opinion Agreement Index (OAI)* is introduced to quantify the influence of u_i on u_j:

$$\begin{aligned} OAI(u_i, u_j) &= a \cdot Influence(u_i) + b \cdot Tightness(u_i, u_j) \\ &+ c \cdot Similarity(u_i, u_j) \end{aligned} \quad (6)$$

where $Influence(u_i)$ is a function of the number of u_i's followers, $Tightness(u_i, u_j)$ is a function of the frequency of interactions (i.e. replying) between u_i and u_j, and $Similarity(u_i, u_j)$ is a function of the cosine similarity between $\theta_{i,*}$ and $\theta_{j,*}$. a, b and c are assigned as 0.6, 0.3 and 0.1 based on empirical knowledge, respectively.

If u_j replies to u_i's one message and their opinion agreement can not be determined, then $OAI(u_i, u_j)$ can be used to approximate the probability that u_j agrees with u_i in this replying action on a certain topic; this method combines additional attributes into Topic level Opinion Influence model, which could help to improve the performances.

3.2 Gibbs Sampling

We use the gibbs sampling to estimate Θ and Φ, with two prior hyperparameters α and β, respectively. Assuming that u_i posted a message and u_j replied to u_i by adding a comment. If the lth noun found in u_i's message is n_h, we sampled a topic for u_i based on Equation 8.

$$\begin{aligned} &P(z^l = t_k | x = u_i, w = n_h, \mathbf{z}^{-\mathbf{l}}) \propto \\ &\frac{C^{-l}_{xz} + \alpha}{\sum_{z \in T} C^{-l}_{xz} + K\alpha} \frac{C^{-l}_{zw} + \beta}{\sum_{w \in W_N} C^{-l}_{zw} + N\beta} \end{aligned} \quad (7)$$

where $z^l = t_k$ denotes the assignment of the lth noun in to topic t_k and $\mathbf{z}^{-\mathbf{l}}$ denotes all topic assignments not including n_h. C^{-l}_{xz} and C^{-l}_{zw} denote the number of times topic z is assigned to user x, and noun w is assigned to topic z respectively, not including the current assignment for the lth noun. For user u_j, if n_h also occurs in u_j's replying message, n_h is also assigned to topic t_k and t_k is assigned to user u_j. For all other nouns in u_j's replying message, the assignment of words and topics are the executed as the same probability as shown in Equation 8. The final Θ and Φ can be estimated by:

$$\theta_{xz} = \frac{C_{xz} + \alpha}{\sum_{z \in T} C_{xz} + K\alpha}, \phi_{zw} = \frac{C_{zw} + \alpha}{\sum_{w \in W_N} C_{zw} + N\beta} \quad (8)$$

3.3 Opinion Prediction

Our ultimate goal is to predict a user's opinion about a topic given his/her own opinion preference and his/her neighbor's opinion. First, we need estimate four parameters Θ, Φ, Ψ and Ω. A gibbs-sampling based parameter estimation algorithm is proposed, where topic modeling, sentiment analysis and influence analysis are interwoven together. Note that Pre1 to Pre5 should be executed before entering

Table 1: Summary of experimental data

	# of post messages	# of reply messages	# of users
Total	2,350,372	959,918	145,327
O_1	320,176	114,382	24,382
O_2	591,433	243,876	31,432
O_3	742,853	298,764	38,796

the loop. In each iteration, Gibbs sampling is used to assign nouns to topics and topics to users, and parse tree and $NOAI$ is used to detect the opinion polarity. When the iteration is done, the four parameters are calculated.

Based on the learning results, we would like to predict users' opinion towards some object with different topic distributions (eg, a new movie, the trend of stock price, a famous person et al.). Two factors are taken into consideration for opinion prediction. First, the historical records of topic preference and opinion distribution learned from TOIM; Second, the historical opinions of neighbors and their influence type and strength learned from TOIM. The prediction result is sampled from a sum-of-weighted probability combing the two factor together as a random process. Details are omitted due to the space limitation.

4. EXPERIMENTATION

4.1 Experiment Setup

The whole data set from Tencent Weibo is crawled from Oct 07, 2011 to Jan 05, 2012, which contains about 40,000,000 daily messages. Three objects that are popular among Tencent Weibo during the 3 months are selected: *Muammar Gaddafi, The Flowers of War (chinese movie),Chinese economics*, which are denoted by O_1 to O_3. The statistics are summarized in Table 1:

For each object, all messages are ranked based on temporal order and the last 200 hundred are selected as testing data. Then we have total number of 1,000 messages as testing data. The rest messages are used for training.

4.2 Prediction Performance

Three algorithms are selected as baseline methods for comparison with TOIM: SVM (Support Vector Machine), CRF (Conditional Random Field) and JST (Joint Sentiment Topic). SVM and CRF are supervised algorithms, we use parse tree and opinion detection technology to auto label micro-blogs with standard grammar structures as train data, which totally contributes 5,746 micro-blogs; the attributes include user name, key words and their grammar related words, such as verbs, adjectives, topic information. None of the above three algorithms consider the topic-level social influences as TOIM does. We also apply Map-Reduce strategy to improve the efficiency of TOIM. The *precision* is used to compare the prediction performance for all four algorithms. Figure 2 shows that the precision of TOIM correlates with the number of topics. Specifically, the precision rises as the the number of topics becomes larger, with the maximum value around 75%. By contrast the precisions of the other three algorithms are lower than TIOM and do not exhibit correlation with the number of topics.

4.3 Qualitative Case Study

TOIM can be applied to analyze users' opinions on two different levels: microscopic level and macroscopic level. In the microscopic level, individual users' opinion distribution over different topics can be observed and the opinion propagation from a seed user to all other neighbouring users can be visualized; active opinion leaders from different topics could also be detected during that process, for example,user 1516940869 is a battlefield journalist, and often reports the newest global event from spot at the first time, which attracts 957,444 followers, he also opens a special column to discuss with his followers. In macroscopic level, TOIM could detect opinion of public, as can be seen in Figure 3,

Figure 3 shows the relation between public opinion and Chinese economic. Specifically, Figure 3(a) exhibits the positive/negative opinion distribution of random selected 1,000 active users over the economics topic under O3. Obviously, many users are more concerned about development of Chinese economics, although China has achieved great economic success. Such concern corresponds to some important problems of Chinese economics, like extremely high house price and larger gap between rich and poor. Figure 3(b) shows that the changes of all users' positive attitude toward the topic finance market under O3, has a high correlation with China Hushen-300 Index (An important Financial Index in China) shown in Figure 3(c). It implies that the public opinion can reflect the real financial situation.

5. RELATED WORK

5.1 Topic Model

Since the introduction of LDA model [2], various extended LDA models have been used for topic extraction from large-scale corpora. Rosen-Zvi et al. [10] introduced the Author-Topic (AT) model, which to include author ship as a latent variable. Ramage et al. [9] 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 [8].

5.2 Sentiment Analysis and Opinion Mining

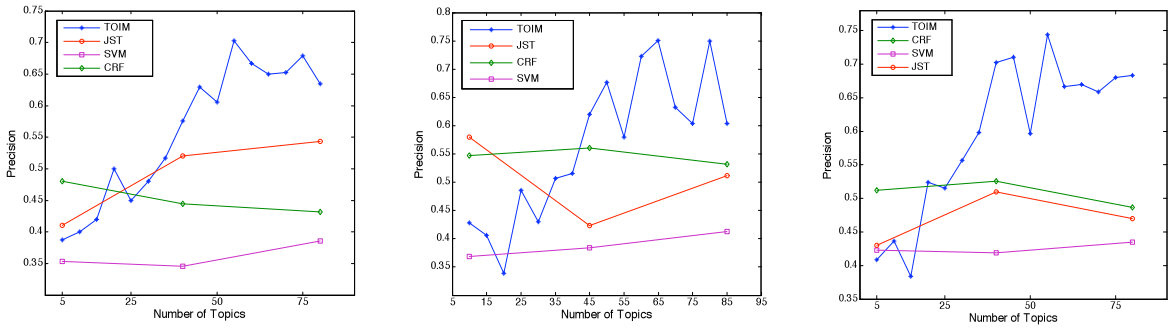
Most of related researches are mainly focused on identification of sentimental object [6], or detection of object's sentimental polarity [12] without considering the topic aspects. Mei et. al [8] and Lin et. al [5] incorporate topic models and sentiment analysis without considering the social influence. Our work attempts to integrate topic models, sentiment analysis and social influence into a two-stage probability model.

5.3 Social Influence Analysis

Social influence analysis is a hot research topic in social network research, including the existence of social influence [1], the influence maximization problem [4], the influence at topic level [7]. Those researches provide us a new perspective to investigate opinion mining from influence perspective.

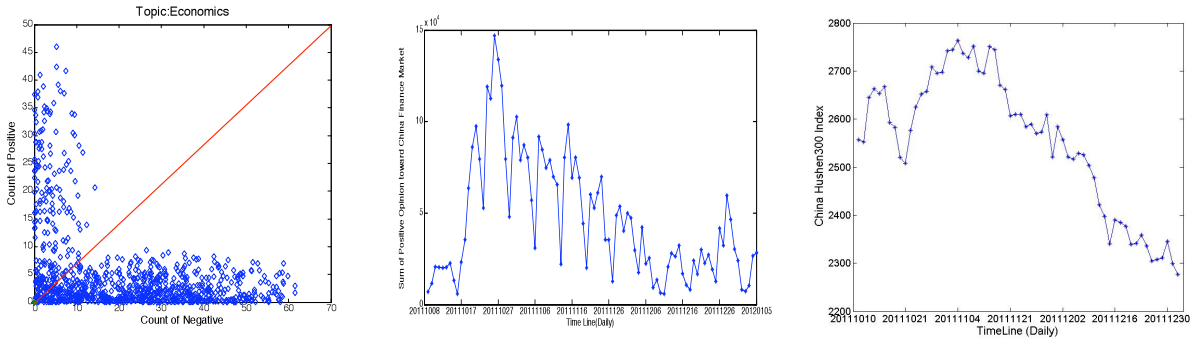
6. CONCLUSIONS

In this paper, we study a novel problem of social opinion influence on different topics in microblog. We proposed a Topic-level Opinion Influence Model (TOIM) to formalize this problem in a unified framework. Users' historical messages and social interaction records are leveraged by TOIM



(a) O_1 (b) O_2 (c) O_3

Figure 2: Opinion Prediction of O_1, O_2, O_3



(a) opinion distribution (b) positive opinion trend (c) financial index trend

Figure 3: Correlation between collective opinions and economic activity.

to construct their historical opinions and neighbors' opinion influence through a statistical learning process, which can be further utilized to predict users' future opinions towards some specific topic. Gibbs sampling method is introduced to train the model and estimate parameters. We experimented on Tencent Weibo and the results show that the proposed TIOM can effectively model social influence and topic simultaneously and clearly outperforms baseline methods for opinion prediction.

7. ACKNOWLEDGEMENT

This paper is supported by China Post Doc Funding (2012-M510027). He Gaoji Project, Tencent Company (No. 2011ZX-01042-001-002). The National Natural Science Foundation of China (NSFC Program No.71072037)

8. REFERENCES

[1] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. KDD '08, pages 7–15, 2008.

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

[3] P. H. C. Guerra, A. Veloso, W. M. Jr., and V. Almeida. From bias to opinion: a transfer-learning approach to real-time sentiment analysis. KDD '11, pages 150–158, 2011.

[4] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. KDD '03, pages 137–146, 2003.

[5] C. Lin and Y. He. Joint sentiment/topic model for sentiment analysis. CIKM '09, pages 375–384, 2009.

[6] H. LIU, Y. ZHAO, B. QIN, and T. LIU. Comment target extraction and sentiment classification. *Journal of Chinese Information Processing*, 24:84–88, 2010.

[7] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang. Mining topic-level influence in heterogeneous networks. CIKM '10, pages 199–208, 2010.

[8] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. WWW '07, pages 171–180, 2007.

[9] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. EMNLP '09, pages 248–256, 2009.

[10] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. UAI '04, pages 487–494, 2004.

[11] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li. User-level sentiment analysis incorporating social networks. KDD '11, pages 1397–1405, 2011.

[12] Z. Zhai, B. Liu, H. Xu, and P. Jia. Constrained lda for grouping product features in opinion mining. PAKDD'11, pages 448–459, 2011.