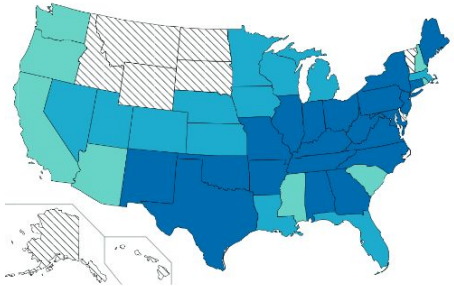
Geolocalización de usuarios en Twitter utilizando redes convolucionales de grafos

F. Funes[1], *J-I Alvarez Hamelin*[1,2], *M. G. Beiró*[1,2]

[1] Facultad de Ingeniería (Universidad de Buenos Aires)
[2] INTECIN (Universidad de Buenos Aires-CONICET)

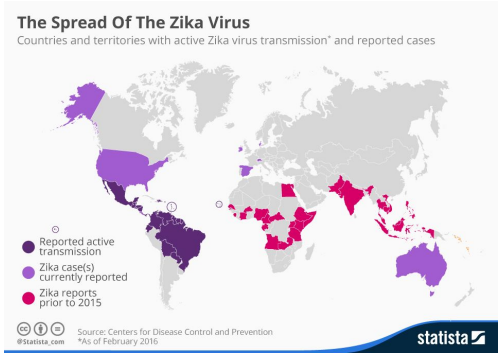
1. OBJETIVO Y MOTIVACION

- El objetivo es **predecir** con cierto grado de resolución la **ubicación de los usuarios** de una red social como Twitter.
- La predicción de geolocalización encuentra diversas **aplicaciones** en áreas como:
  - Detección temprana de emergencias y catástrofes
  - Monitoreo de salud pública
  - Personalización de contenido



(b) April

"You Are What You Tweet: Analyzing Twitter for Public Health", Paul & Dredze, 5th Int. AAAI Conf. on Weblogs and Soc. Media, 2011

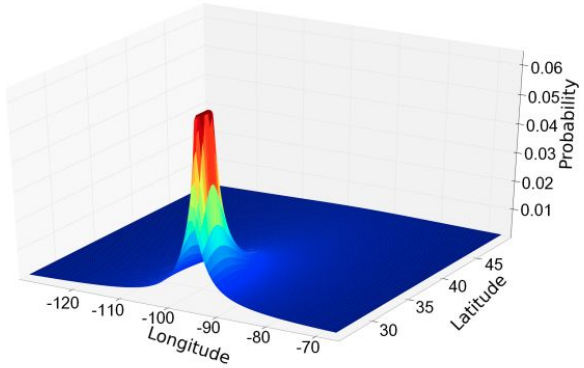


"Forecasting Zika Incidence in the 2016 Latin America Outbreak Combining Traditional Disease Surveillance with Search, Social Media, and News Report Data", McGough et al., PLoS Neglected Tropical Diseases, 2017.

2. ESTADO DEL ARTE

- **[CHENG ET AL., 2010]** *"A Content-Based Approach to Geo-locating Twitter Users"*, Cheng et al., Proc. of the 19th ACM Int. Conf. on Inf. & Know. Man. (CIKM), 2010.
  - Se basa en el **contenido del tweet** para predecir su ubicación
  - Define el concepto de **Local Indicative Words (LIW)**, es decir, palabras que son determinantes en cierta medida de una ciudad.
  - Propone la Accuracy@100 como métrica adaptada al problema.

- 130k usuarios en Estados Unidos
- 4.1M de tweets
- Etiquetas: ciudades con ≥ 5000 hab.
- Obtienen una Acc@100 de 0.51



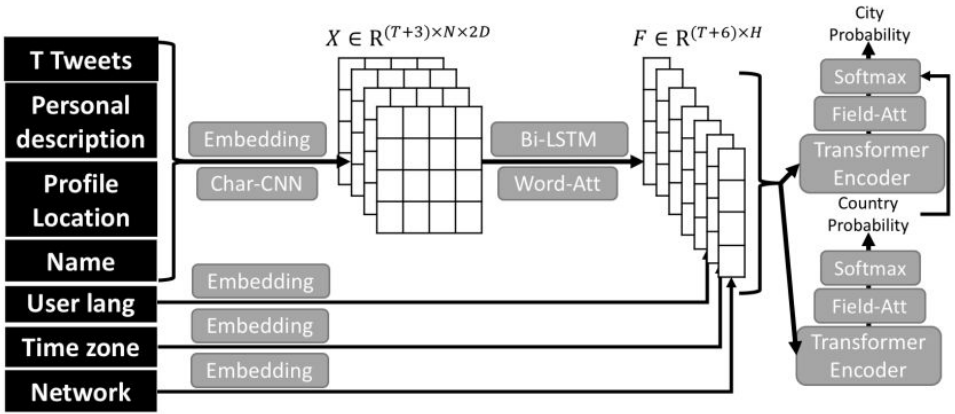
2. ESTADO DEL ARTE

- **[RAHIMI ET AL., 2015]** *"Twitter User Geolocation Using a Unified Text and Network Prediction Model"*, Rahimi et al., Proc. of the 53rd Annual Meet. of the ACL, 2015.
  - El grupo de Rahimi introdujo el uso de las interrelaciones entre usuarios para la predicción de ubicación, construyendo un **grafo de menciones.**
  - Utiliza un **modelo de propagación de etiquetas** *(Modified Adsorption)*.
  - El texto producido por cada usuario se utiliza como *prior*.
  - Las ciudades se agrupan en nodos de un ***k-d tree*** para balancear la cantidad de usuarios en cada nodo.

    - Dataset **Twitter-US [ROLLER ET AL., 2012]**
    - 450k usuarios en Estados Unidos
    - 38M de tweets
    - Etiquetas: 378 ciudades de Estados Unidos
    - Obtienen una Acc@100 de 0.60

2. ESTADO DEL ARTE

- **[HUANG AND CARLEY, 2019]** *"A hierarchical location prediction neural network for Twitter user geolocation"*, Huang & Carley, Proc. of the 2019 Conference on Empirical Methods in NLP, 2019.
  - Utilizan un método que combina embeddings del contenido con embeddings del grafo de menciones.
  - Superan el estado del arte obteniendo una **Acc@100 de 0.70** en el dataset **Twitter-US**., utilizando ciudades como etiquetas.



3. CONJUNTO DE DATOS

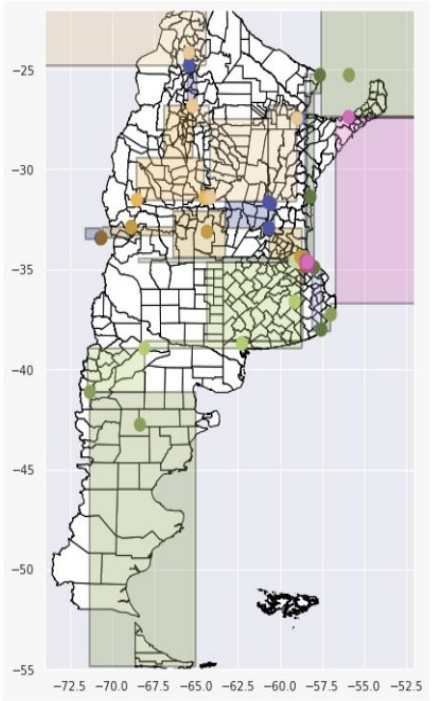
- Obtenido a partir de un dataset de 900 millones de tweets (2M usuarios) capturado durante 2019 [**MUSSI REYERO, 2021**].
- No todos los usuarios cuentan con datos de geolocalización.
- A través de un proceso de limpieza construímos dos datasets curados de usuarios con localización exacta o aproximada en base a sus tweets:

	Twitter-ARG-Exact	Twitter-ARG-Bbox
Cantidad de usuarios	37.146	141.209
Cantidad de tweets	27.574.343	124.192.146

"Evolution of the political opinion landscape during electoral periods", Mussi Reyero *et al.*, EPJ Data Science, 10(1), 31, 2021.

4. MODELO DE APRENDIZAJE

- Cada usuario del conjunto posee una etiqueta que indica la ciudad en donde vive (clase).
- Cantidad de clases:
  - Twitter-ARG-Exact: 95 ciudades con al menos 100 muestras cada una.
  - Twitter-ARG-BBox: 229 ciudades con al menos 100 muestras cada una.
- Ambos conjuntos de datos poseen clases muy desbalanceadas.
- Planteamos el problema como un problema de clasificación multi-etiqueta.
- Comparamos la performance de utilizar nodos vs. ciudades como etiquetas



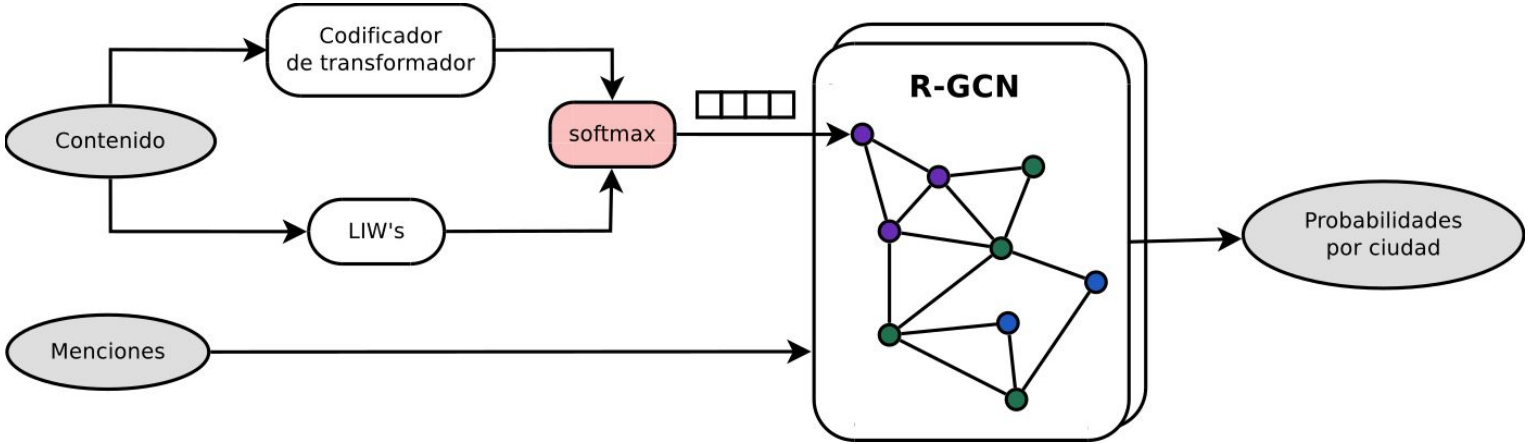
4. MODELO DE APRENDIZAJE

- Nuestro modelo utiliza como información:
  - El **contenido** publicado por los usuarios
  - Las **interrelaciones** que los usuarios establecen con otros usuarios, ya sea al mencionarlos o seguirlos.

- Para combinar esta información, ajustamos un modelo de aprendizaje basado en el contenido, y difundimos el resultado de ese modelo en un grafo de usuarios.

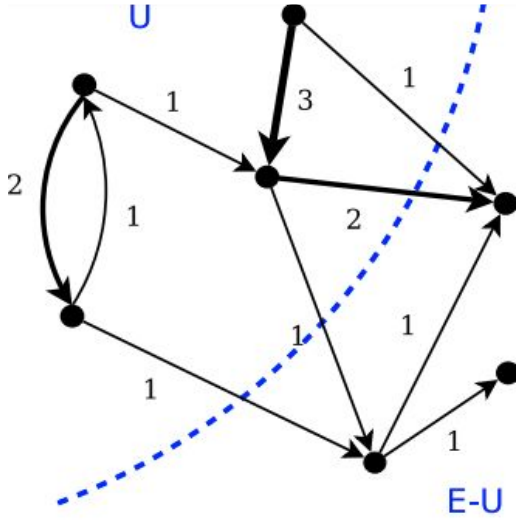
4. MODELO DE APRENDIZAJE

- *La siguiente figura resume la arquitectura de la red:*



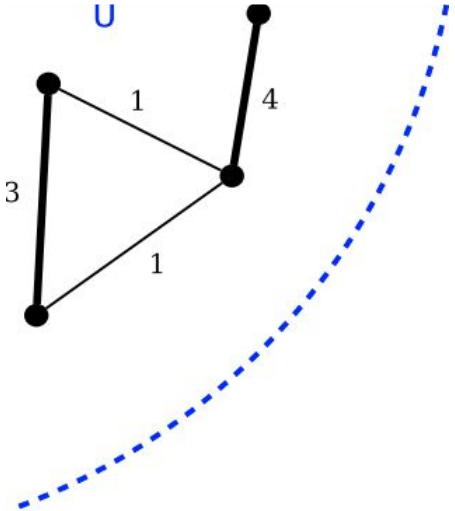
4. MODELO DE APRENDIZAJE

- Grafos de menciones y de seguidores extendidos:



(a)

Grafo original de menciones, incluyendo menciones a usuarios externos.



(b)

Grafo de menciones extendido con comenciones.

4. MODELO DE APRENDIZAJE

- LIWs (*local indicative words*):
  - Utilizamos pruebas χ^2 para hallar palabras que sean significativamente indicativas de una ciudad.

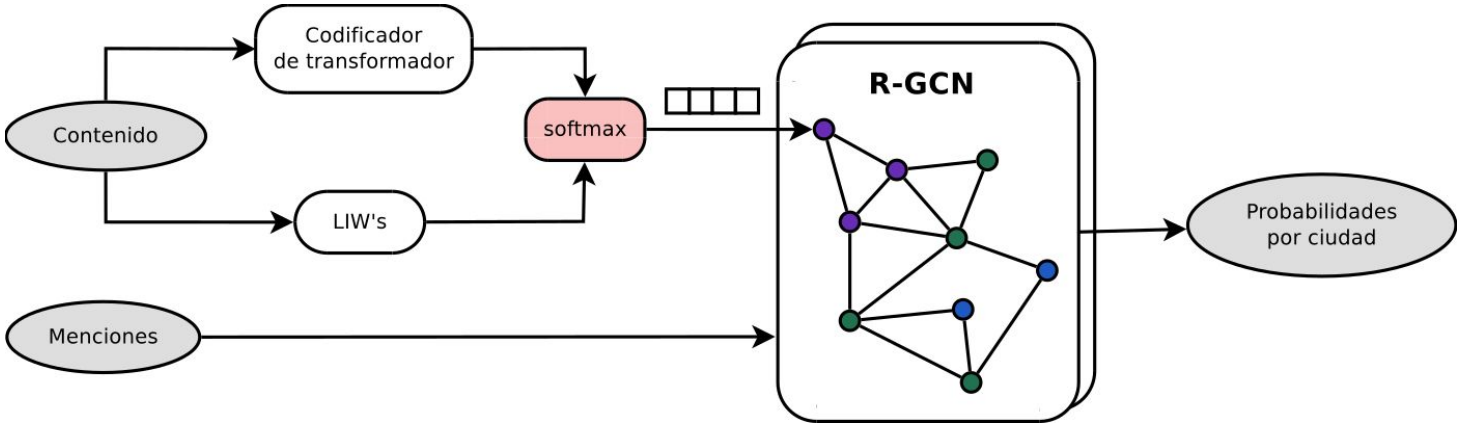
Ciudad	**LIWs principales**
Almirante Brown	‘burzaco’, ‘burzaco en’, ‘foto en almirante’, ‘almirante’, ‘en claypole’, ‘en jose marmol’
Villa Gesell	‘en gesell’, ‘villa gesell con’, ‘villa gesell 2019’, ‘en carilo’, ‘villa gesell hoy’, ‘dixit’, ‘pueblo limite’, ‘le brique oficial’, ‘foto en villa’
Santa Fé	‘santa fe mi’, ‘en esperanza santa’, ‘estanislao’, ‘norte salta’, ‘estanislao lopez’, ‘estadio brigadier general’, ‘santa fe no’, ‘santa fe acaba’, ‘en santa fe’

4. MODELO DE APRENDIZAJE

- Transformer:
    - El contenido producido por cada usuario se pasa por el codificador de un *transformer*, para obtener una representación reducida del mismo.
    - Esta codificación y las LIW's se concatenan para entrenar una regresión logística que dará una primera estimación de la ubicación del usuario.
    - La arquitectura del *transformer* permite obtener una buena codificación del contenido en mucho menor tiempo que utilizando redes recurrentes (RNN's).

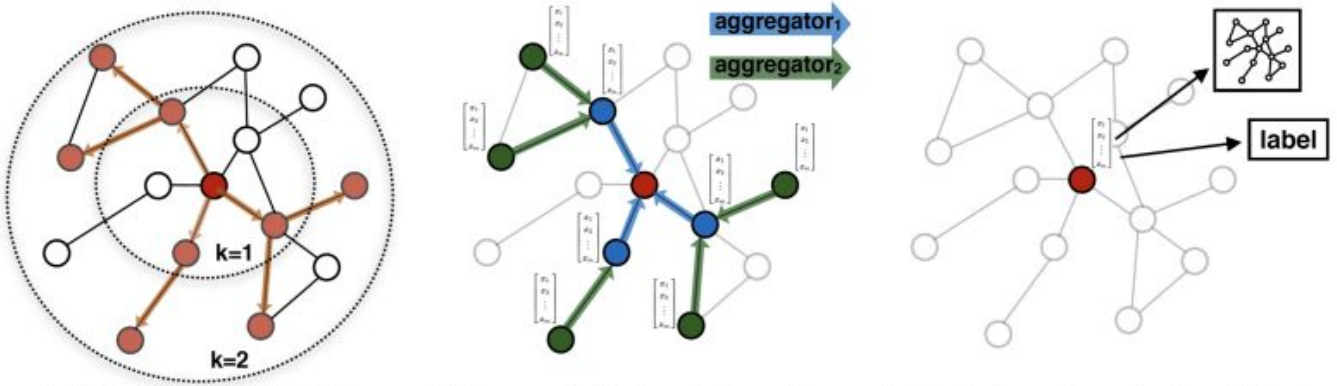
4. MODELO DE APRENDIZAJE

- Entrenamiento de la red convolucional:
  - La estimación producida por la regresión logística se utiliza como vector de features del usuario.
  - Esta información se propaga a través de una red convolucional de grafos (GNN).
  - Evaluamos dos tipos de GNN's: GraphSAGE y R-GCN.



4. MODELO DE APRENDIZAJE

- GraphSAGE **[HAMILTON ET AL., 2017]**
  - Modelo inductivo → permite aplicar el modelo a nuevos nodos
  - Propaga información aprendiendo una función agregadora que se aplica a los nodos vecinos.



1. Sample neighborhood

2. Aggregate feature information from neighbors

3. Predict graph context and label using aggregated information

*(extraído de **[HAMILTON ET AL., 2017]**)*

4. MODELO DE APRENDIZAJE

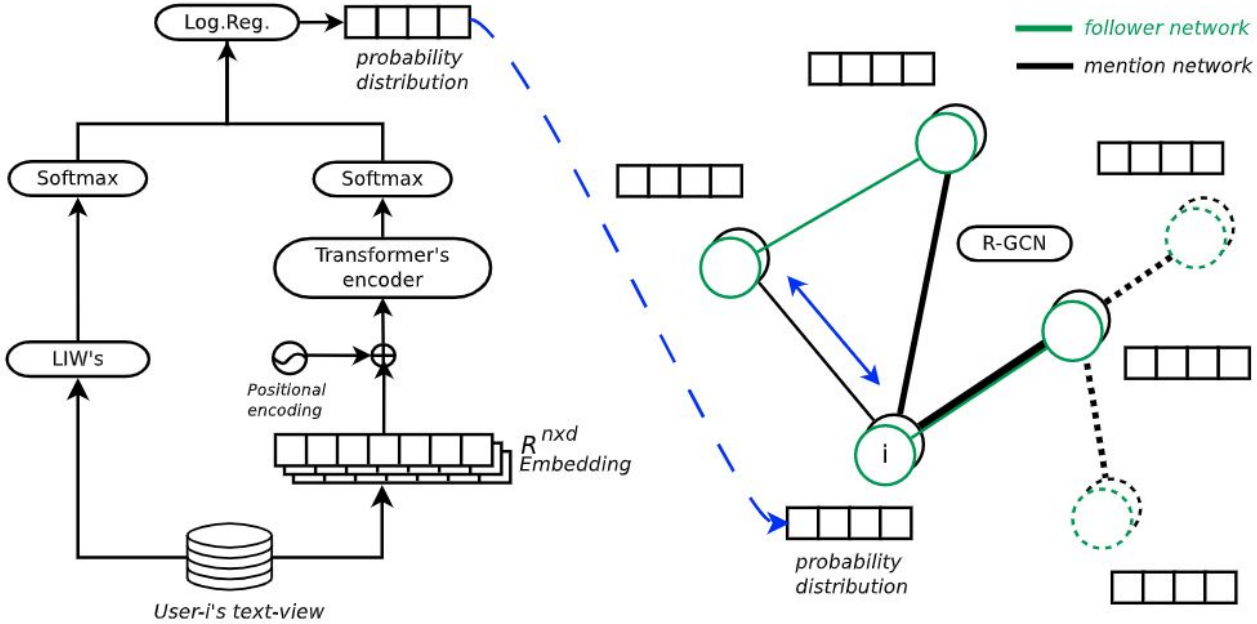
- R-GCN **[SCHLICHTKRULL ET AL., 2018]:**

  - Son una extensión de las GCN tradicionales y permiten trabajar con nodos que presentan distintos tipos de interrelaciones
  - La red convolucional así obtenida es una red multicapa (multilayer).

$$h_i^{(l+1)} = \delta\left(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_i^r}\frac{1}{c_{i,r}}W_r^{(l)}h_j^{(l)} + W_0^{(l)}h_i^{(l)} + b\right)$$

4. MODELO DE APRENDIZAJE

- Esquema completo:



5. RESULTADOS

Modelo	Acc.	Acc@100	Media Err. (km)	Mediana Err.(km)
Twitter-ARG-Exact (ciudades)				
RGCN-EXT-Preds	73,1%	82,9%	367,5	3,8
GraphSAGE-EXT-Preds	72,7%	82,3%	369,1	3,9
Twitter-ARG-Bbox (ciudades)				
RGCN-EXT-Preds	48,6%	64,4%	690,6	6,4
GraphSAGE-EXT-Preds	46,3%	57,8%	798,4	23,2

6. CONTRIBUCIONES

- Evaluamos distintos métodos basados en grafos que permiten aprovechar la información del grafo de menciones y del grafo de seguidores a través de distintos mecanismos.
- Propusimos una metodología para extender los grafos de menciones y de seguidores sin generar un exceso de conexiones.
- Mostramos que es factible utilizar datos de *bounding box* para el entrenamiento y un enfoque inductivo para la red convolucional.
- Pusimos a disposición dos nuevos datasets con un total de 150M de *tweets*, centrado en Argentina*.

* https://github.com/fedefunes96/twitter-location-data

PREGUNTAS

EXTRA: RESULTADOS TWITTER-US

Modelo	Acc@100	Media Err.(km)	Mediana Err.(km)
Twitter-US (ciudades)			
RGCN-EXT-Preds	66,6%	408,7	43,3
GraphSAGE-EXT-Preds	64,3%	432,1	50,3
(Huang and Carley, 2019)	**70,8%**	**361,5**	**31,6**
Twitter-US (nodos)			
RGCN-EXT-Preds	**67,0%**	**384,0**	57,2
GraphSAGE-EXT-Preds	63,2%	424,3	70,2
(Rahimi et al., 2018)	66,0%	420,0	**56,0**