User Behavior Analysis for Sentiment Classification

- **User Modeling with Neural Network for Review Rating Prediction.**
  
  *Duyu Tang, Bing Qin, Ting Liu. IJCAI 2015, full paper.*

- **Learning Semantic Representations of Users and Products for Document Level Sentiment Classification.**
  
  *Duyu Tang, Bing Qin, Ting Liu. ACL 2015, full paper.*
Sentiment Classification

• Given a piece of text, the task aims to determine its polarity as
  • Positive / Negative 🙌👎
  • 1-5 stars ⭐⭐⭐⭐⭐

• The task can be at
  • Word/phrase level, sentence level, document level

• We target at document-level sentiment classification in this work
Standard Supervised Learning Pipeline

<table>
<thead>
<tr>
<th>Feature</th>
<th>Feature Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-Caps</td>
<td>the number of words with all characters in upper case</td>
</tr>
<tr>
<td>Emoticon</td>
<td>the presence of positive (or negative) emoticons, whether the last unit is emoticon</td>
</tr>
<tr>
<td>Hashtag</td>
<td>the number of hashtag</td>
</tr>
<tr>
<td>Elongated units</td>
<td>the number of basic computational containing elongated words (with one character</td>
</tr>
<tr>
<td></td>
<td>repeated more than two times), such as $\text{goood}$</td>
</tr>
<tr>
<td>Sentiment lexicon</td>
<td>the number of sentiment words, the score of last sentiment words, the total sentiment</td>
</tr>
<tr>
<td></td>
<td>score and the maximal sentiment score for each lexicon</td>
</tr>
<tr>
<td>Negation</td>
<td>the number of negations as individual units in a segmentation</td>
</tr>
<tr>
<td>Bag-of-Units</td>
<td>an extension of bag-of-word for a segmentation</td>
</tr>
<tr>
<td>Punctuation</td>
<td>the number of contiguous sequences of dot, question mark and exclamation mark.</td>
</tr>
<tr>
<td>Cluster</td>
<td>the presence of units from each of the 1,000 clusters from Twitter NLP tool (Gimpel</td>
</tr>
<tr>
<td></td>
<td>et al., 2011)</td>
</tr>
</tbody>
</table>
Feature Learning Pipeline

1. Training Data
2. Feature Representation
3. Learning Algorithm
4. Sentiment Classifier

Learn text representation/feature from data!
Deep Learning Pipeline

- Training Data
- Feature Representation
- Learning Algorithm
- Sentiment Classifier

Semantic Composition
Word Representation
Words

$w_1, w_2, \ldots, w_{n-1}, w_n$
User Behavior is important for SA

- From a sentiment analysis perspective, users have different habits to:
  - Assign ratings on IMDB, Yelp...
  - Use diverse sentiment words to express one’s sentiment

<table>
<thead>
<tr>
<th>User 1</th>
<th>Rating Histories</th>
<th>Frequently used words</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>★★★★★ ★★★★☆☆</td>
<td>good, ok, just soso, disgusting...</td>
</tr>
<tr>
<td></td>
<td>★★★★☆☆ ★★★★★</td>
<td></td>
</tr>
<tr>
<td></td>
<td>★★★☆☆ ★★★★☆</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User 2</th>
<th>Rating Histories</th>
<th>Frequently used words</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>★★★★★ ★★★★★</td>
<td>awesome, amazing, excellent, not bad...</td>
</tr>
<tr>
<td></td>
<td>★★★★★ ★★★★☆☆</td>
<td></td>
</tr>
<tr>
<td></td>
<td>★★★★☆☆ ★★★★★</td>
<td></td>
</tr>
</tbody>
</table>
The Approach

• Text semantics

**Softmax**

**Pooling**

**Convolution**

**Tanh**

**Linear**

**Lookup**

\[ w_j: \text{word} \]
The Approach

• Text semantics
  • +User-Text Associations

*Softmax*

*Pooling*

*Convolution*

*Tanh*

*Linear*

*Lookup*

\[ U_k: \text{user} \quad w_i: \text{word} \]
The Approach

• Text semantics
  • +User-Text Associations
  • +User-Sentiment Associations
The Approach

- Text semantics
  - +User-Text Associations +Product-Text Associations
  - +User-Sentiment Associations +Product-Sentiment Associations

The model in ACL 2015
Experiments

• We conduct supervised learning on three datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#users</th>
<th>#products</th>
<th>#reviews</th>
<th>#docs/user</th>
<th>#docs/product</th>
<th>#sents/doc</th>
<th>#words/doc</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMDB</td>
<td>1,310</td>
<td>1,635</td>
<td>84,919</td>
<td>64.82</td>
<td>51.94</td>
<td>16.08</td>
<td>394.6</td>
</tr>
<tr>
<td>Yelp 2014</td>
<td>4,818</td>
<td>4,194</td>
<td>231,163</td>
<td>47.97</td>
<td>55.11</td>
<td>11.41</td>
<td>196.9</td>
</tr>
<tr>
<td>Yelp 2013</td>
<td>1,631</td>
<td>1,633</td>
<td>78,966</td>
<td>48.42</td>
<td>48.36</td>
<td>10.89</td>
<td>189.3</td>
</tr>
</tbody>
</table>

• Some details
  • We build the datasets by ourselves.
  • Split each corpus into training, development and testing sets with a 80/10/10 split
  • IMBD is from Diao et al. 2014
  • Yelp 2013 & 2014 come from Yelp Dataset Challenge
Experimental Results

• On IMDB dataset

![Bar Chart](image-url)
Experimental Results

• On Yelp 2014 dataset
Experimental Results

• On Yelp 2013 dataset

Our Approach
In Summary

• We encode users (and products) in semantic vector space, and apply them to sentiment classification.

• User and product representations can improve the sentiment classification accuracy.
Thanks!

codes and resources will be publicly available at:
http://ir.hit.edu.cn/~dytang