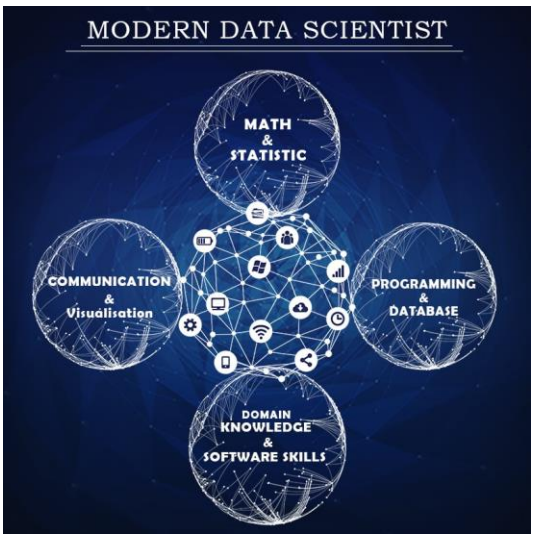
La Ciencia de Datos y sus Aplicaciones

Dr. Carlos Adolfo Piña García
Twitter: @Piniisima

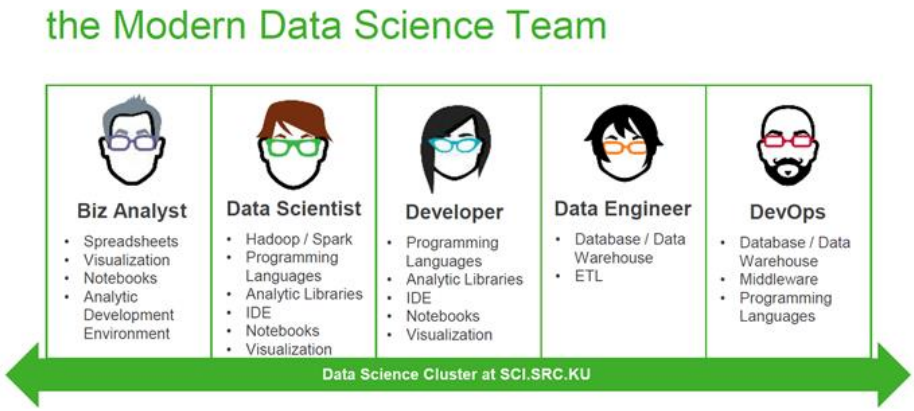
Centro de Estudios de Opinión y Análisis (CEOA)
Universidad Veracruzana

Ciencia de Datos

A continuación una de las muchas versiones que existen entre la relación de los roles en la Ciencia de Datos y como interactúan entre ellos.



Ciencia de Datos



Elecciones

Conoce el trabajo de análisis de redes que hizó la Universidad Veracruzana a través del #LARSI para el Tercer #DebateINE en las #Elecciones2018 #VotoLibre @INEMexico

Elecciones

LARSI-UV (https://uv.mx/larsi/) participó en monitoreo de elecciones en Panamá

- Scraping Global Threats in Facebook Through Movement Patterns Generated by Random Walks

- Buscar, muestrear y monitorear los mensajes de Facebook es un área de aplicación importante, utilizable para varios propósitos. Se puede utilizar, por ejemplo, para intentar detectar personas o grupos que engañan deliberadamente a la policía, o para detectar vulnerabilidades de grupos sospechosos (terrorismo).
- Fuente: https://ieeexplore.ieee.org/document/6375376?arnumber=6375376

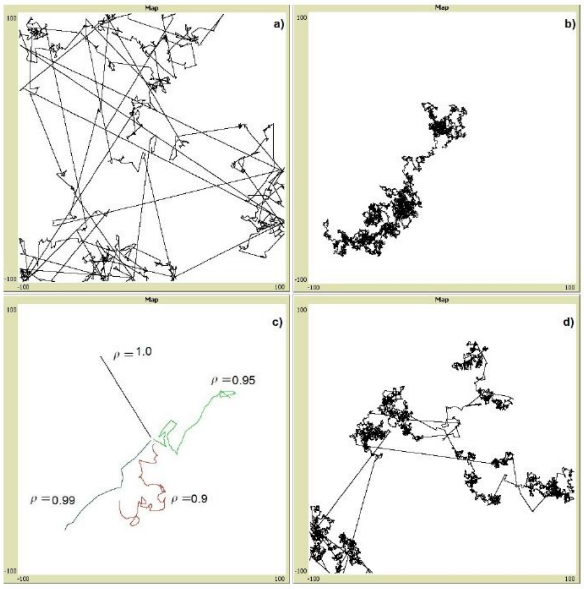


Fig. 1 Plots showing simulated random walks: (a) Lévy walk with $\mu = 2$; (b) Brownian walk with $\mu = 3$; (c) Correlated Random Walk (CRW) with $0 \leq \rho \leq 1$ and (d) Adaptive or Composite strategy switching between Lévy and brownian motion.

TABLE I

LIST OF KEY WORDS USED AS SEARCHING PARAMETERS. THEY ARE CLASSIFIED BY BREAKING NEWS OR MOOD ANALYSIS.

From breaking news reports	From mood analysis
anonymous	happy
Afghanistan	sad
Syria	hello
Korea	revolution
Barack Obama	riots
Government	-
al-Qaeda	-

Patrones de recolección de información

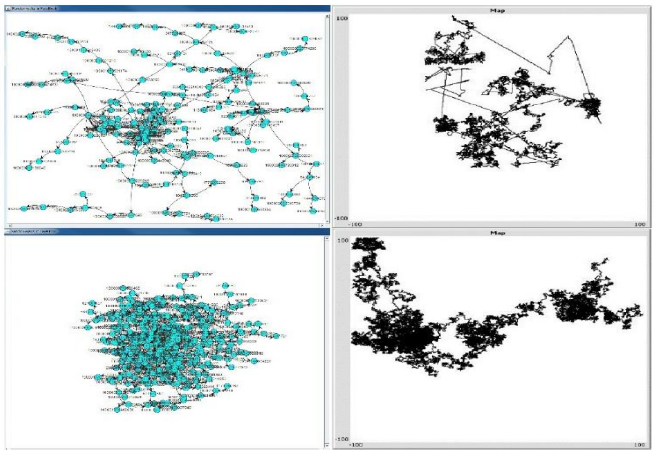


Fig. 3. Examples of movement paths generated by the social forager. On the top left-hand side we show an Adaptive strategy compared with a simulated animal movement path using the same walk (on the top right-hand side). On the bottom left-hand side we show a Brownian motion in a social environment compared with a simulated particle displacement (bottom right-hand side).

Patrones de recolección de información

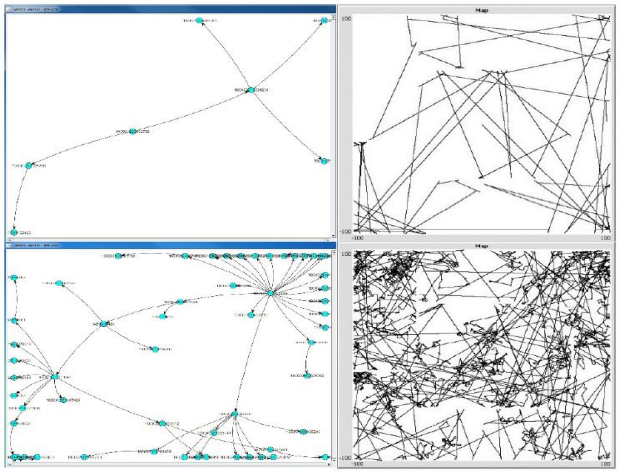
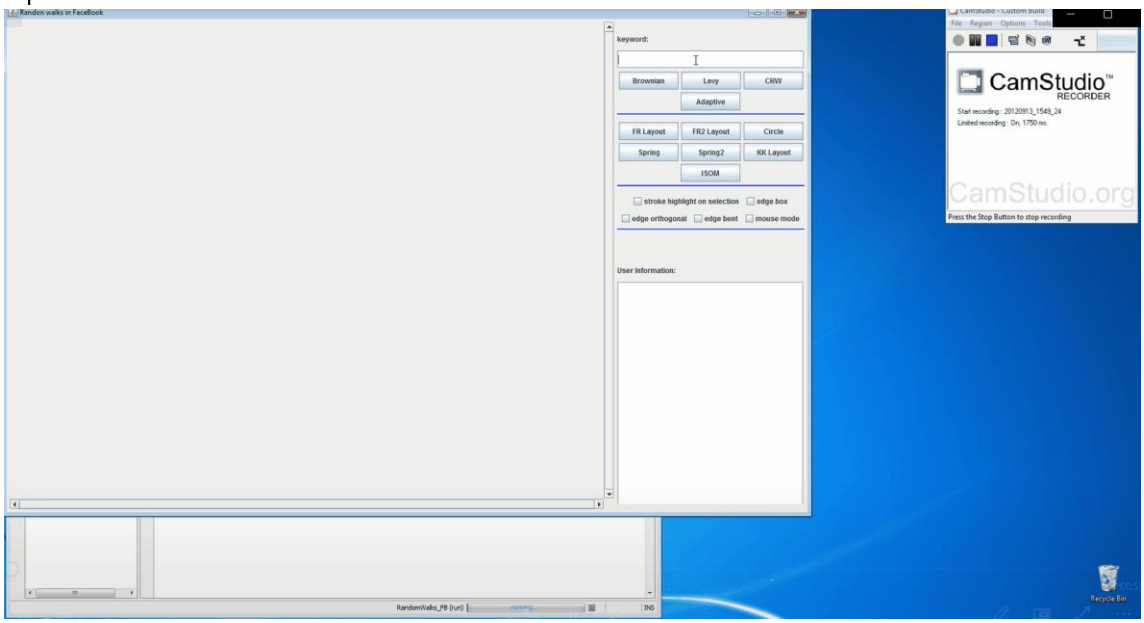


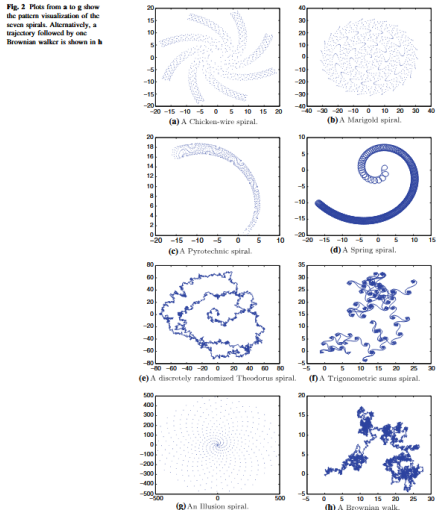
Fig. 4. Examples of movement paths generated by the social forager. On the top left-hand side we show a CRW with $\rho = 1$ compared with a simulated ballistic movement with $\mu = 1$ (on the top right-hand side). On the bottom left-hand side we show a Lévy walk with $\mu \approx 2$ compared with a simulated foraging trajectory with $\mu = 2$ (bottom right-hand side).

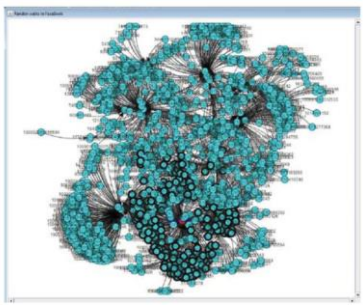
Hasta donde sabemos, nuestro estudio es el primer análisis de patrones que compara la ruta de movimiento de los animales con un modelo conceptual de búsqueda y muestreo para investigar posibles amenazas emergentes planteadas en tiempo real.

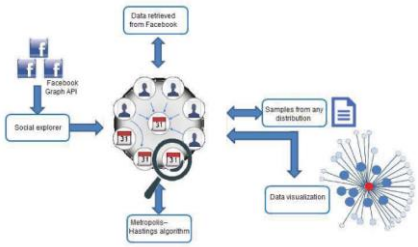


Ciencia

En el artículo: "**Spiraling Facebook: an alternative Metropolis–Hastings random walk using a spiral proposal distribution**" presentamos un nuevo algoritmo de muestreo en redes sociales, el **Alternative Metropolis–Hasting Random Walk (AMHRW)**, basado en una espiral en lugar de la distribución normal tradicional.
https://link.springer.com/article/10.1007/s13278-013-0126-8







Algorithm 1: Alternative Metropolis-Hastings Random Walk.

initialization:
$t \leftarrow 0$ initial time;
$v_0 \leftarrow 0$ initial node;
stopping criterion $\leftarrow$ EOF;
$\varrho \leftarrow q(.|X_t)$ a candidate node from a given distribution;

while *stopping criterion not met* **do**

Generate a candidate node w in terms of ϱ from $q(.|X_t)$;

Generate U from a uniform $(0, 1)$ distribution;

if $U \leq \frac{deg(v)}{deg(w)}$ **then**

$v \leftarrow w$;

else

Stay at v;

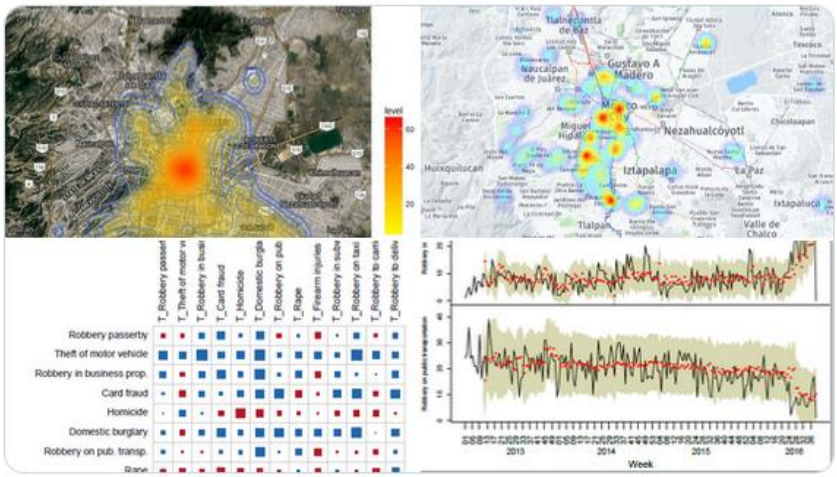
end

Set $t = t + 1$;

end

Ciencia

En el artículo: "**Exploring crime patterns in Mexico City**" presentamos un estudio BigData utilizando Google Trends y Twitter como predictores del delito en CDMX
https://rdcu.be/bLn40



Ciencia

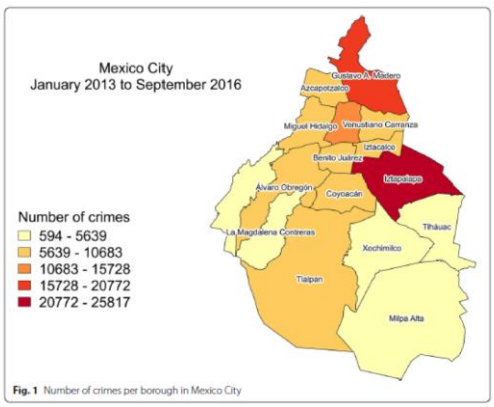


Fig. 1 Number of crimes per borough in Mexico City

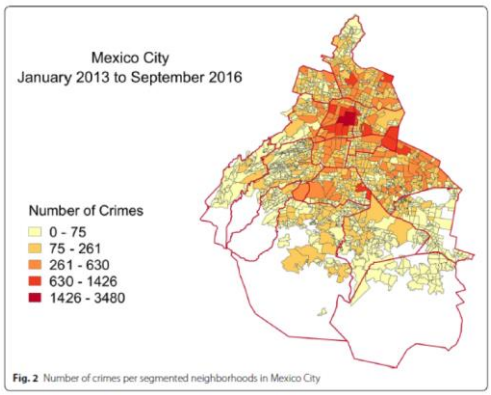


Fig. 2 Number of crimes per segmented neighborhoods in Mexico City

Ciencia

The degree of intensity of relationship in this case implies a **limited** correlation.

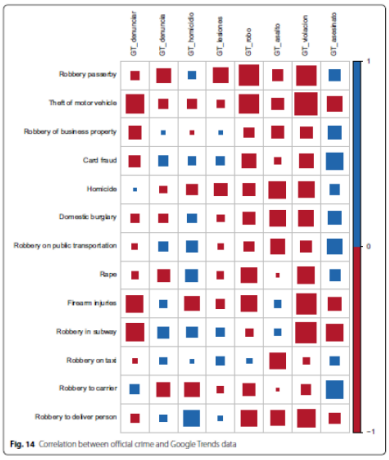


Fig. 14 Correlation between official crime and Google Trends data

Ciencia

- This model is explained as follows: Let $Y_1, \ldots, Y_m$ be the weekly official data series of a specific crime, and let $\boldsymbol{X}_t = X_{t,1}, \ldots, X_{t,r}$ be the number of Google Trends series for keywords or terms $i = 1, \ldots, r$. The considered autoregressive model for week t is
-

$$Y_t^* = X_{t-1}'\beta + \sum_{i=1}^{p} \phi_i Y_{t-i}^* + \sum_{j=1}^{q} \psi_j v_{t-j} + v_t, \qquad (1)$$

- where β is a vector of regression coefficients, $v_t \sim \text{WN}(0, \sigma^2)$; $Y_t^* = (1-B)^d Y_t$, $p, d, q \in \mathbb{Z}^+$, $t = 1, \ldots, m$; $\phi(\lambda) = 1 - \phi_1\lambda - \ldots - \phi_p\lambda^p$; $\psi(\lambda) = 1 + \psi_1\lambda + \ldots + \psi_q\lambda^q$, and $\phi(\lambda), \psi(\lambda) \neq 0$, $\forall |\lambda| \leq 1$.

Ciencia

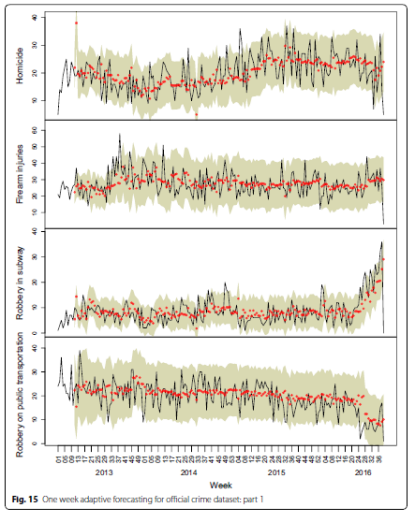


Fig. 15 One week adaptive forecasting for official crime dataset: part 1

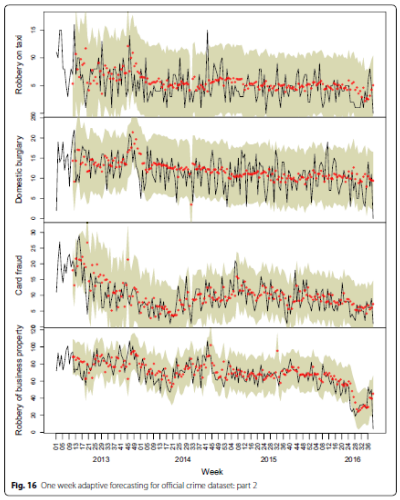


Fig. 16 One week adaptive forecasting for official crime dataset: part 2

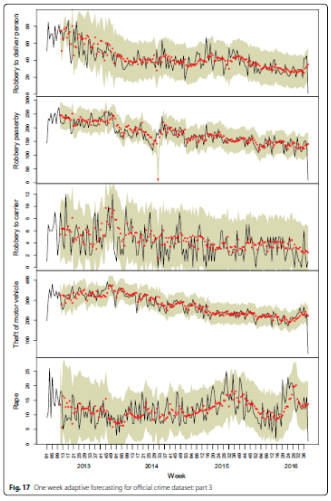
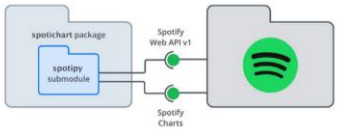
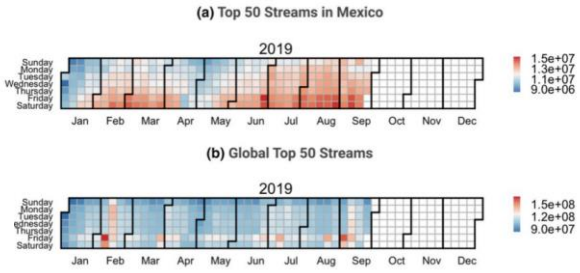


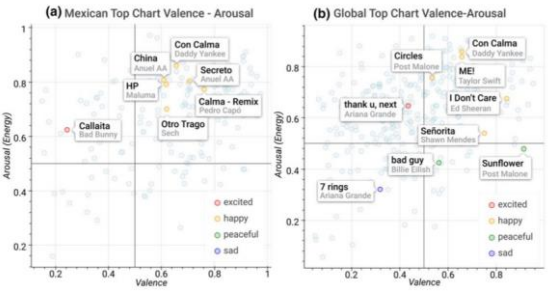
Fig. 17 One week adaptive forecasting for official crime dataset: part 3

Ciencia

En el artículo: "The rhythm of Mexico: an exploratory data analysis of Spotify's top 50" presentamos una comparación de los gustos internacionales y nacionales en la plataforma Spotify: https://bit.ly/3lzl0vd







Ciencia

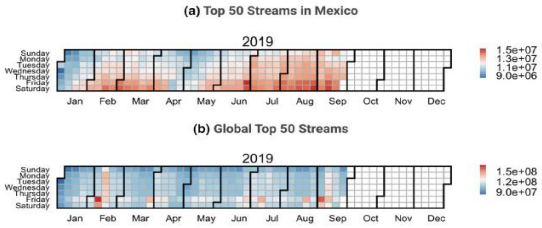


Fig. 2 Heatmap of the streaming of songs in the top 50

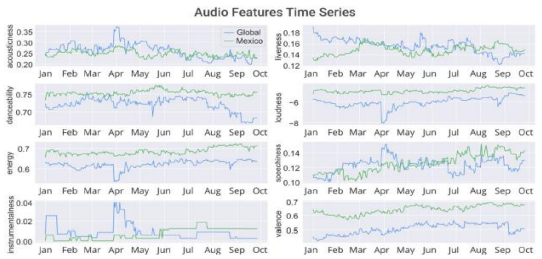


Fig. 3 Audio feature time series comparison

Ciencia

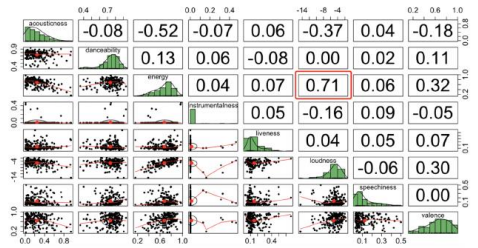


Fig. 4 Pairwise scatter plot matrix, histogram, and Pearson correlation coefficients of Mexico top chart. It can be seen a moderate correlation between energy and loudness

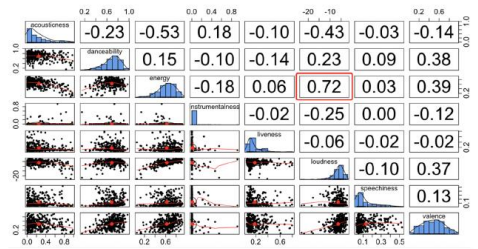


Fig. 5 Pairwise scatter plot matrix, histogram, and Pearson correlation coefficients of the global top chart. It can be seen a moderate correlation between energy and loudness

Ciencia

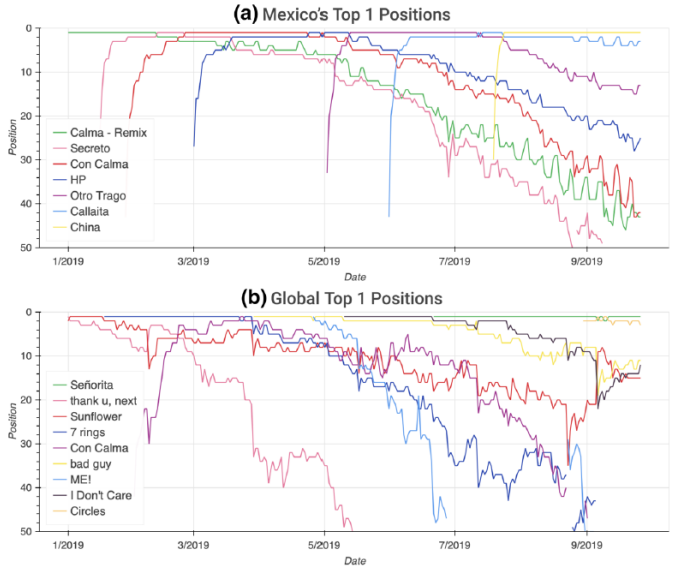
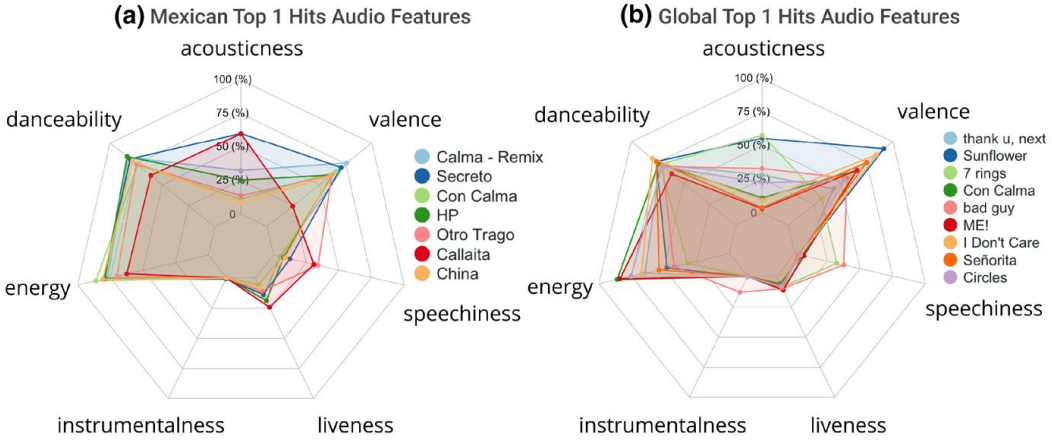


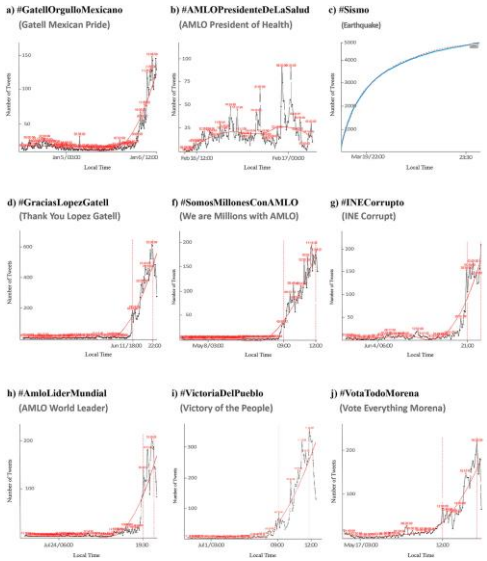
Fig. 10 Top 1 hits time series

Ciencia

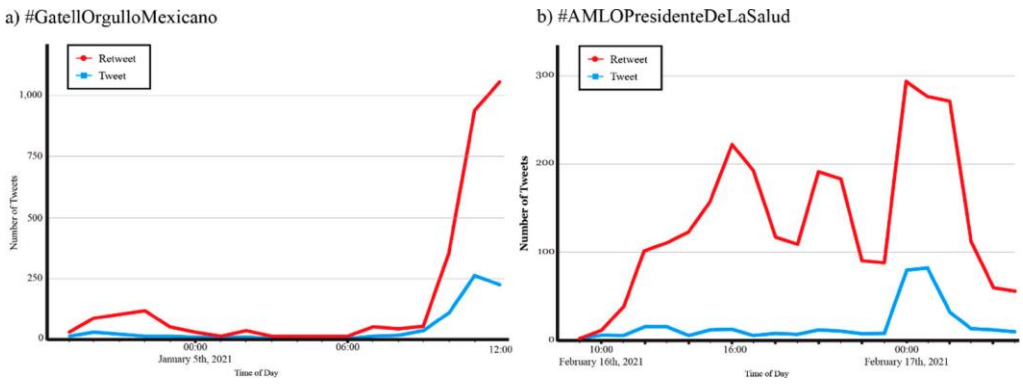


Ciencia

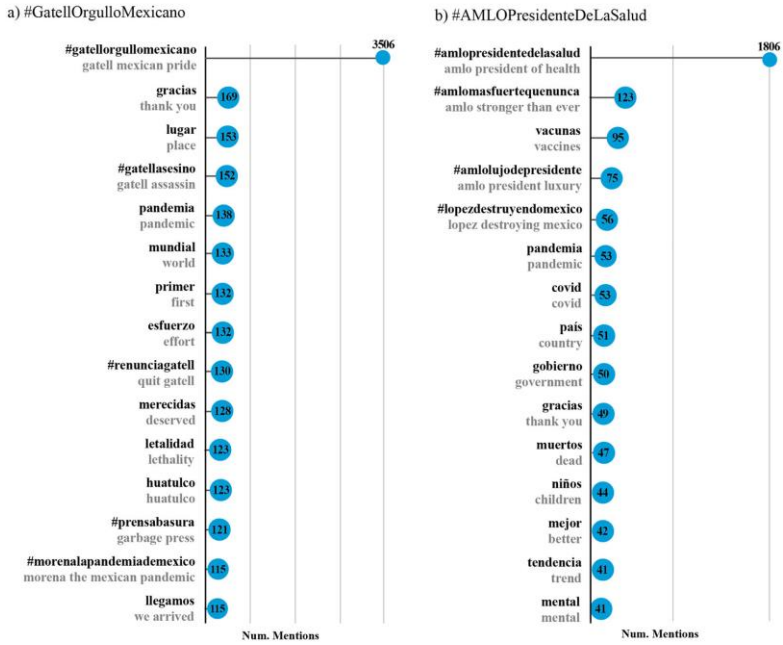
En el artículo: "Coordinated campaigns on Twitter during the coronavirus health crisis in Mexico " presentamos un análisis del astroturfing en México: https://www.tandfonline.com/doi/full/10.1080/25729861.2022.2035935?src=



Ciencia



Ciencia



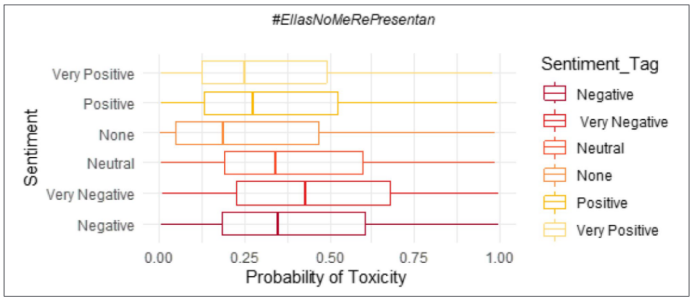
Ciencia

En el artículo: "Toxic Social Media: Affective Polarization After Feminist Protests " presentamos un análisis de polarización en movimientos feministas en México: https://journals.sagepub.com/doi/full/10.1177/20563051221098343?fbclid=IwAR34WdNmOwjB97Ksl_5CnTjdcLpFckM6XGRRo4oUV83-VREvu0d3-mNmIhg

Table 2. Toxicity and Sentiment: Measures of Central Tendency and Dispersion of the Hashtags.

Sentiment	#TheyDoNotRepresentMe			#TheyDoRepresentMe		
	Mean	Median	Variance	Mean	Median	Variance
Very positive	0.330	0.249	0.0628	0.292	0.239	0.0561
Positive	0.350	0.277	0.0689	0.287	0.232	0.0546
Neutral	0.404	0.343	0.0677	0.346	0.295	0.0579
Negative	0.407	0.347	0.0716	0.356	0.301	0.0582
Very negative	0.458	0.429	0.0738	0.429	0.386	0.0622
None	0.277	0.190	0.0713	0.234	0.158	0.0489

Source: Own elaboration.



Ciencia

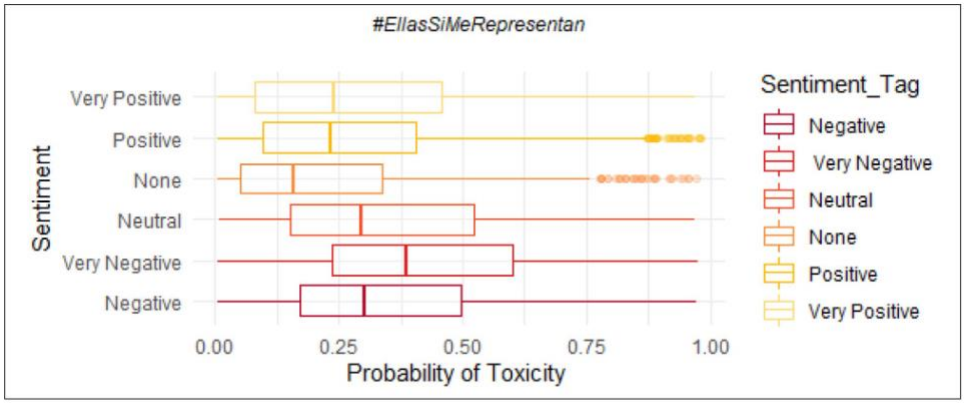


Figure 2. Toxicity of the tweets in relation to sentiment #TheyDoRepresentMe.

Ciencia

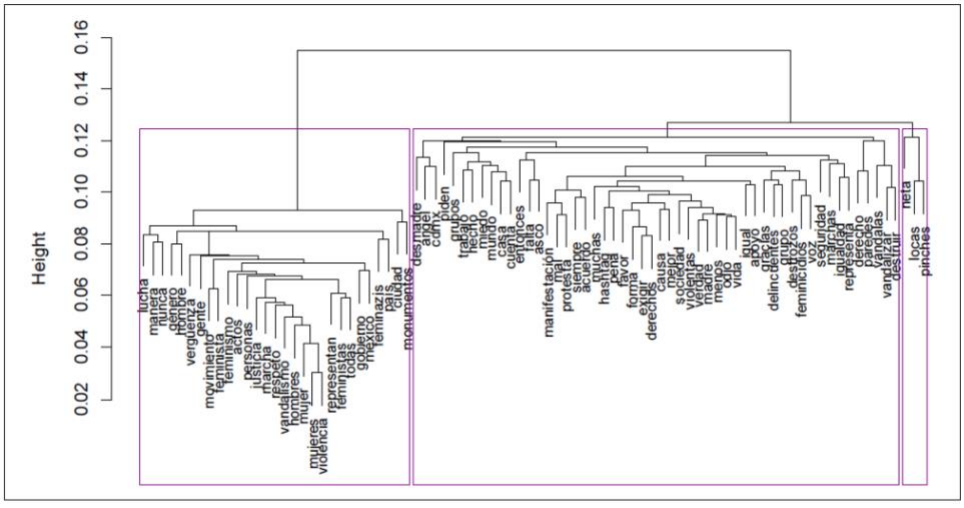


Figure 3. Dendrogram of #TheyDoNotRepresentMe own elaboration.

Ciencia

I don't remember independence or revolution being peaceful, but because they were led by men they were called heroes. Really still don't see it? #TheyDoRepresentMe

#TheyDoRepresentMe When tyranny is law, revolution is order. ✊💚

THE REVOLUTION WILL BE FEMINIST. #TheyDoRepresentMe and represent all those who no longer have a voice. If you are not going to support, don't get in the way!

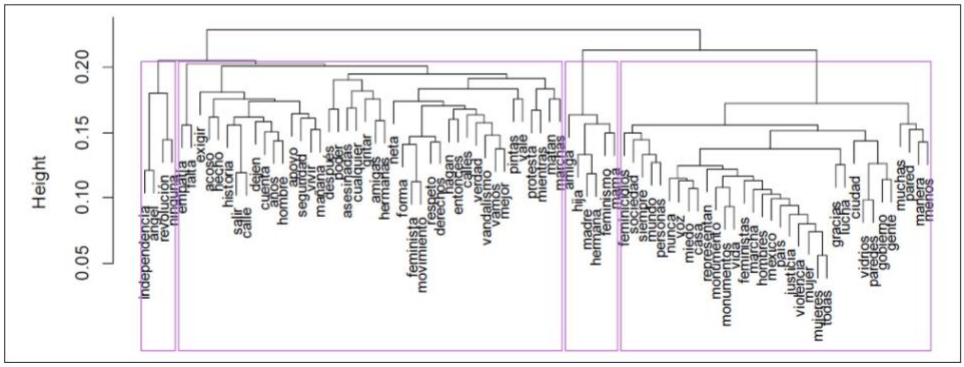
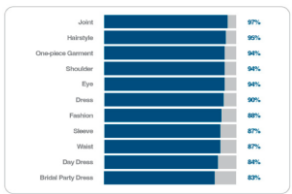
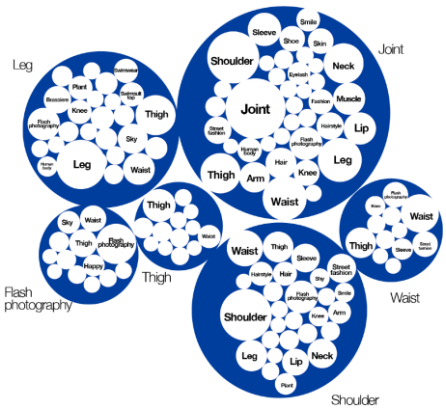


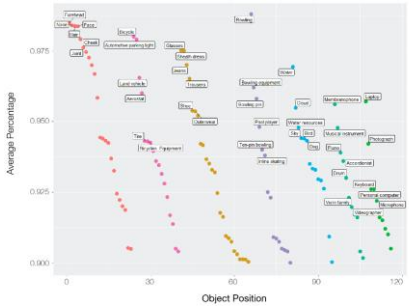
Figure 4. Dendogram of #TheyDoRepresentMe.

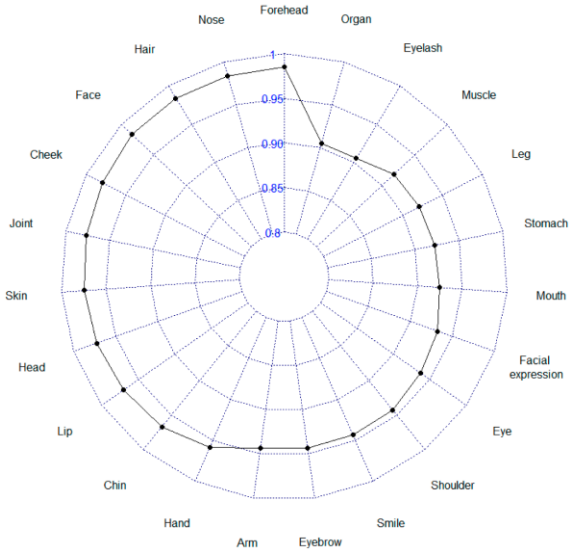
Ciencia

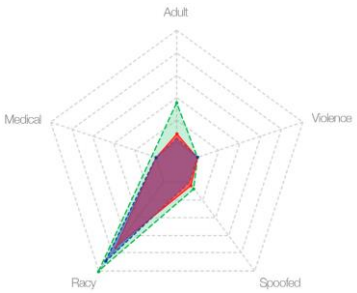
En el artículo: "The Impact of Social Media on Sexual Self-Concept: A Qualitative Analysis of Instagram in Mexico" El estudio analiza la autorrepresentación de mujeres en Instagram y sus efectos negativos ligados a estereotipos y objetivación. Se aplica Google Vision API para identificar contenido sugestivo y se encuentra que más del 90% de las imágenes incluyen desnudez parcial o poses provocativas. https://www.mdpi.com/2411-5118/6/1/9





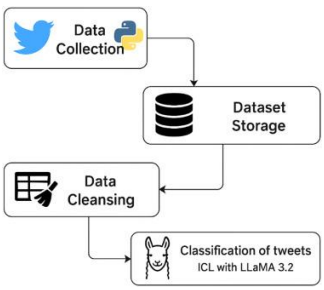


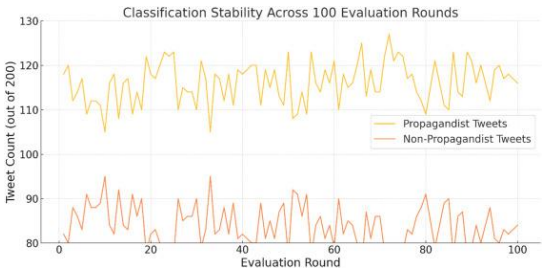


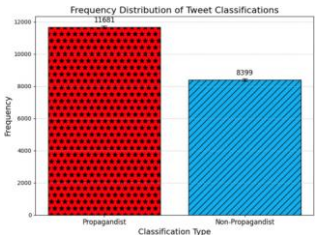


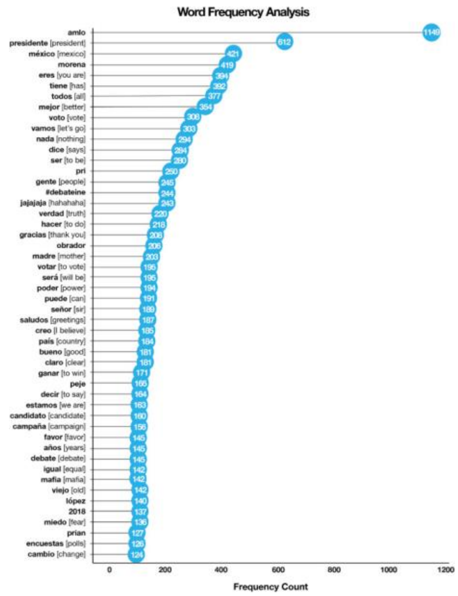
Ciencia

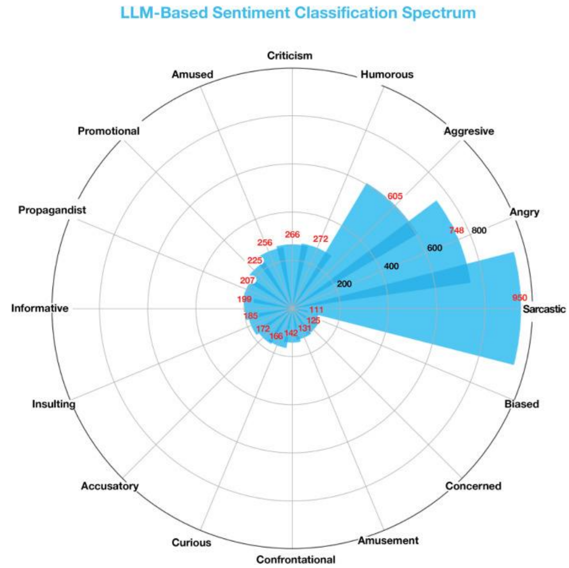
En el artículo: "**In-context learning for propaganda detection on Twitter Mexico using large language model meta AI"** Se aplicaron LLMs (LLaMA 3.2) para detectar propaganda política en Twitter durante la elección presidencial mexicana de 2018. El modelo identificó que 58.4% de los tuits contenían narrativas propagandistas, con predominio de tono negativo y ataques personales.
https://www.sciencedirect.com/science/article/pii/S2772503025000465

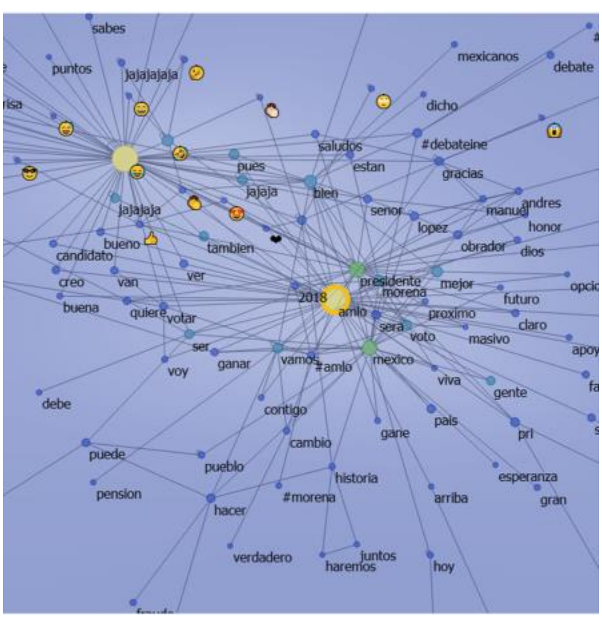


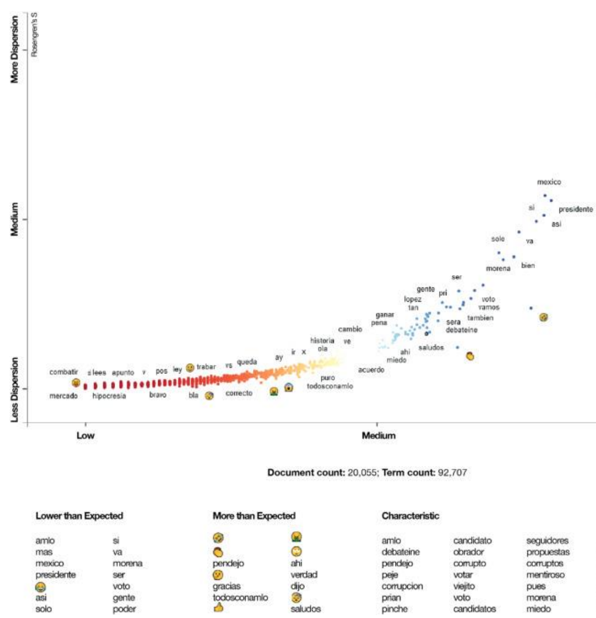





Word Frequency Analysis
amlo 1149
presidente [president] 612
méxico [mexico] 421
morena 419
eres [you are] 394
tiene [has] 392
todos [all] 377
mejor [better] 354
voto [vote] 308
vamos [let's go] 303
nada [nothing] 294
dice [says] 284
ser [to be] 280
pri 250
gente [people] 245
#debateine 244
jajajaja [hahahaha] 243
verdad [truth] 220
hacer [to do] 218
gracias [thank you] 208
obrador 206
madre [mother] 203
votar [to vote] 195
será [will be] 195
poder [power] 194
puede [can] 191
señor [sir] 189
saludos [greetings] 187
creo [I believe] 185
país [country] 184
bueno [good] 181
claro [clear] 181
ganar [to win] 171
peje 165
decir [to say] 164
estamos [we are] 163
candidato [candidate] 160
campaña [campaign] 156
favor [favor] 145
años [years] 145
debate [debate] 145
igual [equal] 142
mafia [mafia] 142
viejo [old] 142
lópez 140
2018 137
miedo [fear] 136
prian 127
encuestas [polls] 126
cambio [change] 124
0
200
400
600
800
1000
1200
Frequency Count


LLM-Based Sentiment Classification Spectrum
Criticism
Humorous
Aggresive
Angry
Sarcastic
Biased
Concerned
Amusement
Confrontational
Curious
Accusatory
Insulting
Informative
Propagandist
Promotional
Amused
266
272
605
748
950
111
125
131
142
166
172
185
199
207
225
256
200
400
600
800


sabes
puntos
jajajajaja
mexicanos
debate
dicho
saludos
#debateine
pues
estan
gracias
jajaja
bien
jajajaja
senor
lopez
manuel
andres
honor
bueno
candidato
tambien
obrador
dios
creo
van
ver
presidente
morena
mejor
2018
futuro
buena
quiere
votar
amlo
proximo
claro
sera
voto
masivo
ser
voy
ganar
vamos
#amlo
mexico
viva
gente
debe
contigo
pais
cambio
gane
prl
puede
pueblo
historia
esperanza
pension
hacer
#morena
arriba
gran
verdadero
juntos
haremos
hoy


More Dispersion
Medium
Less Dispersion
Low
Medium
mexico
si
presidente
asi
solo
ya
morena
bien
ser
gente
pri
lopez
voto
vamos
tambien
sera
debateine
ganar
pena
cambio
historia
ola
ve
ahi
miedo
saludos
acuerdo
puro
todosconamio
combatir
mercado
hipocresia
bravo
bla
correcto
Document count: 20,055; Term count: 92,707
Lower than Expected
amlo
mas
mexico
presidente
asi
solo
si
va
morena
ser
voto
gente
poder
More than Expected
pendejo
gracias
todosconamio
ahi
verdad
dijo
saludos
Characteristic
amlo
debateine
pendejo
peje
corrupcion
prian
pinche
candidato
obrador
corrupto
votar
viejito
voto
candidatos
seguidores
propuestas
corruptos
mentiroso
pues
morena
miedo

¡Gracias!

Dr. Carlos Adolfo Piña García

Twitter: @Piniisima

Mi página personal:
https://www.uv.mx/personal/cpina/

Email: cpina@uv.mx