A Semi-Supervised Bayesian Network Model for Microblog Topic Classification

Yan Chen[1,2] Zhoujun Li[1] Liqiang Nie[2] Xia Hu[3] Xiangyu Wang[2] Tat-seng Chua[2] Xiaoming Zhang[1]

[1]State Key Laboratory of Software Development Environment, Beihang University, China
[2]School of Computing, National University of Singapore, Singapore
[3]Arizona State University, United States

11-12-2012

Outline

Outline

Background

- Microblogging services are becoming immensely popular in breaking-news disseminating, information sharing, and events participation.

Background

- Microblogging services are becoming immensely popular in breaking-news disseminating, information sharing, and events participation.
- The most well known one is Twitter, which has more than 140 million active users with 1 billion Tweets every 3 days as of March 2012.

Background

- Microblogging services are becoming immensely popular in breaking-news disseminating, information sharing, and events participation.
- The most well known one is Twitter, which has more than 140 million active users with 1 billion Tweets every 3 days as of March 2012.
- In China, Weibo (www.weibo.com) has accumulated more than 300 millions users in less than three years. Every second, more than 1000 Chinese tweets are posted in Weibo.

Background

- Microblogging services are becoming immensely popular in breaking-news disseminating, information sharing, and events participation.
- The most well known one is Twitter, which has more than 140 million active users with 1 billion Tweets every 3 days as of March 2012.
- In China, Weibo (www.weibo.com) has accumulated more than 300 millions users in less than three years. Every second, more than 1000 Chinese tweets are posted in Weibo.

With the large volume and multi-aspect messages, how do users locate the specific messages that they are interested in?

Motivation

Example 1:

Query

Motivation

Example 1:

DBS Bank

Motivation

Example 1:

DBS Bank

Motivation

Example 1:

DBS Car

Motivation

Example 2:

Motivation

Example 2:

Motivation

Example 2:

Motivation

Example 2:

Motivation

Example 2:

How do we provide users an overviews of search results based on meaningful and structural categories.

Motivation

Example 2:

Topic Classification!

Outline

Related Work

1. Topic Model based Methods
   - [Hong and Davison, 2010] employs latent dirichlet allocation (LDA) [Blei et al., 2003] and author-topic model [Rosen-Zvi et al., 2010] to deeply investigate to automatically find hidden topic structures on Twitter.
   - Several variants of LDA to incorporate supervision have been proposed by [Ramage et al., 2009, Ramage et al., 2010], and have been shown to be competitive with strong baselines in the microblogging environment.
2. Traditional Classification Methods
   - [Lee et al., 2011] classified tweets into pre-defined categories such as sports, technology, politics, *etc*. They constructed word vectors with tf-idf weights and utilized a Naive Bayesian Multinomial classifier to classify tweets.
   - [Sriram et al., 2010] proposed to use a small set of domain-specific features extracted from the author's profile and text to represent short messages. Their method requires extensive pre-processing to conduct effectively feature analysis.

Challenges and Contribution

1. Challenges
   - Sparseness: lack sufficient word co-occurrence or shared contexts for effective similarity measure-[Hu et al., 2009].
   - Informal: not well conformed as standard structures of documents.
   - Lack of label information. It is time and labor consuming to label the huge amount of messages.

Challenges and Contribution

1. Challenges
   - Sparseness: lack sufficient word co-occurrence or shared contexts for effective similarity measure-[Hu et al., 2009].
   - Informal: not well conformed as standard structures of documents.
   - Lack of label information. It is time and labor consuming to label the huge amount of messages.
2. Contribution
   - to handle data sparseness problem, we employ query related external resources from Google Search Engine to enrich the short messages.
   - to alleviate negative effect brought by informal words, we utilize linguistic corpus to detect informal words and correct them.
   - to require less labelled data, we attempt to use a semi-supervised learning approach for microblog categorization task.

Outline

the General Framework

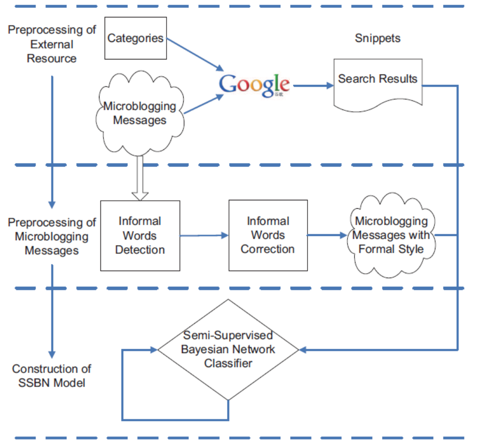


Figure: The General Framework.

Semi-Supervised Bayesian Network Graph Model

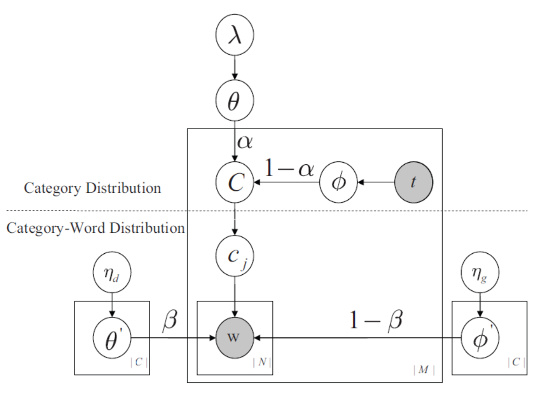


Figure: Probabilistic graphical representation of semi-supervised Bayesian network model.

Parameter Inference

The maximum likelihood category label for a given message m_i is,

$$y_i = \arg\max_{c_j} P(c_j|m_i, \hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}') = \frac{P(c_j|\hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}')P(m_i|c_j, \hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}')}{P(m_i|\hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}')}$$

Parameter Inference

The maximum likelihood category label for a given message m_i is,

$$y_i = \arg\max_{c_j} P(c_j|m_i,\hat{\theta},\hat{\phi},\hat{\theta}',\hat{\phi}') = \frac{\boxed{P(c_j|\hat{\theta},\hat{\phi},\hat{\theta}',\hat{\phi}')}P(m_i|c_j,\hat{\theta},\hat{\phi},\hat{\theta}',\hat{\phi}')}{P(m_i|\hat{\theta},\hat{\phi},\hat{\theta}',\hat{\phi}')}$$

$$P(c_j|\hat{\theta},\hat{\phi},\hat{\theta}',\hat{\phi}') = P(c_j|\hat{\theta},\hat{\phi}) = \hat{\alpha}P(c_j|\hat{\theta}) + (1-\hat{\alpha})P(c_j|\hat{\phi})$$

Parameter Inference

The maximum likelihood category label for a given message m_i is,

$$y_i = \arg\max_{c_j} P(c_j|m_i, \hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}') = \frac{P(c_j|\hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}')P(m_i|c_j, \hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}')}{\boxed{P(m_i|\hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}')}}$$

$$P(c_j|\hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}') = P(c_j|\hat{\theta}, \hat{\phi}) = \hat{\alpha}P(c_j|\hat{\theta}) + (1-\hat{\alpha})P(c_j|\hat{\phi})$$

$$P(m_i|\hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}') = \sum_{c_j} P(c_j|\hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}')P(m_i|c_j, \hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}')$$

Parameter Inference

The maximum likelihood category label for a given message m_i is,

$$y_i = \arg\max_{c_j} P(c_j|m_i, \hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}') = \frac{P(c_j|\hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}')\boxed{P(m_i|c_j, \hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}')}}{P(m_i|\hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}')}$$

$$P(c_j|\hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}') = P(c_j|\hat{\theta}, \hat{\phi}) = \hat{\alpha}P(c_j|\hat{\theta}) + (1-\hat{\alpha})P(c_j|\hat{\phi})$$

$$P(m_i|\hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}') = \sum_{c_j} P(c_j|\hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}')P(m_i|c_j, \hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}')$$

$$\begin{aligned} P(m_i|c_j, \hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}') &= P(m_i|c_j, \hat{\theta}', \hat{\phi}') = \prod_{k=1}^{|m_i|} P(w_k|c_j, \hat{\theta}', \hat{\phi}') \\ &= \prod_{k=1}^{|m_i|} \{\beta P(w_k|c_j, \hat{\theta}') + (1-\beta)P(w_k|c_j, \hat{\phi}')\} \end{aligned}$$

Estimating

1. Estimating θ:

$$\hat{\theta}_{c_j} \equiv P(c_j|\hat{\theta}) = \frac{1 + \sum_{i=1}^{|M|} \Lambda(i) P(y_i = c_j|m_i)}{|C| + |M^l| + \lambda|M^u|} \tag{1}$$

Estimating

1. Estimating θ:

$$\hat{\theta}_{c_j} \equiv P(c_j|\hat{\theta}) = \frac{1 + \sum_{i=1}^{|M|} \Lambda(i) P(y_i = c_j|m_i)}{|C| + |M^l| + \lambda|M^u|} \tag{1}$$

2. Estimating ϕ:

$$\hat{\phi}_{c_j} \equiv P(c_j|\hat{\phi}) = \frac{\frac{1}{NGD(t,c_j)} + \mu}{\sum_{j=1}^{|C|} \frac{1}{NGD(t,c_j)} + |C|\mu} \tag{2}$$

Estimating

1. Estimating θ:

$$\hat{\theta_{c_j}} \equiv P(c_j|\hat{\theta}) = \frac{1 + \sum_{i=1}^{|M|} \Lambda(i) P(y_i = c_j|m_i)}{|C| + |M^l| + \lambda |M^u|} \quad (1)$$

2. Estimating ϕ:

$$\hat{\phi_{c_j}} \equiv P(c_j|\hat{\phi}) = \frac{\frac{1}{NGD(t,c_j)} + \mu}{\sum_{j=1}^{|C|} \frac{1}{NGD(t,c_j)} + |C|\mu} \quad (2)$$

3. Estimating θ' and ϕ':

$$\hat{\theta}'^{w_k}_{c_j} \equiv P(w_k|c_j, \hat{\theta}') = \frac{n_{d}{}^{w_k}_{c_j} + \eta_d}{\sum_{p'=1}^{|N|} n_{d}{}^{w_{p'}}_{c_j} + |N|\eta_d} \quad (3)$$

$$\hat{\phi}'^{w_k}_{c_j} \equiv P(w_k|c_j, \hat{\phi}') = \frac{n_{g}{}^{w_k}_{c_j} + \eta_g}{\sum_{q'=1}^{|N|} n_{g}{}^{w_{q'}}_{c_j} + |N|\eta_g} \quad (4)$$

Outline

Datasets and Evaluation Metrics

Twitter		Sina Weibo	
Total	16935	Total	15811
Sports	2720	Sports	2602
Entertainment	2816	Movies	2694
Business	2912	Games	2605
Science&Tech	2827	Science&Tech	2647
Politics	2937	Politics	2654
Education	2723	Music	2609

Table: The distribution of different categories over two datasets.

Datasets and Evaluation Metrics

Twitter		Sina Weibo	
Total	16935	Total	15811
Sports	2720	Sports	2602
Entertainment	2816	Movies	2694
Business	2912	Games	2605
Science&Tech	2827	Science&Tech	2647
Politics	2937	Politics	2654
Education	2723	Music	2609

Table: The distribution of different categories over two datasets.

1. Apple, stock business
   - iBenApple Mon Jan 24 13:50:42 +0000 2011 #IHateItWhen Apple's stock continue to fall!
2. Apple, ipad science
   - Kericox3 Tue Feb 01 12:34:55 +0000 2011 Apple iphone 4g 32gb and blackberry bold 9700 Unlocked. - Anything ...: Apple Tablet iPad 64GB (Wi-Fi + 3G) http://bit.ly/gbbW1J

Datasets and Evaluation Metrics

Twitter		Sina Weibo	
Total	16935	Total	15811
Sports	2720	Sports	2602
Entertainment	2816	Movies	2694
Business	2912	Games	2605
Science&Tech	2827	Science&Tech	2647
Politics	2937	Politics	2654
Education	2723	Music	2609

Table: The distribution of different categories over two datasets.

- accuracy
- precision
- recall
- F_1

SSBN Model Performance

Twitter				Sina Weibo			
Category	*Precision*	*Recall*	*F1*	*Category*	*Precision*	*Recall*	*F1*
Sports	**0.9322**	0.9483	**0.9402**	Sports	**0.9318**	0.8747	**0.9023**
Entertainment	0.9000	0.5625	0.6923	Movies	0.8848	0.8207	0.8515
Business	0.8043	0.5323	0.6382	Games	0.8090	0.9283	0.8646
Science&Tech	0.6937	**0.9801**	0.8124	Science&Tech	0.8688	0.8323	0.8502
Politics	0.9096	0.9640	0.9360	Politics	0.8661	**0.9324**	0.8980
Education	0.5000	0.5519	0.5165	Music	0.8819	0.8699	0.8759
Micro-average	0.7979	0.7979	0.7979	Micro-average	0.8798	0.8798	0.8798
Macro-average	0.7934	0.6043	0.6128	Macro-average	0.8737	0.8764	0.8738

Table: Performance of SSBN model on two datasets with 5% training data and 95% testing data, respectively.

Baselines

- SVM
- Naive Bayesian
- K Nearest Neighbors
- Rocchio
- Labeled LDA
- Transductive SVM
- Semi-Naive Bayesian classifier

Comparison Performance

Classifier	*Accuracy*	*MicroP*	*MicroR*	*MicroF1*	*MacroP*	*MacroR*	*MacroF1*
SSBN	**0.8875**	**0.8875**	**0.8875**	**0.8875**	0.8282	0.7627	0.7845
SVM	0.8670	0.8670	0.8670	0.8670	0.8768	0.7611	**0.7860**
NB	0.8722	0.8696	0.8722	0.8722	**0.8879**	0.7329	0.7587
KNN	0.7268	0.7268	0.7268	0.7268	0.6721	0.6471	0.6516
Rocchio	0.8180	0.8204	0.8180	0.8192	0.7361	**0.8384**	0.7605
L-LDA	0.8605	0.8605	0.8605	0.8605	0.8467	0.7223	0.7532

Table: Performance comparison among SSBN and other supervised baseline methods on twitter with 90% training data.

Classifier	*Accuracy*	*MicroP*	*MicroR*	*MicroF1*	*MacroP*	*MacroR*	*MacroF1*
SSBN	**0.7979**	**0.7979**	**0.7979**	**0.7979**	**0.7934**	**0.6043**	**0.6128**
Trans-SVM	0.6707	0.6707	0.6707	0.6707	0.6602	0.5108	0.4491
Semi-NB	0.7156	0.7156	0.7156	0.7156	0.7308	0.5653	0.549

Table: Performance comparison among SSBN and other semi-supervised baseline methods on Twitter with 5% training data.

Comparison Performance

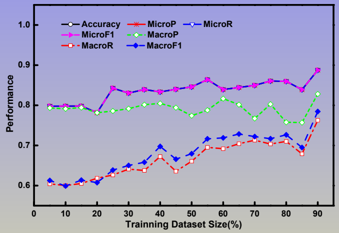
Classifier	*Accuracy*	*MicroP*	*MicroR*	*MicroF1*	*MacroP*	*MacroR*	*MacroF1*
SSBN	**0.9020**	**0.9020**	**0.9020**	**0.9020**	0.8976	**0.9045**	**0.9004**
SVM	0.8991	0.8991	0.8991	0.8991	**0.9017**	0.8971	0.8991
NB	0.9015	0.9015	0.9015	0.9015	0.8990	0.9024	0.9003
KNN	0.8565	0.8565	0.8565	0.8565	0.8589	0.8486	0.8526
Rocchio	0.8802	0.8803	0.8802	0.8802	0.8769	0.8832	0.8781
L-LDA	0.8905	0.8905	0.8905	0.8905	0.8876	0.8989	0.8932

Table: Performance comparison among SSBN and other supervised baseline methods on Sina Weibo with 90% training data.

Classifier	*Accuracy*	*MicroP*	*MicroR*	*MicroF1*	*MacroP*	*MacroR*	*MacroF1*
SSBN	**0.8798**	**0.8798**	**0.8798**	**0.8798**	**0.8737**	**0.8764**	**0.8738**
Trans-SVM	0.8084	0.8084	0.8084	0.8084	0.8049	0.8085	0.8052
Semi-NB	0.8198	0.8198	0.8198	0.8198	0.8225	0.8217	0.8204

Table: Performance comparison among SSBN and other semi-supervised baseline methods on Sina Weibo with 5% training data.

On the Sensitivity of Training Data Size



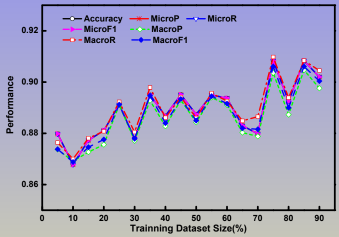
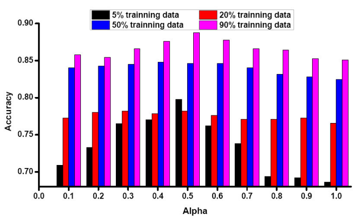


Figure: Performance sensitivity of training set size on Twitter and Sina Weibo

Effect of α

The trade-off parameter α is used to balance the effects of two kinds of prior knowledge at category level: microblogging data collection and external resources.



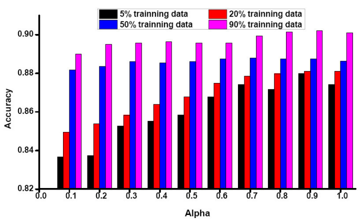
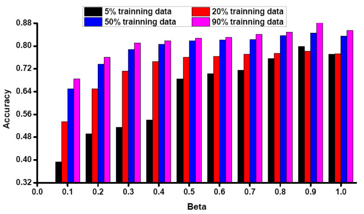


Figure: The Performance with varying α and training data size when other parameters are fixed.

Effect of β

There are two category-word distributions, θ' and ϕ', which are respectively generated from our data collection and google search results; and parameter β is utilized to adjust the contribution between these two different resources in category-word level.



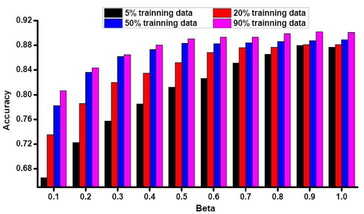


Figure: The Performance with varying β and training data size when other parameters are fixed.

Effect of λ

λ indicates the contribution from unlabeled data points, between 0 and 1.

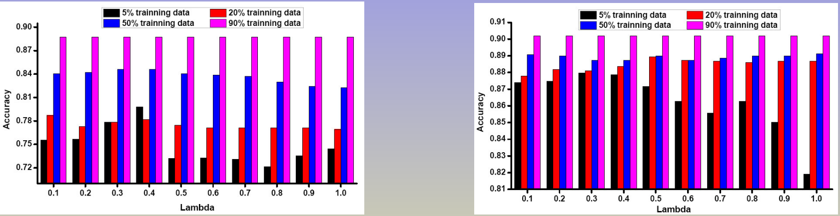


Figure: The Performance with varying λ and training data size when other parameters are fixed.

Outline

1. Conclusion
   - the incorporation of external resources to supplement the short microblogs well compensates the data sparseness issue;
   - the semi-supervised classifier seamlessly fuse labeled data structure and external resources into the training process, which reduced the requirement for manually labeling to a certain degree;
   - we model the category probability of a given message based on the category-word distribution, and this successfully avoided the difficulty brought about by the spelling errors that are common in microblogging messages.

1. Conclusion
    - the incorporation of external resources to supplement the short microblogs well compensates the data sparseness issue;
    - the semi-supervised classifier seamlessly fuse labeled data structure and external resources into the training process, which reduced the requirement for manually labeling to a certain degree;
    - we model the category probability of a given message based on the category-word distribution, and this successfully avoided the difficulty brought about by the spelling errors that are common in microblogging messages.
2. Future Work
    - the incorporation of social network structure can improve the performance of microblogging classification;
    - the use of external resources such as Wikipedia and WordNet might be valuable for understanding microblogging messages;
    - the provision of category summarization can help to organize microblogging messages.

References I

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

Hong, L. and Davison, B. D. (2010).
Empirical study of topic modeling in twitter.
In *Proceedings of KDD Workshop on Social Media Analytics*.

Hu, X., Sun, N., Zhang, C., and Chua, T.-S. (2009).
Exploiting internal and external semantics for the clustering of short texts using world knowledge.
In *Proceedings of the ACM conference on Information and knowledge management*.

Lee, K., Palsetia, D., Narayanan, R., Patwary, M. M. A., Agrawal, A., and Choudhary, A. (2011).
Twitter trending topic classification.
In *Proceedings of ICDM Workshop on Optimization Based Methods for Emerging Data Mining Problems*.

Ramage, D., Dumais, S., and Liebling, D. (2010).
Charaterizing microblog with topic models.
In *Proceedings of International AAAI Conference on Weblogs and Social Media*.

References II

Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009).
Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora.
In *Proceedings of International Conference on Empirical Methods in Natural Language Processing*.

Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., and Steyvers, M. (2010).
Learning author-topic models from text corpora.
ACM Transactions on Information Systems, 28:1–38.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., and Demirbas, M. (2010).
Short text classification in twitter to improve information filtering.
In *Proceedings of Annual ACM Conference on Research and Development in Information Retrieval*.

Thank you!