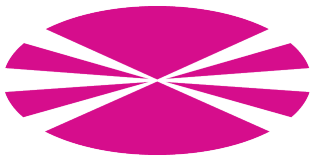
Ranking de usuarias de Reddit aplicando Modelos de Relevancia para trastornos depresivos

Eliseo Bao Souto
Dirección: Álvaro Barreiro García e Miguel Anxo Pérez Vila

Facultade de Informática
Mención en Computación
Grao en Enxeñaría Informática

5 de xullo do 2022



UNIVERSIDADE DA CORUÑA

Índice

Índice

Motivación e obxectivos

Que ocorre?

- 3.8 % da poboación mundial padece depresión (280 millóns).
- Emerxencias bélicas e médicas aumentan a prevalencia.
- Tendencia alcista no grupo de adolescentes e adultos novos.

Que sabemos?

- Detecciónss adecuadas e temperás reducen os efectos negativos.
- Forte relación entre o uso da linguaxe e o estado da persoa.

Que podemos explotar?

- Actividade en redes sociais (Reddit[1]).
- Modelos de Linguaxe baseados na Relevancia.

[1] `https://www.reddit.com/`

Motivación e obxectivos

Cales son os obxectivos?

- Adecuado estudo do problema e posibilidades.
- Planificación axustada.
- Cómputo dos *RMs* e obtención dos vocabularios de depresión.
- Comparativa dos vocabularios con *lexicons* de referencia.
- Optimización do *ranking* e comparativa cos *baselines*.
- Análise global e busca de posibles melloras e/ou ampliacións.

Fundamentos e conceptos básicos

Área de estudo

- Recuperación da Información
  - Satisfacer as necesidades de información das persoas usuarias.

Tarefa clásica

- Recuperación *ad hoc*.
  - Atopar documentos relevantes para necesidade de información (*query*).
    - Comparar representación de documentos e *queries*.
  - Obter un *ranking* cos mesmos.

Modelos de *retrieval*

- *Booleano*, Espazo Vectorial, etc.
- *Language Models*.
  - Probabilísticos e con base estatística sólida.
  - *Relevance Models*: introducen explicitamente o concepto de relevancia.

Aportación novidosa

En Recuperación da Información, con *LMs*, para *ranking sinxelo* de documentos:

- *Query-Likelyhood*[2]
  - *Usado por RMs para relevance feedback:*

$$P(w|R) = \sum_{D \epsilon C} P(w|D) \cdot \underbrace{\prod_{i=1}^{n} P(q_i|D)}_{Query\text{-}Likelyhood} \tag{1}$$

Modelizamos problema con métodos utilizados para outras tarefas :

- Non existe *query* $\rightarrow$ Simular con formulario clínico.
  - ***Query-Likelyhood*** $\rightarrow$ *Score* **BDI**[3]
    - O modelo dará máis *peso* á linguaxe dos BDIs altos.

$$P(w|R) = \sum_{D \epsilon C} P(w|D)^4 \cdot BDI(D) \tag{2}$$

[2] Estimación da probabilidade de que a *query* sexa unha mostra do modelo de linguaxe do documento.

[3] Test psicométrico composto por 21 cuestións que permite medir a severidade da depresión nunha persoa.

[4] Con ***Smoothing*** de probabilidades: Dirichlet, Jelinek-Mercer.

This questionnaire consists of 21 groups of statements. Please read each group of statements carefully, and then pick out the one statement in each group that best describes the way you feel. If several statements in the group seem to apply equally well, choose the highest number for that group.

1. Sadness
    0. I do not feel sad.
    1. I feel sad much of the time.
    2. I am sad all the time.
    3. I am so sad or unhappy that I can't stand it.

2. Pessimism
    0. I am not discouraged about my future.
    1. I feel more discouraged about my future than I used to be.
    2. I do not expect things to work out for me.
    3. I feel my future is hopeless and will only get worse.

...

20. Tiredness or Fatigue
    0. I am no more tired or fatigued than usual.
    1. I get more tired or fatigued more easily than usual.
    2. I am too tired or fatigued to do a lot of the things I used to do.
    3. I am too tired or fatigued to do most of the things I used to do.

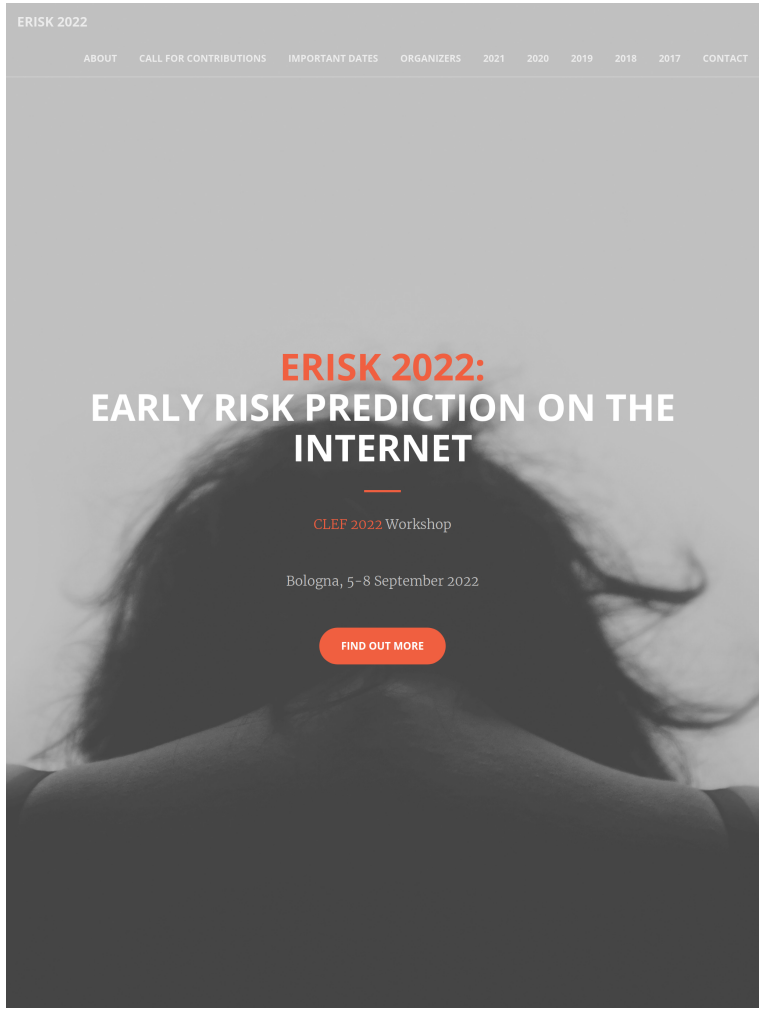
21. Loss of Interest in Sex
    0. I have not noticed any recent change in my interest in sex.
    1. I am less interested in sex than I used to be.
    2. I am much less interested in sex now.
    3. I have lost interest in sex completely at all.

Índice

Colecciónes

eRisk[5], enmarcado na Iniciativa CLEF[6].

- Participa IRLab (UDC).
  - Tamén: USC, USI (Lugano).
- Técnicas e algoritmos para a detección temperá de riscos.
- Metodoloxías e métricas.



[5] https://erisk.irlab.org/

[6] https://www.clef-initiative.eu/

Colecciónes

- Proposta anual de tarefas.
  - Trastornos depresivos, anorexia, ludopatía, etc.
    - Liberación de colecciónes asociadas.

Tarefas	**2017**	**2018**	**2019**	**2020**	**2021**
Detección de depresión	X	X			
Anorexia		X	X		
Comportamentos autolesivos			X	X	X
Severidade dos signos de depresión			X	X	X
Ludopatía					X

Cadro: Distribución da proposta de tarefas ao longo das edicións de eRisk.

Coleccióńs

Tipos de tarefas de depresión en CLEF eRisk

Detección

- Datos binarios.
  - Suxeito deprimido?
- eRisk 2017 e 2018.

Estimación da severidade

- Datos discretos.
  - Puntuación BDI suxeito.
- eRisk 2019, 2020 e 2021.

- Igual formato a excepción do valor de verdade.
- **Desbalanceadas**: moitos máis suxeitos de control que con depresión.

Lexicons

Pedesis

- Léxico recompilado mediante métodos manuais e automáticos.

Choudhury

- A través de Twitter[7], etiquetación de palabras frecuentes entre suxeitos depresivos.

	Pedesis	**Choudhury**
Adexctivos únicos non ambiguos orixinais	153	7
Axd. únicos non amb. expandidos Dist	312	13
Adx. únicos non amb. expandidos WN	549	16
Termos únicos totais orixinais	636	106

Cadro: Detalle das categorizacións usadas nos *lexicons*.

David E. Losada e Paulo Gamalho

- Análise dos dous *lexicons* anteriores e proposta de técnicas para poderen ser aumentados.

[7] `https://twitter.com/`

Índice

Análise de vocabularios de depresión - Preámbulo

Coleccións usadas para os grupos de RMs computados

- Grupo A – Depresión
  - CLEF eRisk 2019, 2020 e 2021.
- Grupo B – Depresión
  - CLEF eRisk 2019, 2020 e 2021 + 2017 e 2018 para *background* dos modelos.
- Grupo C – Control
  - Conxunto de datos propio construído no mesmo dominio (Reddit).
  - Proceso aleatorio de *crawling*.
    - Condicións mínimas aceptables.
    - Repetición en diferentes franxas horarias.

Análise de vocabularios de depresión - Individual

Obxectivo: Contar ocorrencias das familias de pronomes persoais no *top* 20 dos vocabularios de depresión obtidos.

Hipótese: A *I-family:* { *me, my, mine, myself* } é reveladora de signos de depresión.

Vocabulario depresión		*I*	*You*	*He/She*	*We*	*They*
Dirichlet (2500)	A	3	2	0	0	0
	B	3	2	0	1	0
	C	3	2	1	1	0

Resultado: Non determinante, pode existir marca lingüística no dominio.

Análise de vocabularios de depresión - Comparativo

Obxectivo: Estudar solapamento entre os *lexicons* e os vocabularios de depresión, vendo as coincidencias de termos dentro de diferentes *top*.

Vocabulario depresión		Pedesis		Choudhury	
		Top 636	*Top* 1000	*Top* 106	*Top* 1000
Dirichlet (2500)	A	18	37	5	37
	B	20	40	5	36
	C	18	34	4	34

Resultado: Solapamento inferior ao agardado, pero *RMs* valiosos.

Ranking de usuarias - Definición

Obxectivo: Formulación da *query* de depresión que maximice a calidade do *ranking*.

- Alternativas para a selección dos termos.
  - Pedesis e Choudhury ordenados cos vocabularios de depresión.
  - Vocabularios de depresión.
- Optimización *num* termos e suavización.
  - Métrica de referencia.

$$AP@100 = \frac{\sum_{r=1}^{100} P@r}{|R|} \tag{3}$$

- Definición de *baselines* de referencia.
  - *Queries* optimizadas con categorizacións dos *lexicons*.

Training

- Optimización de parámetros.
- eRisk 2017.

Test

- Validación de métricas.
- eRisk 2018.

Ranking de usuarias - Termos das *queries*

Query **baseline** gañador (106 termos).

`addictive, adhd, agree, amaze, answer, antidepressant, anxiety, appetite, attack, beautiful, bible, blur, boy, care, chemical, church, clinical (...)`

Query **Pedesis** ordenado con vocabularios de depresión (10 termos).

`feel, hurt, small, afraid, sad, deep, difficult, weight, numb, night`

Query **Choudhury** ordenado con vocabularios de depresión (20 termos).

`like, girl, want, love, help, friend, date, weight, life, talk, social, discuss, game, doctor, relationship, care, answer, adhd, home, man`

Query **vocabularios de depresión** (50 termos).

`i, you, my, have, so, like, me, just, your, do, what, get, can, about, on, would, think, all, know, we, peopl, becaus, he, time, http, more, when, go, out (...)`

Ranking de usuarias - Resultados

Resultado: Moi exitoso, posto de manifesto na *P*@5.

	nº Termos	*Smoothing*	*AP*@100	*P*@5	*P*@10
Baseline	106	Dirichlet (500)	0.21458	0.4	0.5
Pedesis ordenado	10	Dirichlet (1000)	0.24783	0.6	0.5
Choudhury ordenado	20	Dirichlet (2000)	0.27216	0.6	0.7
Vocabularios	**50**	**Dirichlet (500)**	**0.35163**	**0.8**	**0.7**

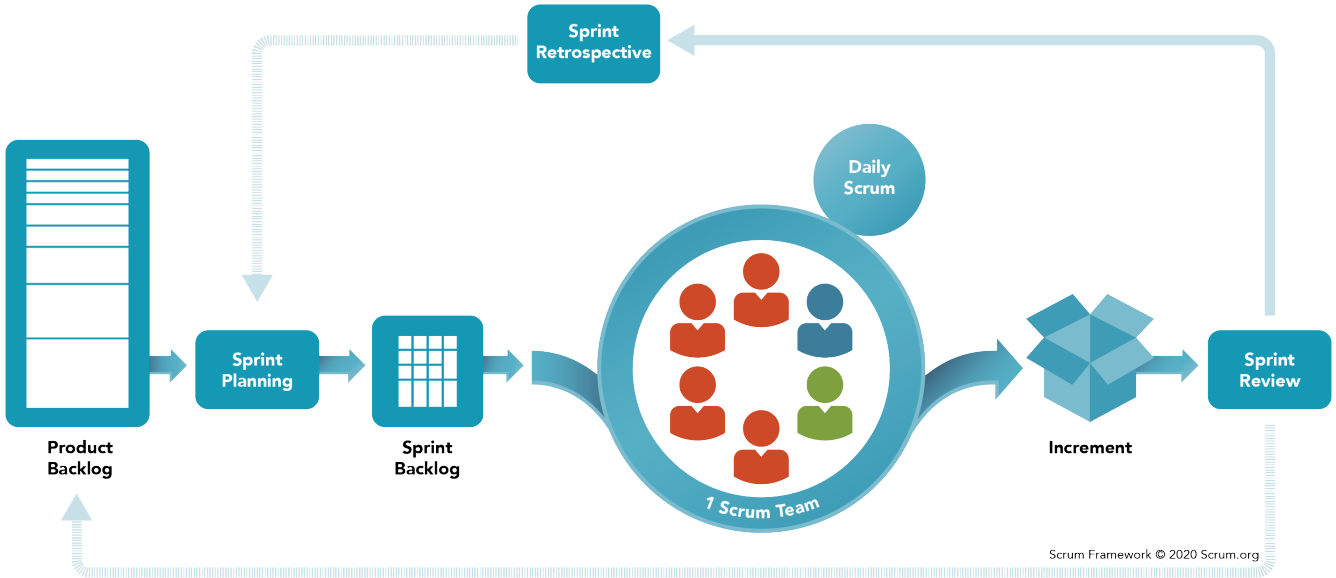
Cadro: Mellores resultados de *ranking* e *baseline*[8] de referencia.

[8] Mellor dos 9 *rankings* coas categorizacións dos *lexicons*.

Índice

Metodoloxía

- **Scrum** con adaptacións.



- Equipo: autor e directores.
- 4 semanas/*Sprint*, a razón de 60 Puntos de Historia.

Planificación

Fase inicial

- Estudo da propia metodoloxía.
- Análise e asimilación da problemática.
- Xestión de riscos.

56 Puntos/*Sprint* ao desenvolvemento.

4 Puntos/*Sprint* á análise.

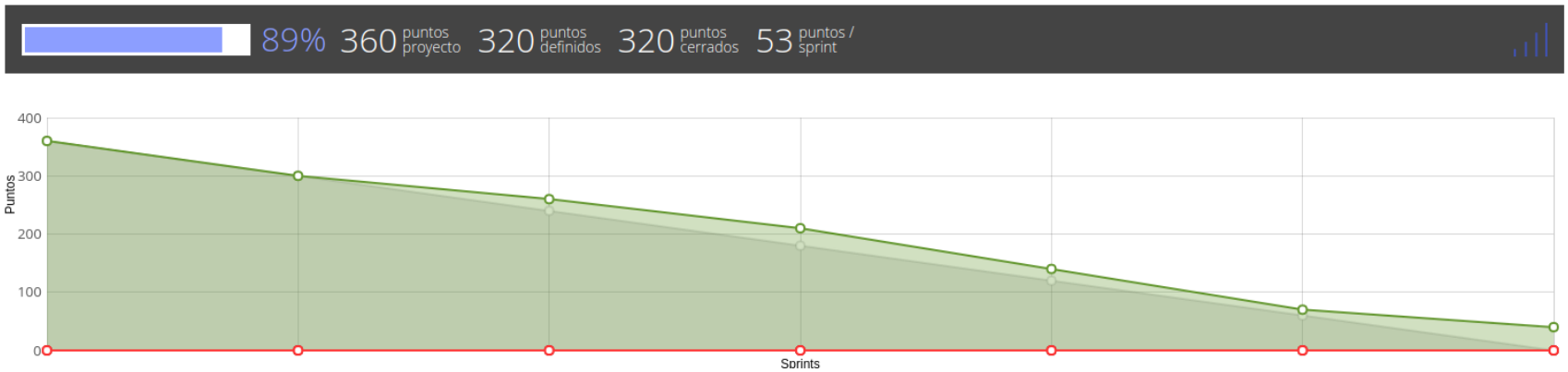


Figura: 320 horas[9] adicadas ao desenvolvemento e a análise.

Git para o control de versións.

[9] 1 Punto de Historia ≈ 1 hora de dedicación

Planificación - Custes

Recursos humanos

Perfil	Tempo (h/*Sprint*)	Dedicación (nº *Sprints*)	Custe (€/h)	Total (€)
Director	10 (ambos)	6	45	2700
Analista	4	6	25	600
Desenvolvedor	56	6	20	6200
			Total	**9500**

Recursos *software* e materiais

Recurso	Unidades	Custe (€/ud)	Vida (meses)	Uso (meses)	Total (€)
Licenzas					0
Ordenador persoal	1	800	48	6	100
				Total	**100**

Custe total aproximado do proxecto: 9600€

Índice

Conclusións

- Obtención exitosa dos **vocabularios de depresión**.
- Análise dos mesmos.
  - Individual.
  - Respecto a *lexicons* de referencia.
- *Ranking* de usuarias explotando vocabularios de depresión.
  - Satisfactorio: **Mellora dos resultados** con respecto ao *baseline*.
    - Valores respetables nas medicións $AP@100$, $P@5$ e $P@10$.

Traballo futuro

- Aplicación doutro tipo de *Language Models*.
  - *Maximum Entropy Divergence Minimization Model* (**MEDMM**).

- Uso de ***contextualized word-embeddings***.
  - Uso de modelos neuronais, por exemplo BERT.
  - Representación vectorial dos termos.
  - Reformulación da estimación dos *Relevance Models*.

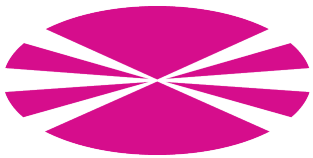
- Presentación do método para a **tarefa real** de CLEF eRisk.
  - Comparativa con laboratorios e grupos de prestixio.

Ranking de usuarias de Reddit aplicando Modelos de Relevancia para trastornos depresivos

Eliseo Bao Souto
Dirección: Álvaro Barreiro García e Miguel Anxo Pérez Vila

Facultade de Informática
Mención en Computación
Grao en Enxeñaría Informática

5 de xullo do 2022



UNIVERSIDADE DA CORUÑA