Cross-Domain Depression Detection via Harvesting Social Media

Tiancheng Shen[1], Jia Jia[1], Guangyao Shen[1], Fuli Feng[2], Xiangnan He[2], Huanbo Luan[1], Jie Tang[1], Thanassis Tiropanis[3], Tat-Seng Chua[2], Wendy Hall[3]

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
- $\boldsymbol{X} \in \mathbb{R}^{N \times M}$: feature matrix
- $\boldsymbol{y} \in \mathbb{R}^{N}$: binary depression states of users
- $\mathcal{D} = \{\boldsymbol{X}, \boldsymbol{y}\}$: dataset of the social media

- **$\boldsymbol{\mathcal{D}_S}$ / $\boldsymbol{\mathcal{D}_T}$: datasets of the source / target domain**
- **$\boldsymbol{\mathcal{D}_T}$: limited labeled in model training**
- $\mathcal{D}_T = \mathcal{D}_{TL} \cup \mathcal{D}_{TU}$, $\mathcal{D}_{TL} = \{\boldsymbol{X}_{TL}, \boldsymbol{y}_{TL}\}$, $\mathcal{D}_{TU} = \{\boldsymbol{X}_{TU}\}$

  $N_T = N_{TU} + N_{TL}$, $N_{TL} \ll N_{TU}$
- **objective function $\boldsymbol{f}$: $\{\boldsymbol{\mathcal{D}_S}, \boldsymbol{\mathcal{D}_{TL}}, \boldsymbol{\mathcal{D}_{TU}}\} \rightarrow \boldsymbol{y_{TU}}$**

Dataset Construction

- **Each sample includes：**
  - 4 weeks of tweets data + user profile

Table 2: Datasets. $+(-)$ denotes (non-) depressed samples.

Dataset	$\mathcal{D}_T(+)$	$\mathcal{D}_T(-)$	$\mathcal{D}_S(+)$	$\mathcal{D}_S(-)$
Users	580	580	1,394	1,394
Tweets	45,461	30,920	290,886	1,119,466

- **Weibo dataset $\mathcal{D}_T$:**
  - 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][^不点]{0,5}抑郁”
  - 580 non-depressed samples that have no tweets containing “depress”.

- **Twitter dataset $\mathcal{D}_S$ [Shen et al., 2017]:**
  - 1394 depressed samples & 1394 non-depressed samples

Feature Extraction

- $\mathcal{D}_T$: 78 features
- $\mathcal{D}_S$: 115 features in [Shen et al., 2017], including 60 features in common

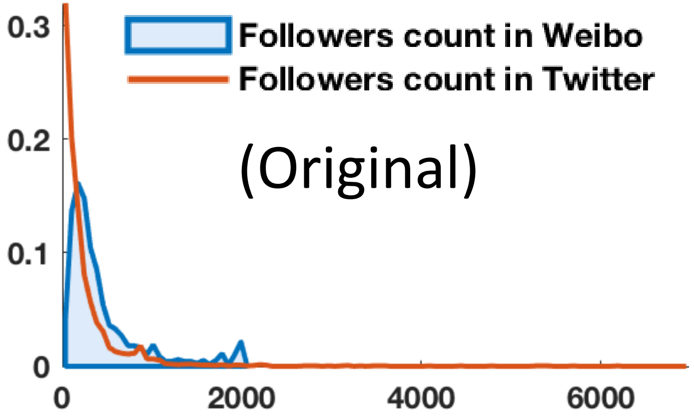
- $\mathcal{D}_T$: 78 features (including 18 $\mathcal{D}_T$-exclusive features).
- $\mathcal{D}_S$: 115 features in [Shen et al., 2017], including 60 features in common

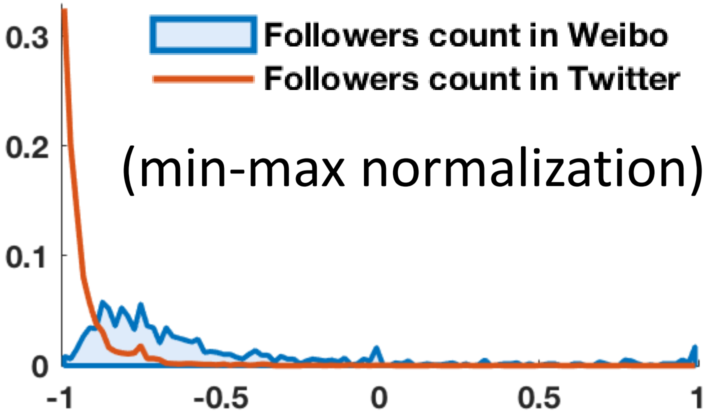
Table 1: Summary of features, where $\#_S$ and $\#_T$ denote the feature dimensionality of Twitter and Weibo, respectively.

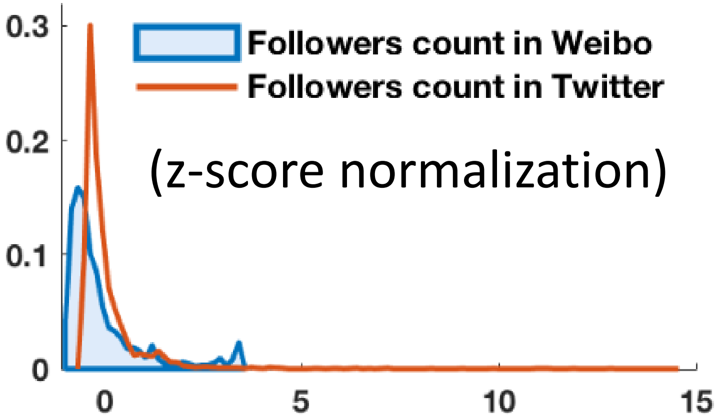
Group	Feature	$\#_S$	$\#_T$	Description
Textual	Emotional Word Count	2	2	The number of positive and negative emotional words.
	Emoticon Count	3	3	The number of positive, neutral and negative emoticons.
	Pronoun Count	2	3	The number of first-person singular / plural pronouns, and **other personal pronouns**.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

Data Analysis: Isomerism

- One feature may follow distinctive **integral** distributions in different domains.
- unrelated to specific user groups (de-pressed / 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

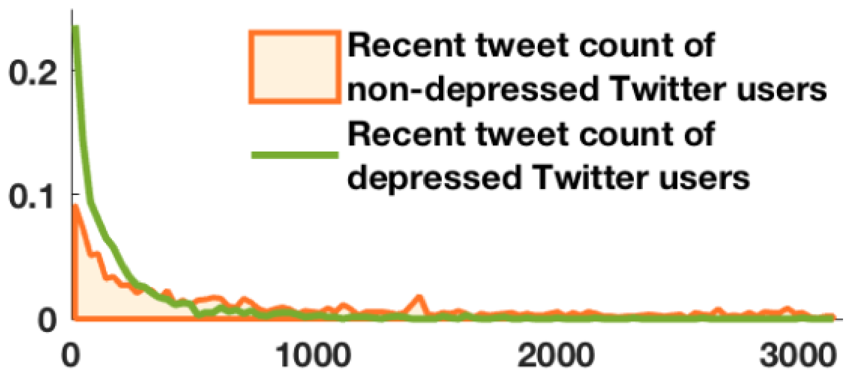


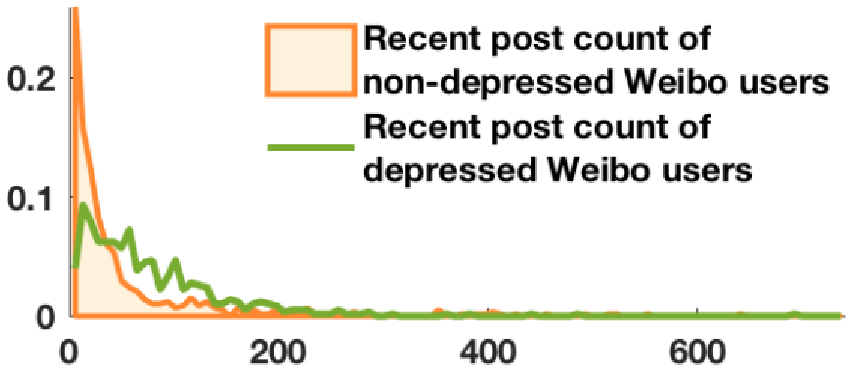




Data Analysis: Divergency

- 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.

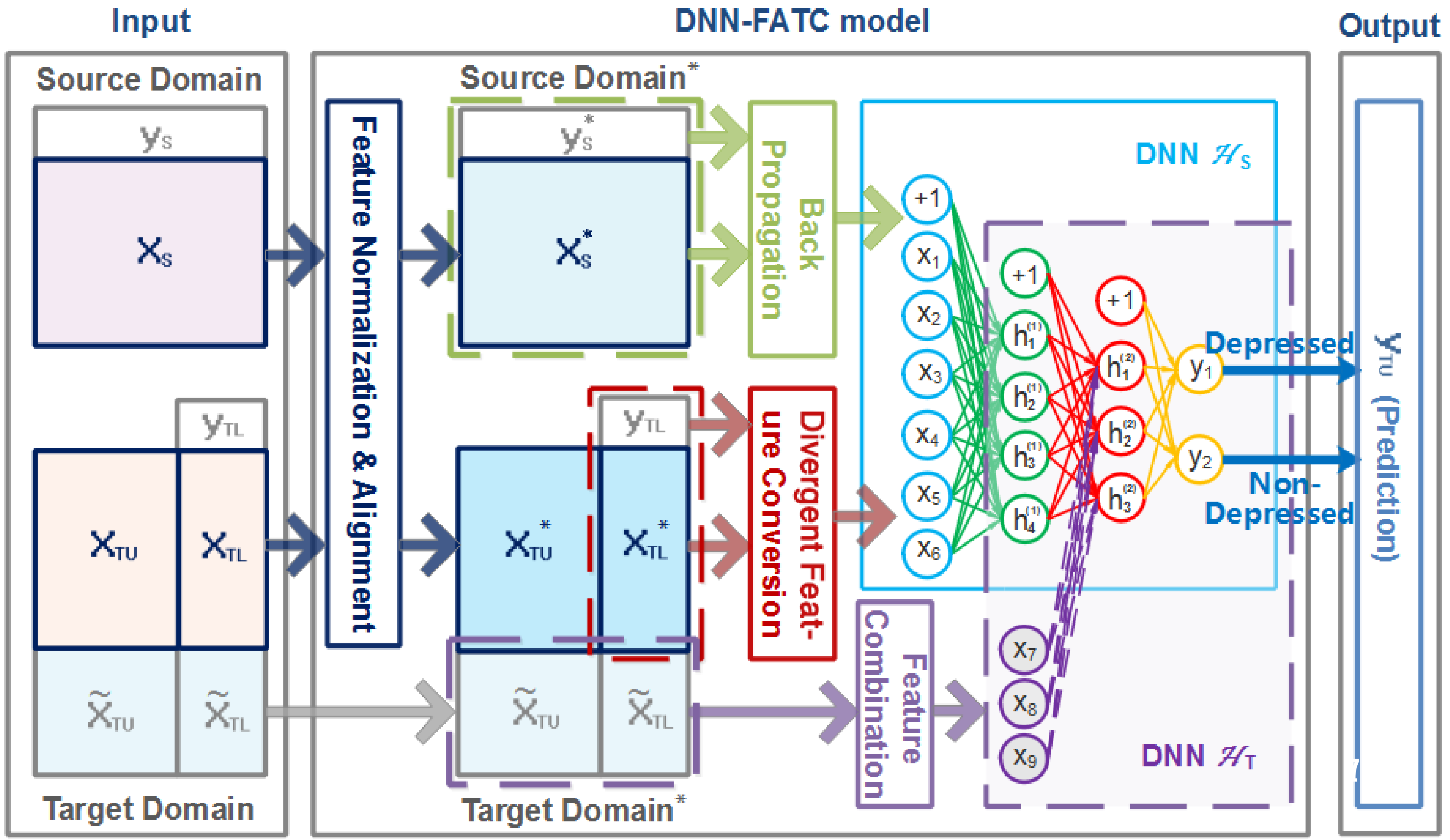




Methodology

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

- Based mainly on $\mathcal{D}_S$
- **Shared features:**
  - Against isomerism
  - Against divergency
- $\mathcal{D}_T$**-exclusive features :**
  - Integrated in the end



Feature Normalization & Alignment (FNA)

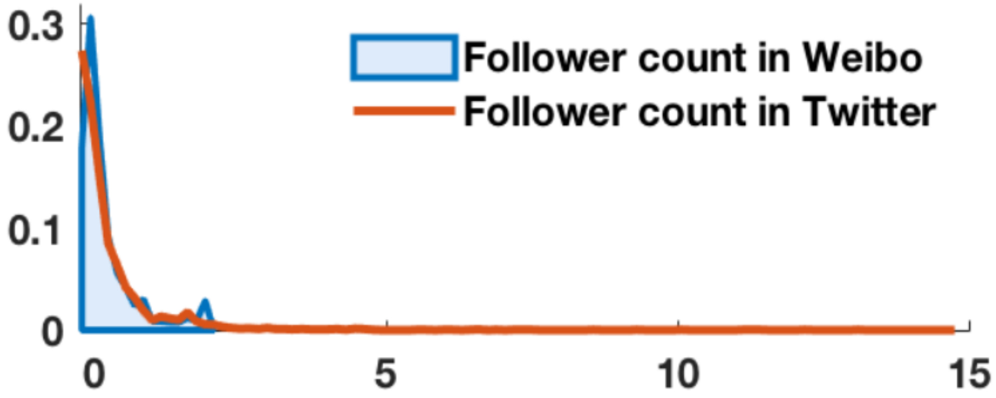
- Targeted on isomerism
- A linear transformation to fill the distributional gap by minimizing the Bhattacharyya distance.

$$\mathbf{x}_S^* = a_S \mathbf{x}_S + b_S, \quad \mathbf{x}_T^* = a_T \mathbf{x}_T + b_T,$$

$$s.t. \quad a_S, a_T, b_S, b_T = \underset{a_S, a_T, b_S, b_T}{\arg\min} \; -\ln \sum_{i=1}^{K} \sqrt{p_{Si}^* p_{Ti}^*}$$

- Two more constraints:

$$a_T \, Q(\mathbf{x}_T, 50) + b_T = 0,$$
$$a_S \, [Q(\mathbf{x}_S, q_2) - Q(\mathbf{x}_S, q_1)] = l$$



$K = 100, q_1 = 25, q_2 = 25$ and $l = 0.5$

- 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

$$\mathbf{x}_{T_i}^{**} = \alpha_i \mathbf{x}_{T_i}^{*} + \beta_i, \quad \mathbf{X}_T^{**} = \boldsymbol{\alpha}\mathbf{X}_T^{*} + \boldsymbol{\beta}\boldsymbol{I}$$

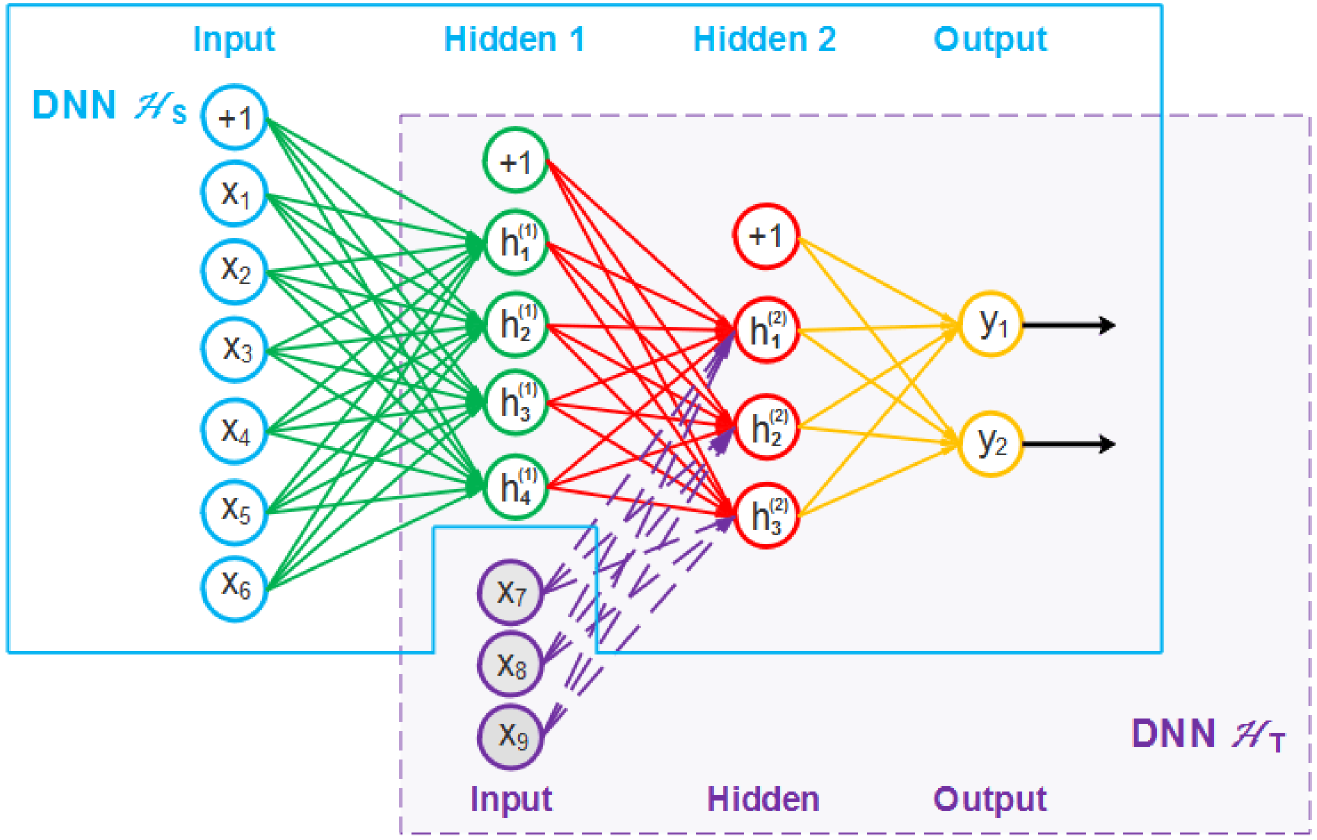
- Identify singular features: better performance is achieved with $\alpha_i < 0$

$$\boldsymbol{\alpha}^*, \boldsymbol{\beta}^* = \arg\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \mathscr{F}(\mathcal{H}_S, \{\boldsymbol{\alpha}\mathbf{X}_{TL}^{*} + \boldsymbol{\beta}\boldsymbol{I}, \mathbf{y}_{TL}\})$$

- 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 α_i, β_i orderly; record α^*, β^* for the best performance in W trials.
- Extra constraint: $\alpha_i \in \{-1,1\}$, $\beta_i = 0$, centrosymmetric transformation
- σ: 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

$\mathcal{D}_S$: 2,788 samples (Twitter Dataset)

$\mathcal{D}_T$: 1,160 samples (Weibo Dataset)

$\mathcal{D}_{TL}$: 280 samples (≈10% the size of $\mathcal{D}_S$)

$\mathcal{D}_{TU}$: 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.

	DL_s	DL	DT	BP	DFC	DFC+FC
MN	$60.5_{\pm 7.9}$	$64.2_{\pm 5.2}$	$34.3_{\pm 12.1}$	$61.2_{\pm 6.9}$	$66.6_{\pm 1.6}$	$67.4_{\pm 4.5}$
ZN	$68.2_{\pm 4.8}$	$70.1_{\pm 2.2}$	$58.6_{\pm 2.2}$	$73.0_{\pm 2.3}$	$72.3_{\pm 2.2}$	$73.9_{\pm 2.1}$
FNA	$72.0_{\pm 3.2}$	$73.3_{\pm 2.7}$	$68.0_{\pm 1.3}$	$75.9_{\pm 1.8}$	$77.6_{\pm 1.1}$	$\mathbf{78.5}_{\pm 1.2}$

Table 4: F1-measure of heterogeneous transfer methods.

FNA+DL	ARC-t	MMDT	HFA	DNN-FATC
$73.3_{\pm 2.7}$	$73.7_{\pm 1.1}$	$73.9_{\pm 1.8}$	$75.1_{\pm 0.8}$	$\mathbf{78.5}_{\pm 1.2}$

FNA vs **MN / ZN:** effectiveness in reducing isomerism

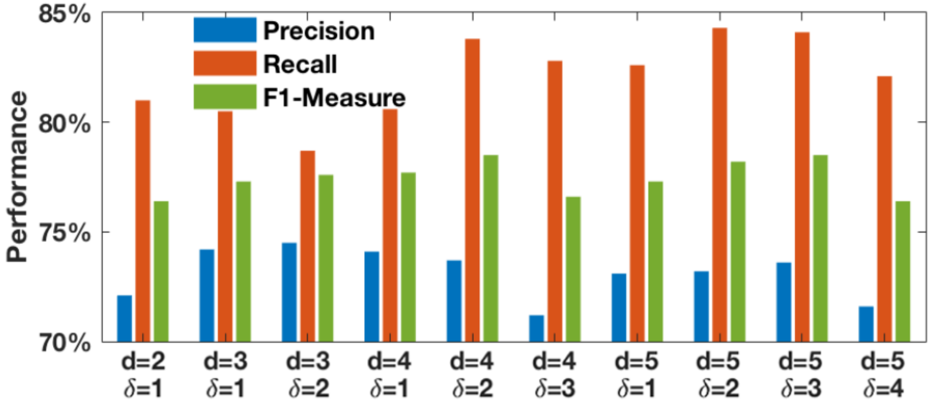
FSC vs **BP:** effectiveness in handling divergency

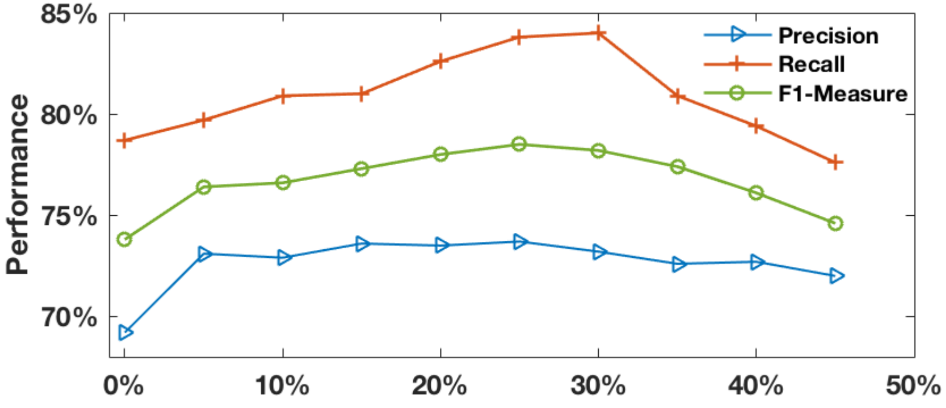
DL vs **DL$_S$**, **DFC+FC** vs **DFC:** effectiveness of $\mathcal{D}_T$-exclusive features, utilization method

$\mathcal{D}_S$ is useful to enhance detection in $\mathcal{D}_T$ and DNN-FATC best fits the assignment.

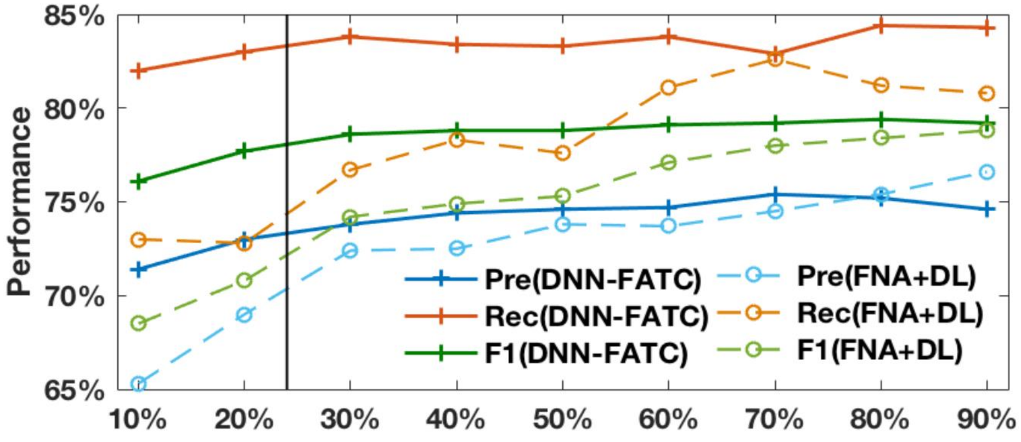
Experimental Results: Further Analysis

- 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

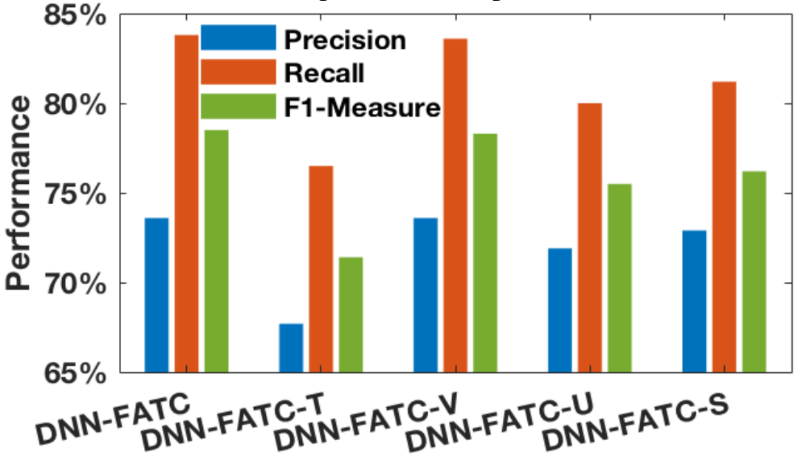




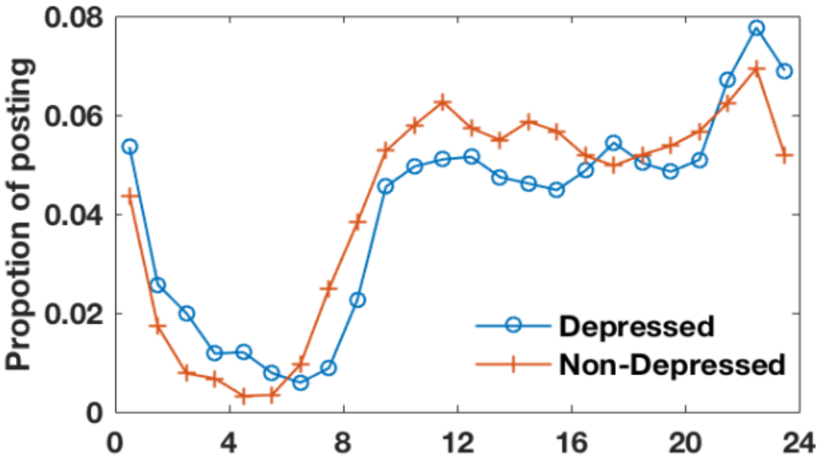
- **Data Scalability Analysis:**



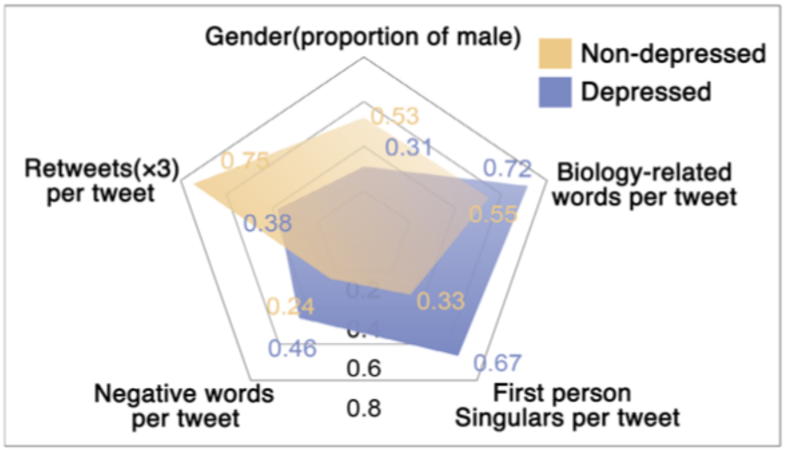
- **Feature Group Analysis:**



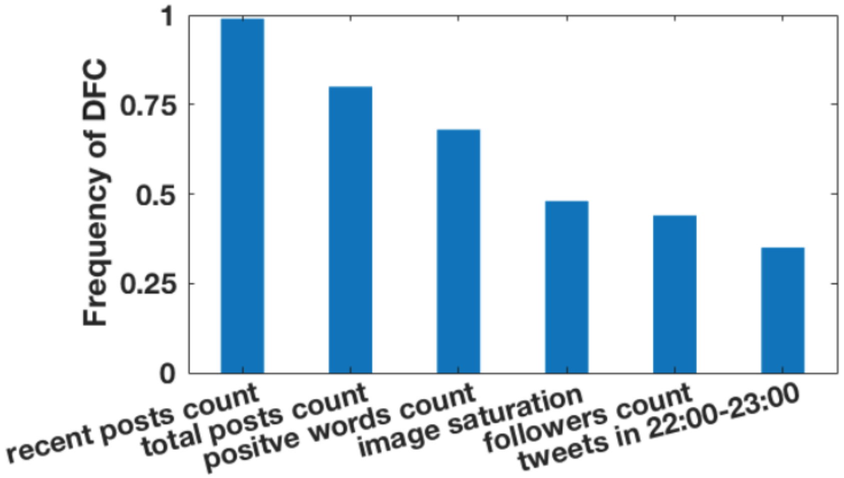
Case Study: Depressive Behavior Discovery



(e) Temporal comparison



(f) Feature comparison



(g) Divergent features

- **Depressive Behavior:**

  Tweet time

  Gender

  Linguistic pattern

  Retweet count

- **Divergent Features:**

  Tweet count

  Image saturation

  follower count

  positive word count

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!