Geolocalización de usuarios en Twitter utilizando redes convolucionales de grafos *

Federico M. Funes[1], J. Ignacio Alvarez-Hamelin[1,2], y Mariano G. Beiró[1,2]

[1] Universidad de Buenos Aires, Facultad de Ingenería, Paseo Colón 850, C1063ACV, Buenos Aires, Argentina {ffunes,ihameli,mbeiro}@fi.uba.ar
[2] CONICET – Universidad de Buenos Aires, INTECIN.

Resumen En este trabajo utilizamos un conjunto de datos recolectados de la red social Twitter con el objetivo de analizar el desempeño de distintos modelos que proponemos para determinar la geolocalización de los usuarios de la plataforma. También realizamos un análisis sobre los perfiles de los usuarios para verificar qué tan fiable puede ser la determinación de su residencia. En el artículo detallamos distintas formas de construir las redes que modelan las relaciones entre los usuarios a fin de mejorar la estimación de su ubicación, con sus respectivas ventajas y desventajas. Por último, explicamos nuestro procedimiento para la detección de términos locales, y la conformación de secuencias para los métodos basados en redes neuronales.

Palabras Clave: aprendizaje automático · redes sociales · procesamiento del lenguaje natural · geolocalización

1. Introducción

En los últimos años se ha visto un gran crecimiento de las redes sociales: cada vez más usuarios adoptan una o más plataformas para comunicarse entre sí o para conocer las últimas noticias que acontecen a sus familiares, amigos, su país o al mundo en sí. Desde las Ciencias Sociales Computacionales puede verse a las redes sociales como un mundo de información que se puede aprovechar de diversas formas. Una de ellas consiste en determinar la ubicación de sus usuarios, información que sería útil para recomendarles contenido personalizado a los usuarios en función de dónde se encuentran, como resúmenes de noticias, anuncios regionales o información comercial local. También puede aprovecharse esta información para el monitoreo de la salud pública (p. ej., monitoreo de síntomas y dolencias generales [13]; pronóstico de la incidencia del virus del Zika [12]) o para informar sobre catástrofes locales (p. ej., terremotos [19], incendios forestales [4] o inundaciones [10]) teniendo en cuenta que estos medios de comunicación pueden llegar a ser más rápidos que los medios tradicionales. Otro uso interesante es el análisis de comunidades o grupos sociales y su comportamiento, que resulta útil para el análisis de la opinión pública.

* Este trabajo fue financiado por el proyecto PICT 2019 de la Agencia Nacional de Promoción Científica y Tecnológica (PICT 2019- 01031), y por el proyecto UBACyT-2018 20020170100421BA de la Universidad de Buenos Aires.

2. Datos utilizados

Para este trabajo utilizamos un conjunto de datos de tweets y usuarios capturados durante la campaña electoral del 2019 en Argentina, descripto en [16]. En total disponemos de aproximadamente 900 millones de *tweets* de los cuales 1 millón se encuentran geolocalizados con coordenadas específicas y 14 millones poseen un *bounding box* indicando una ciudad especificada por los mismos usuarios al momento de realizarse el tweet. Estos últimos tweets no garantizan ser tan certeros como los geolocalizados, ya que los usuarios pueden indicar la ciudad que ellos deseen desde la propia interfaz de usuario de Twitter; pero decidimos utilizarlos para tener una mayor base de usuarios para poder trabajar y porque, al día de hoy, no existen trabajos que aprovechen este tipo de tweets que son más abundantes. Nos proponemos analizar y verificar qué tan fiable puede llegar a ser esta información para que, a futuro y si es posible, más trabajos puedan utilizar estos datos sin tener que recurrir a la escasa cantidad de tweets geolocalizados. Al obtener los tweets de la API de Twitter, cada tweet se representa a través de un documento JSON con campos como `author_id`, `text` y `geo`.

Respetando la políticas de Twitter publicamos únicamente los identificadores de los tweets en el siguiente repositorio de GitHub [5]. A través de estos identificadores es posible rehidratar el conjunto de datos.

2.1. Análisis de geolocalización de los tweets

Dado el conjunto de 14 millones de tweets que contienen una ciudad especificada por los usuarios y determinada por medio de un *bounding box*, procedimos a validar y unificar las ciudades utilizando el *gazetteer* GeoNames[3], y a unificarlos en un par latitud/longitud más certero. De los 14 millones de tweets, 12 millones poseen ciudades con nombres válidos según GeoNames; estos tweets conformarán el conjunto con ubicación especificada por el usuario.

Con respecto a los tweets con geolocalización exacta, realizamos el análisis utilizando las mismas ciudades determinadas por Twitter, quedando registradas en el documento JSON junto a las coordenadas exactas.

En total disponemos de datos de aproximadamente 2 millones de usuarios, de los cuales 1 millón de usuarios especificaron una ubicación en su perfil, que podrá ser verdadera o falsa. En contrapartida, para escoger sus respectivas ciudades de residencia con mayor fiabilidad según sus tweets, decidimos quedarnos con la ciudad en la que cada usuario publicó la mayoría de sus tweets.

En resumen, obtuvimos dos conjuntos de datos: uno basado en tweets con geolocalización exacta, el cual llamaremos `Twitter-ARG-Exact`, y otro con restricciones más relajadas que llamaremos `Twitter-ARG-BBox`, el cual consiste en tweets con ubicación especificada por el usuario. En la Tabla 1 vemos un resumen de ambos conjuntos, sus respectivos tamaños, y los tweets totales.

En el conjunto `Twitter-ARG-BBox` disponemos de 229 ciudades con al menos 100 muestras cada una, distribuidas a lo largo de 22 países, siendo en su mayoría

[3] `https://www.geonames.org/`

Tabla 1. Conjuntos de datos generados durante este trabajo.

Conjuntos	#Usuarios	#Tweets con ubicación	#Tweets totales
`Twitter-ARG-Exact`	37.146	625.180	27.574.343
`Twitter-ARG-Bbox`	141.209	9.298.954	124.192.146

Tabla 2. Ejemplos de perfiles de usuarios para el conjunto `Twitter-ARG-Exact`.

Ubicación perfil	Ubicación real
Txalapa, veracruz	xalapa,méxico
ushuaia	ushuaia, argentina
buenos aires	mar del plata, argentina
santiago, chile	san carlos de bariloche, argentina
perez - santa fe - argentina	perez, argentina
buenos aires, argentina	ciudad autónoma de buenos aires, argentina
palermo hollywood	ciudad autónoma de buenos aires, argentina

de usuarios residentes en Argentina (72, 2 %), y luego de Brasil (4, 3 %), Ecuador (3,5 %), Chile (3,4 %), y otros. Para el conjunto `Twitter-ARG-Exact` disponemos de 95 ciudades con al menos 100 muestras distribuidas a lo largo de 14 países (Argentina: 76, 8 %, Brasil: 4, 7 %, Ecuador: 2, 8 %, Chile: 2, 7 %).

Se observa un gran desbalance entre las muestras de países y también de ciudades. Por ejemplo, para Argentina la mayoría de las muestras corresponden a la Ciudad Autónoma de Buenos Aires (más de 8k usuarios), y le siguen Córdoba (≈2k), Rosario (≈1k), Mar del Plata (≈1k) y La Plata (≈1k).

Durante el resto de este trabajo nos basaremos en el conjunto de datos `Twitter-ARG-Exact`, mientras que el conjunto `Twitter-ARG-BBox` será utilizado sólo para realizar comparaciones en cuanto a los resultados, y para analizar si este conjunto es válido con respecto a las ubicaciones provistas por los usuarios.

2.2. Análisis de perfiles

Hemos observado que una gran parte de los usuarios incluye una ubicación en su perfil, aunque los estudios [2] y [9] afirman que un bajo porcentaje de usuarios informa la ubicación real. Para verificarlo, usaremos las ubicaciones de los perfiles de los usuarios con el fin de detectar una ciudad y/o un país. La Tabla 2 muestra algunos ejemplos de perfiles de usuarios y la ambigüedad con la que se manejan los nombres de las ubicaciones, así como también la dificultad al utilizar distintos niveles de granularidad de ubicación.

Para procesar este campo, realizamos un pasaje a minúsculas del texto en el perfil, y separamos las palabras por medio de caracteres especiales. Luego analizamos cada palabra y determinamos por medio de un *gazetteer* (GeoNames), si con la totalidad de palabras disponibles, se hace referencia a una única ciudad, a varias ciudades o simplemente a un país.

Obtuvimos un 42, 6 % de usuarios con *ubicación exacta* (es decir, cuya información de perfil sólo coincide con una ciudad existente en GeoNames), 21, 9 %

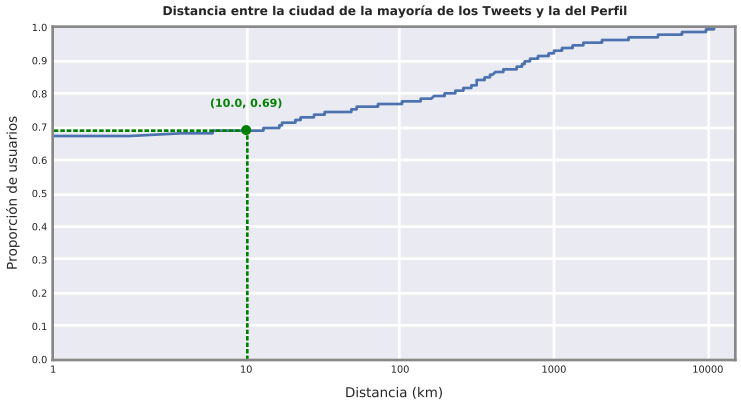

Figura 1. Distribución acumulada de distancias entre la ciudad nombrada en el perfil y la elegida por medio de la mayoría de los tweets de los usuarios con ubicación exacta.

con *ubicación ambigua*, $20,7\,\%$ *sin ubicación*, y un $14,6\,\%$ donde la *ubicación es un país*.

En particular, encontramos que en $51\,\%$ de los perfiles con ubicación exacta, su ciudad coincide con aquella de la mayoría de sus tweets. De aquí deducimos que aproximadamente en el $21,7\,\% \simeq 42,6\,\% \cdot 51\,\%$ de usuarios con perfil disponible coinciden la ciudad de sus perfiles y la seleccionada en la mayoría de sus tweets. Para el $78,3\,\%$ restante, podemos analizar la distribución de distancias entre las ciudades de los perfiles y las obtenidas de la mayoría de los tweets.

Los resultados obtenidos para dicha distribución se muestran en la Figura 1, en donde observamos que el $69\,\%$ de los usuarios con ubicación exacta se encuentra, según sus perfiles, a menos de $10km$ de la ciudad en la que mayormente publican tweets. Definiendo entonces un máximo de $10km$, podríamos resolver los problemas de granularidad (ejemplo: barrios en ciudades). De esta manera, concluimos que el $29,4\,\% \simeq 42,6\,\% \cdot 69\,\%$ de los usuarios con perfil disponible puede ser localizado con un error de hasta $10km$.

Luego, para los usuarios con ubicación ambigua, pudimos determinar que en el $45\,\%$ de los casos, dichas ubicaciones se encuentran a menos de $10km$ de la ubicación determinada por sus tweets. Combinando estos resultados podríamos ubicar a aproximadamente el $39,3\,\% \simeq 21,9\,\% \cdot 45\,\% + 29,4\,\%$ de los usuarios con perfil disponible a menos de $10km$. Estos resultados son consistentes con [3].

3. Metodología

Para que nuestro modelo pueda propagar información entre los usuarios, hemos construido y evaluado distintas redes basadas en las menciones entre los usuarios y sus relaciones de seguidor/seguido. En esta sección describiremos las técnicas que desarrollamos para construir estos grafos y para procesar el texto en los tweets. Está será la información la entrada para nuestros modelos de predicción de ubicación basados en redes convolucionales de grafos.

3.1. Construcción de redes de menciones

En Twitter los usuarios pueden escribir tweets nombrando a otros usuarios (utilizando @usuario) para que a estos les llegue una notificación y puedan así estar atentos al nuevo contenido publicado. En nuestro conjunto de datos poseemos aproximadamente 30 millones de menciones realizadas por los usuarios. Las distribuciones de menciones suelen seguir una ley de cola larga, en la que existen muchos usuarios con pocas menciones y unos pocos usuarios muy mencionados. Al querer utilizar las menciones como *proxy* para la localización, notamos que los usuarios más populares o mencionados no serían los más adecuados adecuados (el presidente de Argentina tiene muchas menciones, pero estas poco nos dicen sobre la ubicación precisa de los usuarios que lo mencionan). Usuarios con pocas menciones posiblemente nos aportarán precisión. Es difícil considerar hasta qué punto un usuario es popular: por ejemplo, las radios locales de una ciudad pueden ser muy predictivas sobre la ubicación de los usuarios que las siguen, a pesar que estas radios pueden tener muchas menciones.

Adicionalmente, otro problema a resolver es que los usuarios de nuestro conjunto de datos pueden mencionar a usuarios por fuera de ese conjunto, lo cuál agrega muchos nodos externos sobre los que no poseemos información (sólo su mención). Por lo tanto no tiene sentido predecir la ubicación de nodos externos por no tener ningún otro dado. Entonces, una opción posible es compactar nuestra red de menciones de forma de sólo trabajar con usuarios de nuestro conjunto de datos, pero tratando de preservar la mayor cantidad de información posible.

Muchos trabajos de la literatura simplemente descartan esas conexiones con usuarios externos. Sin embargo, dado que gran parte de las menciones son de este tipo, de esta forma se pierde mucha información valiosa que puede ayudar a la geolocalización de los usuarios en nuestro conjunto de datos.

Entonces, en este trabajo proponemos incluir esta información sobre *comenciones*, representada por medio de una matriz de comenciones cuyo elemento (i, j) nos indica la cantidad de menciones en común que existen entre los nodos i y j. Un ejemplo sobre cómo conformar esta red puede verse en la Figura 2.

Dado que los nodos externos a nuestro conjunto poseen únicamente aristas entrantes, la red de *comenciones* va a tener a dichos nodos aislados y va a conectar a nuestros nodos sólo si poseen una mención en común, independientemente de si mencionan usuarios externos o de nuestro conjunto. Sin embargo, debemos notar dos inconvenientes: el primero es que alguno de nuestros nodos podría incurrir en un camino más largo hacia un nodo cercano del que poseía con anterioridad (ver nodos 5 y 6 en la Figura 2, que dejan de tener una conexión directa en la red de *comenciones*); el segundo inconveniente está relacionado con la cantidad de aristas que existen en la red de *comenciones*. Consideremos que tenemos un nodo i de grado entrante n (esto quiere decir que n nodos mencionan al nodo i), entonces, cuando generemos la red de *comenciones*, dichos n nodos tendrán todos una arista en común entre ellos, siendo $\frac{(n-1)^2+(n-1)}{2}$ aristas en total. Para resolver el primer inconveniente poseemos dos soluciones: una consiste en generar lazos para cada nodo, con lo cual cualquier mención directa entre dos nodos se conservaría, mientras que otra solución sería combinar ambos enfoques,

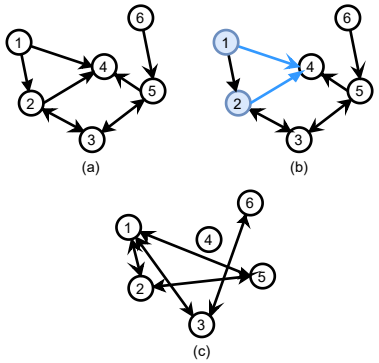

Figura 2. Obtención de la red de *comenciones* por medio de la red de menciones. (a) representa un grafo de menciones dirigida; en (b) vemos que los nodos 1 y 2 poseen una mención en común (*comención*); en (c) ya vemos la red de *comenciones*, sobre la cual notamos que todas las aristas son bidireccionales, con lo que de ahora en adelante consideraremos la matriz como no dirigida.

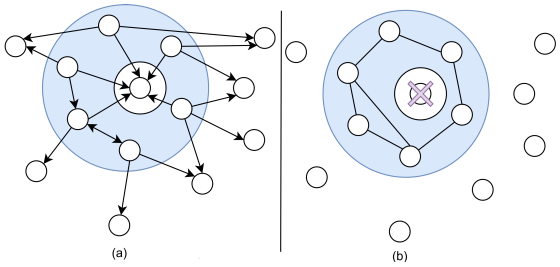

Figura 3. Ejemplo de red de menciones donde el conjunto azul representa usuarios pertenecientes a nuestro conjunto. En (a) observamos la red de menciones dirigida; en (b) la red de *comenciones* resultante al eliminar nodos populares.

conservando la red de menciones para los usuarios de nuestra red y generando *comenciones* entre ellos utilizando únicamente usuarios externos a nuestro conjunto. Dicho de otra forma, uniendo a nuestros nodos entre sí si existe un camino entre ellos por medio de un nodo externo. A la primera solución la llamaremos "*red de menciones y comenciones*", y a la segunda "*red de menciones extendida*". Para reducir los efectos del segundo inconveniente podemos eliminar a los nodos populares como mencionamos antes, ya que estos son los que mayor cantidad de aristas generarían y menos información sobre la ubicación del usuario aportarían.

Construcción de los grafos El grafo de *menciones y comenciones* se obtiene calculando $(M + I) \cdot (M + I)^{\mathrm{T}}$, en donde M es la matriz de menciones, que representa que el usuario i mencionó m_{ij} veces al usuario j. De esta matriz C nos quedamos con las primeras N filas y columnas, que corresponden a los usuarios en nuestro conjunto de datos. Nótese que se agregan lazos a cada nodo al sumar la matriz identidad I.

Tabla 3. Distintos tipos de grafos construídos a partir de las menciones, y sus tamaños para `Twitter-ARG-Exact`.

Grafo	#Nodos	#Aristas	Filtro nodos populares
Menciones completo	1.670.106	9.107.669	-
Menciones cortado	37.146	206.608	-
Menciones y comenciones	37.146	4.597.803	> 20 menciones únicas
Menciones extendido	37.146	4.524.003	> 20 menciones únicas

Para generar el grafo de *menciones extendido* partimos de matriz M y una matriz correspondiente a un grafo bipartito $X_M \in \mathbf{1}^{n\times(m-n)}$ que representa las menciones realizadas por los usuarios de nuestro conjunto hacia usuarios externos. Conformamos la matriz de menciones extendida $F = (M + M^T) + (X_M \cdot X_M^T - diag(X_M \cdot X_M^T))$, en donde el término $(M + M^T)$ corresponde a un grafo de menciones no dirigido entre los usuarios de nuestro conjunto, y el término $(X_M \cdot X_M^T - diag(X_M \cdot X_M^T))$ a un grafo de *comenciones* con respecto a usuarios externos. La Tabla 3 resume los tamaños de cada uno de los grafos obtenidos para trabajar con las menciones.

3.2. Construcción de las redes de seguidores

A diferencia de las menciones, obtener el conjunto de seguidos (*followees*) y seguidores (*followers*) de los usuarios es más costoso por los tiempos mínimos entre peticiones a la API de Twitter. Debido a esto, existen muy pocos trabajos que aprovechen esta información [17], que creemos puede ser un determinante clave para detectar la ubicación de los usuarios. Debemos notar que para esta red, los usuarios poseen aristas entrantes de nodos por fuera de nuestro conjunto de datos, y como en el caso de la red de menciones, debemos acotar la red para poder trabajar con diversos métodos. Por esto seguiremos los mismos conceptos para reducir la cantidad de nodos de nuestra red de seguidores, es decir, conformar una red de *seguidores y coseguidores* y una red de *seguidores extendida*. Observar entonces que, para el grafo de *coseguidores*, hay que tener en cuenta que al disponer tanto de seguidos como seguidores nos indica que poseemos muchas más aristas entre usuarios de nuestro conjunto. Entonces, al conformar el grafo de *coseguidores*, incurriríamos en una masiva cantidad de aristas entre los usuarios, por lo que aún poniendo un filtro más estricto sobre la cantidad máxima de seguidores únicos por nodo el problema persistiría. La Tabla 4 resume los grafos obtenidos para los seguidores; hacemos notar que no trabajaremos con el grafo de *seguidores y coseguidores* dada su gran cantidad de aristas.

3.3. Análisis de contenido

Analizar el contenido de los tweets de los usuarios es una tarea fundamental para determinar sus correspondientes ciudades de residencia. Este es un trabajo que recae dentro del campo conocido como Procesamiento del Lenguaje Natural

Tabla 4. Distintos tipos de grafos construídos a partir de las relaciones de seguidor/seguido, y sus tamaños para `Twitter-ARG-Exact`.

Grafo	#Nodos	#Aristas	Filtro nodos populares
Seguidores completo	686.694	5.903.440	-
Seguidores cortado	37.146	164.376	-
Seguidores y coseguidores	37.146	18.882.006	> 20 seguidores únicos
Seguidores extendido	37.146	1.441.746	> 20 seguidores únicos

Tabla 5. Ejemplos de tweets procesados y hashtags extraídos.

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

(NLP). En esta sección describiremos aquellos métodos que utilizamos, indicando sus correspondientes ventajas y desventajas.

Antes de analizar el contenido de los tweets de los usuarios, se requiere un pre-procesamiento a fin de eliminar términos redundantes, unificar términos similares y reducir la cantidad de información a procesar. Nuestro enfoque consiste en primero pasar todos los tweets a minúsculas, eliminando las menciones (ya que estas son analizadas por medio de otros enfoques desarrollados en secciones anteriores), extraer los *hashtags* utilizados por al menos 10 usuarios distintos, que se identifican por medio del carácter #, y finalmente eliminar símbolos, emojis, saltos de línea y direcciones a páginas web. De esta forma obtenemos un conjunto de tweets filtrados y hashtags tal como se muestra en la Tabla 5.

Búsqueda de términos locales Los usuarios de redes sociales como Twitter suelen publicar contenido utilizando modismos, términos populares en un momento determinado, y abreviaciones de lugares. Trabajos como [2,7,11] mostraron que es posible reconocer dentro de estos términos algunos que se encuentran asociados a una zona geográfica en específico. Por ejemplo, es muy probable que la gente de Almirante Brown hable del Boulevard Shopping. A estos términos los llamamos *Local Indicative Words* (LIW) tal como se hizo en los trabajos previamente mencionados. Encontrar LIW's puede ser una tarea costosa por la enorme cantidad de términos que se encuentran en un conjunto de datos; sin embargo, hoy en día existe una gran variedad de métodos para identificarlos. En esta sección haremos hincapié en algunos de métodos que utilizamos. Para nuestros ensayos, consideramos un término a cualquier unigrama, bigrama o trigrama con al menos 10 usos realizados por usuarios distintos, y con un máximo de uso de hasta el 20 % de los usuarios, para filtrar palabras comúnmente usadas y que

Tabla 6. Ejemplo de tabla de contingencia para el cálculo de χ^2 para un término. La totalidad de muestras M equivale a $\sum_{i,j}^{n} O_{\mathrm{i,j}}$.

	Ciudad	Otra ciudad
Término	$O_{\mathrm{t,c}}$	$O_{\mathrm{t,oc}}$
Otro término	$O_{\mathrm{ot,c}}$	$O_{\mathrm{ot,oc}}$

poco aporten a la búsqueda de LIW's. También consideramos que los *hashtags* podrían ser un fuerte indicativo de localidad, por lo que los tendremos en cuenta a la hora de seleccionar los más significativos.

Nuestro primer enfoque consiste en el uso de métodos estadísticos para realizar pruebas y determinar, con un cierto nivel de confianza, si ciertos eventos ocurren de forma no aleatoria. En particular, vamos a usar la distribución de Pearson χ^2 tal como se hizo en trabajos como [8]. Como primer paso debemos definir las hipótesis a verificar y los eventos: los eventos con los que trataremos consisten en si un término es local a una ciudad, definiendo como hipótesis nula la independencia entre el término y la ciudad, y la dependencia como la hipótesis alternativa. Como segundo paso, construimos una tabla de contingencia como se muestra en la Tabla 6. Luego, calculamos la frecuencia esperada para cada caso como $E_{\mathrm{i,j}} = P(i) \cdot P(j) \cdot M$, siendo M la cantidad de observaciones. Finalmente calculamos $\chi^2 = \sum_{i,j}^{n} \frac{(O_{\mathrm{i,j}} - E_{\mathrm{i,j}})^2}{E_{\mathrm{i,j}}}$, con un solo grado de libertad, y buscamos su valor en la tabla de valores de χ^2 según el nivel de confianza que deseamos tener. A partir de este momento podemos tomar varios enfoques para trabajar con los resultados obtenidos de χ^2 para cada término y ciudad: por ejemplo, conservar los N términos con mayor valor para cada ciudad, o aquellos que para una ciudad en específico estén dentro del nivel de confianza que planteemos. Si decidimos quedarnos con los K términos más significativos podemos ver cuáles serían esos términos para cada ciudad. En la Tabla 7 mostramos un resumen de términos más significativos encontrados para algunas ciudades, en donde observamos que la mayoría de dichos términos son el nombre de la ciudad en sí.

Nuestro segundo enfoque para encontrar LIW's consiste en usar la información mutua para determinar si la presencia o ausencia de un término ayuda a clasificar correctamente a una ciudad; una vez calculado el valor para cada término y ciudad, nos quedamos con los K términos más significativos por ciudad igual que como hicimos con las pruebas χ^2.

Estos métodos de búsqueda de LIW's tienen algunas limitaciones: son incapaces de captar relaciones entre términos, y si queremos considerar el uso de unigramas, bigramas y trigramas como LIW's, la cantidad de términos a evaluar se vuelve excesivamente grande a medida que incorporamos más muestras.

Análisis de secuencias Como complemento a la búsqueda de términos locales, también es posible analizar los tweets como una secuencia para capturar relaciones entre los términos. De esta forma nuestro modelo podrá determinar en dónde viven los usuarios en función de cómo escriben y qué términos combi-

Tabla 7. Ejemplo de términos más significativos según pruebas χ^2 para algunas ciudades para el conjunto `Twitter-ARG-Exact`.

Ciudad	Términos significativos
Almirante Brown, Argentina	'burzaco', 'burzaco en', 'foto en almirante', 'almirante', 'en claypole', 'en jose marmol'
Villa Soldati, Argentina	'villa soldati', 'bidegain nuevo gasometro', 'gasometro club', 'en villa soldati', 'villa soldati distrito', 'polideportivo roberto', 'champagne'
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'
Vicente Lopez, Argentina	'tecnopolis buenos aires', 'en tecnopolis', 'vte lopez', 'gral manuel belgrano', 'en florida vicente', 'ensaladita', 'vicente lopez buenos', 'vicente lopez'
Santa Fé, Argentina	'santa fe mi', 'en esparanza santa', 'estanislao', 'norte salta', 'estanislao lopez', 'estadio brigadier general', 'santa fe no', 'santa fe acaba', 'en santa fe'

nan. Dado que muchos métodos para entrenar sobre secuencias de texto como las LSTM (Long-Short Term Memory) requieren que las secuencias de entrada sean de longitud fija, hemos acotado el contenido de los tweets a una longitud razonable para capturar la mayor cantidad de información posible, manteniendo a su vez el tiempo de entrenamiento acotado. Analizando la distribución de la longitud de los tweets concatenados de los usuarios en la Figura 4, y sabiendo que muchos los usuarios utilizan algunas palabras en forma repetida, consideramos tomar una longitud que se encuentre por encima de la mediana y sea menor a 500 palabras. Finalmente representamos los tweets concatenados de los usuarios como una secuencia fija de las primeras N palabras utilizadas, representada cada una con un número único, y para aquellas secuencias con menos de N palabras, completamos las mismas agregando un símbolo especial en su comienzo.

Tratamiento de datos desbalanceados Uno de los principales inconvenientes para el entrenamiento es el desbalance de muestras a nivel ciudad. Existen múltiples formas de minimizar esta problemática. Uno de los enfoques que consideramos es balancear el conjunto de datos generando y eliminando muestras de forma aleatoria a través de *SMOTE* [1] para mantener un balance a nivel ciudad según cada país. Otro enfoque comúnmente utilizado es dividir al espacio por medio de *kd-trees* [18], generando regiones en donde la cantidad de muestras es similar, esto trae como ventaja que ciudades con muy baja cantidad de muestras pueden juntarse dentro de una misma región mejorando los resultados a simple vista. Sin embargo, este método presenta dos inconvenientes: el primero es que ciudades con una gran cantidad de muestras con respecto al resto se dividirán en regiones, dificultando la tarea del clasificador al colocar a usuarios muy cercanos en clases distintas; y el segundo es que la API de Twitter nos ofrece coordena-

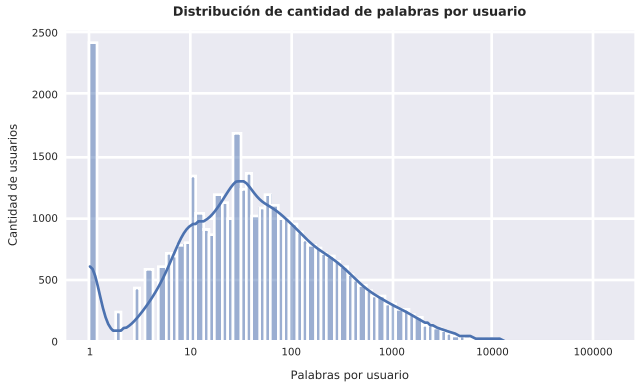

Figura 4. Distribución de cantidad de palabras utilizadas por los usuarios del conjunto `Twitter-ARG-Exact`.

das para los tweets geolocalizados con una precisión de decimales, que hace que muchos usuarios terminen localizados en el mismo punto. En estas situaciones el uso de *kd-trees*, según el valor mínimo de muestras que se tome por región, puede no generar regiones balanceadas. En este trabajo consideramos utilizar regiones de como mínimo 255, 500 ó hasta 1255 muestras para generar distintas resoluciones de regiones posibles, sin embargo, las regiones no se encontraron balanceadas para los dos primeros casos. El último enfoque que consideramos es ajustar pesos en la función de costo, para darles más valor a las ciudades con menor cantidad de muestras. De todos los enfoques para lidiar con el desbalance de las ciudades en nuestro conjunto de datos, el que mejor resultó es el ajuste de pesos por ciudad, que es el que mostraremos en los resultados.

Modelo de entrenamiento Nuestra arquitectura general de entrenamiento se basa en una red neuronal profunda con una componente de procesamiento de texto y otra de difusión a través de redes de usuarios. Construimos para cada usuario una representación vectorial de su contenido que combina Local Indicative Words (LIW's) con un embedding del contenido que luego ingresará a una LSTM o a un transformador. La representación obtenida a la salida de esta etapa es propagada a través de una red convolucional de grafos (*graph convolutional network*, o GCN). Las GCN's son modelos de redes neuronales profundas en las que en cada capa un nodo combina la información de estado proveniente de sus nodos vecinos, utilizando una función de agregación para calcular su nuevo estado. Esta técnica permite implementar un proceso de difusión de información en la red neuronal, y su versión original ha sido aplicada a redes de menciones en trabajos previos como [15]. Aquí hemos evaluado dos arquitecturas de GCN's: R-GCN [20] y GraphSAGE [6]. A la salida de este modelo se obtiene en cada nodo un vector que indica la probabilidad de que el usuario esté ubicado en una de las ciudades objetivo. Utilizamos la *cross entropy* como función de costo, ajustada por el peso de cada ciudad.

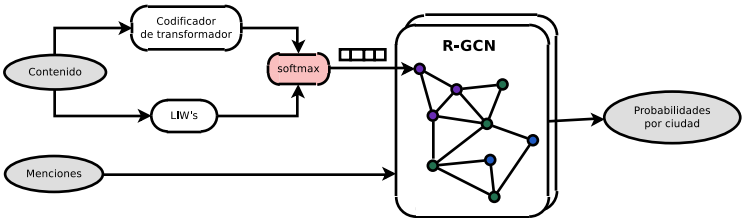

Figura 5. Arquitectura del modelo `RGCN-EXT-Preds`, que ofrece el mejor desempeño entre las arquitecturas evaluadas.

4. Resultados

Hemos evaluado los distintos métodos propuestos de entrenamiento a partir del contenido: el uso de LIW's rankeadas por el método de χ^2, el de LIW's rankeadas por información mutua, y la inyección directa del texto en una LSTM o en el codificador de un transformador. Las métricas de evaluación que utilizamos, conforme a otros trabajos en la literatura, son: la *accuracy (Acc.)* o proporción de muestras correctamente clasificadas, la *Acc@161* [2,14], que representa la proporción de muestras clasificadas en ciudades a menos de 100 millas ($\approx 161km$) de su ubicación etiquetada, y la media y mediana del error, en kilómetros.

De entre todos los métodos evaluados para contenido, los mejores resultados se obtuvieron a partir del método basado en χ^2 y el uso de un transformador: por ejemplo, al intentar predecir un vector de probabilidades de ubicación únicamente a partir del contenido, el desempeño con el método de χ^2 arrojó una Acc@161 de $76,9\,\%$, mientras que a través de información mutua obtuvimos $71,2\,\%$, y utilizando una LSTM $75\,\%$, y utilizando un codificador de transformador, $76,4\,\%$.

En base a esto, se escogió combinar las LIW's extraídas por el método basado en χ^2 con la representación generada por el transformador, y se evaluaron las arquitecturas R-GCN y GraphSAGE con los grafos extendidos de menciones y seguidores. Esta arquitectura se ilustra para el caso de la R-GCN en la Figura 5.

Todas las mediciones de desempeño se realizaron mediante un proceso de validación cruzada anidada: en una primera instancia utilizamos *Stratified 5-fold cross validation* para generar conjuntos de entrenamiento y prueba rotativos; luego, dentro de cada conjunto de entrenamiento volvemos a dividir en 3 *folds* para encontrar los mejores hiperparámetros del modelo utilizando 2 *folds* y evaluando sobre el *fold* restante. De esta forma, cada una de nuestras predicciones es realizada sobre un modelo que no observó esa muestra durante el entrenamiento.

Los resultados obtenidos se muestran en la Tabla 8. Aquí vemos que una R-GCN con una capa para las menciones extendidas y otra para los seguidores extendidos resultó ser la mejor arquitectura para `Twitter-ARG-Exact`, obteniendo una Acc@161 de $82,9\,\%$ y una mediana de error de $3,8$km, siendo ligeramente superior a la performance obtenida por GraphSAGE.

Por otra parte, observamos que utilizando el conjunto de datos `Twitter-ARG-Bbox` el desempeño resulta bastante inferior, practicamente duplicando el error medio en kilómetros. Sin embargo, vistos los resultados, consideramos que representa

Tabla 8. Tabla resumida de resultados de modelos aplicados a los conjuntos de datos `Twitter-ARG-Exact` y `Twitter-ARG-Bbox`.

Modelo	Acc.	Acc@161	Media Err.(km)	Mediana Err.(km)
`Twitter-ARG-Exact`				
`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`				
`RGCN-EXT-Preds`	48.6 %	64.4 %	690.6	6.4
`GraphSAGE-EXT-Preds`	46.3 %	57.8 %	798.4	23.2

una alternativa válida utilizar los datos de *bounding box* cuando no se dispone de datos precisos suficientes de geolocalización.

5. Conclusiones

Hemos presentado un análisis de un conjunto de datos extraído de Twitter para estimar la ubicación de un usuario a nivel de ciudad. A partir del mismo, hemos propuesto construcciones de grafos aumentados que permitan incluir información externa a los usuarios que forman parte de nuestro conjunto, para mejorar las predicciones. A partir de estas construcciones, y combinando dos arquitecturas de aprendizaje automático como los transformadores y las redes convolucionales de grafos, hemos mostrado que es posible la mejora de la estimación de la ubicación a partir de información extraída de las redes de usuarios. Como trabajo futuro se espera explorar otros métodos de aprendizaje automático, y aplicar esta metodología a otras bases de datos similares.

Referencias

1. Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.
2. Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proc. of the 19th ACM international conference on Information and knowledge management*, pages 759–768, 2010.
3. Ruben Cuevas, Roberto Gonzalez, Angel Cuevas, and Carmen Guerrero. Understanding the locality effect in twitter: measurement and analysis. *Personal and Ubiquitous Computing*, 18(2):397–411, 2014.
4. Bertrand De Longueville, Robin S Smith, and Gianluca Luraschi. "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. In *Proc. of the 2009 international workshop on location based social networks*, pages 73–80, 2009.

5. Federico M Funes. Tweet ids for dataset hydration, 2021. https://github.com/fedefunes96/twitter-location-data.
6. William L Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proc. of the 31st International Conference on Neural Information Processing Systems*, pages 1025–1035, 2017.
7. Bo Han, Paul Cook, and Timothy Baldwin. Geolocation prediction in social media data by finding location indicative words. In *Proc. of COLING 2012*, pages 1045–1062, 2012.
8. Bo Han, Paul Cook, and Timothy Baldwin. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49:451–500, 2014.
9. Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H Chi. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *Proc. of the SIGCHI conference on human factors in computing systems*, pages 237–246, 2011.
10. Brenden Jongman, Jurjen Wagemaker, Beatriz Revilla Romero, and Erin Coughlan De Perez. Early flood detection for rapid humanitarian response: harnessing near real-time satellite and twitter signals. *ISPRS International Journal of Geo-Information*, 4(4):2246–2266, 2015.
11. Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):1–21, 2014.
12. Sarah F McGough, John S Brownstein, Jared B Hawkins, and Mauricio Santillana. 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*, 11(1):e0005295, 2017.
13. Michael J Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. In *Fifth international AAAI conference on weblogs and social media*, 2011.
14. Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. A neural model for user geolocation and lexical dialectology. In *Proc. of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 209–216, Vancouver, Canada, July 2017. Association for Computational Linguistics.
15. Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. Semi-supervised user geolocation via graph convolutional networks. In *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2009–2019, Melbourne, Australia, July 2018.
16. Tomás Mussi Reyero, Mariano G Beiró, J Ignacio Alvarez-Hamelin, Laura Hernández, and Dimitris Kotzinos. Evolution of the political opinion landscape during electoral periods. *EPJ Data Science*, 10(1):31, 2021.
17. Erica Rodrigues, Renato Assunção, Gisele L Pappa, Diogo Renno, and Wagner Meira Jr. Exploring multiple evidence to infer users' location in twitter. *Neurocomputing*, 171:30–38, 2016.
18. Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. Supervised text-based geolocation using language models on an adaptive grid. In *Proc. of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 1500–1510, 2012.
19. Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proc. of the 19th international conference on World wide web*, pages 851–860, 2010.
20. Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer, 2018.