Topic-Level Sentiment Analysis in Social Networks with Pair-wise User Influence

JIAQI WANG
School of Foreign Languages and Cultures
Nanjing Normal University
CHINA
jqw413@126.com

Abstract: Inspired by the principle of Homophily which suggests that opinions are influenced by connection, we introduce relations into sentiment analysis in the context of social networks, which also helps to reduce the content sparsity by utilizing the networked SNS data. We propose a model which utilized textual content and link structure simultaneously to evaluate pair-wise social influence on topic level between users. The framework depicts the topic distribution for each user by LDA based on text information; and model the pair-wise influence between users on topic level by measuring their centralities and interactive weights. The learned influence is then applied into sentiment classification as supplementary features. The experiment results on two datasets show that the model incorporating user relations outperforms the methods which based on textual features only.

Key–Words: Sentiment Analysis, Social Network, LDA, Topic Model

1 Introduction

Social network sites (SNS) such as micro blog and BBS are popular among netizen for their freedom and convenient interaction. SNS break the patterns in which public opinions are monopolized by television and newspaper. They also help the individuals to become the information transmitters. Meanwhile, as propelled by the growth of opinion data, it is urgent to utilize automated tools to monitor the user relationship and topic trend in social networks. As SNS data is short, non-normative and networked, it is not recommended to directly utilize the methods which are traditionally used in text.

Social networks exhibit small-world network characteristics. The hub users with high number of connections greatly promote the information spreading and orientation by retweeting and commenting the posts. In this paper we introduce social relations into user sentiment analysis by quantifying pair-wise influence between users. We do this for two reasons. First, relations in social networks are easy to obtain and they help to reduce content sparsity. Second, homophily [1] holds the opinion that *Similarity breeds connection* and people's personal networks are homogeneous with regard to many social behaviors and intrapersonal characteristics [2].

At present, many studies interpret a user's influence as its node in-degree in the network and ignore the fact that it is the interest that affects the way users influence one another. Users' interest varies in different topics. For example, a machine learning expert A can have high influence on his follower B on topic ”Machine Learning” while what he says on ”economics” or ”politics” may be discounted. Identifying influential users on specific topics will benefit the study.

In this paper, we detect the users' sentiment orientation in social networks on topic-level with pair-wise influence. For each user, the text information and link structure are given; the purpose is to assign a sentiment label to the user on a specific topic. We first validate that connections and shared opinions tend to co-occur in our dataset. Then we study about the details of the model which evaluates the pair-wise influence between users. The motivation is that SNS data is short and non-normative while the link structure is easy to obtain and also help to reduce the content sparsity. Then model is applied to sentiment classification and the results show that social relations are helpful to sentiment analysis.

2 Related Work

In this paper, we introduce social relationship into sentiment analysis by quantifying pair-wise user influence on topic-level. To some extent, social network sites can be regarded as map of real human society. Every node is embedded in relation networks and user influence evaluation can be regarded as node ranking.

Many researchers measure a user's influence or social status as Centrality. PageRank [3] is a random surfer model. At the beginning the PR value of each node is set to the same initialization value. In each round, PR is updated by repeatedly dividing current PR among its forward links evenly and summing up PR of its back links. PageRank takes the number of links and the quality of nodes into consideration simultaneously and protect ranking from noise. TwitterRank [4] is proposed to find influential users in twitter. The model first applies LDA to identify latent topic distribution and calculate the similarity for users on specific topic. Then it evaluates the influence by capturing link structure and number of posts. Jing Zhang [5] gives a formal definition of "social influence locality" which denoted as $Q(S_v, G_v^\tau)$, $G_v^\tau$is user v's $\tau - ego$ network, S_v is the collection of active neighbors in G_v^τ. Then measure pair-wise influence with random walk and structural influence with connected circles; finally use the model to predict users' retweet behaviors by training a classifier based on the defined functions.

The mentioned methods above can effectively measure user (node) influence. However, the influence of a user on his friends about a specific topic has been largely ignored [6]. In social networks, users form circles when share similar interests and their influence vary greatly in different topics.

Topic-sensitive PageRank [7] first computes general PageRank values for each node; then modifies the results according to topic correlation. The model can generate more accurate rankings than with a single general PageRank. Topical Affinity Propagation (TAP) [8] models the social influence in large networks on topic-level. In particular, TAP takes the results of topic model LDA and the network structure to perform topic-level influence propagation. Chenhao Tan [9] first shows that "links and shared sentiment are clearly correlated"; then defines a factor-graph-based model based on a given topic q, which believes that a user's sentiment is influenced by the sentiment labels of his tweets and of his neighbors. FLDA [10] is a Bernoulli-Multinomial mixture model which contains two levels of mixtures: an upper-level Bernoulli mixture with one of the components being a Multinomial mixture.

Inspired by the models above, taking both homophily and the observations of our datasets into consideration, we consider proposing a model which learns the topic distribution and social relations jointly and also improve the sentiment classification tasks.

3 Observations

We first study that given a topic q, does there exits a correlation between network structure and user opinion. The analysis is conducted on two datasets: *xiciEdu*[1] and *sinaWeibo*[2].

Homophily is a phenomenon that people's social networks are homogeneous with regard to many socio-demographic, behavioral and intrapersonal characteristics[4, 2]. People tend to communicate with those who share the same action and attitude. That is "similarity breeds connection"[1].

Although built on virtual accounts, social network sites can be regarded as map of real human society. We adapt [9]'s work into our datasets and conduct a statistical analysis to study that whether homophily presents in the context of social network.

We chose 1000 pairs of linked users and 1000 pairs of randomly selected users respectively from *sinaWeibo* and in *xiciEdu* the number is 500. In *sinaWeibo*, user can choose who he wants to follow without requiring a permission first, so "link" is defined as "follow"; In *xiciEdu*, user finds the posts which they are interested in and replies back, there is no explicit "follow" relationship, so we define "link" as reply. We calculate the sentiment consistency on three randomly selected topics which distilled by LDA. In *xiciEdu* the three topics are: "family education", "school choosing" and "extra-curricular training"; in *sinaWeibo* the three topics are: "foreign affairs", "wearable devices" and "military". Details about topic extraction will be discussed in 4.

Table 1 shows that in our dataset, linked users have higher sentiment consistency on topic level. In the table, "link" represents pairs formed by connected users; "random" represents pairs formed by randomly selected users. The result suggests that connected users are more likely to hold similar opinions than randomly selected users. This verifies the statement of Hatfield that the sentiment orientation of two messages posted by friends are more likely to be similar than those randomly selected messages[11].

Table 1: Topic Level Sentiment Consistency on Two Types of Connections

Type	***xiciEdu***			***sinaWeibo***		
	Topic1	Topic2	Topic3	Topic1	Topic2	Topic3
Link	0.562	0.566	0.538	0.556	0.502	0.548
Random	0.5	0.498	0.462	0.506	0.46	0.498

Table 2 shows that among all connected pairs, the

[1] crawled from http://www.xici.net.

[2] crawled by http://bigdataopc.ihep.ac.cn.

ratio of users who share the sentiment with their partners is larger than those who hold different opinions.

Table 2: Topic Level Sentiment Consistency on Linked Users

	xiciEdu	*sinaWeibo*
Topic1	0.546	0.541
Topic2	0.517	0.554
Topic3	0.488	0.543

Our study verifies the intuition that in our dataset users' sentiment orientation is correlated with link structure. So we consider introducing homophily into sentiment analysis by exploiting social relations and text information simultaneously on topic level.

4 Measuring Pair-wise User Influence on Topic-level

In this section, we propose a model which measures pair-wise social influence between users on topic-level. For each pair of users i and j, the model evaluates user i's influence on user j in topic k by incorporating topic model and user relations.

Problem Description: A panorama $G_z = \{V, E, P_{v_i} | v_i \in V\}$ on topic z is formed with the text information and the link structure of the users in the dataset. V is a set of users. The edge set $E \subseteq V \times V$ denotes the connections between users. P_{v_i} is the text information posted by user v_i. For each user $v_i \in V$, our task is to get a ranked list R of influential users, item R_{v_j} in the list represents the user v_j's influence on user v_i in topic z and the items are sorted by the influence.

Topic-level social influence measurement is composed of two steps: (1) topic distribution distilling; (2) influence weight evaluation. First we identify the topic distribution by LDA for each user, and then estimate the pair-wise influence weight between users. The influence weight that user j exerts on user i in topic z is evaluated from two aspects: the centrality of user j in topic z which denoted by $UR_z(j)$, and interactive strength between user j and i which denoted as $UI_z(i,j)$.

4.1 Topic Extraction

Topic extraction can automatically recognize the topics that users are interested in. We apply LDA to assign a topic to each word so as to depict the topic distribution for each user and then analyze user interest.

LDA is a three-level statistical model which uses a "bag of words" assumption. It captures the intuition that each document exhibits the topics in different proportion, which is denoted as θ; each topic is associated with a multinomial distribution over a fixed vocabulary, which is denoted as φ[12, 13, 14, 15]. Furthermore, topic model is unsupervised and do not require any prior annotation of the corpus.

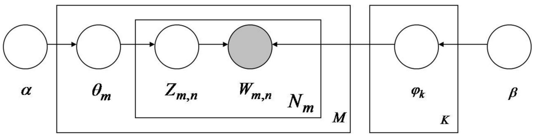


Figure 1: LDA Model

More formally, M is the number of documents in dataset, V is number of distinct words in vocabulary, K is the number of topics, N_m is the length of m_{th} document, φ_k is the word distribution for topic k, θ_m is the topic distribution for document m. The procedure of generating a word w of document m boils down to two stages. For each word w in each document m, first a topic z is drawn from the multinomial distribution θ_m, and then word w is sampled from the multinomial distribution $\varphi_{z_{m,n}}$. The definition of the generation process of LDA is shown as follows.

- For each topic $k \in [1, K]$,draw $\varphi_k \sim Dir(\beta)$
- For each document $m \in [1, M]$,
    1. Draw $\theta_m \sim Dir(\alpha)$
    2. For each word $n \in [1, N_m]$ in document m,
        (a) Draw $Z_{m,n} \sim Mult(\theta_m)$
        (b) Draw $w_{m,n} \sim Mult(\varphi_{z_{m,n}})$

Learning from [4, 16, 17, 18, 19, 20], the results of topic distilling is shown as follows:

- ○ matrix $UK(U \times K)$,where U is the number of documents and K is the number of topics. $UK_{i,j}$ is the proportion of topic j in document i.
- ○ matrix $KW(K \times W)$,where K is the number of topics and W is the number of distinct words. $KW_{i,j}$ is the proportion of word j in topic i.
- ○ vector $Z_m(1 \times N_m)$,where N_m is the length of m_{th} document. $z_m(i)$ is the topic assignment of word i in document m.

4.2 Pair-wise User Influence Evaluation on Topic Level

We measure user influence on topic-level under the principle of Homophily. Define $w_z(i,j)$ as user i's influence on user j in topic z. Incorporating text information, $w_z(i,j)$ is evaluated concerning two aspects: the centrality of user j in topic z and interactive strength between user j and i in topic z.

Definition 1 *Given topic z, the centrality of user i is defined as:*

$$UR_z(i) = \lambda \cdot norm(fan\_num(i)) \\ +(1-\lambda)[\theta_z(i) \cdot ConLength(i) \cdot norm(UW_z(i))] \quad (1)$$

All other things be equal, a user has high influence if his centrality value is high. Definition 1 is composed of two parts. $norm(fan\_num(i))$ is the outline of the user, it normalizes user i's fans number and reflects the amount of attention that user i gets in social networks. The second part captures two notions: $\theta_z(i) \cdot ConLength(i)$ denotes the normalized user's topic-level authority which calculated by the amount of content that associated with topic z, and $norm(UW(i))$ is the sum of influence of i's followers. We adapted [3]'s idea to the context of social network. $UW(i)$ can be calculated iteratively by:

$$UW(i) = \sum_{j followsi} \frac{UW(j)}{C(j)} \quad (2)$$

$C(j)$ is the number of users that user j follows. $UW(j)$ is divided among j's followees evenly to contribute to the rank of user i.

Definition 2 *Given topic z, the centrality of user i is defined as:*

$$UI_z(i,j) = norm(\alpha_{i,j} \cdot contentWeight(i) \cdot sim_z(i,j)) \quad (3)$$

We adapt [4]'s formulation of TwitterRank to evaluate pair-wise interactive strength in this paper. When other factors are not taken into account, user i has high influence on user j if their interactive strength is high. Definition 2 captures the intuition that user i's influence on j is determined by the number of times that i communicates with j which is denoted as $\alpha i, j$; on the other hand, the influence is high if the proportion of content that j received from i is high; at the same time, topic similarity between i and j in topic z which is denoted by $sim_z(i,j)$ also contributes to the influence measurement.

$weiboCount(i)$ is the amount of text published by user i. $\sum weiboCount(k)$ is the sum of contents published by all the users that j follows. $contentWeight(i)$ captures the intuition that the more user i publishes, the higher his influence on j is, because j reads much from i. For example, user a follows user b and c, b and c publish 20 and 30 messages respectively. All other things be equal, c's influence on a is 1.5 times of that of b.

$$contentWeight(i) = \frac{weiboCount(i)}{\sum_{j followsk} weiboCount(k)} \quad (4)$$

Homophily suggests that $UI_z(i,j)$ is also related with topical similarity $sim_z(i,j)$. Users have different interests in different topics. $sim_z(i,j)$ is defined as:

$$sim_z(i,j) = 1 - |DT(i,z) - DT(j,z)| \quad (5)$$

DT is a row-normalized matrix, it represents user's topic distribution which extracted by LDA. $DT(j,z)$ is the probability of user j's interest in topic z. We evaluate the similarity between i and j in topic z by comparing the probability that two users are interested in the topic z.

Definition 3 *Above all, user i's pair-wise influence on user j in topic z is defined as:*

$$w_z(i,j) = UR_z(i) \cdot UI_z(i,j) \quad (6)$$

$UR_z(i)$ is user i's centrality in topic z and $UI_z(i,j)$ is the interactive strength between user i and user j in topic z.

4.3 Algorithm Description

The framework is illustrated as Algorithm1. First LDA is applied to automatically recognize the topic distribution θ; then $UW_z(i)$ of each node i is initialized to 1. In each round, $UW_z(j)$ is updated iteratively to convergence. Last, we calculate $w_z(i,j)$ by evaluating $UR_z(i)$ and $UI_z(i,j)$.

5 Sentiment Classification

In this section, we introduce the learned influence into sentiment classification as supplementary features.

Sentiment orientation prediction can be considered as a classification problem: given the piece of text of user u, a relation network G on topic z, the goal is to assign a sentiment label $s_{u,z}$ to user u on topic z, $s_{u,z} = 1$ represents that u holds positive opinion on topic z and vice versa. Unlike the traditional text classification problem which only based on text, in SNS users are influenced by their neighborhood, so it is expected that incorporating relation networks can improve the sentiment classification.

Algorithm 1 Pair-wise Social Influence on Topic Level

```
Input:
    panorama $G_z = \{V, E, P_{v_i} | v_i \in V\}$
    user topic distribution $\theta$
    number of iterators $nIters$
Output:
    pair-wise topic-level user influence Matrix $W_z$
$UW_z \leftarrow 1; iIters \leftarrow 0$
while $iIters < nIters$ do
    for all users $j \in V$ do
        $UW_z(j) = \sum_{kfollowsj} \frac{UW_z(k)}{C(k)}$
        $iIters \leftarrow iIters + 1$
    end for
end while
for all users $i \in V$ do
    $UR_z(i) = \lambda \cdot norm(fan\_num(i)) + (1 - \lambda)[\theta_z(i) \cdot ConLength(i) \cdot norm(UW_z(i))]$
end for
for all edges $e(i,j) \in E$ do
    $UI_z(i,j) = \alpha_{i,j} \cdot \frac{weiboCount(i)}{\sum_{jfollowsk} weiboCount(k)} \cdot sim_z(i,j)$
end for
for user $j \in V$ do
    for user $i$ who $j$ follows do
        $w_z(i,j) = UR_z(i) \cdot UI_z(i,j)$
    end for
end for
```

Our model takes the user relationship and text information into consideration simultaneously. The feature vector can be constructed as the combination of text and relationship:

$$\boldsymbol{f}_i = (f_1, f_2, \ldots, f_V, w_1, w_2, \ldots, w_n) \qquad (7)$$

Items $f_1, f_2, \ldots, f_V$ are features calculated by text, while $w_1, w_2, \ldots, w_n$ are calculated based on the top n most influential users on user i. Pair-wise influence between users is used as feature expansion for vectors. The advantage of this method is that we can integrate text and relationship information into one model conveniently. We run classifications on vectors which only based on text and on vectors which incorporating user relations, in the last we compare their effectiveness.

6 Experiments

In this section, several groups of experiments are presented. We first analyze the users' behavioral features in our datasets and discuss the correlation between opinions and connections; then depict topic distribution for each user by LDA; calculate pair-wise user influence on topic-level; introduce the learned influence into sentiment categorization as supplementary features. We finally evaluate the efficiency of the proposed approach.

6.1 Datasets

The two datasets used in the experiment are *xiciEdu* and *sinaWeibo*. Weibo is a microblog site on which users can publish and re-tweet like twitter. Xici is the earliest online community in China. Learning from TwitterRank, we first lay out the panoramic pictures of the two datasets, then analyze the distribution of the posts per user. The results are illustrated in Figure 2 and Figure 3.

xiciEdu We crawled 332660 posts from Xici.net between November 2010 and December 2011 and also 942290 replies. Totally we collected 126162 users and there are as much as 74698 users who even did not send any post, which means most users are onlookers and the very few active users lead the topic and sentiment tendency of the community. Specifically, we choose the forum "the way to middle school" of Xici as example and depict a panorama, which contains 1770 users, 8646 posts and 4445 relationships.

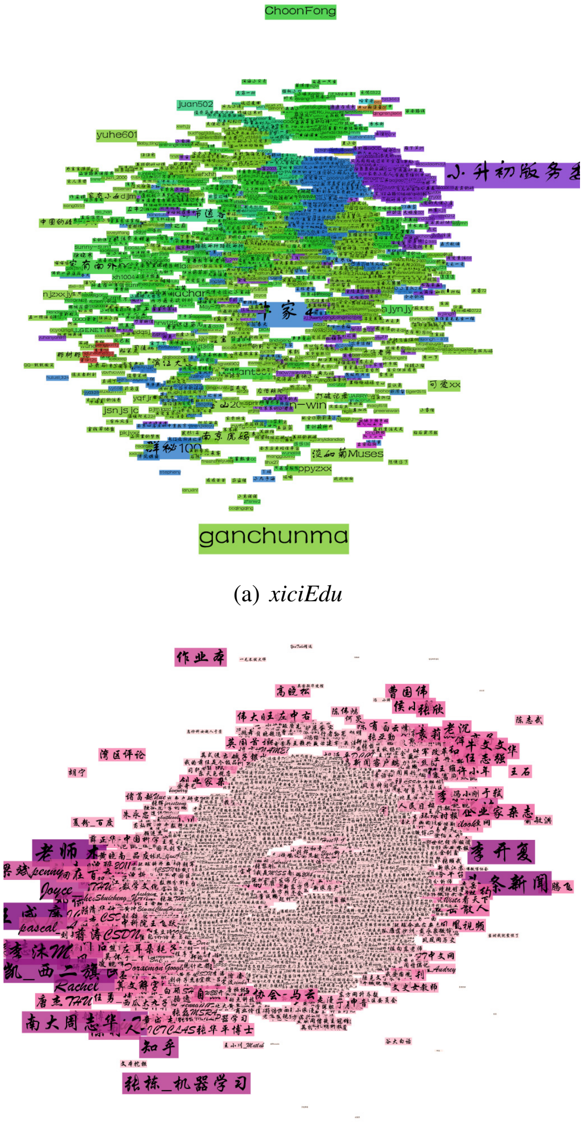
sinaWeibo To prepare the dataset, we obtained 100 influential users as seeds and crawled all the followees of each individual. We finally obtain a set of 3656 users and 4633 relationships by removing all the users whose post number is less than 500 and fans number less than 1000. For each user, we obtained messages he had posted between July 2013 and January 2014. Apparently users in *sinaWeibo* are clustered in many groups meanwhile they share the same central users. Then we depict the distribution of the posts per user. Most users published only very few messages and minority users published most of the messages.

6.2 Topic Extraction

This experiment is to extract topics via Latent Dirichlet Allocation. As discussed in section 4.2, the learned topics will be used to measure user influence.

For each user, we collect all the text he has posted as a document, which means users and documents are one to one. LDA is conditioned on three parameters, in this paper, we empirically set topic number $k = 50$, $\alpha = 50/k$,$\beta = 0.1$.

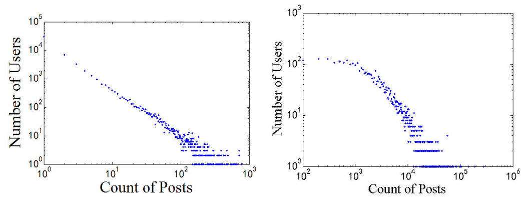
Table 3 and Table 4 show five topics respectively with their associated top words extracted from *xiciEdu* and *sinaWeibo*(in English). Topics are inferred by the order of the probabilities of words. For example, in Table 3, Topic1 refers to 'family education', Topic2 corresponds to 'junior school entrance examination', Topic4 corresponds to 'immigrant', and Topic 5 corresponds to 'selecting major and university'. In Table4, topic 1 corresponds to 'international affairs', topic2 corresponds to 'wearable devices' and topic5 corresponds to 'national defense'.



(a) *xiciEdu*

(b) *sinaWeibo*

Figure 2: Panoramas for Two Datasets



(a) *xiciEdu* (b) *sinaWeibo*

Figure 3: Distributions of the Posts per User on Two Datasets

Table 3: Topics extracted from *xiciEdu*

Topic1	Topic2	Topic3	Topic4	Topic5
parents	random	internet	parents	university
education	news	marketer	happiness	Nanjing
family	quota	charity	immigrant	college
father	cry	show	China	Communication
wolf	comfort	hype	foreigner	Normal
mother	sprint	society	Canada	Southeast
highschool	admitted	planning	persistent	Technology
home	fair	undercover	English	Polytechnic
establish	branch	scandal	study abroad	Telecom

Table 4: Topics extracted from *sinaWeibo*

Topic1	Topic2	Topic3	Topic4	Topic5
presentation	domestic	moksa	achieve	specify
mission	IOS	achieve	player	nextgeneration
GA	wearable	chant	MWC2014	Changé
place	questioning	help	NokiaWorld	Intercontinental
intern	ranking	bad	purchase	Chiefs
countries	topic	do devil	WP8	fire
properly	promising	monk	Surface	ball firing
consulate	second	wisdom	interface	warship
mansion	watch	epiphany	design	defend

In *xiciEdu*, posts are tightly correlated with the theme of the forum, which means users share the same interests, in here is ”education”. We obtain 1770 topic distributions during the experiment and plot 6 randomly selected ones. As shown in Fig.4, we find that the curve shapes of users are much different. The curve of user NO.180 peaks in Topic2 which implies that he talks much about Topic2 while user No.227 prefers Topic5. User NO.244 and NO.245 mainly post about Topic1. The phenomenon is a strong indicator of topic variation between users. We should differentiate the user influence in different topics while measuring pair-wise social influence.

6.3 Pair-wise User Influence Evaluation on Topic Level

In this experiment we focus on pair-wise user influence evaluation. We compare our method which is denoted as UR, with $normFansCount$ algorithm, which measures user influence by the normalized number of fans. The details of user influence are plotted in Figure5. In *sinaWeibo*, ”weibo Secretary”,

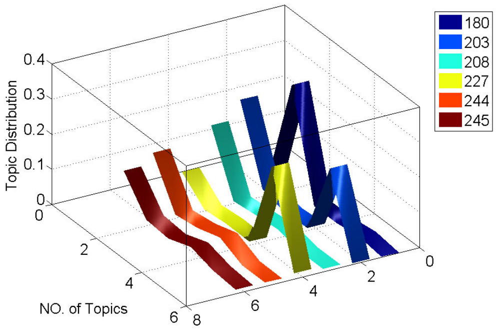
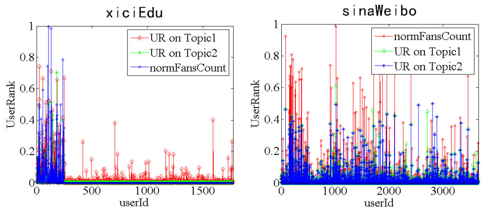


Figure 4: Topic Distributions of Six Random Selected Users in *xiciEdu*



(a) *XiciEdu* (b) *sinaWeibo*

Figure 5: User Ranks Evaluated by Different Methods

”Chen Kun” and ”Guo Degang” have highest number of followers, while the most influential users on Topic1 are ”SinaTech”, ”36Kr” and ”Android Market”. In *xiciEdu*, ”Niu Jiatun”, ”ganchunma” and ”The Executive Committee” are identified as the most influential users by $normFansCount$. The most influential users on Topic1 are ”bant007”, ”ganchunma” and ”zf219”.

As discussed above, Table 5 lists top 3 influential users of the user on specific topic. The results of our model are reasonable. For example, although ”Niu Jiatun” and ”ganchunma” have highest number of fans in *xiciEdu*, the highest score of UR on ”david” in topic ”family education” are ”jlsj200”, ”samBaby” and ”English teacher Chen”. The three users talks much about the topic.

Table 5: Examples of Influential Users

DataSet	Topic	user	top 3 influential users on the topic
xiciEdu	Family Education	david	jlsj200, samBaby, English teacher Cheng
	extra-curriculum	ChoonFong	dxllovelgw, oasisdew, cnwuhao
sinaWeibo	News Media	Jiang Shengyang	Phoenix video, Audrey, Xie Nan
	CS	Researcher July	Li Kaifu, Internet Matters, Jiang Tao of CSDN

6.4 Sentiment Prediction

Based on the results obtained during pair-wise user influence measurement, we introduce the learned influence into sentiment prediction in this section.

6.4.1 KNN-TEXT Classification on Social Datasets

We first conduct a comparison between *Stan_News*[3], *xiciEdu* and*sinaWeibo* with KNN-TEXT which means without any link structure features.

Table 6 lists the performance of KNN-TEXT on the three datasets. It is observed that the classifier performs well on *Stan_News* while *xiciEdu* and *sinaWeibo* have significantly lower Recall value. The phenomenon implies that the classifier works better when the text information is words-formal, classes-explicit and length-moderate, while social network text is colloquial, short and non-normative. What is more, in this paper the two classes are sentimental positive and negative, which are not as explicit as common classes. To get the most use of social relationship and reduce the content sparsity, we consider improving the effect of sentiment classification in social network by introducing pair-wise user influence.

Table 6: KNN-TEXT on Different Datasets

DataSet	Precition	Recall	F_Score
Stan_News	0.746	0.745	0.732
xiciEdu_Topic1	0.734	0.141	0.177
xiciEdu_Topic2	0.564	0.166	0.256
sinaWeibo_Topic1	0.619	0.168	0.262
sinaWeibo_Topic2	0.685	0.176	0.269

6.4.2 Comparisons between KNN-TEXT and KNN-NETWORK

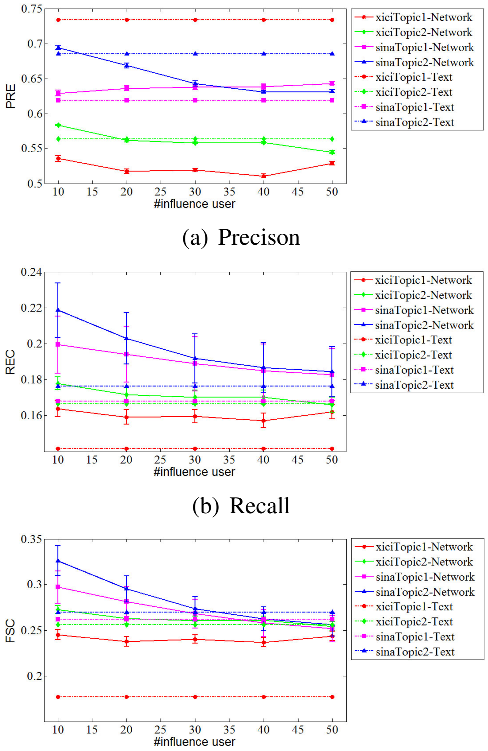
In this section, we construct features by taking into account both textual information and user relationship. We study that in the context of social media; can user relations improve sentiment classification? We utilize KNN as the classifier and compare the traditional text-based methods with the proposed network-based methods which incorporating user relationships.

- xiciTopici-Text & sinaTopici-Text: the method is a KNN classification which based on word features only. The experiment is conducted on *xiciEdu* and *sinaWeibo* on Topic i.
- xiciTopici-Network & sinaTopici-Network: the method is a KNN classification which incorporating text and relationship features. The experiment is conducted on *xiciEdu* and *sinaWeibo* on Topic i.

Experimental results are plotted in Fig.6. X-axis stands for the number of influential users which are introduced as supplementary features, Y-axis stands for the value of Precision, Recall and F_Score respectively. For example, the second data point on the blue line in Fig.b which denoted as ’sinaTopic2-Network’ implies that the recall value is 0.203 when the number of influential users is 20. We count the improvement of the proposed methods compared to the traditional methods and draw the following observations.

In 6.4.1, it is observed that text-based methods on *xiciEdu* and *sinaWeibo* have significantly lower Recall

[3] crawled by finallyliuyu, some of them was provided by www.cnblogs.com, Netease news center, tencent news center at no charge. The dataset covers eight genres including history, military, culture, reading, education, IT, entertainment, society and legal system. The training set size is 13026 and testing set size is 3254.



(a) Precison

(b) Recall

(c) F_score

Figure 6: comparisons between text-based and network-based methods

value compared with that on *Stan_News*. In this section, network-based methods have consistently higher recall than text-based methods on several different number of expended features except when n = 40 and 50 on sinaTopic2 and when n=50 on xiciTopic2. Fig.6 shows that exploiting social relations contributes to the improvement of recall and further contributes to the final F-Score as compared to the text-based methods.

6.4.3 Comparisons on Three Algorithms

Table 7: Comparisons on Three Algorithms

		xiciTopic1	xiciTopic2	sinaTopic1	sinaTopic2
KNN	text-based	0.734	0.563	0.618	0.684
	network-based	0.541 (-26.2%)	0.582 (+3.37%)	0.640 (+3.56%)	0.687 (+0.44%)
BPNN	text-based	0.542	0.544	0.483	0.499
	network-based	0.541 (-1.8%)	0.582 (+3.37%)	0.640 (+32.5%)	0.687 (+37.6%)
SVM	text-based	0.532	0.544	0.531	0.512
	network-based	0.537 (+9.40%)	0.546 (+0.37%)	0.531 (+0.0%)	0.512 (+0.0%)

To further evaluate our proposed framework, in this section we use KNN, BPNN and SVM as classifiers, each method is conducted on two types of features: text-based and network-based.

The performances of the methods are reported in Table 7. The classifications based on network outperform the text-based methods other than KNN on xiciTopic1. KNN-NETWORK achieves better performance on three other occasions. BPNN-NETWORK achieves better performance on xiciTopic2 and significantly better than BPNN-TEXT by more than thirty percent. SVM-NETWORK performs better on xiciTopic1 and no changes on *sinaWeibo*. The results indicate that in general user relations are helpful to sentiment analysis in our datasets.

7 Conclusion

In this paper, we focus on sentiment analysis in the context of social networks. It is not recommended to directly utilize the methods which traditionally used in text, for SNS data are short, semi-structured and non-normative. We consider improving the sentiment classification in social networks by taking both text information and relations into account. First we have linked user pairs and randomly selected user pairs to verify that sentiment similarity and connections are tend to co-occur in our dataset. Then we depict the topic distribution for each user by LDA to further understand users' interests. Based on the topics, we model User A's pair-wise influence on User B in topic T to be associated with User A's centrality in T and interactive weight between A and B in T. The learned influence is then applied to sentiment analysis as supplementary features. The experiment results verify the effectiveness of our model. For future work, the model can be implied to larger scale and different kinds of datasets.

Acknowledgements: This research was supported by National Natural Science Foundation of China (No. 41471371).

References:

[1] Miller McPherson, Lynn Smith-Lovin and James M Cook, Birds of a feather: Homophily in social networks, *Annual review of sociology,* 2001, pp. 415–444.

[2] Mike Thelwall, Emotion homophily in social net-work site messages, *First Monday,* 4, 2010.

[3] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd, The pagerank citation ranking: bringing order to the web, *Technical*

report, Stan-ford Digital Library Technologies Project, 1999.

[4] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He, Twitterrank: finding topic-sensitive influential twitterers, *Proceedings of the third ACM international conference on Web search and data mining,* 2010, pp. 261–270.

[5] Jing Zhang, Biao Liu, Jie Tang, Ting Chen, and Juanzi Li, Social influence locality for modeling retweeting behaviors, *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence, AAAI Press,* 2013, p-p. 2761–2767.

[6] Lu Liu, Jie Tang, Jiawei Han, Meng Jiang, and Shiqiang Yang, Mining topic-level influence in heterogeneous networks, *Proceedings of the 19th ACM international conference on Information and knowledge management,* 2010, p-p. 199–208.

[7] Taher H Haveliwala, Topic-sensitive pagerank, *Proceedings of the 11th international conference on World Wide Web,* 2002, pp. 517–526.

[8] Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang, Social influence analysis in large-scale network-s, *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining,* 2009, pp. 807–816.

[9] Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li, User-level sentimen-t analysis incorporating social networks, *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining,* 2011, pp. 1397–1405.

[10] Bin Bi, Yuanyuan Tian, Yannis Sismanis, Andrey Balmin and Junghoo Cho, "Scalable topic-specific influence analysis on microblogs, *Proceedings of the 7th ACM international conference on Web search and data mining,* 2014, p-p. 513–522.

[11] Elaine Hatfield, John T Cacioppo, and Richard L Rapson, Emotional contagion, *Cambridge university press,* 1994.

[12] David M Blei, Andrew Y Ng, and Michael I Jordan, Latent dirichlet allocation, *the Journal of machine learning research,* 3, 2003, pp. 993–1022.

[13] Thomas K Landauer, Peter W Foltz, and Darrell Laham, An introduction to latent semantic analysis, *Discourse processes,* 25, 1998, pp. 259–284.

[14] Thomas Hofmann, Probabilistic latent semantic indexing, *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval,* 1999, pp. 50–57.

[15] David M Blei, Probabilistic topic models, *Communications of the ACM,* 55, 2012, pp. 77–84.

[16] Thomas L Griffiths and Mark Steyvers, Finding scientific topics, *Proceedings of the National A-cademy of Sciences,* 101, 2004, pp. 5228–5235.

[17] Gregor Heinrich, Parameter estimation for text analysis, *Technical report,* 2005.

[18] Tom Griffiths, Gibbs sampling in the generative model of latent dirichlet allocation, 2002.

[19] GW Peters and SA Sisson, Bayesian inference, monte carlo sampling and operational risk, *Journal of Operational Risk,* 1, 2006, pp. 27–50.

[20] Walter R Gilks, Sylvia Richardson, and David JSpiegelhalter, Introducing markov chain monte carlo, *Markov chain Monte Carlo in practice,* 1996, pp. 1–19.