Contexto del Curso - Administrativas

Marcelo Errecalde[1,2]

[1]Universidad Nacional de San Luis, Argentina
[2]Universidad Nacional de la Patagonia Austral, Argentina

Curso: Minería de Textos
Facultad de Informática - Universidad Nacional de La Plata
23 al 27 de Septiembre de 2019

Administrativas

Profesor Responsable:

Marcelo Luis Errecalde

Contacto:

merreca@unsl.edu.ar

o bien

merrecalde@gmail.com

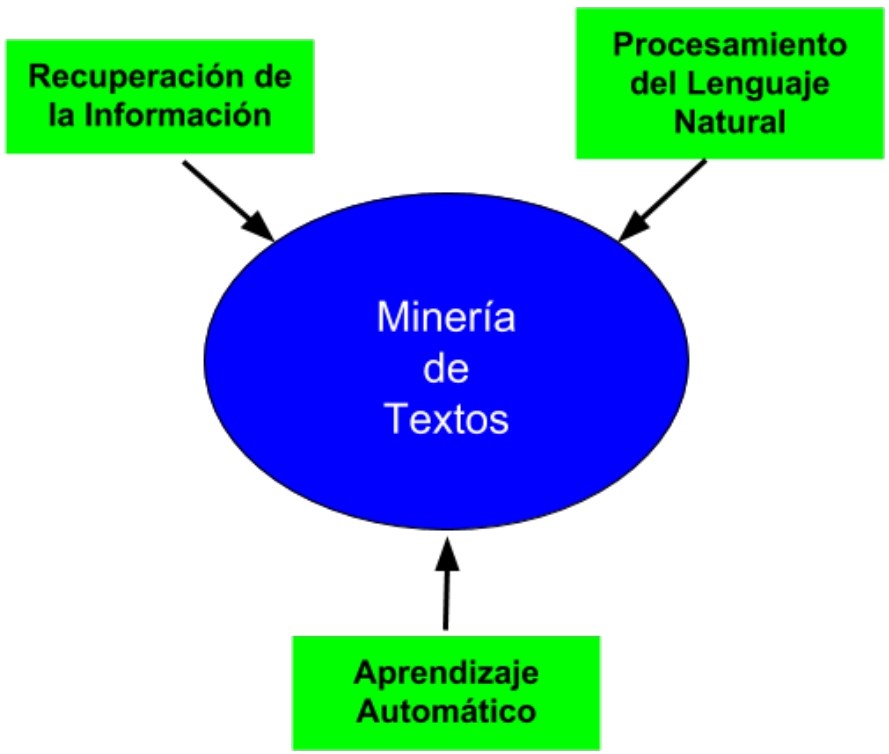
Material del curso:

Se enviará via e-mail o se hará disponible en links específicos a lo largo del desarrollo del curso

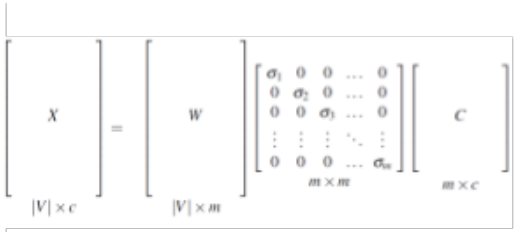
¿Qué se pretende con este curso?

- Que sea útil para ustedes .... (no fácil, por los distintos perfiles)
- Introducir los conceptos y herramientas básicas fundamentales vinculados a la minería de textos (MT)/análisis automático de textos.
- Analizar y entender las dificultades, posibilidades y principales aplicaciones de este tipo de técnicas.
- Analizar con una mirada crítica los trabajos realizados en el área y/o tener las herramientas para implementar soluciones a problemas específicos.

Dificultad de balancear las áreas

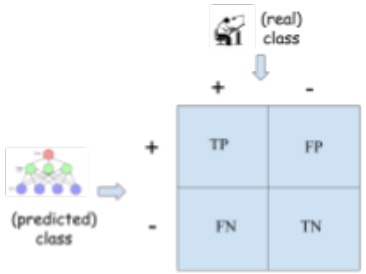


Dificultad de balancear las áreas

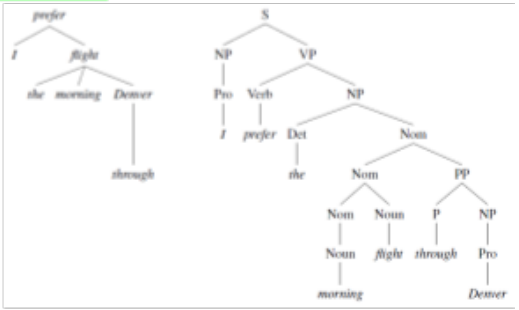


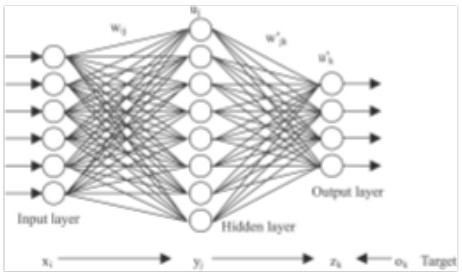
Recuperación de la Información

Procesamiento del Lenguaje Natural

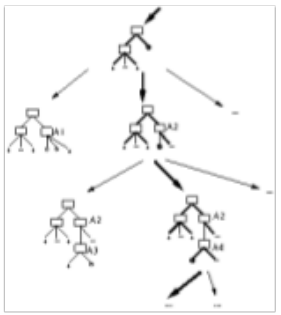


Minería de Textos





Aprendizaje Automático



Características del curso

- ¿Teoría?: la necesaria para entender los conceptos subyacentes (más intuitiva que definiciones)
- ¿Práctica?: haremos algunos ejercicios como parte de la evaluación del curso
- ¿Implementación/máquina?: se verán ejemplos en Python (y unos pocos en Weka). Notebooks disponibles para práctica más intensiva.

Organización del Curso

6 clases:

- Clase 1: Aspectos generales (conceptos, más bien teórica)
- Clase 2: Pre-procesamiento de textos
- Clase 3: Representación de documentos
- Clase 4: Reducción de la dimensionalidad
- Clase 5: Categorización supervisada y no supervisada de textos
- Clase 6: Aplicaciones y aspectos avanzados

Evaluación del Curso

Alternativas:

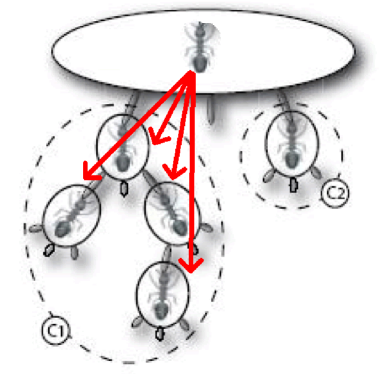
- Un cuestionario teórico/práctico básico, a ser enviado en un plazo de una a dos semanas.
- Un trabajo a consensuar dependiendo de las inquietudes e intereses del asistente.

¿En qué temas trabajamos/investigamos?

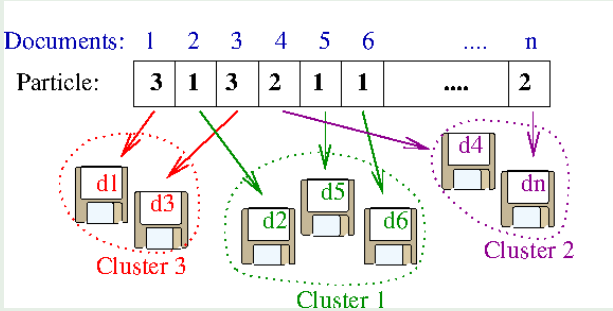
- Agrupamiento de textos cortos
- Calidad de información en la Web (esp. en Wikipedia)
- Perfil del autor en medios sociales (edad, género, rasgos de personalidad, etc)
- Detección temprana de riesgos

(Agrupamiento de textos cortos - técnicas bio-inspiradas

Técnicas basadas en hormigas



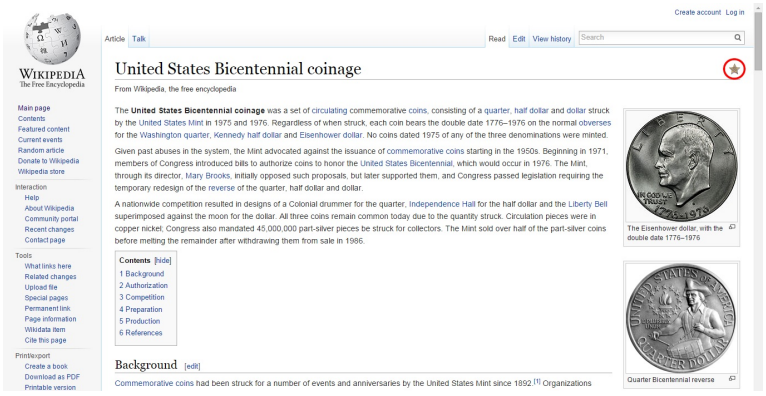
Técnicas basadas en PSO



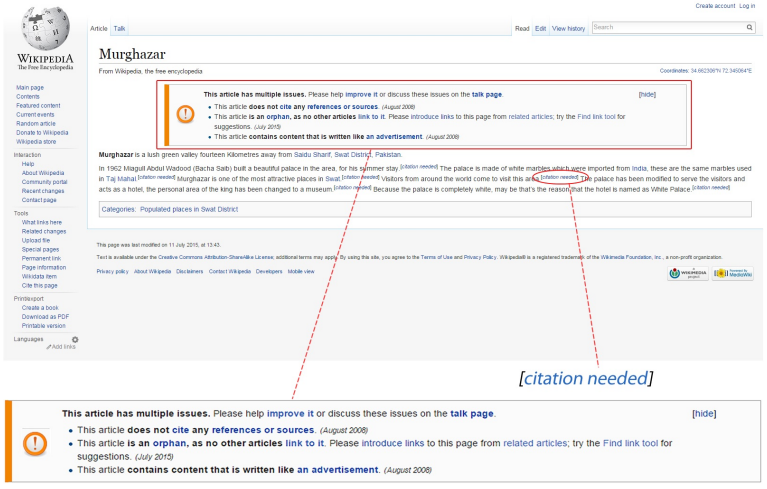
Publicaciones recientes

Calidad de Información en Wikipedia

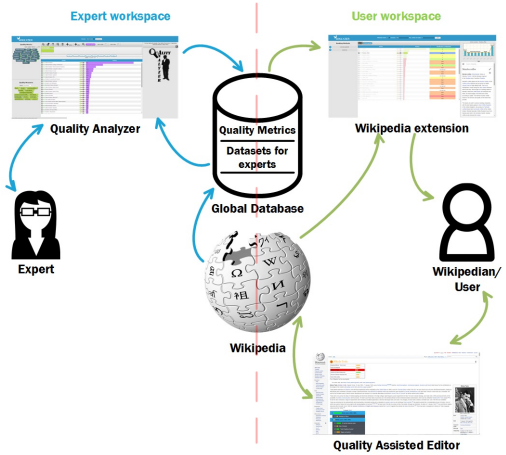
Artículos Destacados



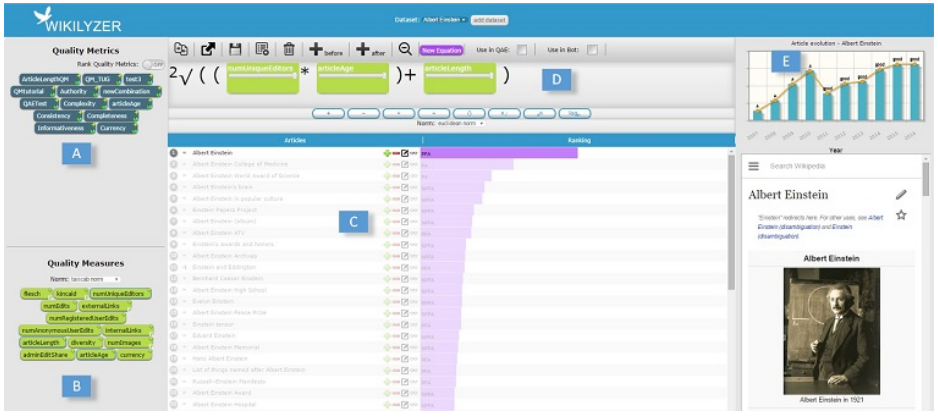
Detección de Fallas



Visualización



Métricas de Calidad



Perfil de autor

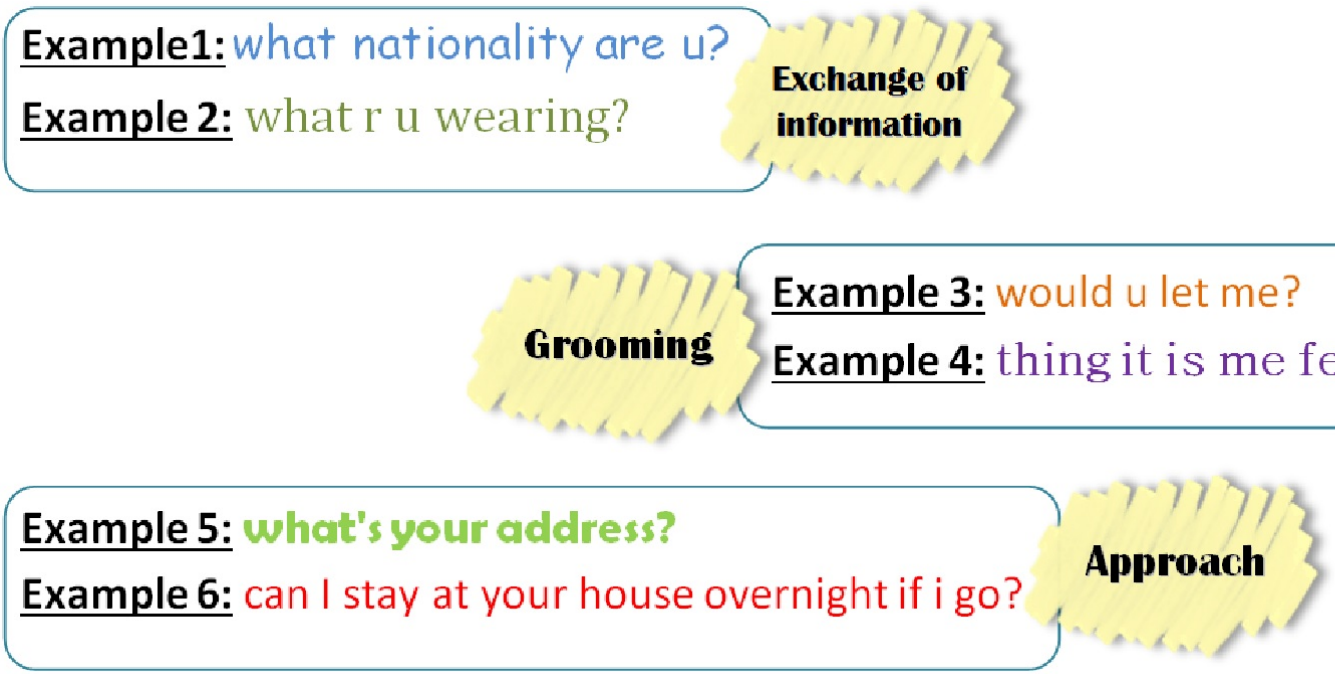
Tareas: determinar

1. edad
2. género
3. personalidad
4. orientación política
5. ...

Detección de pedófilos en la Web

- Datos de entrenamiento en
  www.perverted-justice.com
- Competencias recientes
  pan.webis.de/clef12/pan12-web

Ejemplo:



Detección anticipada de riesgos en contexto

Detección anticipada de riesgos (DAR)

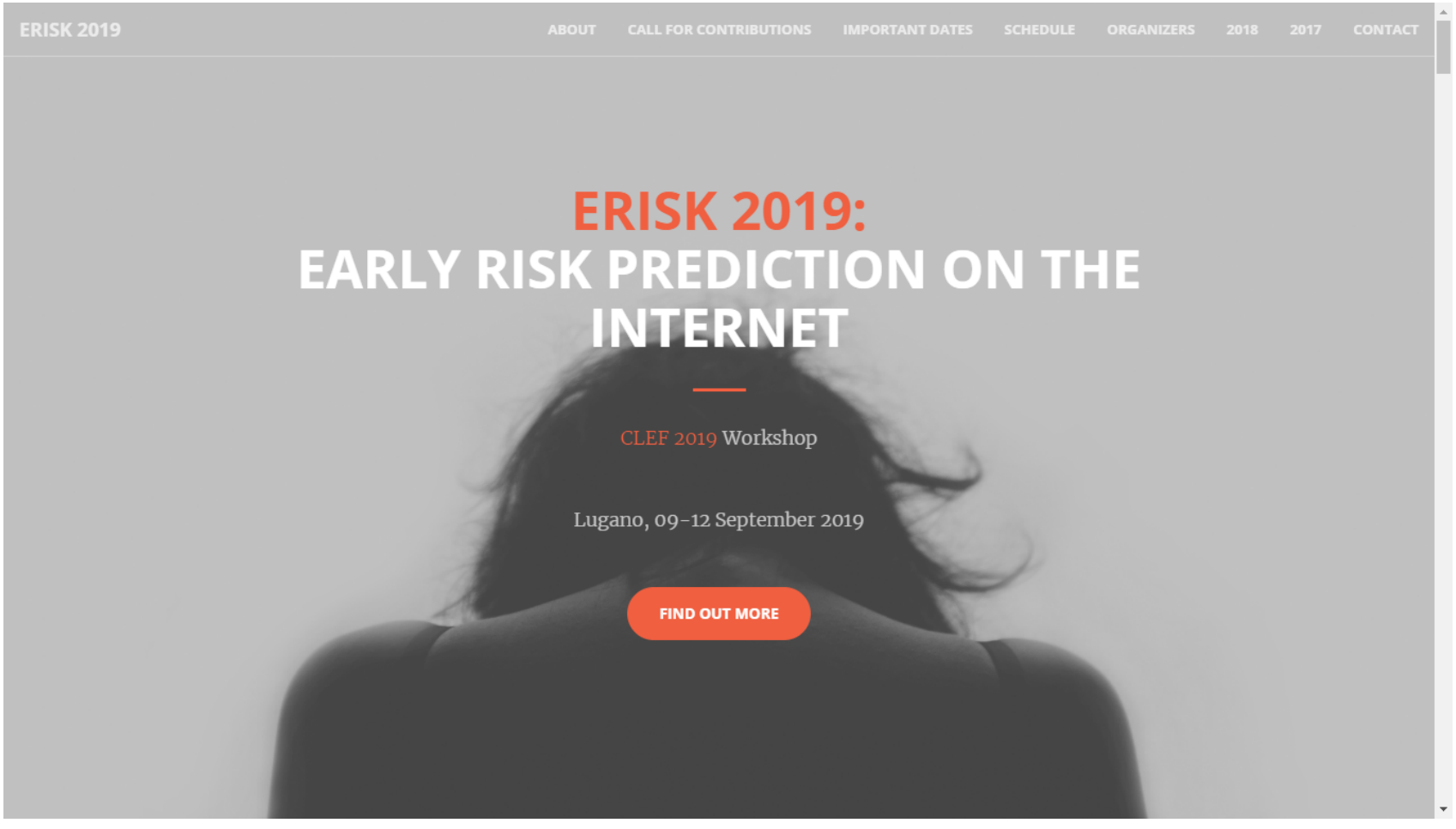
Tarea donde los datos son leídos secuencialmente como un flujo de datos y el desafío consiste en detectar casos de riesgo, tan pronto como sea posible.

Aplicaciones Potenciales

Detección temprana de:

- predadores sexuales
- personas con inclinaciones suicidas
- personas con signos de depresión
- comportamiento agresivo en las redes sociales
- actividades criminales y terroristas
- abandono en cursos académicos

Una tarea internacional dedicada al tema .....



Ediciones de eRisk

eRisk 2017 (https://early.irlab.org/2017/index.html)

- Task: Early Detection of Depression

eRisk 2018 (https://early.irlab.org/2018/index.html)

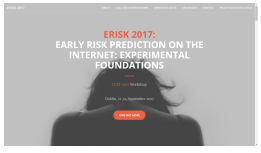
- Task 1: Early Detection of Signs of Depression
- Task 2: Early Detection of Signs of Anorexia

eRisk 2019 (https://early.irlab.org/)

- Task 1: Early Detection of Signs of Anorexia
- Task 2: Early Detection of Signs of Self-harm
- Task 3: Measuring the severity of the signs of depression

La tarea ERISK 2017

Primera edición de la tarea



Erisk 2017 - CLEF 2017 Workshop

Conference and Labs of the Evaluation Forum

2) Nuestra participación se realiza con un método diseñado específicamente para este tipo de tarea (TVT)

3) Se obtiene el mejor valor en la medida $ERDE_{50}$

Conjuntos de Datos

Conj. de Entrenamiento

486 usuarios

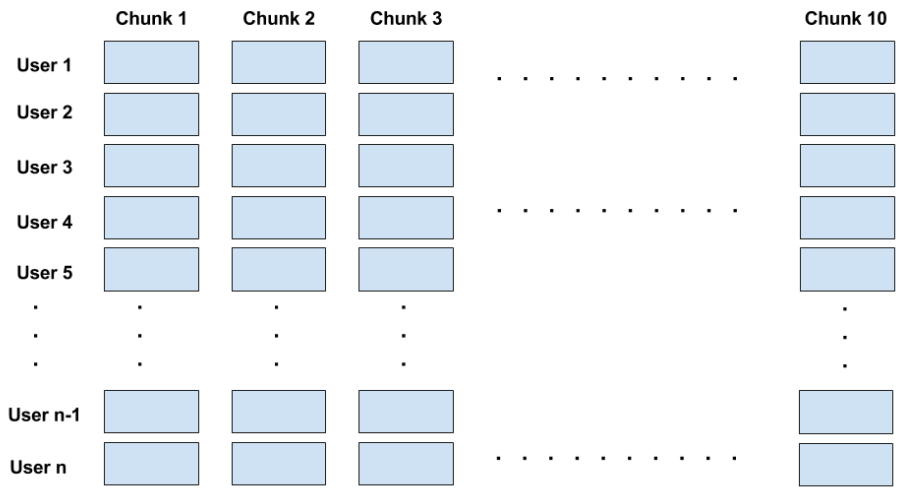
- 83 “deprimidos” (+)
- 403 “no deprimidos” (-)

Conjunto de Prueba

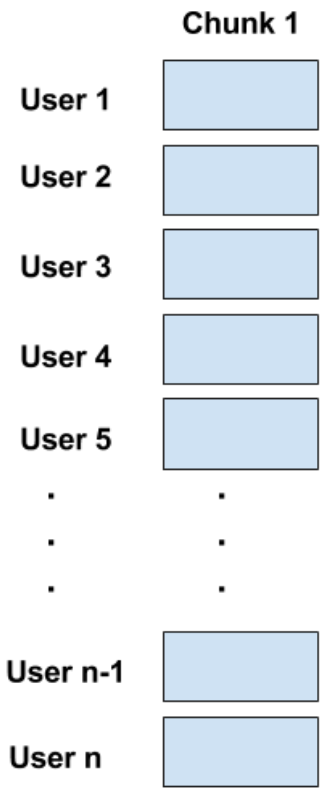
401 usuarios

- 52 “deprimidos” (+)
- 349 “no deprimidos” (-)

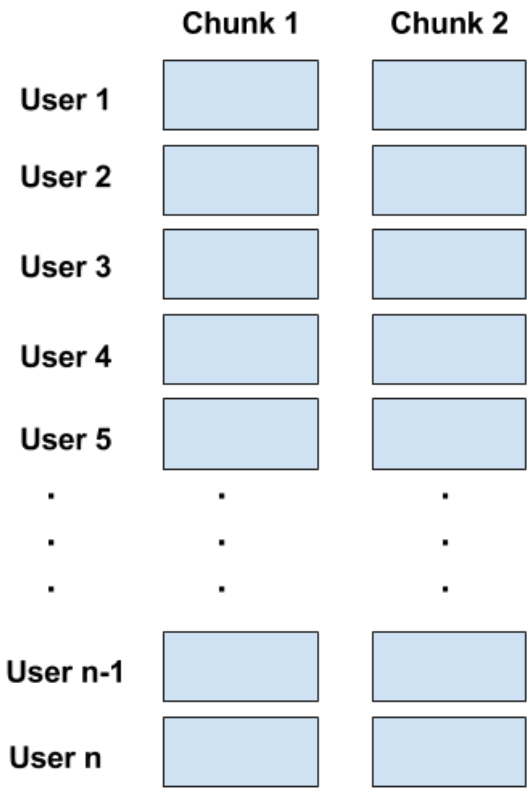
Divididos en 10 bloques (ordenados cronológicamente):



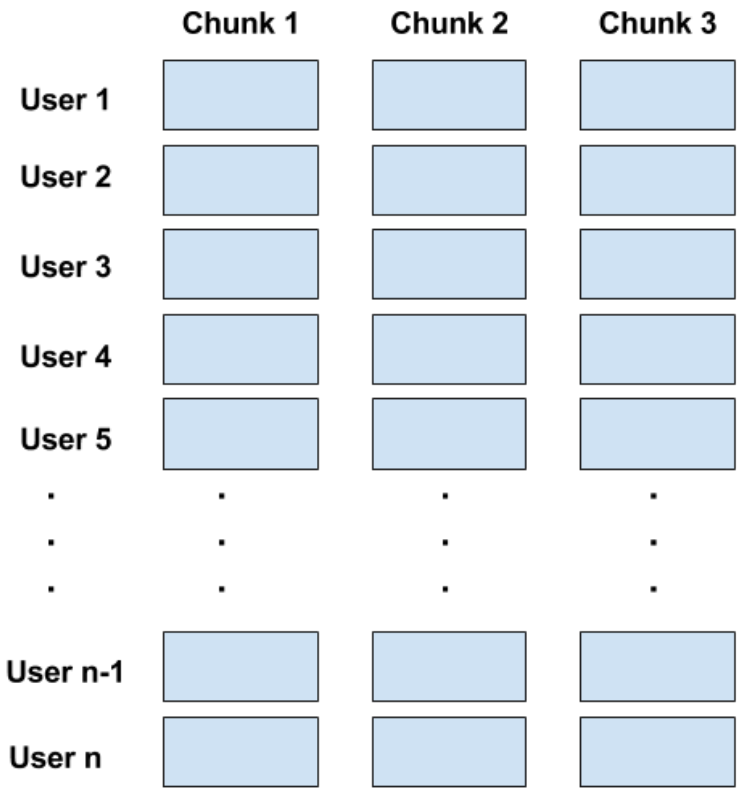
Conjunto de prueba, provisto bloque por bloque ...



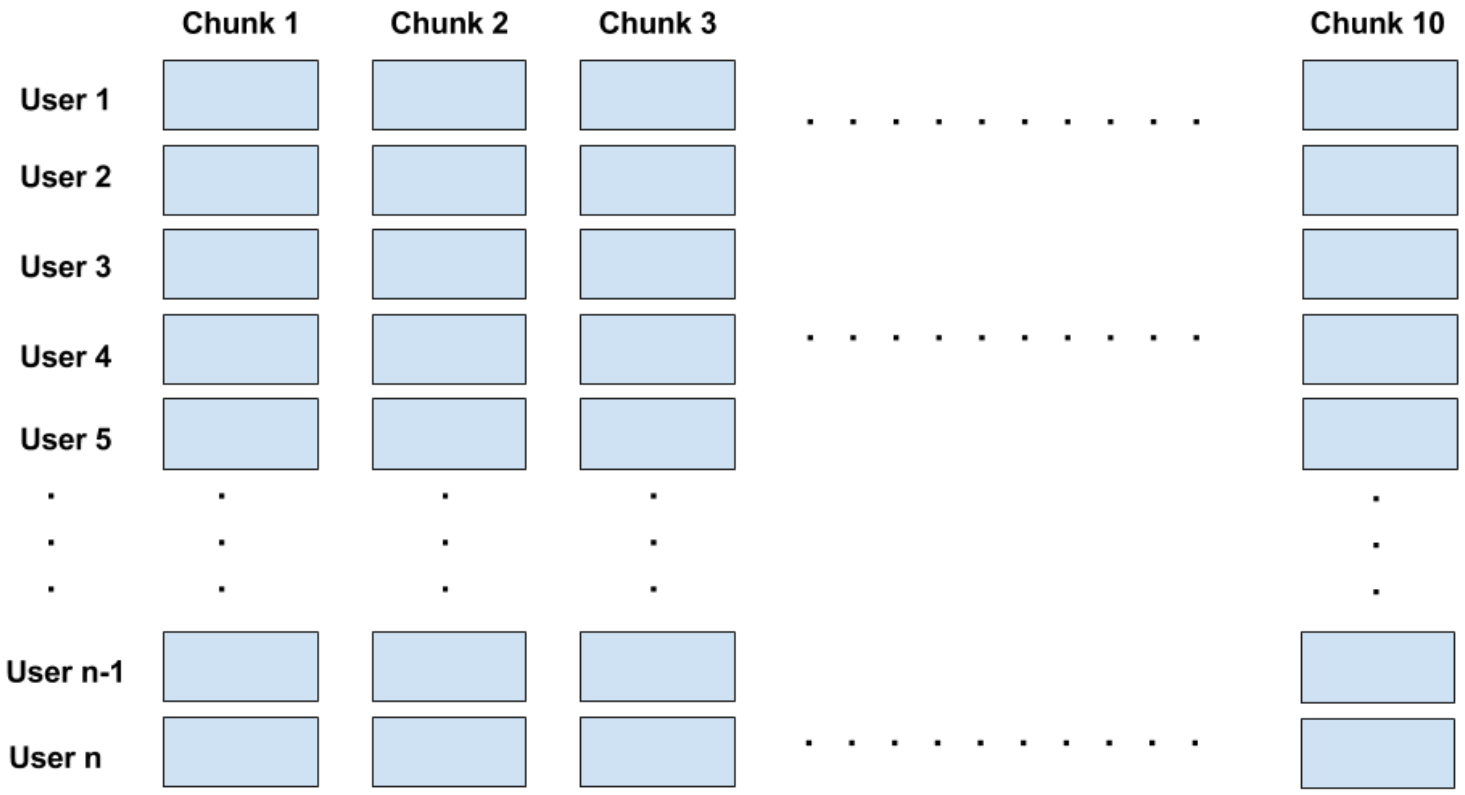
Conjunto de prueba, provisto bloque por bloque ...



Conjunto de prueba, provisto bloque por bloque ...



Conjunto de prueba, provisto bloque por bloque ...



La tarea eRisk 2018

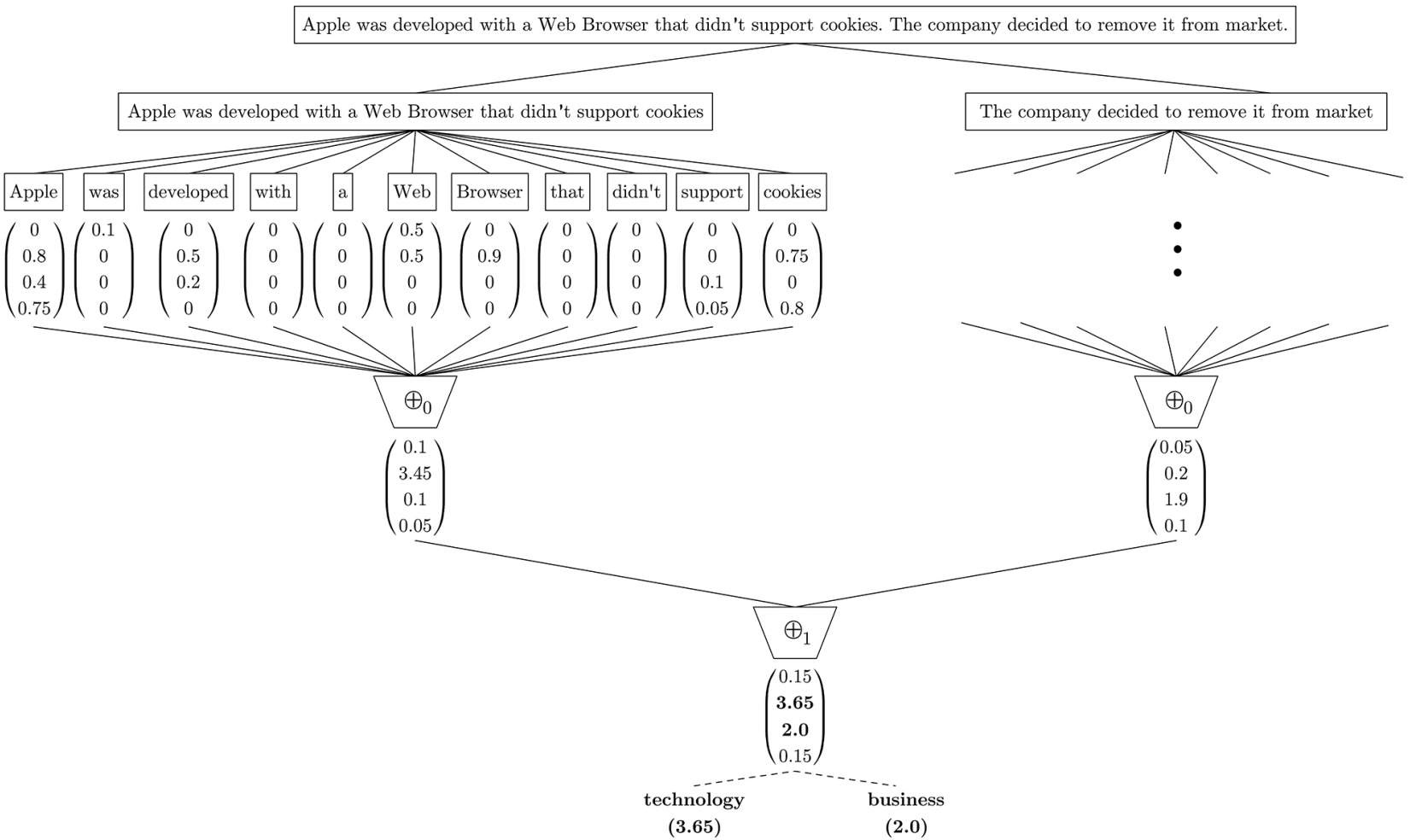
Enfoques utilizados

- FTVT
- Sequential-Incremental Classification (SIC)

Resultados obtenidos

- Mejores valores de $ERDE_5$ tanto en anorexia como depresión
- Valor más alto de precisión en anorexia (0.91)

Nuestro enfoque actual para DAR: SS3



Explicado en ...

Expert Systems With Applications 133 (2019) 182–197

Contents lists available at ScienceDirect

Expert Systems With Applications

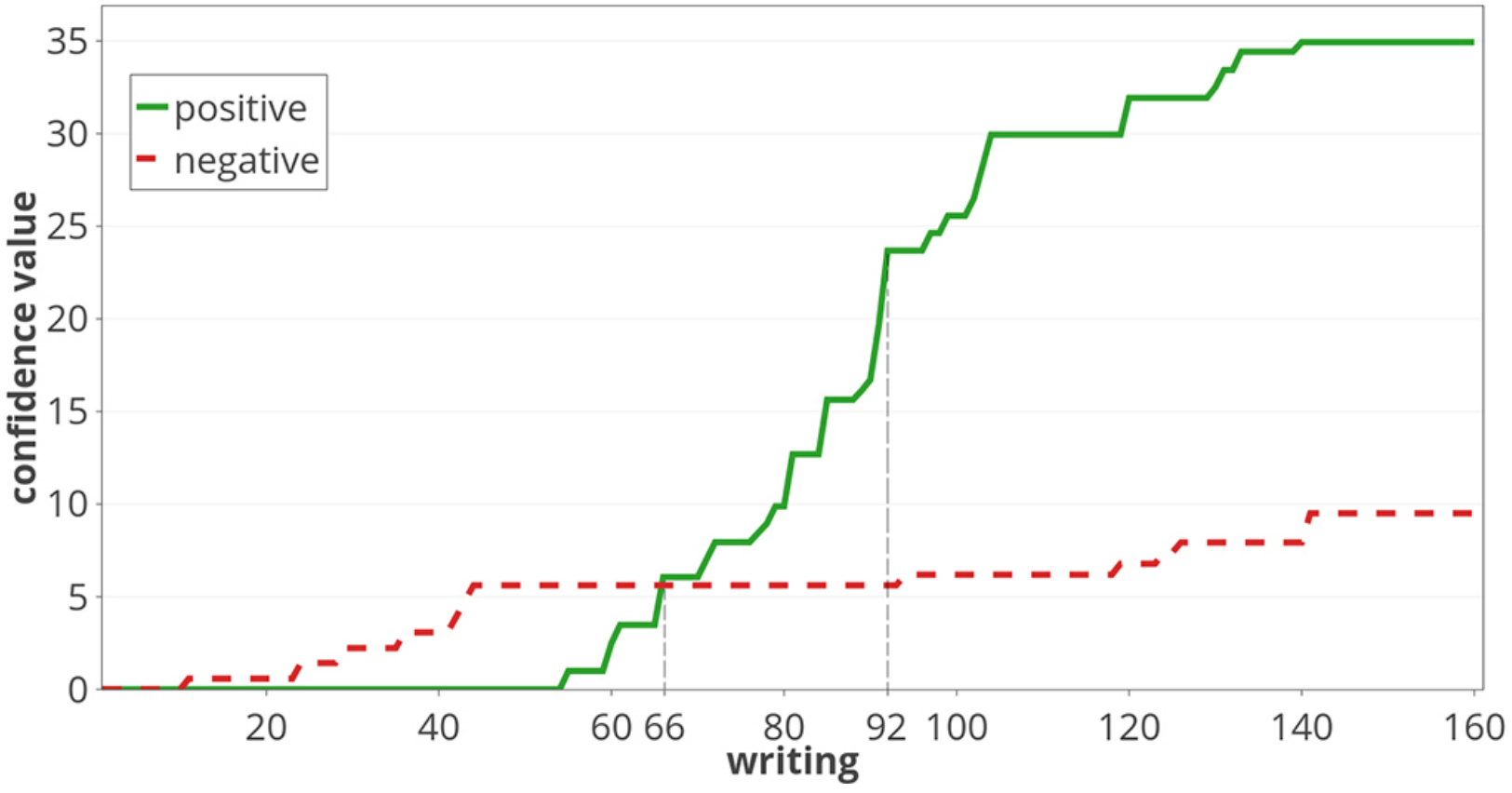
journal homepage: www.elsevier.com/locate/eswa

A text classification framework for simple and effective early depression detection over social media streams

Sergio G. Burdisso[a,b,*], Marcelo Errecalde[a], Manuel Montes-y-Gómez[c]

[a] Universidad Nacional de San Luis (UNSL), Ejército de Los Andes 950, San Luis, San Lius C.P. 5700, Argentina
[b] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina
[c] Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Luis Enrique Erro No. 1, Sta. Ma. Tonantzintla, Puebla C.P. 72840, Mexico

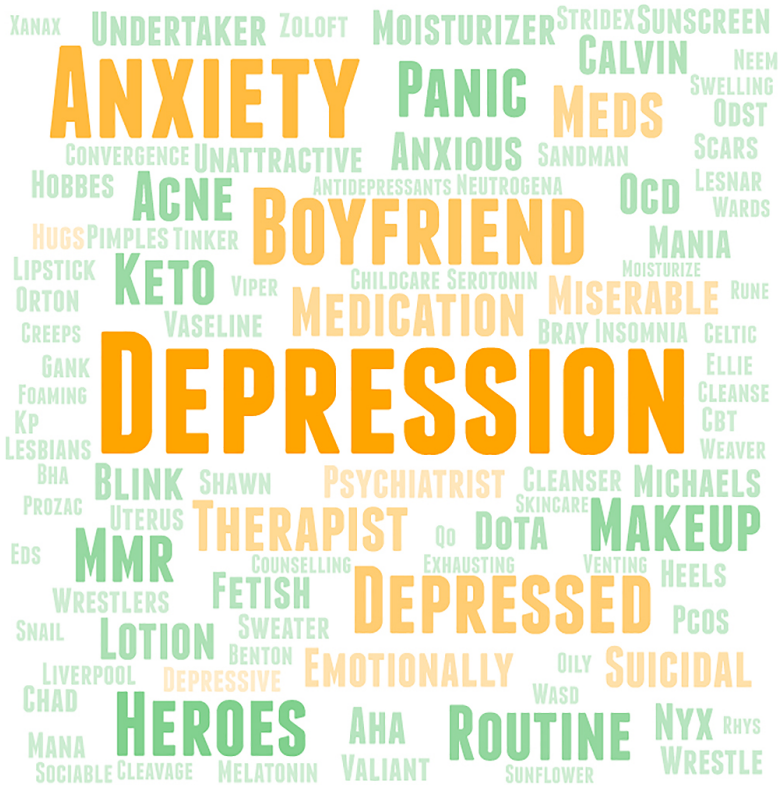
Método flexible para clasificación de casos de riesgo



Reconocimiento de “gránulo fino” de palabras relevantes

Tamaño por valor global (SS3)

Tamaño por frecuencia común



Interpretación gráfica de resultados (caja blanca)

...

Writing 54 ▶ I'm going to agree with everyone else and say you definitely need a lawyer. Get in touch with the[...]

Writing 55 ▶ You don't mention what the fertility issue is (and you don't have to) but his feelings may stem fr[...]

...

Writing 59 ▶ Thankfully I was able to realize that I was in a bad place and get help. My sister has been awesom[...]

Writing 60 ▶ I have been seeing a therapist which I think is helping a little. Fact is, I was feeling really depressed[...]

Writing 61 ▶ My Wife Wants a Divorce . This will be long, sorry in advance. My wife told me shortly after the[...]

Writing 62 ▶ the Earth Arena coming up I have: Zelnite x2 Dilma Ophelia For my last spot, should I use Miku [...]

...

(a) Subject 9579's history - writing level

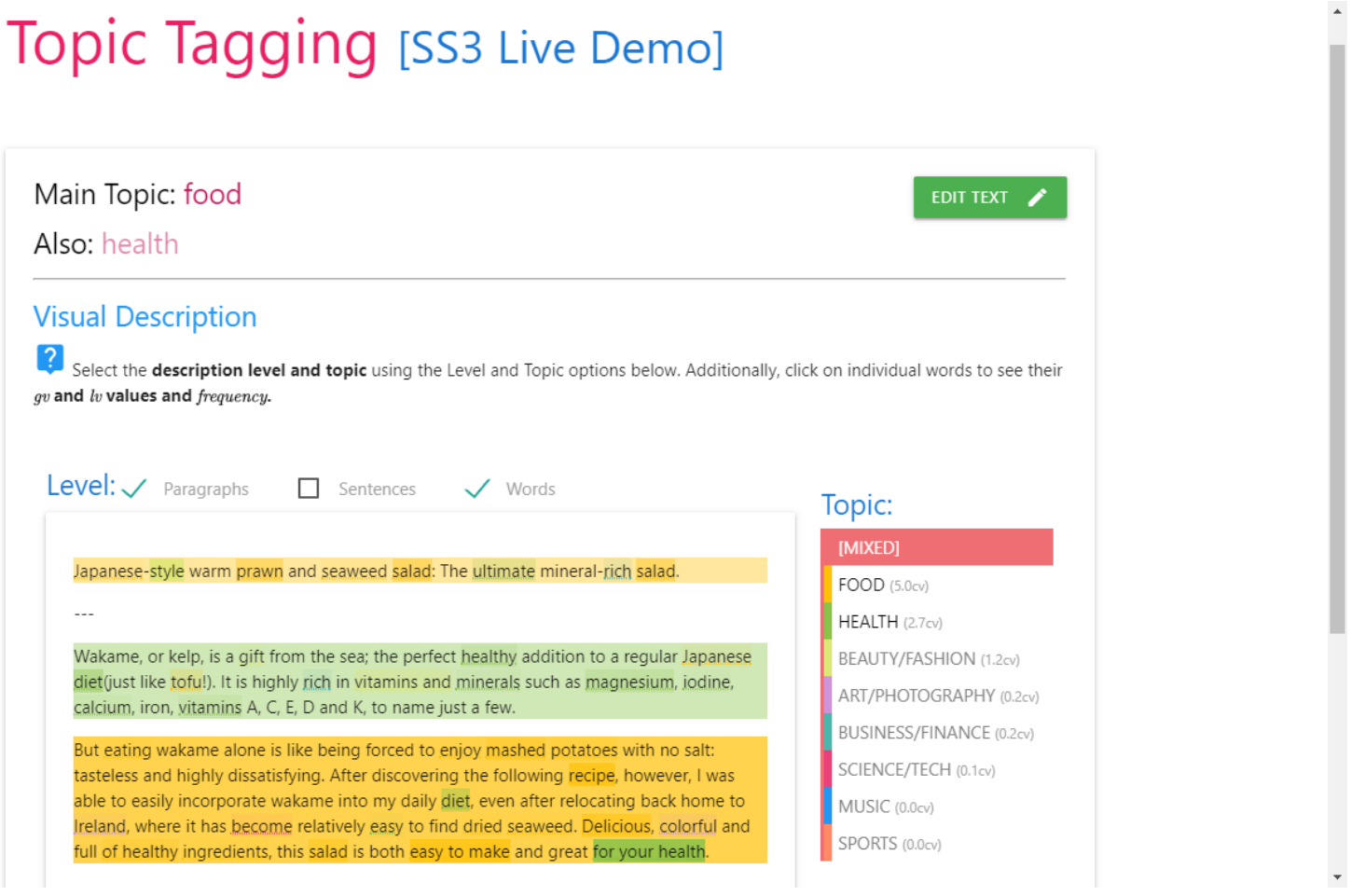
I have been seeing a therapist which I think is helping a little. Fact is, I was feeling really depressed and wanting to kill myself. I spent basically all of Feb in the hospital[...]

(b) Writing 60 - sentence level

I have been seeing a therapist which I think is helping a little. Fact is, I was feeling really depressed and wanting to kill myself. I spent basically all of Feb in the hospital[...]

(c) Writing 60 - word level

SS3 en vivo (`http://tworld.io/ss3/`)



La tarea eRisk 2019

Enfoques utilizados

- Clasificador de texto SS3
- Estimador de la severidad de los síntomas de depresión (basado en SS3)

Resultados obtenidos

- SS3 fue el método más rápido
- Obtuvo el mejor *ERDE* y las mejores medidas basadas en ranking en todas las tareas.
- La mejor precisión, F_1 y $F_{latency}$ en la tarea T2
- En la tarea T3, obtuvo los mejores valores de AHR y ACR, y el segundo mejor valor de ADODL y DCHR.

Nueva tarea: completar el cuestionario Beck's Depression Inventory (BDI)

```
Instructions:

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.

3. Past Failure
0. I do not feel like a failure.
1. I have failed more than I should have.
2. As I look back, I see a lot of failures.
3. I feel I am a total failure as a person.

4. Loss of Pleasure
0. I get as much pleasure as I ever did from the things I enjoy.
1. I don't enjoy things as much as I used to.
2. I get very little pleasure from the things I used to enjoy.
3. I can't get any pleasure from the things I used to enjoy.
```