Hierarchical Attention Network for Explainable Depression Detection on Twitter Aided by Metaphor Concept Mappings

Sooji Han, Rui Mao, and Erik Cambria.

COLING2022

정 시 열

목차

1. Introduction
2. Methodology
3. Experiments
4. Results
5. Conclusion

Introduction

- 최신 Depression Detection연구들은 성능 향상에 초점을 맞추어 SOTA 모델들을 사용하기에 설명이 어렵다는 단점 존재

- Depression 환자 특징
  - 본인의 감정과 정신 상태를 전문의에게 표현하기 앞서 Social Media에 표현하는 경향이 있음.
  - 본인의 감정과 경험을 ***은유**적으로 표현함.

  *__Metaphor__ is not only a linguistic phenomenon, but also a reflection of cognitive **mappings of source and target concept**

  - 환자의 은유 표현은 심리치료에서 중요함.

- "You are ***wasting*** my time" => "**Time** is **Money**"
- "This is the ***core*** of the matter" => "**importance** is **interiority**"

➢MCM을 활용한 설명 가능한 모델 **HAN**을 소개함.

*MCM = Metaphor Concept Mappings (" A is B ") / A: target, B: Source

Methodology

MCM 만들기

- MI 과정 (Metaphor Identification)

$$r_\epsilon, \rho_\epsilon = MI(x_\epsilon),$$

Input:	I	have	already	passed	the	two	written	exams	.
SMI:	lit	lit	lit	met	lit	lit	lit	lit	lit
PoS:	PRON	AUX	ADV	VERB	DET	NUM	VERB	NOUN	PUNCT

Pron 대명사, aux 조동사, adv 형용사 verb 동사, det: 한정사, num: 수사 noun 명사

Tweet 1개

$x_\epsilon = \{\tau_{\epsilon,1}, \tau_{\epsilon,2}, \ldots, \tau_{\epsilon,g}\}$

↓

Metaphor label

$r_\epsilon = \{r_{\epsilon,1}, r_{\epsilon,2}, \ldots, r_{\epsilon,g}\}$

$r_{\epsilon,j} = \{metaphor, literal\}$

Pos label

$\rho_\epsilon = \{\rho_{\epsilon,1}, \rho_{\epsilon,2}, \ldots, \rho_{\epsilon,g}\}$

Methodology

MCM 만들기

- MP 과정 (Metaphor paraphrasing)

은유 표현을 대신하는 단어 찾기

$$\omega^{L}_{\epsilon,j} = MP(\tau_{\epsilon,j}, \rho_{\epsilon,j}).$$

$$p(m_i|w_1, ..., w_{i-1}, w_{i+1}, ..., w_n)$$
$$= \mathrm{BERT}([\mathrm{CLS}], w_1, ..., [\mathrm{MASK}]_i, ..., w_n, [\mathrm{SEP}]),$$

해당 SENTENCE에서 은유표현을 마스킹 한 이후에 가장 적합한 단어를 찾도록 한다. 혹여 은유 표현이 여러 개이면 하나씩 찾는다.

진행 과정

1. 은유 표현을 표제화 시키기

$\tau_{\epsilon,j}$ ← 은유 표현 토큰
↓
$\tau^{L}_{\epsilon,j}$ ← 표제화 된 은유 표현 토큰

2. 해당 단어의 동의어, 상위어 중 적합한 단어 찾기

3. 적합한 단어를 표제화 시켜서 리턴

Methodology

MCM 만들기

- CG 과정 (Concept mapping Generation)

$$\mathrm{A}_{\epsilon,j} = CG(\tau^{L}_{\epsilon,j}),$$
$$\mathrm{B}_{\epsilon,j} = CG(\omega^{L}_{\epsilon,j}).$$

$$\mathrm{MCM}_{\epsilon,j} = \mathrm{B}_{\epsilon,j}\ \mathrm{IS}\ \mathrm{A}_{\epsilon,j}.$$

컨셉은 해당 단어를 잘 드러내는 WordNet의 Hypernym
=> 해당 단어를 포함하는 상위 개념

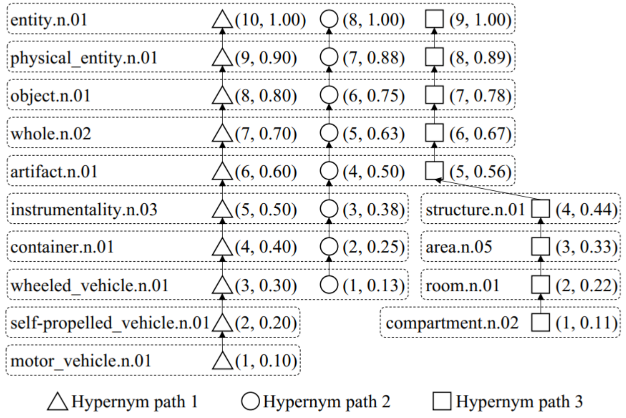
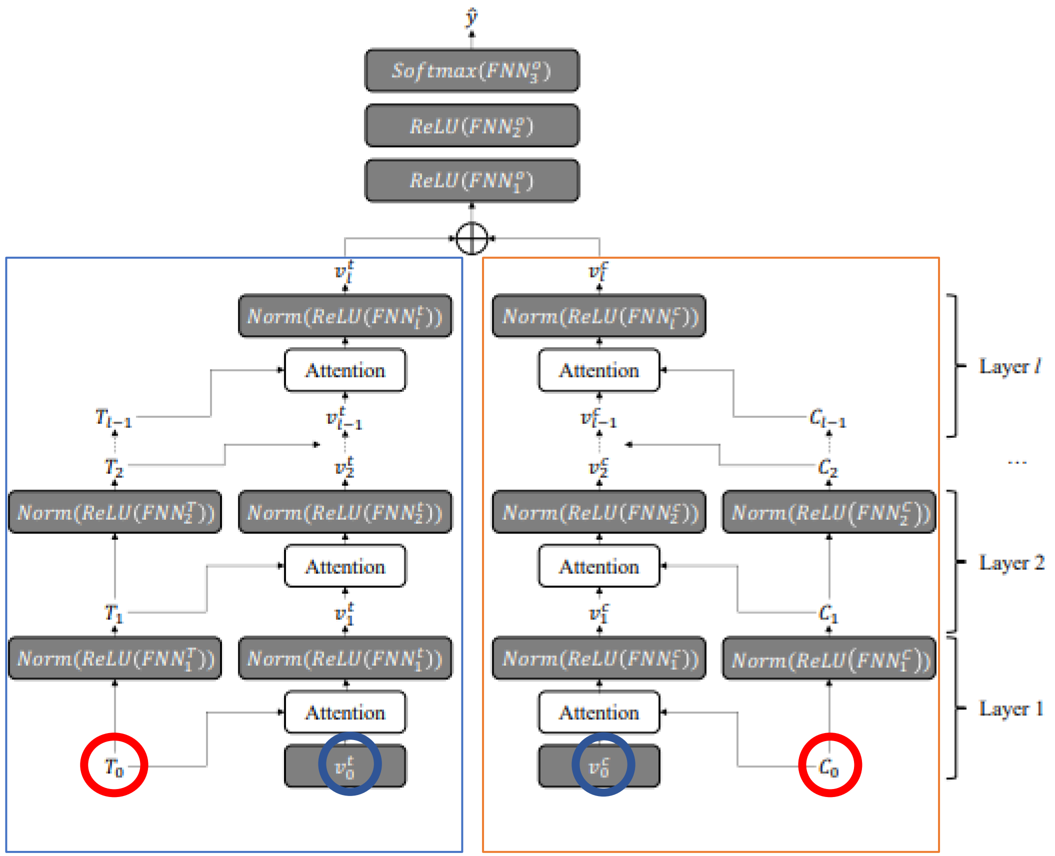


Figure 2: The rating scores of hypernyms of "car" in separate paths. The nodes in the same box denote the same synset in WordNet. The numbers in the parentheses besides a node denote the index of a node in a hypernym path (left), and the rating score in the path (right), respectively.

	Word pair	DR	Target	Source
M1	*steamroller* bill	vo	DOCUMENT	WAY
M2	son *drift*	sv	MALE_OFFSP.	VESSEL
M3	wine *breathe*	sv	ALCOHOL	ADULT
M4	story *lend*	sv	FICTION	ADULT
M5	government *bow*	sv	POLITY	ADULT
T1	*blind* alley	an	STREET	ADULT
T2	*raw* emotion	an	FEELING	ARTIFACT
T3	*weak* password	an	POSITIVE_ID.	ARTIFACT
T4	*rough* draft	an	WRITING	ARTIFACT
T5	*steep* discount	an	DECREASE	GEO._FORM.
G1	*bitter* night	an	TIME_PERIOD	SENSATION
G2	*sour* trade	an	TRANSACTION	SENSATION
G3	*clear* definition	an	EXPLANATION	MATERIAL
G4	*warm* gratitude	an	FEELING	MATERIAL
G5	*clean* datum	an	INFORMATION	MATERIAL

Table 7: Case study. M, T, and G are MOH, TSV, and GUT datasets. DR denotes dependency relationship, where sv, vo, and an are subjective-verb, verb-direct object, and adjective-noun dependencies, respectively.

Methodology



Tweet

MCM

User: u_k

Set of tweets: $X_k = \{x_{k,1}, \dots, x_{k,n}\}$

Set of MCM in tweets: $M_k = \{m_{k,1}, \dots, m_{k,n}\}$

$$u_k = [X_k, M_k]$$

$$T_0 = BERT(\mathbb{X}).$$

$$C_0 = BERT(\mathbb{M}).$$

$$v_i^t, T_i = HAN_i^t(v_{i-1}^t, T_{i-1}).$$

$$v_i^c, C_i = HAN_i^c(v_{i-1}^c, C_{i-1}).$$

Layer 내부 전개

$K_{i-1} \in \mathbb{R}^{o \times d}$

$(q_{i-1} \in \mathbb{R}^{1 \times d})$

$$w_i = Softmax\left(\frac{q_{i-1} \otimes K_{i-1}^{\top}}{\sqrt{d}}\right)$$

$$q_i = LN(ReLU(FNN_i^{query}(w_i \otimes K_{i-1}))).$$

$$K_i = LN(ReLU(FNN_i^{key}(K_{i-1}))).$$

Experiment

Dataset

- MDL (Twitter에서 Depression Detection을 위한 데이터 셋)

  직접적으로 우울증에 걸렸다고 tweet을 올린 유저: positive
  직접적으로 우울증에 걸렸다고 tweet을 올리지 않은 유저: negative

- IMDL

  MDL 데이터셋의 모든 트윗에서 직접적으로 우울증에 걸렸다고 적힌 트윗을 전부 삭제한 데이터셋

Dataset	Total # of tweets		Mean # of tweets per user	
	Positive	Negative	Positive	Negative
D1	156,013	153,328	72	75
D2	151,538	119,188	71	58
D3	142,057	118,611	66	58
D4	143,725	124,925	66	61
D5	148,039	134,700	69	66

Table 1: Statistics of the five randomly sampled datasets. The number of positive users and that of negative users are 2,159 and 2,049 for all the datasets.

layer: 2 / Batch size:64 /
Optimizer: Adam/
input length:200 / Dropout: 0.2

Results

Model	P	R	F1	Acc.
Gui et al. (2019)	0.900	0.901	0.900	0.900
Lin et al. (2020)	0.903	0.870	0.886	0.884
Zogan et al. (2021)	0.909	0.904	0.912	0.901
$\text{HAN}_{\text{ours}}\text{-Avg}_{\text{D1-D5}}$	0.975	0.969	**0.972**	**0.971**
D1	0.981	0.965	0.973	0.973
D2	0.988	0.956	0.972	0.971
D3	0.972	0.972	0.972	0.971
D4	0.968	0.970	0.969	0.968
D5	0.964	0.981	0.972	0.971

Table 2: Depression detection results. Our model result is averaged over the five testing sets (D1-D5).

Model	F1 on MDL-validation						F1 on IMDL-validation					
	D1	D2	D3	D4	D5	Avg	D1	D2	D3	D4	D5	Avg
HAN	0.985	0.960	0.971	0.976	0.975	**0.973**	0.939	0.911	0.933	0.927	0.931	**0.928**
HAN-MCM	0.972	0.947	0.963	0.967	0.962	**0.962**	0.914	0.897	0.914	0.905	0.918	**0.909**
Δ	0.013	0.013	0.008	0.009	0.013	**0.011**	0.025	0.014	0.019	0.022	0.014	**0.019**

Table 3: Ablation study results on validation sets, measured by F1 score. Δ is defined by $F1_{\text{HAN}} - F1_{\text{HAN-MCM}}$.

# of HAN layers	1	2	4	8
MDL D1 validation	0.542	**0.985**	0.983	0.979

Table 4: F1 scores for different numbers of encoder layers.

	LSTM	BiLSTM	GRU	BiGRU	TF-first	HAN-TF	HAN
F1 on MDL D1 val. $\uparrow$	0.966	0.965	0.961	0.959	0.898	<u>0.976</u>	**0.985**
# of param. per layer $\downarrow$	4.72M	9.45M	<u>3.54M</u>	7.09M	3.55M	4.14M	**1.18M**

Table 5: Comparison results of different encoder layers. $\uparrow$ denotes that the higher the value is, the better the model is. $\downarrow$ denotes that the lower the value is, the better the model is.

Results

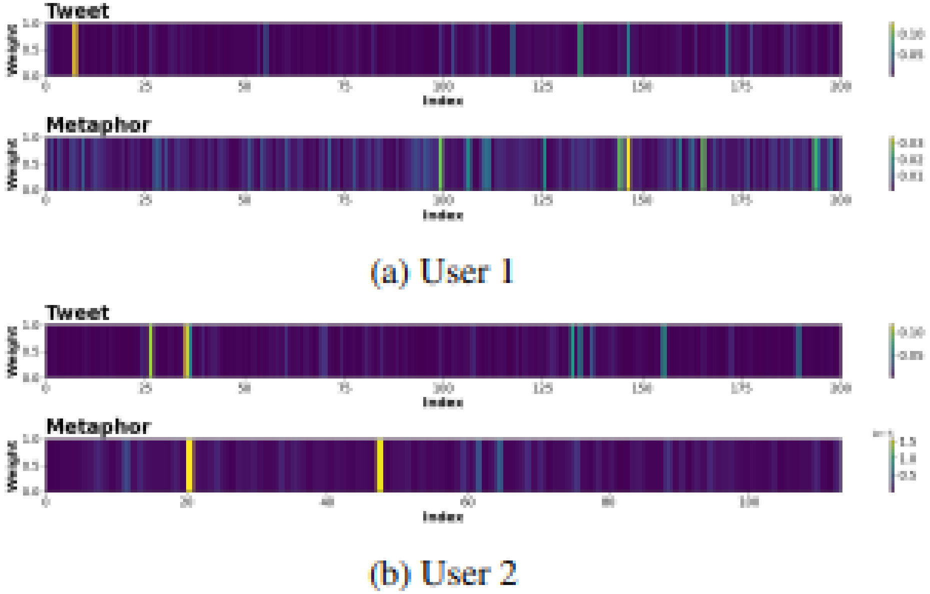


Figure 3: Visualization of attention weights for two depressed users. The lighter the color bar of an instance (tweet or MCM) is, the higher its attention weight is.

User	Tweet	Metaphor
1	1. I hate how I can't tell if I have allergies or I'm getting sick.	1. LEVEL IS IMPORTANCE
	2. get better, I love you	2. PERSON IS EXTREMITY
	3. I'm slightly allergic to cats but I still have them and I don't CARE IF I SNEEZE	3. SITUATION IS HAPPENING
	4. I'm having a bad night	4. ATHLETE IS AREA
	5. So I'm so nervous for my MAC interview tomorrow but I know I'll do great. Everything will be okay	5. MORPHEME IS EXTREMITY
2	1. Today is not a good day: Driver, teen shot to death after vehicle hits and kills -year-old	1. CONCERN IS STATE
	2. Autistic th Grader Assaulted by School Cop, Now He is a Convicted Felon	2. POSITION IS DISAPPEARANCE
	3. Thank you Father, GM FB! I gotta start taking My butt to bed at night, woke late again	3. LEVEL IS IMPORTANCE
	4. Cellphone Video Surfaces Showing Moments After Police Shot -Year-Old Boy in the Back	4. FEELING IS ILL_HEALTH
	5. Freddie Gray dies one week after Baltimore arrest	5. ARTIFACT IS SUPPORT

Table 6: The top 5 tweets and metaphors, selected based on attention weights, for two example users.

Conclusion

- Depression Detection을 위한 설명 가능한 모델을 만들었다.
- MCM을 활용하여 우울증 환자들이 그들의 감정 등을 표현하는 방법을 소개 가능했다.
- Depression Detection에 은유를 사용하는 것의 이점을 설명하였다.

End