

Clase 6 - Recursos y aplicaciones

Marcelo Errecalde^{1,2}

¹Universidad Nacional de San Luis, Argentina 

²Universidad 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

Resumen

1 Recursos para PLN

2 Aplicaciones

- Atribución de Autoría
- Author Profiling
- Detección de Plagios

3 Aplicaciones: Nuestros trabajos recientes en el área

- Detección anticipada de riesgos: casos de depresión en la Web

Sitios conocidos

- **Guía de recursos para NLP de Stanford:**
<http://nlp.stanford.edu/links/statnlp.html>
- **National Centre for Text Miner:**
<http://www.nactem.ac.uk>
- **NLP research group de U. Sheffield:**
<http://nlp.shef.ac.uk/>
- **NLP en U. Columbia:**
<http://www1.cs.columbia.edu/nlp/index.cgi>

Plataformas genéricas

- **NLTK**. Basada en Python. Provee lectores de corpus, tokenizers, stemmers, taggers, parsers, etc.
<http://www.nltk.org/>
 - **GATE**. Basada en Java. Permite crear “pipelines” de componentes de PLN
<http://gate.ac.uk/>
 - **Freeling**. Conjunto de herramientas para análisis de lenguajes.
<http://nlp.lsi.upc.edu/freeling/>

Machine Learning y PLN

- **scikit-learn.**
<http://scikit-learn.org/>
- **Apache OpenNLP.**
<http://opennlp.apache.org/>
- **MALLET.**
<http://mallet.cs.umass.edu/>

Algunos recursos interesantes para Python

- **gensim**

<https://radimrehurek.com/gensim/>

- **spaCy**

<https://spacy.io/>

- **TextBlob**.

<https://textblob.readthedocs.io/en/dev/>

- **SyntaxNet**.

<https://opensource.google.com/projects/syntaxnet>

<https://ai.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html>

- **solidscraper**.

<https://github.com/sergioburdisso/solidscraper>

APIs y recursos on-line

- **MorphAdorner**

[http://morphadorner.northwestern.edu/
morphadorner/](http://morphadorner.northwestern.edu/morphadorner/)

- **NetSpeak, Altools, etc.**

[https://www.uni-weimar.de/en/media/chairs/
computer-science-department/webis/home/](https://www.uni-weimar.de/en/media/chairs/computer-science-department/webis/home/)

- **Text Analysis Tools.** Buena info de herramientas ...

- **Topic Tagging (ss3).**

<http://tworld.io/ss3/>

- **Reverb (Open IE).** Extracción de información abierta

<http://reverb.cs.washington.edu/>

Problema de atribución de autoría

- Dado un **conjunto de autores** candidatos de los que se dispone ejemplos de textos de autoría indiscutida, el problema consiste en asignar un texto de **autoría desconocida** a **uno** de los autores candidatos.
- Problema de categorización de **múltiples clases** y **único rótulo** (single-label).
- Tarea también conocida como **identificación de autoría/autor**.

Otras tareas de análisis de autoría

- **Verificación de autor:** dado un texto, corresponde a un autor específico o no? (caso **binario**).
- **Detección de plagio:** encontrar similitudes entre dos textos.
- **Caracterización/perfil del autor:** extraer información sobre la edad, nivel de educación, sexo, nacionalidad, etc) del autor de un texto dado.
- **Detección de inconsistencias estilísticas**

Features

Están orientadas a cuantificar el **estilo** de escritura.

Stamatatos organiza estas features estilísticas/estilométricas en:

- Caracter
 - Léxicas
 - Sintácticas
 - Semánticas
 - Específicas de la aplicación

Métodos de Atribución

En todo problema de identificación de autoría hay:

- un conjunto de **autores candidatos**
- un conjunto de **ejemplos de muestra** de autoría conocida cubriendo todos los autores candidatos (training corpus)
- un conjunto de **textos** de autoría desconocida que deben ser atribuídos a un autor candidato (test corpus)

Los métodos de atribución de autoría se diferencian de acuerdo a si tratan cada ejemplo de entrenamiento **individualmente** (basados en **instancia**) o **acumulativamente** (por autor) (basados en **perfil**)

Enfoques basados en perfil

Arquitectura típica (Stamatatos):

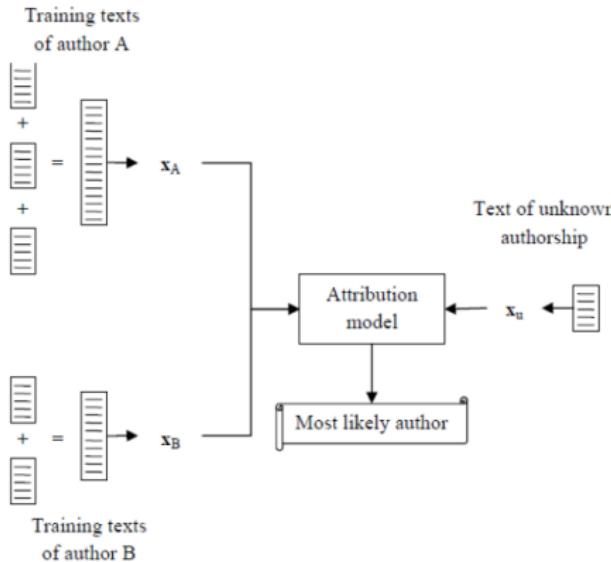


FIG 1. Typical architecture of profile-based approaches.

Enfoques basados en perfil

- todos los textos de entrenamiento correspondientes a cada autor se concatenan en un único archivo (**perfil del autor**)
- Este enorme archivo es usado para extraer las propiedades del estilo del autor
- un documento a clasificar es comparado con cada perfil en base a una medida de **distancia** o **similitud**
- el autor más probable es aquel cuyo perfil es **más cercano** al documento a clasificar

Enfoques basados en instancias

Arquitectura típica (Stamatatos):

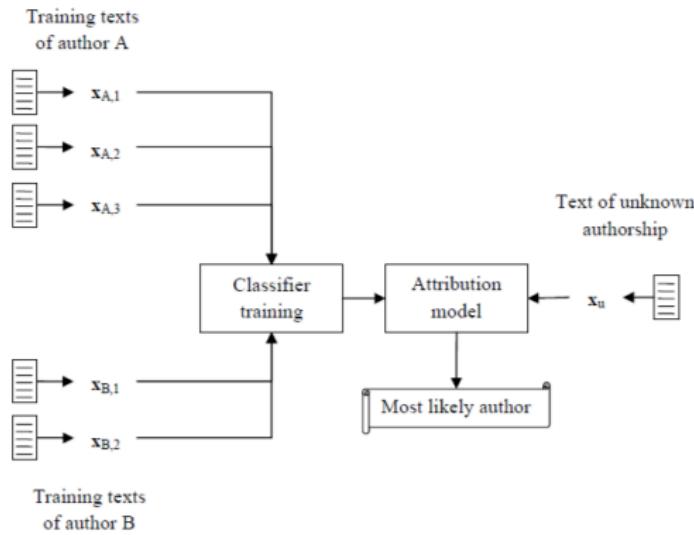


FIG 2. Typical architecture of instance-based approaches.

Enfoques basados en instancias

- Corresponden al enfoque usual en categorización de textos, con cada ejemplo representado **individualmente**
- Enfoques usuales:
 - SVM
 - Bayes
 - Árboles de decisión y NN
- También existen propuestas de enfoques **híbridos** que los combinan con métodos basados en perfiles.

Software Disponible para atribución de autoría:

JGAAP (Java Graphical Authorship Attribution Program)

http://ev11labs.com/jgaap/w/index.php/Main_Page

Author Profiling

Author Profiling (Caracterización/perfil del autor)

- Tarea también denominada **identificación del perfil del autor**
- Consiste en identificar las **características** o **rasgos** que integran el perfil de una persona como por ejemplo:
 - la **edad**
 - el **nivel de educación**
 - **sexo**
 - tipo de **personalidad**
 - **nacionalidad**
 - **lenguaje de origen**, etc.
- Tarea importante en **seguridad**, **marketing** y área **forense**

Author Profiling (Caracterización/perfil del autor)

- Ejemplos:
 - Determinar si un mensaje de Tweeter, fue escrito por un hombre o una mujer
 - Determinar si un documento fue escrito por una persona entre 13-17 años, 23-27 años o más de 33 años
 - Determinar el lenguaje nativo de un e-mail en inglés con una amenaza

Author Profiling (Caracterización/perfil del autor)

- Involucra determinar rasgos estilísticos y de contenido de un grupo de autores
 - Ejemplos
 - la longitud de las sentencias en los blogs tiende a ser más corta en los adolescentes que en los adultos
 - los pronombres personales tienen alta ganancia de información para identificar mujeres.
 - Los hombres tienden a hablar más de política y de deportes
 - las mujeres en los blogs usan la palabra friday el triple que los hombres, y los adolescentes cinco veces la frecuencia de los adultos

Features utilizadas

Son similares a las usadas en **atribución de autoría**

- Palabras de función (stopwords) y POS features (textos formales)
- las dos previas + palabras “non-dictionary”, palabras muy informales/coloquiales (“slang”), longitud promedio de sentencias + palabras de alta IG (contenido) (en blogs)
- En la competencia de PAN:
 - feat. **estilísticas**: frecuencias de signos de puntuación, mayúsculas, comillas, POS tags, índices de legibilidad, emoticones, *n*-gramas, etc.
 - feat. de **contenido**: BOW, entidades nombradas, palabras de emoción, de sentimiento, slang, contracciones, etc.
 - El ganador, usó una representación de **segundo orden** (Análisis de Semánticas Concisas)

¿Qué es Plagiar?

- Copiar en lo sustancial obras ajenas, dándolas como propias (RAE).
- Plagiar es robar el crédito por el trabajo realizado por otra persona (Barrón Cedeño).
- El simple **reúso** de información (RI) **no implica** necesariamente plagio (Platón de Sócrates, investigación científica, etc).
- Pero el RI debe tener como motivación el objetivo de enriquecerla. No se trata simplemente de un proceso de **extracción de información**, sino que implica su **obtención, raciocinio y generación** de una nueva versión, ya sea totalmente diferente o enriquecida de manera significativa.

Principales enfoques a la detección de plagios

- **Análisis intrínseco de plagio (AIP)**: el único recurso utilizado es el texto sospechado. Se pueden determinar fragmentos sospechosos pero no determinar la fuente.
- **Detección de plagio con referencia (DPR)**: se requiere de un conjunto de documentos originales con el objetivo de buscar el origen de los fragmentos potencialmente plagiados dentro de un texto sospechoso. Es posible obtener la fuente original del texto plagiado.

Características útiles para detectar plagios (Clough 2000)

- Nuevo vocabulario utilizado (respecto a documentos previos).
- Cambio de vocabulario (a través de un documento).
- Uso similar de los signos de puntuación.
- Cantidad de texto común entre documentos.
- Errores en común.
- Distribución del uso de palabras.
- Estructura sintáctica del texto.
- Largas secuencias de texto en común.
- Preferencia por el uso de sentencias cortas o largas.
- Legibilidad del texto.
- Referencias incongruentes.

Análisis Intrínseco de Plagios

Idea principal: capturar el estilo y la complejidad de un documento sospechoso, en base a un conjunto de parámetros:

- Longitud de sentencias.
- Categorías gramaticales (POS) utilizadas
- Número promedio de stopwords.
- Índice de Confusión de **Gunning**: mide qué tan comprensible es un texto escrito.
- Índice de **Flesch-Kincaid**: similar a la anterior.
- Índice de **Dale-Chall**: similar a la anterior.
- Función **R**: intenta capturar la variedad en el vocabulario de un autor.
- Función **K**: otra medida para la riqueza de vocabulario.

Las medidas se calculan sobre **todo el texto** y se comparan con las obtenidas en **los párrafos**

Detección de plagio con referencia

- Análisis a **nivel de documentos**: tipo **SCAM**. Detecta relaciones de **plagio, subconjunto, copia y relación**. Usa una adaptación de la medida de similitud coseno.
- Análisis basados en comparación de *n*-gramas: comparación exhaustiva de *n*-gramas. Prototipo **Ferrett**. Mejores resultados con tri-gramas. Comparaciones con **Jaccard** sobre *n*-gramas y variante para comparación de longitud no comparable.
- Análisis a nivel de sentencia: se compara sentencia sospechosa a cada una de las sentencias de referencia. Ejemplo: **PPChecker**. Considera:
 - 1 Copia **exacta**.
 - 2 Copia con **inserción** de palabras.
 - 3 Copia con **eliminación** de palabras.
 - 4 **Reescritura**: (se realiza expansión de vocabulario de sentencias con Wordnet)

Algunos sistemas para detección de plagio

- **Sherlock** (University of Warwick, UK).
<http://www2.warwick.ac.uk/fac/sci/dcs/research/ias/software/sherlock/>
- **WCopyFind** (University of Virginia, US).
<http://plagiarism.bloomfieldmedia.com/z-wordpress/software/wcopyfind/>
- **Antiplagiarist** (ACNP Software).
www.anticutandpaste.com/antiplagiarist/
- **seeSources.com**
<http://www.plagscan.com/seesources/>
- **EVE2 (The Essay Verification Engine)** (Canexus).
<http://www.canexus.com/>

Algunos sistemas para detección de plagio (II)

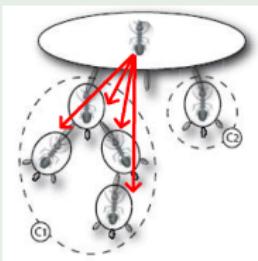
- **turnitin** (IParadigms).
<http://www.turnitin.com/>
- **Plagiarism-Finder** (Mediaphor).
<http://www.plagiarismfinder.de/>
- **Dupli Checker**
<http://www.duplichecker.com/>
- **Article Checker**
<http://www.articlechecker.com/>
- **Ferret** (University of Hertfordshire, UK).
- **SafeAssign** (Blackboard).

Nuestros trabajos recientes en Minería de Textos

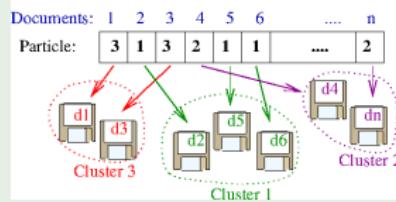
- Agrupamiento de textos cortos
- Calidad de información en la Web (esp. en [Wikipedia](#))
- Perfil del autor en medios sociales (edad, género, rasgos de personalidad, etc)
- Clasificación temprana de textos/Detección temprana de riesgos

(Agrupamiento de textos cortos - técnicas bio-inspiradas

Técnicas basadas en hormigas



Técnicas basadas en PSO



Publicaciones recientes

 Information Sciences
Volume 265, 1 May 2014, Pages 36–49

An efficient Particle Swarm Optimization approach to cluster short texts

Leticia Cagnina^a , Marcelo Erracalde^a , Diego Ingaramo^a, Paolo Rosso^b 

 Natural Language Engineering

Article

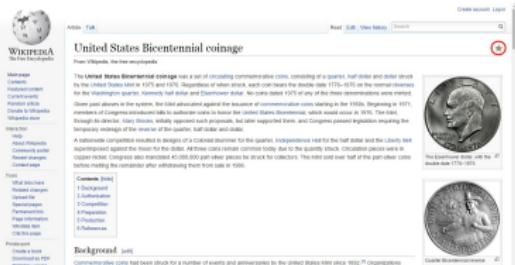
Get access

Natural Language Engineering, Volume 22, Issue 5
September 2016, pp. 687-726

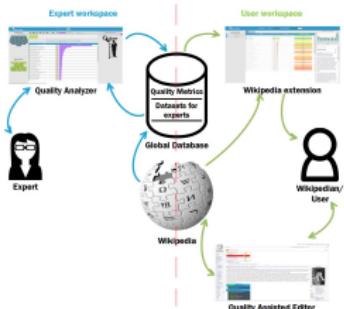
Silhouette + attraction: A simple and effective method for text clustering¹

Calidad de Información en Wikipedia

Artículos Destacados



Visualización



Detección de Fallas

Murghazar

From Wikipedia, the free encyclopedia

What article has multiple issues? Please help improve it or discuss these issues on the talk page. [edit]

August 2009

This article has multiple issues. Please help improve it or discuss these issues on the talk page. [\[help\]](#)

1. This is a list of cities in any country or territory. August 2009

2. This is an [eponym](#), as no other articles link to it. Please introduce links to this page from related articles; try the [Find link](#) tool for suggestions. August 2009

3. This article contains content or writing like an [advertisement](#). August 2009

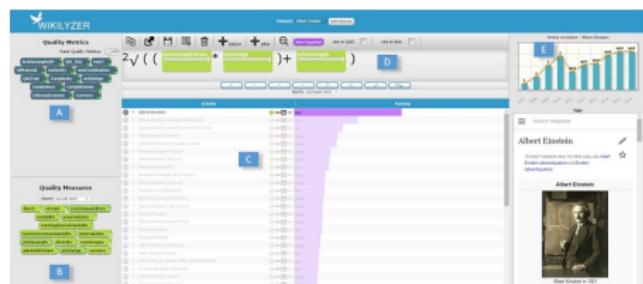
Discussion [Edit this box](#) [View history](#) [Page information](#) [Contributions](#) [Log in](#)

Categories: [People from Saudi Arabia](#)

The page you are viewing has been automatically created by Wikipedia's [Bot](#). The page is made of whole or partially machine-generated code. Please use the same machine used to generate this page to view any other places or regions. If you have any questions about how this page was generated, please ask the [Machine-readable page](#) or the [Machine-readable version](#). If you have any questions about the content of this page, please ask the [Machine-readable version](#).

[citation needed]

Métricas de Calidad



Detección de fallas en Wikipedia en español

Information Processing and Management 54 (2018) 1169–1181



Contents lists available at ScienceDirect

Information Processing and Management

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



Quality flaw prediction in Spanish Wikipedia: A case of study with verifiability flaws

Edgardo Ferretti^{a,b}, Leticia Cagnina^{a*,a,b,c}, Viviana Paiz^a, Sebastián Delle Donne^a, Rodrigo Zacagnini^a, Marcelo Errecalde^{a,b}

^a Departamento de Informática, Universidad Nacional de San Luis (UNSL), Ejército de los Andes 950, San Luis, Argentina

^b Laboratorio de Investigación y Desarrollo en Inteligencia Computacional (UNSL), Argentina

^c Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina



Detección de fallas en Wikipedia en español: features

Table 1
Features that comprise the document model.

Feature	Description
<i>Content-based</i>	
Character count	Number of characters in the text (no spaces).
Word count	Number of words in the plain text.
Sentence count	Number of sentences in the plain text.
Word length	Average word length in characters.
Sentence length	Average sentence length in words.
Paragraph count	Number of paragraphs.
Paragraph length	Average paragraph length in sentences.
Longest word length	Length in characters of the longest word.
Longest sentence length	Number of words in the longest sentence.
Shortest sentence length	Number of words in the shortest sentence.
Long sentence rate	Percentage of long sentences. A long sentence is defined as containing at least 30 words.
Short sentence rate	Percentage of short sentences. A short sentence is defined as containing at most 15 words.
Longest subsection length	Length in words of the longest subsection.
Shortest subsection length	Length in words of the shortest subsection.
Subsections length	Total number of words in the article's subsections.
Average subsection length	Average number of words per subsection.
Longest subsubsection	Length in words of the longest subsubsection.
Shortest subsubsection	Length in words of the shortest subsubsection.
Subsubsections length	Total number of words in the article's subsubsections.
Average subsubsections	Average number of words per subsubsection.
<i>Structure-based</i>	
Section count	Number of sections.
Subsection count	Number of subsections.
Subsubsection count	Number of subsubsections.
Heading count	Number of sections, subsections and subsubsections.
Section nesting	Average number of subsections per section.
Subsection nesting	Average number of subsubsections per subsection.
Reference Sections Count	Number of reference sections, e.g. "References", "Footnotes", "Sources", "Bibliography".
Mandatory Sections Count	Number of mandatory sections, e.g. "See also".
Related page count	Number of related pages, e.g. "Further reading", "See also", etc.
Lead length	Number of words in the lead section (text before the first heading).
Lead rate	Percentage of words in the lead section.
Image count	Number of images.
Image rate	Ratio of image count to section count.
Link count	Every occurrence of a link (introduced with two open square brackets) in the unfiltered text.
Link rate	Percentage of links.
Table count	Number of tables.
Reference count	Number of all references using the <code><ref>...</ref></code> syntax.
Reference section rate	Ratio of reference count to the accumulated section, subsection and subsubsection count.
Reference word rate	Ratio of reference count to word count.
Unique reference count	Number of unique references using the <code><ref>...</ref></code> syntax.
Reference ratio	Ratio between the reference word rate of the article and the maximum reference word rate found in the dataset.
Templates-count	Number of (different) Wikipedia templates.

Detección de fallas en Wikipedia: PU-Learning

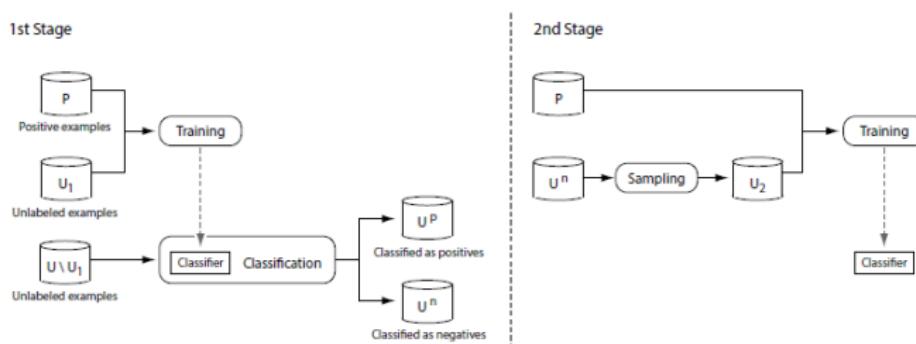
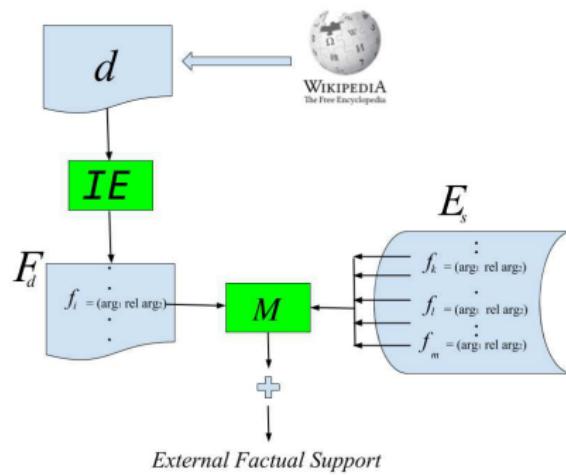


Figure 2. Non-iterative proposed approach for the two-step strategy to PU Learning

"On the Use of PU Learning for Quality Flaw Prediction in Wikipedia". Edgardo Ferretti, Donato Hernández, Rafael Guzman, Manuel, Montes-Y-Gómez, Marcelo Errecalde and Paolo Rosso. CLEF2012 - Labs. Track: PAN - Quality Flaw Prediction in Wikipedia. 2012

"On the Use of Reliable-Negatives Selection Strategies in the PU Learning Approach for Quality Flaws Prediction in Wikipedia". Edgardo Ferretti, Marcelo Errecalde, Maik Anderka and Benno Stein. In IEEE Proceedings of DEXA-2014 Publisher: IEEE, Munich, Germany, 2014.

Métricas de Calidad en Wikipedia:: Información factual



- "Measuring the Quality of Web Content using Factual Information". E. Lex, M. Voelske, M. Errecalde, E. Ferretti, L. Cagnina, Ch. Horn, M. Granitzer, B. Stein. Proc. of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality, 2012. ACM, New York, NY, USA.
- "On the Feasibility of External Factual Support as Wikipedia's Quality Metric", C. G. Velázquez, L. C. Cagnina, M. L. Errecalde. Revista del Procesamiento del Lenguaje Natural (SEPLN), Vol. 58, pp. 93-100. 2017.

Análisis de Calidad Interactivo: Wikilizer

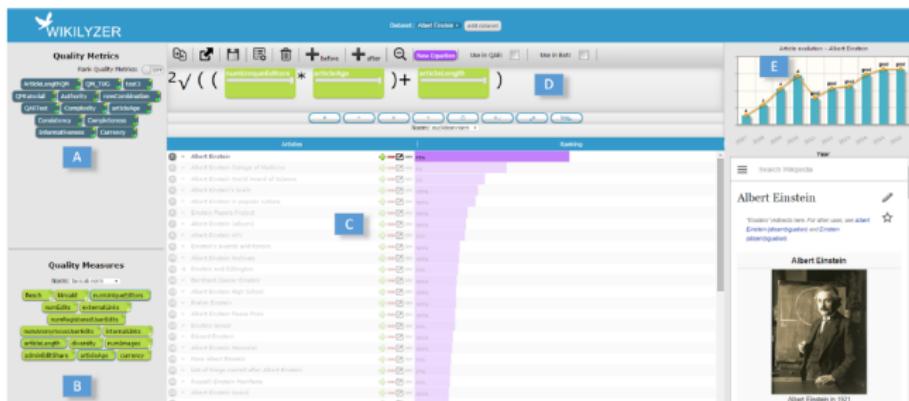


Fig. 1. *Quality Agent UI.* (A) QM Panel: contains built-in and custom QMs. (B) Attributes Panel: displays all available attributes to create QMs. (C) Ranking View: ranks articles in the current dataset by the selected QM(s). (D) Equation Composer: area where experts can create QMs through mathematical combinations of attributes and/or other QMs. (E) Article Panel: shows quality-based evolution of the selected article (top) and its mobile version (below).

“Wikilizer: Interactive Information Quality Assessment in Wikipedia”. Cecilia di Sciascio, David Strohmaier, Marcelo Errecalde,. Proceedings of the 22nd International Conference on Intelligent User Interfaces. Pages 377-388. Limassol, Cyprus - March 13 - 16, 2017. ACM New York, NY, USA, 2017.

Perfil de autor



Tareas: determinar

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

Perfil de autor



"Cross Domain Author Profiling Task in Spanish Language: An Experimental Study".
María José Garcíarena Ucelay, María Paula Villegas, Leticia C. Cagnina and Marcelo L. Errecalde. Journal of Computer Science and Technology (JCST). Vol. 15. Nro. 2, pp. 122-128. Iberoamerican Science & Technology Education Consortium.
México/Argentina. ISSN: 1666-6038, November 2015.

Detección de pedófilos en la Web

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

Ejemplo:

Example 1: what nationality are u?

Exchange of
information

Example 2: what r u wearing?

Grooming

Example 3: would u let me?

Example 4: thing it is me feeling u

Example 5: what's your address?

Example 6: can I stay at your house overnight if i go?

Approach

Detección de pedófilos en la Web

- Enfoque: utilizar un **perfil de n-gramas de caracteres** con información de las etapas y una adecuada función de **similitud** para detectar los cambios de escritura.

“Character n-grams profiles for predatory lines detection”. Leticia C. Cagnina, Marcelo L. Errecalde and Efstathios Stamatatos. Proceedings del Workshop de Procesamiento Automatizado de Textos y Corpus 2016 (WOPATEC-2016),

Detección de pedófilos en la Web

Detección de depredadores sexuales en chats

- 1) Elaboración del perfil de la etapa (perfil del depredador):

Grooming

Líneas catalogadas extraídas de conversaciones reales de pedófilos.

Hi how are you?
Have you been? Have you any plan?
How old are you? Are you alone?
Give me your address.
I want see you tonight
You are perfect for me. I like you
Do you want to see me tomorrow?
Give me you opinion about this picture.
Where do you think is a good place?



Hi_
i_h
ho_
how
ow_
w_a
ar_
are
re_
....

3gramas de caracteres

- 2) Elaboración del perfil de la conversación

Procesar la conversación utilizando “ventanas móviles”.

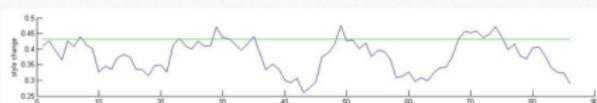
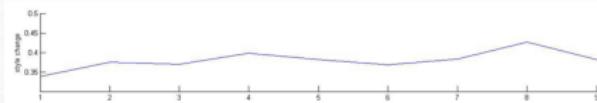
Elaborar el perfil de la ventana.

[Adamou217]	whats ur screen name..miles adamou217..on either
[superboy_93]	superboy93
Adamou217	who's tim
DFW Superbo	my best friend
Adamou217	ah the bl one
Adamou217	u boys ever play together..lol
DFW Super	yea we have ;)
Adamou217	
Adamou217	when do i get a chance..lol
Adamou217	is she home tonight
DFW Superboy 93	so far
DFW Superboy 93	she has to work tomorrow
Adamou217	ah well maybe she'll go to be early..:)

Detección de pedófilos en la Web

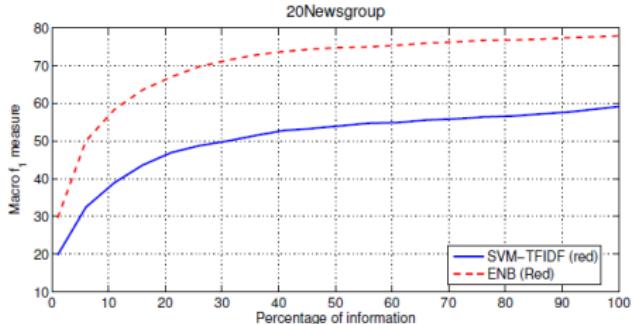
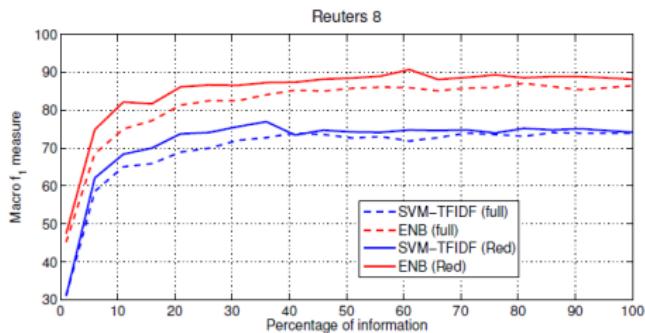
Detección de depredadores sexuales en chats

-
- 3) Buscar similitudes entre el perfil de la ventana y el perfil del depredador



Clasificación temprana/anticipada

- Entrenar con información secuencial completa.
- Luego clasificar, tan pronto como sea posible



Clasificación temprana/anticipada

- Entrenar con información secuencial **completa**.
- Luego clasificar, **tan pronto como sea posible**

“Early text classification: a Naïve solution”. Hugo Jair Escalante, Manuel Montes y Gómez, Luis Villaseñor-Pineda and Marcelo Luis Errecalde. Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2016), pp. 91-99, ISBN 978-1-941643-82-2, NAACL-HLT 2016, Association for Computational Linguistics, San Diego, California, June 2016. URL: www.aclweb.org/anthology/W16-0416

Detección anticipada de riesgos: casos de depresión en la Web

Depresión: un flagelo de nuestros días

La depresión en la salud pública

- En 2015, se estimó un **4,4 %** de la población mundial con depresión (más de **332 millones** de personas).
- Más allá de los efectos perjudiciales que tiene la depresión, ésta puede llevar al **suicidio**
- Más de 800.000 muertes por suicidio por año y segunda causa de muerte en el rango de 15 a 29 años.

En este contexto, la investigación de métodos que asistan en la **detección** de posibles casos de **depresión** se vuelve de **gran importancia**

Detección anticipada de riesgos: casos de depresión en la Web

Algunos enfoques de detección automática de depresión (DAD)

- Stirman y Pennebaker, en 2001, usan **Linguistic Inquiry and Word Count** (LIWC) para caracterizar depresión a partir del uso del lenguaje natural (patrones interesantes con pronombres)
- De Choudhury y otros, en 2013, analizan tweets de sujetos diagnosticados de depresión, entrenando con SVM. Cuentan con scores de tests de depresión (CES-D y BDI), información de la red de contactos, índices de insomnio, etc.)

Detección anticipada de riesgos: casos de depresión en la Web

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

Detección anticipada de riesgos: casos de depresión en la Web

Una tarea internacional dedicada al tema

The background image is a grayscale photograph of a person's profile, facing right. The person has long hair that is blowing in the wind. The lighting is soft, creating a contemplative atmosphere. Overlaid on this image is the text for the ERISK 2019 workshop.

Detección anticipada de riesgos: casos de depresión en la Web

Ediciones de eRisk

eRisk 2017 (<https://early.irrlab.org/2017/index.html>)

- Task: Early Detection of Depression

eRisk 2018 (<https://early.irrlab.org/2018/index.html>)

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

eRisk 2019 (<https://early.irrlab.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

Detección anticipada de riesgos: casos de depresión en la Web

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 $ERDE_{50}$

Detección anticipada de riesgos: casos de depresión en la Web

Conjuntos de Datos

Conj. de Entrenamiento

486 usuarios

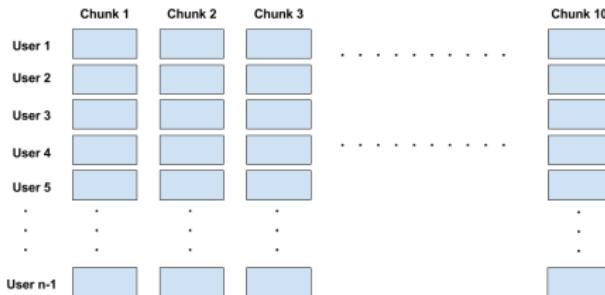
- 83 positivos (deprimidos)
- 403 negativos (no-dep.)

Conjunto de Prueba

401 usuarios

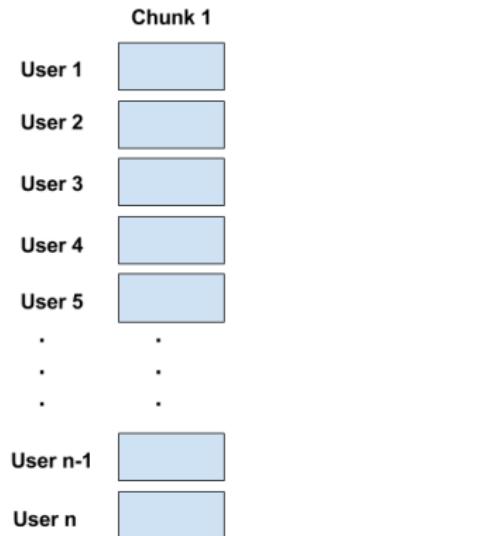
- 52 positivos (deprimidos)
- 349 negativos (no-dep.)

Divididos en 10 bloques (ordenados cronológicamente):



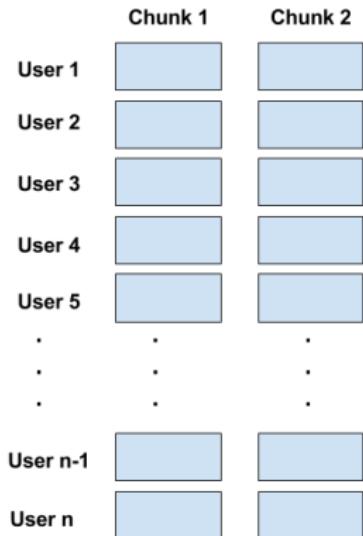
Detección anticipada de riesgos: casos de depresión en la Web

Conjunto de prueba, provisto bloque por bloque ...



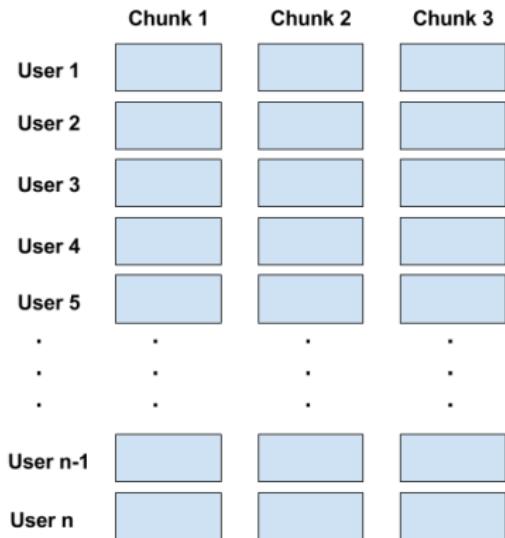
Detección anticipada de riesgos: casos de depresión en la Web

Conjunto de prueba, provisto bloque por bloque ...



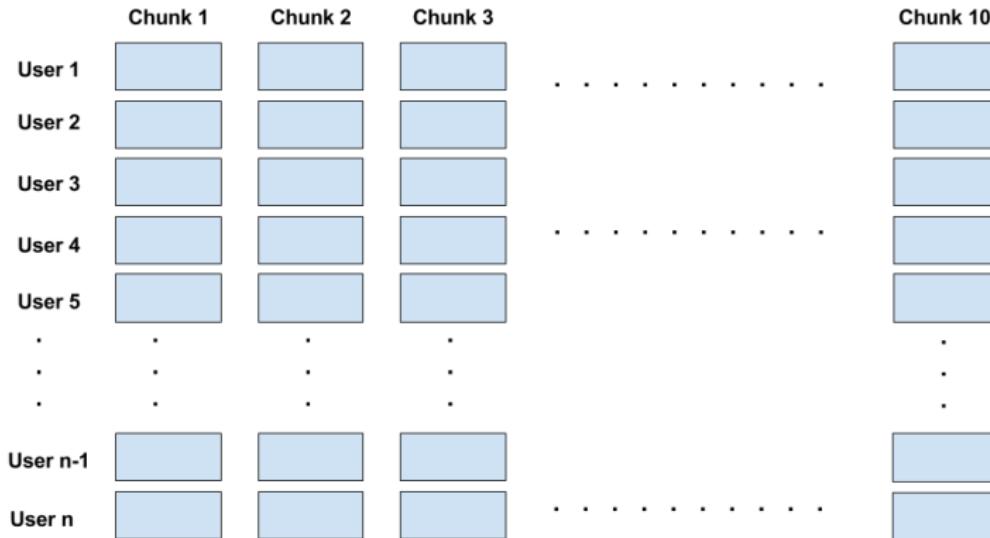
Detección anticipada de riesgos: casos de depresión en la Web

Conjunto de prueba, provisto bloque por bloque ...



Detección anticipada de riesgos: casos de depresión en la Web

Conjunto de prueba, provisto bloque por bloque ...



Detección anticipada de riesgos: casos de depresión en la Web

La tarea eRisk 2018

Enfoques utilizados

- FTVT
- Sequential-Incremental Classification (SIC)

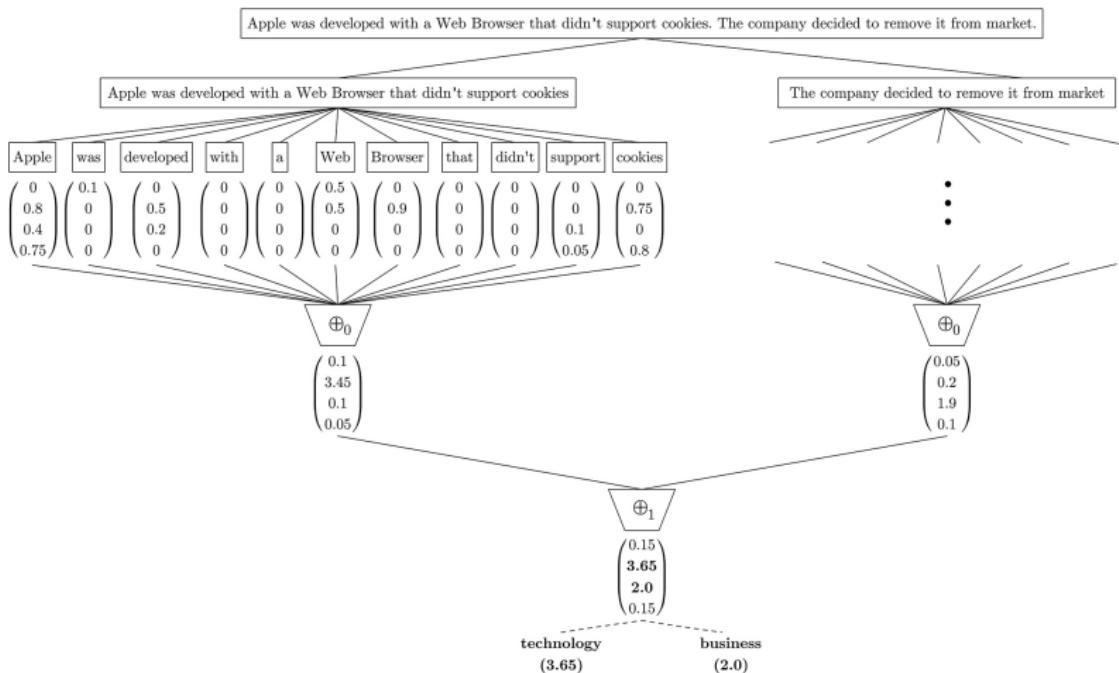
Resultados obtenidos

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

oooooo
ooooooooooooooooooooo
oooooooooooooooooooo●oooooooo

Detección anticipada de riesgos: casos de depresión en la Web

Nuestro enfoque actual para DAR: SS3



Detección anticipada de riesgos: casos de depresión en la Web

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

A circular icon with a red and blue design, with the text "Check for updates" below it.

Sergio G. Burdisso^{a,b,*}, Marcelo Errecalde^a, Manuel Montes-y-Gómez^c

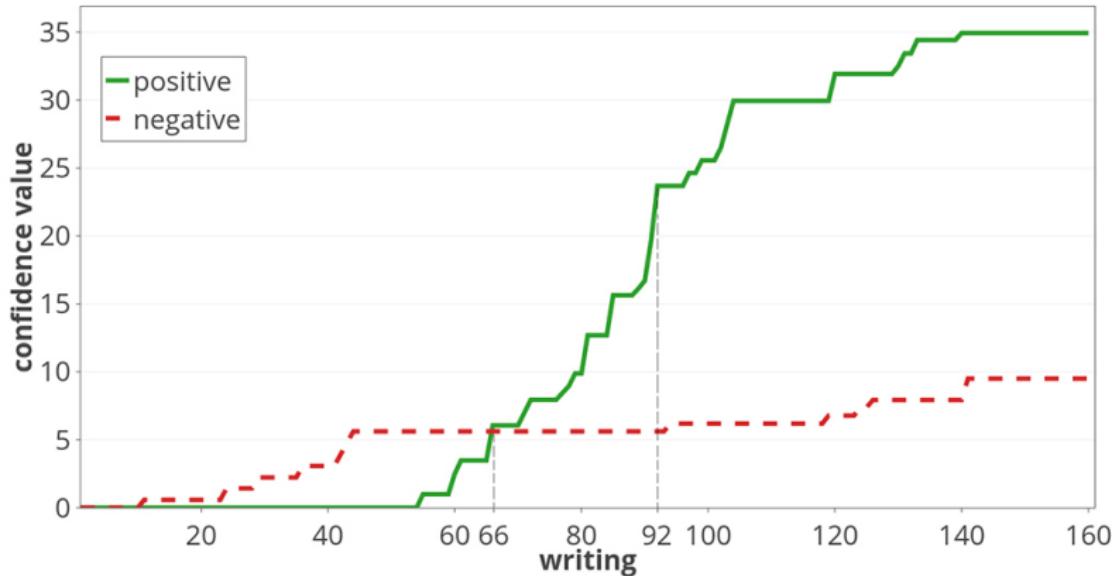
^a Universidad Nacional de San Luis (UNSL), Ejército de Los Andes 950, San Luis, San Luis C.P. 5700, Argentina

^bConsejo 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

Detección anticipada de riesgos: casos de depresión en la Web

Método flexible para clasificación de casos de riesgo



Detección anticipada de riesgos: casos de depresión en la Web

Reconocimiento de “gránulo fino” de palabras relevantes

Tamaño por valor global (SS3)

Tamaño por frecuencia común



Detección anticipada de riesgos: casos de depresión en la Web

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

Detección anticipada de riesgos: casos de depresión en la Web

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

Topic Tagging [SS3 Live Demo]

Main Topic: **food** EDIT TEXT

Also: **health**

Visual Description

?

Select the **description level and topic** using the Level and Topic options below. Additionally, click on individual words to see their *gv* and *lv* values and frequency.

Level: Paragraphs Sentences Words

Japanese-style warm prawn and seaweed salad: The ultimate mineral-rich salad.

Wakame, or kelp, is a gift from the sea; the perfect healthy addition to a regular Japanese diet (just like tofu!). It is highly rich in vitamins and minerals such as magnesium, iodine, calcium, iron, vitamins A, C, E, D and K, to name just a few.

But eating wakame alone is like being forced to enjoy mashed potatoes with no salt: tasteless and highly dissatisfying. After discovering the following recipe, however, I was able to easily incorporate wakame into my daily diet, even after relocating back home to Ireland, where it has become relatively easy to find dried seaweed. Delicious, colorful and full of healthy ingredients, this salad is both easy to make and great for your health.

Topic:

[MIXED]

- FOOD (5.0cv)
- HEALTH (2.7cv)
- BEAUTY/FASHION (1.2cv)
- ART/PHOTOGRAPHY (0.2cv)
- BUSINESS/FINANCE (0.2cv)
- SCIENCE/TECH (0.1cv)
- MUSIC (0.0cv)
- SPORTS (0.0cv)

Detección anticipada de riesgos: casos de depresión en la Web

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.

Detección anticipada de riesgos: casos de depresión en la Web

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.