Extracción de información y clasificación automática de textos

Helena Gómez Adorno
Investigadora Titular A
Departamento de Ingeniería de Sistemas Computacionales y Automatización

Trayectoria Académica

Licenciatura, 2001- 2005
Universidad Nacional de Asunción, Paraguay.

Maestría, 2011-2013
Benemérita Universidad Autónoma de Puebla, México.

Doctorado, 2014-2018
Instituto Politécnico Nacional, México.

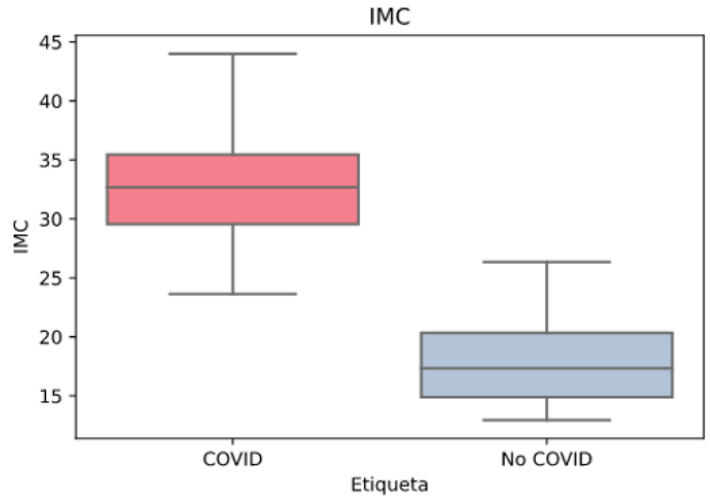
EXTRACCIÓN DE INFORMACIÓN

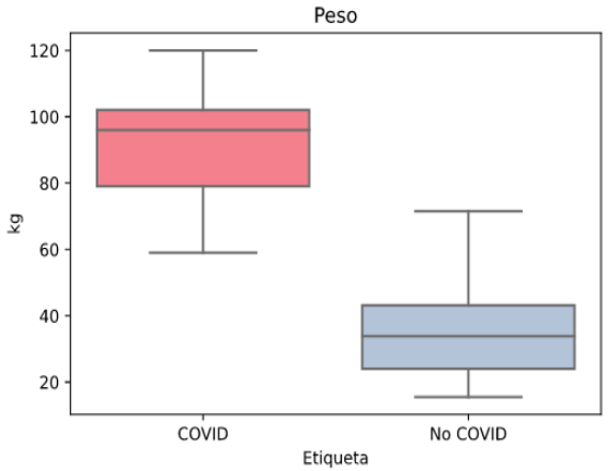
Identificación de entidades médicas

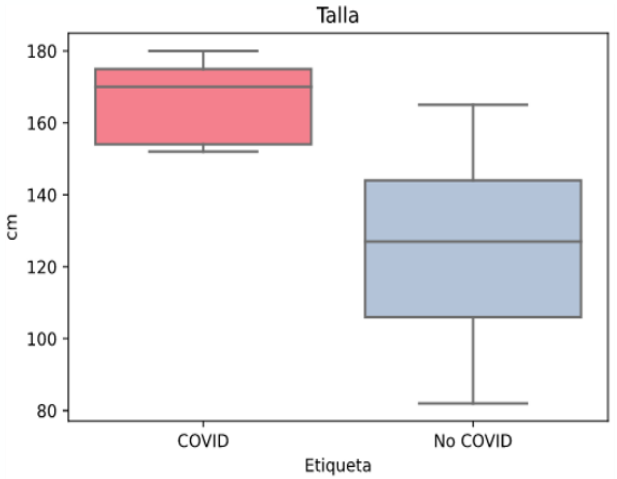
Extracción de información de expedientes médicos electrónicos

Identificación de datos antropométricos

- Son mediciones técnicas sistematizadas que expresan, cuantitativamente, las dimensiones del cuerpo humano: edad, peso, talla, IMC.
- Se ocuparon expresiones regulares para la extracción de los datos antropométricos, con las siguientes complicaciones:
  - *Cantidad de variaciones:*
  - *Diferentes escalas (metros, cm, etc.)*
  - *Errores ortográficos*







Detección de Síntomas

- Hipotético: El contexto puede cambiar la temporalidad a hipotética.
  - *“El paciente debe regresar en caso de fiebre”*
  - *“Paracetamol 1gr VO en caso de DOLOR”*
- Negaciones: El contexto determina si una condición se niega.
  - *“Sin fiebre”*
  - *“No dolor”*
- Faltas de ortografía
  - *“dificultadrespiratoria”*

Detección de entidades médicas

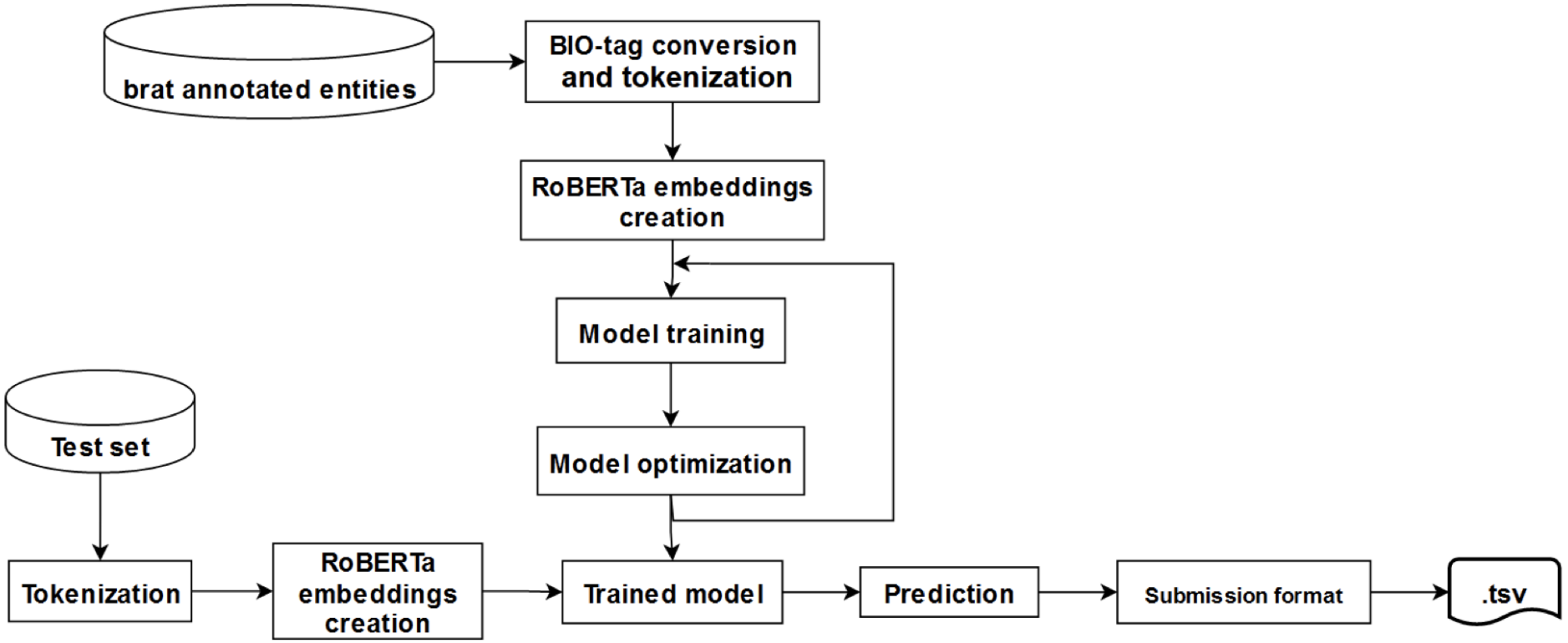


Figure 1: Named Entity Recognition System Diagram (Subtask 1)

Créditos: Orlando Ramos, Rodrigo del Moral y Javier Reyes

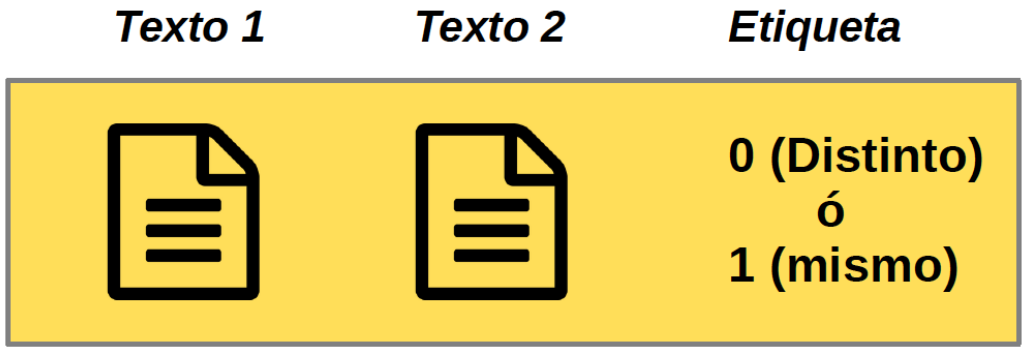
CLASIFICACIÓN AUTOMÁTICA DE TEXTOS

Verificación de autoría, Verificación de hechos e Identificación de noticias falsas, Minería de opiniones en Twitter, Análisis de sentimientos en repositorios de Software

Verificación de autoría

- Para entrenar el modelo, necesitamos parejas de textos etiquetados a forma de conocer si ambos textos fueron o no escritos por el mismo autor.

Una instancia:

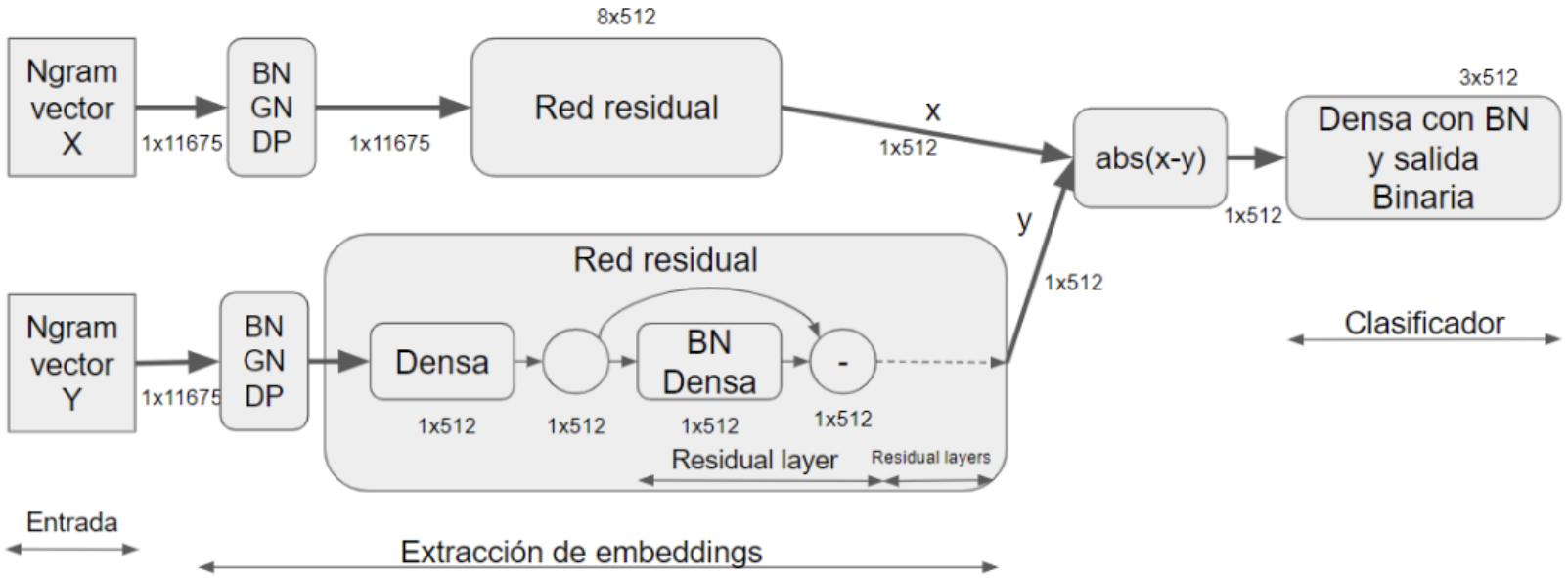


- Entrenamiento: conjunto de datos de provistos por PAN 2020 para la tarea de Verificación de Autoría

Corpus	Muestras	Muestras Positivas	Textos Diferentes	Caracteres máximos	Caracteres mínimos	Media de caracteres
Grande	275,565	147,778	494,257	943,947	20,355	21,426
Pequeño	52,601	27,834	93,667	296,887	20,670	21,425

Tabla 4.2: Mediciones estadísticas del conjunto de datos PAN 2020 de verificación de autoría (Kestemont et al., 2020).

Aprendizaje profundo aplicado a la verificación de autoría



Tesis de Maestría: Emir Araujo Pino (Enero, 2021)

Aprendizaje profundo aplicado a la verificación de autoría

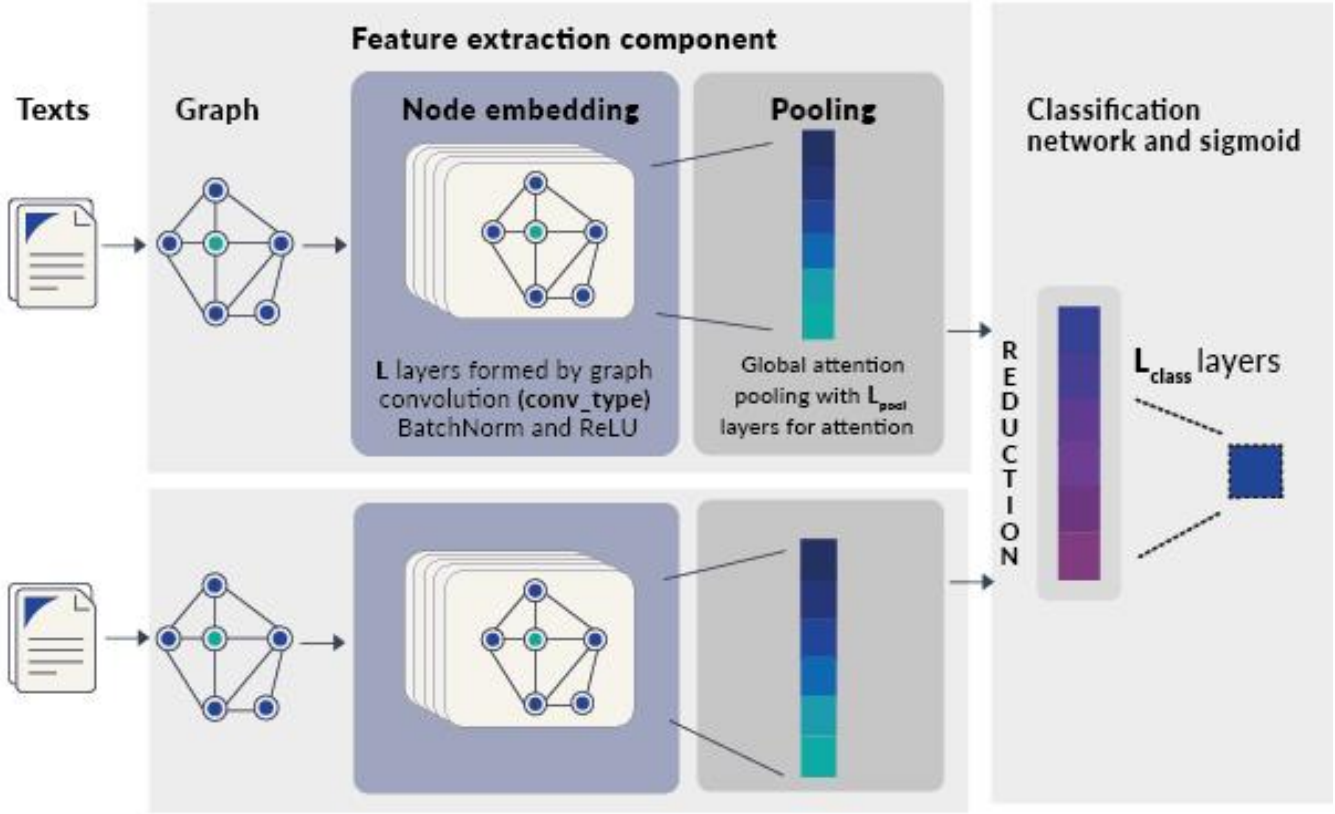
Nombre del Modelo	AUC	c@1	f_05_u	F1	Promedio
Red Siamesa Residual Grande	0.859	0.751	0.745	0.800	**0.789**
Red Siamesa Residual Pequeño	0.874	0.770	0.762	0.811	**0.804**
Baseline Naive (Kestemont, 2020)	0.780	0.723	0.716	0.767	0.747
Baseline Compresión (Stamatatos (2020))	0.778	0.719	0.703	0.770	0.742

Tabla 5.9: Resultados finales de la tarea compartida de verificación de autoría del PAN 2020 (Kestemont et al., 2020).

Tesis de Maestría: Emir Araujo Pino (Enero, 2021)

Araujo-Pino, E., Gómez-Adorno, H., & Pineda, G. F. (2020). Siamese Network applied to Authorship Verification. In *CLEF (Working Notes)*.

Graph-based Siamese Network applied to the Authorship Verification Task



Tesis de Maestría: Daniel Embarcadero (Diciembre, 2021)

Graph-based Siamese Network applied to the Authorship Verification Task

Table 5.8: System rankings for all PAN 2021 submissions. Taken from [21].

Team	Dataset	AUC	c@1	F1	F0.5u	Brier	Overall
boenninghoff21	large	**0.9869**	**0.9502**	**0.9524**	**0.9378**	**0.9452**	**0.9545**
embarcaderoruiz21	large	0.9697	0.9306	0.9342	0.9147	0.9305	0.9359
weerasinghe21	large	0.9719	0.9172	0.9159	0.9245	0.9340	0.9327
weerasinghe21	small	0.9666	0.9103	0.9071	0.9270	0.9290	0.9280
menta21	large	0.9635	0.9024	0.8990	0.9186	0.9155	0.9198
peng21	small	0.9172	0.9172	0.9167	0.9200	0.9172	0.9177
embarcaderoruiz21	small	0.9470	0.8982	0.9040	0.8785	0.9072	0.9170
menta21	small	0.9385	0.8662	0.8620	0.8787	0.8762	0.8843
rabinovits21	small	0.8129	0.9129	0.8094	0.8186	0.8129	0.8133
ikae21	small	0.9041	0.7586	0.8145	0.7233	0.8247	0.8050
unmasking21	small	0.8298	0.7707	0.7803	0.7466	0.7904	0.7836
tyo21	large	0.8275	0.7594	0.7911	0.7257	0.8123	0.7832
naive21	small	0.7956	0.7320	0.7856	0.6998	0.7867	0.7600
compressor21	small	0.7896	0.7282	0.7609	0.7027	0.8094	0.7581

Tesis de Maestría: Daniel Embarcadero (Diciembre, 2021)

Embarcadero-Ruiz, D., Gómez-Adorno, H., Reyes-Hernández, I., García, A., & Embarcadero-Ruiz, A. (2021). Graph-based Siamese network for authorship verification. In *CLEF*.

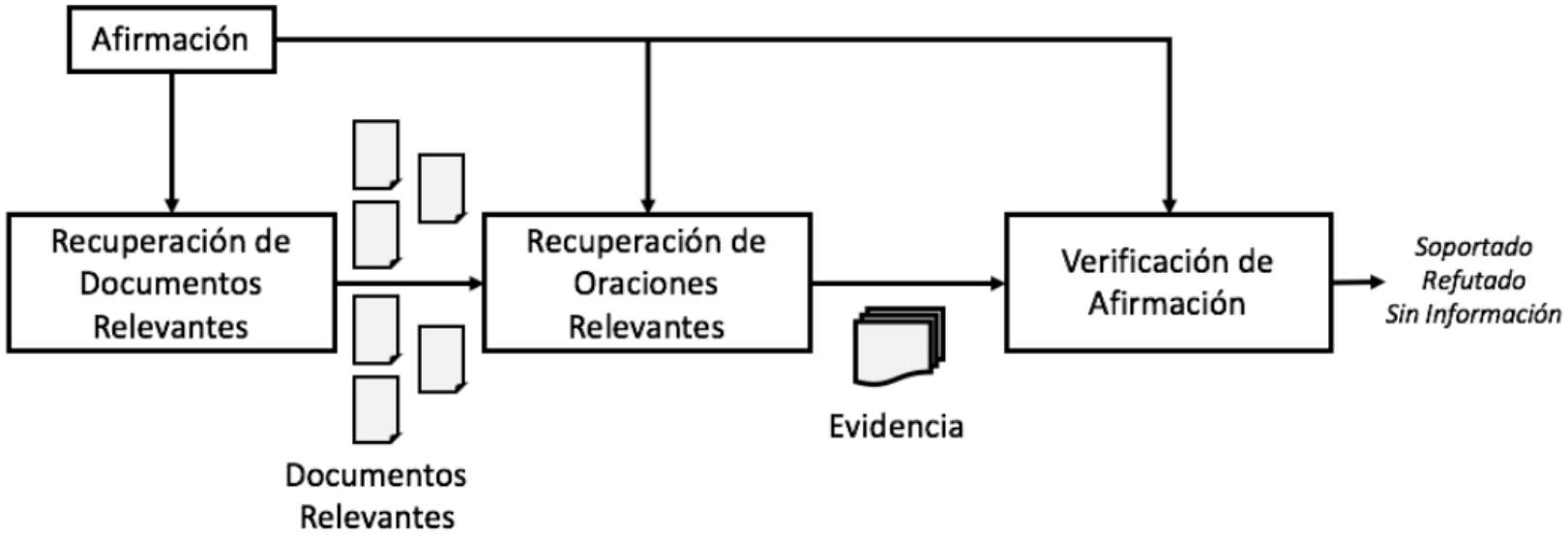
Detección de noticias falsas y verificación de hechos

- Distribución de noticias falsas no es un fenómeno nuevo
- Rápida distribución por medio de redes sociales
- Para lograr la detección de información falsa se han propuesto diferentes soluciones:
  - *Métodos basados en el conocimiento*
  - *Métodos basados en el estilo*
  - *Métodos basados en aspectos sociales*

Verificación de hechos

Método basado en el conocimiento

Figura 3.3: Verificación de hechos en 3 pasos.

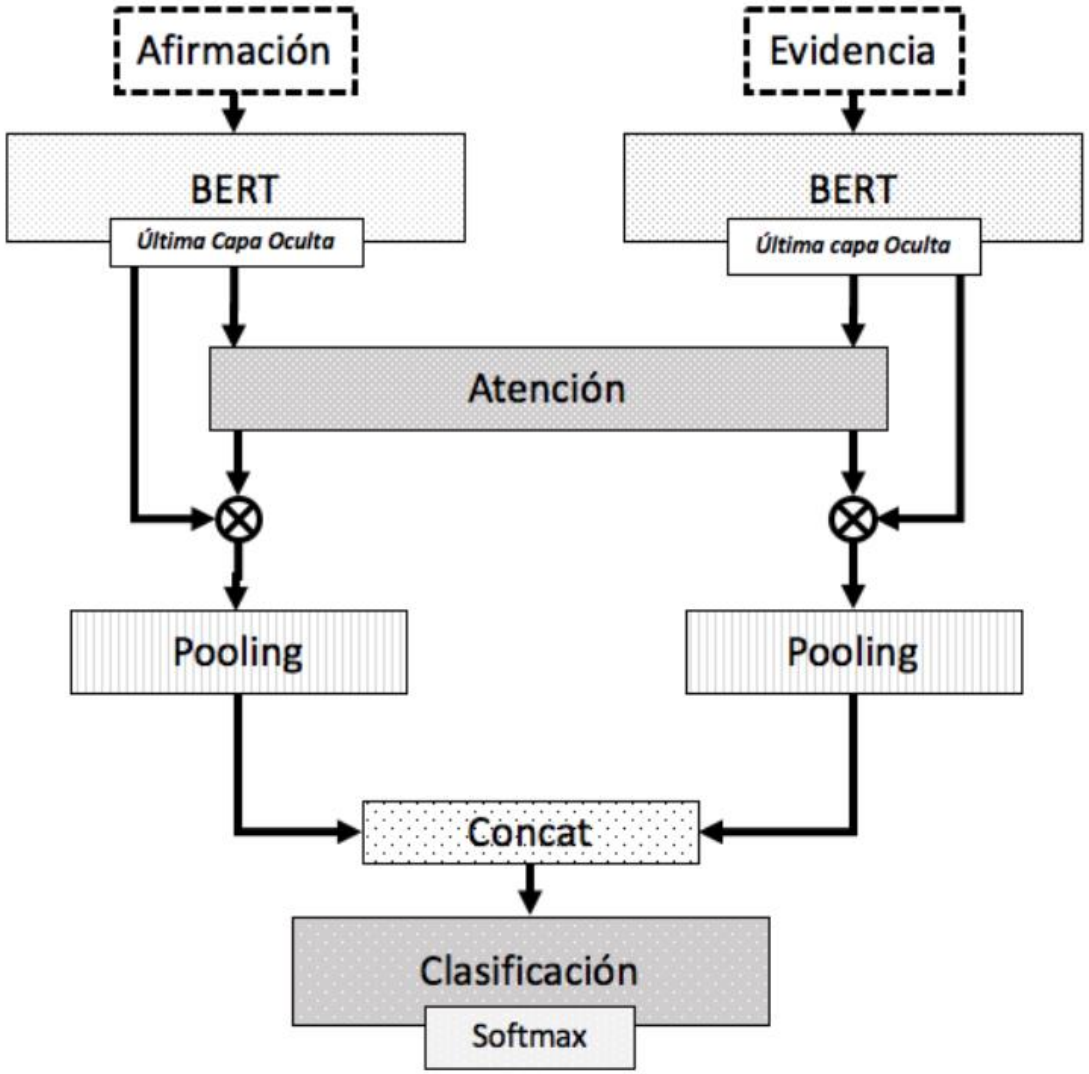


Tesis de Maestría: Ramón Casillas (Octubre, 2021)

Verificación automática de hechos utilizando un modelo de aprendizaje profundo interpretable

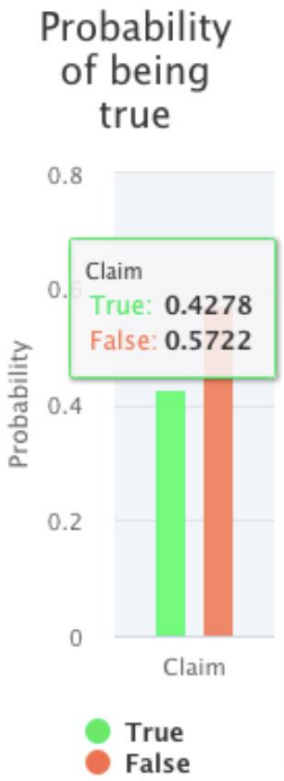
Se utilizó **Covid-19 Open Research Dataset** (CORD-19, por sussiglas en inglés). Este conjunto de datos surgió como una iniciativa del gobierno de EstadosUnidos junto con varios grupos de investigación. Contiene más de 300 mil artículos de inves-tigación sobre COVID-19 y temas relacionados.

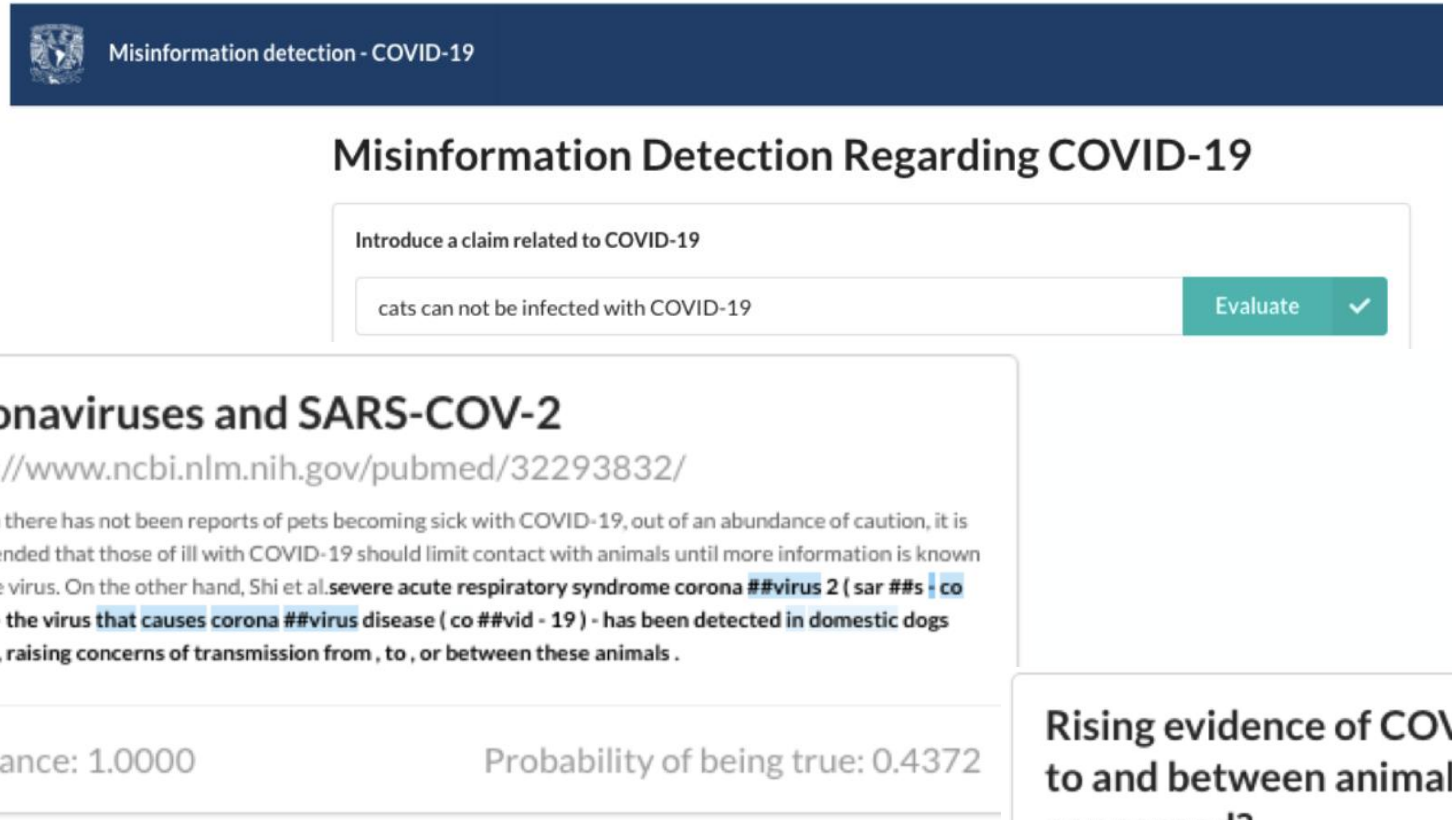
Dataset	Afirmaciones	Evidencia
Original	112	759
Original + negaciones	224	1518



Tesis de Maestría: Ramón Casillas (Octubre, 2021)

Verificación deinformación sobre COVID-19





Tesis de Maestría: Ramón Casillas (Octubre, 2021)

Detección de noticias falsas en Español

- Organización del FakeDES 2021 @ Iberlef
- **Objetivo:** Clasificar un conjunto de noticias en Español como verdaderas o falsas
- **Recolección** de corpus durante 2 años: 1,636 noticias recolectadas, 47% Falsas.
- Participants through the Codalab platform, 21 systems submitted

Approach	Best Accuracy	MPA	CFD	Number of Systems
Transformers	0.7657	0.9528	0.3834	5
Traditional Deep Neural Networks	0.6224	0.6224	-	1
BoW, n-grams, Stylometrics	0.7535	0.9580	0.3941	5
All teams (with submission)	0.7657	0.9895	0.3732	9
All teams	0.7657	0.9965	0.3382	21

Gómez-Adorno, H., Posadas-Durán, J. P., Enguix, G. B., & Capetillo, C. P. (2021). Overview of FakeDeS at IberLEF 2021: Fake News Detection in Spanish Shared Task. Procesamiento del Lenguaje Natural, 67, 223-231.

Detección de noticias falsas en Español

- **Métodos basados en el estilo**
  - Modelo de clasificación entrenado mediante la representación de bolsa de palabras (BoW).
  - Modelo de clasificación entrenado mediante la representación de n-gramas.
  - Modelo de clasificación basado en arquitecturas Transformers (BERT).

Baseline	Fake	True	F_{macro}	Accuracy
baseline-BERT	0.7321	0.7432	0.7321	0.7378
baseline-BOW-SVM	0.7217	0.7359	0.7217	0.729
baseline-CHAR-3-GRAMS-SVM	0.7063	0.6883	0.7063	0.6976

Tesis de Maestría: Claudia Porto Capetillo

Análisis de opiniones

iimas

Lingeniería lística

INSTITUTO DE INGENIERÍA UNAM

COVID-19 México Monitor de síntomas Análisis de emociones

Análisis de Twitter para COVID-19

Este es un sistema automático de vigilancia de COVID19 mediante Twitter. Se busca evaluar el comportamiento de las personas, estados de ánimo, la popularidad de las medidas del tomadas por el gobierno y síntomas de coronavirus.

Palabras clave de hoy

A continuación se enlistan los (#) hashtags, (@) menciones y palabras más frecuentes relacionados con la pandemia.

Última actualización: Thu May 26 00:00:20 2022

Hashtags de hoy

8 #covid19
4 #mexico
3 #salud
3 #puebla
3 #slp

Menciones de hoy

45 @hlgatell
17 @lopezobrador
10 @catrina_nortena
8 @who
5 @ssalud_mx

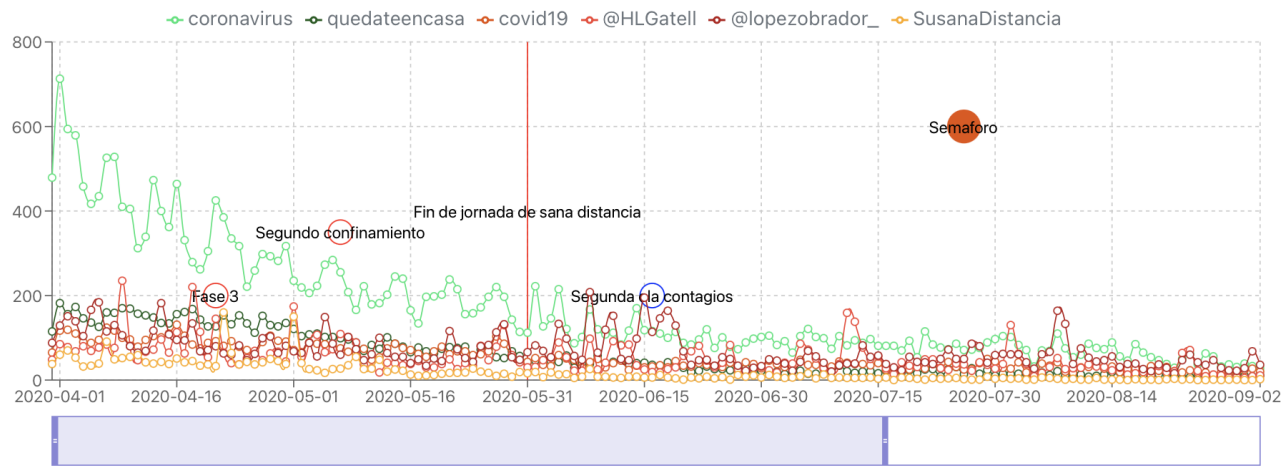
Palabras de hoy

42 pandemia
37 crisis
23 años
21 salud
14 🤣

Línea del tiempo de palabras clave

La siguiente línea del tiempo presenta 6 palabras clave dentro de los tweets obtenidos cada semana. Las primeras 4 palabras son acerca del virus y el aislamiento social: **coronavirus**, **quedateencasa**, **covid19** y **Susanadistancia**. Se tomaron en cuenta las menciones más frecuentes referentes a **Lopez Gatell** y **Lopez Obrador**. Esto para conocer cómo se van mencionando a medida que pasa el tiempo y el virus avanza.

Última actualización: Thu May 26 00:00:19 2022



http://www.miopers.unam.mx/covid/#/

Análisis de opiniones

Monitor de síntomas Análisis de emociones

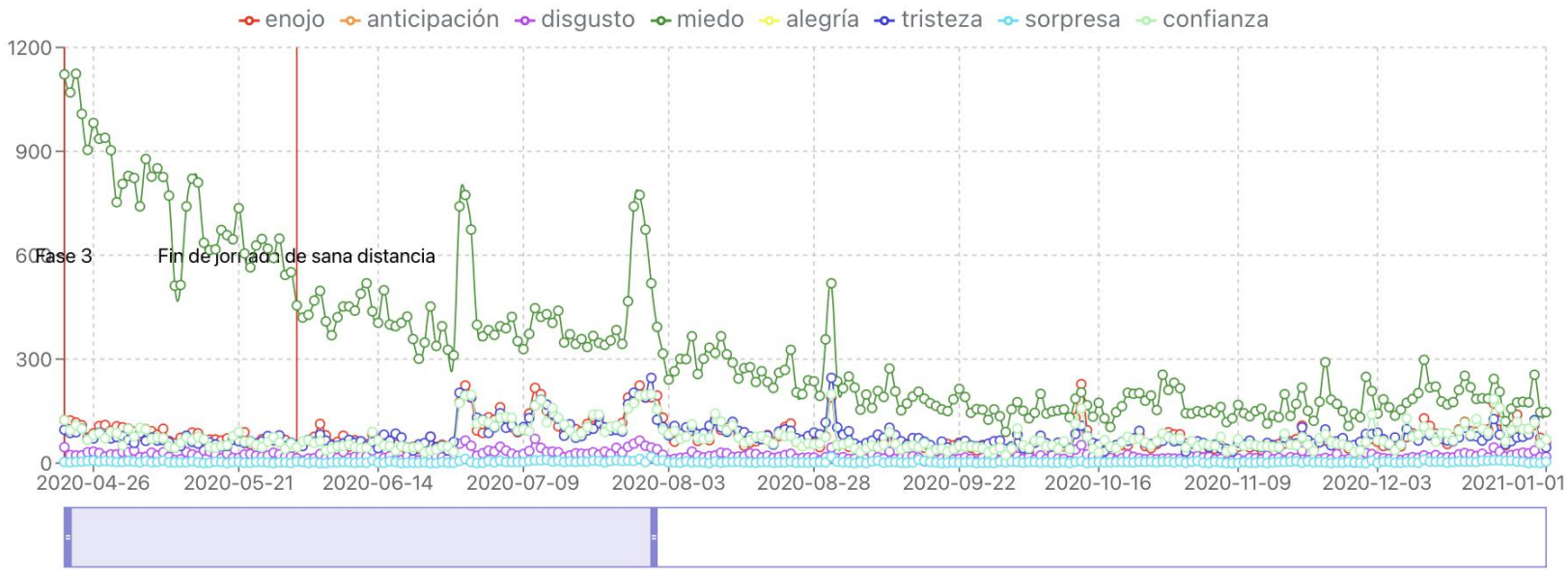
Análisis de emociones

Este es un sistema que monitorea las emociones provocadas por el COVID19 en Twitter

Línea del tiempo de emociones

La siguiente línea del tiempo presenta las emociones causadas en los usuarios de la red social.

Última actualización: Thu May 26 00:00:22 2022



http://www.miopers.unam.mx/covid/#/emociones

Análisis de opiniones

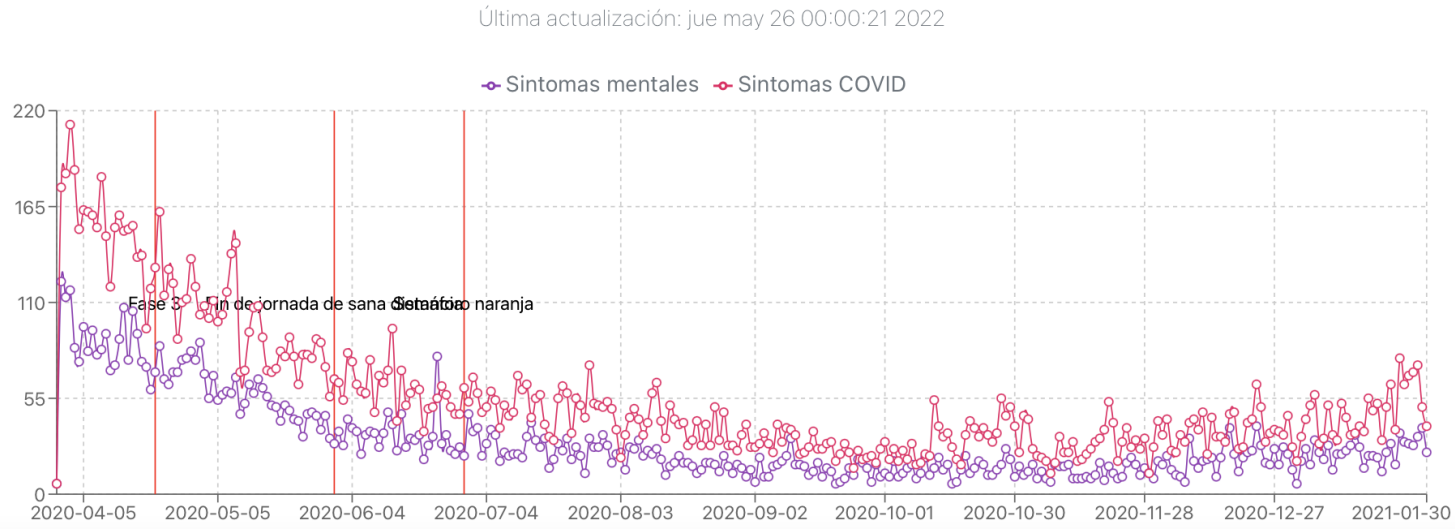
Monitor de síntomas Análisis de emociones

Análisis de síntomas para COVID-19

Este es un sistema automático de vigilancia de síntomas de COVID19 para México Los síntomas presentados tienen 2 categorías principales: los físicos causados por COVID19 como fiebre, tos o gripe y los transtornos ocasionados por el asilamiento social como ansiedad, depresión, insomnio, entre otros.

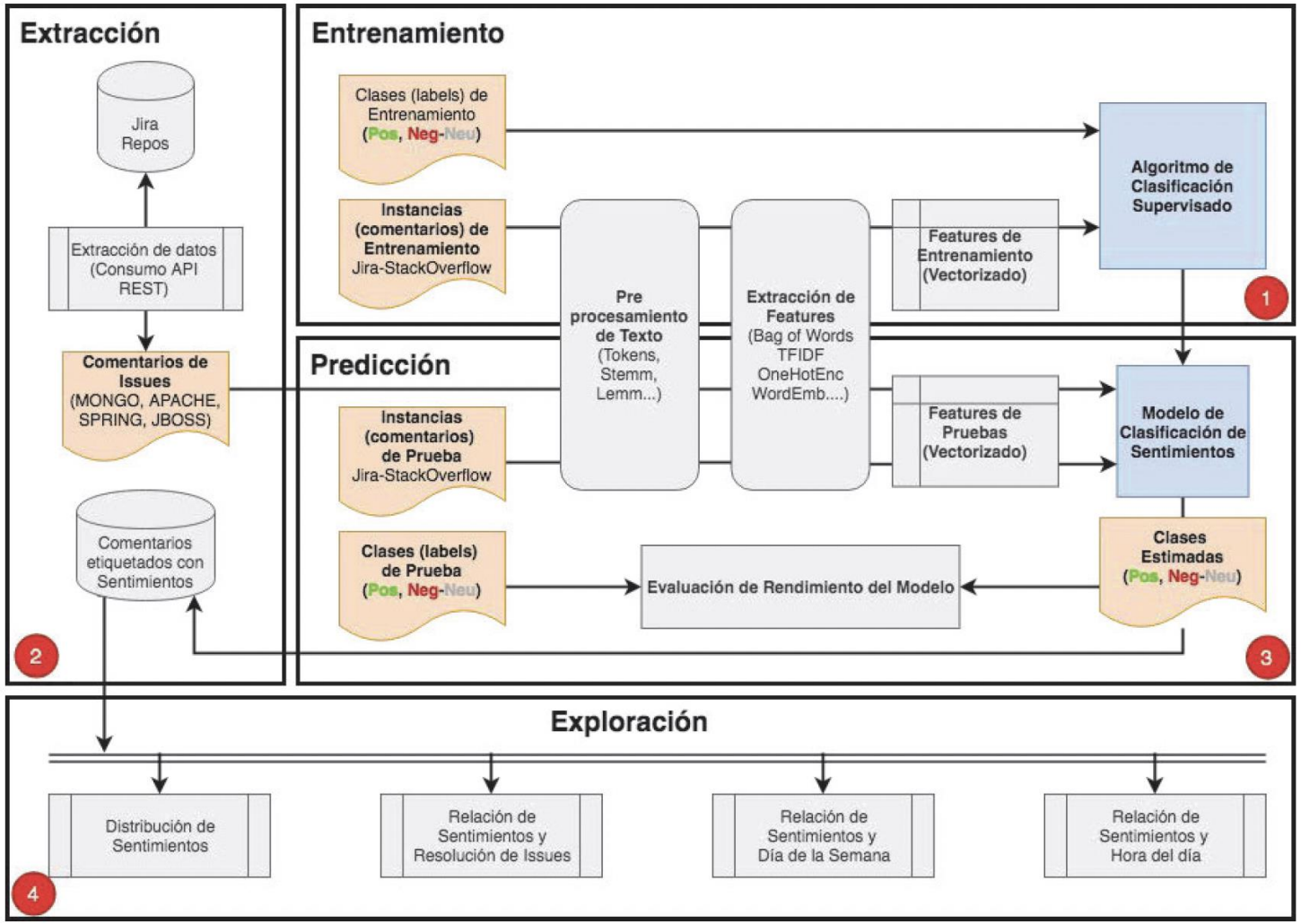
Línea del tiempo de síntomas COVID vs salud mental

El siguiente gráfico representa la cantidad de síntomas relacionados al COVID-19 y los estados de salud mental que se pueden presentar en la población debido al aislamiento social y demás factores. Por cada día se presentan dos valores, cada uno de ellos indica la cantidad de apariciones de síntomas de cada tipo (COVID-19 y estados de salud mental o psicológicos), esto con el fin de conocer cómo afecta el paso del tiempo a la frecuencia de estos dos valores.



http://www.miopers.unam.mx/covid/#/sintomas

Análisis de sentimientos en repositorios de Software



Tesis de Maestría: Andric Valdez Valenzuela (diciembre, 2020)

Análisis de sentimientos en repositorios de Software

#	Comentario	Sentimiento	#	Comentario	Sentimiento
1	thanks for reporting the issue. Processing-time based windows internally uses a ScheduledExecutor.	Positivo	7	I made some minor tweak to process the tuples with some time offset, please pull the current changes and test.	Neutral
2	Sorry for late response, call on getStartTimestamp() and getEndTimestamp()	Negativo	8	I pull and test again, yes, no tuple lost now. Thanks.	Positivo
3	You might want to try your tests with the patch	Neutral	9	when you get some time can you pull the updated changes and test again ?	Neutral
4	I tested several times with your patch, now it seems all right, no tuples overlap. Thank you	Positivo	10	I test again, as the same arguments as before, it seems all right. But when I make the spout emit more frequently, the ahead tuples are expired	Neutral
5	Eh, it still lose data, following is one test result, and you can see that the 15073 is lost.	Neutral	11	The initial tuples expired because the trigger was not fired exactly after the window interval but after a delay	Neutral
6	thanks for the update. I made some minor changes to the patch today. If you could pull the latest changes and run your tests a few times will be great.	Positivo	12	So that's it, I got, thanks !	Positivo

Tabla 5.4: Secuencia de comentarios para una Incidencia Resuelta del proyecto *STORM-APACHE*. Sentimientos: 41.67 % positivos, 8.33 % negativos y 50 % neutrales.

Tesis de Maestría: Andric Valdez Valenzuela (diciembre, 2020)

Análisis de sentimientos en repositorios de Software

#	Comentario	Sentimiento	#	Comentario	Sentimiento
1	Sorry, but anyone had a chance to look at this? The error gets quite frequent after primary/secondary switch, but after restart it becomes fine.	Negativo	7	If you see the log I sent you there's some lines logging setting this option. Sorry, but I think that's not the root cause of the problem.	Negativo
2	Sorry - not yet. I'll have a look tomorrow unless somebody beats me to it.	Negativo	8	Hi guys, am I right about above comment?	Neutral
3	Hi Derick, thanks for looking into this. Any luck about this issue?	Positivo	9	Sorry but we've both been on vacations so haven't had the time to look into it again. Last I tried I couldn't reproduce it though.. I'll see if I can look into it tomorrow	Negativo
4	I got as far as setting up an SSL replicated environment, but this is causing me some issues that I haven't been able to nail down. I have not forgotten about this yet!	Neutral	10	Thanks for your reply. We found that the issue only happens when accessing the test script via apache, Let me know if you need any more information.	Positivo
5	Sorry to ask, but this issue is really blocking us from moving forward for hot-standby with PHP . Is there any way we can do or any resource we can reach to speed it up? Thanks	Negativo	11	You are not hitting anything similar to PHP-329, your issues are SSL specific.	Neutral
6	I cannot reproduce this at all. Can you create a short reproduce script that reproduces this issue for you?	Neutral	12	Closing this due to inactivity.	Neutral

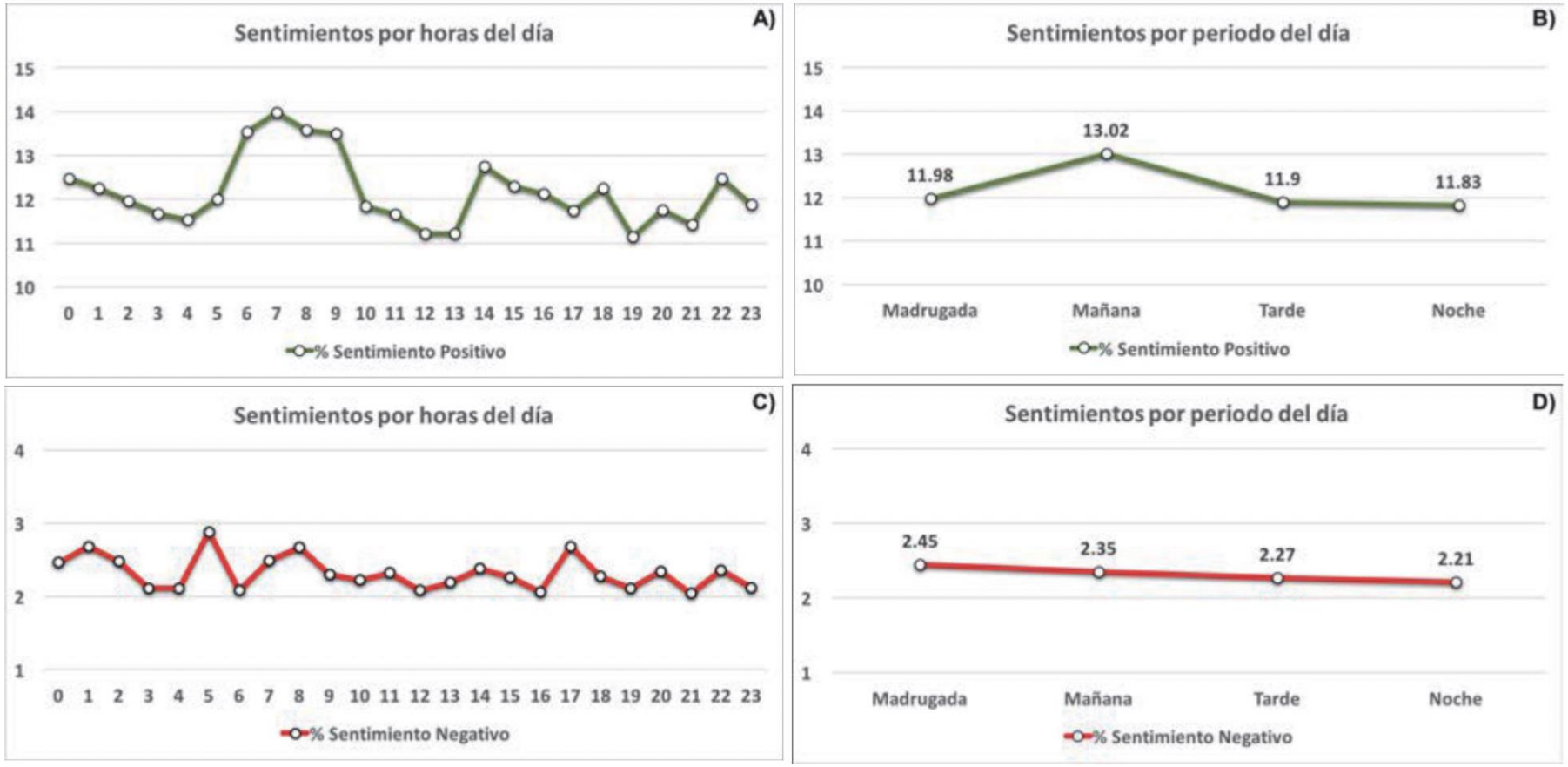
Tabla 5.5: Secuencia de comentarios para una Incidencia No Resuelta del proyecto *PHP-MONGO*. Sentimientos: 11.76 % positivos, 29.41 % negativos y 58.82 % neutrales.

Tesis de Maestría: Andric Valdez Valenzuela (diciembre, 2020)

Análisis de sentimientos en repositorios de Software

Tesis de Maestría: Andric Valdez Valenzuela (diciembre, 2020)

Análisis de sentimientos en repositorios de Software



Tesis de Maestría: Andric Valdez Valenzuela (diciembre, 2020)

GRACIAS

Dra. Helena Gómez Adorno
helena.gomez@iimas.unam.mx

Créditos:

Alan Emir Araujo, Daniel Embarcadero, Ramón Casillas,
Claudia Porto, Javier Reyes, Rodrigo del Moral, Andric Valdez

iimas