ActNeT: Active Learning for Networked Texts in Microblogging

Xia Hu* Jiliang Tang* Huiji Gao* Huan Liu*

Abstract

Supervised learning, e.g., classification, plays an important role in processing and organizing microblogging data. In microblogging, it is easy to mass vast quantities of unlabeled data, but would be costly to obtain labels, which are essential for supervised learning algorithms. In order to reduce the labeling cost, active learning is an effective way to select representative and informative instances to query for labels for improving the learned model. Different from traditional data in which the instances are assumed to be independent and identically distributed (i.i.d.), instances in microblogging are networked with each other. This presents both opportunities and challenges for applying active learning to microblogging data. Inspired by social correlation theories, we investigate whether social relations can help perform effective active learning on networked data. In this paper, we propose a novel Active learning framework for the classification of Networked Texts in microblogging (**ActNeT**). In particular, we study how to incorporate network information into text content modeling, and design strategies to select the most representative and informative instances from microblogging for labeling by taking advantage of social network structure. Experimental results on Twitter datasets show the benefit of incorporating network information in active learning and that the proposed framework outperforms existing state-of-the-art methods.

1 Introduction

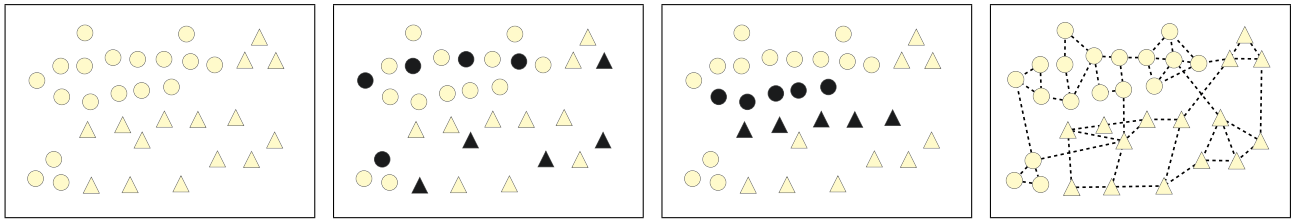
With the massive amounts of data produced by microblogging services, substantial efforts have been devoted to processing and understanding microblogging data via supervised learning techniques, e.g., text classification [4] and sentiment classification [15] of microblogging messages. Supervised learning methods aim to learn a model based on training data, which involves a basic assumption that a large number of labeled instances are available. However, labels can be expensive and time consuming to obtain, especially for microblogging messages, which presents great challenges to the application of supervised learning algorithms.

One effective approach to reducing the cost of labeling is *active learning* [8]. Active learning aims to determine which data instances should be selected to query for labels such that the classifier could achieve high accuracy using as few labeled instances as possible, thereby minimizing the cost of obtaining labeled data [8]. The objective of active learning is to maximize information gain given a fixed budget of labeling efforts. Active learning has been shown to be useful in many real-world applications, including node classification [19], document classification [31], etc. However, traditional active learning methods often assume that data instances are independent and identically distributed (i.i.d.). This is not the case with microblogging data, in which texts are networked with each other. To the best of our knowledge, use of active learning to handle the labeling bottleneck in networked microblogging data has not been well studied yet.

We illustrate their differences using an example in Figure 1. Figure 1(a) shows a binary classification example with classes represented by different shapes (circle and triangle). Traditional active learning methods select instances to label according to two main criteria, i.e., representativeness and informativeness [18]. Representativeness measures whether an instance can well represent the overall input patterns of unlabeled data, and informativeness is the ability of an instance to reduce the uncertainty of a statistical model [27]. Examples of the selection criteria are shown in Figures 1(b) and 1(c). Unlike traditional data, as shown in Figure 1(d), microblogging data provides information beyond text. A distinct feature of texts in microblogging is that they can be correlated through user connections, which could contain useful information that is lost in purely text-based metrics. Besides content information, relations between messages can be represented via user-message relations and user-user relations. As indicated by Figures 1(b) and 1(c), traditional methods tend to select instances to learn the decision boundary by analyzing their content information. It necessitates the investigation of active learning in handling microblogging messages with their relation information.

In this paper, we investigate issues of active learning for microblogging data as illustrated in Figure 1(d). To utilize both content information and relation infor-

*Computer Science and Engineering Department, Arizona State University, Tempe, AZ 85287, USA. {xiahu, jiliang.tang, huiji.gao, huan.liu}@asu.edu



(a) Traditional Data (b) Selection of Representative Instances (c) Selection of Informative Instances (d) Networked Texts in Microblogging

Figure 1: A Toy Example for Selecting Representative and Informative Instances in Microblogging

mation in an active learning framework for networked data, we have two challenges: 1) how do we incorporate relation information into text content modeling? 2) how do we select the most representative and informative instances by taking advantage of relation information? To tackle these two challenges, we propose a novel active learning framework (*ActNeT*) for networked microblogging data. The main contributions of this paper are summarized as follows:

- We formally define the problem of active learning for networked microblogging data;
- We formally model relation information between messages, and integrate the information into content modeling;
- We present a novel active learning framework by proposing two selection strategies to make use of the network structure of microblogging data;
- We empirically evaluate the proposed *ActNeT* framework on two real-world Twitter datasets and elaborate the effects of selection strategies on active learning.

2 Problem Statement

Given a microblogging corpus $\mathbf{G} = (\mathbf{X}, \mathbf{S})$, where $\mathbf{X}$ is a text content matrix and $\mathbf{S}$ is a social context matrix. For the text content matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$, n is the number of messages, and m is the number of features. For the social context matrix $\mathbf{S} = (\mathbf{P}, \mathbf{F})$, $\mathbf{P} \in \mathbb{R}^{d \times n}$ is a user-message matrix, and $\mathbf{F} \in \mathbb{R}^{d \times d}$ is a user-user matrix. $\mathbf{u} = \{u_1, u_2, \ldots, u_d\}$ is the user set, where d is the number of distinct users in the corpus. In the user-message matrix, $\mathbf{P}_{ij} = 1$ denotes that message $\mathbf{t}_j$ is posted by user u_i. In the user-user friendship matrix, $\mathbf{F}_{ij} = 1$ indicates that user u_i is connected by user u_j. The graph is a directed graph, thus $\mathbf{F}$ is asymmetric.

Now we formally define active learning in microblogging as:

Given a corpus of microblogging messages $\mathbf{G}$ *with their text content information* $\mathbf{X}$*, and social context information* $\mathbf{S}$*, including the user-message matrix* $\mathbf{P}$ *and user-user matrix* $\mathbf{F}$*, and a budget* B*, the task is to select* B *instances from* $\mathbf{X}$ *to be labeled by an oracle (e.g., human annotator), so that the learned classifier* $\mathbf{W}$ *based on the labeled data can achieve maximal accuracy on unseen data (i.e., test data).*

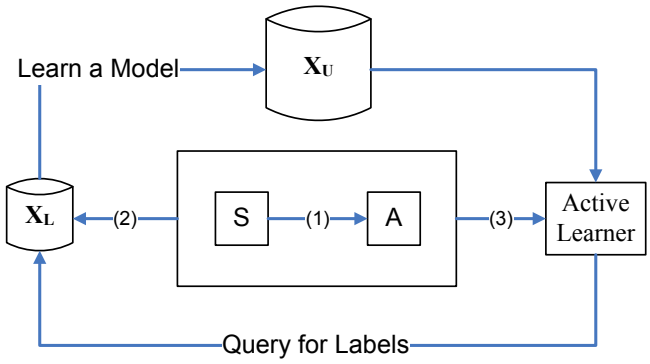


Figure 2: ActNeT Framework: (1) relation modeling; (2) text content modeling; (3) selection strategies for networked data.

3 A New Framework – ActNeT

We plot the work flow of our proposed framework in Figure 2. In the figure, $\mathbf{X}_L$ represents a dataset with label information, $\mathbf{X}_U = \mathbf{X} \backslash \mathbf{X}_L$ is an unlabeled dataset, $\mathbf{S}$ is the social context matrix, and $\mathbf{A}$ is a message-message relation matrix.

In Figure 2, the outer cycle illustrates a traditional pool-based active learning work flow [27]. In the beginning, we have a small (or empty) labeled dataset $\mathbf{X}_L$. A learner may request labels for one or more carefully selected instances, learn from the query results, and then leverage its updated knowledge to choose instances from $\mathbf{X}_U$ to query next.

To leverage social context information, our proposed framework *ActNeT* consists of three more com-

ponents than traditional active learning, as shown in the inner part of Figure 2. (1) In the *relation modeling* component (Section 3.1), we extract and formally model message-message relations by analyzing social context information. (2) We incorporate the built relation information as a regularization into the *text content modeling* (Section 3.2). (3) We study two *selection strategies for networked data* (Section 3.3) to help the active learner choose the most representative and informative instances, in terms of network structure, to query for labels. The second component incorporates relation information into content modeling, while the third analyzes the social network structure directly. The second and third components correspondingly tackle the two challenges posed in Section 1. We elaborate these three components in the next three subsections.

3.1 Relation Modeling Following previous work [15, 22] on social media data, we extract two kinds of relations between messages based on social context – user-message and user-user relations. Given the user-message matrix $\mathbf{P}$ and user-user matrix $\mathbf{F}$, the first message-message matrix $\mathbf{A}p$ is defined as $\mathbf{A}p = \mathbf{P}^T \times \mathbf{P}$, where $\mathbf{A}p_{ij} = 1$ indicates that $\mathbf{t}_i$ and $\mathbf{t}_j$ are posted by the same user. The second message-message matrix $\mathbf{A}f$ is defined as $\mathbf{A}f = \mathbf{P}^T \times \mathbf{F} \times \mathbf{P}$, where $\mathbf{A}f_{ij} = 1$ indicates that the author of $\mathbf{t}_i$ is a friend of the author who wrote $\mathbf{t}_j$. The message-message matrix can be considered as either the relation $\mathbf{A}p$, $\mathbf{A}f$, or the combination $\mathbf{A} = \mathbf{A}p + \theta\mathbf{A}f$, where θ controls the weight of two different relations. In this paper, we focus on incorporating the constructed relations between messages in text content modeling for active learning, but not optimal ways of combining them. Thus, we combine these two relations with equal weight $\theta = 1$ to construct a combined relation matrix.

To incorporate the extracted relation information, the basic idea is to build a latent connection to make two messages as close as possible if they are posted by the same user or two users who are friends with each other. Thus we form a regularization term, which can be mathematically formulated as minimizing the following objective function,

$$\begin{aligned} &\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\mathbf{A}_{ij}\sum_{k=1}^{c}(\hat{\mathbf{Y}}_{ik} - \hat{\mathbf{Y}}_{jk})^2\right] \\ &= tr(\mathbf{W}^T\mathbf{X}\mathcal{L}\mathbf{X}^T\mathbf{W}) \end{aligned} \tag{3.1}$$

where $\hat{\mathbf{Y}} = \mathbf{X}^T\mathbf{W}$ is the fitted value of the true class label $\mathbf{Y}$. $\mathcal{L} = \mathbf{D} - \mathbf{A}$ is the Laplacian matrix [28], where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is the message-message relation matrix to represent a direct graph, $\mathbf{D} \in \mathbb{R}^{n\times n}$ is a diagonal matrix with $\mathbf{D}_{ii} = \sum_{j=1}^{n}\mathbf{A}_{ij}$.

3.2 Text Content Modeling For multi-class classification problems, ridge regression has been widely used in different domains. It learns multiple linear classifiers by solving the following optimization problem,

$$\min_{\mathbf{W}} \quad \frac{1}{2}||\mathbf{X}_L^T\mathbf{W} - \mathbf{Y}||_F^2 + \frac{\lambda_R}{2}||\mathbf{W}||_F^2, \tag{3.2}$$

and it has a closed-form solution,

$$\hat{\mathbf{W}}_{Ridge} = (\mathbf{X}_L\mathbf{X}_L^T + \lambda_R\mathbf{I})^{-1}\mathbf{X}_L\mathbf{Y}, \tag{3.3}$$

where $\mathbf{I}$ is an $n \times n$ identity matrix.

To integrate relation information into a classification model, we propose a least squares model with Laplacian regularization (LSLap), which is formulated by solving the following optimization problem:

$$\begin{aligned} \min_{\mathbf{W}} \quad & f(\mathbf{W}) = \frac{1}{2}||\mathbf{X}_L^T\mathbf{W} - \mathbf{Y}||_F^2 \\ & + \frac{\lambda_L}{2}tr(\mathbf{W}^T\mathbf{X}\mathcal{L}\mathbf{X}^T\mathbf{W}) + \frac{\lambda_R}{2}||\mathbf{W}||_F^2, \end{aligned} \tag{3.4}$$

where λ_L is the regularization parameter to control the contribution of relation information, and λ_R is the ridge regularization parameter.

We take the derivative of $f(\mathbf{W})$,

$$\begin{aligned} \frac{\partial f(\mathbf{W})}{\partial \mathbf{W}} &= \mathbf{X}_L\mathbf{X}_L^T\mathbf{W} - \mathbf{X}_L\mathbf{Y} \\ &+ \lambda_L\mathbf{X}\mathcal{L}\mathbf{X}^T\mathbf{W} + \lambda_R\mathbf{I}\mathbf{W}. \end{aligned} \tag{3.5}$$

Matrices $\mathbf{X}_L\mathbf{X}_L^T$ and $\mathbf{X}\mathbf{L}\mathbf{X}^T$ are semi-positive definite, therefore, $\mathbf{X}_L\mathbf{X}_L^T + \lambda_L\mathbf{X}\mathbf{L}\mathbf{X}^T + \lambda_R\mathbf{I}$ is a positive definite matrix. By setting the derivative to zero, we can obtain the solution as follows:

$$\hat{\mathbf{W}}_{LSLap} = (\mathbf{X}_L\mathbf{X}_L^T + \lambda_L\mathbf{X}\mathcal{L}\mathbf{X}^T + \lambda_R\mathbf{I})^{-1}\mathbf{X}_L\mathbf{Y}, \tag{3.6}$$

The Laplacian regularization incurs a penalty to force two connected messages $\mathbf{x}_i$ and $\mathbf{x}_j$ to have similar labels. Compared with original ridge regression, LSLap integrates explicit relation information among messages.

3.3 Selection Strategies for Networked Data Traditional active learning methods select *representative* [25] or *informative* [1] instances to query for labels according to their content information only. Given the social network information available in microblogging data, in this section, we further explore particular features of the network topology to help select instances to query for labels.

In particular, based on the constructed message-message relation network, we examine two selection strategies for active learning.

3.3.1 Strategy 1: Global Selection As we know, representativeness-based active learning methods aim to select instances which can well represent the overall pattern of unlabeled data. For the networked data, we ask if we can select representative nodes to capture topological patterns of the whole network.

In social network analysis, many methods have been proposed to capture particular features of the network topology. The proposed methods quantify network structure with various metrics [24]. We use one of the widely used methods, PageRank [26], to select representative nodes in a network. The key idea of this selection strategy is that the nodes in the network with high PageRank scores could represent the overall patterns of the social network topology. In other words, by labeling highly representative nodes, the label information will propagate through the whole network [19].

The PageRank score can be calculated as:

$$\mathbf{x} = \alpha \mathbf{A}\mathbf{O}^{-1}\mathbf{x} + \beta \mathbf{1}, \tag{3.7}$$

where $\mathbf{x}$ is a vector of PageRank scores of all the nodes, α and β are two positive constants, $\mathbf{A}$ is the adjacency matrix, and $\mathbf{1}$ is the vector (1,1,1, ...). $\mathbf{O}$ is the diagonal matrix with elements $\mathbf{O}_{ii} = max(k_i^{out}, 1)$, and k_i^{out} is the out-degree of node i.

3.3.2 Strategy 2: Local Selection As discussed above, we select representative nodes from the whole network according to their PageRank scores. An alternative selection strategy is to consider both representativeness and informativeness of network topology in the active learning framework.

As we know, nodes sharing certain properties in a network tend to form groups with more within-group connections, which is related to a fundamental task in social network analysis – community detection [30]. Community detection algorithms aim to partition nodes in a network into different communities that have more within-group connections than between-group connections. Thus a natural choice of selecting representative nods in the network is to sample locally representative nodes from different communities. Modularity [23] is a popular community measure that explicitly takes the degree distribution into consideration and has been shown to be an effective quantity by which to measure community structure in many social network applications [10]. Here, we use modularity maximization [23] to partition the social network into communities.

After obtaining community membership information, we then select nodes with high PageRank scores in each community. We consider the messages selected from different communities as the ones that are informative in terms of network topology. The idea of this strategy is that finding locally representative nodes in each community takes both representativeness and informativeness into account.

Our work focuses on studying the impact of social network information to facilitate the performance of active learning framework. It is possible to use other alternative community detection methods and network metrics in the selection procedure.

Algorithm 1 *ActNeT*: Active Learning for Networked Texts in Microblogging

Input: $\{B, b, \mathbf{X}, \mathbf{P}, \mathbf{F}, k\}$
Output: $\mathbf{X}_L$

1: Construct Laplacian matrix $\mathcal{L}$ from $\mathbf{P}$ and $\mathbf{F}$
2: Compute Selection Score $SS(\mathbf{x}), \mathbf{x} \in \mathbf{X}$
3: Initialize $\mathbf{X}_L$ with b instances
4: Train $\hat{\mathbf{W}}_{LSLap}$ and $\hat{\mathbf{W}}_{Ridge}$
5: $\mathbf{C}^k \leftarrow$ Pick k instances based on $SS(\mathbf{x})$
6: **while** $|\mathbf{X}|_L < B$ **do**
7: $\quad \mathbf{x}_* = \arg\max_{\mathbf{x} \in C^k} Entropy(\mathbf{x}, \hat{\mathbf{W}}_{LSLap}, \hat{\mathbf{W}}_{Ridge})$
8: $\quad$ Remove $\mathbf{x}_*$ from $\mathbf{C}^k$, add $\mathbf{x}_*$ to $\mathbf{X}_L$
9: $\quad$ Update $\hat{\mathbf{W}}_{LSLap}$ and $\hat{\mathbf{W}}_{Ridge}$
10: **end while**

3.4 Active Learning Algorithm After elaborating the three components plotted in Figure 2, we introduce the detailed algorithm of *ActNeT* in Algorithm 1.

In line 2, we compute the selection scores $SS(x)$ for all the networked instances. The selection scores can be computed with either global or local selection strategies, discussed in Sections 3.3.1 and 3.3.2. In line 3, a small number (b) of instances with highest selection scores are selected to query for labels. These instances are used to train the base learners $\hat{\mathbf{W}}_{LSLap}$ and $\hat{\mathbf{W}}_{Ridge}$ in line 4. This step presents challenges for traditional active learning method, in which they have to randomly select some instances to label as initialization. The classification result is sensitive to the initialization to some extent. As we discussed above, the two selection strategies can be applied to the readily available message-message network directly. Thus our proposed method can avoid the initialization problem.

In line 5, k instances with highest selection scores are selected from $\mathbf{X}_U$ as candidates. In lines 6 to 10, *ActNeT* proceeds in iterations until the budget B is exhausted. In each iteration, we select the most informative instances from the candidates pool $\mathbf{C}^k$ based on their vote entropy [9] evaluated by a committee of base

learners. In line 7, the instance with highest entropy is defined as:

$$\mathbf{x}_* = \arg\max_{\mathbf{x}\in C^k} \; -\sum_{i=1}^{\mathcal{K}} \frac{V(y_i)}{\mathcal{K}} log \frac{V(y_i)}{\mathcal{K}}, \tag{3.8}$$

where $\mathcal{K}$ is the number of classifiers in the committee, y_i is class label provided by the ith classifier in the committee, and $V(y_i)$ is the number of occurrences of the class label y_i. In particular, we utilize LSLap and LS as base classifiers of the committee in the experiment. This step is to select the most informative nodes based on their content information. Then the selected instances are queried for labels, added to $\mathbf{X}_L$, and used to update the base classifiers.

4 Experiments

We present experimental results to assess the effectiveness of our proposed active learning framework.

4.1 Datasets We now introduce two real-world Twitter datasets used in our experiment.

TRECTopic: Similar to experimental settings in [14, 4], topics (hashtags) are considered to be class labels of tweets in our experiment. According to the topics of the tweets, we construct a ten-class Twitter dataset, which is a subset of TREC2011 data[1]. We balance the number of tweets in each class to avoid bias brought by skewed class distribution.

We further refined the tweets according to the social network information of users, which is crawled during July 2009 [20]. We filter tweets whose author has no friends or published less than two tweets. All the hashtags in the original tweets are removed during training to avoid bias brought by class labels.

TwitterStream: Following the data construction process in [4], based on the selected ten topics, tweets are crawled using Twitter Search API[2]. Tweets retrieved by the same topic are considered to be in the same category. Then we have tweets belonging to ten categories. In order to obtain the relation information, the tweets are filtered according to the same rules used in refining the TRECTopic dataset.

We remove stop-words and perform stemming for all the tweets. The statistics of the two datasets are presented in Table 1.

4.2 Experimental Setup In the experiment, the dataset is divided into two groups of equal size for training and testing. The active learner selects instances from the training data to query for labels. LibSVM [3] is used to train a SVM classifier based on the labeled data, and used to classify the instances in the testing data. The testing data is separate with an active learning process. Testing is done on unseen instances, but not on the remaining part of $\mathbf{X}$ in Algorithm 1. We apply different active learning methods to select B instances, and train a SVM classifier based on the selected labeled instances. Following the ratio of selection budget to the whole data size used in active learning literature [27, 19], we set $B = 500$ for general experiment purposes. Classification accuracy is employed as the performance metric to evaluate the quality of selected instances for classification. In order to demonstrate the effectiveness of our proposed active learning framework, we compare the proposed framework with following methods:

Table 1: Summary of Experimental Datasets

	TRECTopic	*TwitterStream*
# of Tweets	119,448	7,138
# of Unigrams	90,388	12,233
# of User	38,467	2412
# of Classes	10	10
Max Class Size	12,012	766
Min Class Size	11,885	688
Max Degree of Users	1,244	426
Min Degree of Users	1	1
Ave. Tweets per User	3.11	2.96

- *Random*: this method randomly selects instances to query for labels.
- *Uncertainty* [21]: the key idea of this method is to select the instances with least prediction margin between the first and second most probable class labels under the model, which is defined as:

$$\mathbf{x}^* = \arg\min_{x\in\mathbf{X}_U} \; P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x),$$

  where $\hat{y}_1$ and $\hat{y}_2$ are the first and second most probable labels. In this framework, the instances with small margins are considered to be ambiguous, thus knowing the true label would help the classification model discriminate more effectively between them.
- *QBC* [9]: this method selects the instances with highest disagreement level evaluated by a committee of several learners. In the experiment, entropy is used to combine the votes provided by the committee members in the experiment.
- *CLUSTER* [11]: this method samples instances with hierarchical clustering of unlabeled data.

[1] http://trec.nist.gov/data/tweets/
[2] http://search.twitter.com/api/

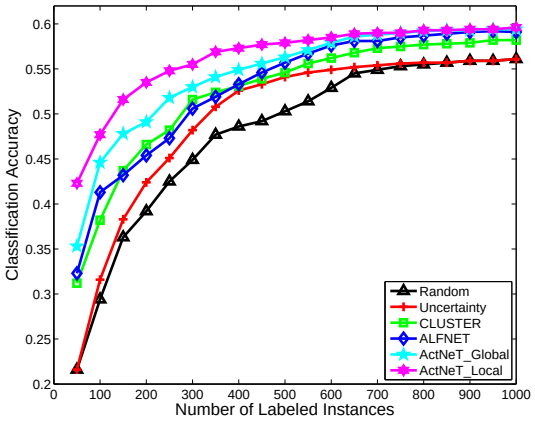
- *ALFNET* [2]: this method clusters the nodes of a graph into several groups, and then randomly samples nodes from each cluster. The selected instances are utilized to train a collective classifier to incorporate the network information.

- *ActNeT_Global*: our proposed method with a global selection strategy.

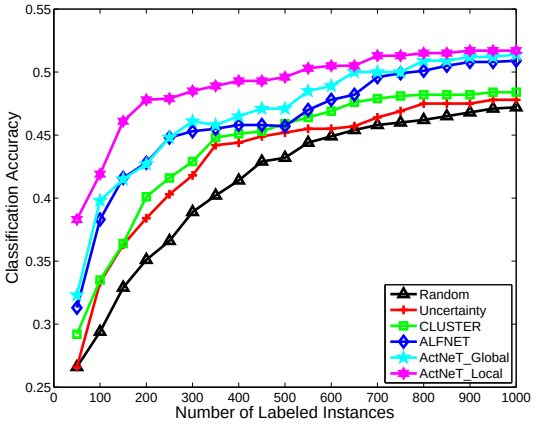
- *ActNeT_Local*: our proposed method with a local selection strategy.

Among the baseline methods, *Random* is the way many supervised methods in social media mining used to build the training data for learning, *Uncertainty* and *QBC* are traditional content-based active learning methods, *CLUSTER* and *ALFNET* are the state-of-the-art active learning methods on content and graph information, respectively. Some methods, i.e. *Uncertainty* and *QBC*, need a small number (b) of labeled instances to train the base learners for initialization. Following experimental settings in [18], we set $b = 50$, which is very small in 10-class classification tasks. Thus, 50 instances are randomly selected for initialization of the two methods in the general experiment.

There are four important parameters involved in our experiments, including λ_R, λ_L in Eq. (3.3) and (3.6), number of communities c in Section 3.3.2, and number of selected instances k in Algorithm 1. All four parameters are positive. As a common practice, λ_R and λ_L can be tuned via cross-validation. In the experiments, we set $\lambda_R = 0.005$ and $\lambda_L = 0.01$ for all the methods. We simply set $c = 10$, $k = 2 \times B$ (i.e., $k = 1000$) for general experiment purposes, and the impacts of the two parameters c and k will be further discussed in Section 4.4.



(a) Classification Accuracy on TRECTopic



(b) Classification Accuracy on TwitterStream

Figure 3: Classification Accuracy vs. Number of Labels Used on the Two Datasets

4.3 Performance Evaluation Experimental results of the baseline methods on the two Twitter datasets are respectively plotted in Figures 3(a) and 3(b). For each classification task, we keep increasing the number of instances selected to label (budget B) from 50 to 1000, and compare the accuracy of classifiers trained based on the labeled data with different numbers of instances. From the figures, we draw the following observations:

(1) *ActNeT_Local* performs consistently better than other baselines. It demonstrates the significance of our proposed framework by exploiting the explicit network structure. *Uncertainty* and *QBC* are two classical content-based methods, and they turn out to perform similarly to each other. *ActNeT_Global* has comparable results with *ALFNET*, which further demonstrates that the representativeness and informativeness in a network are both important criteria for active learning.

(2) Specifically, the methods *ActNeT_Local*, *ActNeT_Global*, *ALFNET*, and *CLUSTER* achieve significant improvement compared with other baselines when the number of labeled instances is small ($B = 50$). This is because *Uncertainty* and *QBC* randomly select a portion of data to label for training base learners. Quality of the randomly selected instances is unreliable. This property has its significance for various applications in social media when the labeling budget is small.

We further provide detailed results of the baseline methods on the two Twitter datasets with the budget $B = 500$ in Tables 2 and 3. To reduce bias brought by class distribution, experiments on the two Twitter datasets are conducted by choosing different numbers of classes (i.e., 3, 5, 8, 10 classes) from the original dataset. For each given class number, the average accuracy of 10 repeat randomly chosen classes is reported. For

Table 2: Classification Accuracy on TRECTopic Dataset

	3-class (gain)	5-class (gain)	8-class (gain)	10-class (gain)
Random	0.717 (N.A.)	0.652 (N.A.)	0.581 (N.A.)	0.503 (N.A.)
Uncertainty	0.761 (+6.13%)	0.682 (+4.60%)	0.613 (+5.51%)	0.541 (+7.55%)
QBC	0.778 (+8.51%)	0.732 (+7.67%)	0.623 (+7.23%)	0.547 (+8.75%)
CLUSTER	0.768 (+7.11%)	0.701 (+7.52%)	0.635 (+9.29%)	0.546 (+8.55%)
ALFNET	0.778 (+8.51%)	0.706 (+8.28%)	0.631 (+8.61%)	0.557 (+10.74%)
ActNeT_Global	0.793 (+10.60%)	0.719 (+10.28%)	0.645 (+11.02%)	0.563 (+11.93%)
ActNeT_Local	**0.801 (+11.72%)**	**0.731 (+12.12%)**	**0.673 (+15.84%)**	**0.591 (+17.90%)**

Table 3: Classification Accuracy on TwitterStream Dataset

	3-class (gain)	5-class (gain)	8-class (gain)	10-class (gain)
Random	0.627 (N.A.)	0.566 (N.A.)	0.510 (N.A.)	0.432 (N.A.)
Uncertainty	0.629 (+0.32%)	0.568 (+0.35%)	0.518 (+1.57%)	0.452 (+4.63%)
QBC	0.636 (+1.44%)	0.577 (+1.94%)	0.525 (+2.94%)	0.443 (+2.55%)
CLUSTER	0.649 (+3.51%)	0.591 (+4.42%)	0.541 (+6.08%)	0.459 (+6.25%)
ALFNET	0.665 (+6.06%)	0.594 (+4.95%)	0.543 (+6.47%)	0.457 (+5.79%)
ActNeT_Global	0.672 (+7.18%)	0.606 (+7.07%)	0.555 (+8.82%)	0.471 (+9.03%)
ActNeT_Local	**0.699 (+11.48%)**	**0.632 (+11.66%)**	**0.582 (+14.12%)**	**0.501 (+15.97%)**

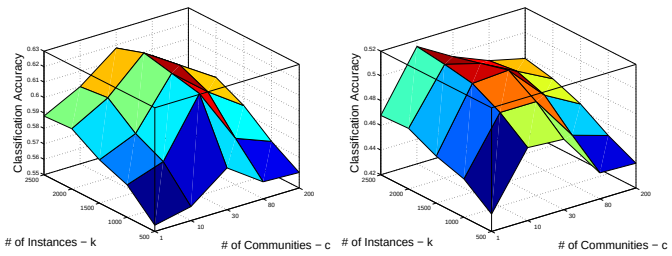
the 10-class dataset, experiment is done on the whole dataset. In the tables, "gain" represents the percentage of relative improvement of the methods as compared to the *Random* method, which is the strategy used in many supervised methods for obtaining labeled training data. By comparing the classification accuracy of different methods, we draw the following observations:

(1) Generally, methods *Uncertainty*, *QBC*, and *CLUSTER* achieve better performance than the *Random*. This demonstrates the effectiveness of active learning methods to achieve better results with limited labeling efforts. These content-based methods achieve comparable performance. This indicates that, only considering content information of tweets, neither of the methods can achieve much better performance than the other, which is consistent with previous findings [18].

(2) In most cases, especially on the TwitterStream dataset, *ALFNET* achieves better performance than traditional methods that only utilize the content information of the microblogging data. This demonstrates the usefulness of utilizing explicit network information for active learning. *ActNeT_Global* achieves comparable performance to *ALFNET*. While *ALFNET* is a typical method to exploit informative instances, *ActNeT_Global* aims to select representative instances from the relation network. The results show that, similar to the content-based methods, active learners can have relatively good results by exploring either informative or representative instances from the network.

(3) Compared with the baselines, *ActNeT_Local* achieves the best performance on both datasets with different class settings, indicating that the proposed framework successfully utilizes the network information to facilitate active learning. The highest improvement (17.90%) with respect to *Random* is obtained on the TRECTopic dataset when we have 10-class data for experiment. We conduct a two-sample one-tail t-test at 95% significance level to compare *ActNeT_Local* with the best baselines *ALFNET* and *ActNeT_Global*. The results demonstrate that our approach significantly outperforms the two methods with p-value $<< 0.01$.

4.4 Parameter Selection As discussed in Section 4.2, two important parameters, i.e., c and k, are involved in our proposed framework. The number of communities c is a tradeoff of informativeness and representativeness for sampling in the network. When $c = 1$, the local selection strategy becomes similar to the global measure, i.e., *ActNeT_Local* evaluates representativeness in one community. The setting k is to control the search area during sampling. A larger k indicates that the model is able to find more informative instances in terms of discriminative content information for learning the classifiers, but not only depending on the network



(a) TRECTopic (b) TwitterStream

Figure 4: Impact of Number of Communities (c) and Search Area (k) to Classification Accuracy

structure. Hence, we analyze the effects of those two important parameters on our proposed active learning framework. In particular, we conduct experiments to compare the classification accuracy of *ActNeT_Local* on the two datasets with different parameter settings.

The classification results of *ActNeT_Local* with different parameter settings on the two datasets are plotted in Figure 4. In Figure 4(a), performance of *ActNeT_Local* improves as the number of communities increases, and reaches a peak at $c = 30$. When $c > 30$, as the number of communities grows, the performance of *ActNeT_Local* declines. The results demonstrate that the proposed framework can achieve a relatively good performance when choosing a reasonable number of communities to partition. In practice, setting c to 10 - 30 achieves good performance in both datasets.

In general, performance of *ActNeT_Local* increases when we select more instances to evaluate. This is because we utilize more content information, with the combination of network information. Apparently, it is time consuming, especially for large-scale social media data, when we utilize more instances to perform entropy comparison. Thus we need to make a tradeoff between accuracy and time efficiency to set the parameter. When $k > 1500$, the improvement is not as significant as that when k is small. Thus, choosing $k < 5B$ (budget) is a good choice in practice. Similar patterns can be observed from the classification results on the TwitterStream dataset.

5 Related Work

Supervised learning algorithms [5] have been extensively used in microblogging applications, including sentiment classification [33], event analysis [16], tweet classification [4], topic modeling [17], etc. The "labeling bottleneck" widely exists in microblogging, which presents great challenges to achieving good performance for supervised learning methods. However, to the best of our knowledge, as an effective way to tackle the "labeling bottleneck", active learning on networked microblogging data has not been considered yet.

Active learning has been extensively studied in various domains for years. Most of the existing methods focus on the data represented by feature vectors [27], and they can be generally categorized into three groups. First, active learners select either the most uncertain instances determined by a single classifier [1, 31] or a committee of classifiers [9, 12]. These approaches always evaluate the data instances separately, thus can not utilize the structure of the data. The second group of methods exploit cluster structure in data, and select instances in each cluster to avoid sampling bias [11, 25, 32]. The key idea of these approaches is to identify a sophisticated cluster structure based on content information. The key limitation of these methods is that they cannot well utilize information from labeled data. Different from traditional approaches, our proposed framework incorporates relation information into the content modeling, and further selects instances by taking advantage of the social network structure.

It has been found to be useful in various applications by studying the topological characteristics of social networks, including feature selection [13], phenotype classification [7], trust prediction [29], anomaly detection [6], etc. While many active learning approaches have been proposed on attribute-value data that is independent and identically distributed, efficient active learners that utilize the explicit social network structure in the data have not been considered until recently [2]. Different from our task, [2] focuses on dealing with citation networks, in which the instances are directly connected. In addition, they randomly sample nodes from different clusters of the network, but ignore another important criterion in active learning – representativeness of the selected nodes in social networks.

6 Conclusions

To overcome the labeling bottleneck in microblogging, we propose to utilize active learning to select the most representative and informative texts to query for labels. Unlike texts in traditional media, microblogging texts are embedded with social relations, which presents great challenges for active learning. In this paper, we develop a novel active learning framework to handle the networked texts in microblogging. In particular, we extract relations between texts based on social theories, and model the relations using graph Laplacian, which is employed as a regularization to ridge regression. Thus the relations between messages can be naturally embedded into the active learning process to effectively select informative instances from the data. We further propose global and local selection strategies for

networked instances. Experimental results show that message-message relations are helpful for active learning on microblogging messages. Empirical evaluations demonstrate that our framework *ActNeT_Local*, which considers representativeness and informativeness in active learning, significantly outperforms the representative baselines on two real-world datasets.

This work suggests some interesting directions for future work. For example, it would be interesting to investigate the potential impact of additional information, like geographical information, temporal information, etc. We can further explore different metrics in social network analysis to select influential nodes in the network from other perspectives.

Acknowledgments

We thank the anonymous reviewers for their comments. This work is, in part, supported by ONR (N000141110527) and (N000141010091).

References

[1] M. Balcan, A. Broder, and T. Zhang. Margin based active learning. *Learning Theory*, pages 35–50, 2007.

[2] M. Bilgic, L. Mihalkova, and L. Getoor. Active learning for networked data. In *ICML*, 2010.

[3] C. Chang and C. Lin. Libsvm: a library for support vector machines. *TIST*, 2011.

[4] Y. Chen, Z. Li, L. Nie, X. Hu, X. Wang, T.-S. Chua, and X. Zhang. A semi-supervised bayesian network model for microblog topic classification. In *Proceedings of COLING*, 2012.

[5] Z. Chen, W. Hendrix, H. Guan, I. K. Tetteh, A. Choudhary, F. Semazzi, and N. F. Samatova. Discovery of extreme events-related communities in contrasting groups of physical system networks. *DMKD*, 2012.

[6] Z. Chen, W. Hendrix, and N. Samatova. Community-based anomaly detection in evolutionary networks. *JIIS*, 2011.

[7] Z. Chen, K. Padmanabhan, A. Rocha, Y. Shpanskaya, J. Mihelcic, K. Scott, and N. Samatova. SPICE: discovery of phenotype-determining component interplays. *BMC Syst Biol*, 2012.

[8] D. Cohn, Z. Ghahramani, and M. Jordan. Active learning with statistical models. *Arxiv*, 1996.

[9] I. Dagan and S. Engelson. Committee-based sampling for training probabilistic classifiers. In *ICML*, 1995.

[10] L. Danon, J. Duch, A. Arenas, and A. Daz-guilera. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005.

[11] S. Dasgupta and D. Hsu. Hierarchical sampling for active learning. In *ICML*, 2008.

[12] Y. Freund, H. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine learning*, 28(2):133–168, 1997.

[13] Q. Gu and J. Han. Towards feature selection in network. In *CIKM*, 2011.

[14] X. Hu, N. Sun, C. Zhang, and T.-S. Chua. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In *Proceedings of CIKM*, 2009.

[15] X. Hu, L. Tang, J. Tang, and H. Liu. Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of WSDM*, 2013.

[16] Y. Hu, A. John, D. Seligmann, and F. Wang. What were the tweets about? topical associations between public events and twitter feeds. In *Proceedings of ICWSM*, 2012.

[17] Y. Hu, A. John, F. Wang, and S. Kambhampati. Et-lda: Joint topic modeling for aligning events and their twitter feedback. In *Proceedings of AAAI*, 2012.

[18] S. Huang, R. Jin, and Z. Zhou. Active learning by querying informative and representative examples. *NIPS*, 2010.

[19] M. Ji and J. Han. A variance minimization criterion to active learning on graphs.

[20] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of WWW*, 2010.

[21] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *SIGIR*, 1994.

[22] Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi. Exploiting social context for review quality prediction. In *WWW*, 2010.

[23] M. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[24] M. Newman. *Networks: an introduction*. Oxford University Press, Inc., 2010.

[25] H. Nguyen and A. Smeulders. Active learning using pre-clustering. In *ICML*, 2004.

[26] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. *Technical Report, Stanford*, 1999.

[27] B. Settles. Active learning literature survey. *University of Wisconsin, Madison*, 2010.

[28] A. Smola and R. Kondor. Kernels and regularization on graphs. *Learning theory and kernel machines*, 2003.

[29] J. Tang, H. Gao, X. Hu, and H. Liu. Exploiting homophily effect for trust prediction. In *WSDM*, 2013.

[30] L. Tang and H. Liu. Relational learning via latent social dimensions. In *SIGKDD*, 2009.

[31] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2002.

[32] Z. Xu, K. Yu, V. Tresp, X. Xu, and J. Wang. Representative sampling for text classification using support vector machines. *Information Retrieval*, 2003.

[33] K. Zhang, Y. Xie, Y. Cheng, D. Honbo, D. Downey, A. Agrawal, W. Liao, and A. Choudhary. Sentiment identification by incorporating syntax, semantics and context information. In *Proceedings of SIGIR*, 2012.