UNIVERSIDAD DE BUENOS AIRES

FACULTAD DE CIENCIAS EXACTAS Y NATURALES
FACULTAD DE INGENIERÍA

La pandemia en palabras: dinámica del cambio semántico a partir de redes semánticas

Tesis Presentada para optar al Título de Magíster en Explotación de Datos y Descubrimiento de Conocimiento

Bioq. Vanesa Judith Copa

Director: Dra. Laura Kaczer

Co-Director: Dr. Sebastián Pinto

Fecha de defensa: 27 de marzo de 2024

Índice general

1. Introducción **3**
1.1. Introducción . . . 3
1.1.1. Redes sociales para estimar cambios semánticos . . . 5
1.1.2. Redes semánticas . . . 6
1.1.3. Trabajos previos sobre impacto del COVID-19 en el lenguaje . . . 7
1.2. Motivación y Objetivos . . . 8

2. Datos y métodos **9**
2.1. Descripción del dataset . . . 9
2.2. Preprocesamiento . . . 10
2.3. Construcción de redes semánticas . . . 11
2.3.1. Redes semánticas ponderadas . . . 11
2.3.2. Red semántica agregada . . . 15
2.3.3. Binarización y reducción del tamaño de la red . . . 15
2.3.4. Redes resultantes a analizar . . . 16

3. Análisis de las redes semánticas **18**
3.1. Propiedades topológicas . . . 18
3.1.1. Métricas generales . . . 18
3.1.2. Medidas de centralidad . . . 19
3.1.3. Resultados y análisis . . . 19
3.2. Detección de comunidades . . . 20
3.2.1. Resultados y análisis . . . 21

4. Evaluación de términos en el tiempo **25**
4.1. Listado de palabras seleccionadas . . . 25
4.2. Seguimiento de comunas asociadas a términos de interés . . . 26
4.3. Variación de la posición en la red de los términos de interés . . . 26
4.3.1. Medidas de centralidad . . . 27
4.3.2. Vecindad de los nodos . . . 28
4.4. Visualización de los cambios de significado . . . 29

5. Conclusiones **32**

A. Apéndice **37**
A.1. Algoritmo Backbone (Columna vertebral): ejemplo de cálculo 37
A.2. Comunidades 38
A.2.1. Selección del algoritmo de detección de comunidades 38
A.2.2. Otras comunidades 39

La pandemia en palabras: dinámica del cambio semántico a partir de redes semánticas

Resumen

Comprender el cambio en el significado de las palabras en diferentes contextos y períodos de tiempo es crucial para revelar el papel del lenguaje en la evolución social y cultural. En este estudio, examinamos el efecto diacrónico de la pandemia COVID-19 en el uso y significado de las palabras a través de la lente de la teoría de redes complejas. En concreto, analizamos cómo se expresan en Twitter un conjunto de usuarios de forma tal de realizar un seguimiento longitudinal y analizamos cómo su uso del lenguaje se vio afectado por la pandemia. Utilizando el análisis de redes, construimos redes correspondientes a tres años diferentes (2019, 2020, 2021) para representar la relación entre las palabras utilizadas por los usuarios de Twitter, lo que nos permite rastrear cuantitativamente los cambios semánticos. Nuestros resultados demuestran que se pueden detectar cambios semánticos tanto en la aparición de nuevas comunidades en la red semántica como en el cambio de la vecindad de palabras específicas. Proponemos entonces que el uso de redes complejas puede verse como una nueva metodología útil para rastrear cambios semánticos producidos en diferentes períodos de tiempo y, en particular, ayudarnos a comprender el impacto de la pandemia en el uso del lenguaje.

Palabras claves: Cambios semánticos diacrónicos, Covid-19, redes complejas, Twitter, procesamiento de lenguaje natural

The pandemic in words: dynamics of semantic change from semantic networks.

Abstract

Understanding the change in the meaning of words in different contexts and time periods are crucial to revealing the role of language in social and cultural evolution. In this study, we examine the diachronic effect of the COVID-19 pandemic on the use and meaning of words through the lens of complex network theory. Specifically, we analyze how a group of Twitter users exprees themselves in order to carry out longitudinal monitoring and how their language use was affected by the pandemic. Using network analysis, we construct networks corresponding to three different years (2019, 2020, 2021) to represent the relationship between words used by Twitter users, allowing us to track semantic changes quantitatively. Our results demonstrate that semantic changes can be detected both in the emergence of new communities in the semantic network and in the change of the neighborhood of targeted words. Therefore, we propose that the use of complex networks can be seen as a new methodology useful for tracking semantic changes produced in different periods of time, and, in particular, it can help us understand the impact of the pandemic in the use of language.

Keywords: Diachronic semantic shifts, Covid-19, Complex networks, Twitter, natural lenguage processing

Capítulo 1

Introducción

1.1. Introducción

El lenguaje no es estático, sino que va evolucionando para adaptarse a nuevas ideas, tecnologías y cambios sociales. A nivel individual, a lo largo de nuestra vida estamos expuestos a situaciones que requieren un cambio en el significado de palabras que ya conocemos. Por ejemplo, recientemente aprendimos que "nube" también hace referencia al almacenamiento de datos a través de Internet, y una "historia" puede referirse a una herramienta de las redes sociales. Estos nuevos significados no reemplazaron a los anteriores, sino que ampliaron el rango de aplicación de estas palabras [Wilkins, 1996]. Es decir que la memoria de las palabras que almacenamos en nuestro léxico mental es muy maleable y sujeta a ser actualizada o modificada en diferentes dimensiones del espacio semántico [Shtyrov et al., 2010].

En cuanto al cambio colectivo del lenguaje, en lingüística diacrónica (o histórica), el cambio semántico se refiere a cualquier cambio en el significado de una palabra a lo largo del tiempo [Hamilton et al., 2016a]. La mayoría de las palabras tienen una variedad de sentidos, que pueden agregarse, eliminarse o modificarse con el tiempo, manteniendo su forma. Por ejemplo, en inglés la palabra "gay" cambió de significar "alegre" o "jovial" a referirse a la homosexualidad, y la palabra "awful" experimentó un proceso de peyoración, al pasar de significar "lleno de admiración" a significar "terrible o espantoso" [Simpson and Proffitt, 1997]. También podemos mencionar en español el caso de la palabra "bizarro" que, según el diccionario de 1786 recopilado en el Nuevo Tesoro Lexicográfico de la Lengua Española [NTLLE, nd], una de sus acepciones era "magnánimo"; en el diccionario de la Real Academia Española RAE (2014) poseía dos acepciones: "valiente (arriesgado)" y "generoso, lúdico, espléndido"; y en la actualización del año 2022 se incorporó una tercera acepción: "raro, extravagante o fuera de lo común" [Española, 2022].

Como cualquier cambio lingüístico, un cambio semántico suele ser lento, no adquirido simultáneamente por todos los miembros de una comunidad lingüística y difícil de predecir [Wilkins, 1996]. Sin embargo, acontecimientos históricos de envergadura global pueden conllevar

asociados cambios rápidos en el significado de las palabras. Por ejemplo, los ataques del 11 de septiembre de 2001 han cambiado significativamente el panorama general de interpretación de la palabra "terrorismo" [Rababah, 2015]. En el presente estudio estamos interesados en abordar tal plasticidad de las representaciones léxico-semánticas en el contexto de la pandemia de COVID-19.

Detectado por primera vez en Wuhan, China, en diciembre 31 de 2019, el brote de COVID-19 creció rápidamente en escala y severidad, y fue declarado oficialmente como pandemia el 11 de marzo de 2020. La pandemia no sólo irrumpió de manera abrupta en la salud, la cotidianidad familiar, social, cultural, económica, laboral y académica, sino que también lo hizo en la lengua oral y escrita, en su uso especializado y no especializado. En tan solo unos meses, algunas palabras de baja frecuencia ya existentes se volvieron cotidianas (por ejemplo, "cuarentena", "inmunidad") y otras como "burbuja" fueron adquiriendo nuevos sentidos ("una burbuja social"). En este sentido, la Real Academia Española en "Crónica de la lengua española" [Española, 2020], indicó: "La pandemia ha provocado cambios radicales en nuestras vidas que, por supuesto, han dejado una huella imborrable en la lengua". Con estas palabras, el equipo de la Real Academia Española (RAE) hacía balance de cómo todos los hispanohablantes, al igual que el resto del mundo, hemos incorporado a nuestro vocabulario palabras que ni existían o apenas se usaban hace unos años como "covid-19", "coronavirus", "antígenos" o "confinamiento".

Razonamos que esta era una ocasión única para investigar en terrenos cuantitativos la deriva de los conceptos de palabras [Carrillo et al., 2015] que eventualmente podría conducir a un cambio semántico, y también determinar la dinámica de este proceso. Pero, ¿cómo capturamos mejor estos cambios en el significado de las palabras? En un trabajo reciente, Laurino et al. [Laurino et al., 2023] usaron como herramienta un experimento de asociación de palabras a gran escala, donde se les pidió a un grupo de participantes que respondan las asociaciones frente a un grupo de palabras targets, relacionadas o no con la pandemia (por ej, "vacuna"). A pesar de su sencillez, esta tarea constituye una herramienta metodológica poderosa para estudiar las representaciones internas y los procesos involucrados en el significado de las palabras [De Deyne and Storms, 2008], dado que las asociaciones están libres de las demandas básicas de comunicación en lenguaje, lo que permite revelar aspectos que no pueden reducirse a patrones de uso léxico [Steyvers et al., 2006]. En este sentido, el proyecto "Small World of Words" [De Deyne et al., 2019] es un estudio colaborativo a gran escala que recientemente se generó también para el español rioplatense [Cabana et al., 2023], que permitió construir una serie de redes semánticas para ciertas palabras y determinar los cambios diacrónicos, como se ejemplifica en la figura 1.1.

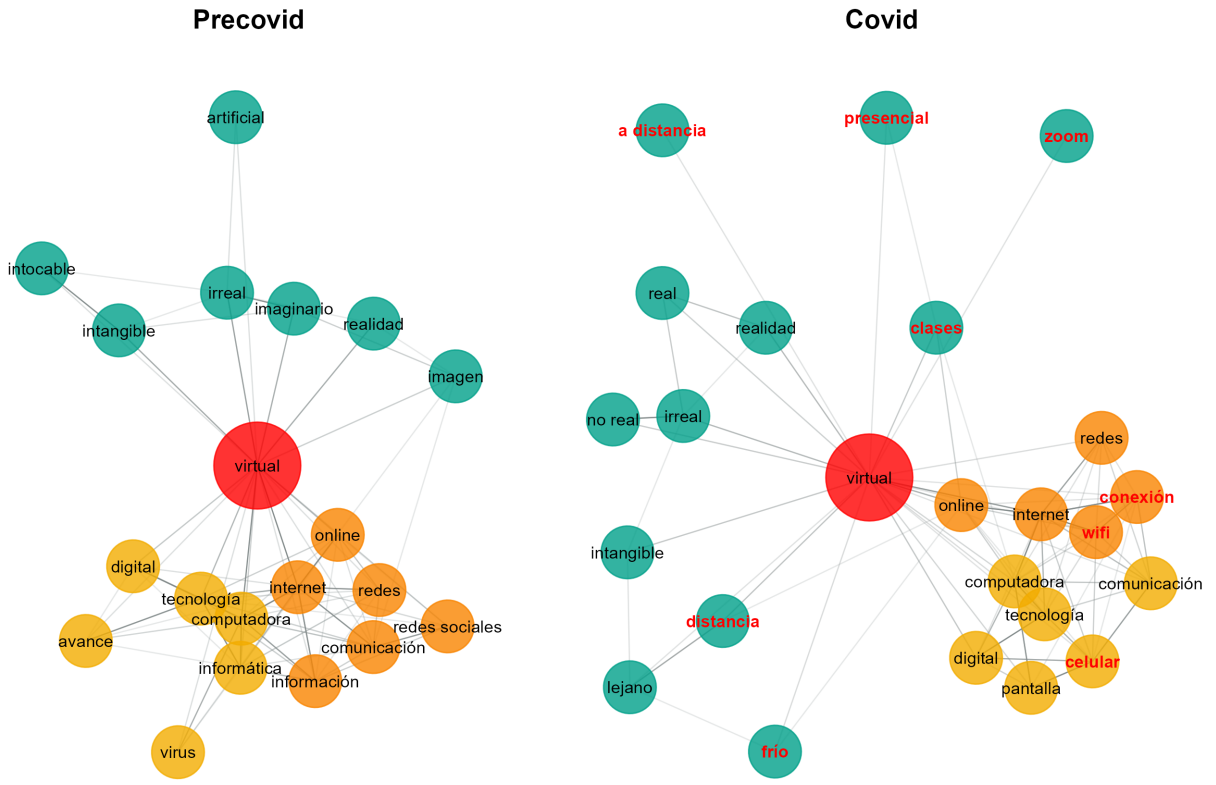


Figura 1.1: Visualización de la red correspondiente a la palabra "virtual", con sus correspondientes vértices adyacentes. Los bordes más gruesos representan mayor peso de la asociación (más veces dado como respuesta uno a otro). Las nuevas respuestas durante el período Covid se muestran en rojo. Las comunidades de nodos se muestran en el mismo color. (Adaptado de [Laurino et al., 2023])

1.1.1. Redes sociales para estimar cambios semánticos

Otra manera de analizar los cambios semánticos rápidos es a partir del análisis del lenguaje volcado en distintos textos en la web, tales como los que podemos encontrar en las redes sociales. Es innegable que vivimos en una era digital, y hoy se habla de "nativo digital" [Prensky, 2001] para referirse al grupo demográfico más dominante en el mundo, denominando así a las personas que nacieron en una nueva cultura, una cultura digital en donde los teléfonos celulares, las computadoras, los videojuegos, resultan ser "una extensión del propio cuerpo humano cuya actividad con la tecnología configura sus nociones sobre lo que es la comunicación, el conocimiento, el estudio/aprendizaje e, incluso, sus valores personales". [García et al., 2007]

En este sentido, las redes sociales han promovido nuevas formas de interacción social y han cambiado la forma en que las personas se comunican e interactúan. Redes sociales tales como Facebook, Twitter, Reddit, entre otras contienen una gran cantidad de información pública sobre las interacciones, los gustos, disgustos, e intereses de cientos de millones de individuos

que forman grandes segmentos de la sociedad global.

Desde una perspectiva de investigación, las redes sociales pueden entenderse como una especie de laboratorio viviente, que permite a los investigadores recopilar grandes cantidades de datos generados en un entorno del mundo real. [Stieglitz et al., 2014, Agrawal et al., 2014]. Aprovechando toda esta información disponible, es posible la construcción de redes semánticas similares a las obtenidas en experimentos de asociación de palabras, definiendo, por ejemplo, que dos palabras están asociadas si se utilizan en un mismo post o tweet. En particular, los datos recopilados de estas plataformas pueden servir como recursos valiosos al estudiar el efecto de la pandemia en el lenguaje, dada la disponibilidad de los mismos durante distintos momentos del desarrollo de la pandemia.

1.1.2. Redes semánticas

Las redes semánticas son una representación gráfica declarativa que puede usarse para representar conocimiento o para soportar sistemas automatizados para razonar sobre el conocimiento. Antiguamente este tipo de representación se ha utilizado durante mucho tiempo en filosofía, psicología y lingüística [Sowa et al., 1992]. Para investigar el cambio semántico, consideraremos el significado de las palabras integradas en grafos semánticos. Esto sigue el trabajo seminal de Collins y Quillian [Collins and Quillian, 1969], que representa la memoria semántica como una red o grafo en la que los nodos representan palabras, conceptos o características específicas conectadas entre sí. Desde esta perspectiva, el significado de una palabra se puede entender siempre y cuando se estudie su relación con las otras palabras en la red semántica. En las redes semánticas, dos conceptos están relacionados semánticamente si están juntos o cercanos en la red. En este estudio, obtendremos redes semánticas a partir de los datos derivados de la red social Twitter, donde dos palabras estarán asociadas si se utilizan en un mismo tweet.

Cabe mencionar que hay una extensa cantidad de literatura que demuestra que las palabras que ocurren en contextos similares tienden a tener significados similares. Este vínculo entre la similitud en cómo se distribuyen las palabras y la similitud en lo que significan se denomina hipótesis de distribución. La hipótesis distribucional fue formulada por primera vez en la década de 1950 por lingüistas como Joos en 1950 [Joos, 1950], Harris en 1954 [Harris, 1954] y Firth en 1957 [Firth, 1957], quienes notaron que las palabras que son sinónimos tendían a ocurrir en el mismo entorno. Utilizando este principio y otras metodologías, varios estudios han derivado incrustaciones de palabras (como LSA o word2vec) a través de cambios históricos conocidos (por ejemplo, Garg en 2018 [Garg et al., 2018], Hamilton en 2016 [Hamilton et al., 2016a] y Stella en 2020 [Stella et al., 2020]). En el trabajo de Hamilton [Hamilton et al., 2016b] se desarrolló una metodología robusta para cuantificar el cambio semántico mediante la evaluación de incrustaciones de palabras (se comparó PPMI, SVD y word2vec) frente a cambios históricos conocidos. Luego utilizaron dicha metodología para revelar las leyes estadísticas de la evolución semántica. Propusieron dos leyes cuantitativas del cambio semántico: (i) la ley de conformidad (la tasa de escalas de cambio semántico con una ley de potencia inversa de frecuencia de palabras); (ii) la ley de la innovación (independientemente de la frecuencia, las palabras que son más polisémicas

tienen mayores tasas de cambio semántico). (ej., figura 1.2)

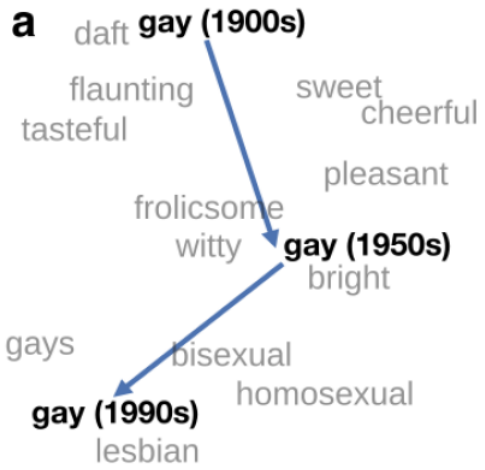


Figura 1.2: Visualización bidimensional del cambio semántico diacrónico (Adaptado de Hamilton et al., 2016b)

1.1.3. Trabajos previos sobre impacto del COVID-19 en el lenguaje

Se han realizado algunos estudios sobre cómo impactó la pandemia de COVID-19 en el significado de las palabras, utilizando corpus de redes sociales, por ejemplo, en el trabajo de Guo del 2021 [Guo et al., 2021] se explora la estabilidad semántica de las palabras utilizando tweets en inglés. Basándose en el enfoque de mapeo rotacional (alineación), demostraron que se cumple la ley de conformidad [Hamilton et al., 2016b] para cambios semánticos inducidos por eventos históricos específicos y propusieron una nueva ley de cambio semántico que indica que las palabras que aparecen con más frecuencia en un conjunto de datos específico de un evento tienen más probabilidades de cambiar su significado. También podemos mencionar otro estudio donde se investigó el cambio de uso de palabras provocado por la pandemia, en el cual se analiza el cambio de uso a corto plazo en la comunidad italiana de Reddit. En este trabajo [Signoroni et al., 2022], particularmente se habla de cambio de uso en lugar de cambio de significado. Utilizando una metodología basada en el vecindario de las palabras descrito por Gonen [Gonen et al., 2021] para el idioma italiano, se identificó que se produjo algún grado de cambio de uso, dejando como incógnita si este cambio de uso detectado se traducirá en mutaciones duraderas en el idioma.

1.2. Motivación y Objetivos

Como mecionamos anteriormente, trabajos previos estudiaron las modificaciones en la memoria de las palabras provocadas por la pandemia basándose en experimentos de asociación de palabras [Laurino et al., 2023], donde la memoria semántica se representa como una red o grafo de palabras. Como motivación de este trabajo se estudia si es posible reproducir este tipo de ensayos controlados con datos obtenidos de redes sociales como, por ejemplo, Twitter.

El objetivo de este trabajo es producir una metodología análoga que nos permita explorar cómo afectó la pandemia en el uso y el significado de un conjunto de términos, con datos obtenidos a partir de redes sociales y mediante el uso de redes semánticas. La hipótesis del trabajo es que los cambios en el uso y el significado de diferentes términos se verán reflejados en la variación de propiedades topológicas de las redes semánticas construidas en diferentes etapas del desarrollo de la pandemia. Para ello, se estudiarán las redes semánticas asociadas a diferentes términos (relacionados o no al COVID-19), a partir de una correcta definición de vínculo entre términos. En particular, el objetivo del trabajo se dividirá en los siguientes sub-objetivos:

- construir redes semánticas a partir de los datos obtenidos en redes sociales para un conjunto de usuarios en diferentes períodos de tiempo,
- analizar las redes semánticas, evaluando ciertas propiedades topológicas y comunidades emergentes,
- determinar los cambios en las redes semánticas asociados a la incorporación de nuevos significados en el marco de la pandemia y explorar la estabilidad semántica de aquellos términos cuyos significados hayan sido igualmente modificados. Nos interesa particularmente determinar este efecto a diferentes tiempos del transcurso de la pandemia.

Capítulo 2

Datos y métodos

2.1. Descripción del dataset

El conjunto de datos analizado se compone de tweets públicos originales y geolocalizados cuya fecha de publicación pertenece a los años 2019, 2020 y 2021, años que comprenden el período previo, de surgimiento y de transcurso, respectivamente, de la pandemia de COVID-19. Los tweets fueron limitados a aquellos escritos en el idioma español y correspondientes a Argentina.

Se seleccionaron 200 usuarios de estudio con el objeto de hacer un seguimiento longitudinal. En nuestro caso, consideramos que seguir a un usuario de Twitter es en cierta medida análogo a la actividad de un participante en un experimento de tipo longitudinal en el cual se analizan los cambios a través del tiempo, brindando una aproximación con mayor validez "ecológica" a nivel cognitivo [Schmuckler, 2001] .

El criterio de selección de los usuarios fue que tuvieran cierta actividad en los tres periodos mencionados (por lo menos 40 tweets/año) y que hayan utilizado alguna vez en alguno de los períodos la palabra testigo "vacuna", término relacionado con la pandemia de COVID-19 (ya que queremos medir el efecto de la pandemia, buscamos usuarios que hayan tuiteado en algún momento sobre este tema). Una vez seleccionados los usuarios, se recopilaron todas sus producciones entre el 01 de enero de 2019 y el 31 de diciembre de 2021. Los tweets y los metadatos que los acompañan se descargaron de la API de Twitter v2 que permite el acceso a todo el contenido histórico de la plataforma con fines académicos. (`https://developer.twitter.com/en/products/twitter-api/academic-research`)

En la figura 2.1 se muestra la distribución inicial de la cantidad de tweets en los tres períodos. El conjunto de datos consta de 3477422 tweets en total. Se puede observar que la distribución de los tres datasets es relativamente balanceada. Esto indica que la cantidad de tweets para cada año es del mismo orden de magnitud y por lo tanto comparable. Cabe desta-

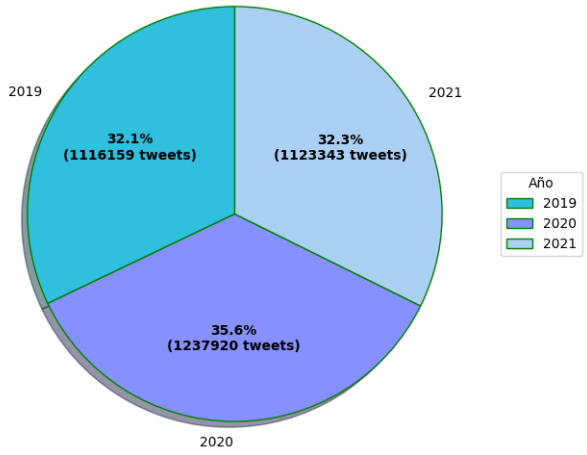


Figura 2.1: Distribución inicial de la cantidad de tweets en los tres períodos.

car que la cantidad de usuarios de Twitter en Argentina aumentó paulatinamente en los tres períodos, pasando de 5,7 a 6,8 y 7,2 millones aproximadamente en los años 2019, 2020 y 2021, respectivamente. Por otro lado, en la figura 2.1 se puede observar también un leve aumento en la cantidad de tweets producidos en el año 2020, lo cual puede deberse al hecho de que en Argentina las medidas de aislamiento social fueron mas restrictivas a lo largo del año 2020 comparado con el año 2021 y los usuarios pudieron haber utilizado un poco más de lo habitual las diferentes redes sociales como medio de expresión. (ver Stringency Index, Argentine: `https://ourworldindata.org/covid-stringency-index`)

2.2. Preprocesamiento

El preprocesamiento de los datos prepara los textos para traducir la información textual a información numérica necesaria para, por ejemplo, la construcción de las redes semánticas, entre otros análisis que se realizarán durante todo este trabajo. En general los datos de textos publicados informalmente, como es habitual en redes sociales, suelen ser muy ruidosos y, por lo tanto, procesar los datos es una tarea importante para poder realizar cualquier análisis posterior de calidad. Al aplicar los pasos detallados mas abajo intentamos reducir la cantidad de información diferente que contienen los textos, ayudando a mejorar la eficiencia de su posterior procesamiento.

A continuación se detallan los pasos de preprocesamiento realizados:

- Conversión a minúsculas: todas las letras de cada tweet se convierten a minúsculas.
- Eliminación de URL y menciones: se eliminaron todas las menciones a otros usuarios y

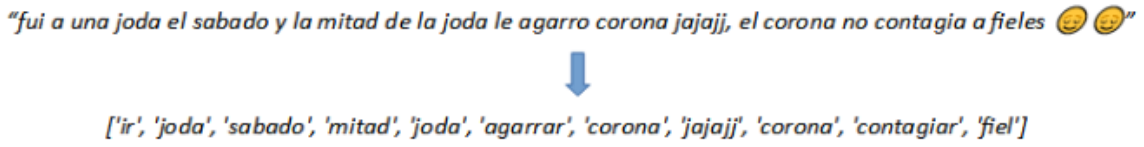


Figura 2.2: Resultado de aplicar los pasos de preprocesamiento a un texto.

URLs que aparecen en los tweets.

- Eliminación de símbolos emoji y caracteres especiales.
- Eliminación de palabras vacías: las palabras vacías son palabras frecuentes que contienen poca información, como preposiciones, pronombres y conjunciones. Se eliminaron aquellas palabras pertenecientes a la lista de palabras vacías en español de la biblioteca NLTK. [Bird et al., 2009]
- Tokenización: la tokenización consiste en dividir los tweets en entidades más pequeñas llamadas tokens. Se puede pensar al token como la unidad para procesamiento semántico. En principio tomamos como tokens a las palabras presentes en los tweets eliminando las palabras vacías descritas en el punto anterior.
- Normalización de tokens mediante lematización: la lematización implica realizar un análisis del vocabulario y su morfología para retornar a la forma básica de la palabra (sin conjugar, en singular, etc). Se utiliza la librería del departamento de NLP de la universidad de Stanford *Stanza* [Qi et al., 2020]. Normalizar el texto significa convertirlo a una forma estándar más conveniente para su procesamiento. La técnica de lematización reduce la cardinalidad del vocabulario, por lo que se asocia un único token para diferentes formas flexionadas (entreno, entrenarás, entrenaría → entrenar). De este modo, permite eliminar cierta redundancia en el análisis.

En la figura 2.2 se muestra un texto luego de aplicar los pasos de preprocesamiento. Se puede observar la disminución en el número de tokens respecto de la cantidad de palabras presentes en el tweet original. También podemos destacar que se elimina cierta redundancia al colapsar palabras con significado similar durante el paso de normalización. Luego de aplicar las tareas de preprocesamiento, obtenemos finalmente que cada usuario produjo en promedio 5512 tweets para el año 2019, 6125 para el año 2020 y 5526 para el año 2021.

2.3. Construcción de redes semánticas

2.3.1. Redes semánticas ponderadas

Luego de preprocesar los tweets buscamos generar redes semánticas que reflejen el contenido de los mismos y construyendo un objeto de estudio análogo al explorado en los experimentos

de asociación de palabras [Laurino et al., 2023].

Una red semántica, o esquema de representación de red, busca representar el conocimiento lingüístico (los conceptos y sus interrelaciones) mediante un grafo. Un grafo (G) es un conjunto no vacío de objetos llamados vértices (V) o nodos, que en nuestro caso representa los términos, y una selección de pares de vértices llamados aristas o enlaces (E), que en nuestro caso representan la relación entre dos términos. En este trabajo utilizaremos grafos ponderados, en los cuales los enlaces tienen asociados una ponderación o peso ω. Por ejemplo, en nuestro caso se define como $\omega(i, j)$ al peso del enlace entre los términos w_i y w_j.

Para este trabajo, buscaremos que el peso del enlace sea proporcional a qué tan vinculados están dos términos mencionados en el conjunto de datos de tweets. Para ello exploramos dos propuestas para la definición del peso de los enlaces, el conteo de coocurrencias (Co) y la *normalized pointwise mutual information* ($NPMI$):

1. Conteo de coocurrencias (Co): para el conjunto de tweets de un dado usuario, el peso de conexión entre dos palabras va a estar dado por la cantidad relativa de tweets en la cual dichas palabras coocurren. En nuestro caso, esta coocurrencia se define a partir de una variable binaria que indica si dos palabras aparecen juntas en un dado tweet, independientemente de si se repiten dentro del mismo tweet. Matemáticamente, la coocurrencia entre dos palabras se define como:

$$Co(i, j) = \frac{f_{ij}}{N} \equiv p(w_i, w_j)$$

   donde f_{ij} mide la cantidad de tweets en los que la palabra w_i aparece en el mismo tweet que la palabra w_j y N es el número total de tweets del usuario (a lo largo del trabajo definiremos esta cantidad para distintos periodos estudiados). Notar que esta cantidad es una estimación de la probabilidad conjunta $p(w_i, w_j)$ de que las palabras w_i y w_j aparezcan en un mismo tweet de un dado usuario.

2. Normalized Pointwise Mutual Information ($NPMI$) [Bouma, 2009]: debido a la distribución tipo *power-law* [Piantadosi, 2014] que presentan la frecuencia de palabras en el lenguaje natural, algunas palabras pueden a menudo ser observadas en la vecindad de otras simplemente porque son muy frecuentes, resultando posiblemente en una coocurrencia alta sin que esto sea indicativo de una relación en particular. Es por ello que una medida más adecuada para medir la relación entre palabras está dada por la *pointwise mutual information* (PMI). Esta es una medida de cuánto difiere la probabilidad real de una coocurrencia particular de eventos $p(w_i, w_j)$ respecto de lo que esperaríamos sobre la base de las probabilidades de los eventos individuales y la suposición de la independencia $p(w_i)p(w_j)$. Partiendo de la PMI, se define a la $NPMI$ como una versión normalizada, definida matemáticamente como:

$$NPMI(i, j) = \frac{1}{-log(p(w_i, w_j))} log\Big(\frac{p(w_i, w_j)}{p(w_i)p(w_j)}\Big)$$

donde en nuestro caso $p(w_i, w_j)$ es la cantidad de tweets en los que w_i y w_j coocurren dividido sobre los N tweets de un usuario, $p(w_i)$ es el número total de tweets que contienen la palabra w_i dividido por los N y $p(w_j)$ es el número total de tweets que contienen la palabra w_j dividido por N.

Bajo esta última definición, $NPMI$ toma valores entre -1 y 1. Por ejemplo, algunos valores orientativos que se obtienen son:

- Cuando las palabras w_i y w_j sólo aparecen juntas, $NPMI(w_i, w_j) = 1$. Este es el caso de una asociación perfecta.
- Cuando las palabras no tienen ninguna asociación entre si, $p(w_i, w_j) \sim p(w_i)p(w_j)$, resultando en $NPMI(i, j) \sim 0$. En este caso la aparición de la palabra w_i está completamente descorrelacionada con la aparición de la palabra w_j.
- Finalmente, cuando dos palabras aparecen siempre por separado y nunca lo hacen juntas, se define $NPMI(i, j) = -1$, ya que se aproxima a este valor cuando $p(w_i, w_j)$ se aproxima a 0 para $p(w_i)$ y $p(w_j)$ distintos de cero. Este es el caso de palabras que son mutuamente excluyentes.

A modo de ejemplo, a continuación se muestran los tokens procesados pertenecientes a 4 tweets, sobre los cuales se ensayan los diferentes tipos de pesos definidos:

- $[protocolo, legal]$
- $[quedate, casa, covid, quedate, casa, covid, quedate, casa, covid]$
- $[casa, cuidado, covid]$
- $[aislado, casa, covid]$

Por ejemplo, el cálculo de NPMI para los pares $legal - protocolo$ y $casa - cuidado$ resulta en:

$$NPMI(legal, protocolo) = \frac{log(\frac{p(legal,protocolo)}{p(legal).p(protocolo)})}{-log(p(legal, protocolo))} = \frac{log(\frac{1/4}{1/4 \cdot 1/4})}{-log(1/4)} = 1$$

$$NPMI(casa, cuidado) = \frac{log(\frac{p(casa,cuidado)}{p(casa).p(cuidado)})}{-log(p(casa, cuidado))} = \frac{log(\frac{1/4}{3/4 \cdot 1/4})}{-log(1/4)} = 0{,}2075$$

En la tabla 2.1 se muestra el resultado de los diferentes pesos de los posibles pares de tokens correspondientes a los 4 tweets dados de ejemplo. En este ejemplo, destacamos que el par $legal - protocolo$ posee un conteo de baja ocurrencia respecto del par $casa - cuidado$. Sin embargo, ambos pares son igual de informativos, debido a que las palabras *protocolo* y *legal* sólo aparecen una única vez en los tweets, pero esa única vez aparecen juntas, mientras que *casa* y *covid* aparecen en 3 de 4 tweets y también siempre que aparecen lo hacen juntas.

Pesos de conexión entre nodos			
w_i	w_j	Co	NPMI
legal	protocolo	0.25	1.0000
casa	covid	0.75	1.0000
casa	quedate	0.25	0.2075
covid	quedate	0.25	0.2075
covid	cuidado	0.25	0.2075
casa	cuidado	0.25	0.2075
aislado	covid	0.25	0.2075
aislado	casa	0.25	0.2075

Cuadro 2.1: Diferentes pesos ensayados

En la figura 2.3 se observa el grafo obtenido a partir del ejemplo de los 4 tweets arriba mencionados, utilizando como peso de los enlaces el $NPMI$ calculado en la tabla 2.1. La figura 2.3 representa el primer ejemplo de red semántica construida a partir de un conjunto de tweets. A veces los conteos bajos de coocurrencias son muy informativos y los conteos altos de coocurrencias no lo son. Por ejemplo, cualquier palabra va a tener una coocurrencia relativamente alta si tiene contextos muy comunes (por ej. ser, tener, etc.), pero esto no nos dirá mucho sobre lo que significa esa palabra [Jurafsky and Martin, 2008]. Por ello es necesario identificar cuándo los conteos de coocurrencia son más probables de lo que esperaríamos por azar.

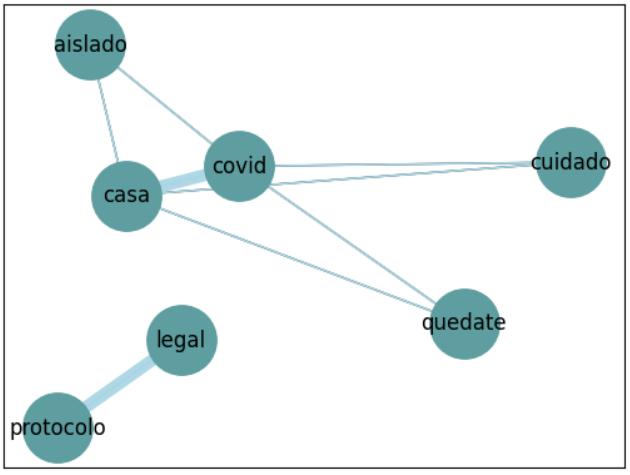


Figura 2.3: Grafo generado a partir de los 4 tweets ejemplo, a mayor grosor de las aristas mayor valor de peso $NPMI$.

2.3.2. Red semántica agregada

Con el objetivo de describir las relaciones semánticas del conjunto de usuarios, se procede a agregar los grafos creados para cada usuario, a fin de obtener un grafo resultante que represente la red semántica del total de la muestra. Específicamente, se procedió de la siguiente manera:

- Se construyen las redes semánticas para cada usuario (tanto con el peso calculado con $NPMI$, como con la coocurrencia Co). Recordemos que por definición, estas redes son pesadas, con los pesos entre -1 y 1 para $NPMI$ (entre 0 y 1 para Co).
- Se agregan los resultados para todos los usuarios sumando los pesos pertenecientes al mismo par de palabras. Notar que esta suma incluye tanto términos positivos (que refuerzan la vinculación entre palabras) y negativos (que inhiben la relación entre las palabras).
- Los pesos resultantes se dividen por la cantidad total de usuarios $M = 200$. De esta manera, los pesos del grafo resultante también caen en el intervalo $[-1, 1]$ para $NPMI$ ($[0, 1]$ para Co).

La metodología descrita para la construcción del grafo agregado tiene varias ventajas. En principio, en caso que existiera una desbalanceo en la cantidad de tweets publicados por cada usuario, esperamos que su efecto esté atenuado debido a que el grafo calculado para cada usuario ya contempla una normalización respecto de la cantidad de tweets publicados. La existencia de un mayor número de tweets para un dado usuario permite obtener una red semántica con más detalle, pero no afecta al momento de agregarla a la red resultante. Segundo, en la red resultante queda naturalmente ponderada la relación entre palabras que utilizan varios usuarios, por lo que consideramos que los pesos más grandes resultantes son representativos del conjunto. En este sentido, si solamente en unos pocos usuarios aparece una relación entre palabras, esta quedará atenuada al normalizar por M.

Como se verá en el siguiente capítulo, las redes semánticas resultantes del conjunto de usuarios se calcularon para los años 2019, 2020 y 2021 por separado, a fin de estudiar la evolución temporal de la red.

Debido a la capacidad de *normalized pointwise mutual information* ($NPMI$) de detectar mejor las relaciones entre palabras respecto de Conteo de coocurrencias (Co), a partir de ahora nos enfocaremos en trabajar principalmente con grafos ponderados utilizando a $NPMI$ como peso de los enlaces.

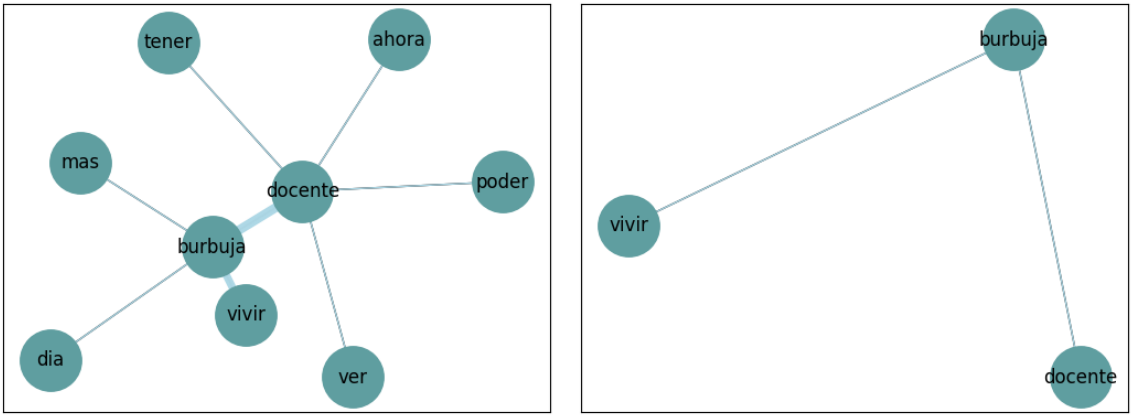
2.3.3. Binarización y reducción del tamaño de la red

A fin de simplificar el análisis de las redes, se procedió a eliminar enlaces que consideramos no informativos y transformar las redes pesadas construidas en redes binarias. En principio, se filtraron los enlaces resultantes con $NPMI < 0$, debido a que valores negativos indican básicamente que la coocurrencia de dos palabras se da con menos frecuencia que lo esperado

por azar.

Para binarizar y reducir el tamaño de las redes se aplicó el algoritmo propuesto en [Serrano et al., 2009] que propone una manera de identificar las conexiones verdaderamente relevantes que forman la "columna vertebral"de la red, preservando los enlaces cuyo peso corresponde a desviaciones estadísticamente significativas respecto de un modelo nulo. Los detalles de este algoritmo están descriptos en la sección A.1. Básicamente, se preservan los enlaces en los cuales se descarta que su peso se haya dado por azar respecto de un modelo nulo ($p_{val} = 0{,}05$).

Finalmente, trabajaremos con la componente gigante de las redes resultantes. En la figura 2.4 se puede observar el resultado de un ejemplo antes y después de aplicar el algoritmo de filtrado de enlaces, extrayendo las conexiones significativas, y quedándonos con la componente gigante. En este ejemplo, las conexiones significativas que permanecen son entre los nodos "burbuja", "docente" y "vivir", que constituyen la componente gigante de la red binaria resultante (descartamos el peso de los enlaces al momento de decidir si quedan en la red o no).



(a) Antes de aplicar algoritmo, mayor grosor de las aristas indican mayor valor de peso $NPMI$

(b) Componente gigante de la red binaria resultante luego de aplicar algoritmo

Figura 2.4: Ejemplo del subgrafo que contiene la palabra "burbuja"del año 2021, antes y después de aplicar algoritmo presentado en [Serrano et al., 2009]. Notar que solo preservamos los nodos que caen en la componente gigante de la red fitrada y binarizada.

2.3.4. Redes resultantes a analizar

En la tabla 2.2 se muestra el resultado de reducir el tamaño, indicando el valor inicial de nodos de los grafos agregados, luego de aplicar el algoritmo de filtrado y finalmente la cantidad de nodos de que quedan en la componente gigante. Obsérvese que en todos los casos la componente gigante posee aproximadamente el 12 % de la cantidad de nodos originales. El esquema de reducción utilizado codifica una gran cantidad de información relevante, debido a que la red reducida no sólo contiene los enlaces con mayor peso, sino también los enlaces que son significativos respecto de una hipótesis nula.

Cantidad de nodos			
Año	Iniciales	Luego del algoritmo de filtrado	Componente gigante
2019	173579	21790	20702
2020	176465	22893	21640
2021	162575	21376	20225

Cuadro 2.2: Resultado de reducir el tamaño de la red.

Capítulo 3

Análisis de las redes semánticas

El análisis de los grafos ayuda a investigar la estructuras de los mismos e identificar las características destacadas de los datos que los componen. A continuación se detallan la propiedades topológicas de las redes semánticas y las comunidades emergentes.

3.1. Propiedades topológicas

3.1.1. Métricas generales

Evaluamos diferentes métricas para caracterizar las redes semánticas, entre ellas, la cantidad de nodos total, el grado medio (número de aristas promedio), el diámetro de la red, el coeficiente de clustering y la asortatividad.

Como primera característica general de la red podemos destacar el numero total de nodos. Cabe recordar que los nodos representan palabras y las aristas representan la relación semántica entre ellas. El grado medio o número de aristas promedio, se define como el número de aristas totales sobre el número total de nodos. El diámetro de la red se define como la máxima excentricidad, donde la excentricidad de un nodo V es la distancia máxima de dicho nodo a todos los demás nodos en un grafo.

Por otro lado, el coeficiente de clustering es una medida de la cohesividad local, definida como la probabilidad de que dos nodos que tienen un vecino común también estén conectados entre ellos. Se calcula como el cociente entre el número de aristas entre vecinos observadas para V_i (E_i) y el número máximo de aristas posibles (E_{max}) entre dichos vecinos. Finalmente, la asortatividad mide la similitud de las conexiones en el grafo con respecto al grado del nodo. La asortatividad se describe como la tendencia de los nodos a conectarse a nodos con propiedades similares, propiedad conocida como homofilia.

3.1.2. Medidas de centralidad

Las medidas de centralidad definidas sobre una red se utilizan para identificar nodos estructuralmente importantes. En particular, la medida más simple y más utilizada es el grado de un nodo. El grado k_i de un nodo i se define simplemente como el número total de sus conexiones.

Por otro lado, podemos señalar que otras medidas de centralidad como por ejemplo: intermediación (*betweenness*), cercanía (*closeness*), centralidad de autovectores (*eigenvector centrality*)[Newman, 2018]. La utilidad analítica de cada una depende en gran medida del contexto del cual surge la red y el tipo de relaciones que se analizan. Una medida puede ser mas útil que otra según los objetivos de cada análisis. En particular, nosotros utilizamos la medida de *eigenvector centrality*. Esta medida define que un nodo es importante si está vinculado a otros nodos importantes.

La *eigenvector centrality* se calcula de la siguiente manera. Sea $A = (a_{i,j})$ la matriz de adyacencia de un grafo, la centralidad del vector propio x_i del nodo i viene dada por:

$$x_i = \frac{1}{\lambda} \sum_k a_{k,i}\, x_k$$

donde $\lambda \neq 0$ es una constante. En forma matricial tenemos:

$$\lambda x = xA$$

Utilizaremos esta medida de centralidad en las siguientes secciones para filtrar los principales nodos de las comunidades detectadas y de esta manera identificar el tema al cual se podría referir cada comunidad (ver por ejemplo figura 3.1).

3.1.3. Resultados y análisis

En el cuadro 3.1 se observa el resultado de las diferentes métricas sobre las redes semánticas definidas para los tres períodos estudiados.

Caracterización de la red			
Medida	Año 2019	Año 2020	Año 2021
Cantidad de nodos	20702	21640	20225
Grado medio	22.7349	23.7645	22.5031
Diámetro de la red	10	11	11
Clustering global	0.4249	0.4347	0.4191
Asortatividad	-0.2860	-0.2870	-0.2821

Cuadro 3.1: Distintas medidas que caracterizan las redes semánticas

Según se detalla en el Cuadro 3.1 se observa que el número de nodos totales se mantuvo relativamente constante, observándose un leve aumento en el año 2020. Por otro lado, también

se observa un leve aumento en el grado medio (número de aristas promedio) para el año 2020. Aunque la cantidad de nodos fue levemente mayor en el 2020 la cantidad de interacciones entre ellos aumentó más que en los otros años.

El diámetro de la red pasó de 10 en el año 2019 a 11 para los años 2020 y 2021. Esto puede ser debido a que en el año 2020 el tamaño de la red aumentó, por lo que es posible que esto haya provocado el aumento del diámetro. Por otro lado para el año 2021 el numero de nodos totales disminuyó junto con el grado medio sin observarse una disminución en el diámetro de la red. Según [Leskovec et al., 2007] es esperable que en redes reales se observe una disminución del diámetro a medida que aumenta el tamaño de la red. Sin embargo, las variaciones observadas aquí en el tamaño de la red en función del tiempo son prácticamente despreciables y por lo tanto tampoco observamos grandes variaciones en el diámetro de las mismas.

En sintonía con lo anterior, la red del año 2020 posee un coeficiente de clustering levemente mayor que las redes de los años 2019 y 2021, indicando que los términos poseen una interrelación levemente mayor en 2020 que las registradas en los otros períodos. Finalmente, en los tres casos los valores de asortatividad sugieren que no hay asociaciones preferenciales por un lado entre nodos de alto grado y por el otro entre los de bajo grado. Esto puede deberse a como son redes formadas por palabras, palabras comunes suelen asociarse a palabras no comunes.

Como conclusión del análisis de las propiedades topológicas de las redes podemos observar que no hay diferencias mayores al 10 % en las métricas medidas para los distintos años. Esto por un lado indica que no hay grandes variaciones en la estructura global de la red en función del tiempo, resultando en que los diferentes períodos son en algún modo comparables, dado que todas las medidas ensayadas son del mismo orden de magnitud. Como observaremos en la siguiente sección, observaremos diferencia entre los tres períodos al realizar un análisis de comunidades sobre las redes.

3.2. Detección de comunidades

La detección de comunidades es importante no sólo para caracterizar el grafo, sino que además proporciona información sobre la forma y funcionalidad de la red. Una comunidad es un grupo de vértices que comparten propiedades comunes y en ciertos contextos tienen funciones similares dentro del grafo. En este trabajo vamos a utilizar el algoritmo de Louvain [Blondel et al., 2008] para la detección de comunidades, el cual está basado en la maximización de la modularidad. La modularidad es una función que mide la bondad de la división en un grafo en comunidades [Newman and Girvan, 2004]. El algoritmo de Louvain es ampliamente utilizado en problemas de detección de comunas y es una de las técnicas más eficientes entre las basadas en optimización de modularidad. En este trabajo, utilizamos este algoritmo para la detección de comunidades dentro de la componente gigante de cada período. Para ver mas de detalles de éste y otros algoritmos ensayados dirigirse a la sección A.2.

3.2.1. Resultados y análisis

Luego de aplicar el algoritmo de Louvain para las redes de cada período, se seleccionaron las diez comunidades que contienen mayor cantidad de nodos a fin de identificar las estructuras emergentes más importantes. Por ejemplo para el año 2019, las diez comunidades más grandes contenían el 97 % del total de los nodos. A fin de caracterizar cada comunidad, se seleccionaron los primeros 30 nodos de cada una de ellas el valor de *eigenvector centrality* obtenido. Con base en estos nodos, cada comunidad fue etiquetada con un tema al que podría referir.

Por ejemplo, en los cuadros 3.2 y 3.3 se muestran la principales palabras relacionadas a los temas que llamamos *alimentación y nutrición* y *Pandemia COVID-19*. Como se puede observar, las palabras más importantes guardan una estrecha relación ente sí, permitiendo etiquetarlas con un tema bien definido. Más aún, es posible seguir ciertos temas a través de los años debido a la repetición de relaciones entre palabras en cada período.

Año	Principales palabras, orden según eigenvector centrality
2019	comer, comprar, leche, queso, pan, carne, agua, huevo, rico, dulce, precio, cafe, fruta, vino, arroz, asado, papas, aceite, pollo, harina
2020	poner, salir, dia, quedar, buen, casa, comprar,bien, venir, agua, comer, despues, tomar, llevar, usar, leche, comida, huevo, rico, blanco
2021	comer, comprar, carne, queso, rico, leche, pan, quedar, agua, comida, huevo, asado, fruta, cafe, vino, pollo, aceite, horno, fideo,dulce

Cuadro 3.2: Palabras en la comunidad relacionas a *alimentación y nutrición*

Año	Principales palabras, orden según eigenvector centrality
2020	coronaviru, caso, covid, salud, persona, llegar, cuarentena, virus, medico, primero, vacuna, contagio, hospital, chino, paciente, via, enfermedad, contagiar, control, viajar
2021	vacuna, argentino, covid, pais, vacunar, llegar, dosis, primero, salud, caso, vacunacion, nuevo, vacunado, usar, mayor, medico, sputnik, riesgo, virus, aplicar, coronaviru

Cuadro 3.3: Palabras en la comunidad relacionadas a *Pandemia COVID-19*

En la figura 3.1 se presentan 3 de las comunidades mas relevantes para cada período. Se observa que no existen heterogeneidades dentro de las comunas. Esto se refleja en la homogeneidad de los enlaces de las palabras más centrales dentro de cada comuna. Al ver las principales palabras representadas en cada comunidad se puede observar también que es adecuado la definición tema al cual se refieren. Por ejemplo, en el caso de la comunidad relacionada al tema política, que se muestra en el panel (h) de la figura 3.1, se puede observar que contiene un conjunto de palabras relacionadas a nombres de políticos, procesos electorales, distritos, entre otras palabras.

En el cuadro 3.4 se muestran algunas de las principales comunidades encontradas y el porcentaje de la red que representan para cada período. Se puede observar que ciertos temas se repiten a lo largo del tiempo, mientras que otras emergen o desaparecen según el período. Por ejemplo, que el tema relacionado a la pandemia del $COVID-19$ lógicamente no se encuentra en el año 2019 previo al estallido de la pandemia. También se puede ver que el en año 2020 el tema *fútbol* no se encontraba dentro de las principales comunidades, quizás esto debido a que gran parte de ese año se encontraban todas las actividades deportivas prohibidas.

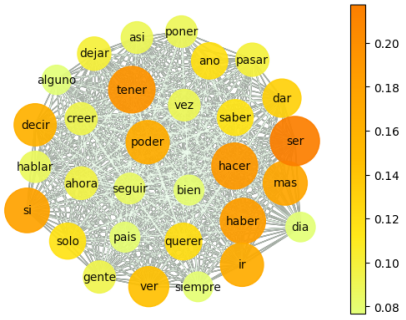
Tema	Año 2019	Año 2020	Año 2021
Palabras comunes	40,1 %	38,9 %	33,5 %
Pandemia, Covid-19	-	8,9 %	7,4 %
Vida, familia	14.5 %	14,4 %	13,0 %
Alimentación, nutrición	6,4 %	14,9 %	6,8 %
Política	17,3 %	-	17,6 %
Fútbol	6,8 %	-	2.8 %
Palabras en inglés	8,5 %	9.6 %	7,4 %
% Total de nodos	93,6 %	86,7 %	88,5 %

Cuadro 3.4: Principales temas y porcentaje de nodos respecto a la componente gigante que contiene esa comunidad

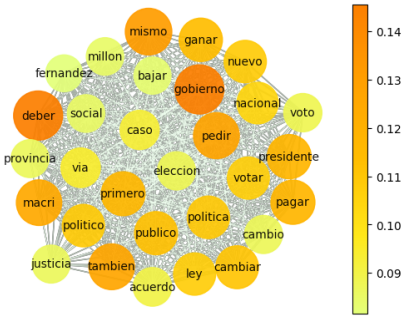
Siguiendo con el análisis del cuadro 3.4, podemos ver que el tema *vida, familia*, que se refiere a cuestiones cotidianas (ver panel (c) de la figura 3.1, se mantiene durante los tres períodos con un porcentaje relativamente similar. Esto mismo ocurre con la comunidad de *palabras en inglés* debido a que esta comunidad tampoco posee ningún motivo particular por el cual podría variar. Como se mencionó anteriormente, el tema *Pandemia COVID-19* emerge en el año 2020 debido al impacto mundial de este suceso. Notar que a su vez empieza a disminuir el porcentaje en el año 2021. Por otro lado, el tema *política* se encuentra entre las primeras comunidades de los años 2019 y 2021 con porcentajes similares. Esto podría ser debido a que dichos años fueron electorales, mientras que en el año 2020 no hubo elecciones. En relación al *fútbol* y la diferencia de porcentajes entre 2019 y 2021 podríamos decir que dicho tema en el año 2019 competía con la comunidad *política*, pero en el año 2021 competía además con el tema *Pandemia COVID-19*. Cabe recordar que el porcentaje es prácticamente nulo para el año 2020 posiblemente por la falta de actividades deportivas. Podemos también mencionar la variación que sufrió el tema *alimentación, nutrición* observándose que aparece con más fuerza en el año de mayor confinamiento 2020. Los cambios en los patrones de consumo, las dificultades financieras y la reducción en la movilidad (disminución de la actividad física) pueden desencadenar este aumento respecto de los años 2019 y 2021.

Notar que en el cuadro 3.4 podemos ver que las 7 comunidades analizadas contienen entre el 86 y el 93 % del total de los nodos de la componente gigante del año al que pertenecen, constituyendo un conjunto de temas bien representativos de los sucedido en cada período (ver en el apéndice A.2.2 otras comunidades encontradas).

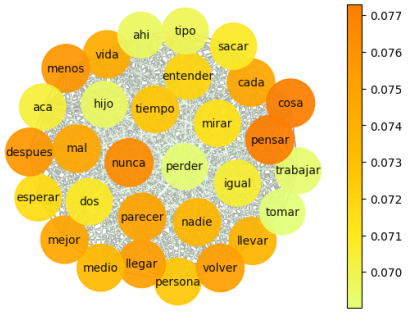
Finalmente, podemos concluir de la sección que si bien no encontramos cambios significativos en las estructuras topológicas de las redes en general, los cambios semánticos sí pueden detectarse en la aparición o desaparición de comunidades en la red semántica, debido a que las mismas dan idea del uso de un conjunto de palabras para referirse a un objeto en particular. Ejemplo de esto se observó en la emergencia de la comunidad relacionado a la pandemia del *COVID-19* o la desaparición de la comunidad etiquetada como *fútbol* durante el año 2020.



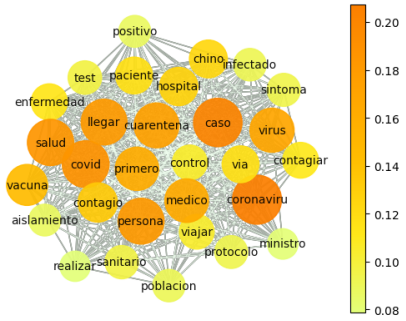
(a) 2019, Comunidad *palabras comunes*



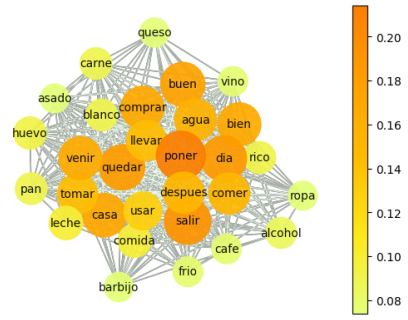
(b) 2019, Comunidad *política*



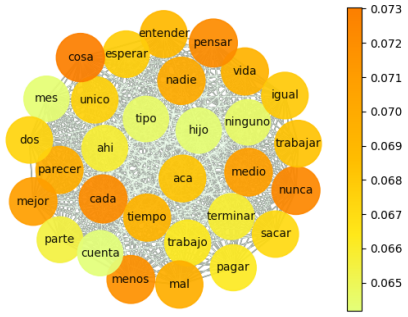
(c) 2019, Comunidad *vida, familia*



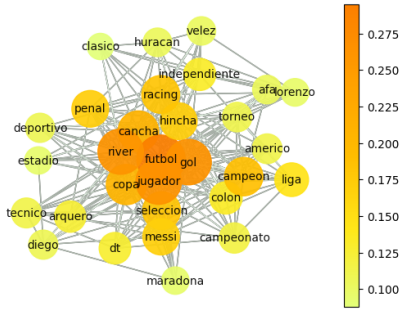
(d) 2020, Comunidad *pandemia COVID-19*



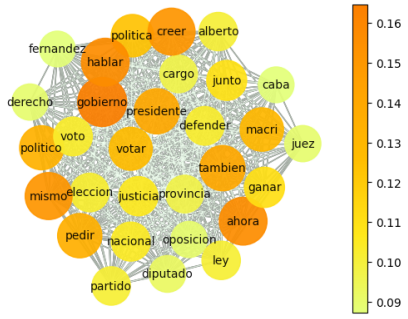
(e) 2020, Comunidad *alimentación, nutrición*



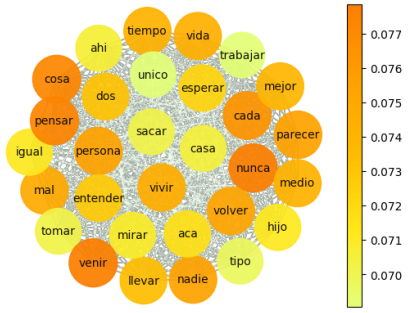
(f) 2020, Comunidad *vida, familia*



(g) 2021, Comunidad *fútbol*



(h) 2021, Comunidad *política*



(i) 2021, Comunidad *vida, familia*

Figura 3.1: Principales comunidades obtenidas en los tres períodos, el tamaño y color indican el valor de *eigenvector centrality* de los nodos. Las tres comunidades son representativas del total del sub conjuto de 10 comunidades analizadas

Capítulo 4

Evaluación de términos en el tiempo

Hasta este punto, en el capítulo anterior, podríamos decir que realizamos un análisis macroscópico de la red semántica, el cual corresponde al análisis de las características topológicas, luego realizamos un análisis mesoscópico, en el cual se realizó la detección de comunidades, y a continuación, en el presente capítulo realizaremos un análisis microscópico, que es el seguimiento de ciertas palabras en el tiempo.

4.1. Listado de palabras seleccionadas

Para determinar y cuantificar los cambios transcurridos en los diferentes períodos de tiempo en las redes semánticas asociados a la incorporación de nuevos significados debido a la emergencia de la pandemia de COVID-19, definimos un listado de palabras que denominamos pandémicas, las cuales esperamos que se vean afectadas por la misma, y un grupo de palabras control que esperamos su significado se mantenga estable en el tiempo.

Basados en el experimento de asociación de palabras [Laurino et al., 2023], a continuación se detallan las palabras seleccionadas a las que en particular les hicimos un seguimiento:

- Palabras control: *acuerdo, claro, cosa, palabra, historia, lluvia, rueda.*
- Palabras pandémicas: *aislado, alcohol, burbuja, corona, virtual, inmunidad, protocolo.*

A su vez, se analizó la variación de ciertas métricas en el tiempo de dos conjuntos de palabras mas extensos. Se utilizó un conjunto de 95 palabras pandémicas y un conjunto de 102 palabras control, también definidas en [Laurino et al., 2023].

4.2. Seguimiento de comunas asociadas a términos de interés

En primera instancia se evaluó en qué comunidades se encuentran las palabras control y pandémicas en función del tiempo. En el cuadro 4.1 se puede observar en qué comunidad se encontraban los términos analizados para los tres períodos estudiados. Por ejemplo, las palabras "aislado" e "inmunidad" en el año 2019 se encontraban en las comunidades *palabras comunes* y *política*, respectivamente, y luego ambas pasaron a la comunidad *pandemia COVID-19* en los años 2020 y 2021. Por otro lado, palabras que tomamos de control, tales como "claro", "historia" y "cosa", entre otras, se mantuvieron en la misma comunidad en los tres períodos.

Temas de Comunidades donde se encuentra la palabra analizada			
Palabra	Año 2019	Año 2020	Año 2021
acuerdo	política	vida, familia	vida, familia
claro	vida, familia	vid, familia	vida, familia
cosa	vida, familia	vida, familia	vida, familia
palabra	vida, familia	vida, familia	vida, familia
historia	vida, familia	vida, familia	vida, familia
lluvia	vida, familia	alimentación, nutrición	palabras comunes
rueda	palabras comunes	palabras comunes	palabras comunes
aislado	palabras comunes	pandemia COVID-19	pandemia COVID-19
alcohol	palabras comunes	alimentación, nutrición	alimentación, nutrición
burbuja	palabras comunes	alimentación, nutrición	pandemia COVID-19
corona	palabras comunes	pandemia COVID-19	palabras comunes
virtual	palabras comunes	vida, familia	palabras comunes
inmunidad	política	pandemia COVID-19	pandemia COVID-19
protocolo	política	pandemia COVID-19	pandemia, COVID-19

Cuadro 4.1: Comunidades donde se encuentran palabras control y pandémicas

Estos resultados muestran que los cambios semánticos pueden detectarse en el cambio de vecindad de las palabras, en este caso viendo a groso modo la comunidad en la que caen para cada período.

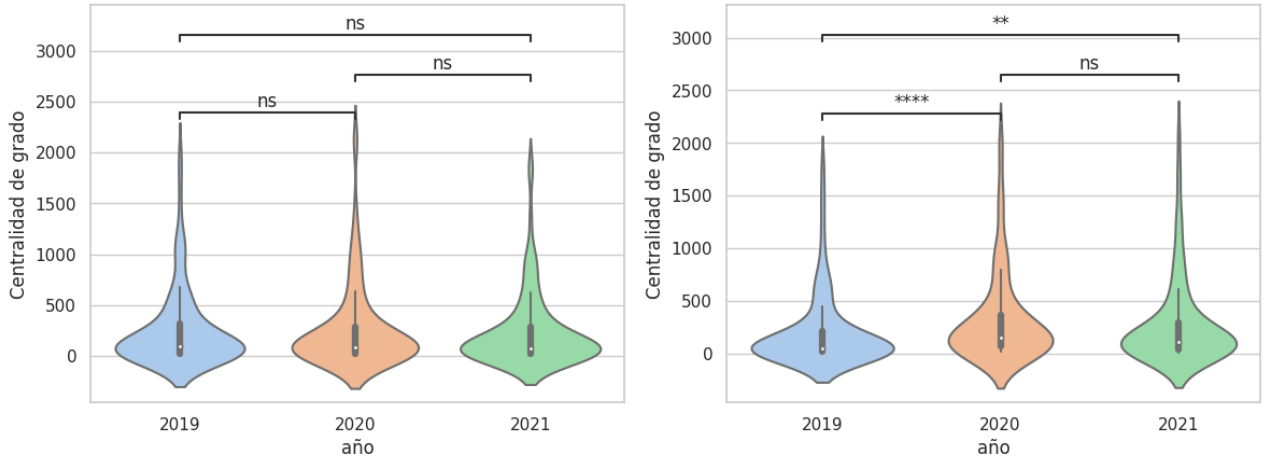
4.3. Variación de la posición en la red de los términos de interés

En esta sección, evaluamos los cambios en medidas de centralidad y variaciones en el conjunto de primeros vecinos en función del tiempo para el conjunto extendido de palabras control y pandémicas.

4.3.1. Medidas de centralidad

Para cada palabra de los conjuntos de control y pandémicas se midió el valor de centralidad de grado y centralidad de *eigenvector* en los tres períodos estudiados. Para evaluar las diferencias entre los distintos períodos de las distribuciones de las centralidades obtenidas se utilizó el test estadístico no paramétrico Mann Whitney Wilcoxon bilateral con corrección de Bonferroni para contrastar si las centralidades de grado de palabras pandémicas y control en los diferentes períodos provienen de poblaciones equidistribuidas.

En la figura 4.1 se observa las distribuciones de las centralidad de grado de los dos conjuntos en los diferentes períodos. En el panel (a) se observa que, razonablemente, la centralidad de grado de las palabras control no exhiben un cambio significativo en los diferentes períodos. Por otro lado, en el panel (b) se observa que las centralidades de grado de las palabras pandémicas sí exhiben un cambio significativo entre los años 2019-2020 y 2019-2021, sin observarse un cambio significativo en la transición 2020-2021. También se observa que las centralidades de grado de las palabras pandémicas de los años 2020 y 2021 en promedio son mayores que las del año 2019. Esto demuestra que estas palabras se vieron afectadas por la emergencia de la pandemia (2019-2020), volviéndose más centrales, que en este caso se refiere a un aumento en la cantidad de vecinos en la red semánticas, lo que significa un aumento en la cantidad de palabras vinculadas, sin volver en el 2021 a su estado basal del 2019.



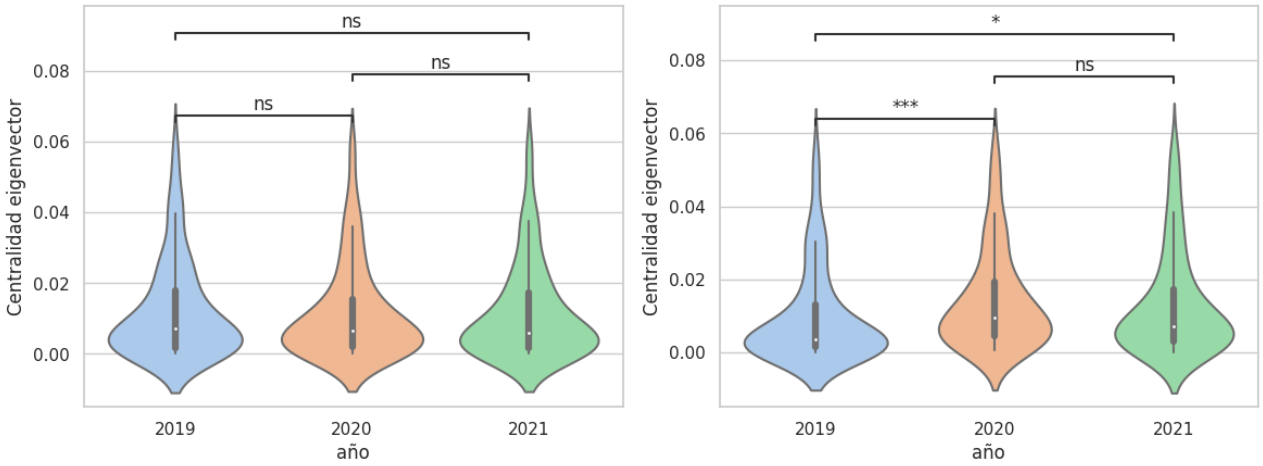
(a) Centralidad de grado de palabras control, años 2019, 2020 y 2021

(b) Centralidad de grado de palabras pandémicas, años 2019, 2020 y 2021

Figura 4.1: Comparación de las Centralidades de grado entre palabras pandémicas y control en los tres períodos. Los asteriscos indican diferencias estadísticas significativas ($p<0.05$) entre los períodos indicados

De la misma manera, en la figura 4.2 se observa que la centralidad de *eigenvector* de las palabras control no exhiben un cambio significativo en los diferentes períodos detallados (panel (a)), que las centralidades *eigenvector* de las palabras pandémicas exhiben también un cambio

significativo entre los años 2019-2020 y 2019-2021 (panel (b) de la figura). Se observa, al igual que en la figura 4.1, que las centralidades de *eigenvector* de las palabras pandémicas de los años 2020 y 2021 en promedio son mayores que las del año 2019.



(a) Centralidad *eigenvector* de palabras control, años 2019, 2020 y 2021

(b) Centralidad *eigenvector* de palabras pandémicas, años 2019, 2020 y 2021

Figura 4.2: Comparación de centralidad de *eigenvector* entre palabras pandémicas y control en los tres períodos. Las estrellas o np indican diferencias estadísticas entre los períodos indicados

4.3.2. Vecindad de los nodos

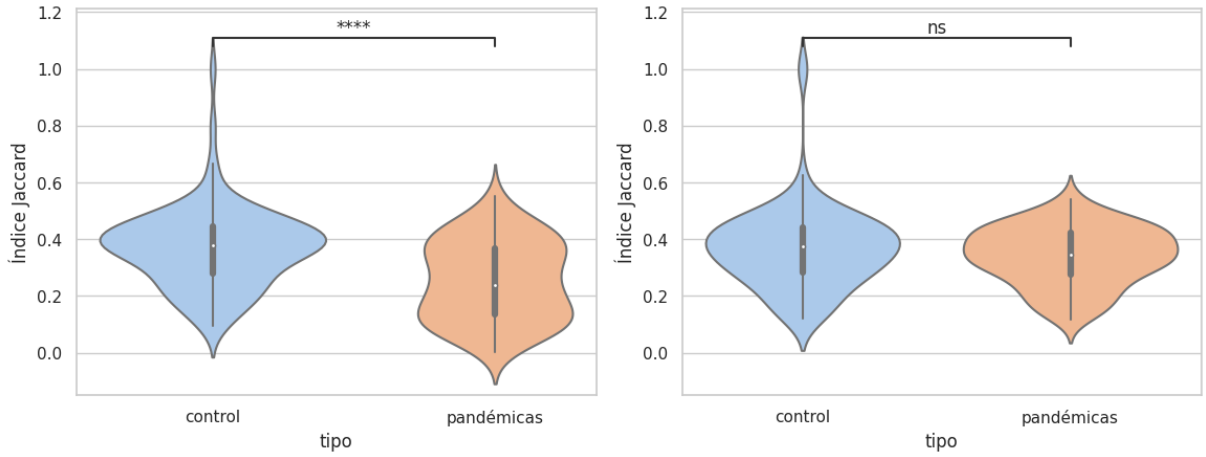
Análogamente a lo estudiado con la centralidad, en esta sección estudiamos cómo varío el entorno del conjunto de palabras control y pandémicas en los distintos períodos. Para tal fin, se extrajo el conjunto de primeros vecinos de cada palabra (nodos para los que existe una arista que los conecta a la palabra en cuestión) y se utilizó el índice de Jaccard [Real and Vargas, 1996] para calcular las similitudes de los conjuntos de vecinos entre los períodos analizados, específicamente comparando la transición entre períodos.

El índice de similitud de Jaccard compara los miembros de dos conjuntos para ver qué miembros se comparten y cuáles son distintos. Es una medida de similitud entre dos conjuntos de datos, con un rango entre 0-1. Cuanto mayor, más similares son las dos poblaciones. Nuevamente, se utilizó el test estadístico no paramétrico Mann Whitney Wilcoxon para contrastar si los coeficientes de Jaccard de palabras pandémicas y control en las diferentes transiciones de los períodos provienen de poblaciones equidistribuidas.

En la figura 4.3 se observa el grado de similitud entre los dos conjuntos de palabras. En el panel (a) se puede observar que la distribución de los índices de Jaccard de las palabras control para los años 2019 y 2020 son significativamente distintos de las pandémicas, es decir, se rechaza la hipótesis de igualdad de promedios y se concluye que los grupos definidos proceden de poblaciones con distinto promedio. También se puede observar que para el conjunto de palabras

control los índices se distribuyen en promedio con valores mayores que las pandémicas, indicando que los conjuntos de las palabras control son más semejantes en la transición de los años 2019-2020 respecto de las palabras pandémicas. Esto significa que los vecinos de palabras control son más similares, es decir, no cambió tanto el entorno de las palabras control en la transición 2019-2020, respecto de los cambios observados por las palabras pandémicas, que lógicamente vieron una alteración en el conjunto de palabras con las que se vinculan.

Por otro lado, en la comparación de los índices entre los años 2020 y 2021 se observa que las distribuciones del índice de Jaccard de las palabras control y pandémicas no son significativamente distintas, es decir, no podemos rechazar la hipótesis de igualdad de promedios y se concluye que los grupos definidos proceden de poblaciones con similar promedio. Esto refleja el hecho de que una vez establecido el nuevo significado de las palabras pandémicas en el 2020, este se mantiene en el 2021, al igual que que las palabras control que no vieron alterado significado en los tres períodos estudiados.

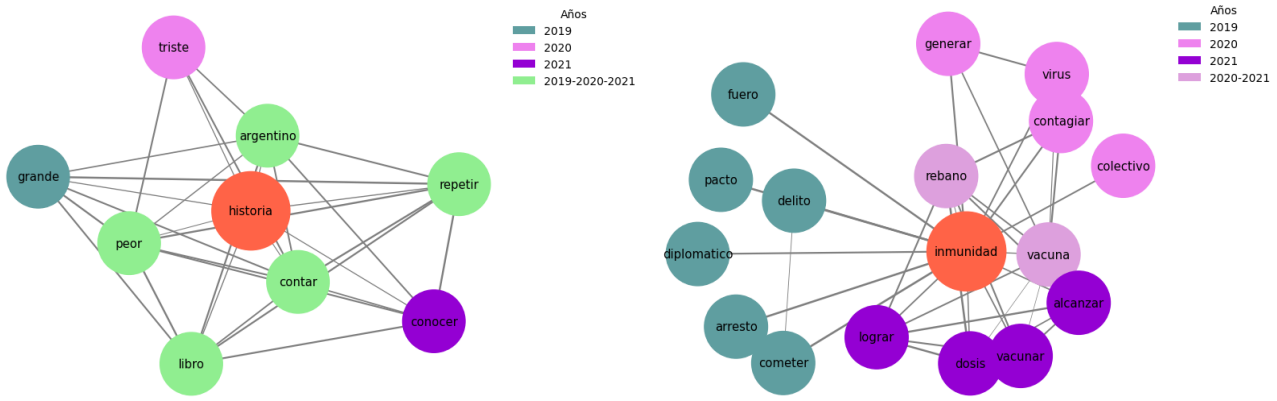


(a) Índice de Jaccard entre los años 2019-2020 (b) Índice de Jaccard entre los años 2020-2021

Figura 4.3: Comparación de los índices de Jaccard entre palabras pandémicas y control en la transición de los períodos detallados. Las estrellas o np indican diferencias estadísticas entre los tipos de palabras para las transiciones entre los períodos indicados

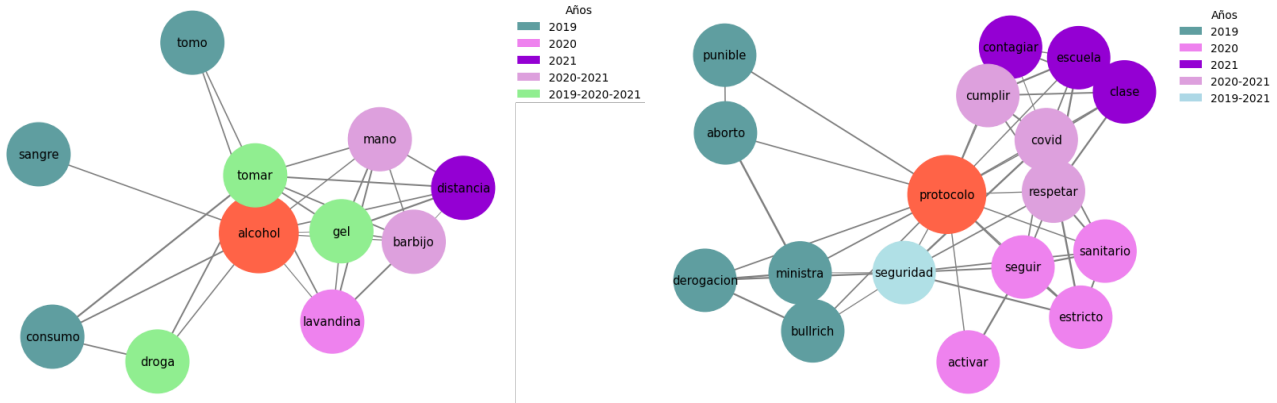
4.4. Visualización de los cambios de significado

Finalmente, a modo de conclusión de las métricas calculadas en las secciones anteriores, visualizamos una muestra de los subgrafos asociados a las palabras pandémicas y de control para los distintos períodos estudiados. En la figura 4.4 se muestra el subgrafo de las 6 palabras mas cercanas según su valor de NPMI a la palabra control “historia”, y a las palabras pandémicas “inmunidad”, “alcohol” y “protocolo”.



(a) Palabra control "historia", años 2019, 2020 y 2021

(b) Palabra pandémica "inmunidad", años 2019, 2020 y 2021



(c) Palabra pandémica "alcohol", años 2019, 2020 y 2021

(d) Palabra pandémica "protocolo", años 2019, 2020 y 2021

Figura 4.4: Subgrafos de las palabras mas cercanas según su valor de $NPMI$ al término control "historia" y a los términos pandémicos "inmunidad" , "alcohol" y "protocolo" en los tres períodos. Cada período se señala en la figura. Se puede observar el cambio de vecindad de la palabra asociado a un cambio en su uso semántico

En relación a la palabra control "historia"(panel (a) de la figura 4.4) se ve que la vecindad del nodo no presenta grandes cambios repitiéndose para los tres periodos 4 de las 6 palabras analizadas en cada período. En cambio, se puede observar cualitativamente que la vecindad de los nodos de estudio de las palabras pandémicas cambian en los tres períodos. En relación a la palabra pandémica "inmunidad"(panel (b) de la figura), en el período 2019 se puede observar que la palabra se acerca más a temas de delitos o temas diplomáticos, mientras que en 2020 se acerca semánticamente a temas relacionados con la salud, virus y contagio, y finalmente en 2021 el término comienza a relacionarse con vacuna. Del mismo modo en la palabra "alcohol"(panel (c) de la figura) se observa que en el año 2019 se relaciona a consumo y sangre (temas de desinfección), en el año 2020 se comienza a relacionar a temas de limpieza (como pue-

de verse con la aparición del término “lavandina”), y finalmente en el año 2021 se la relaciona como una barrera ante el posible contagio y se acerca al concepto distancia, es decir, a temas relacionados con la prevención de enfermedades. Finalmente, respecto de la palabra “protocolo”(panel (d) de la figura), en el período 2019 se puede observar que la palabra se acerca más a temas de ley de aborto y seguridad. En el año 2020, se observa que el término “protocolo”se acerca semánticamente a temas relacionados con la salud (sanitario); y finalmente en 2021 el término “seguridad”vuelve a aparecer, se observa que el término comienza a relacionarse con educación. Estos ejemplos muestran que las redes semánticas obtenidas permiten detectar cambios semánticos léxicos ocurridos en las palabras pandémicas en los periodos estudiados que justamente asociamos a la emergencia de la pandemia de COVID-19.

Capítulo 5

Conclusiones

Entender cómo las palabras cambian sus significados a lo largo del tiempo es relevante para los estudios sobre la evolución del lenguaje y la cultura. En muchos casos, estos cambios se producen a partir de la emergencia de un evento extremo. En particular, la pandemia de COVID-19 ha influido drásticamente en nuestro discurso diario en un corto período de tiempo. Trabajos previos estudiaron las modificaciones en la memoria de las palabras provocadas por la pandemia basándose en experimentos de asociación de palabras [Laurino et al., 2023], donde la memoria semántica se representa como una red o grafo de palabras. Como motivación de este trabajo se estudió si es posible reproducir este tipo de ensayos controlados con datos obtenidos de redes sociales como, por ejemplo, Twitter. Buscamos producir una metodología análoga que nos permita explorar cómo afectó la pandemia en el uso y el significado de un conjunto de términos, con datos obtenidos a partir de redes sociales y mediante el uso de redes semánticas.

Redes sociales como Twitter contienen una gran cantidad de información pública sobre las interacciones, los gustos, disgustos, e intereses de cientos de millones de individuos. Esto permite recolectar datos masivos de situaciones naturales no controladas. Es interesante destacar que Twitter posee características únicas como, por ejemplo, poca extensión de los textos e informalidad en los mismos. Todas estas características nos permite hacer una analogía entre los textos escritos en la red social con las respuestas espontáneas que se dan en experimentos de asociación de palabras, pero a su vez nos permiten extrapolar dichos resultados a una mayor cantidad de datos.

En el capítulo 2 de esta tesis se presentó la metodología del trabajo. Como se mencionó anteriormente, la metodología se basó en la construcción de redes semánticas para detectar sistemáticamente el cambio diacrónico de conceptos utilizando todos los textos de los usuarios seleccionados volcados en Twitter. El enfoque novedoso de este trabajo fue la selección de un conjunto de 200 usuarios de forma tal de realizar un seguimiento longitudinal a lo largo de 3 años consecutivos. Esto es importante pues permite asemejar a un experimento controlado (entrevistando a un mismos sujeto experimental en diferentes etapas), reduciendo la variabilidad

intrínseca que se observa en la generación de información en redes sociales. Para la construcción de las redes semánticas ponderadas se definió una medida de similaridad entre palabras basado en la coocurrencia de términos e información mutua ($NPMI$) y se agregó en forma aditiva la información proveniente de cada usuario. Finalmente, se aplicaron técnicas de filtrado de redes a fin de obtener redes binarias que constituyeron nuestro objeto de estudio, representando la relación entre palabras utilizadas por el conjunto total de usuarios para cada período.

El criterio de selección de los usuarios fue que tuvieran cierta actividad en los tres periodos (2019, 2020 y 2021) y que hayan utilizado alguna vez en alguno de los períodos la palabra testigo "vacuna"', término relacionado con la pandemia de COVID-19, ya que queremos medir el efecto de la pandemia. Algunos cuestiones relacionadas a este punto, y que se puede mejorar en futuros trabajos, es que por la forma de seleccionar a los participantes se podría estar incurriendo en un sesgo de selección, siendo no representativa la muestra tomada de la población objetivo.

Por otro lado, también hay que considerar el sesgo cultural, ya que el trabajo se basó en usuarios de Argentina que se expresan habitualmente en Twitter y que poseen cierto nivel de educación formal. Si bien Twitter exhibe características distintivas en comparación con otras redes sociales, por ser un espacio de interacción [Gruzd et al., 2018, Jaidka et al., 2018] donde usuarios conversan de manera natural unos con otros, hay cierto consenso acerca de que en Twitter se expresan usuarios más polarizados y con ciertos recursos culturales [Stier et al., 2020]. Cabe mencionar que en esta red también se puede ver que predomina en cierta medida la afinidad política, siendo el rasgo principal la interpelación en general, donde se puede verificar que hay una mayor proporción de mensajes con toxicidad [Albanese et al., 2023]. A pesar de todas estas características propias de Twitter, podemos considerar que el tipo de participantes del experimento de asociación de palabras que utilizamos como referencia [Laurino et al., 2023], ya sea por recursos culturales o accesibilidad y participación en este tipo de redes sociales, podrían poseer características similares (ciertas similitudes) a los usuarios seleccionados. También es importante considerar que se conoce que en Twitter hay actores no auténticos, como robots sociales maliciosos y trolls, por lo que no todos los mensajes son una expresión genuina de un usuario legítimo. Esto hace que cuentas hiperactivas podrían estar sobrerrepresentadas en las muestras e introducir sesgos y señales no auténticas. Sin embargo, la prevalencia de actividades no auténticas en los flujos de datos sociales aún no está del todo clara, lo que dificulta evaluar los sesgos de los análisis basados en dichos datos [Yang et al., 2022]. En nuestro caso al ser un número relativamente bajo de usuarios estudiados este sesgo se pudo controlar. Sin embargo, en caso de extrapolar nuestro trabajo a una mayor cantidad de usuarios sería conveniente aplicar algún método que identifique usuarios genuinos.

En el capítulo 3, analizamos las redes semánticas mediante el formalismo de redes complejas junto con su abanico de técnicas de análisis [Barabási, 2013, Newman, 2003]. El análisis de la redes obtenidas para cada periodo demostró que las propiedades topológicas, medidas a partir de métricas generales tales como el grado medio, diámetro de la red, coeficiente de clustering y asortatividad, no presentaron grandes variaciones en función del tiempo. Esto significa que la estructura global de la red se mantuvo relativamente estable. Esto fue relevante porque

posibilita que las redes de los distintos periodos sean comparables entre si.

Los resultados más interesantes vinieron de estudiar las comunas emergentes. Esto se realizó mediante el algoritmo de Louvain y se identificaron los nodos más importantes según la centralidad *eigenvector*. Los nodos más centrales permitieron etiquetar a las comunidades según el tema al cual se refieren. Con este análisis, vimos la aparición y desaparición de comunidades en las redes semánticas en función del tiempo. El ejemplo más notorio de esto se observó en la emergencia de la comunidad relacionada a la pandemia del *COVID-19* en los años 2020 y 2021. Por otro lado, se observó la desaparición de la comunidad etiquetada como *fútbol* durante el año 2020, posiblemente por la falta de actividades deportivas durante dicho año y la aparición de la comunidad relacionada con *política* en los años 2019 y 2021, relacionado al hecho de que dichos años fueron años electorales. Además se pudo evidenciar que los temas detectados compiten entre si por el interés de los usuarios como, por ejemplo, se observó en el aumento del porcentaje en tema *alimentación y nutrición*, que aparece con más fuerza en el año de mayor confinamiento 2020 debido a los cambio de consumo, dificultades financieras y reducción de movilidad, o en la comunidad *fútbol* que presenta diferencia de porcentajes en su tamaño entre 2019 y 2021, donde se podría decir que este tema en el año 2019 competía solo con la comunidad *política*, pero en el año 2021 competía además con el tema *Pandemia COVID-19*.

En el capítulo 4, se evaluaron los cambios en el tiempo de palabras ligadas a la pandemia (esperábamos que se vean afectadas por la misma) y control (esperábamos que su significado se mantenga estable en el tiempo), análogo al estudio realizado en [Laurino et al., 2023]. Aquí se corroboró que los cambios semánticos se detectan además en el cambio de vecindad de las palabras. En primera instancia se evaluó en qué comunidades se encuentran algunas de las palabras control y pandémicas en función del tiempo. Por ejemplo, las palabras "aislado" e "inmunidad" en el año 2019 se encontraban en las comunidades *palabras comunes* y *política*, respectivamente, y luego ambas pasaron a la comunidad *Pandemia COVID-19* en los años 2020 y 2021, evidenciando un cambio en el significado del uso habitual de dichas palabras. Por otro lado, se observó que palabras que tomamos de control, tales como "claro", "historia" y "cosa", entre otras, se mantuvieron en sus respectivas comunidades durante los tres períodos.

En segunda instancia, vimos que los cambios semánticos se pueden detectar estudiando cambios en la vecindad de las palabras. Tomando un conjunto extenso de términos pandémicos y de control, observamos un aumento de centralidad *eigenvector* de las palabras pandémicas respecto de las control en los años 2020 y 2021 respecto del año 2019. Esto demuestra un aumento en la cantidad de vecinos de las palabras pandémicas en la red semántica, aumento que se inició en el 2020 y se mantuvo en el 2021, evidenciando un enriquecimiento del contexto en el cual se utilizan dichas palabras. Continuando con el análisis, se utilizó el índice de similitud de Jaccard, que compara los miembros de dos conjuntos, para ver qué miembros se comparten y cuáles son distintos. Específicamente se utilizó para comparar la transición entre períodos de cada palabra analizada y su conjunto de vecinos. Se demostró que los índices de las palabras pandémicas se distribuyen en promedio con valores menores que las controles indicando que estas palabras pandémicas variaron más su vecindario respecto de las control, esto relacionado

a nuevos significados de las palabras. Por otro lado, en la comparación de los índices entre los años 2020 y 2021 se observa que las distribuciones del índice de Jaccard de las palabras control y pandémicas no son significativamente distintas, esto evidencia que una vez establecido el nuevo significado de las palabras pandémicas en el 2020, este se mantiene en el 2021. Para futuros trabajos, sería necesario analizar si estos nuevos significados de las palabras pandémicas se mantienen a lo largo de otros períodos, por ejemplo, durante los años 2022 y 2023.

Finalmente, en el capítulo 4 visualizamos los cambios de significado de las palabras investigadas, apoyando los resultados de las métricas calculadas. Se mostraron los subgrafos de palabras más cercanas según su valor de $NPMI$ de ciertas palabras control y pandémicas. Se pudo observar como era de esperar que cualitativamente la vecindad de los nodos de palabras pandémicas cambian en los tres períodos. Por ejemplo, la palabra pandémica "inmunidad"se observó que en el período 2019 se acercaba más a temas de delitos o temas diplomáticos, mientras que en 2020 se acercaba semánticamente a temas relacionados con la salud, virus y contagio, y en 2021 comienzan a relacionarse con vacuna. La palabra "protocolo"se observa que en el período 2019 se acerca más a temas de ley de aborto y seguridad, en 2020 se acerca semánticamente a temas relacionados con la salud y finalmente en 2021 el término comienza a relacionarse con educación. En cambio, la palabra control "historia"no presentaba grandes cambios en los tres períodos analizados. Estos ejemplos muestran que las redes semánticas obtenidas permiten detectar cambios semánticos léxicos ocurridos en las palabras pandémicas en los periodos estudiados asociados a la emergencia de la pandemia de COVID-19.

Por todo lo detallado anteriormente, proponemos que el uso de redes complejas, construidas a partir del lenguaje utilizado en la red social Twitter, puede verse como una nueva metodología que podría ser útil para rastrear cambios semánticos rápidos producidos en diferentes períodos de tiempo en un escenario colectivo [Carrillo et al., 2015]. Se pudo detectar cambios semánticos tanto en la aparición de nuevas comunidades en la red semántica como en el cambio de la vecindad de las palabras específicas. Esta nueva metodología nos permitió generar un nuevo conjunto de datos correspondientes a los periodos 2019, 2020 y 2021 el cual fue útil para la detección del cambio semántico y además nos permitió realizar una aproximación al análisis lingüístico del impacto de la pandemia de COVID-19 en el idioma español, más específicamente en su versión rioplatense.

Discusiones y trabajos a futuro

A continuación detallamos algunas puntos no cubierto en esta tesis y que podrían constituir futuras direcciones de investigación:

- Un enfoque alternativo a la construcción de redes semánticas es el uso de *embeddings* de palabras. Diferentes estudios sobre cuantificar cambios en el significado de palabras están basados en métodos tales como SVD y word2vec [Garg et al., 2018, Hamilton et al., 2016a, Stella et al., 2020]. Esta tesis se enfocó en el uso de redes debido a tomar como punto de partida los experimentos de asociación de palabras, donde la representación natural de

los datos eran las redes semánticas, con la ventaja de que las representaciones en redes permite utilizar las métricas y los algoritmos desarrollados dentro de la teoría de redes. Sin embargo, métodos como word2vec también están basados en el análisis del entorno de las palabras, lo cual podría ser semejante a nuestro enfoque. Como trabajo a futuro, sería interesante analizar en forma exhaustiva hasta qué punto los métodos de redes semánticas y *embeddings* son similares y las ventajas y desventajas de cada uno de ellos.

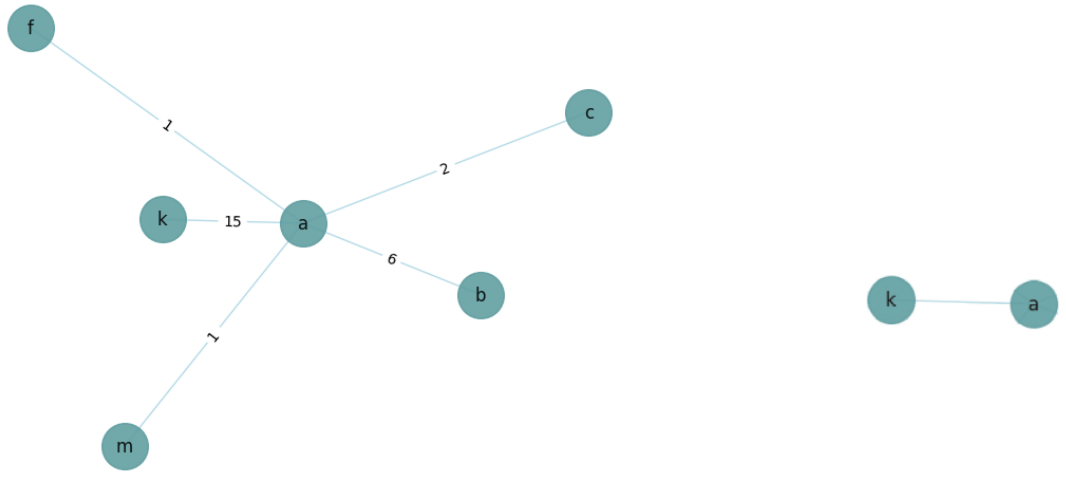
- En la tesis no realizamos una comparación cuantitativa entre las redes construidas con información de redes sociales y las redes obtenidas en los experimentos de asociación realizados por el grupo de investigación [Laurino et al., 2023]. Se propone como trabajo a futuro, una discusión más extensa de hasta qué punto ambos enfoques son comparables (redes a partir de datos de redes sociales y experimentos controlados) y la comparación cuantitativa de los resultados obtenidos en cada caso.

- La adquisición de datos para la realización de esta tesis se hizo a partir de la API de Twitter, cuya política de accesibilidad a los datos cambió durante el desarrollo de la tesis. Esto pone en jaque la reproducibilidad de los estudios realizados aquí y la replicación de los resultados en otro conjunto de datos. Se podrían utilizar otras redes sociales para adquirir los datos tales como Reddit, Instagram y Facebook. Sin embargo, la disponibilidad de nuevos datos para la replicación de experimentos es de suma importancia para la ciencia en general y la inconveniencia de depender de cambios en las políticas de cada plataforma es un problema que trasciende esta tesis.

Apéndice A

Apéndice

A.1. Algoritmo Backbone (Columna vertebral): ejemplo de cálculo



(a) Grafo antes de aplicar el algoritmo, los pesos se indican en las aristas

(b) Grafo luego de aplicar el algoritmo

Figura A.1: Ejemplo de un grafo, antes y después de aplicar algoritmo backbone [Serrano et al., 2009]. Notar que solo se preservan los nodos significativos.

En la Figura A.1 se muestra un ejemplo para demostrar el cálculo del algoritmo Backbone. Se indica un nivel de significancia α de 0,050.

Se analiza para cada nodo $(_i)$, sus nodos vecinos $(_j)$:

Si $1 - p_{i,j}^{(k_n-1)} < \alpha \rightarrow$ se conservan el enlace del nodo$(_j)$

Siendo:
$suma - pesos$=suma de todos los pesos de los vecinos $(_j)$ del nodo $(_i)$

$p_{i,j} = \frac{peso-nodo_{i,j}}{suma-pesos}$

$k_n =$ cantidad de nodos vecinos de i

Ejemplo de cálculo de los nodos vecinos de a: c y k
$k_n = 5$
$suma - pesos = 15 + 6 + 1 + 2 + 1 = 25$

Cálculo para el nodo c:
$p_{i,j} = \frac{2}{25}$

$(1 - p_{i,j})^{(k_n-1)} = (1 - 0{,}08)^{(5-1)}$

0,716 no es $< 0{,}050 \rightarrow$ no se conserva el enlace del nodo c

Cálculo para el nodo k:
$p_{i,j} = \frac{15}{25}$

$(1 - p_{i,j})^{(k_n-1)} = (1 - 0{,}6)^{(5-1)}$

0,026 es $< 0{,}050 \rightarrow$ se conserva el enlace del nodo k

A.2. Comunidades

A.2.1. Selección del algoritmo de detección de comunidades

En la tabla se puede observar el resultado de aplicar distintos algortimos de detección de comunidades. En el trabajo se adoptó el algortimo de Louvain dado que presenta el valor más

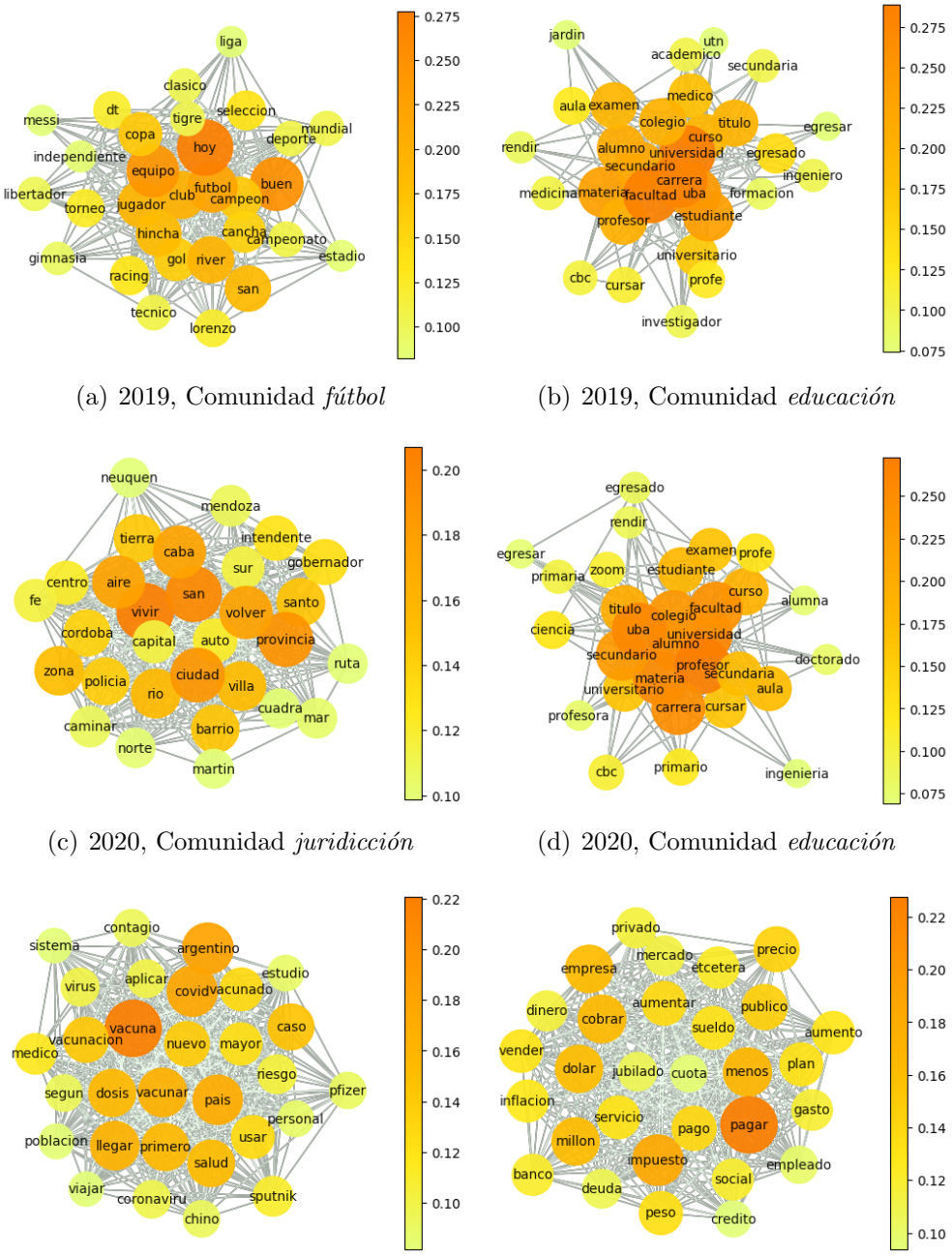
alto de modularidad A.1

Detección de comunidades			
Algoritmo	tiempo	Modularidad	Clusters
Infomap	77.096	0.052475	1593
Louvain (multilevel)	2.274	0.188673	118
Leading_eigenvector	9.0233	0.125526	2
Leiden	0.2408	-0.001021	2072
Walktrap	163.5277	0.040569	11872

Cuadro A.1: Resultado de algoritmos aplicados a la red del año 2019

A.2.2. Otras comunidades

En la figura A.2 se pueden observar otros ejemplos de comunidades detectadas.



(a) 2019, Comunidad *fútbol*

(b) 2019, Comunidad *educación*

(c) 2020, Comunidad *juridicción*

(d) 2020, Comunidad *educación*

(e) 2021, Comunidad *Pandemia, COVID-19*

(f) 2021, Comunidad *dinero, pagos*

Figura A.2: Comunidades obtenidas en los tres períodos, el tamaño y color indican el valor de centralidad *eigenvector* de los nodos

Bibliografía

[Agrawal et al., 2014] Agrawal, D., Budak, C., El Abbadi, A., Georgiou, T., and Yan, X. (2014). Big data in online social networks: user interaction analysis to model user behavior in social networks. In *Databases in Networked Information Systems: 9th International Workshop, DNIS 2014, Aizu-Wakamatsu, Japan, March 24-26, 2014. Proceedings 9*, pages 1–16. Springer.

[Albanese et al., 2023] Albanese, F., Feuerstein, E., Kessler, G., and Ortiz de Zárate, J. M. (2023). Aprendizaje automático para el análisis crossplataforma de la comunicación política: Gobierno y oposición argentinos en facebook, instagram y twitter. *Cuadernos. info*, n. v.(55):256–280.

[Barabási, 2013] Barabási, A.-L. (2013). Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987):20120375.

[Bird et al., 2009] Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. .º'Reilly Media, Inc.".

[Blondel et al., 2008] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

[Bouma, 2009] Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40.

[Cabana et al., 2023] Cabana, Á., Zugarramurdi, C., Valle-Lisboa, J. C., and De Deyne, S. (2023). The"small world of words"free association norms for rioplatense spanish. *Behavior Research Methods*, pages 1–18.

[Carrillo et al., 2015] Carrillo, F., Cecchi, G. A., Sigman, M., and Slezak, D. F. (2015). Fast distributed dynamics of semantic networks via social media. *Computational intelligence and neuroscience*, 2015:50–50.

[Collins and Quillian, 1969] Collins, A. M. and Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, 8(2):240–247.

[De Deyne et al., 2019] De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., and Storms, G. (2019). The "small world of words" english word association norms for over 12,000 cue words. *Behavior research methods*, 51:987–1006.

[De Deyne and Storms, 2008] De Deyne, S. and Storms, G. (2008). Word associations: Network and semantic properties. *Behavior research methods*, 40(1):213–231.

[Española, 2020] Española, R. A. (2020). *Crónica de la lengua española 2020*. Espasa.

[Española, 2022] Española, R. A. (2022). Diccionario. *Madrid: Espasa*.

[Firth, 1957] Firth, J. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, pages 10–32.

[García et al., 2007] García, F., Portillo, J., Romo, J., and Benito, M. (2007). Nativos digitales y modelos de aprendizaje. In *SPDECE*.

[Garg et al., 2018] Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

[Gonen et al., 2021] Gonen, H., Jawahar, G., Seddah, D., and Goldberg, Y. (2021). Simple, interpretable and stable method for detecting words with usage change across corpora. *arXiv preprint arXiv:2112.14330*.

[Gruzd et al., 2018] Gruzd, A., Lannigan, J., and Quigley, K. (2018). Examining government cross-platform engagement in social media: Instagram vs twitter and the big lift project. *Government Information Quarterly*, 35(4):579–587.

[Guo et al., 2021] Guo, Y., Xypolopoulos, C., and Vazirgiannis, M. (2021). How covid-19 is changing our language: Detecting semantic shift in twitter word embeddings. *arXiv preprint arXiv:2102.07836*.

[Hamilton et al., 2016a] Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016a). Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the conference on empirical methods in natural language processing. Conference on empirical methods in natural language processing*, volume 2016, page 2116. NIH Public Access.

[Hamilton et al., 2016b] Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016b). Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

[Harris, 1954] Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

[Jaidka et al., 2018] Jaidka, K., Guntuku, S., and Ungar, L. (2018). Facebook versus twitter: Differences in self-disclosure and trait prediction. In *Proceedings of the International AAAI*

Conference on Web and Social Media, volume 12.

[Joos, 1950] Joos, M. (1950). Description of language design. *The Journal of the Acoustical Society of America*, 22(6):701–707.

[Jurafsky and Martin, 2008] Jurafsky, D. and Martin, J. H. (2008). Speech and language processing: An introduction to speech recognition, computational linguistics and natural language processing. *Upper Saddle River, NJ: Prentice Hall*.

[Laurino et al., 2023] Laurino, J., De Deyne, S., Cabana, A., and Kaczer, L. (2023). The pandemic in words: tracking fast semantic changes via a large-scale word association task. *Open mind*, pages 1–19.

[Leskovec et al., 2007] Leskovec, J., Kleinberg, J., and Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2–es.

[Newman, 2018] Newman, M. (2018). *Networks*. Oxford university press.

[Newman, 2003] Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.

[Newman and Girvan, 2004] Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113.

[NTLLE, nd] NTLLE, R. A. E. (n.d.). Nuevo tesoro lexicográfico de la lengua española [en línea].

[Piantadosi, 2014] Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21:1112–1130.

[Prensky, 2001] Prensky, M. (2001). Digital natives, digital immigrants part 2: Do they really think differently? *On the horizon*, 9(6):1–6.

[Qi et al., 2020] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

[Rababah, 2015] Rababah, A. G. (2015). Corpus linguistic analysis of the connotative meaning of some terms used in the context of'the war on terror'. *International Journal of English Linguistics*, 5(1):113.

[Real and Vargas, 1996] Real, R. and Vargas, J. M. (1996). The probabilistic basis of jaccard's index of similarity. *Systematic biology*, 45(3):380–385.

[Schmuckler, 2001] Schmuckler, M. A. (2001). What is ecological validity? a dimensional analysis. *Infancy*, 2(4):419–436.

[Serrano et al., 2009] Serrano, M. Á., Boguná, M., and Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. *Proceedings of the national academy of sciences*, 106(16):6483–6488.

[Shtyrov et al., 2010] Shtyrov, Y., Nikulin, V. V., and Pulvermüller, F. (2010). Rapid cortical plasticity underlying novel word learning. *Journal of Neuroscience*, 30(50):16864–16867.

[Signoroni et al., 2022] Signoroni, E., Jezek, E., and Sprugnoli, R. (2022). Word usage change and the pandemic: A computational analysis of short-term usage change in the italian reddit community. *IJCoL. Italian Journal of Computational Linguistics*, 8(8-2).

[Simpson and Proffitt, 1997] Simpson, John Andrew, E. S. W. and Proffitt, M. (1997). *Oxford English dictionary additions series*, volume 3. Clarendon Press.

[Sowa et al., 1992] Sowa, J. F. et al. (1992). Semantic networks. *Encyclopedia of artificial intelligence*, 2:1493–1511.

[Stella et al., 2020] Stella, M., Restocchi, V., and De Deyne, S. (2020). # lockdown: Network-enhanced emotional profiling in the time of covid-19. *Big Data and Cognitive Computing*, 4(2):14.

[Steyvers et al., 2006] Steyvers, M., Griffiths, T. L., and Dennis, S. (2006). Probabilistic inference in human semantic memory. *Trends in cognitive sciences*, 10(7):327–334.

[Stieglitz et al., 2014] Stieglitz, S., Dang-Xuan, L., Bruns, A., and Neuberger, C. (2014). Social media analytics: An interdisciplinary approach and its implications for information systems. *Business & Information Systems Engineering*, 6:89–96.

[Stier et al., 2020] Stier, S., Bleier, A., Lietz, H., and Strohmaier, M. (2020). Election campaigning on social media: Politicians, audiences, and the mediation of political communication on facebook and twitter. In *Studying Politics Across Media*, pages 50–74. Routledge.

[Wilkins, 1996] Wilkins, D. P. (1996). Semantic change and the search for cognates. *The comparative method reviewed: Regularity and irregularity in language change*, pages 264–304.

[Yang et al., 2022] Yang, K.-C., Hui, P.-M., and Menczer, F. (2022). How twitter data sampling biases us voter behavior characterizations. *PeerJ Computer Science*, 8:e1025.