Modelos para la detección anticipada de riesgos en flujos de datos

Juan Martín Loyola
Marcelo Errecalde

Universidad Nacional de San Luis

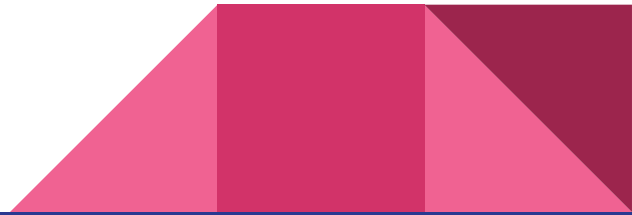
Temas a tratar

- Clasificación anticipada de texto
- Detección anticipada de riesgos
- Modelos para detección anticipada de riesgo

https://jmloyola.github.io/files/talks/2021_encuentro_posgrado.pdf

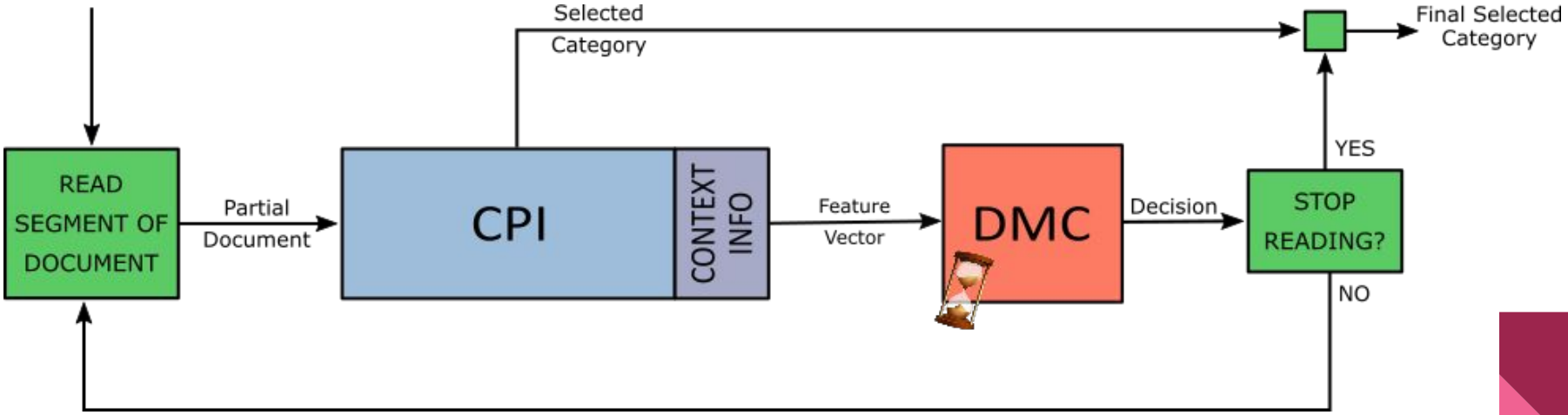
Clasificación anticipada de texto

- Desarrollo de modelos predictivos que determinen la categoría de un documento lo antes posible.
- Encontrar equilibrio entre:
  - precisión de la clasificación;
  - tiempo necesario para clasificar.
- Se puede conceptualizar en dos partes:
  - Clasificación con Información Parcial (CPI);
  - Decisión del Momento de Clasificación (DMC).



Clasificación anticipada de texto

- CPI → Clasificación con Información Parcial
- DMC → Decisión del Momento de Clasificación

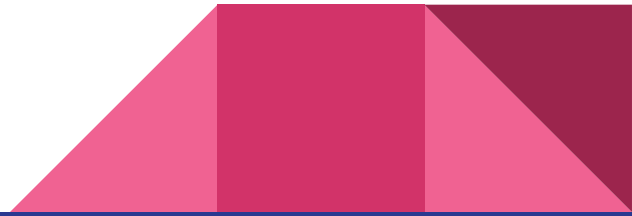


Detección anticipada de riesgo

- Caso especial de la clasificación anticipada de texto.
- Sólo nos preocupa predecir lo antes posible un subconjunto de las categorías (categoría de riesgo).
- Si la entrada parcial actual se clasifica como clase sin riesgo, el modelo sigue acumulando información en caso de que, en el futuro, el usuario comience a mostrar patrones de riesgo.
- Es fundamental recuperar a tantos usuarios en riesgo como sea posible ya que sus vidas podrían estar en peligro.
- Ejemplos: detección anticipada de depresión, detección anticipada de pedófilos.

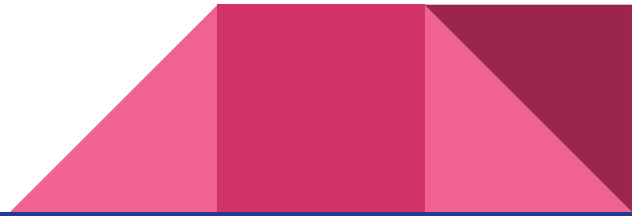
Medidas de desempeño

- Precisión, exhaustividad (recall), medida F1



Medidas de desempeño

- Precisión, exhaustividad (recall), medida F1
- ERDE
- F latency



ERDE (error de detección de riesgo temprano)

$$ERDE_o(d,k) = \begin{cases} c_{fp} & \text{if } d\text{=positive AND ground truth=negative (FP)} \\ c_{fn} & \text{if } d\text{=negative AND ground truth=positive (FN)} \\ lc_o(k) \cdot c_{tp} & \text{if } d\text{=positive AND ground truth=positive (TP)} \\ 0 & \text{if } d\text{=negative AND ground truth=negative (TN)} \end{cases}$$

$$lc_o(k) = 1 - \frac{1}{1 + e^{k-o}}$$

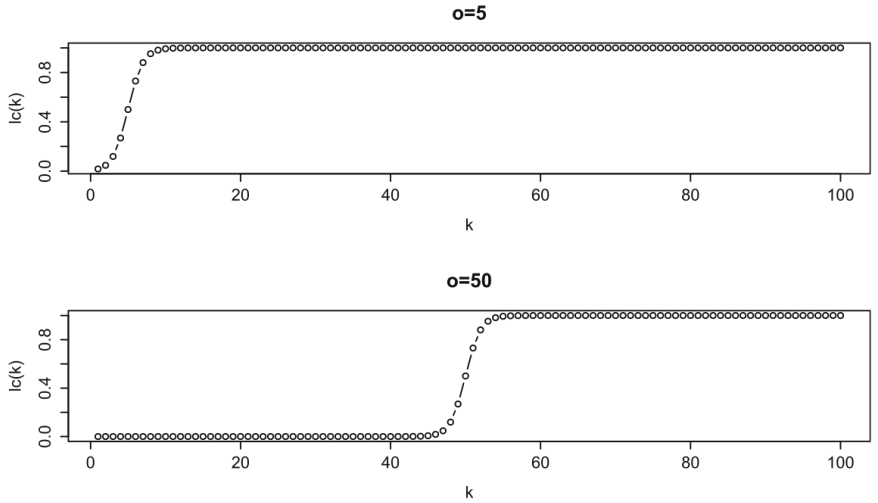


Fig. 1. Latency cost functions: $lc_5(k)$ and $lc_{50}(k)$

Losada et al. 2018

ERDE (error de detección de riesgo temprano)

$$ERDE_o(d,k) = \begin{cases} c_{fp} & \text{if } d\text{=positive AND ground truth=negative (FP)} \\ c_{fn} & \text{if } d\text{=negative AND ground truth=positive (FN)} \\ lc_o(k) \cdot c_{tp} & \text{if } d\text{=positive AND ground truth=positive (TP)} \\ 0 & \text{if } d\text{=negative AND ground truth=negative (TN)} \end{cases}$$

ERDE (error de detección de riesgo temprano)

$$ERDE_o(d,k) = \begin{cases} c_{fp} & \text{if } d\text{=positive AND ground truth=negative (FP)} \\ c_{fn} & \text{if } d\text{=negative AND ground truth=positive (FN)} \\ lc_o(k) \cdot c_{tp} & \text{if } d\text{=positive AND ground truth=positive (TP)} \\ 0 & \text{if } d\text{=negative AND ground truth=negative (TN)} \end{cases}$$

Umbral de penalización

Losada et al. 2018

ERDE (error de detección de riesgo temprano)

$$ERDE_o(d,k) = \begin{cases} c_{fp} & \text{if } d\text{=positive AND ground truth=negative (FP)} \\ c_{fn} & \text{if } d\text{=negative AND ground truth=positive (FN)} \\ lc_o(k) \cdot c_{tp} & \text{if } d\text{=positive AND ground truth=positive (TP)} \\ 0 & \text{if } d\text{=negative AND ground truth=negative (TN)} \end{cases}$$

Categoría predicha para el usuario

Losada et al. 2018

ERDE (error de detección de riesgo temprano)

$$ERDE_o(d,k) = \begin{cases} c_{fp} & \text{if } d\text{=positive AND ground truth=negative (FP)} \\ c_{fn} & \text{if } d\text{=negative AND ground truth=positive (FN)} \\ lc_o(k) \cdot c_{tp} & \text{if } d\text{=positive AND ground truth=positive (TP)} \\ 0 & \text{if } d\text{=negative AND ground truth=negative (TN)} \end{cases}$$

Cantidad de posts leídos del usuario

Losada et al. 2018

ERDE (error de detección de riesgo temprano)

$$ERDE_o(d,k) = \begin{cases} c_{fp} & \text{if } d\text{=positive AND ground truth=negative (FP)} \\ c_{fn} & \text{if } d\text{=negative AND ground truth=positive (FN)} \\ lc_o(k) \cdot c_{tp} & \text{if } d\text{=positive AND ground truth=positive (TP)} \\ 0 & \text{if } d\text{=negative AND ground truth=negative (TN)} \end{cases}$$

Costo falsos positivos

(#positivos) / (#positivos + #negativos)

ERDE (error de detección de riesgo temprano)

$$ERDE_o(d,k) = \begin{cases} c_{fp} & \text{if } d\text{=positive AND ground truth=negative (FP)} \\ c_{fn} & \text{if } d\text{=negative AND ground truth=positive (FN)} \\ lc_o(k) \cdot c_{tp} & \text{if } d\text{=positive AND ground truth=positive (TP)} \\ 0 & \text{if } d\text{=negative AND ground truth=negative (TN)} \end{cases}$$

Costo falsos negativos

Costo verdaderos positivos

1

ERDE (error de detección de riesgo temprano)

$$ERDE_o(d,k) = \begin{cases} c_{fp} & \text{if } d\text{=positive AND ground truth=negative (FP)} \\ c_{fn} & \text{if } d\text{=negative AND ground truth=positive (FN)} \\ lc_o(k) \cdot c_{tp} & \text{if } d\text{=positive AND ground truth=positive (TP)} \\ 0 & \text{if } d\text{=negative AND ground truth=negative (TN)} \end{cases}$$

Penalización por demora

ERDE (error de detección de riesgo temprano)

$$ERDE_o(d,k) = \begin{cases} c_{fp} & \text{if } d\text{=positive AND ground truth=negative (FP)} \\ c_{fn} & \text{if } d\text{=negative AND ground truth=positive (FN)} \\ lc_o(k) \cdot c_{tp} & \text{if } d\text{=positive AND ground truth=positive (TP)} \\ 0 & \text{if } d\text{=negative AND ground truth=negative (TN)} \end{cases}$$

$$lc_o(k) = 1 - \frac{1}{1 + e^{k-o}}$$

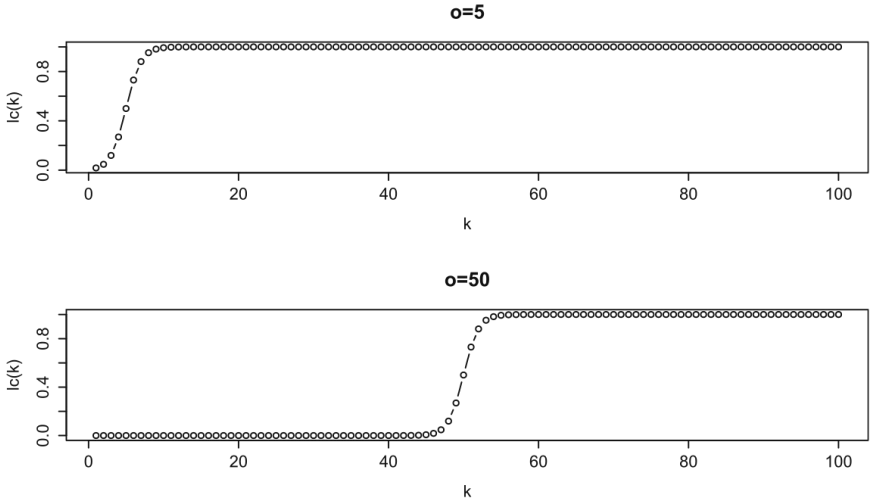


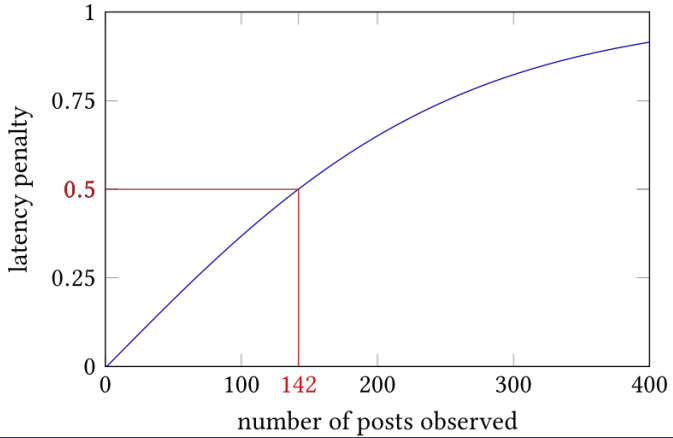
Fig. 1. Latency cost functions: $lc_5(k)$ and $lc_{50}(k)$

Losada et al. 2018

F latency

$$F_{latency} = F \cdot speed$$

$$speed = (1 - \text{median}\{penalty(k_u) : u \in U, d_u = g_u = 1\})$$



$$penalty(k_u) = -1 + \frac{2}{1 + \exp^{-p \cdot (k_u - 1)}}$$

Sadeque et al. 2018

F latency

$$F_{latency} = F \cdot speed$$

F latency

$$F_{latency} = F \cdot speed$$

Medida F1

F latency

$$F_{latency} = F \cdot speed$$

$$speed = (1 - \text{median}\{penalty(k_u) : u \in U, d_u = g_u = 1\})$$

F latency

$$F_{latency} = F \cdot speed$$

$$speed = (1 - \text{median}\{penalty(k_u) : u \in U, d_u = g_u = 1\})$$

Usuarios

F latency

$$F_{latency} = F \cdot speed$$

$$speed = (1 - \text{median}\{penalty(k_u) : u \in U, d_u = g_u = 1\})$$

Verdadera categoría del usuario u

Categoría predicha para el usuario u

F latency

$$F_{latency} = F \cdot speed$$

$$speed = (1 - \text{median}\{penalty(k_u) : u \in U, d_u = g_u = 1\})$$

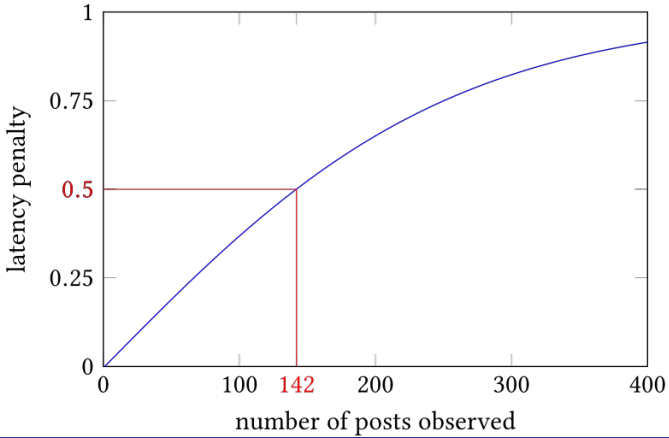
Cantidad de posts leídos del usuario u

Sadeque et al. 2018

F latency

$$F_{latency} = F \cdot speed$$

$$speed = (1 - \text{median}\{penalty(k_u) : u \in U, d_u = g_u = 1\})$$



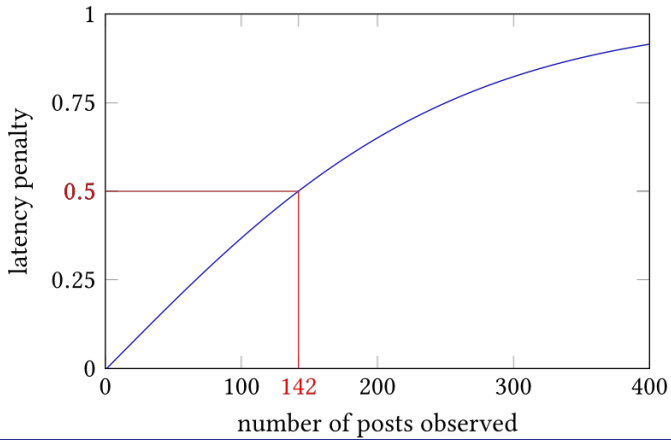
$$penalty(k_u) = -1 + \frac{2}{1 + \exp^{-p \cdot (k_u - 1)}}$$

Sadeque et al. 2018

F latency

$$F_{latency} = F \cdot speed$$

$$speed = (1 - \text{median}\{penalty(k_u) : u \in U, d_u = g_u = 1\})$$



El valor de p se elige de forma tal que la penalidad sea 0,5 para la mediana de cantidad de posts

$$penalty(k_u) = -1 + \frac{2}{1 + \exp^{-p \cdot (k_u - 1)}}$$

Sadeque et al. 2018

Modelos para detección anticipada de riesgo

- EarlyModel
- SS3
- EARLIEST

UNSL at eRisk 2021: A Comparison of Three Early Alert Policies for Early Risk Detection

Juan Martín Loyola[1,3], Sergio Burdisso[1,2], Horacio Thompson[1,2], Leticia Cagnina[1,2] and Marcelo Errecalde[1]

[1]*Universidad Nacional de San Luis (UNSL), Ejército de Los Andes 950, San Luis, C.P. 5700, Argentina*
[2]*Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)*
[3]*Instituto de Matemática Aplicada San Luis (IMASL), CONICET-UNSL, Av. Italia 1556, San Luis, C.P. 5700, Argentina*

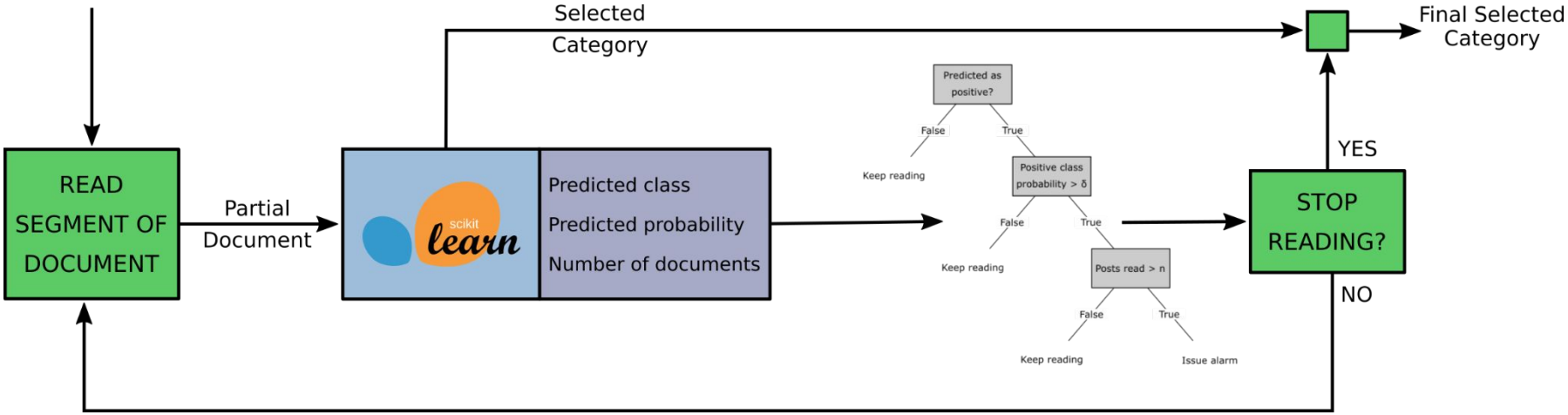
https://github.com/jmloyola/unsl_erisk_2021

Modelos para detección anticipada de riesgo

Para cada modelo se puede identificar:

- Representación de la entrada
- Modelo de clasificación con información parcial (CPI)
- Política de alerta temprana (DMC)

EarlyModel


READ SEGMENT OF DOCUMENT
Partial Document
scikit learn
Predicted class
Predicted probability
Number of documents
Selected Category
Predicted as positive?
False
True
Keep reading
Positive class probability > δ
False
True
Keep reading
Posts read > n
False
True
Keep reading
Issue alarm
STOP READING?
YES
NO
Final Selected Category

EarlyModel

Representación de la entrada:

- Bolsa de palabras
- Linguistic Inquiry and Word Count (LIWC)
- Latent Dirichlet Allocation (LDA)
- Latent Semantic Analysis (LSA)
- doc2vec

EarlyModel

Modelo de clasificación con información parcial:

- Árboles de decisión
- K-Vecinos más cercanos
- Máquinas de vectores de soporte (SVM)
- Regresión logística
- Perceptrón multicapa (MLP)
- Random forests

- LSTM

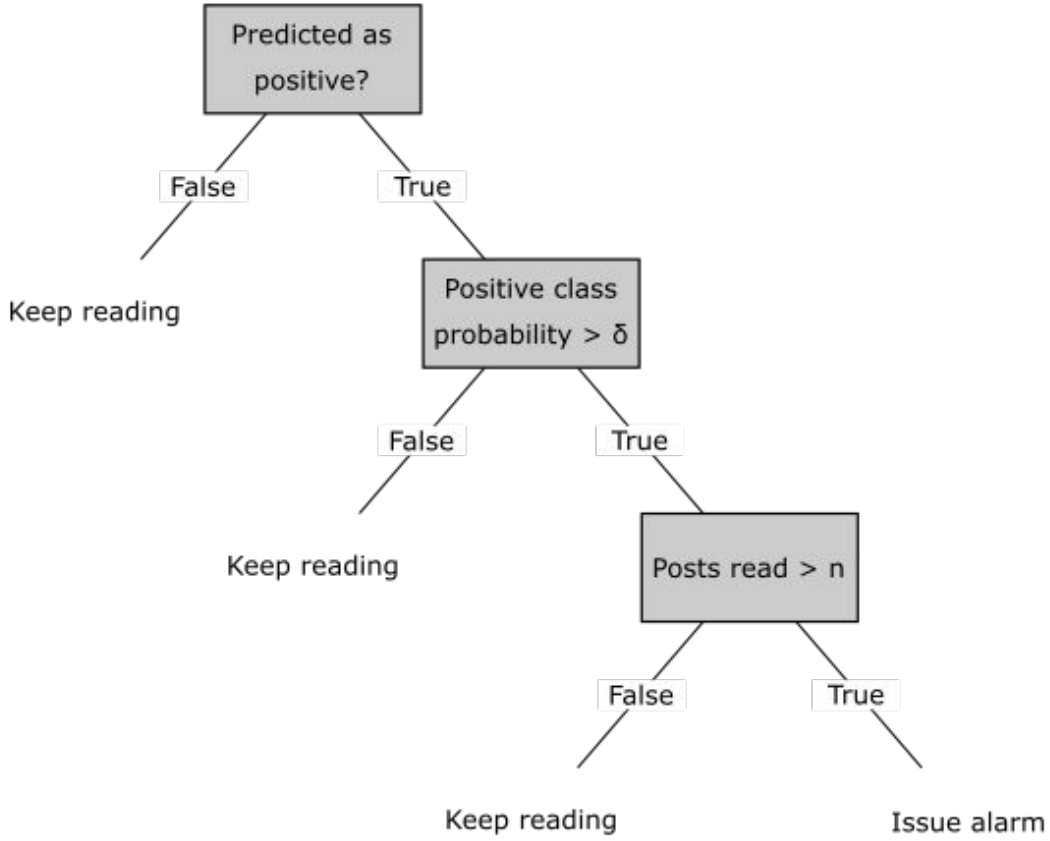
PyTorch

- BERT

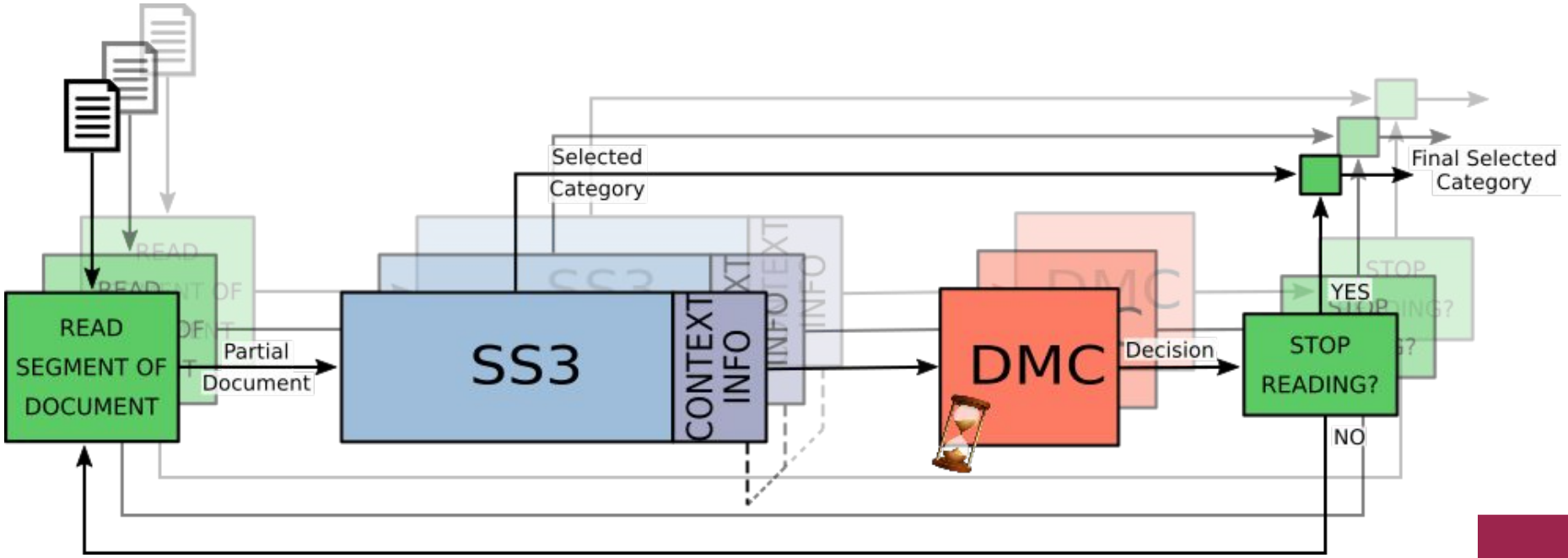
Transformers

EarlyModel

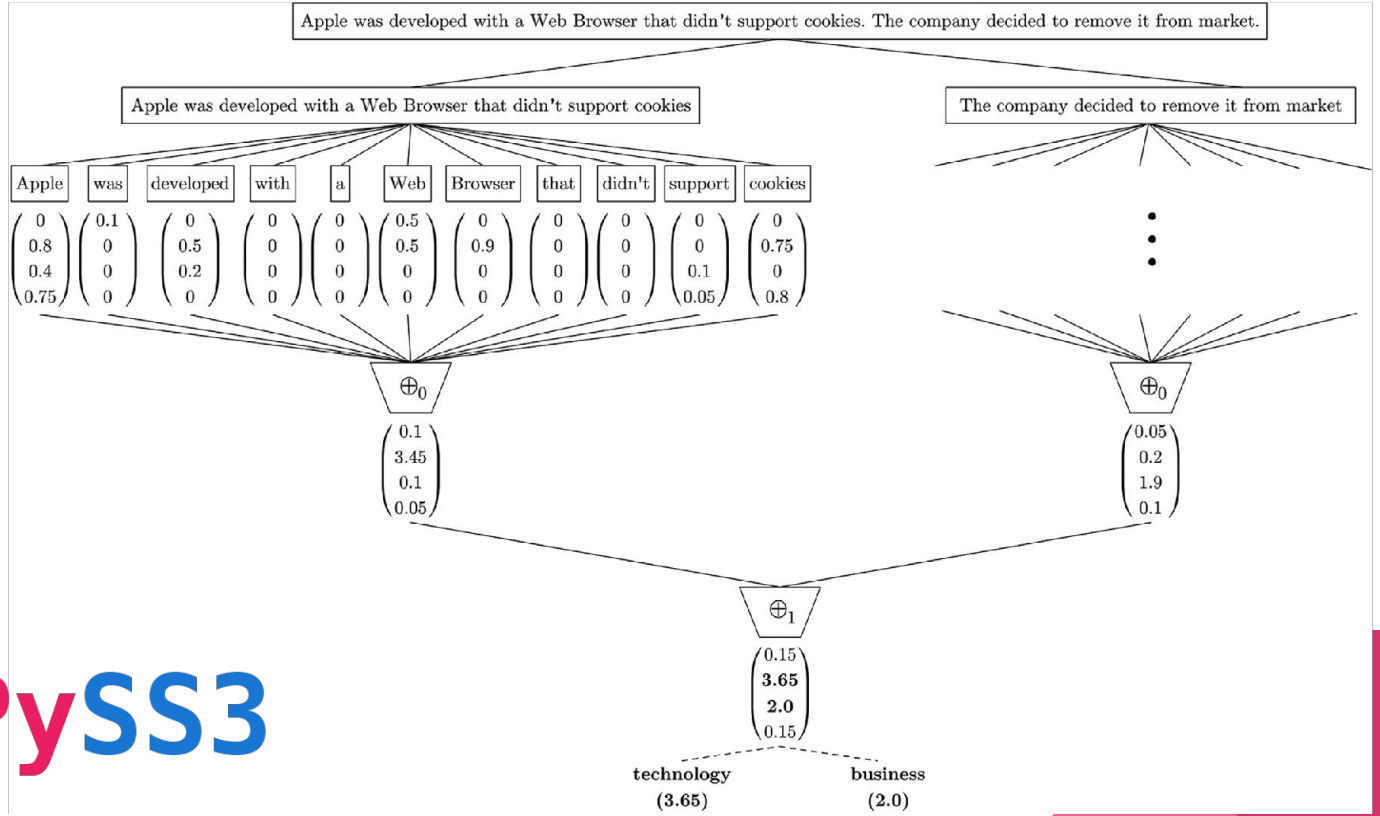
Política de alerta temprana:



SS3


READ
SEGMENT OF
DOCUMENT
Partial
Document
SS3
CONTEXT
INFO
DMC
Decision
STOP
READING?
YES
NO
Selected
Category
Final Selected
Category

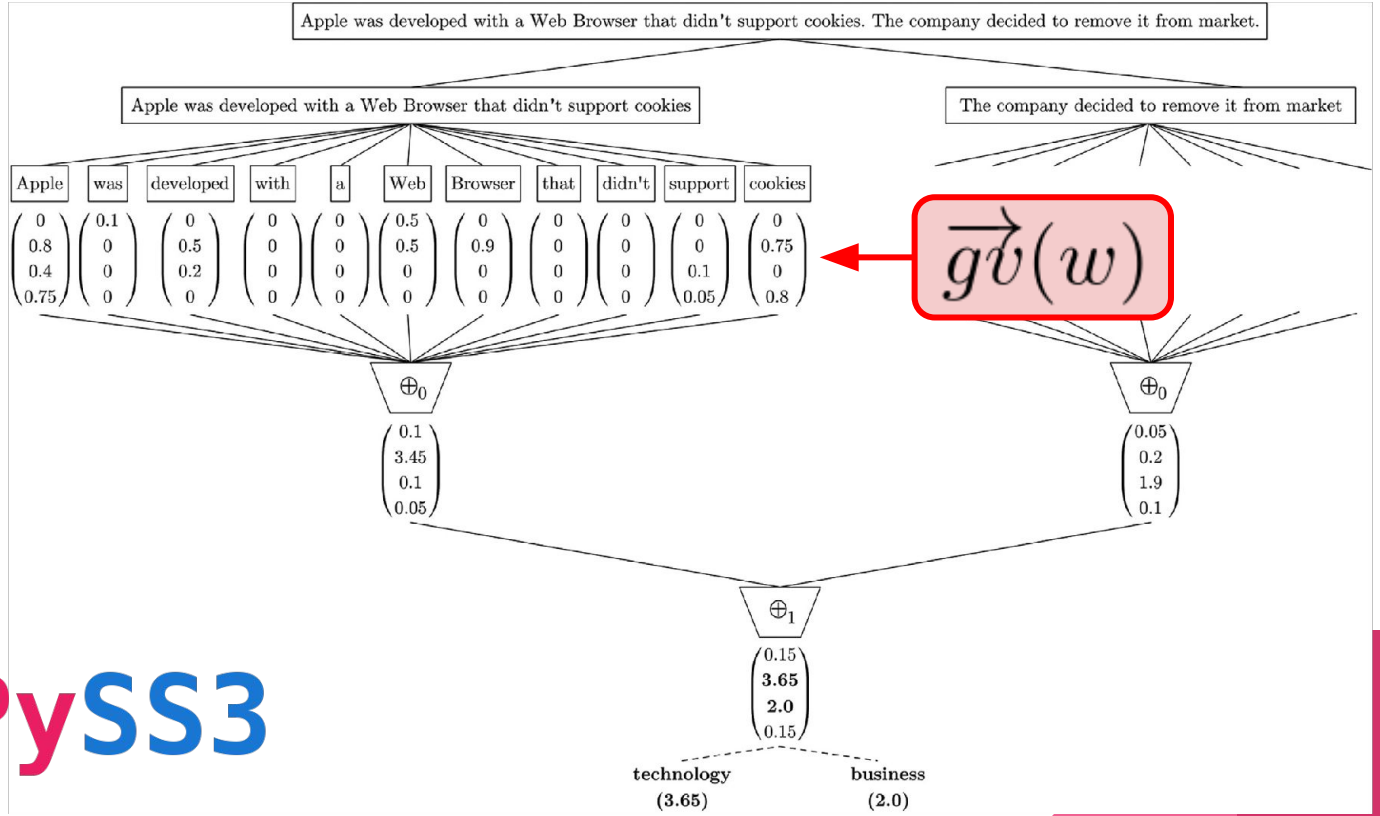
SS3



PySS3

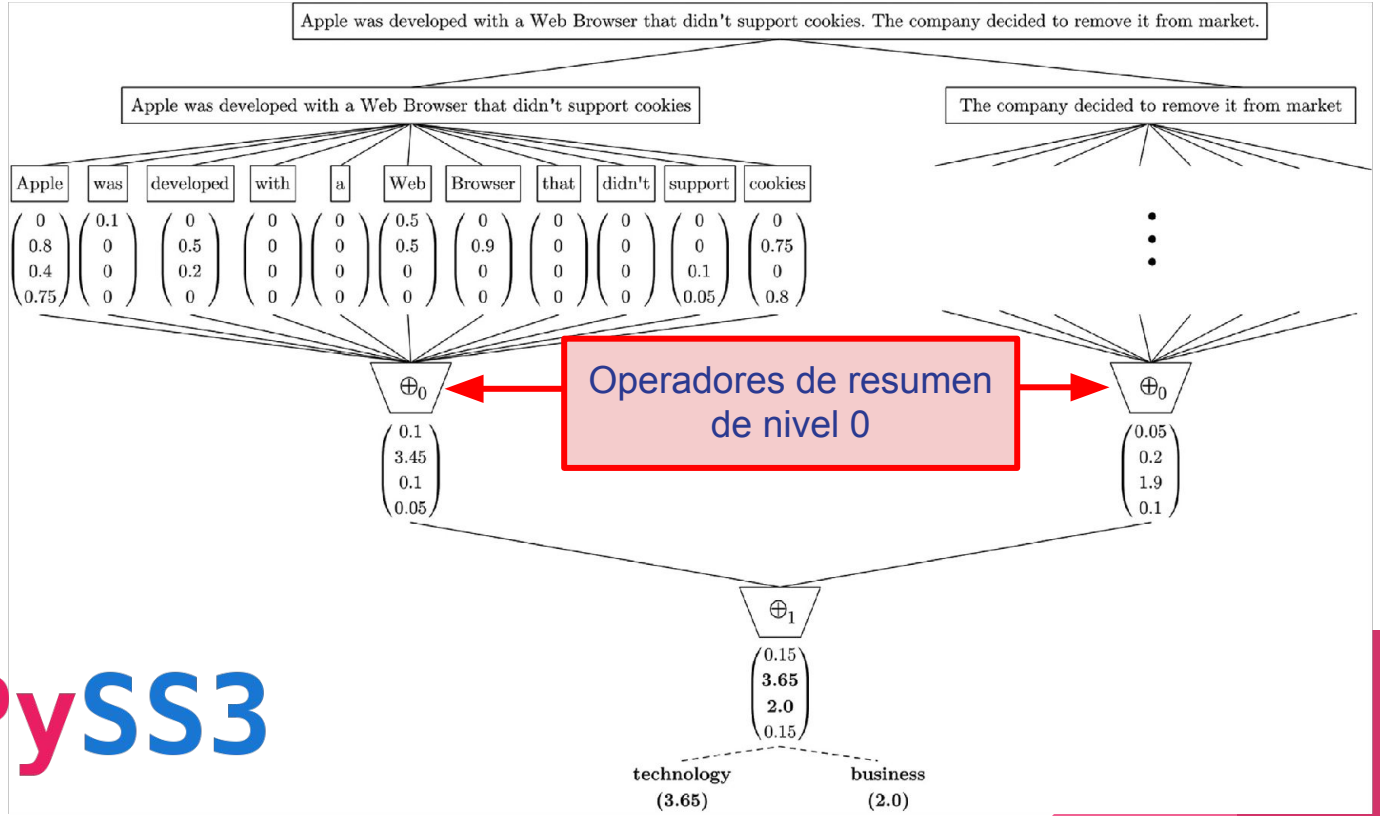
Burdisso et al. 2019

SS3


Apple was developed with a Web Browser that didn't support cookies. The company decided to remove it from market.
Apple was developed with a Web Browser that didn't support cookies
The company decided to remove it from market
Apple
was
developed
with
a
Web
Browser
that
didn't
support
cookies
$\vec{gv}(w)$
$\oplus_0$
$\oplus_0$
$\oplus_1$
technology
(3.65)
business
(2.0)

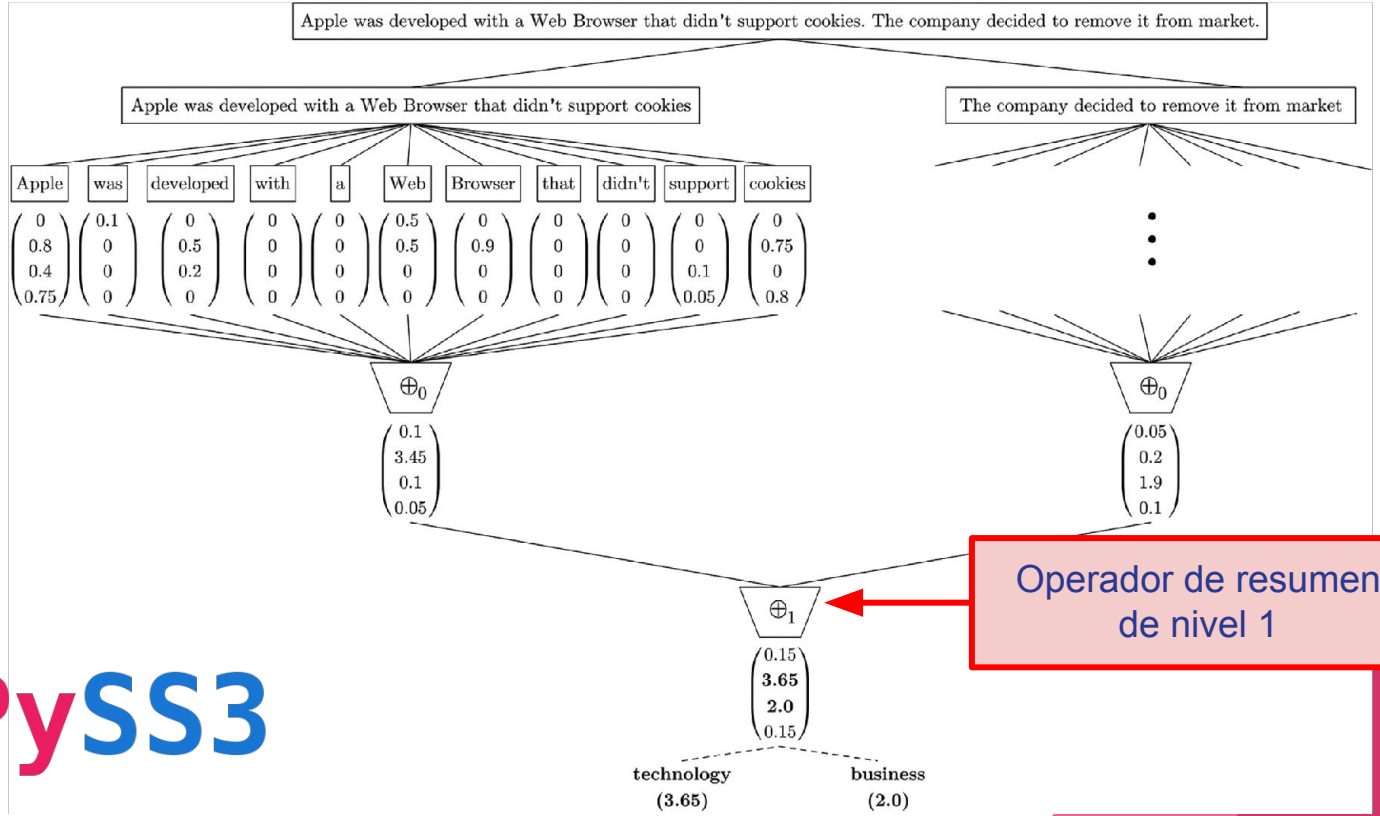
PySS3

SS3


Apple was developed with a Web Browser that didn't support cookies. The company decided to remove it from market.
Apple was developed with a Web Browser that didn't support cookies
The company decided to remove it from market
Apple
was
developed
with
a
Web
Browser
that
didn't
support
cookies
Operadores de resumen de nivel 0
technology
(3.65)
business
(2.0)

PySS3

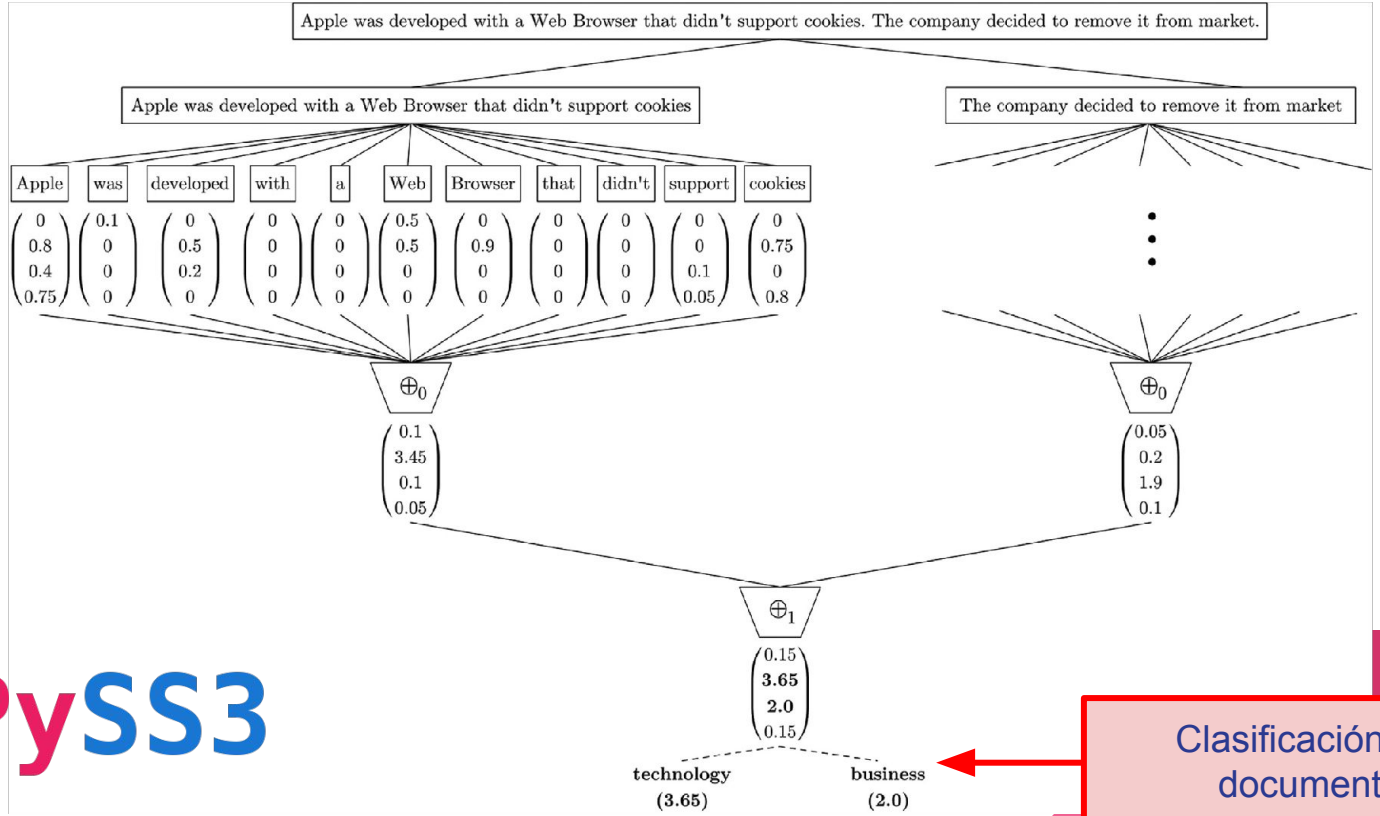
SS3



Operador de resumen de nivel 1

PySS3

SS3



PySS3

Clasificación del documento

SS3

Política de alerta temprana:

$$decision_u = \begin{cases} 1, & \text{if } score_u > \text{median}(scores) + \gamma \cdot \text{MAD}(scores) \\ 0, & \text{otherwise.} \end{cases}$$

SS3

Política de alerta temprana:

$$decision_u = \begin{cases} 1, & \text{if } score_u > \text{median}(scores) + \gamma \cdot \text{MAD}(scores) \\ 0, & \text{otherwise.} \end{cases}$$

Puntaje clase riesgo
-
Puntaje clase no riesgo

SS3

Política de alerta temprana:

$$decision_u = \begin{cases} 1, & \text{if } score_u > \text{median}(scores) + \gamma \cdot \text{MAD}(scores) \\ 0, & \text{otherwise.} \end{cases}$$

$$scores = \{score_u | u \in \text{Users}\}$$

SS3

Política de alerta temprana:

$$decision_u = \begin{cases} 1, & \text{if } score_u > \text{median}(scores) + \gamma \cdot \text{MAD}(scores) \\ 0, & \text{otherwise.} \end{cases}$$

Mediana de la desviación absoluta

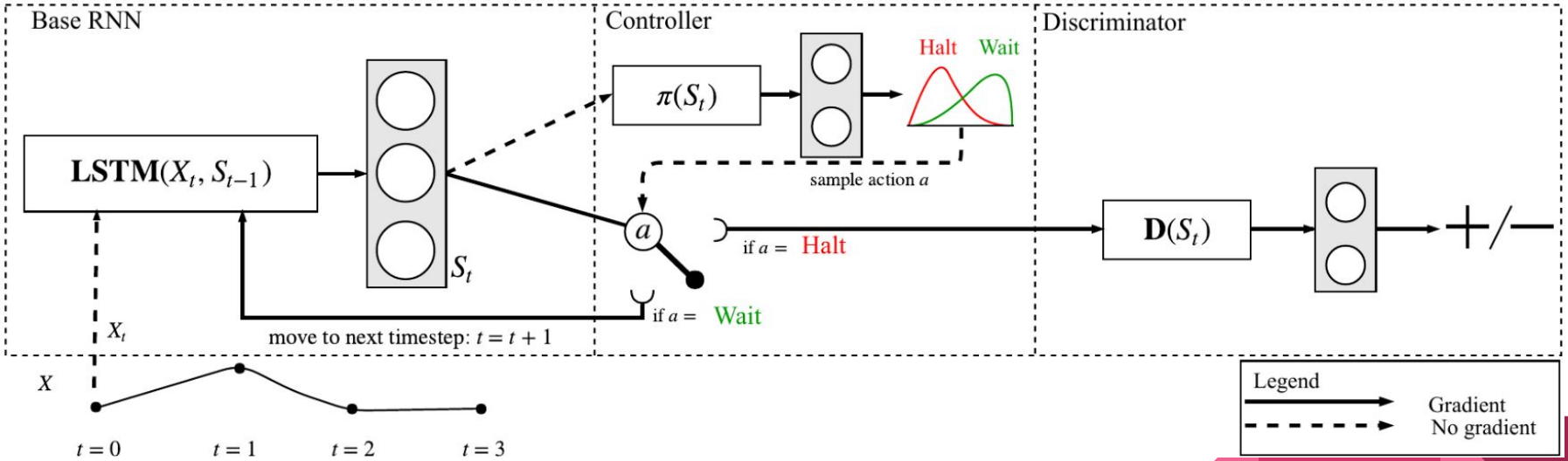
SS3

Política de alerta temprana:

$$decision_u = \begin{cases} 1, & \text{if } score_u > \text{median}(scores) + \gamma \cdot \text{MAD}(scores) \\ 0, & \text{otherwise.} \end{cases}$$

Hiper-parámetro de la política de alerta temprana

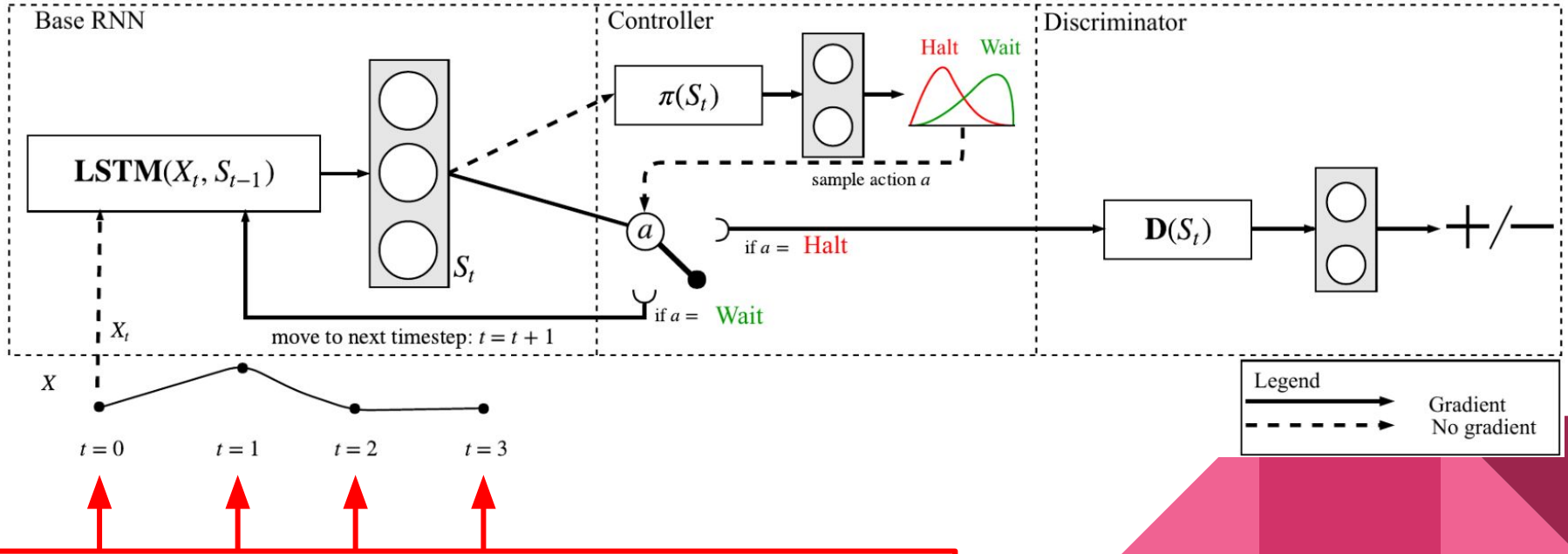
EARLIEST



Hartvigsen et al. 2019

EARLIEST

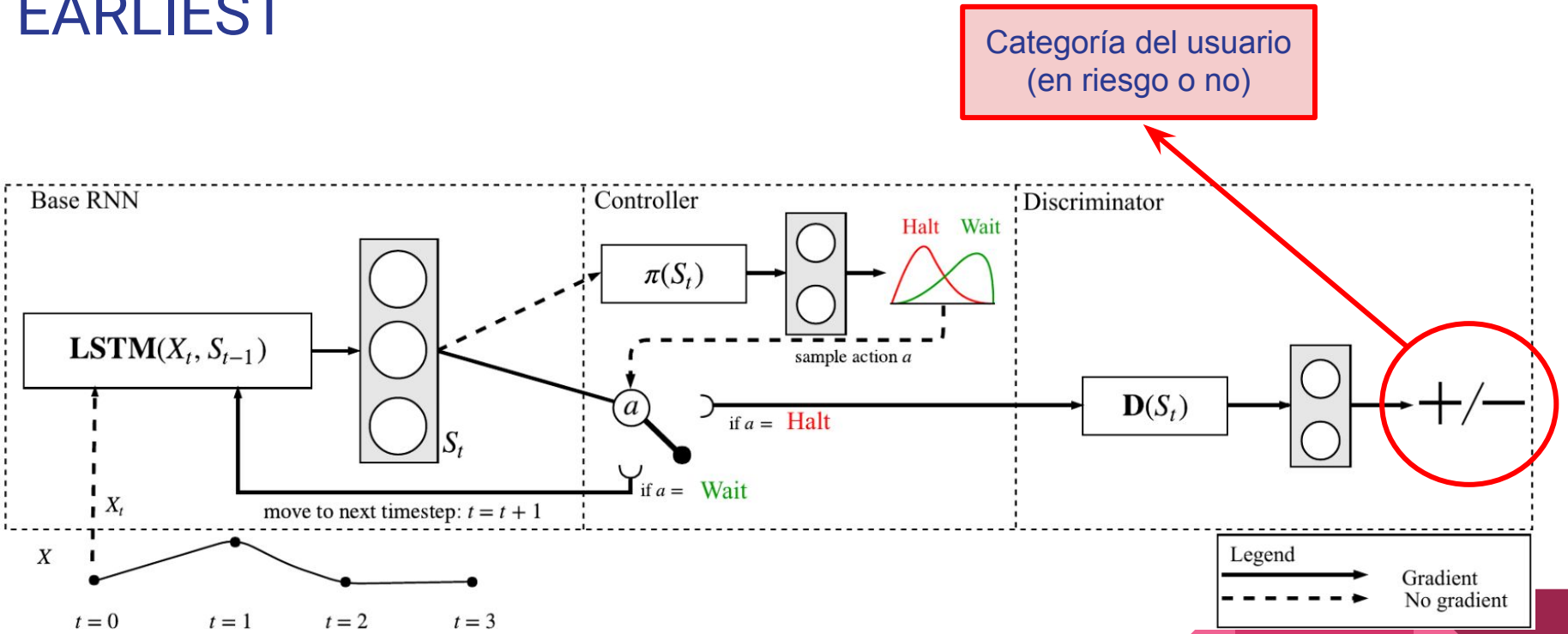
Early and Adaptive Recurrent Label ESTimator



Secuencia de documentos de entrada para un usuario (doc2vec)

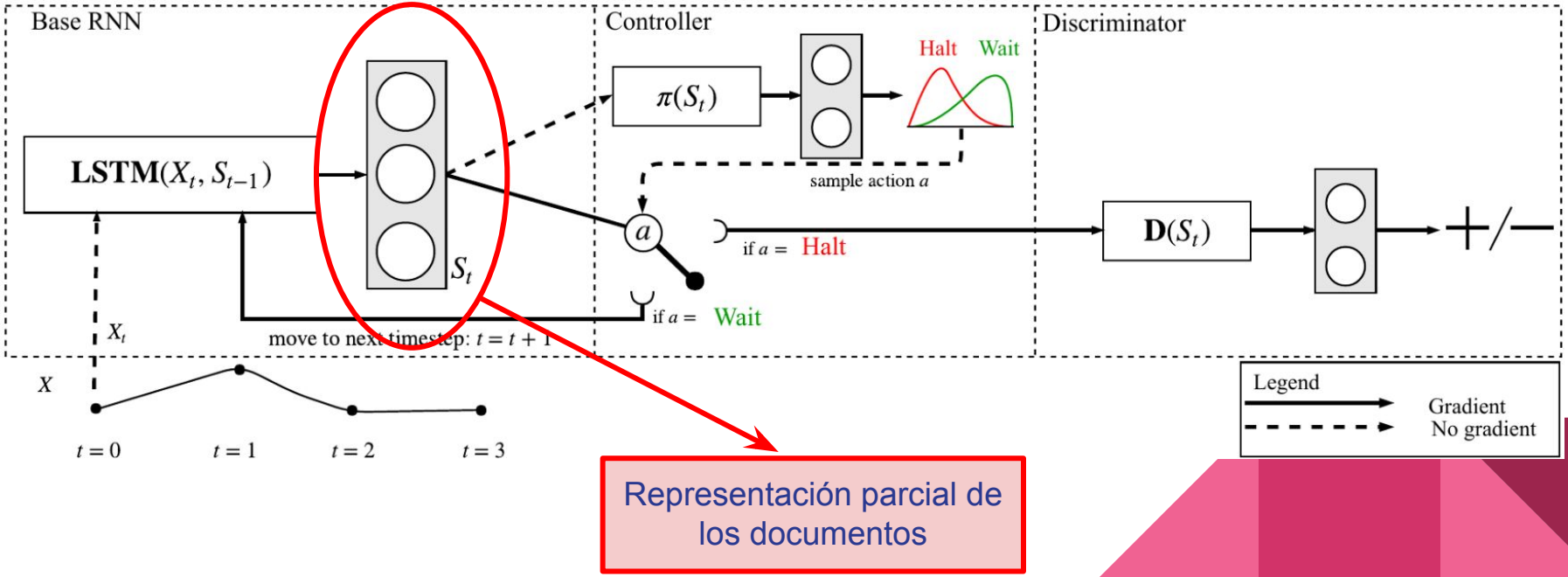
Hartvigsen et al. 2019

EARLIEST



Hartvigsen et al. 2019

EARLIEST

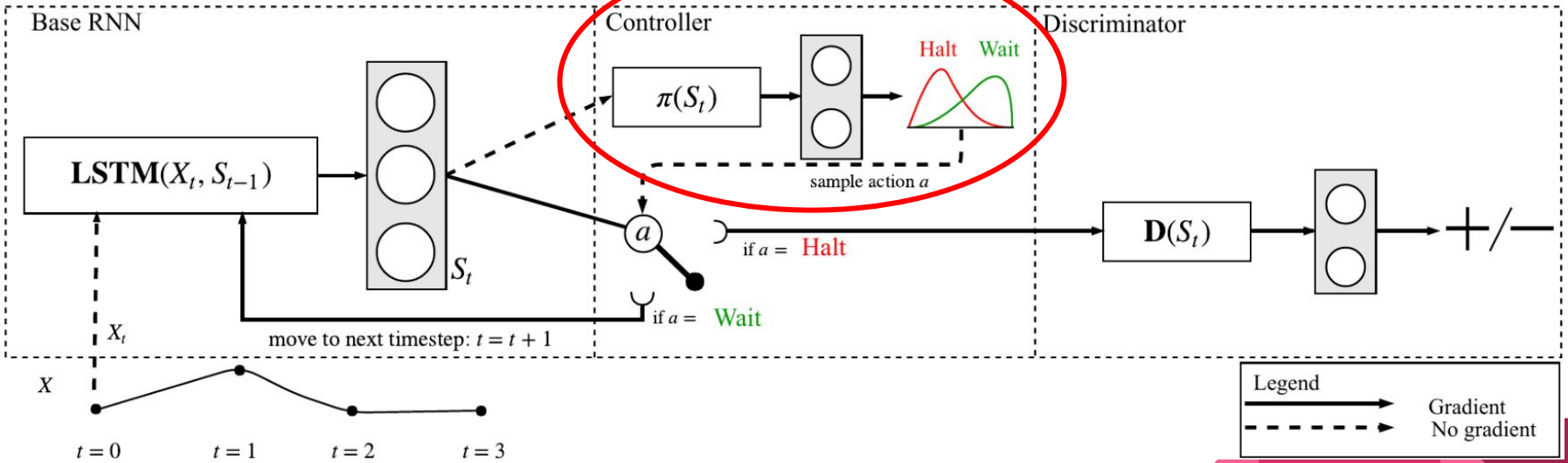


Representación parcial de los documentos

Hartvigsen et al. 2019

EARLIEST

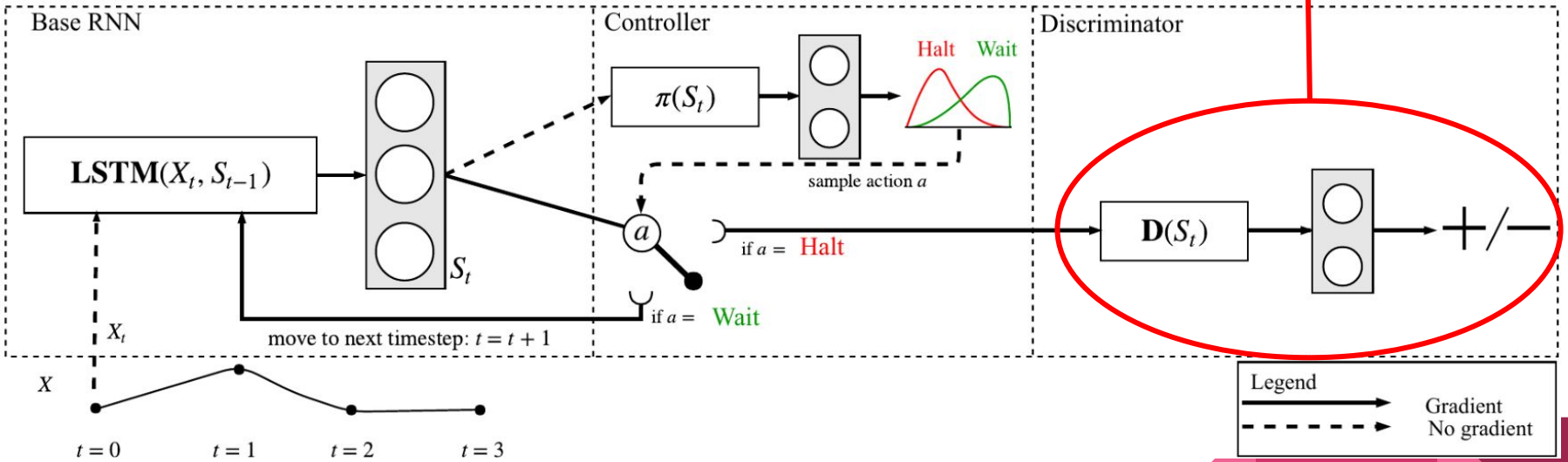
Parte encargada de decidir cuándo dejar de procesar la entrada



Hartvigsen et al. 2019

EARLIEST

Parte encargada de clasificar el documento parcial



Hartvigsen et al. 2019

EARLIEST

Para controlar la velocidad de la clasificación, el modelo cuenta con el hyper-parámetro λ que penaliza al modelo por la tardanza en la clasificación.

eRisk 2021

- Laboratorio parte de CLEF 2021.
- Predicción temprana de riesgos en Internet.
- Explora la metodología de evaluación, las métricas de efectividad y las aplicaciones prácticas de la detección temprana de riesgos en Internet.
- Tareas:
  - Tarea 1 - Detección temprana de signos de juego patológico.
  - Tarea 2 - Detección temprana de signos de autolesión.

Modelos para T1

- UNSL#0 (EarlyModel):
  - Representación ⟶ bolsa de palabras (unigrama de palabras con tf-idf)
  - Modelo ⟶ regresión logística
  - Política de decisión ⟶ umbral = 0,7 y número mínimo de posts = 10
- UNSL#1 (EarlyModel):
  - Representación ⟶ doc2vec
  - Modelo ⟶ regresión logística
  - Política de decisión ⟶ umbral = 0,85 y número mínimo de posts = 3
- UNSL#2 (EarlyModel):
  - Representación ⟶ bolsa de palabras (4-gramas de caracteres con tf-idf)
  - Modelo ⟶ SVM
  - Política de decisión ⟶ umbral = 0,75 y número mínimo de posts = 10

Modelos para T1

- UNSL#3 (EARLIEST):
  - Representación ⟶ doc2vec
  - Modelo ⟶ LSTM
  - Política de decisión ⟶ λ = 0,000001
- UNSL#4 (EARLIEST):
  - Representación ⟶ doc2vec
  - Modelo ⟶ LSTM
  - Política de decisión ⟶ λ = 0,00001

Resultados de tarea “Detección temprana de signos de juego patológico”

team name	run id	P	R	$F1$	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	$speed$	$latency\text{-}weighted\ F1$
UNSL (EarlyModel)	0	.326	.957	.487	.079	.023	11	.961	.468
UNSL (EarlyModel)	1	.137	.982	.241	.060	.035	4	.988	.238
UNSL (EarlyModel)	2	**.586**	.939	**.721**	.073	**.020**	11	.961	**.693**
UNSL (EARLIEST)	3	.084	.963	.155	.066	.060	1	**1**	.155
UNSL (EARLIEST)	4	.086	.933	.157	.067	.060	1	**1**	.157
RELAI	0	.138	.988	.243	**.048**	.036	1	**1**	.243
BLUE	1	.157	.988	.271	.054	.036	2	.996	.270
UPV-Symanto	0	.042	.415	.077	.088	.087	1	**1**	.077
CeDRI	0	.076	**1**	.142	.079	.060	2	.996	.141
EFE	2	.233	.750	.356	.082	.033	11	.961	.342

Table 2. Decision-based evaluation

Modelos para T2

- UNSL#0 (EarlyModel):
  - Representación ⟶ doc2vec
  - Modelo ⟶ MLP
  - Política de decisión ⟶ umbral = 0,7 y número mínimo de posts = 10
- UNSL#1 (EARLIEST):
  - Representación ⟶ doc2vec
  - Modelo ⟶ LSTM
  - Política de decisión ⟶ λ = 0,000001
- UNSL#2 (EARLIEST):
  - Representación ⟶ doc2vec
  - Modelo ⟶ LSTM
  - Política de decisión ⟶ λ = 0,00001

Modelos para T2

- UNSL#3 (SS3):
  - Representación ⟶ texto
  - Modelo ⟶ SS3
  - Política de decisión ⟶ γ = 2
- UNSL#4 (SS3):
  - Representación ⟶ texto
  - Modelo ⟶ SS3
  - Política de decisión ⟶ γ = 2,5

Resultados de tarea “Detección temprana de signos de autolesión”

team name	run id	P	R	$F1$	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	$speed$	$latency\text{-}weighted\ F1$
UNSL (EarlyModel)	0	.336	.914	.491	.125	**.034**	11	.961	.472
UNSL (EARLIEST)	1	.11	.987	.198	.093	.092	**1**	**1.0**	.198
UNSL (EARLIEST)	2	.129	.934	.226	.098	.085	**1**	**1.0**	.226
UNSL (SS3)	3	.464	.803	.588	.064	.038	3	.992	.583
UNSL (SS3)	4	.532	.763	**.627**	.064	.038	3	.992	**.622**
NLP-UNED	4	.453	.816	.582	.088	.04	9	.969	.564
Birmingham	0	.584	.526	.554	.068	.054	2	.996	.551
Birmingham	2	**.757**	.349	.477	.085	.07	4	.988	.472
EFE	2	.366	.796	.501	.12	.043	12	.957	.48
BLUE	2	.454	.849	.592	.079	.037	7	.977	.578
UPV-Symanto	1	.276	.638	.385	**.059**	.056	1	1.0	.385

Table 5. Decision-based evaluation

Gracias por su atención. ¿Preguntas?

jmloyola@unsl.edu.ar

Referencias

- Loyola, J. M., Errecalde, M. L., Escalante, H. J., & y Gomez, M. M. (2017, October). Learning when to classify for early text classification. In Argentine Congress of Computer Science (pp. 24-34). Springer, Cham.
- Losada, D. E., Crestani, F., & Parapar, J. (2018, September). Overview of eRisk: early risk prediction on the internet. In International conference of the cross-language evaluation forum for european languages (pp. 343-361). Springer, Cham.
- Sadeque, F., Xu, D., & Bethard, S. (2018, February). Measuring the latency of depression detection in social media. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (pp. 495-503).
- Loyola, J. M., Burdisso, S. G., Thompson, H., Cagnina, L. & Errecalde, M. (2021, September). UNSL at eRisk 2021: A comparison of three early alert policies for early risk detection. In Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucarest, Romania, September 21-24, 2021.
- Burdisso, S. G., Errecalde, M., & Montes-y-Gómez, M. (2019). A text classification framework for simple and effective early depression detection over social media streams. Expert Systems with Applications, 133, 182-197.
- Hartvigsen, T., Sen, C., Kong, X., & Rundensteiner, E. (2019, July). Adaptive-halting policy network for early classification. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 101-110).

eRisk corpus (T1)

- Basado en publicaciones y comentarios en Reddit (https://www.reddit.com/).
- No se suministró corpus para entrenar.

Corpus	#users			#posts	#posts per user			#words per post		
	Total	Pos	Neg		Med	Min	Max	Med	Min	Max
T1_test	2,348	164	2184	1,130,792	244	10	2,001	12	0	10,175
T1_train	726	176	550	71,187	54	31	740	20	1	4,516
T1_valid	726	176	550	74,507	55	31	1,234	19	1	7,479

eRisk corpus (T2)

- Basado en publicaciones y comentarios en Reddit (https://www.reddit.com/).
- Se suministró un corpus de entrenamiento y validación.

Corpus	#users			#posts	#posts per user			#words per post		
	Total	Pos	Neg		Med	Min	Max	Med	Min	Max
T2_test	1,448	152	1296	746,098	275.5	10	1,999	12	0	18,064
T2_train	340	41	299	170,698	282.0	8	1,992	10	1	6,700
T2_valid	423	104	319	103,837	95.0	9	1,990	7	1	2,663
redd_train	1,051	494	557	118,452	61.0	31	1,466	18	1	5,971
redd_valid	1,051	494	557	119,651	59.0	31	1,781	18	1	4,382
comb_train	1,391	535	856	289,150	73.0	8	1,992	13	1	6,700
comb_valid	1,474	598	876	223,488	63.0	9	1,990	11	1	4,382
ilab_train	26,256	10319	15937	259,297	5.0	1	1,825	19	1	11,933

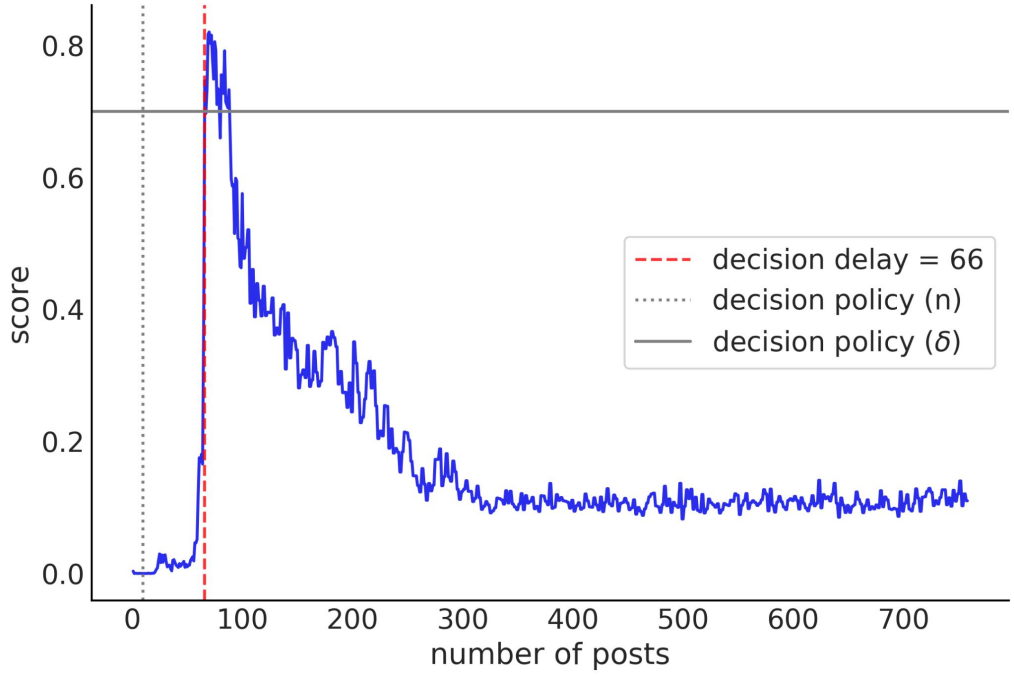
Procedimiento de generación de corpus

- Basado en publicaciones y comentarios en Reddit (https://www.reddit.com/).
- Los casos positivos fueron obtenidos de subreddits particulares
  - T1: https://www.reddit.com/r/problemgambling/
  - T2: https://www.reddit.com/r/selfharm/
- Los casos negativos fueron obtenidos de subreddits generales, sports, jokes, gaming, politics, news, y LifeProTips.
- Se descartaron todos los usuarios con menos de 31 publicaciones o comentarios, o con un promedio de palabras por publicacion menor a 15.

Pre-procesamiento de entrada

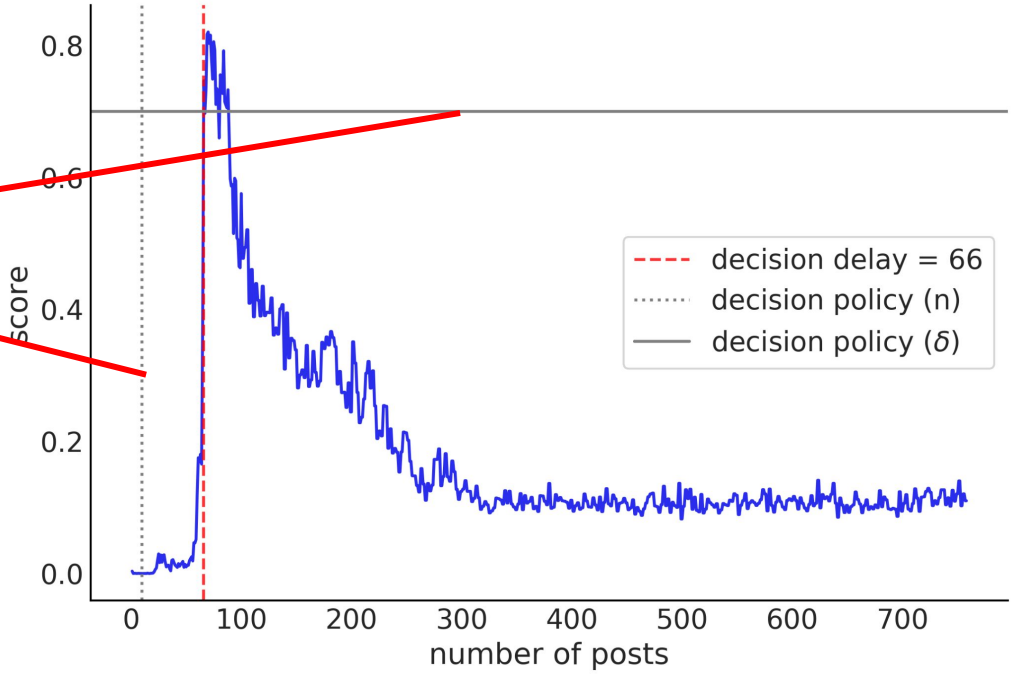
1. Convertir a minúscula.
2. Convertir códigos HTML y Unicode en su símbolos.
3. Reemplazar enlaces a la web con un token.
4. Reemplazar enlaces internos a reddit por sitio al que dirigen.
5. Eliminar símbolos que no sean letras y números.
6. Reemplazar números con un token.
7. Eliminar espacios en blanco consecutivos, nuevas líneas y tabs.

Visualización puntaje y política de decisión (EarlyModel)



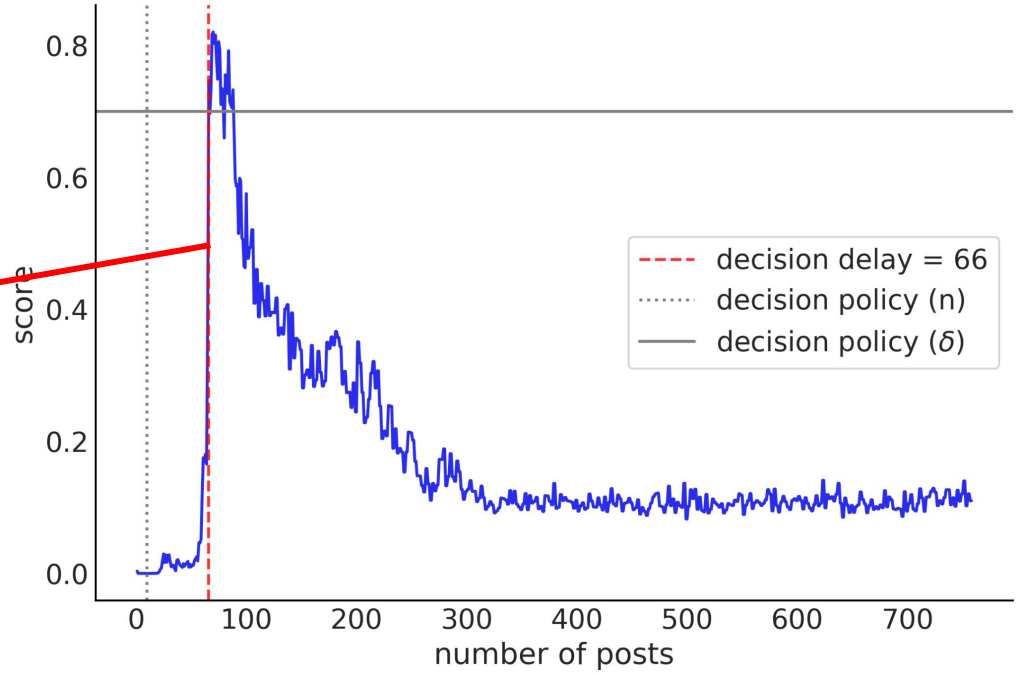
Visualización puntaje y política de decisión (EarlyModel)

Política de decisión estática

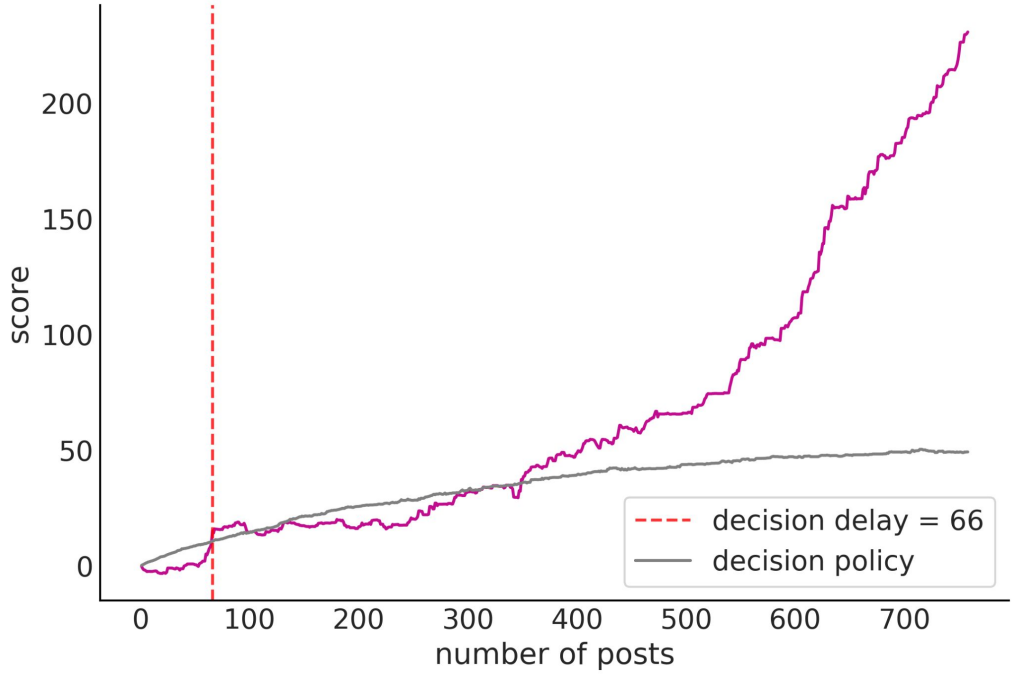


Visualización puntaje y política de decisión (EarlyModel)

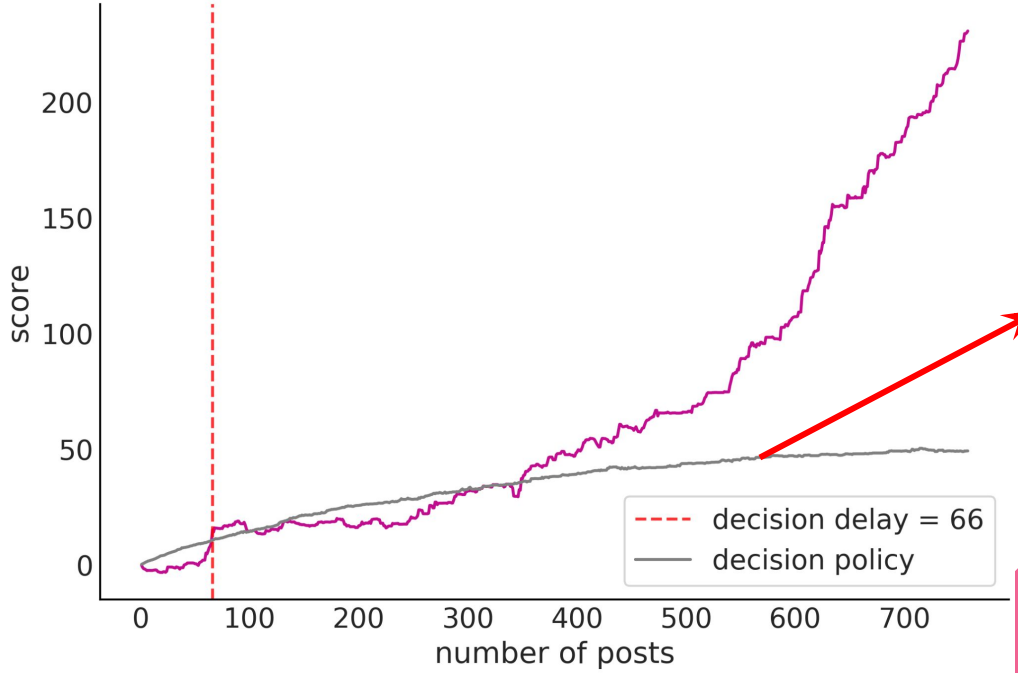
Momento en el que se clasifica como positivo



Visualización puntaje y política de decisión (SS3)



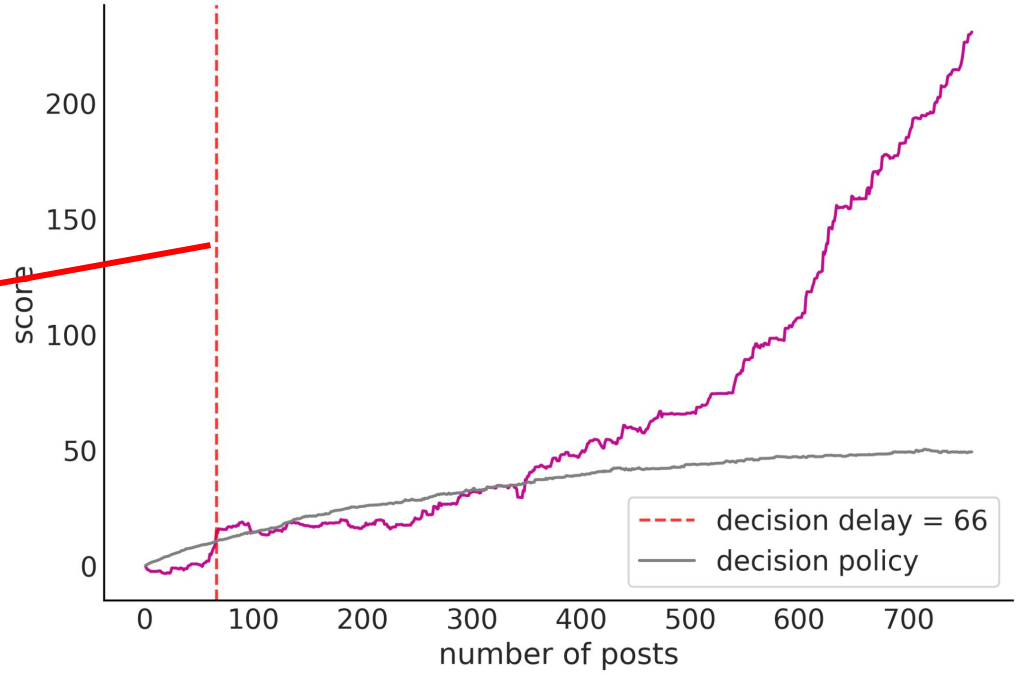
Visualización puntaje y política de decisión (SS3)



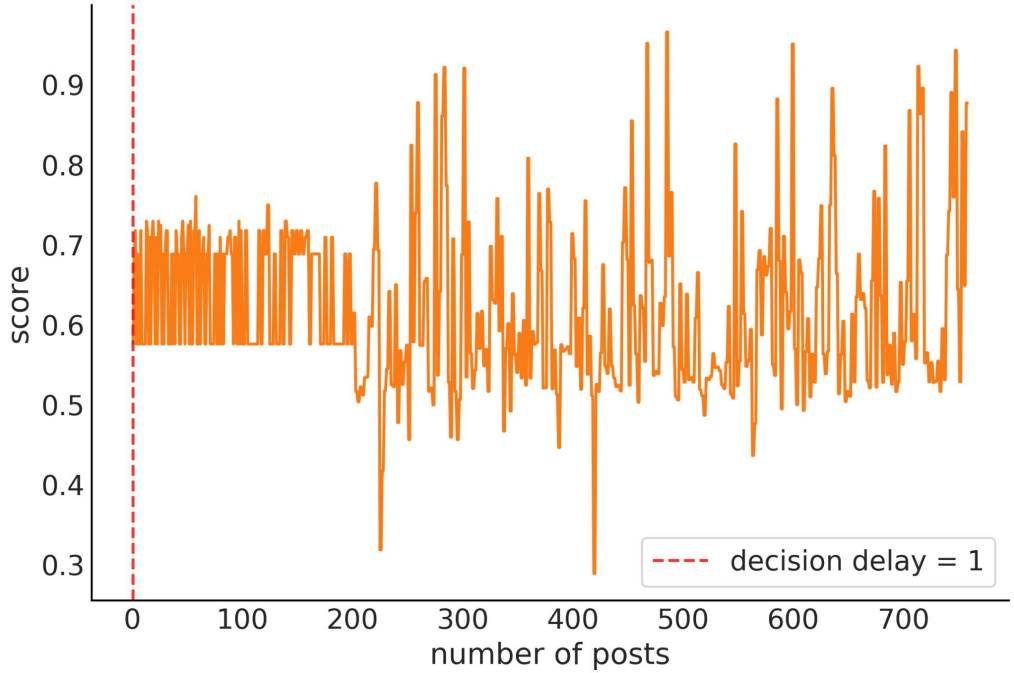
Política de decisión dinámica (depende de los scores de todos los usuarios)

Visualización puntaje y política de decisión (SS3)

Momento en el que se clasifica como positivo



Visualización puntaje y política de decisión (EARLIEST)



Visualización puntaje y política de decisión (EARLIEST)

Momento en el que se clasifica como positivo