Adaptation of IDPT system based on patient-authored text data using NLP: Supplement resources

5 juni 2020

Supplement Documents

1 Introduction

This document is created as the supplement resources for the paper entitled `Adaptation of IDPT system based on patient-authored text data using NLP: Supplement resources` by Mukhiya et. al. (`itsmeskm99@gmail.com`).

2 Prepossessing algorithm

3 PHQ-9 questionnaire symptoms

4 Seed term generation algorithm

Algorithm 1: Preprocessing test text sentences

```
input  : A patient authored test text
output : Processed test text

text ← encode(text, UTF-8);
text ← lower(text);
text ← strip(text);
symbols = [#,$, +, -, =, http, https)];
slangs =[(åon't", åill not "),("isn’t", "is not "),("can't", "can
  not "),(n't", not "),("i'm", "i am "),("'re", äre "),("'d",
  åould "),("'ll", åill ")]
foreach word w in text do
    if contains(word, symbol) then
        text ← remove_symbol(word, text);
    end if
    if contains(word, slangs) then
        text ← replace_slangs(word, slangs);
    end if
end foreach
return text
```

ID	PHQ-9 symptoms	Extracted Lexicons
S1	Little interest or pleasure in doing things	[interest, interested]
S2	Feeling down, depressed, or hopeless	[feeling, depressed, hopeless]
S3	Trouble falling or staying asleep, or sleeping too much	[sleep, asleep]
S4	Feeling tired or having little energy	[tired, energy]
S5	Poor appetite or overeating	[appetite, overeating]
S6	Feeling bad about yourself or that you are a failure or have let yourself or your family down	[failure, family]
S7	Trouble concentrating on things, such as reading the newspaper or watching television	[concentration, reading, watching]
S8	Moving or speaking so slowly that other people could have noticed. Or the opposite being so figety or restless that you have been moving around a lot more than usual	[moving, speaking, restless]
S9	Thoughts that you would be better off dead, or of hurting yourself	[dead, hurt, suicide]

Tabell 1: PHQ-9 symptoms (original) and the extracted lexicons (created)

Algorithm 2: Algorithm to generate lexicons

```
input  : $extracted\_lexicons$ from PHQ-9
output : Domain specific contextual-aware lexicons

seed_terms ← [];
foreach symptom $s \in extracted\_lexicons$ do
    terms ← [];
    foreach word $w \in s$ do
        synonyms ← wordnet.synonyms(s);
        foreach synonym $w \in synonyms$ do
            terms ← wordnet.hyperonym(s);
            terms ← wordnet.hyponym(s);
            terms ← wordnet.antonyms(s);
        end foreach
    end foreach
    sWord ← [];
    if $Model{=}{=}Depression2Vec$ then
        foreach term $t$ in $\in terms$ do
            sWord ← $word2vec_{dep}(t, nWord = 5)$
        end foreach
    end if
    if $Model{=}{=}Glove2vec$ then
        foreach term $t$ in $\in terms$ do
            sWord ← $glove2vec(t, word_{sim} > 80\%)$
        end foreach
    end if
    if $Model{=}{=}Wordnet$ then
        foreach term $t$ in $\in terms$ do
            sWord ← $t$
        end foreach
    end if
    terms ← sWord;
    seed_terms ← terms;
end foreach
return seed_terms
```

Tabell 2: Dataset details

Type	Statistics
Corpus size (Number of posts)	15044
Number of sentences	133524
Average sentences per post	8.87
Number of words	3502245
Average words per post	232
Training set size (Number of posts)	14944
Testing set size (Number of posts)	100
Online availability	On request

Tabell 3: Model embedding details

Model	Embedding Corpus	Embedding Size	Source
Depression2vec	15043 (training set)	300	Link
Universal sentence encoder	Pre_trained	512	Link
Glove	Pre_trained	300	Link
Wordnet	-	-	Link