A Semi-Supervised Bayesian Network Model for Microblog Topic Classification

Yan Chen\textsuperscript{1,2} Zhoujun Li\textsuperscript{1} Liqiang Nie\textsuperscript{2} Xia Hu\textsuperscript{3} Xiangyu Wang\textsuperscript{2} Tat-seng Chua\textsuperscript{2} Xiaoming Zhang\textsuperscript{1}

\textsuperscript{1}State Key Laboratory of Software Development Environment, Beihang University, China
\textsuperscript{2}School of Computing, National University of Singapore, Singapore
\textsuperscript{3}Arizona State University, United States

11-12-2012
Outline

1. Background and Motivation
2. Related Work
3. Semi-Supervised Graphical Model
   - The General Framework
   - Probabilistic Graph Model Construction
   - Parameter Inference
4. Experiments
   - Experimental Settings
   - Analysis
   - Parameter Analysis
5. Conclusion and Future Work
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.
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

Yan Chen (Beihang University)
Motivation

Example 1:

[Screen capture of a website showing a tweet by Stoxx about DBS Bank (Taiwan). The tweet reads: "TEXT-S&P summary: DBS Bank (Taiwan) Ltd. (The following statement was released by the rating agency) Nov 30 - ... yhoo.it/11mXpPn."

DBS Bank]
Example 1:

In 1 afternoon’s Casino Royale film shooting, 3 Aston Martin DBS cars valued at $300,000 each were destroyed for the car roll sequence #007.
Motivation

Example 2:

- **#请问玛雅人靠谱吗#**
  - "玛雅体"来了，玛雅人靠谱的话我就XXX
  - PK 你觉得玛雅人靠谱吗？(共有4412条观点PK)
  - 红方：靠谱啊，所... (3004) VS 蓝方：不靠谱，老... (1408)

- **#一九四二#**
  - 《1942》上映了，来写一句话影评吧
  - 微影评 (共有12条相关微博)
  - 2012-11-11 13:25:16 _沈慕白_ 发布了第一条事件相关微博

- **#王的盛宴#**
  - 电影《王的盛宴》
  - 微影评 (共有6条相关微博)
  - 2012-11-11 13:25:16 _沈慕白_ 发布了第一条事件相关微博
Motivation

Example 2:
Motivation

Example 2:

![Image of Weibo page showing popular topics]

- "#请问玛雅人靠谱吗#
  "玛雅体"来了，玛雅人靠谱的话我就XXX
  PK 你觉得玛雅人靠谱吗？ (共有4412条观点PK)

- "#一九四二#
  (1942) 上映了，来写一句话影评吧
  微影评 (共有12条相关微博)
  2012-11-11 13:25:16 _沈易白_ 发布了第一条事件相关微博

- "#王的盛宴#
  电影《王的盛宴》
  微影评 (共有6条相关微博)
Background and Motivation

Motivation

Example 2:

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

Yan Chen (Beihang University)
Example 2:

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

Example 2:

Topic Classification!
Outline

1. Background and Motivation
2. Related Work
3. Semi-Supervised Graphical Model
   - The General Framework
   - Probabilistic Graph Model Construction
   - Parameter Inference
4. Experiments
   - Experimental Settings
   - Analysis
   - Parameter Analysis
5. Conclusion and Future Work
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.

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.
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

1. Background and Motivation
2. Related Work
3. Semi-Supervised Graphical Model
   - The General Framework
   - Probabilistic Graph Model Construction
   - Parameter Inference
4. Experiments
   - Experimental Settings
   - Analysis
   - Parameter Analysis
5. Conclusion and Future Work
the General Framework

Figure: The General Framework.
**Semi-Supervised Bayesian Network Graph Model**

**Figure:** Probabilistic graphical representation of semi-supervised Bayesian network model.
The maximum likelihood category label for a given message $m_i$ is,

$$y_i = \arg \max_{c_j} P(c_j|m_i, \hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}') = \frac{P(c_j|\hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}')P(m_i|c_j, \hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}')}{P(m_i|\hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}')}$$
The maximum likelihood category label for a given message $m_i$ is,

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

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

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

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

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

$$P(m_i|\hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}') = \sum_{c_j} P(c_j|\hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}')P(m_i|c_j, \hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}')$$
The maximum likelihood category label for a given message \(m_i\) is,

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

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

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

\[
P(m_i|c_j, \hat{\theta}, \hat{\phi}, \hat{\theta}', \hat{\phi}') = P(m_i|c_j, \hat{\theta}', \hat{\phi}') = \prod_{k=1}^{m_i} P(w_k|c_j, \hat{\theta}', \hat{\phi}')
\]

\[
= \prod_{k=1}^{m_i} \{\beta P(w_k|c_j, \hat{\theta}') + (1 - \beta)P(w_k|c_j, \hat{\phi}')\}
\]
Estimating

1. Estimating $\theta$:

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

(1)
Estimating

1. Estimating $\theta$:

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

2. Estimating $\phi$:

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

1. Estimating $\theta$:

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

2. Estimating $\phi$:

$$\hat{\phi}_{c_j} \equiv P(c_j|\hat{\phi}) = \frac{1}{NGD(t, c_j)} + \mu$$

$$\frac{1}{\sum_{j=1}^{|C|} \frac{1}{NGD(t, c_j)} + |C| \mu}$$  \hspace{1cm} (2)

3. Estimating $\theta'$ and $\phi'$:

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

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

1. Background and Motivation

2. Related Work

3. Semi-Supervised Graphical Model
   - The General Framework
   - Probabilistic Graph Model Construction
   - Parameter Inference

4. Experiments
   - Experimental Settings
   - Analysis
   - Parameter Analysis

5. Conclusion and Future Work
Datasets and Evaluation Metrics

<table>
<thead>
<tr>
<th>Category</th>
<th>Twitter</th>
<th>Sina Weibo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>16935</td>
<td>15811</td>
</tr>
<tr>
<td>Sports</td>
<td>2720</td>
<td>2602</td>
</tr>
<tr>
<td>Entertainment</td>
<td>2816</td>
<td>2694</td>
</tr>
<tr>
<td>Business</td>
<td>2912</td>
<td>2605</td>
</tr>
<tr>
<td>Science &amp; Tech</td>
<td>2827</td>
<td>2647</td>
</tr>
<tr>
<td>Politics</td>
<td>2937</td>
<td>2654</td>
</tr>
<tr>
<td>Education</td>
<td>2723</td>
<td>2609</td>
</tr>
</tbody>
</table>

Table: The distribution of different categories over two datasets.
Datasets and Evaluation Metrics

<table>
<thead>
<tr>
<th>Category</th>
<th>Twitter</th>
<th>Sina Weibo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>16935</td>
<td>15811</td>
</tr>
<tr>
<td>Sports</td>
<td>2720</td>
<td>2602</td>
</tr>
<tr>
<td>Entertainment</td>
<td>2816</td>
<td>2694</td>
</tr>
<tr>
<td>Business</td>
<td>2912</td>
<td>2605</td>
</tr>
<tr>
<td>Science&amp;Tech</td>
<td>2827</td>
<td>2647</td>
</tr>
<tr>
<td>Politics</td>
<td>2937</td>
<td>2654</td>
</tr>
<tr>
<td>Education</td>
<td>2723</td>
<td>2609</td>
</tr>
</tbody>
</table>

Table: The distribution of different categories over two datasets.

1. **Apple, stock business**
   - iBenApple Mon Jan 24 13:50:42 +0000 2011 #I Hate It When Apple's stock continue to fall!

2. **Apple, ipad science**
## Datasets and Evaluation Metrics

<table>
<thead>
<tr>
<th></th>
<th>Twitter</th>
<th>Sina Weibo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>16935</td>
<td>15811</td>
</tr>
<tr>
<td>Sports</td>
<td>2720</td>
<td>2602</td>
</tr>
<tr>
<td>Entertainment</td>
<td>2816</td>
<td>2694</td>
</tr>
<tr>
<td>Business</td>
<td>2912</td>
<td>2605</td>
</tr>
<tr>
<td>Science &amp; Tech</td>
<td>2827</td>
<td>2647</td>
</tr>
<tr>
<td>Politics</td>
<td>2937</td>
<td>2654</td>
</tr>
<tr>
<td>Education</td>
<td>2723</td>
<td>2609</td>
</tr>
</tbody>
</table>

**Table:** The distribution of different categories over two datasets.

- accuracy
- precision
- recall
- $F_1$
## SSBN Model Performance

<table>
<thead>
<tr>
<th>Category</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Category</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sports</td>
<td>0.9322</td>
<td>0.9483</td>
<td><strong>0.9402</strong></td>
<td>Sports</td>
<td>0.9318</td>
<td>0.8747</td>
<td><strong>0.9023</strong></td>
</tr>
<tr>
<td>Entertainment</td>
<td>0.9000</td>
<td>0.5625</td>
<td>0.6923</td>
<td>Movies</td>
<td>0.8848</td>
<td>0.8207</td>
<td>0.8515</td>
</tr>
<tr>
<td>Business</td>
<td>0.8043</td>
<td>0.5323</td>
<td>0.6382</td>
<td>Games</td>
<td>0.8090</td>
<td>0.9283</td>
<td>0.8646</td>
</tr>
<tr>
<td>Science&amp;Tech</td>
<td>0.6937</td>
<td><strong>0.9801</strong></td>
<td>0.8124</td>
<td>Science&amp;Tech</td>
<td>0.8688</td>
<td>0.8323</td>
<td>0.8502</td>
</tr>
<tr>
<td>Politics</td>
<td>0.9096</td>
<td>0.9640</td>
<td>0.9360</td>
<td>Politics</td>
<td>0.8661</td>
<td><strong>0.9324</strong></td>
<td>0.8980</td>
</tr>
<tr>
<td>Education</td>
<td>0.5000</td>
<td>0.5519</td>
<td>0.5165</td>
<td>Music</td>
<td>0.8819</td>
<td>0.8699</td>
<td>0.8759</td>
</tr>
<tr>
<td>Micro-average</td>
<td>0.7979</td>
<td>0.7979</td>
<td>0.7979</td>
<td>Micro-average</td>
<td>0.8798</td>
<td>0.8798</td>
<td>0.8798</td>
</tr>
<tr>
<td>Macro-average</td>
<td>0.7934</td>
<td>0.6043</td>
<td>0.6128</td>
<td>Macro-average</td>
<td>0.8737</td>
<td>0.8764</td>
<td>0.8738</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Category</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science&amp;Tech</td>
<td>0.8688</td>
<td>0.8323</td>
<td>0.8502</td>
<td>Science&amp;Tech</td>
<td>0.8688</td>
<td>0.8323</td>
<td>0.8502</td>
</tr>
<tr>
<td>Politics</td>
<td>0.8661</td>
<td><strong>0.9324</strong></td>
<td>0.8980</td>
<td>Music</td>
<td>0.8819</td>
<td>0.8699</td>
<td>0.8759</td>
</tr>
<tr>
<td>Micro-average</td>
<td>0.8798</td>
<td>0.8798</td>
<td>0.8798</td>
<td>Micro-average</td>
<td>0.8798</td>
<td>0.8798</td>
<td>0.8798</td>
</tr>
<tr>
<td>Macro-average</td>
<td>0.8737</td>
<td>0.8764</td>
<td>0.8738</td>
<td>Macro-average</td>
<td>0.8737</td>
<td>0.8764</td>
<td>0.8738</td>
</tr>
</tbody>
</table>

### 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

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>MicroP</th>
<th>MicroR</th>
<th>MicroF1</th>
<th>MacroP</th>
<th>MacroR</th>
<th>MacroF1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSBN</td>
<td>0.8875</td>
<td>0.8875</td>
<td>0.8875</td>
<td>0.8875</td>
<td>0.8282</td>
<td>0.7627</td>
<td>0.7845</td>
</tr>
<tr>
<td>SVM</td>
<td>0.8670</td>
<td>0.8670</td>
<td>0.8670</td>
<td>0.8670</td>
<td>0.8768</td>
<td>0.7611</td>
<td>0.7760</td>
</tr>
<tr>
<td>NB</td>
<td>0.8722</td>
<td>0.8696</td>
<td>0.8722</td>
<td>0.8722</td>
<td>0.8879</td>
<td>0.7329</td>
<td>0.7587</td>
</tr>
<tr>
<td>KNN</td>
<td>0.7268</td>
<td>0.7268</td>
<td>0.7268</td>
<td>0.7268</td>
<td>0.6721</td>
<td>0.6471</td>
<td>0.6516</td>
</tr>
<tr>
<td>Rocchio</td>
<td>0.8180</td>
<td>0.8204</td>
<td>0.8180</td>
<td>0.8192</td>
<td>0.7361</td>
<td>0.8384</td>
<td>0.7605</td>
</tr>
<tr>
<td>L-LDA</td>
<td>0.8605</td>
<td>0.8605</td>
<td>0.8605</td>
<td>0.8605</td>
<td>0.8467</td>
<td>0.7223</td>
<td>0.7532</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>MicroP</th>
<th>MicroR</th>
<th>MicroF1</th>
<th>MacroP</th>
<th>MacroR</th>
<th>MacroF1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSBN</td>
<td>0.7979</td>
<td>0.7979</td>
<td>0.7979</td>
<td>0.7979</td>
<td>0.7934</td>
<td>0.6043</td>
<td>0.6128</td>
</tr>
<tr>
<td>Trans-SVM</td>
<td>0.6707</td>
<td>0.6707</td>
<td>0.6707</td>
<td>0.6707</td>
<td>0.6602</td>
<td>0.5108</td>
<td>0.4491</td>
</tr>
<tr>
<td>Semi-NB</td>
<td>0.7156</td>
<td>0.7156</td>
<td>0.7156</td>
<td>0.7156</td>
<td>0.7308</td>
<td>0.5653</td>
<td>0.549</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>MicroP</th>
<th>MicroR</th>
<th>MicroF1</th>
<th>MacroP</th>
<th>MacroR</th>
<th>MacroF1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSBN</td>
<td>0.9020</td>
<td>0.9020</td>
<td>0.9020</td>
<td>0.9020</td>
<td>0.8976</td>
<td>0.9045</td>
<td>0.9004</td>
</tr>
<tr>
<td>SVM</td>
<td>0.8991</td>
<td>0.8991</td>
<td>0.8991</td>
<td>0.8991</td>
<td>0.9017</td>
<td>0.8971</td>
<td>0.9004</td>
</tr>
<tr>
<td>NB</td>
<td>0.9015</td>
<td>0.9015</td>
<td>0.9015</td>
<td>0.9015</td>
<td>0.9017</td>
<td>0.9024</td>
<td>0.9003</td>
</tr>
<tr>
<td>KNN</td>
<td>0.8565</td>
<td>0.8565</td>
<td>0.8565</td>
<td>0.8565</td>
<td>0.8589</td>
<td>0.8486</td>
<td>0.8526</td>
</tr>
<tr>
<td>Rocchio</td>
<td>0.8802</td>
<td>0.8803</td>
<td>0.8802</td>
<td>0.8802</td>
<td>0.8769</td>
<td>0.8832</td>
<td>0.8781</td>
</tr>
<tr>
<td>L-LDA</td>
<td>0.8905</td>
<td>0.8905</td>
<td>0.8905</td>
<td>0.8905</td>
<td>0.8876</td>
<td>0.8989</td>
<td>0.8932</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>MicroP</th>
<th>MicroR</th>
<th>MicroF1</th>
<th>MacroP</th>
<th>MacroR</th>
<th>MacroF1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSBN</td>
<td>0.8798</td>
<td>0.8798</td>
<td>0.8798</td>
<td>0.8798</td>
<td>0.8737</td>
<td>0.8764</td>
<td>0.8738</td>
</tr>
<tr>
<td>Trans-SVM</td>
<td>0.8084</td>
<td>0.8084</td>
<td>0.8084</td>
<td>0.8084</td>
<td>0.8049</td>
<td>0.8085</td>
<td>0.8052</td>
</tr>
<tr>
<td>Semi-NB</td>
<td>0.8198</td>
<td>0.8198</td>
<td>0.8198</td>
<td>0.8198</td>
<td>0.8225</td>
<td>0.8217</td>
<td>0.8204</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>MicroP</th>
<th>MicroR</th>
<th>MicroF1</th>
<th>MacroP</th>
<th>MacroR</th>
<th>MacroF1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSBN</td>
<td>0.9020</td>
<td>0.9020</td>
<td>0.9020</td>
<td>0.9020</td>
<td>0.8976</td>
<td>0.9045</td>
<td>0.9004</td>
</tr>
<tr>
<td>SVM</td>
<td>0.8991</td>
<td>0.8991</td>
<td>0.8991</td>
<td>0.8991</td>
<td>0.9017</td>
<td>0.8971</td>
<td>0.9004</td>
</tr>
<tr>
<td>NB</td>
<td>0.9015</td>
<td>0.9015</td>
<td>0.9015</td>
<td>0.9015</td>
<td>0.9017</td>
<td>0.9024</td>
<td>0.9003</td>
</tr>
<tr>
<td>KNN</td>
<td>0.8565</td>
<td>0.8565</td>
<td>0.8565</td>
<td>0.8565</td>
<td>0.8589</td>
<td>0.8486</td>
<td>0.8526</td>
</tr>
<tr>
<td>Rocchio</td>
<td>0.8802</td>
<td>0.8803</td>
<td>0.8802</td>
<td>0.8802</td>
<td>0.8769</td>
<td>0.8832</td>
<td>0.8781</td>
</tr>
<tr>
<td>L-LDA</td>
<td>0.8905</td>
<td>0.8905</td>
<td>0.8905</td>
<td>0.8905</td>
<td>0.8876</td>
<td>0.8989</td>
<td>0.8932</td>
</tr>
</tbody>
</table>

**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 $\alpha$

The trade-off parameter $\alpha$ 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 $\alpha$ and training data size when other parameters are fixed.
Effect of $\beta$

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

**Figure:** The Performance with varying $\beta$ and training data size when other parameters are fixed.
Effect of $\lambda$

$\lambda$ indicates the contribution from unlabeled data points, between 0 and 1.

Figure: The Performance with varying $\lambda$ and training data size when other parameters are fixed.
Outline

1. Background and Motivation
2. Related Work
3. Semi-Supervised Graphical Model
   - The General Framework
   - Probabilistic Graph Model Construction
   - Parameter Inference
4. Experiments
   - Experimental Settings
   - Analysis
   - Parameter Analysis
5. Conclusion and Future Work
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.

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.
## 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.

## 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

Latent dirichlet allocation.

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

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.

Twitter trending topic classification.
In Proceedings of ICDM Workshop on Optimization Based Methods for Emerging Data Mining Problems.

Characterizing microblog with topic models.
In Proceedings of International AAAI Conference on Weblogs and Social Media.


Thank you!