.UBA fiuba
FACULTAD DE INGENIERÍA

Desarrollo y evaluación de una metodología para la ubicación de usuarios de Twitter

Funes Federico
Padrón: 98372

Director: Dr. Ing. Beiró Mariano
Co-director: Dr. Ing. Alvarez-Hamelin José Ignacio

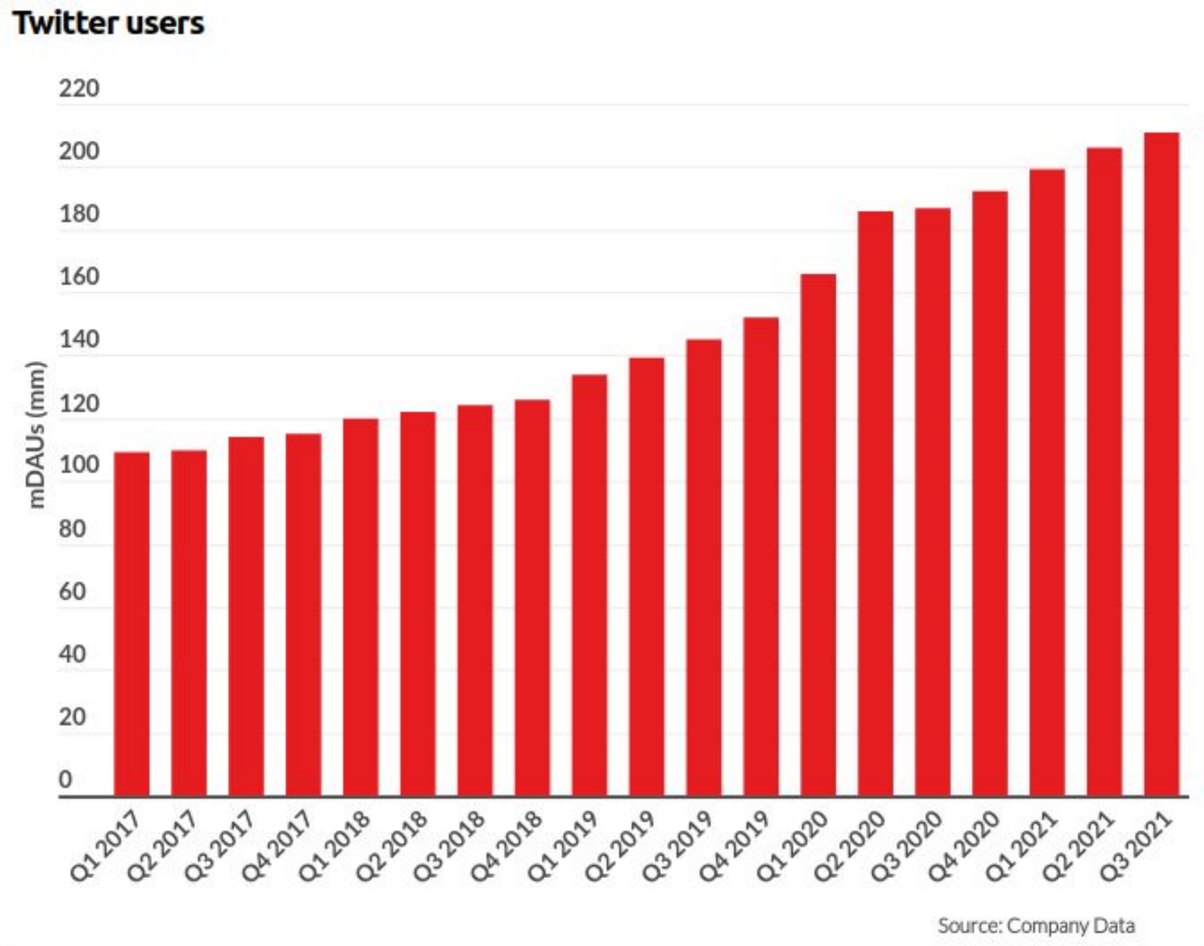
15/12/2021

Contexto

- Gran crecimiento de las redes sociales en los últimos años

Contexto

- Gran crecimiento de las redes sociales en los últimos años



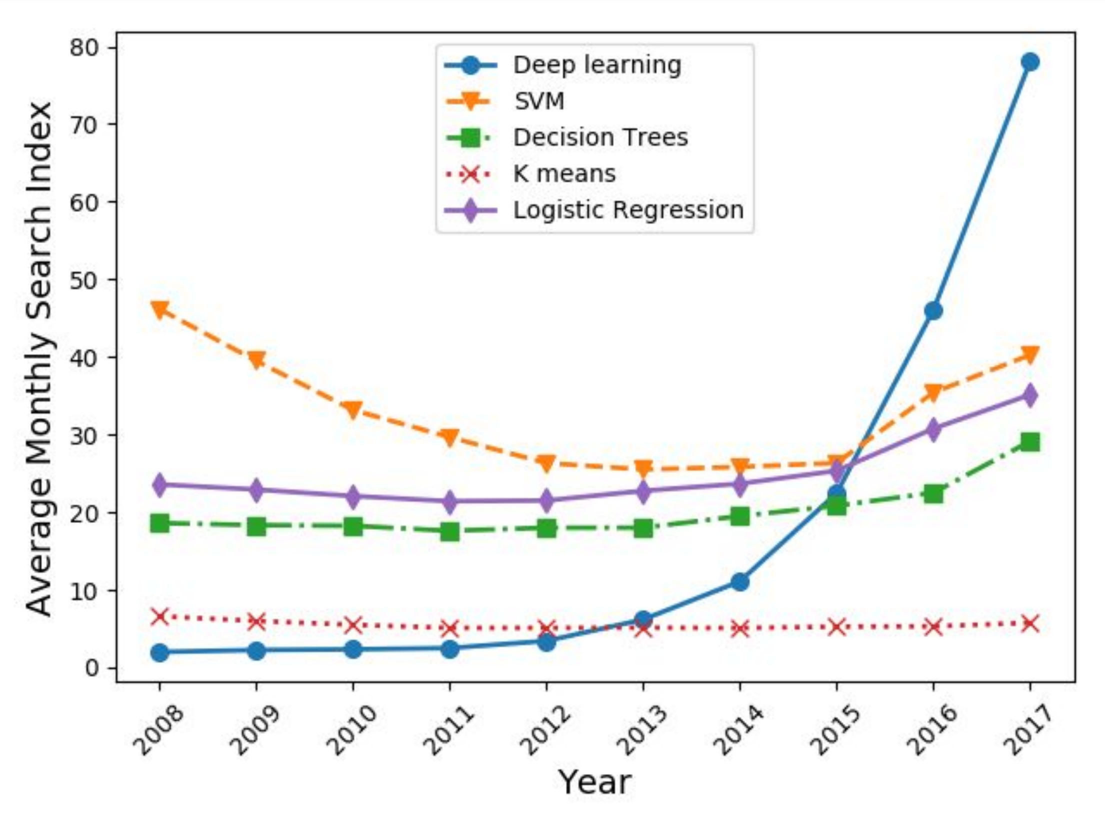
Contexto

- Gran crecimiento de las redes sociales en los últimos años

- Mucha información pública

Contexto

- Gran crecimiento de las redes sociales en los últimos años
- Mucha información pública
- Mayor capacidad de cómputo



Motivación

- Monitoreo de la salud pública

Monitoreo de síntomas y dolencias generales

- *"You are what you tweet: Analyzing twitter for public health", Fifth international AAAI conference on weblogs and social media, 2011*

Pronóstico de la incidencia del virus del Zika

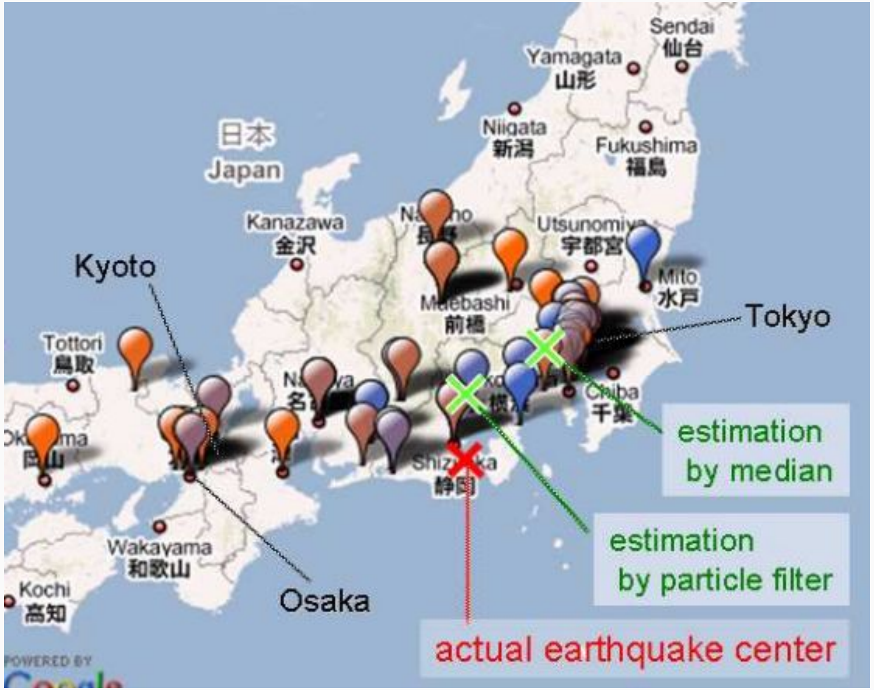
- *"Forecasting Zika incidence in the 2016 Latin America outbreak combining traditional disease surveillance with search, social media, and news report data", PLoS neglected tropical diseases, 2017*

Motivación

- Monitoreo de la salud pública
- Informar sobre catástrofes locales

Detección de terremotos en tiempo real

- *"Earthquake shakes twitter users: real-time event detection by social sensors", Proceedings of the 19th international conference on World wide web, pages 851-860, 2010*



Motivación

- Monitoreo de la salud pública
- Informar sobre catástrofes locales

Detección de terremotos en tiempo real

- *"Earthquake shakes twitter users: real-time event detection by social sensors", Proceedings of the 19th international conference on World wide web, pages 851-860, 2010*

Incendios forestales

- *"OMG, from here, I can see the flames!" a use case of mining location based social networks to acquire spatio-temporal data on forest fires", Proceedings of the 2009 international workshop on location based social networks, pages 73-80, 2009*

Inundaciones

-*"Early flood detection for rapid humanitarian response: harnessing near real-time satellite and Twitter signals", ISPRS International Journal of Geo-Information, pages 2246-2266, 2015*

Motivación

- Monitoreo de la salud pública
- Informar sobre catástrofes locales
- Recomendación de contenido
- Análisis de comunidades y su comportamiento

Terminología

- Tweets
- Menciones
- Hashtags
- Geolocalización
- *Bounding box*
- Perfil

Problemáticas

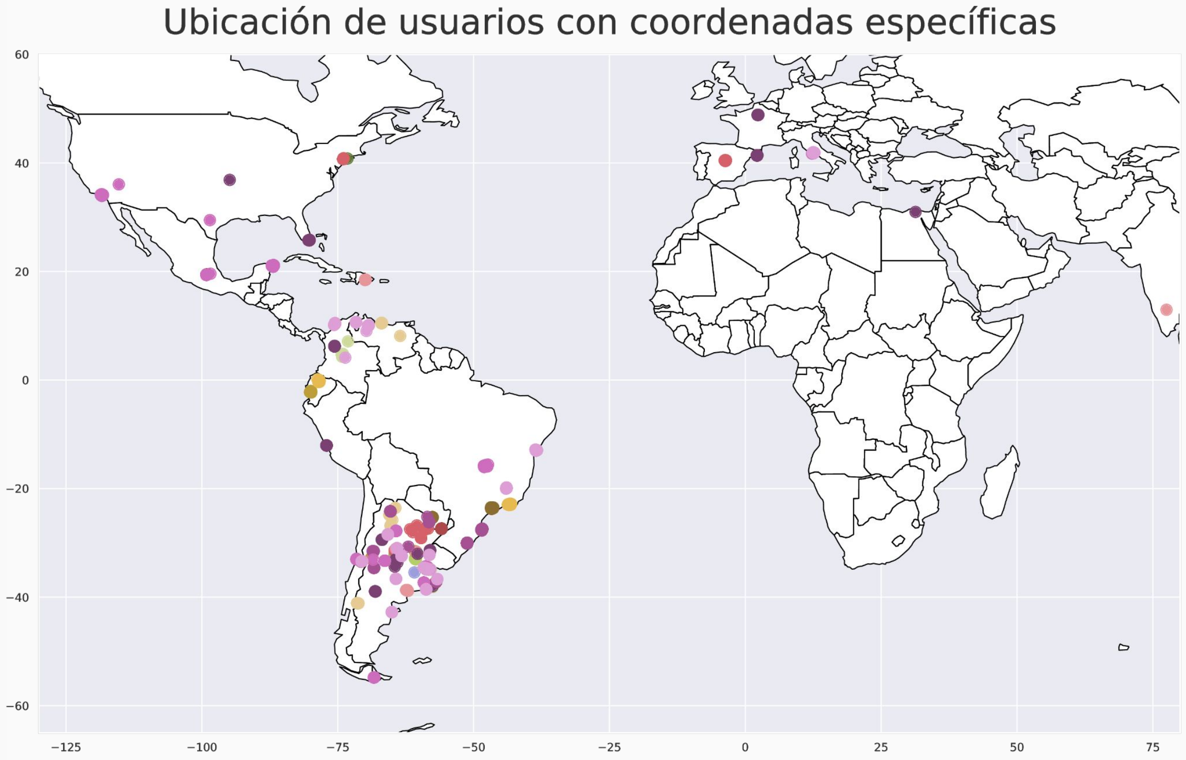
- Tweets geolocalizados < 1%
- Perfiles con ubicación falsa

Alternativas:

- Analizar perfiles
- Analizar publicaciones
- Analizar relaciones

Problemáticas

- Tweets geolocalizados < 1%
- Perfiles con ubicación falsa
- Desbalance de muestras por ciudad



Estado del arte - Cheng 2010

- Local Indicative Words (LIW)
- Nueva métrica: Acc@161
- Solo contenido de tweets

You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users

Zhiyuan Cheng
Department of Computer Science and Engineering
Texas A&M University
College Station, TX, USA
zcheng@cse.tamu.edu

James Caverlee
Department of Computer Science and Engineering
Texas A&M University
College Station, TX, USA
caverlee@cse.tamu.edu

Kyumin Lee
Department of Computer Science and Engineering
Texas A&M University
College Station, TX, USA
kyumin@cse.tamu.edu

ABSTRACT

We propose and evaluate a probabilistic framework for estimating a Twitter user's city-level location based purely on the content of the user's tweets, even in the absence of any other geospatial cues. By augmenting the massive human-powered sensing capabilities of Twitter and related microblogging services with content-derived location information, this framework can overcome the sparsity of geo-enabled features in these services and enable new location-based personalized information services, the targeting of regional advertisements, and so on. Three of the key features of the proposed approach are: (i) its reliance purely on tweet content, meaning no need for user IP information, private login information, or external knowledge bases; (ii) a classification component for automatically identifying words in tweets with a strong local geo-scope; and (iii) a lattice-based neighborhood smoothing model for refining a user's location estimate. The system estimates k possible locations for each user in descending order of confidence. On average we find that the location estimates converge quickly (needing just 100s of tweets), placing 51% of Twitter users within 100

approximately 75 million users as of 2010 [4]. These users actively publish short messages ("tweets") of 140 characters or less to an audience of their subscribers ("followers"). With such a large geographically diverse user base, Twitter has essentially published terabytes of real-time "sensor" data in the form of these status updates.

Mining this people-centric sensor data promises new personalized information services, including local news summarized from tweets of nearby Twitter users [21], the targeting of regional advertisements, spreading business information to local customers [3], and novel location-based applications (e.g., Twitter-based earthquake detection, which can be faster than through traditional official channels [18]).

Unfortunately, Twitter users have been slow to adopt geospatial features: in a random sample of over 1 million Twitter users, only 26% have listed a user location as granular as a city name (e.g., Los Angeles, CA); the rest are overly general (e.g., California), missing altogether, or nonsensical (e.g., Wonderland). In addition, Twitter began supporting per-tweet geo-tagging in August 2009. Unlike user location (which is a single location associated with a user and listed

test set are geo-located within 100 miles to their real locations and the **AvgErrDist** is 1,773 miles, meaning that such a baseline content-based location estimator provides little value. On inspection, we discovered two key problems: (i) most words are distributed consistently with the population across different cities, meaning that most words provide very little power at distinguishing the location of a user; and (ii) most cities, especially with a small population, have a sparse set of words in their tweets, meaning that the per-city word distributions for these cities are underspecified leading to large estimation errors.

In the rest of this section, we address these two problems in turn in hopes of developing a more valuable and refined location estimator. Concretely, we pursue two directions:

- *Identifying Local Words in Tweets:* Is there a subset of words which have a more compact geographical scope compared to other words in the dataset? And can these "local" words be discovered from the content of tweets? By removing noise words and non-local words, we may be able to isolate words that can distinguish users located in one city versus another.
- *Overcoming Tweet Sparsity:* In what way can we overcome the location sparsity of words in tweets? By exploring approaches for smoothing the distributions of words, can we improve the quality of user location estimation by assigning non-zero probability for words to be issued from cities in which we have no word observations?

4.1 Identifying Local Words in Tweets

Our first challenge is to filter the set of words considered by the location estimation algorithm (Algorithm 1) to consider primarily words that are essentially "local". By considering all words in the location estimator, we saw how the performance suffers due to the inclusion of noise words that do not convey a strong sense of location (e.g., "august",

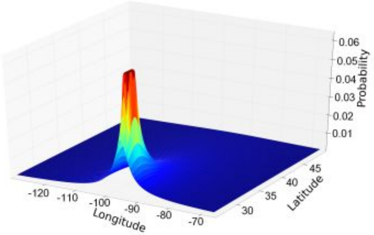


Figure 3: Optimized Model for the Word "rockets"

Recently Backstrom et al. introduced a model of spatial variation for analyzing the geographic distribution of terms in search engine query logs [8]. The authors propose a generative probabilistic model in which each query term has a geographic focus on a map (based on an analysis of the IP-address-derived locations of users issuing the query term). Around this center, the frequency shrinks as the distance from the center increases. Two parameters are assigned for each model, a constant C which identifies the frequency in the center, and an exponent α which controls the speed of how fast the frequency falls as the point goes further away from the center. The formula for the model is $Cd^{-\alpha}$ which means that the probability of the query issued from a place with a distance d from the center is approximately $Cd^{-\alpha}$. In the model, a larger α identifies a more compact geo-scope of a word, while a smaller α displays a more global popular distribution.

Estado del arte - Rahimi 2015-2018

- Red de *comenciones*
- Clases usando hojas de un KD-Tree con un mínimo de muestras
- Redes neuronales basadas en grafos (GCN), no muy buenos resultados

Twitter User Geolocation Using a Unified Text and Network Prediction Model

Afshin Rahimi, Trevor Cohn, and **Timothy Baldwin**
Department of Computing and Information Systems
The University of Melbourne
arahimi@student.unimelb.edu.au
{t.cohn,tbaldwin}@unimelb.edu.au

Abstract

We propose a label propagation approach to geolocation prediction based on Modified Adsorption, with two enhancements: (1) the removal of "celebrity" nodes to increase location homophily and boost tractability; and (2) the incorporation of text-based geolocation priors for test users. Experiments over three Twitter benchmark datasets achieve state-of-the-art results, and demonstrate the effectiveness of the enhancements.

1 Introduction

Geolocation of social media users is essential in applications ranging from rapid disaster response (Earle et al.,

Most previous research on user geolocation has focused either on text-based classification approaches (Eisenstein et al., 2010; Wing and Baldridge, 2011; Roller et al., 2012; Han et al., 2014) or, to a lesser extent, network-based regression approaches (Jurgens, 2013; Compton et al., 2014; Rahimi et al., 2015). Methods which combine the two, however, are rare.

In this paper, we present our work on Twitter user geolocation using both text and network information. Our contributions are as follows: (1) we propose the use of Modified Adsorption (Talukdar and Crammer, 2009) as a baseline network-based geolocation model, and show that it outperforms previous network-based approaches (Jurgens, 2013; Rahimi et al., 2015); (2) we demonstrate that removing "celebrity" nodes (nodes with high in-

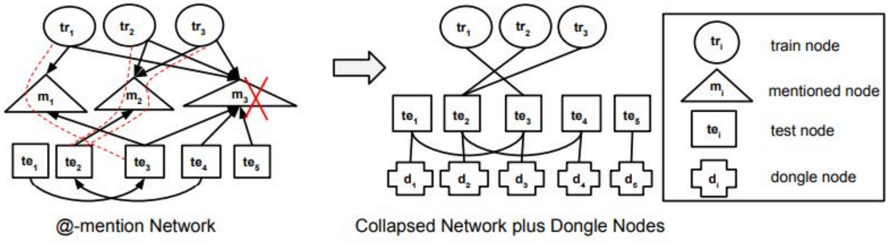


Figure 1: A collapsed network is built from the @-mention network. Each mention is shown by a directed arrow, noting that as it is based exclusively on the tweets from the training and test users, it will always be directed from a training or test user to a mentioned node. All mentioned nodes which are not a member of either training or test users are removed and the corresponding training and test users, previously connected through that node, are connected directly by an edge, as indicated by the dashed lines. Mentioned nodes with more than T unique mentions (celebrities, such as m_3) are removed from the graph. To each test node, a dongle node that carries the label from another learner (here, text-based LR) is added in MADCEL-B-LR and MADCEL-W-LR.

that the following objective function is minimised:

$$C(\hat{Y}) = \sum_l \Big[\mu_1 (Y_l - \hat{Y}_l)^T S (Y_l - \hat{Y}_l) + \mu_2 \hat{Y}_l^T L \hat{Y}_l\Big]$$

which we rewrite with the centre of the map). We redress these two issues as follows.

Celebrity Removal To address the first issue, we target "celebrity" users, i.e. highly-mentioned Twitter users. Edges involving these users often carry little or

Estado del arte - Huang & Carley 2019

- *Embeddings* para cada *feature*
- Clases usando ciudades colapsadas > 100.000 habitantes

A Hierarchical Location Prediction Neural Network for Twitter User Geolocation

Binxuan Huang
School of Computer Science
Carnegie Mellon University
binxuanh@cs.cmu.edu

Kathleen M. Carley
School of Computer Science
Carnegie Mellon University
kathleen.carley@cs.cmu.edu

Abstract

Accurate estimation of user location is important for many online services. Previous neural network based methods largely ignore the hierarchical structure among locations. In this paper, we propose a hierarchical location prediction neural network for Twitter user geolocation. Our model first predicts the home country for a user, then uses the country result to guide the city-level prediction. In addition, we employ a character-aware word embedding layer to overcome the noisy information in tweets. With the feature fusion layer, our model can accommodate various feature

of tweet text-based methods, where the word distribution is used to estimate geolocations of users (Roller et al., 2012; Wing and Baldridge, 2011). In the second type, methods combining metadata features such as time zone and profile description are developed to improve performance (Han et al., 2013). Network-based methods form the last type. Several studies have shown that incorporating friends' information is very useful for this task (Miura et al., 2017; Ebrahimi et al., 2018). Empirically, models enhanced with network information work better than the other two types, but they do not scale well to larger datasets (Rahimi et al.,

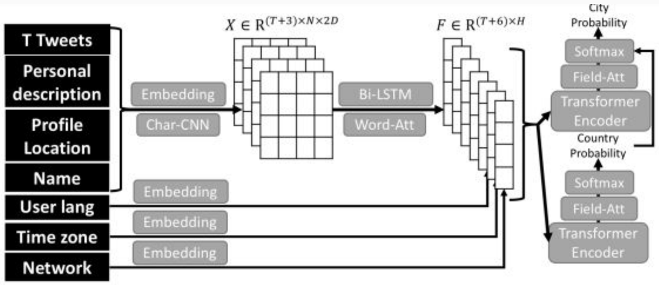


Figure 1: The architecture of our hierarchical location prediction neural network.

get one representation vector $f \in R^{2D}$ after projection by W^O for each text field.

3.3 Feature Fusion

For two categorical features, we assign an embedding vector with dimension $2D$ for each time zone and language. These embedding vectors are learned during training. We pretrain network embeddings for users involved in the mention network using LINE (Tang et al., 2015). Network embeddings are fixed during training. We get a feature matrix $F \in R^{(T+6)\times 2D}$ by concatenating

tweet text and personal description. To overcome this issue, we add feature type embeddings to the input representations F. There are seven features in total. Each of them has a learned feature type embedding with dimension $2D$ so that one feature type embedding and the representation of the corresponding feature can be summed.

Because the input and the output of transformer encoder have the same dimension, we stack L layers of transformer encoders to learn representations for country-level prediction and city-level prediction respectively. These two sets of en-

Estado del arte

- Foco en el contenido y poco análisis de redes
- No proveen explicabilidad
- No hay unificación con respecto a los *labels* -> Comparaciones injustas

Conjuntos de datos

Disponemos de 900 millones de tweets focalizados en Argentina, capturados en 2019:

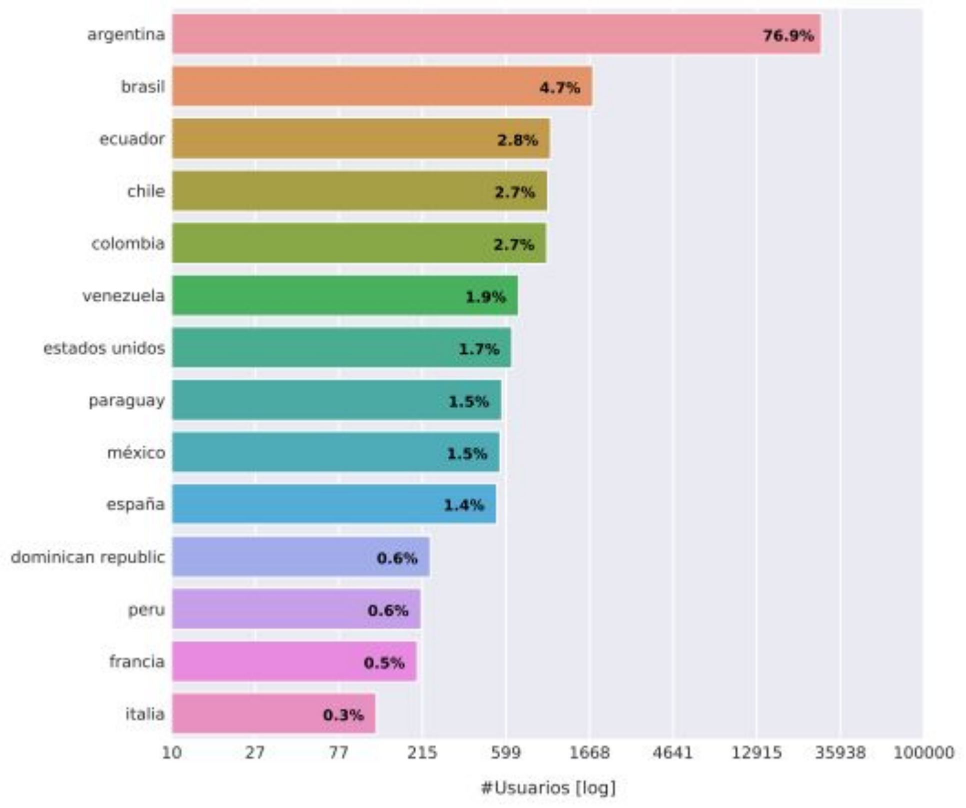
- 1 millón se encuentran geolocalizados (0.11%)
- 14 millones poseen un *bounding box*

Y aproximadamente 2 millones de usuarios con su información pública

Conjuntos de datos

Twitter-ARG-Exact

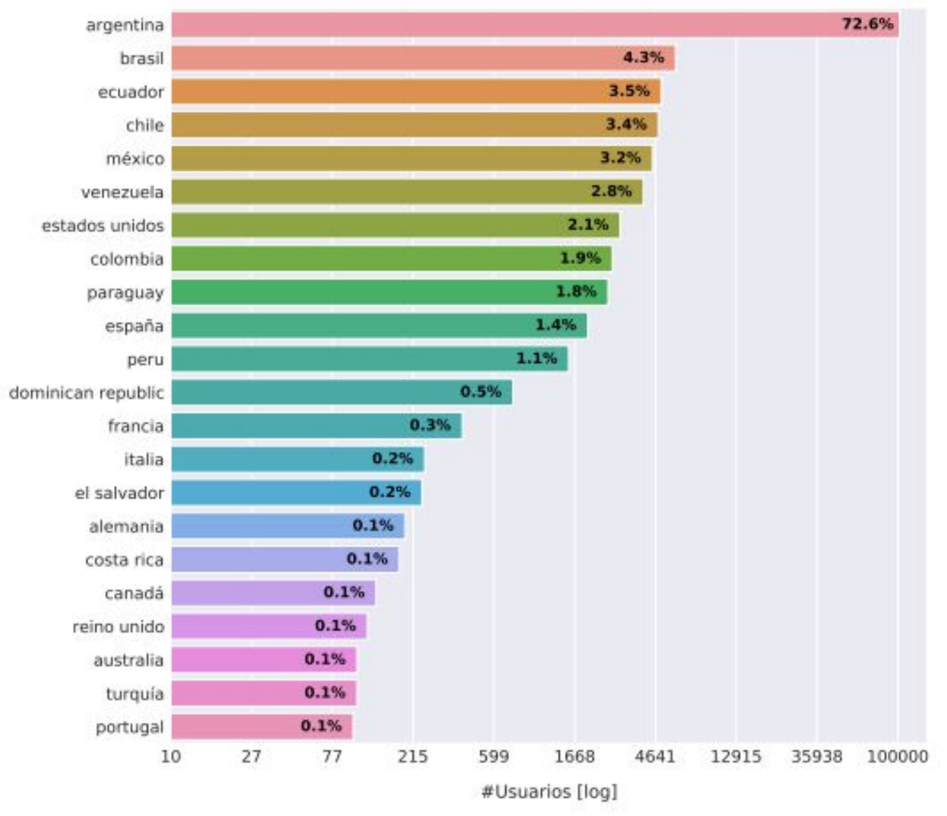
- 14 países
- 95 ciudades con más de 100 muestras
- ~37.000 usuarios
- ~600.000 tweets geolocalizados
- ~27.000.000 tweets



Conjuntos de datos

Twitter-ARG-BBox

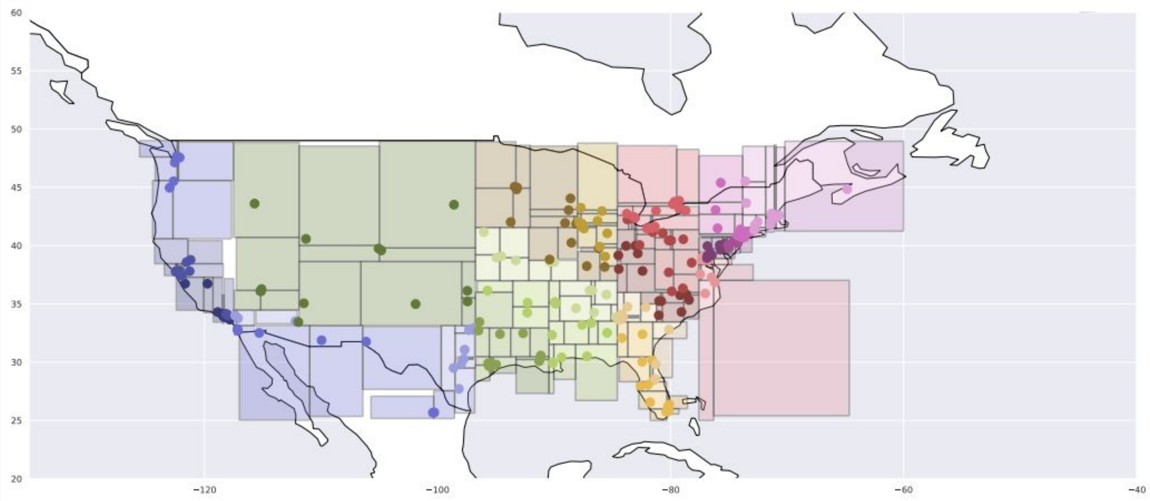
- 22 países
- 229 ciudades con más de 100 muestras
- ~140.000 usuarios
- ~9.000.000 tweets con *bounding box*
- ~124.000.000 tweets



Conjuntos de datos

Conjuntos de otros trabajos:

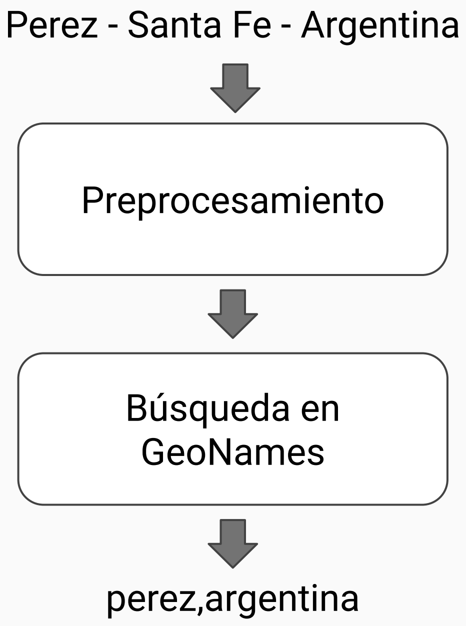
- **Twitter-US**
- Twitter-World
- W-NUT



Análisis de datos

- Análisis de perfiles: ¿Qué tan fiable es la ubicación pública de los usuarios?
- Análisis de redes conformadas por relaciones entre usuarios
- Análisis del contenido de los tweets

Análisis de perfiles



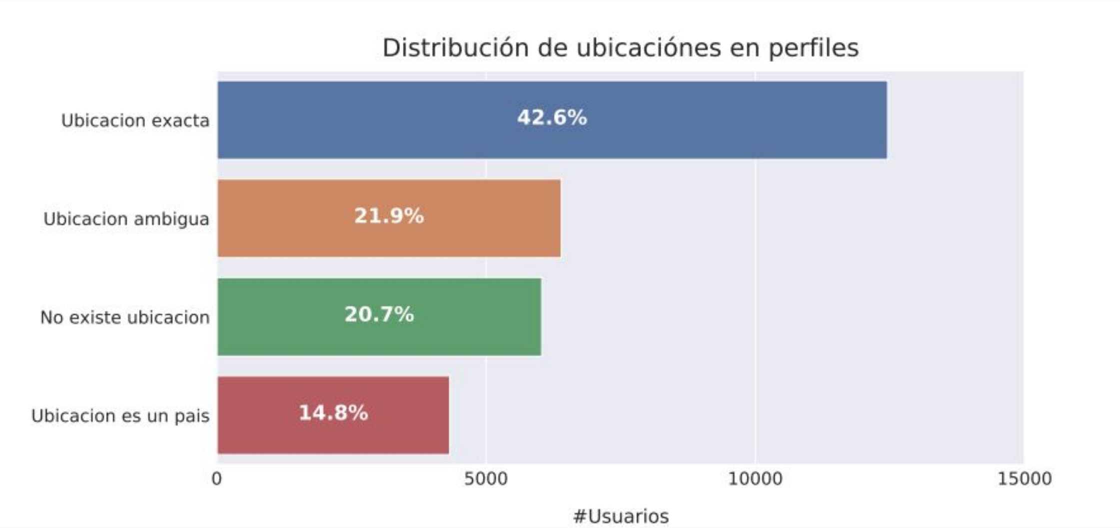
Análisis de perfiles

Perez - Santa Fe - Argentina

Preprocesamiento

Búsqueda en GeoNames

perez,argentina



Análisis de perfiles - Resumen

Podemos determinar la ubicación del:

- 69% usuarios con perfil y ubicación exacta a menos de 10km (29% del total)
- 45% usuarios con perfil y ubicación ambigua a menos de 10km (10% del total)

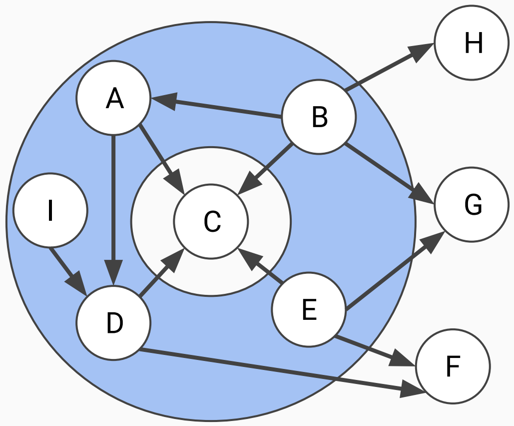
Si combinamos ambos resultados:

- 39% usuarios con perfil a menos de 10km

Análisis de redes - Menciones

Podemos trabajar con la red de:

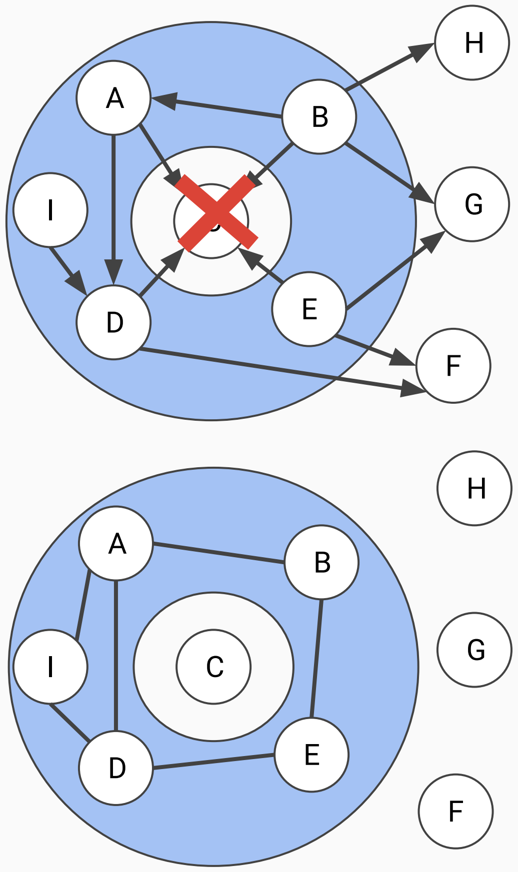
- **Menciones**:
    - Muchos nodos sin información
    - Necesidad de acotar



Análisis de redes - Menciones

Podemos trabajar con la red de:

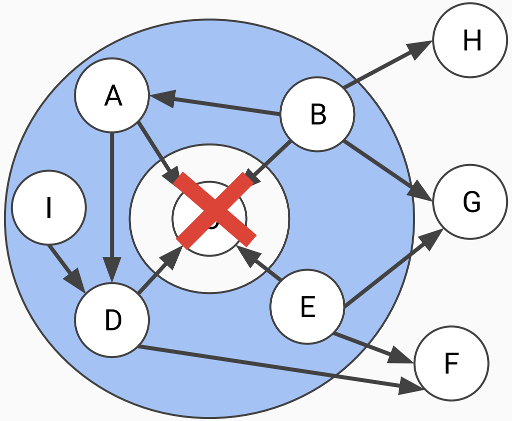
- Menciones:
    - Muchos nodos sin información
    - Necesidad de acotar
- **Comenciones**:
    - Nodos populares -> Muchas aristas
    - Caminos redundantes

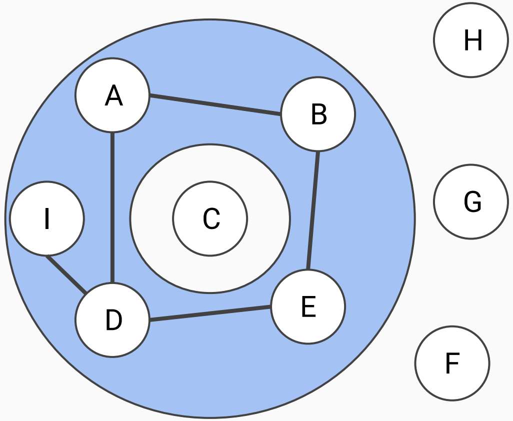


Análisis de redes - Menciones

Podemos trabajar con la red de:

- Menciones:
    - Muchos nodos sin información
    - Necesidad de acotar
- Comenciones:
    - Nodos populares -> Muchas aristas
    - Caminos redundantes
- **Menciones extendida**:
    - Nodos populares -> Muchas aristas
    - < Aristas que *Comenciones*



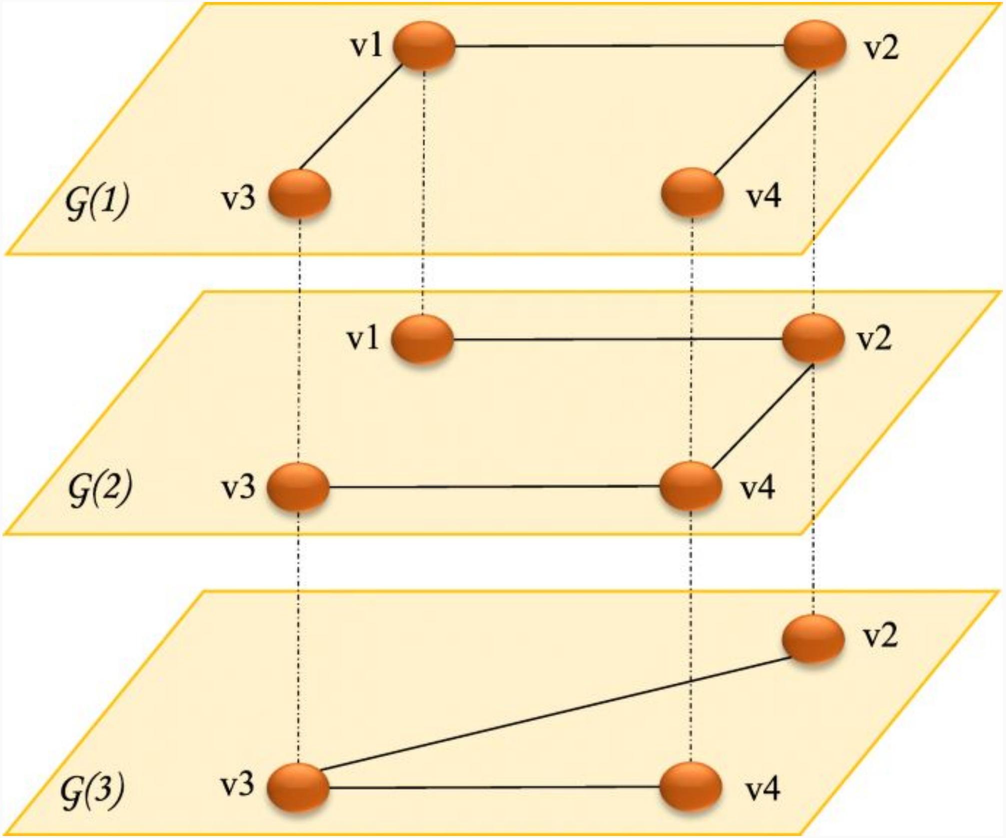


Análisis de redes - Redes multicapa

- Conservar información de las relaciones

- Múltiples formas de representación

- Escasa cantidad de métodos sobre este tipo de redes



Análisis de redes - Resumen

Grafos resultantes:

- Menciones extendido
- Multicapa menciones locales + comenciones externas
- Seguidores extendido
- …
- Multicapa menciones locales + comenciones externas + seguidores locales + coseguidores externos

Análisis de contenido - Preprocesamiento

Tweet original	Tweet procesado	Hashtags
At work #trabajo #toys #juguetes #jugueteria #buenosaires en Ciudad Autónoma de Buenos Aires https://t.co/gud6uoILmQ	at work en ciudad autónoma de buenos aires	trabajo, toys, juguetes, jugueteria, buenosaires
Feliz Aniversario ♡ 23 años #bodasdeagua que sigamos navegando juntos en Cartagena, Colombia https://t.co/8TWfnP3gj4	feliz aniversario 23 años que sigamos navegando juntos en cartagena colombia	bodasdeagua

Análisis de contenido - Local Indicative Words (LIW)

Búsqueda de términos locales (LIW):

Seleccionar *K* términos más significativos por ciudad:

- Pruebas $\chi 2$
- Información mutua

Un término puede ser un unigrama, bigrama o trigrama

Análisis de contenido - Local Indicative Words (LIW)

Búsqueda de términos locales (LIW):

Seleccionar *K* términos más significativos por ciudad:

- Pruebas $\chi 2$
- Información mutua

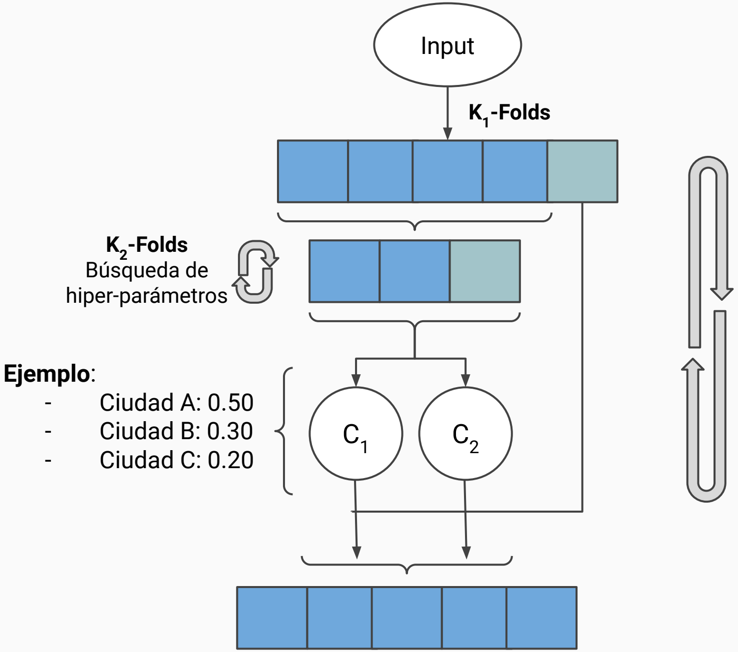
Villa Gesell, Argentina	'en gesell', 'villa gesell con', 'villa gesell 2019', 'en carilo', 'villa gesell hoy', 'dixit', 'pueblo limite', 'le brique oficial', 'foto en villa'

Un término puede ser un unigrama, bigrama o trigrama

Modelo de evaluación

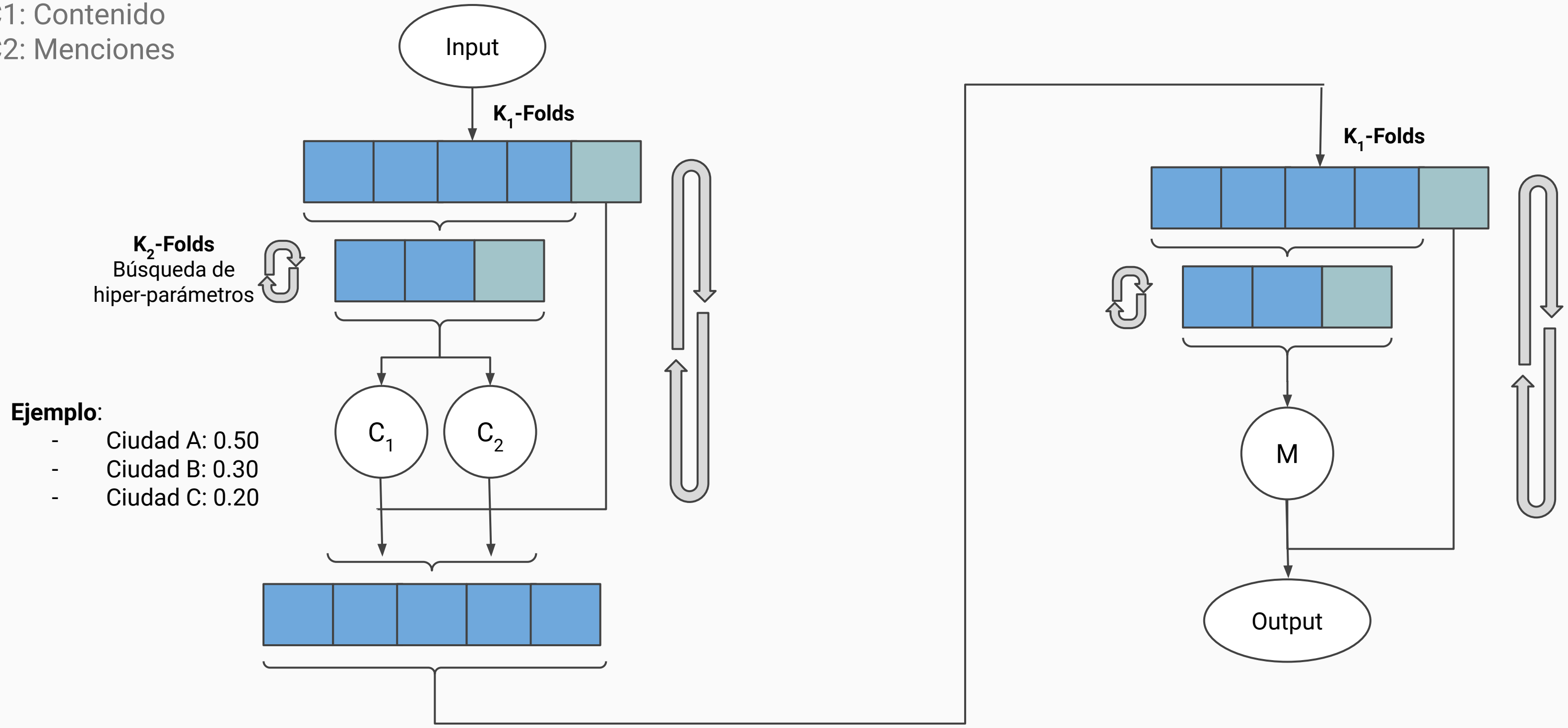
Nested cross validation:

- Búsqueda de hiper-parámetros
- Toda muestra se aprovecha tanto en entrenamiento como en pruebas
- $K_1 . K_2$ entrenamientos por clasificador

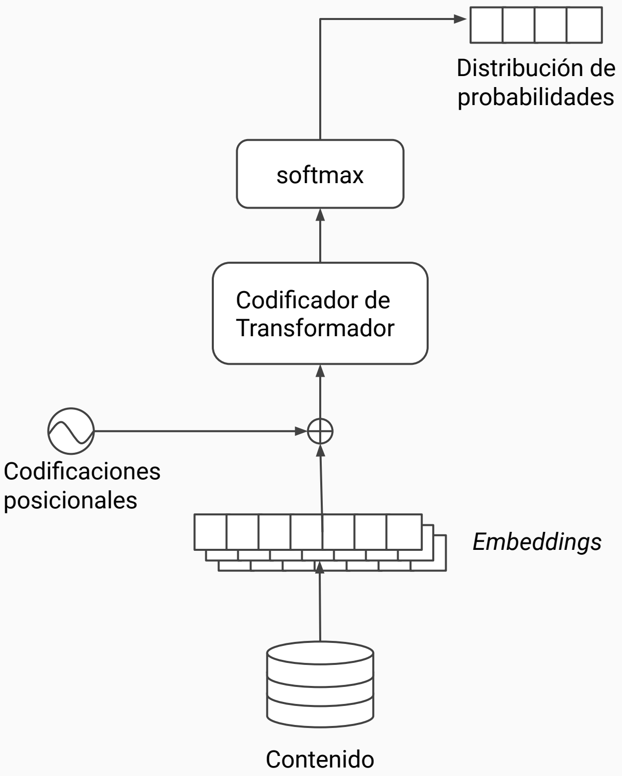


Modelos propuestos - Stacking Meta-clasificador

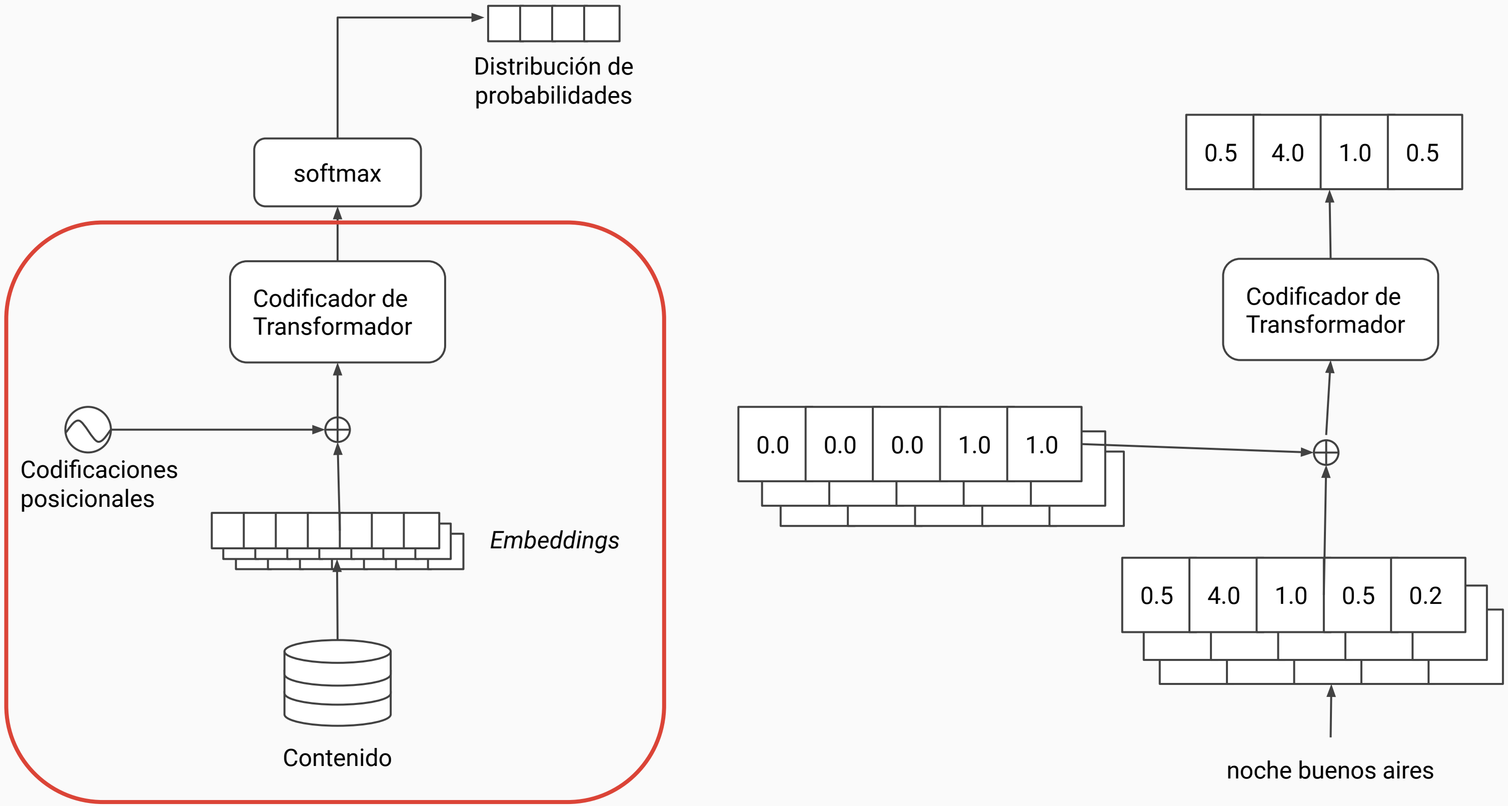
- C1: Contenido
- C2: Menciones



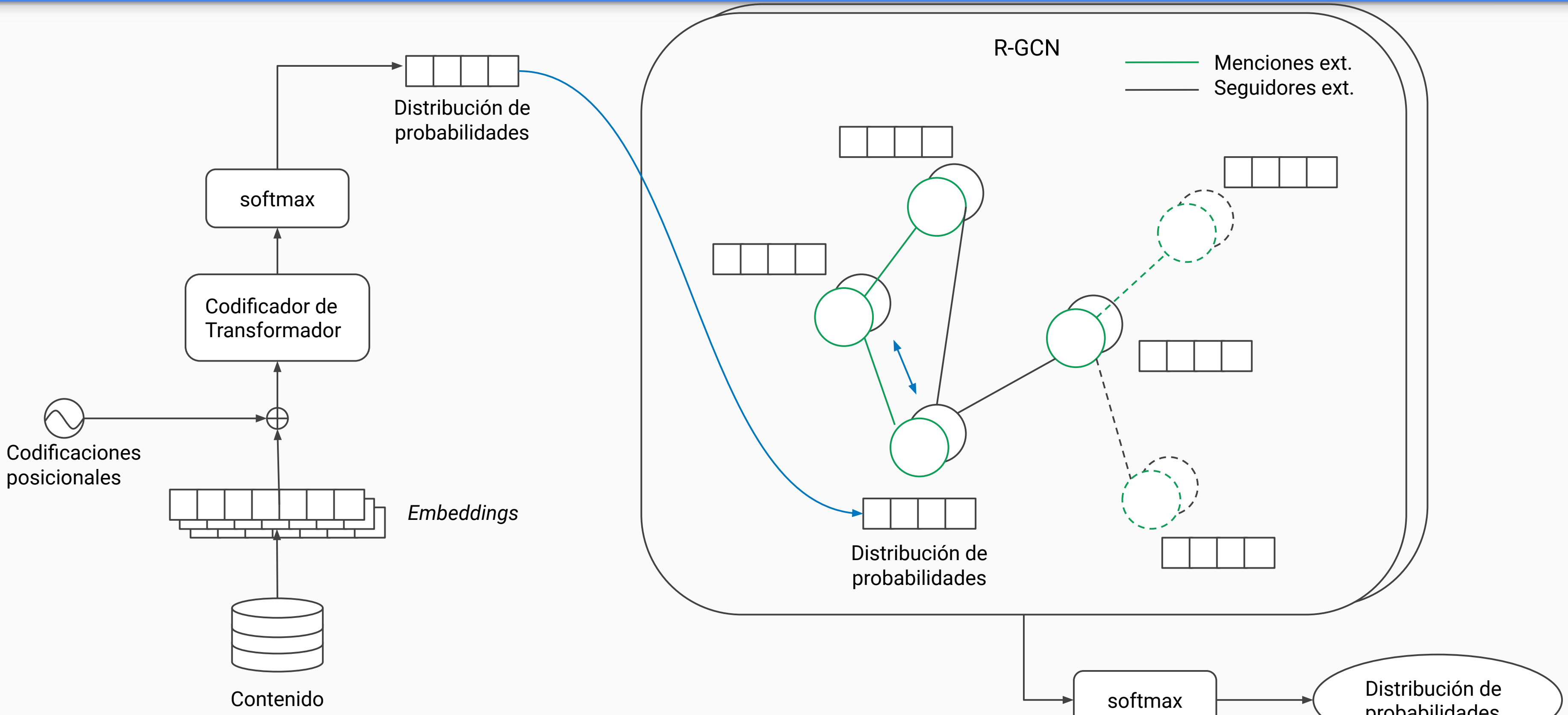
Modelos propuestos - R-GCN



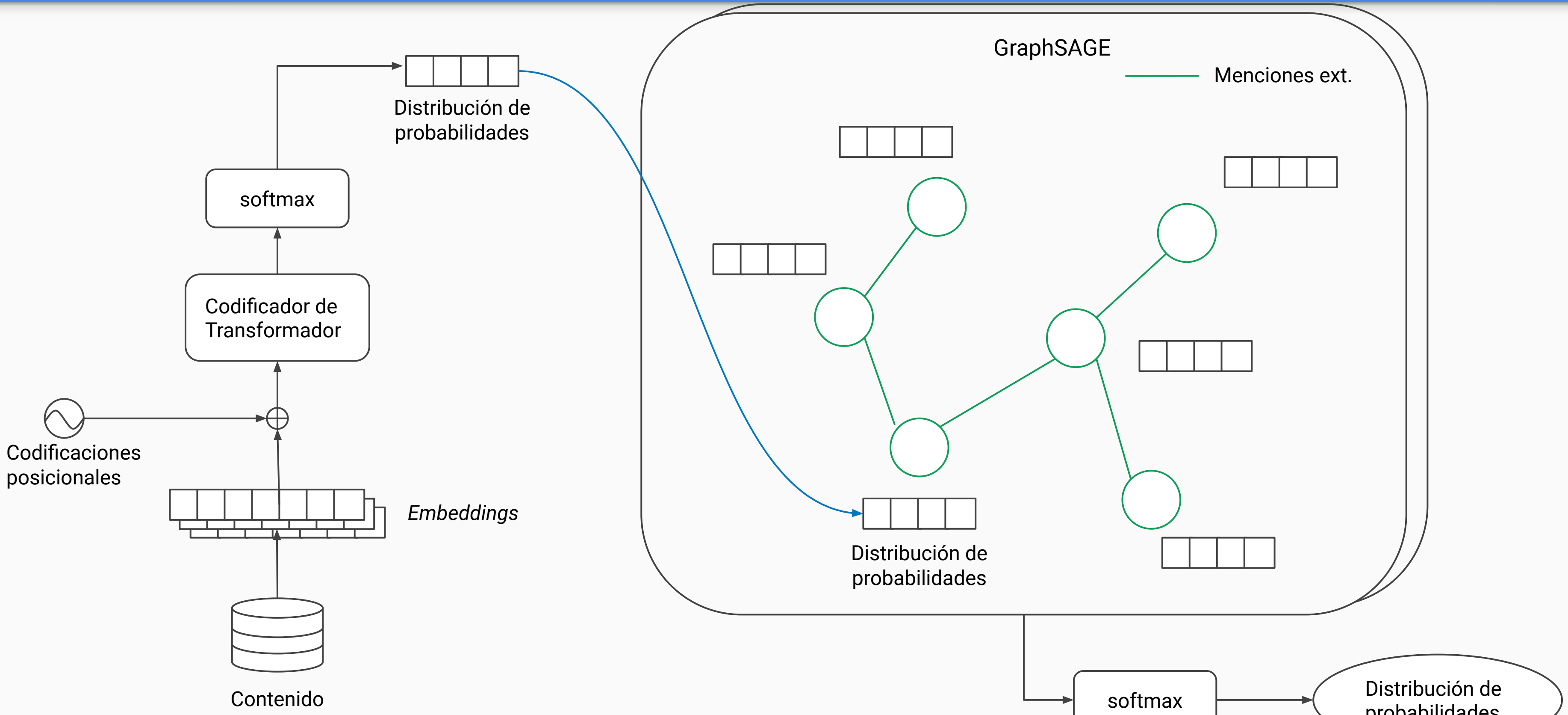
Modelos propuestos - R-GCN



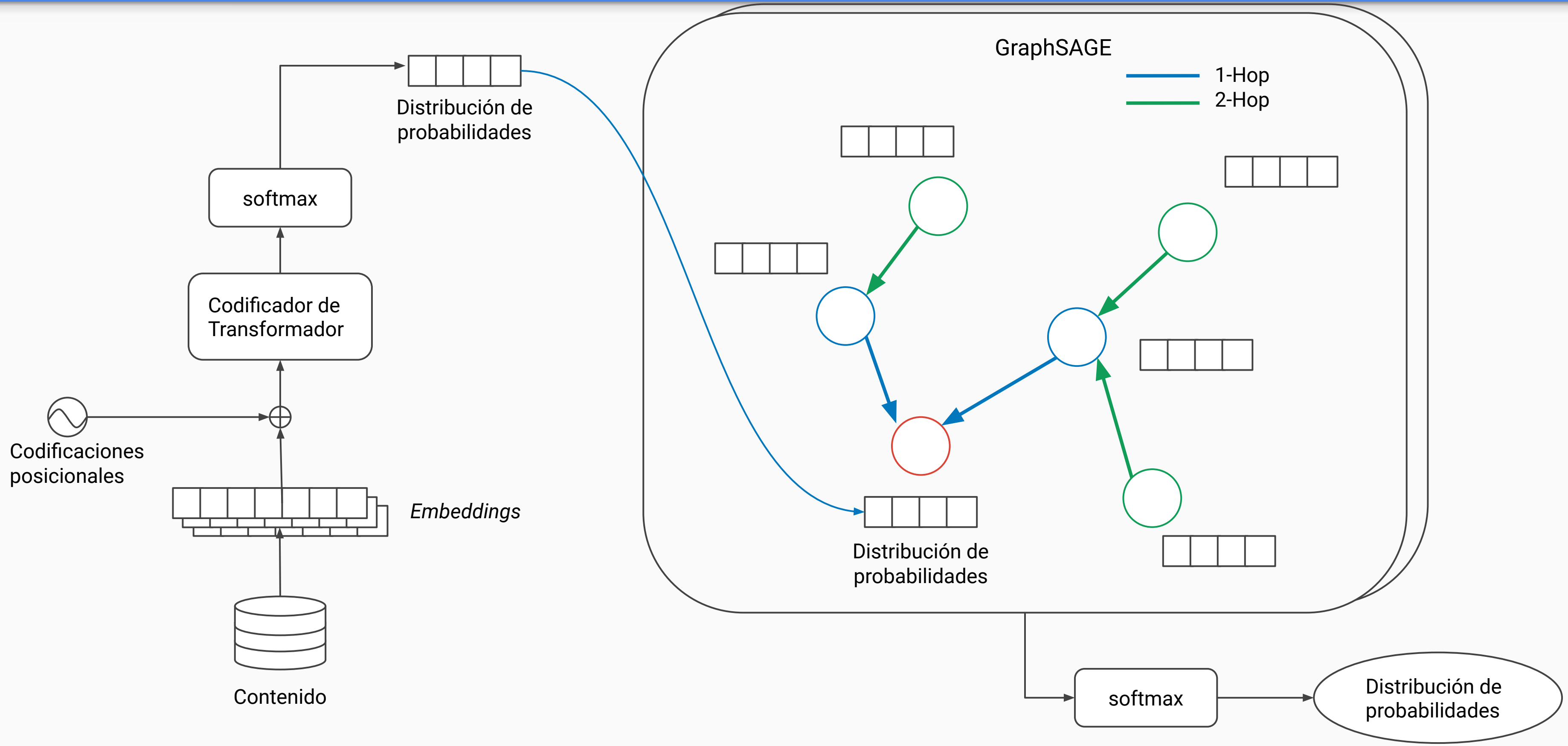
Modelos propuestos - R-GCN



Modelos propuestos - GraphSAGE



Modelos propuestos - GraphSAGE



Resultados - Twitter-ARG-Exact - Contenido

Método	Acc	Acc@161	Error prom. (Km)	Error mediana (Km)
Naive Bayes sobre los términos	59.2%	71.0%	716.8	5.4
Regresor logístico sobre top K términos más significativos por ciudad	**67.2%**	**76.9%**	524.8	**4.7**
LSTM bidireccional sobre los términos	60.4%	75.0%	547.5	5.2
Codificador de Transformador sobre los términos	65.8%	76.4%	**522.5**	4.9

Resultados - Twitter-ARG-Exact - Menciones

Método	**Acc**	**Acc@161**	**Error prom. (Km)**	**Error mediana (Km)**
Naive Bayes sobre grafo de menciones	38.7%	58.5%	925.5	24.5
OSLOM sobre grafo de menciones extendido	42.6%	62.0%	811.2	18.2
Node2vec sobre grafo de menciones extendido + Regresor logístico	**46.0%**	**66.0%**	**711.2**	**13.5**

Resultados - Twitter-ARG-Exact - Seguidores

Método	**Acc**	**Acc@161**	**Error prom. (Km)**	**Error mediana (Km)**
Naive Bayes sobre grafo de seguidores	36.8%	57.6%	904.3	25.4
OSLOM sobre grafo de seguidores extendido	26.1%	43.0%	1417.5	386.5
Node2vec sobre grafo de seguidores extendido + Regresor logístico	**39.2%**	**59.9%**	**865.5**	**21.7**

Resultados - Twitter-ARG-Exact - Completo

Método	**Acc**	**Acc@161**	**Error prom. (Km)**	**Error mediana (Km)**
Menciones + contenido				
RGCN sobre grafo de menciones extendido	**73.1%**	82.9%	**367.5**	**3.8**
GraphSage sobre grafo de menciones extendido	72.7%	82.3%	369.1	3.9
Meta clasificador + mejores clasificadores de cada *feature*	72.2%	82.7%	374.9	3.9
Menciones + seguidores + contenido				
RGCN sobre grafo multicapa	71.7%	82.3%	380.2	4.0
Meta clasificador + mejores clasificadores de cada *feature*	72.7%	**83.2%**	371.4	3.9

Resultados - Twitter-ARG-BBox

Método	Acc	Acc@161	Error prom. (Km)	Error mediana (Km)
Contenido				
Codificador de Transformador sobre los términos	39.7%	53.2%	914.7	51.6
Redes				
Node2vec sobre grafo de menciones extendido + Regresor logístico	35.7%	51.8%	1032.6	66.8
Node2vec sobre grafo de seguidores extendido + Regresor logístico	36.9%	53.2%	1023.2	51.6
Menciones + seguidores + contenido				
Meta clasificador + mejores clasificadores de cada *feature*	**51.5%**	**65.3%**	719.23	**0.0**
RGCN sobre grafo multicapa	48.6%	64.4%	**690.6**	6.4

Resultados - Twitter-US

Resultados considerando clases como regiones de una grilla con KD-Trees

Método	Acc@161	Error prom. (Km)	Error mediana (Km)
Miura et al. (2017)	60.1%	582.8	66.5
Rahimi et al. (2017)	61.0%	515.0	77.0
Rahimi et al. (2018)	66.0%	420.0	56.0
RGCN sobre grafo de menciones extendido	67.0%	384.0	57.2
Meta clasificador + mejores clasificadores de cada *feature*	67.0%	394.8	57.0

Resultados - Twitter-US

Resultados considerando clases como ciudades colapsadas con población > 100.000

Método	**Acc@161**	**Error prom. (Km)**	**Error mediana (Km)**
Miura et al. (2017)	61.5%	481.5	65.0
Huang and Carley (2019)	70.8%	361.5	31.6
RGCN sobre grafo de menciones extendido	66.6%	408.7	43.3
Meta clasificador + mejores clasificadores de cada *feature*	69.0%	421.0	35.3

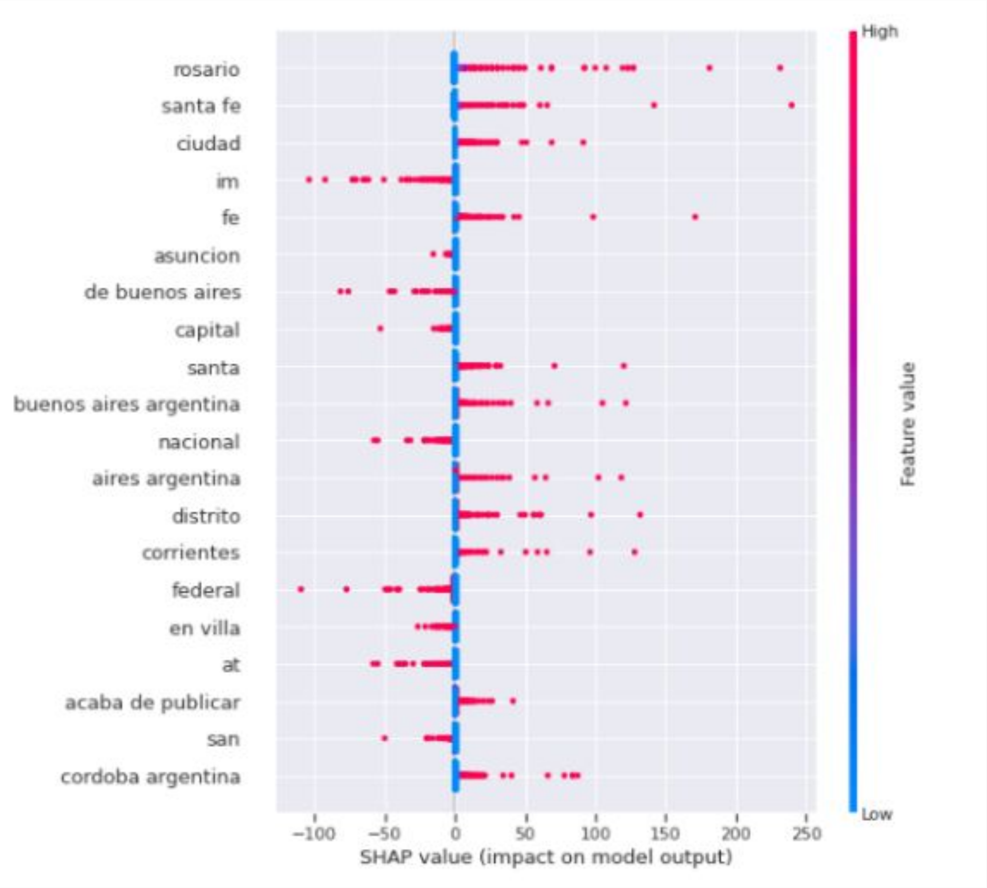
Explicabilidad

- Sin énfasis en la literatura
- Complicado al trabajar con *embeddings* de grafos

Explicabilidad - Contenido

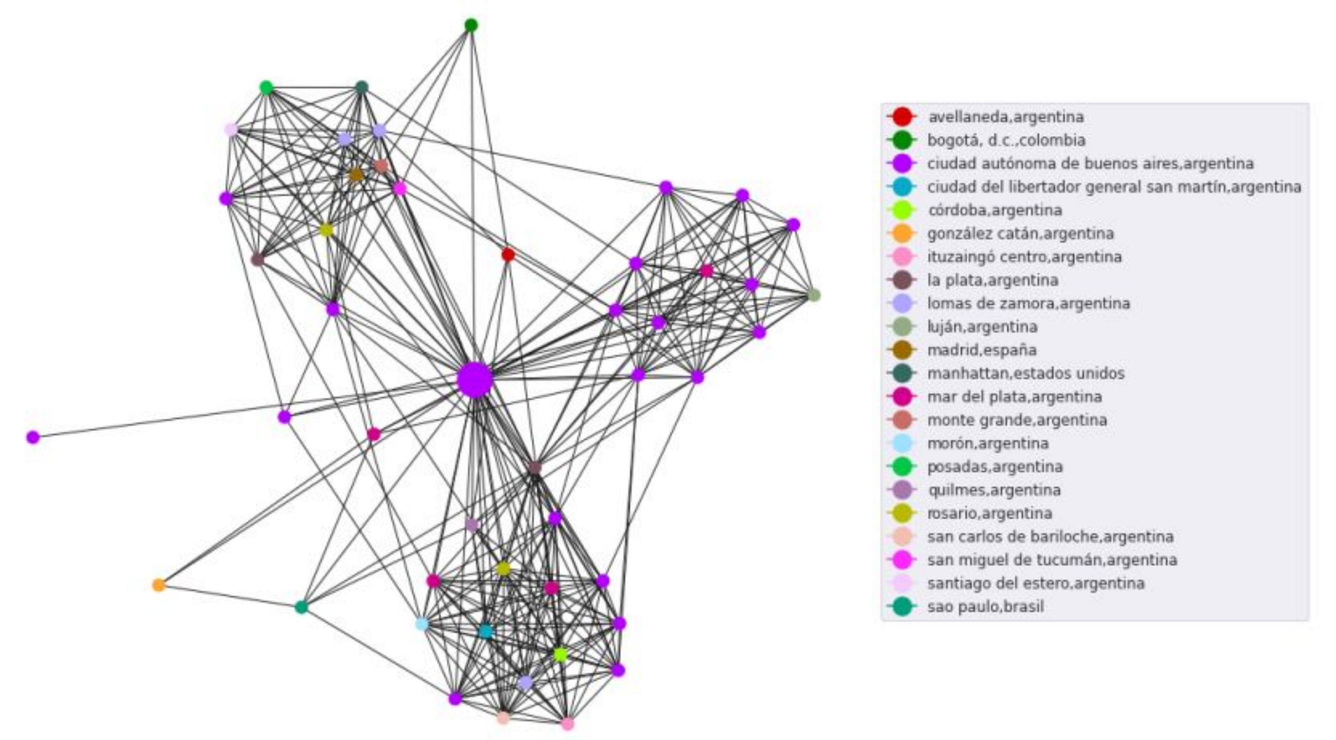
Gráficos SHAP:

- Explicabilidad para diversos modelos
- Fácil de interpretar
- Interactivo



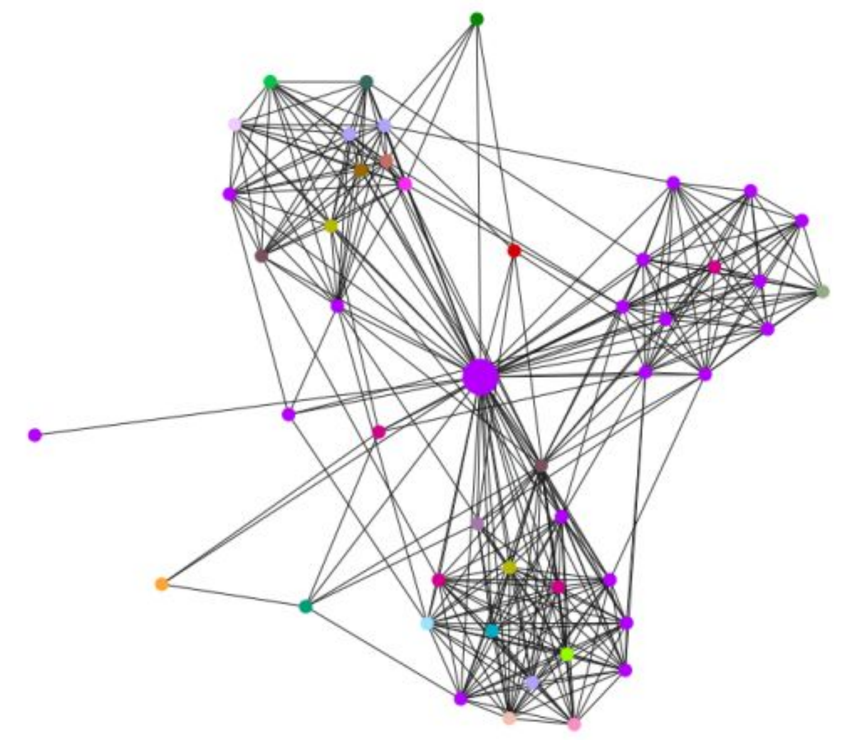
Explicabilidad - Redes

EGO-Nets - Ejemplo para usuario de Buenos Aires clasificado correctamente

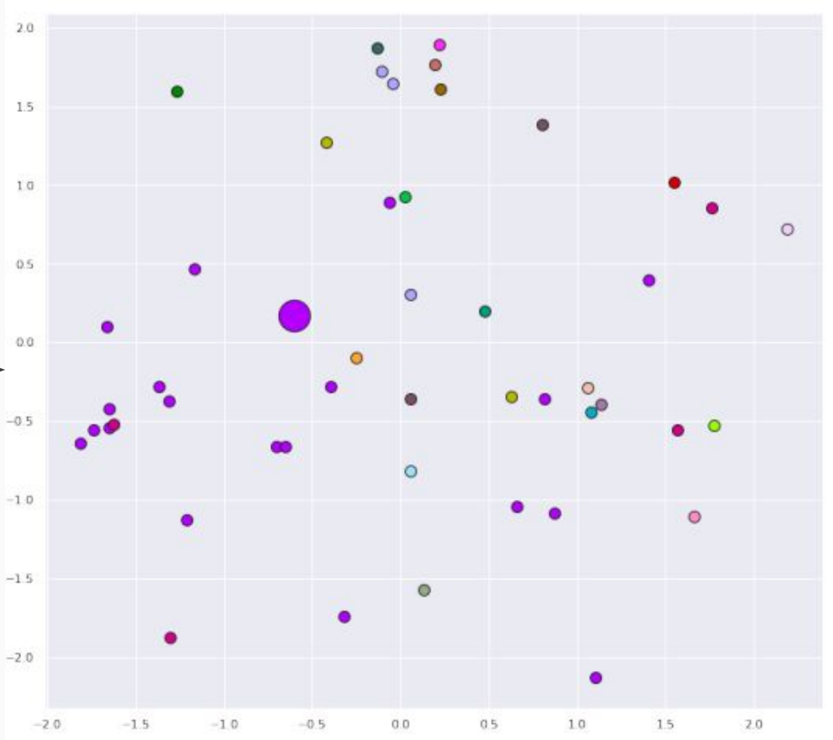


Explicabilidad - Redes

EGO-Nets - Ejemplo para usuario de Buenos Aires clasificado correctamente

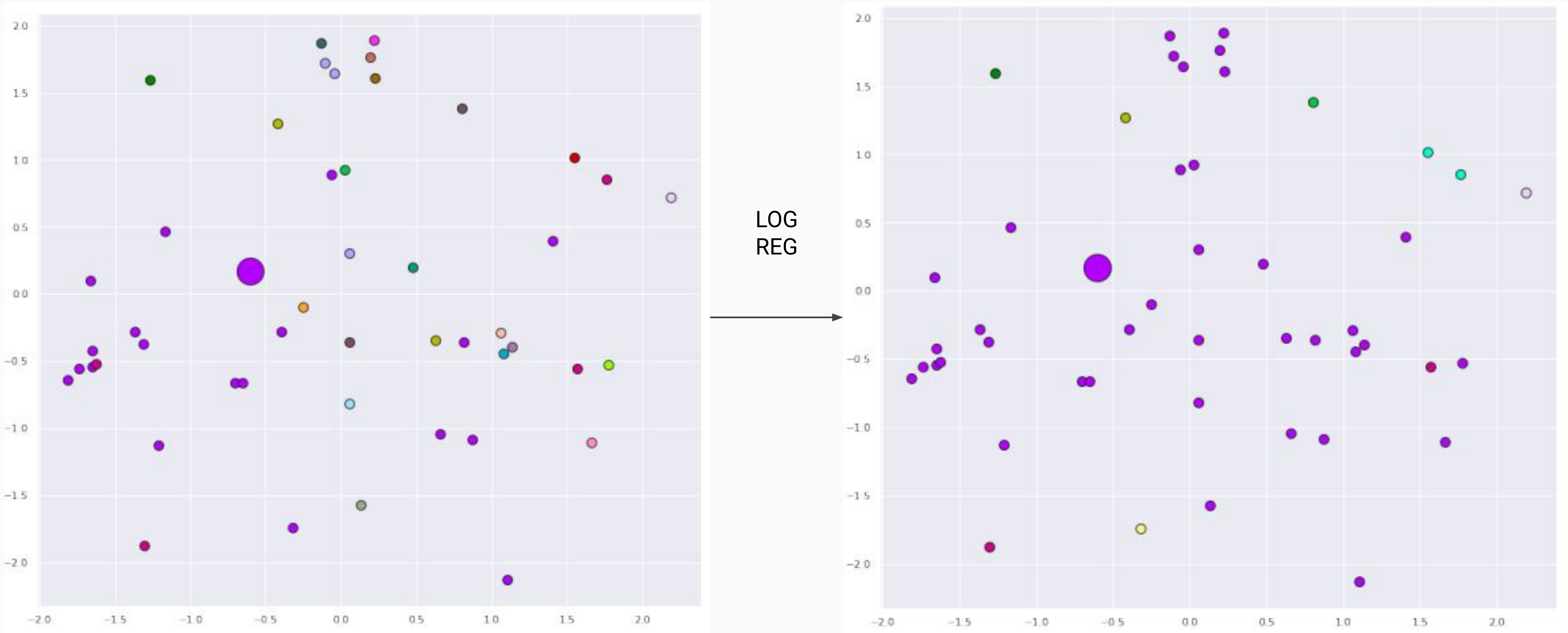


N2V +
TSNE



Explicabilidad - Redes

EGO-Nets - Ejemplo para usuario de Buenos Aires clasificado correctamente



Conclusiones

Propusimos:

- 2 conjuntos de datos centrados en Argentina: **Twitter-ARG-Exact** y **Twitter-ARG-BBox**
- **GraphSAGE**: un método inductivo basado en redes neuronales en grafos
- **R-GCN**: un método basado en redes neuronales en grafos que capta múltiples relaciones entre usuarios
- Métodos de búsqueda de comunidades en grafos, entre ellos **OSLOM** e **Infomap**
- Utilizar la información de seguidores y evaluar diferencias por su uso

Publicación en arXiv

Designing weighted and multiplex networks for deep learning user geolocation in Twitter

Federico M. Funes
Facultad de Ingeniería (UBA)
ffunes@fi.uba.ar

José Ignacio Alvarez-Hamelin
Facultad de Ingeniería (UBA)
INTECIN (UBA-CONICET)
ihameli@cnet.fi.uba.ar

Mariano G. Beiró
Facultad de Ingeniería (UBA)
INTECIN (UBA-CONICET)
mbeiro@fi.uba.ar

Abstract

Predicting the geographical location of users of social media like Twitter has found several applications in health surveillance, emergency monitoring, content personalization, and social studies in general. In this work we contribute to the research in this area by designing and evaluating new methods based on the literature of weighted multigraphs combined with state-of-the-art deep learning techniques. The explored methods depart from a similar underlying structure (that of an

flow of real-time information available through the platform to specific places.

These limitations motivate the study of location prediction in Twitter based on the information that is available on the platform, such as the content published by the users, the hashtags they use, and the users they mention, reply to, or follow. This task can be enriched by combining techniques from a variety of areas, such as Natural Language Pro-

¿Preguntas?