Cross-Domain Depression Detection via Harvesting Social Media

Tiancheng Shen¹, Jia Jia¹, Guangyao Shen¹, Fuli Feng², Xiangnan He², Huanbo Luan¹, Jie Tang¹, Thanassis Tiropanis³, Tat-Seng Chua², Wendy Hall³

1. Department of Computer Science and Technology, Tsinghua University
2. School of Computing, National University of Singapore
3. Electronics and Computer Science, University of Southampton

17 July 2018 @ IJCAI 2018, Stockholm, Sweden
Why Cross-Domain Depression Detection?

- Depression detection: a significant issue for human well-being
- Traditional psychological diagnosis: reliable but reactive
- Online detection via social media: success in Twitter, proactive
- Labeling of depressed users: questionnaire vs self-reported sentence pattern matching
  (matching expressions like “I’m diagnosed with depression” in user-generated content)
- In Twitter: a large well-labeled dataset [Shen et al., 2017]
- Replication in other platforms: cultural differences, insufficient data for model training
- Problem: can we utilize the multi-source datasets to enhance depression detection for a certain platform?
Problem Formulation

- $N$: total number of users
- $M$: dimensionality of feature vector
- $X \in \mathbb{R}^{N \times M}$: feature matrix
- $y \in \mathbb{R}^N$: binary depression states of users
- $D = \{X, y\}$: dataset of the social media
- $D_S / D_T$: datasets of the source / target domain
- $D_T$: limited labeled in model training
- $D_T = D_{TL} \cup D_{TU}$, $D_{TL} = \{X_{TL}, y_{TL}\}$, $D_{TU} = \{X_{TU}\}$
  $N_T = N_{TU} + N_{TL}$, $N_{TL} \ll N_{TU}$
- objective function $f$: $\{D_S, D_{TL}, D_{TU}\} \rightarrow y_{TU}$
Each sample includes:

- 4 weeks of tweets data + user profile

Weibo dataset $D_T$:

- 580 depressed samples by self-report sentence pattern labeling
- "^[^【@如]*[^【@怀疑想似像觉认能怕曾前@]{2,3}(我|自己)[^们会怀疑似觉想认早快怕要易像点能曾前她他你家国的它没不@]*[诊患得有][^点]{0,5}抑郁"
- 580 non-depressed samples that have no tweets containing “depress”.

Twitter dataset $D_S$ [Shen et al., 2017]:

- 1394 depressed samples & 1394 non-depressed samples

<table>
<thead>
<tr>
<th>Dataset</th>
<th>$D_T(+)$</th>
<th>$D_T(-)$</th>
<th>$D_S(+)$</th>
<th>$D_S(-)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users</td>
<td>580</td>
<td>580</td>
<td>1,394</td>
<td>1,394</td>
</tr>
<tr>
<td>Tweets</td>
<td>45,461</td>
<td>30,920</td>
<td>290,886</td>
<td>1,119,466</td>
</tr>
</tbody>
</table>
Feature Extraction

- $\mathcal{D}_T$: **78 features** (including 18 $\mathcal{D}_T$-exclusive features).
- $\mathcal{D}_S$: **60 features** in [Shen et al., 2017], including 60 features in common

### Table 1: Summary of features, where $\#_S$ and $\#_T$ denote the feature dimensionality of Twitter and Weibo, respectively.

<table>
<thead>
<tr>
<th>Group</th>
<th>Feature</th>
<th>$#_S$</th>
<th>$#_T$</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Textual</td>
<td>Emotional Word Count</td>
<td>2</td>
<td>2</td>
<td>The number of positive and negative emotional words.</td>
</tr>
<tr>
<td></td>
<td>Emoticon Count</td>
<td>3</td>
<td>3</td>
<td>The number of positive, neutral and negative emoticons.</td>
</tr>
<tr>
<td></td>
<td>Pronoun Count</td>
<td>2</td>
<td>3</td>
<td>The number of first-person singular/plural pronouns, and other personal pronouns.</td>
</tr>
<tr>
<td></td>
<td>Punctuation Count</td>
<td>3</td>
<td></td>
<td>The number of 3 typical punctuations (',', '?', '...').</td>
</tr>
<tr>
<td></td>
<td>Topic-Related Word Count</td>
<td>8</td>
<td></td>
<td>The number of words related to biology, body, health, death, society, money, work and leisure.</td>
</tr>
<tr>
<td></td>
<td>Text Length</td>
<td>1</td>
<td>1</td>
<td>The mean length of the tweet texts.</td>
</tr>
<tr>
<td>Visual</td>
<td>Saturation &amp; Brightness</td>
<td>4</td>
<td>4</td>
<td>The mean value of saturation and brightness, and their contrasts.</td>
</tr>
<tr>
<td></td>
<td>Warm/Clear Color</td>
<td>2</td>
<td>2</td>
<td>Ratio of colors with hue in [30, 110] and colors with saturation &lt; 0.7.</td>
</tr>
<tr>
<td></td>
<td>Five-Color Theme</td>
<td>15</td>
<td>15</td>
<td>A combination of five dominant colors in HSV color space.</td>
</tr>
<tr>
<td>User Profile &amp; Posting Behaviour</td>
<td>User Profile</td>
<td>2</td>
<td></td>
<td>Gender and length of screen name.</td>
</tr>
<tr>
<td></td>
<td>Tweet Count</td>
<td>2</td>
<td>2</td>
<td>The number of tweets published in the certain 4 weeks and ever since.</td>
</tr>
<tr>
<td></td>
<td>Tweeting Type</td>
<td>1</td>
<td>2</td>
<td>The proportion of original tweets and tweets with pictures.</td>
</tr>
<tr>
<td></td>
<td>Tweeting Time</td>
<td>24</td>
<td>24</td>
<td>The proportion of tweets posted in each hour of the day.</td>
</tr>
<tr>
<td>Social Interaction</td>
<td>Social Engagement</td>
<td>1</td>
<td>3</td>
<td>The number of retweets, comments and mentions per tweet.</td>
</tr>
<tr>
<td></td>
<td>Follow &amp; Favorites</td>
<td>3</td>
<td>4</td>
<td>The number of followers, friends and favorites and proportion of bi-followers.</td>
</tr>
</tbody>
</table>
Data Analysis: Isomerism

- One feature may follow distinctive integral distributions in different domains.
- Unrelated to specific user groups (depressed / non-depressed users).
- Example: follower count
- The same value of feature might have different implications across domains
- Quite common in the dataset
- Normalization methods: min-max & z-score, unsatisfactory
Due to cultural differences, the same feature may have distinctive, or even opposite implications on depression detection in different domains.

Such features: referred to as *divergent features*

Example: recent tweet count.

may tremendously impact the validity of transfer methods.
Methodology

DNN-FATC: A cross-domain Deep Neural Network model with Feature Adaptive Transformation & Combination strategy

- Based mainly on $D_S$

- **Shared features:**
  - Against isomerism
  - Against divergency

- **$D_T$-exclusive features**: Integrated in the end
Feature Normalization & Alignment (FNA)

- Targeted on isomerism
- A linear transformation to fill the distributional gap by minimizing the Bhattacharyya distance.

\[ x_S^* = a_S x_S + b_S, \quad x_T^* = a_T x_T + b_T, \]

subject to \( a_S, a_T, b_S, b_T = \arg \min_{a_S, a_T, b_S, b_T} - \ln \sum_{i=1}^{K} \sqrt{p_S^* p_T^*} \)

- Two more constraints:

\[ a_T Q(x_T, 50) + b_T = 0, \]
\[ a_S [Q(x_S, q_2) - Q(x_S, q_1)] = l \]

- Minimize the distinction between the feature spaces
- Train a DNN \( \mathcal{H}_S \) based on \( \mathcal{D}_S \)
Divergent Feature Conversion (DFC)

- Targeted on divergency

\[ x_{T_i}^{**} = \alpha_i x_{T_i}^* + \beta_i, \quad x_{T}^{**} = \alpha x_T + \beta I \]

- Identify singular features: better performance is achieved with \( \alpha_i < 0 \)

\[ (\alpha^*, \beta^*) = \arg \max_{\alpha, \beta} \mathcal{F}(\mathcal{H}_S, \{\alpha X_{TL}^* + \beta I, y_{TL}\}) \]

- Complexity of enumeration: \( \mathcal{O}( (|\alpha_i| \cdot |\beta_i|)^M_S) \)

- \( W \): an upper bound for times of enumeration.

- In each iteration, traverse all features in a random sequence, determine \( \alpha_i, \beta_i \) orderly; record \( \alpha^*, \beta^* \) for the best performance in \( W \) trials.

- Extra constraint: \( \alpha_i \in \{-1, 1\}, \beta_i = 0 \), centrosymmetric transformation

- \( \sigma \): a threshold of performance improvement to avoid overfitting.
Feature Combination (FC)

- $\mathcal{H}_T$: weights initialized to those of $\mathcal{H}_s$, trained on $\mathcal{D}_{TL}$ via back propagation
Experimental Setup: Dataset

$D_S$: 2,788 samples (Twitter Dataset)
$D_T$: 1,160 samples (Weibo Dataset)
$D_{TL}$: 280 samples ($\approx$10% the size of $D_S$)
$D_{TU}$: 880 samples (for testing)
Experimental Setup: Compared Methods

- **Feature normalization methods:**
  - **MN:** Min-Max Normalization
  - **ZN:** Zero-Mean Normalization
  - **FNA:** Feature Normalization & Alignment

- **Heterogeneous transfer learning methods:**
  - **ARC-t** [Kulis, Saenko, and Darrell 2011]
  - **MMDT** [Hoffman et al., 2013]
  - **HFA** [Li et al., 2014]

- **Utilization of $\mathcal{D}_S$ and $\mathcal{D}_T$:**
  - **Direct Learning (DL).** Learning a DNN merely on $\mathcal{D}_T$.
  - **Direct Learning of Shared features (DL$_S$).** Learning a DNN on $\mathcal{D}_T$ with the shared features.
  - **Direct Transfer (DT).** Learning a DNN on $\mathcal{D}_S$ and directly applying it on $\mathcal{D}_T$.
  - **Back Propagation (BP).** After $\mathcal{H}_S$ is learned on $\mathcal{D}_S$, retrain it on $\mathcal{D}_{TL}$ by back propagation.
  - **Divergent Feature Conversion (DFC).**
  - **Feature Combination (FC).**
Experimental Results: Performance

Table 3: F1-measure of method combinations in DNN-FATC.

<table>
<thead>
<tr>
<th></th>
<th>DL&lt;sub&gt;S&lt;/sub&gt;</th>
<th>DL</th>
<th>DT</th>
<th>BP</th>
<th>DFC</th>
<th>DFC+FC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MN</td>
<td>60.5±7.9</td>
<td>64.2±5.2</td>
<td>34.3±12.1</td>
<td>61.2±6.9</td>
<td>66.6±1.6</td>
<td>67.4±4.5</td>
</tr>
<tr>
<td>ZN</td>
<td>68.2±4.8</td>
<td>70.1±2.2</td>
<td>58.6±2.2</td>
<td>73.0±2.3</td>
<td>72.3±2.2</td>
<td>73.9±2.1</td>
</tr>
<tr>
<td>FNA</td>
<td>72.0±3.2</td>
<td>73.3±2.7</td>
<td>68.0±1.3</td>
<td>75.9±1.8</td>
<td>77.6±1.1</td>
<td>78.5±1.2</td>
</tr>
</tbody>
</table>

Table 4: F1-measure of heterogeneous transfer methods.

<table>
<thead>
<tr>
<th></th>
<th>FNA+DL</th>
<th>ARC-t</th>
<th>MMDT</th>
<th>HFA</th>
<th>DNN-FATC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>73.3±2.7</td>
<td>73.7±1.1</td>
<td>73.9±1.8</td>
<td>75.1±0.8</td>
<td>78.5±1.2</td>
</tr>
</tbody>
</table>

FNA vs MN / ZN: effectiveness in reducing isomerism
FSC vs BP: effectiveness in handling divergency
DL vs DL<sub>S</sub>, DFC+FC vs DFC: effectiveness of $\mathcal{D}_T$-exclusive features, utilization method $\mathcal{D}_S$ is useful to enhance detection in $\mathcal{D}_T$ and DNN-FATC best fits the assignment.
Experimental Results: Further Analysis

• Aforesaid performance: \( K = 100, \ W = 50, \ q_1 = q_2 = 25, \ l = 0.5, \ \sigma = 0.01, \ d = 4, \ \delta = 2 \)

• Parameter Analysis: limited Impact on performance

• Data Scalability Analysis:

• Feature Group Analysis:
Case Study: Depressive Behavior Discovery

• Depressive Behavior:
  Tweet time
  Gender
  Linguistic pattern
  Retweet count

• Divergent Features:
  Tweet count
  Image saturation
  follower count
  positive word count

(e) Temporal comparison
(f) Feature comparison
(g) Divergent features
Conclusion

- Raised the problem of enhancing depression detection via social media with multi-source datasets.

- Proposed a cross-domain Deep Neural Network model with Feature Adaptive Transformation & Combination strategy (DNN-FATC) to transfer the relevant information across heterogeneous domains.

- We expect the research to assist online depression detection for more countries, and contribute to the well-being of more people.

- **Future work:**
  - Beyond binary classification: more fine-grained detection
  - Further improve online detection by combining offline researches
Thank you!