



Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Física

Análisis comparativo de métodos para detección de tópicos en la agenda social

Tesis de Licenciatura en Ciencias Físicas

Franco Eskinazi

Dirección: Pablo Balenzuela
Co-Dirección: Marcos Alberto Trevisán

Marzo 2024

TEMA: Análisis comparativo de métodos para detección de tópicos en textos

ALUMNO: Franco Eskinazi

LU: 205/19

LUGAR DE TRABAJO: SoPhy lab, Departamento de Física, Facultad de Ciencias Exactas y Naturales, UBA.

DIRECCIÓN: Pablo Balenzuela (INFINA/CONICET) y Marcos Alberto Trevisán (INFINA/CONICET)

FECHA DE INICIACIÓN: Marzo 2023

FECHA DE FINALIZACIÓN: Abril 2024

FECHA DE EXAMEN: 20/12/2021

INFORME APROBADO POR:

Autor	Jurado
Director	Jurado
Profesor de la Tesis de Licenciatura	Jurado

Resumen

Ante la abundancia de información que nos llega a través de los medios de comunicación y de las redes sociales, es muy importante detectar y poder seguir en el tiempo distintos tópicos o ejes temáticos en los cuales se organiza la discusión pública. Hacerlo de manera automatizada a partir de analizar grandes cantidades de datos es una línea de investigación que se viene desarrollando en los últimos años. En este trabajo de tesis de licenciatura, se comparan dos métodos de detección de ejes temáticos de la agenda pública analizando la red social Twitter.

El primer método se basa en la hipótesis que un conjunto de tweets hablan del mismo tema porque están etiquetados con un conjunto de hashtags que co-ocurren entre ellas. Para poder detectar estos tópicos, representamos esta co-ocurrencia mediante una red compleja y realizamos sobre la misma una detección de comunidades que nos permitió encontrar grupos de hashtags fuertemente entrelazados. Estos grupos de hashtags se asocian a cada eje temático de la agenda.

El segundo método se basa en la hipótesis que un conjunto de tweets hablan del mismo tema porque ellos son muy parecidos entre sí. Para desarrollarlo representamos los textos de los tweets usando métodos Procesamiento de Lenguaje Natural (NLP) que usan herramientas de machine learning para representar dichos textos como vectores en un espacio n-dimensional. La principal fortaleza de estos métodos es que textos similares semánticamente se encuentran cercanos en el espacio embebido en el cual se representan. Utilizando estas herramientas, representamos el conjunto de tweets mediante una red compleja pesada donde los enlaces entre ellos representan la medida de similaridad entre ambos. En esta red, también identificamos los tópicos emergentes mediante un proceso de detección de comunas.

Una vez determinados el conjunto de tópicos (o agenda) en el cual se representa el conjunto de tweets analizado, analizamos el volumen de interés en cada eje temático obteniendo la serie temporal de cantidad de tweets asociado a cada uno.

Los resultados obtenidos nos muestran que ambas metodologías muestran resultados consistentes, aunque no necesariamente similares, pero que permiten una descripción temporal adecuada de los ejes de discusión que dominan la conversación pública en una red social.

Índice general

1. Introducción	4
2. Métodos	7
2.1. Procesamiento de lenguaje natural (NLP)	7
2.1.1. Niveles de procesamiento	7
2.1.2. VSM	9
2.1.3. Machine Learning	11
2.2. Redes complejas	14
2.2.1. Red de coocurrencias	16
2.2.2. Red de similitud	17
2.3. Detección de comunidades en una red compleja	18
2.3.1. Louvain	18
2.3.2. Infomap	20
3. Datos	23
3.1. Recolección de datos	23
3.2. Timestamp	24
3.2.1. Detección de tweets irrelevantes	24
3.3. Hashtags	24
3.4. Rank por retweets	25
3.5. Texto	26
4. Caracterización de tópicos mediante una red de co-ocurrencia de hashtags	28
4.1. Construcción de la red	28
4.2. Detección e identificación de tópicos	32
4.3. Series temporales	41
4.3.1. Series temporales de JxC y FdT	44
5. Caracterización de tópicos mediante una red de similaridad semántica de textos	46
5.1. Representación vectorial de los textos	46
5.2. Construcción de la red de similitud	47
5.2.1. Métrica de similitud	47
5.2.2. Construcción de la red y selección de enlaces significativos	48
5.3. Detección de tópicos en la red de similitud	49

5.4.	Identificación de los tópicos subyacentes	51
5.4.1.	Bag of Words	51
5.4.2.	Co-ocurrencia de palabras	53
5.4.3.	Análisis de sentimiento	55
5.5.	Series temporales	59
6.	Discusiones y conclusiones	63

Agradecimientos

Primero me gustaría agradecer a la Universidad de Buenos Aires por la oportunidad de tener una educación de tantísimo valor en mi vida. Estos últimos 5 años pude estudiar una ciencia que amo gracias a esta universidad. Estoy muy feliz de haber hecho mi licenciatura de ciencias físicas acá. Conocí a muchos de mis amigos y, más allá de las materias, gané muchísimas herramientas para crecer como persona.

Quiero agradecer a mis papás por todo el amor que me demuestran constantemente. Desde siempre me respetaron para hacer lo que amaba y priorizar mis estudios ante todo, sabiendo que era lo que más me motivaba. Fue por ellos que pude empezar y terminar esta parte de mi vida que tanto me marca. Me mostraron y siguen dando un enorme apoyo día a día.

A mis hermanos, Roma e Ivo, que me motivan con su dedicación y perseverancia para cumplir sus objetivos de vida con tantas ganas. Los amo y ya no aguento de volver a verlos.

A Abi, por la mejor compañía en estos años de facultad, por las mejores enseñanzas y reflexiones que tuve y sigo teniendo con esta increíble persona.

A Juli, por todo lo que me hace crecer y el amor que me da. También por hacerme llorar de la risa todo el tiempo, algo que necesité en momentos de la carrera.

A mis amigos más íntimos, Andy, Marto, Toto, Ale, Bruno, Bati, Ivo, Tobi, Cunha, Joaco e Ioni. Desde 2011 que me enseñan a ser una persona curiosa y reflexiva. Estoy seguro que la pasión que tengo por mi carrera surgió de ellos. Son una parte fundamental de mi vida.

A mis amigos de la facultad, habiendo vivido una carrera en plena pandemia llena de zooms de estudio y juegos. Les agradezco por siempre hacerme sentir parte de un grupo llenísimo de cariño.

Le agradezco a mi director, Pablo, por la paciencia y honestidad a lo largo de este último año. Me dejó aprender las cosas con el juego, la prueba y error, y el tiempo.

A Marcos, que siempre busca la forma de enseñar haciendo las preguntas interesantes y las observaciones afiladas. Asimismo me transmitió en todo el año una serenidad necesaria para poder atravesar esta etapa.

Por último, a todo el grupo de SoPhy. Ellos estuvieron dispuestos a ayudarme en muchísimas etapas de la tesis siempre con mucho cariño y buena onda. Les agradezco un montón por la inclusión al hermoso grupo.

Introducción

La comprensión de los procesos de formación de opinión pública alrededor de temas dominantes de la agenda informativa, así como también su dependencia temporal es un campo que ha sido investigado durante décadas [1, 2, 3].

Hace más de 40 años, McCombs y Shaw introdujeron la teoría del *agenda setting* [4] que sostiene que los temas que interesan a la sociedad están influenciados significativamente por los medios de comunicación y por la manera en que estos presentan y transmiten las noticias. En el estudio que presentaron se investigó el consumo de los medios de comunicación; televisión, diarios, revistas, etc. de 100 personas y se observó cómo impactaban las opiniones de estos medios en los sujetos. Las opiniones de estos se midieron con el uso de encuestas, preguntando sobre distintos problemas sociopolíticos contemporáneos y con qué opinión de las figuras políticas se alineaban ellos. Este trabajo mostró una fuerte correlación en las respuestas de las personas con las de los medios de comunicación que consumían.

El trabajo de McCombs & Shaw dio lugar a estudios sobre la agenda social y la opinión pública para entender qué agentes la influencian y cómo varían en el tiempo [1, 2].

En los últimos años el estudio de la agenda social se desarrolló analizando las opiniones e intereses de las personas en internet y las redes sociales. Por ejemplo, en el trabajo de S. Pinto [2] se realiza un estudio de la agenda social para ver cómo esta se relaciona con la agenda mediática. La agenda social se la define en este trabajo observando las búsquedas más frecuentes de Google con Google Trends, y las tendencias de Twitter. Esta agenda es comparada con la mediática, que se define partir de los tópicos que cubren los artículos de noticias de los medios masivos de comunicación como Clarín, Infobae, Página 12, etc. De cada agenda se estudió la diversidad de tópicos abarcados, la evolución de esta y también el interés relativo que le daba cada agenda a un tópico en particular.

Las técnicas para realizar estudios sobre la opinión pública se modificaron drásticamente desde el trabajo de McCombs & Shaw, en particular con la aparición de las redes sociales [5, 6, 7]. El acceso a estas redes y a los registros de las actividades de las personas en ellas le provee a la comunidad científica un volumen de datos inmenso para realizar estudios estadísticamente relevantes. Más aún, existen estudios que concluyen que el uso de las redes sociales para entender la agenda social es más efectivo que el uso de encuestas tradicionales [6]. Esto se debe a que en las encuestas las personas responden lo que suponen que la sociedad asume que es correcto responder, en oposición a lo que uno mismo realmente cree. Esto se conoce como el sesgo de deseabilidad social.

Uno de los ejemplos de redes sociales estudiadas para entender la agenda social y su com-

portamiento es Twitter [5, 6, 7]. En ella, muchos usuarios publican textos de pocos caracteres llamados *tweets*. Otros usuarios pueden darles 'me gusta' a estos tweets, comentarlos o compartirlos con otros usuarios en forma de *retweet*. Se puede considerar que el hecho de retuitear un tweet es un gesto de estar de acuerdo con la idea compartida [7]. Los usuarios que escriben los tweets pueden usar hashtags para compactar la idea del mismo. Estas son palabras en los tweets que son identificados por empezar con un carácter '#' y buscan resumir el mensaje del tweet asociado en una o pocas palabras clave. Estas unidades de información se destacan en los tweets y suelen aparecer frecuentemente generando tendencias en la red social. Son una gran utilidad para entender los tópicos de interés de las personas y la dinámica de ellos son estudiados para caracterizar la agenda social [5, 6, 8].

El uso de las redes sociales para el estudio de la agenda social trae un nuevo desafío, la cuantificación de las agendas y los tópicos de interés de las personas dentro de una red social. Este desafío ha sido y es atacado de múltiples maneras.

En Twitter, por ejemplo, una de las formas de estudiar la agenda pública es analizando el volumen de los distintos hashtags y como co-ocurren entre ellos [5, 8]. El comportamiento de estos se puede modelar a través de una red compleja que codifica la co-ocurrencia de los distintos hashtags en los tweets. H. Schawe y M. Beiró [5] estudiaron la teoría del *agenda setting* enfocándose en medios de comunicación tradicional como Fox News, The New York Times, etc. y en los usuarios en Twitter que seguían a las cuentas de estos medios en la red social. Este estudio, en parte, detecta los tópicos mencionados por los usuarios a partir de una red de co-ocurrencias de hashtags. Analizando como co-ocurren, se puede entender que grupos de hashtags se relacionan fuertemente a partir de algún algoritmo de detección de comunidades. Estos grupos son los que se suponen que representan a los distintos tópicos de interés de las personas en Twitter.

El trabajo de A. Bovet [8] también utiliza una red de co-ocurrencias de hashtags para entender cuales de estos son representativos al contexto de las elecciones presidenciales de Estados Unidos en 2016, cuando los candidatos eran Hillary Clinton y Donald Trump. Más aún, pudieron reforzar la idea de que en el período de elecciones en los EEUU había una alta polarización entre los dos partidos principales, ya que por un lado se encontró un grupo de hashtags co-ocurrentes mencionando al partido demócrata y otro grupo con hashtags relacionados al partido republicano.

Existe otra forma de caracterizar agendas, tanto sociales como mediáticas, mediante el uso de grandes volúmenes de textos, sean tweets hechos por usuarios como artículos de medios de comunicación [2, 6]. Para analizar los distintos tópicos subyacentes de un conjunto grande de textos se usan varias herramientas del área de Procesamiento de Lenguaje Natural, o NLP por sus siglas en inglés. Estas herramientas son utilizadas extraer información relevante de grandes volúmenes de textos y también permiten cuantificar similitudes entre textos [9, 10]. Esto permite estudiar un corpus masivo de oraciones y reconocer cuales son semánticamente parecidas. Encontrar oraciones de ideas similares sirven para detectar opiniones públicas y tópicos de discusión.

En el trabajo de S. Pinto [2], por ejemplo, se utiliza TF-IDF para representar vectorialmente artículos de noticias. Esta representación vectorial se construye a partir de las ocurrencias de palabras que aparecen en cada artículo y cuan representativas son para cada uno.

Más allá de los métodos frequentistas de palabras para la extracción de información de textos [2, 11], existen otros más modernos como modelos de machine learning especializados en la representación vectorial de textos. Estos modelos convierten palabras o documentos en vectores numéricos que capturan el contenido semántico y carga emocional de un texto, entre otros aspectos. Estos modelos se entrena usando grandes cantidades de texto, ajustando sus parámetros para que los vectores reflejen similitudes semánticas. A partir de la cuantificación de estos atributos, uno es capaz de encontrar grupos de textos dentro de un corpus que hablen de temas similares, asociando estos grupos a ejes temáticos [12, 6].

Se puede ver entonces que dentro del estudio de la agenda social y formación de opiniones, la detección de los tópicos y temas de interés son pasos fundamentales para caracterizar correctamente esta agenda. En este trabajo comparamos dos metodologías distintas para identificar tópicos de discusión en Twitter, Analizaremos similitudes y diferencias y buscaremos identificar los puntos fuertes y débiles de cada una.

Para responder estas preguntas, se propone estudiar una misma base de datos conteniendo tweets que hacen referencia a temas de la política argentina en el año 2019, momento de elecciones presidenciales. Para evaluar las ventajas y desventajas de cada método, y realizar un estudio comparativo entre ambas metodologías, se busca hacer un preprocesamiento de la base de datos para asegurarse que las dos detecciones de tópicos se realicen sobre los mismos tweets. Esto implica que los tweets deben ser capaces de ser analizados por los dos métodos distintos. Por ejemplo, en la red de co-ocurrencia de hashtags se estudian únicamente tweets con hashtags, entre otros criterios.

Una vez curada la base de datos, se desarrollarán las dos metodologías de detección de tópicos en capítulos separados. Más aún, en cada uno de los capítulos se mostrarán las técnicas necesarias para la detección, identificación, y caracterización de la agenda social. En estos capítulos es de esperar que se encuentren las ventajas y desventajas de cada metodología.

Métodos

2.1. Procesamiento de lenguaje natural (NLP)

Un lenguaje se puede definir como un conjunto de símbolos que pueden ser combinados para transmitir información de individuo a individuo. Esta definición es lo suficientemente general para incluir a lenguajes de programación, capaces de instruir a una computadora que pasos debe realizar para ejecutar programas, el lenguaje matemático, y el lenguaje natural. Este último se refiere al lenguaje cotidiano que usamos los humanos para comunicarnos entre nosotros. El campo de la ciencia de la computación conocido como Procesamiento de Lenguaje Natural (PLN), abreviado en inglés como NLP (Natural Language Processing), se enfoca en dotar a las computadoras, que en principio solo procesan lenguajes de programación, la capacidad de comprender y generar el lenguaje natural.

Entender el lenguaje natural es uno de los pilares del NLP y se enfoca en estudiar las palabras, las definiciones de ellas, los contextos en los que son usados, las connotaciones de estas, etc. Otro aspecto del lenguaje natural a entender es la construcción sintáctica de las oraciones, detectar el sujeto, predicado, entre otros. Por último, es importante entender el mensaje general que quiere transmitir una oración, es decir, la semántica [13].

Conociendo estos distintos niveles del lenguaje natural, las computadoras son capaces de realizar una variedad de aplicaciones.

Entre estas aplicaciones, se encuentran la traducción de texto automatizada, la síntesis de un texto, el análisis sentimental, y la extracción de información de un texto o documento. Estas aplicaciones requieren de un entendimiento del lenguaje natural en los distintos niveles.

A continuación se desarrollará cada nivel y sus aplicaciones en NLP.

2.1.1. Niveles de procesamiento

Dentro del ámbito del NLP, existen tres niveles principales para el procesamiento de un texto. Cada nivel de entendimiento es utilizado en distintas aplicaciones de NLP.

a. Morfológico o léxico

Este nivel se encarga del entendimiento de las palabras individuales, su raíz, su prefijo, sufijo, etc. Este nivel de entendimiento del lenguaje natural es aplicable para detectar en las palabras sus raíces de diccionario. Esto se conoce como la lematización de las palabras. [10]

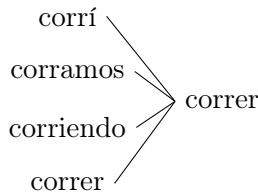


Figura 2.1: Lematización de distintas variantes de 'correr'

En la figura 2.1 se hace un ejemplo de la lematización de algunas palabras asociadas con 'correr'. Estas se asocian a su palabra raíz. De esta manera se simplifica la alta variedad de palabras en un texto y se asocian a sus palabras raíz. La lematización de las palabras se puede encontrar con problemas de ambigüedad en las definiciones. Por ejemplo, la palabra 'banco' puede referirse al banco de una plaza o el ente financiero. A veces depende del contexto que rodea la palabra entender a qué se refiere esta [13, 14].

Otro estudio a nivel léxico es la asignación de los roles gramaticales a ciertas palabras. Nuevamente, dependiendo del contexto, una misma palabra puede cumplir distintos roles. Este proceso es el que se llama *part-of-speech tag* o PoS [13].

b. Sintáctico

Este nivel se encarga de estudiar la estructura gramatical de un texto. Nuevamente, en este nivel también se realizan tareas como el PoS pero ahora para palabras que el nivel léxico no podría estudiar. Estas palabras son las conocidas como *stopwords* y son las palabras que no aportan un contenido de significado al texto. Entre estas están los conectores, las preposiciones, etc. El nivel sintáctico es capaz de detectar los roles gramaticales de estas palabras [13].

Otro de los procesos importantes de este nivel es el análisis sentimental. Este estudio de los textos detecta emociones del texto y encuentra un valor sentimental de la oración. Esto puede hacerse, por ejemplo, contando la cantidad de palabras en una oración de emociones negativas y pesarlas contra la cantidad de palabras de emociones positivas. En modelos aún más elaborados, existe el análisis sentimental dirigido a un sujeto o *targeted sentiment analysis*. Estos modelos son capaces de reconocer uno o múltiples sujetos en una oración y entender la emoción que se expresa hacia cada uno de esos sujetos por separado [15, 16].

c. Semántico

El nivel semántico se enfoca principalmente en entender el contenido del texto o del corpus. Este nivel de entendimiento de NLP sirve para la extracción de información, el resumen o detección del tópico que trabaja el mismo.

Una forma de estudiar el nivel semántico de los textos es a través del uso de la representación vectorial de los mismos [9, 10]. Esta representación es una descripción numérica de un texto. Esta representación tiene múltiples utilidades. Una de ellas es la cuantificación de la importancia de las palabras para caracterizar un texto. Otra utilidad de la representación vectorial es la capacidad de realizar estudios comparativos entre textos. De esta manera se puede describir numéricamente cuán similares son dos textos entre sí [17, 18].

Existen distintas formas de representar vectorialmente los textos, cada uno con sus distintos criterios. Entre ellas, existe la rama del modelo de espacio vectorial, VSM por sus iniciales en inglés. En este modelo, cada texto de un corpus es representado por un vector en un espacio multidimensional donde cada dimensión de este espacio se asocia a una única palabra que aparezca en el corpus. Dado un texto del corpus, las componentes de su vector asociado codifican el peso de las distintas palabras que mejor describan al texto.

Dentro del VSM, existen distintas técnicas de la representación de un texto en vectores. Entre ellas esta la 'Bolsa de palabras' (Bag of Words en inglés o BOW) y *term frequency - inverse document frequency* o TF-IDF. Cada una define el peso de las palabras de distintas formas [9, 10].

Otra forma de representar vectorialmente los textos es con modelos de machine learning. Aquellos modelos que reciben como entrada un texto y devuelven un vector numérico son llamados los *encoder*. Existen muchos modelos de tipo encoder para representar vectorialmente los textos [19, 20, 21]. Entre ellos existen las redes neuronales recurrentes, las conocidas como *long short-term memory* y también el modelo conocido como el *transformer*.

A continuación se desarrollará con mayor profundidad las técnicas mencionadas para la representación vectorial de un texto a nivel semántico y sus distintas utilidades.

2.1.2. VSM

Como dicho antes, las técnicas relacionadas con VSM representan cada texto como un vector en un espacio multidimensional, donde cada dimensión representa el peso de una palabra en el texto. El conjunto de todos estos vectores se pueden unir para formar una matriz término-documento. Estas matrices compactan la relación entre las palabras utilizadas en un corpus con cada uno de sus textos. Dado N palabras únicas y M textos en un corpus, la matriz término-documento A tiene los elementos A_{ij} que representa el peso de la palabra i -ésima en el documento j -ésimo [22].

A continuación se desarrollarán las distintas técnicas de VSM útiles para la tesis

a. Bag of Words

La técnica de Bag of Words (BOW) es una forma simple de la representación vectorial de un texto. Esta representación considera la cantidad de veces que aparecen las distintas palabras dentro de un texto, la frecuencia de las ocurrencias.

En un corpus de N documentos con M palabras, la matriz término-documento A tiene elementos

$$A_{ij} = \text{ocurrencias de la palabra } i \text{ en el texto } j \quad (2.1)$$

Para armar un ejemplo sencillo se analizan las oraciones

- *El gato negro corre rápido*
- *El perro marrón ladra fuerte*
- *Entre una pelea del gato y perro siempre gana el gato*

Antes de hacer la matriz de término-documento, se pueden eliminar palabras que no hacen efecto al análisis de los textos. Entre ellos están 'el', 'del', 'y'. Estas son palabras conocidas como *stopwords*, y suelen ser eliminadas antes de hacer una representación vectorial de los textos de tipo VSM.

	Texto 1	Texto 2	Texto 3
gato	1	0	2
negro	1	0	0
corre	1	0	0
rápido	1	0	0
perro	0	1	1
marrón	0	1	0
ladra	0	1	0
fuerte	0	1	0
entre	0	0	1
una	0	0	1
pelea	0	0	1
siempre	0	0	1
gana	0	0	1

Cuadro 2.1: Matriz término-documento de las tres oraciones.

A partir de la tabla 2.1, uno representa a cada texto como un vector de M componentes (M siendo las palabras relevantes, sin las *stopwords*). La palabra 'gato', aparece en la oración 1 y en la 3. Más aún aparece en la oración 3 dos veces. La frecuencia de las palabras es útil tambien para distinguir el peso de una palabra en las distintas oraciones.

Esta metodología sirve para representar un texto a partir de las palabras que el mismo presenta. BOW puede realizarse a un corpus de textos, o simplemente a un solo texto individual.

b. TF-IDF

TF-IDF, que significa *term frequency - inverse document frequency*, es un método que busca identificar las palabras importantes que diferencia a un texto de los otros en un mismo corpus.. Este método parte de la idea de BOW. La idea detras de BOW es que en un texto las palabras que mejor lo representan son las que más suelen aparecer. TF-IDF agrega la noción de relevancia de las palabras. Esta relevancia dice cuán importante es la palabra para la descripción del texto, no solamente con su frecuencia, si no con la especificidad de la palabra.

Por ejemplo, aunque una palabra como 'que' pueda aparecer muchas veces en un texto individual, si también aparece en la mayoría de los textos del corpus, es probable que no sea representativa del contenido específico del texto en cuestión.

Se considera nuevamente un corpus de N documentos con M tipos de palabras. Cada documento se puede representar con un vector de M dimensiones. Considere el documento i -ésimo. Tomando la componente j -ésima de su vector asociado, esta se construye por 2 factores. Uno de los factores se define como *TF* de *term-frequency* donde

$$TF_{ij} = \frac{\text{Ocurrencias de la palabra } j \text{ en el documento } i}{\text{Cantidad de palabras en el documento}}, \quad (2.2)$$

similar al BOW. A este factor TF se le multiplica el factor de especificidad IDF . Este factor le da mayor especificidad a las palabras que aparezcan en pocos textos. Esto entonces sirven para diferenciar facilmente algunos textos de otros. Este factor de especificidad se calcula como

$$IDF_j = -\log \left(\frac{\text{Cantidad de documentos que contienen la palabra } j}{N} \right) \quad (2.3)$$

Asi se construye, para el vector \vec{F}_i

$$(\vec{F}_i)_j = TF_{ij} * IDF_j \quad (2.4)$$

Se observa que una palabra que aparece en todos los documentos tendrá una contribución muy baja a la especificidad del documento j , ya que no distingue un texto de otro, a menos que aparezca mucho más frecuentemente que otras palabras. Las componentes del vector V_i con valores numéricos altos estarán asociadas a las palabras más representativas de dicho documento [23].

La matriz de término-documento se construye utilizando a los vectores \vec{F}_i como columnas, tal que

$$A_{ij} = (\vec{F}_j)_i. \quad (2.5)$$

En resumen, TF-IDF destaca las palabras que son frecuentes en un documento particular pero poco frecuentes en el conjunto del corpus, lo que ayuda a identificar términos que son distintivos y relevantes para ese documento en particular. Es necesario remarcar que la representación vectorial de un texto con TF-IDF varía dependiendo el corpus que lo contenga. TF-IDF hace una representación de un texto relativa a la del corpus, a partir del factor IDF de las palabras.

Las representaciones de tipo VSM son útiles para realizar estudios de similitud entre los documentos. Bajo la representación de BOW, uno puede ver cuán similares son los textos basado en las palabras que usan y con cuanta frecuencia las usan, mientras que en TF-IDF, comparar los vectores \vec{F}_j representa que textos tienen palabras relevantes similares y con que peso.

2.1.3. Machine Learning

Después de explorar modelos de NLP que utilizan enfoques frequentistas de las palabras, ahora se abordará un proceso basado en modelos de machine learning. Esta área de la computación se enfoca en desarrollar algoritmos y modelos que permiten a las computadoras aprender patrones de manera automatizada y realizar tareas sin ser explícitamente programadas. Estos algoritmos se entrena con grandes volúmenes de datos para descubrir los patrones y luego aplicar este análisis a datos nuevos, incluso aquellos con los que el modelo no ha sido entrenado. Cuando uno se refiere a 'datos', puede estar incluyendo una gran variedad de objetos de estudio. Entre ellos estan las imágenes, audios y textos [24].

El estudio de un texto para NLP con machine learning se puede enfocar en distintas capas de entendimiento del mismo. Un modelo podría enfocarse en preguntas generales del texto;

'¿En que idioma está escrito?' o '¿De qué habla?'. Por otro lado, existen modelos que sirven para responder preguntas más detalladas del contenido del texto, esta funcionalidad se conoce como la extracción de información.

¿Que diferencia el NLP de métodos frecuentistas con los de machine learning? ¿Como estudia uno el lenguaje natural en textos con machine learning? En particular, a lo que se presta atención es como las palabras y su ordenamiento en una frase construyen la semántica de la misma. Para métodos frecuentistas, una oración no es más que un conjunto de palabras, sin importar el orden de estas. Para los modelos de machine learning, el orden de las palabras hace una enorme diferencia en el estudio de la oración. Estos tipos de datos, que dependen del orden que se presenta la información, se llaman datos secuenciales. Las oraciones, audios y videos, entre otros, se pueden considerar como datos secuenciales [24].

Para estudiar las oraciones como datos secuenciales propiamente, se pueden usar redes neuronales que tengan en cuenta el orden de las palabras de una oración para el estudio de esta. Un ejemplo de las redes que cumplen esto son las redes neuronales recurrentes (RNN) y las long short-term memory (LSTM) [22]. Estas redes 'leen' una oración de manera secuencial y consideran el orden de las palabras que la componen. Se entrena leyendo palabra por palabra en la oración, intentando predecir la que le sigue a la actual considerando las anteriores. Aun así, estas arquitecturas presentan algunos problemas. La RNN pierde los efectos de palabras muy lejanas a la hora de predecir palabras, esto se conoce como el problema del desvanecimiento del gradiente. Otro problema con el que cuenta la RNN es con la explosión del gradiente. Esto significa que los resultados que da la red pueden ser muy distintos entre iteraciones. Esto también trae problemas ya que se espera que el entrenamiento de una red vaya actualizandola gradualmente hasta llegar a un resultado estable.

Este problema lo soluciona la LSTM. Las redes LSTM agregan una implementación que reconoce las palabras importantes para la semántica, priorizándolas en la memoria para la predicción de las próximas palabras. A pesar de solucionar el problema del desvanecimiento del gradiente, este modelo trae otras desventajas como la complejidad computacional a la hora de entrenar el modelo.

Una arquitectura que viene a resolver este problema es el transformer. Este modelo enfatiza el uso de una herramienta novedosa a la hora de buscar las palabras relevantes del texto, la 'intra-atención' [25]. Esta herramienta busca encontrar la relevancia de cada palabra para el contexto del texto pero a un costo computacional menor. Esto permite que los modelos de transformers tengan un entrenamiento más rápido que el de las LSTM [26, 25].

Dentro del mundo de machine learning y el procesamiento de los datos secuenciales, existen los modelos unidireccionales y los bidireccionales. La diferencia está en cómo le atribuyen la relevancia de cada palabra en la oración segun el contexto. Los unidireccionales le atribuyen la relevancia a una palabra dependiendo del contexto que tienen únicamente a un lado de la oración. Por ejemplo, tomando únicamente lo que esta a la izquierda de la palabra para atribuirle significado a la misma. El bidireccional usa toda la oración, izquierda y derecha, para atribuirle la relevancia a la palabra analizada. Esta última lleva a introducir a BERT (Bidirectional Encoder Representantion from Transformers). BERT trae mejoras al procesamiento de datos secuenciales, con un costo computacional menor y necesitando un volumen de datos menor a las redes anteriores para ser entrenadas [21].

Con esta nueva arquitectura, se pueden entrenar modelos para realizar representaciones vectoriales de textos, y otras tareas. Entre ellos está el Sentence-BERT, que hace un estudio semántico del texto en un nivel de la oración entera, a diferencia de BERT que originalmente estudia el texto al nivel de cada palabra. Sentence-BERT pasó a ser utilizado en la detección de tópicos, ya que es útil para encontrar tópicos generales viendo aquellos textos cuyas representaciones vectoriales sean similares. Considerar una similitud entre vectores implicaría una similitud semántica de los textos. La similitud de estos textos se cuantifica con la similitud coseno [18]. Otra tarea realizada por una arquitectura de BERT es el análisis de sentimiento, particularmente con `pysentimiento`, capaz de detectar el valor emocional de una oración, sea positivo, negativo, y neutro, y apuntado a un sujeto particular dentro de la oración [15, 16].

El grado de precisión para clasificar y entender un texto puede variar dependiendo de los criterios que uno considera para realizar representaciones vectoriales de un texto. Dependiendo el criterio, uno extrae distintos aspectos de los textos y puede realizar comparaciones cuantitativas con metricas de similitud, por ejemplo la similitud coseno.

Bag of words y TF-IDF pueden dar una noción muy general sobre lo que habla un corpus de textos, pero puede fallar en detalles específicos del mismo. Para resolver preguntas más específicas, existen los distintos modelos de machine learning entrenados para varias tareas como estudios de semántica, sentimiento, entre otros.

2.2. Redes complejas

Las estructuras complejas en forma de redes permiten describir sistemas de numerosos individuos que se relacionan por algún tipo de interacción de pares. Por ejemplo, los animales en un ecosistema presentan una compleja red de alimentación, donde las relaciones predador-presa entre especies tienen una estructura altamente compleja. Otro ejemplo son las modas e ideas se propagan en la red social a partir de interacciones entre usuarios, comentando y dando un simple 'Me gusta' a publicaciones, compartiendo links de noticias, etc. Estos sistemas representan solo unos pocos ejemplos de los muchos que pueden ser estudiados con una red compleja, modelando acordemente a los individuos y sus relaciones entre ellos [27].

Una red compleja se puede entender como un conjunto de puntos o nodos que están interconectados por medio de líneas o enlaces. Dentro de una red, un nodo representa una entidad individual, mientras que los enlaces representan las relaciones o conexiones entre estas entidades. Por ejemplo, en una red social, los nodos podrían representar usuarios individuales, mientras que los enlaces denotarían las amistades que comparten los distintos usuarios.

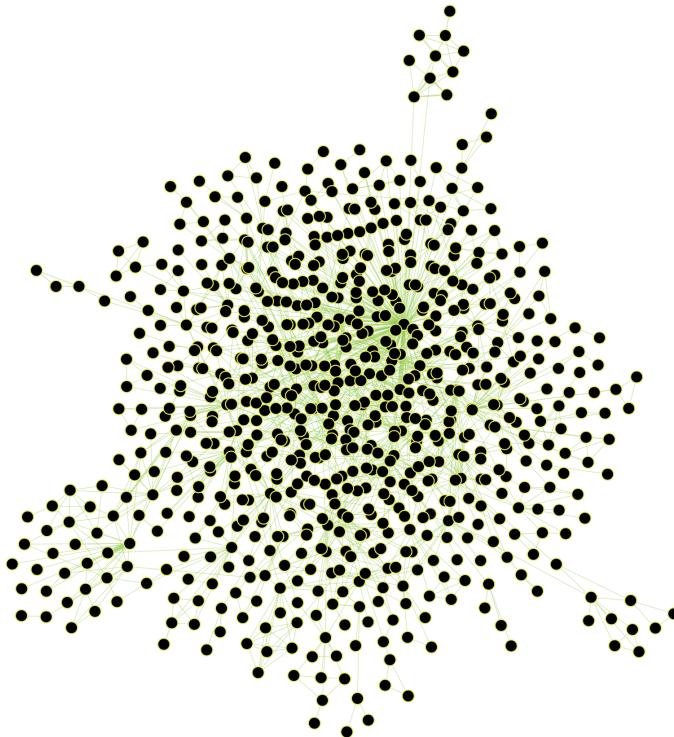


Figura 2.2: Representación gráfica de una red compleja. Un conjunto de nodos $\{n_i\}$ representado por los puntos en el espacio y los enlaces entre ellos como líneas que unen los puntos.

El conjunto de nodos de una red está representado por las variables $\{n_i\}$ y, dado un par de nodos n_i, n_j , el enlace que los une se denomina v_{ij} , el cual esta variable toma un valor numérico. Para la red binaria, los enlaces v_{ij} pueden valer únicamente 1 o 0. Estos valores muestran únicamente dos casos donde, o un par de agentes están relacionados de alguna forma, o no, respectivamente. El caso de tener una red pesada, el enlace v_{ij} puede tomar un rango de valores distintos, mostrando la intensidad de la relación que comparte el par de

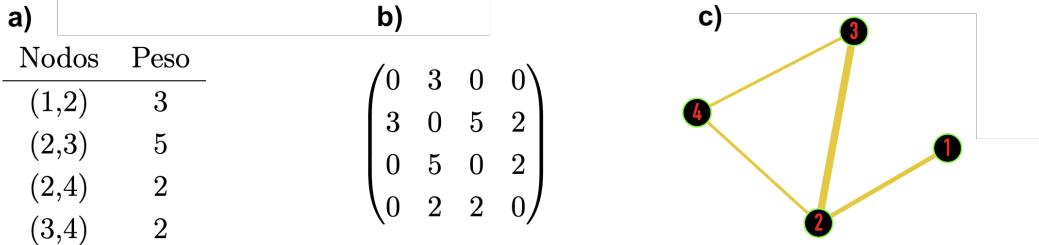


Figura 2.3: Se presentan las 2 formas de describir una red compleja. En la lista de adyacencia (a) la columna 'Nodos' codifica los pares de nodos conectados y la columna 'Peso' el peso del enlace. En la matriz de adyacencia (b) cada fila y columna representa un nodo en particular, el elemento A_{ij} de la matriz de adyacencia muestra el peso entre el par de nodos n_i y n_j . En (c) se visualiza la red resultante.

nodos.

Las redes complejas pueden ser representadas matemáticamente de distintas maneras [28]. El armado de una matriz de adyacencia suele ser utilizado para la representación algebráica de una red compleja. Una red con N nodos se puede representar en una matriz A cuadrada de $N \times N$ con los elementos $A_{ij} = v_{ij}$. Otra forma de representar una red compleja es con una lista de adyacencia. Esta muestra en una columna los pares de nodos conectados y en otra el peso de este enlace. En la figura 2.3 se ejemplifica la diferencia de estas dos representaciones en una red de 4 nodos.

Como se ejemplificó antes, las redes complejas son herramientas fundamentales para modelar una variedad de sistemas biológicos, sociales, etc. Estos modelos sirven para entender patrones del sistema a un nivel macroscópico, definiendo agentes e interacciones a un nivel microscópico. La naturaleza de estos agentes y las reglas que gobiernan sus interacciones varían según el sistema en cuestión. En el estudio de sistemas sociales, como las redes sociales, los nodos pueden representar usuarios mientras que los enlaces denotan las interacciones entre ellos. Dependiendo el estudio que se quiera realizar, los enlaces pueden representar distintas interacciones. Si es una red de amistades, un enlace entre pares de nodos representa que esos usuarios tienen una 'amistad digital' o que uno es 'seguidor' del otro, etc . Otro tipo de red que uno puede modelar es de influencias. En este caso, no necesariamente se buscan aquellos pares de usuarios que sean amigos, si no aquellos usuarios que consumen o comparten publicaciones de otros usuarios. Esta red muestra un flujo de información donde algunos usuarios publican ideas y otros le dan 'Me gusta' o las comparten a otros usuarios [7].

Se puede ver entonces que las redes complejas son una herramienta muy usada en el área de los sistemas complejos. Partiendo de un grupo de objetos individuales y reglas de interacción entre ellos, uno puede registrar comportamientos macroscópicos del sistema completo.

Dependiendo del sistema a modelar, se suelen utilizar distintos criterios para establecer los enlaces en una red pesada. En algunos sistemas, la relaciones entre agentes se determinan por relaciones frequentistas. Por ejemplo, contar la cantidad veces se cruzó un par de personas en la calle, o considerar la cantidad de procesos que aparecen el mismo par de proteínas en un sistema biológico. Este tipo de relaciones se suele llamar la co-ocurrencia de los agentes.

Otra forma de crear enlaces en una red pesada es codificando cuán similares o distintos

son dos pares de agentes. En el caso de que dos agentes comparten muchos atributos o sean similares en ciertos criterios, el enlace entre estos los dos nodos asociados será relevante, a diferencia del enlace entre nodos cuyos agentes no comparten muchos atributos. De esta forma, se crea una red de similitud donde los nodos altamente conectados presentan relaciones fuertes de similitud, mientras que los que no se asemejan tanto están débilmente conectados.

A continuación se desarrollará con mayor profundidad cada uno de los tipos de redes mencionados.

2.2.1. Red de coocurrencias

Las redes de co-ocurrencia permiten encontrar grupos de agentes que están estrechamente relacionados dentro de un conjunto de registros. La idea radica en que cuando se encuentran frecuentemente los mismos agentes juntos en registros, esa co-ocurrencia refleja una relación subyacente que probablemente sea valiosa para este par o pares de agentes. Se considera que la co-ocurrencia de dos agentes es muy alta si estos aparecen con frecuencia juntos en un conjunto de registros y lo hacen raramente separados en el resto de los registros.

Dependiendo de estos agentes que se definen, los registros también pueden cambiar para encontrar relaciones distintas entre los agentes mismos. Por ejemplo, un registro para las personas puede definirse como conversaciones que presencien dos o más personas, encuentros en las calles de una ciudad, o también la cantidad de veces que comentan en una misma publicación en redes sociales.

Dependiendo de esta definición, se encontraran distintos grupos de personas que suelan conversar, tener rutinas similares, consumir los mismos contenidos en redes, etc.

De la misma manera se pueden definir registros en palabras para encontrar relaciones semánticas de estas. Cuando uno registra pares de palabras en distintos textos, puede asumir que estos pares de palabras están altamente relacionados por una temática, un contexto histórico, área de conocimiento, etc.

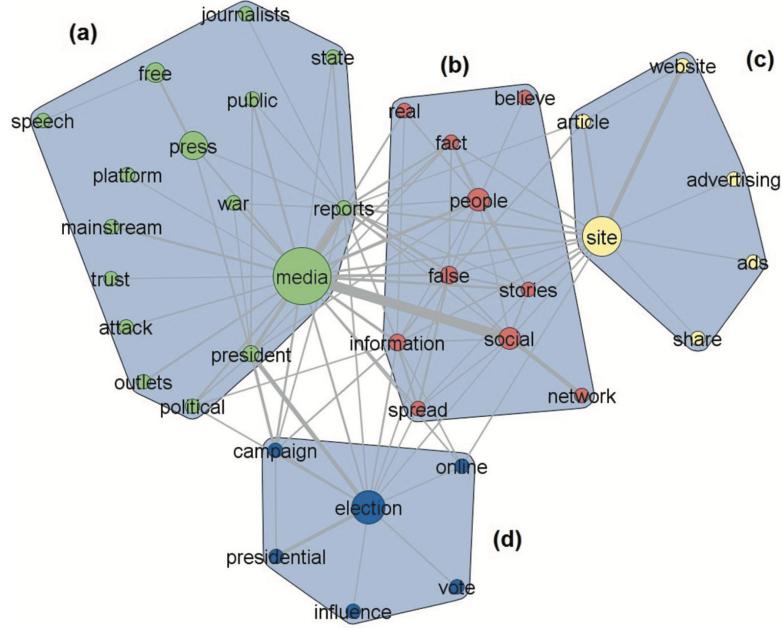


Figura 2.4: Red de co-ocurrencia de palabras clave en noticias del New York Times y The Washington Post que contengan las palabras '*fake news*'. Se detectan tambien los grupos de palabras que suelen co-ocurrir entre ellas.

El ejemplo de la figura 2.4 muestra una red de co-ocurrencia de palabras de un corpus de noticias del New York Times y The Washington Post que contengan las palabras '*fake news*'[29]. Este estudio permite detectar que tópicos aparecen en estas noticias, observando las distintas palabras que suelen co-ocurrir. Estudiando estos grupos de palabras, uno puede identificar a los temas de discusión de las noticias estudiadas. Por ejemplo, el grupo marcado como (a) muestra palabras como "medios", "prensa", "reportes", "periodistas", "libre", "expresión", "publico", y "confianza". Estas palabras pueden asociarse a el contexto de los medios de comunicación dentro de las fake news y la importacia de los periodistas, etc.

Otro grupo de palabras como el (b) presenta términos como 'personas', 'creen', 'falsa', 'información', 'comparten', etc. Estas palabras hacen referencia al comportamiento de los usuarios con las fake news, como hacen que se dispersen en la sociedad estas noticias.

Con una herramienta como la red de co-ocurrencias, existe la oportunidad de detectar estos tópicos emergentes simplemente a partir de estudiar los distintos documentos en los cuales aparecen todos los pares de palabras posibles.

2.2.2. Red de similitud

Una red de similitud es una representación gráfica donde los nodos están interconectados en función de la similitud entre ellos. En esencia, dos nodos estarán unidos por un enlace si comparten atributos o características significativas. [7]

Un ejemplo para construir este tipo de red es tomando un conjunto de vectores. Cada vector dentro de la red lo representa un nodo. Los enlaces pesados entre los vectores se representan por una medida de similitud entre ese par de vectores, estas pueden ser la distancia euclídea entre los dos vectores o la similitud coseno. Cuanto más parecidos sean los dos vectores en

la métrica tomada, más pesado va a ser el enlace de ese par. El peso del enlace suele ser proporcional a la métrica misma. Esta red que se forma asegura que, si uno tiene N nodos en la red, cada nodo en principio va a tener $N - 1$ enlaces. Esto es porque cada nodo tiene su medida de similitud con todos los otros nodos, sea alta o baja. Estos vectores pueden codificar cualquier tipo de atributos de los agentes. Si uno puede realizar una representación vectorial de estos atributos, es capaz de realizar estudios comparativos de similitud entre los agentes para representar los enlaces de la red a partir de estas similitudes.

2.3. Detección de comunidades en una red compleja

Las redes complejas pueden ser construidas mediante diferentes reglas, ya sea a través de relaciones frequentistas entre los individuos o utilizando medidas de similitud entre ellos. Como se discutió en la introducción a las redes complejas, estas reglas de interacción entre individuos permiten observar comportamientos a una escala macroscópica, los cuales son conocidos como fenómenos emergentes. En particular, algunos de los fenómenos emergentes surgen de encontrar grupos de nodos con una alta conectividad entre ellos. Estos grupos se conocen como comunidades. Existen distintos algoritmos de detección de comunidades que buscan particionar una red adecuadamente. La aplicación de estos métodos permite analizar la estructura y dinámica de las redes complejas, revelando patrones subyacentes y proporcionando una comprensión más profunda de cómo emergen los fenómenos a niveles macroscópicos en diversos sistemas [30].

Estos algoritmos dividen la red en grupos de nodos, o comunidades, basándose en ciertos criterios. Estos criterios se asocian a la optimización de una métrica propuesta, es decir, una función a maximizar o minimizar dada una partición. Entre los métodos más utilizados se encuentran el algoritmo de Louvain, donde busca maximizar la modularidad, e Infomap y Stochastic Block Model, que proponen minimizar la longitud de descripción de la red. A continuación se desarrollarán cada uno de estos métodos y sus métricas asociadas.

2.3.1. Louvain

El método de Louvain [31] es un algoritmo de clustering utilizado en muchos trabajos para la detección de comunas. Como dicho antes, cada algoritmo de *clustering* maximiza o minimiza una función o métrica en particular. Para entender cada algoritmo, es esencial entender la naturaleza de su métrica asociada. En el caso de Louvain, la métrica que se busca maximizar es la modularidad. A continuación se explicará esta métrica.

a. Modularidad

Dada una red de N nodos $\{n_i\}$ con sus enlaces $\{v_{ij}\}$, para cada nodo se puede definir su grado pesado k_i calculando

$$k_i = \sum_j v_{ij}, \quad (2.6)$$

es decir, se suman todos los pesos de los enlaces hacia el nodo n_i .

Para esta red, se propone una partición de la red en B comunas. Al i -ésimo nodo se le atribuye su etiqueta de comunidad b_i donde $b_i \in [0 : B - 1]$.

La modularidad se define como

$$Q = \frac{1}{2m} \sum_{ij} \left(v_{ij} - \frac{k_i k_j}{2m} \right) \delta(b_i, b_j) \quad (2.7)$$

donde $m = \frac{1}{2} \sum_{ij} v_{ij}$, es decir, la suma de todos los pesos de los enlaces en la red. La fórmula de modularidad se compone de una sumatoria doble que recorre todos los pares de nodos en la red. Al aplicar la delta de Dirac, se enfoca en aquellos pares de nodos que pertenecen a la misma comunidad. Para cada par de nodos, se calcula la diferencia entre dos términos. El primero representa el peso real del enlace entre los nodos, mientras que el segundo término, $\frac{k_i k_j}{2m}$, establece la esperanza del peso del enlace si los enlaces estuvieran distribuidos de manera aleatoria en la red.

En resumen, la fórmula de modularidad compara el número real de enlaces entre nodos dentro de las mismas comunidades (observado en la red real) con el número esperado de enlaces si los nodos estuvieran conectados de manera aleatoria. Un valor positivo de modularidad indica una estructura de red modular, es decir, una red con comunidades bien definidas y relevantes en comparación a una partición aleatoria. Un valor negativo sugiere una estructura menos modular, mientras que un valor cercano a cero indica que la red no tiene una estructura modular clara.

Con la idea de modularidad explicada, de explicará como funciona el algoritmo de Louvain

b. Algoritmo de Louvain

Para comprender el algoritmo de Louvain, es fundamental entender su enfoque iterativo y los pasos simples que sigue para identificar comunidades en una red. El algoritmo comienza con una asignación inicial donde cada nodo pertenece a una comunidad única. Luego, se realizan iteraciones compuestas por dos pasos básicos:

Asignación de nodos a comunidades vecinas

En este paso, se examina cada nodo y sus vecinos. Para cada nodo, se evalúa el impacto en la modularidad al moverlo a la comunidad de alguno de sus vecinos. Si esta acción aumenta la modularidad, el nodo se reasigna a la comunidad del vecino que maximice este aumento. Este proceso se repite para todos los nodos en la red.

Fusión de comunidades

Después de que todos los nodos hayan sido reasignados, se fusionan las comunidades para formar una nueva red en la que las comunidades son nodos y los enlaces entre ellas representan el flujo de nodos entre comunidades. Los enlace intracomuna se convierten en autoenlaces y los aquellos enlaces de nodos entre comunas se redefinen como un solo enlace pesado entre los nuevos nodos.

El esquema representativo de este proceso iterativo se puede ver en la figura 2.5.

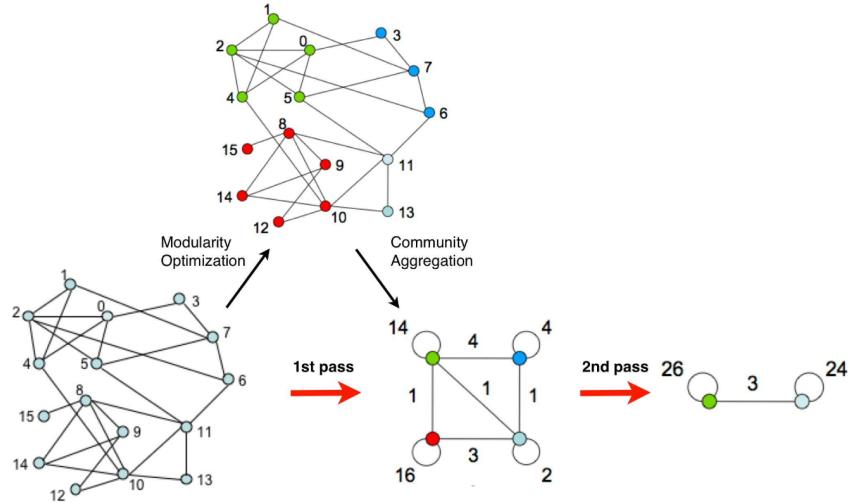


Figura 2.5: Esquema del proceso iterativo que emplea Louvain. El primer paso de asignación de nodos a comunidades vecinas o *modularity optimization* se encarga en reasignar los nodos a nuevas y menos comunidades con el objetivo de mejorar la modularidad de la partición. El segundo paso, fusión de comunidades o *community aggregation* se encarga de representar todos los nodos de una misma comunidad por un único nodo, y representar los enlaces intracomuna e intercomuna con una menor cantidad de enlaces [31].

Este proceso de dos pasos se repite iterativamente hasta que no haya mejoras en la modularidad.

En la última iteración, se obtiene una red donde cada nodo representa un subconjunto de nodos de la red original. Cada subconjunto de nodos es una comunidad definida.

2.3.2. Infomap

Otro algoritmo de clustering de interés para el trabajo es conocido como Infomap [32]. La partición que provee este algoritmo busca optimizar otra métrica llamada la entropía de la red o la longitud de descripción mínima, o MDL por sus iniciales en inglés.

Esta entropía se puede pensar como la de una caminata aleatoria por la red. Esta caminata está influenciada por los pesos de los enlaces de la red. Cada peso determina una probabilidad de transición para el caminante entre nodo y nodo.

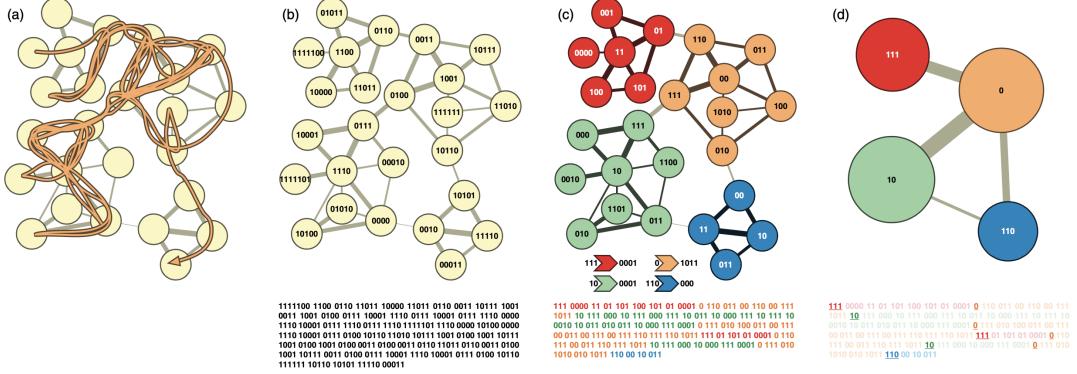


Figura 2.6: Representación gráfica del algoritmo de Infomap. En (a) se puede ver una caminata aleatoria a lo largo de un grafo. La probabilidad transición de un nodo a otro esta determinado por el peso del enlace entre los nodos. La partición final de (c) se encuentra minimizando la longitud de descripción.

Se parte nuevamente de una red no direccional de N nodos $\{n_i\}$ con sus enlaces $\{v_{ij}\}$. Para esta red, se define una partición aleatoria \mathbf{M} de los nodos, donde los grupos o módulos de nodos se definen como $\{U_j\}_{j=1\dots m}$.

Ahora se considera a un caminante que realiza una caminata aleatoria por la red. El caminante tiene una probabilidad

$$p_\alpha = \frac{\sum_j v_{\alpha j}}{\sum_{ij} v_{ij}} \quad (2.8)$$

de caer en el nodo n_α , es decir, la suma de los pesos relativos de los links que conectan al nodo n_α .

Por otro lado, si el nodo en el que se encuentra el caminante pertenece al grupo U_k y, al transicionar a otro nodo, este pertenece al grupo U_l , entonces se dará una transición del caminante de un grupo al otro si $k \neq l$, o se mantendrá en el mismo grupo si $k = l$. La probabilidad de que el caminante salga de la comunidad U_k se define como $q_{k \rightarrow}$. Esta probabilidad se calcula como la relación entre la suma de todos los pesos relativos de los enlaces saliendo de la comunidad U_k .

De esta manera, se pueden entender que para cada transición, el caminante realiza 1 de 2 situaciones. O sale de un módulo U_k a otro U_l , o se mantiene en en mismo U_k . Con esta idea en mente, se le asocia una noción de entropía a esta caminata aleatoria de la siguiente forma

$$L(\mathbf{M}) = q_{\rightarrow} H(\mathbf{Q}) + \sum_k p_{\circlearrowright}^k H(\mathbf{P}^k) \quad (2.9)$$

Esta entropía tiene dos componentes, la primera viene del término $q_{\rightarrow} H(\mathbf{Q})$. Esta es una entropía que hace referencia a las transiciones del caminante de una comunidad a otra distinta, ponderada por la probabilidad de salir de cualquier comunidad $q_{\rightarrow} = \sum_i q_{i \rightarrow}$. La otra componente de (2.9) es de la sumatoria $\sum_k p_{\circlearrowright}^k H(\mathbf{P}^k)$. Cada término de esta sumatoria hace referencia a la entropía asociada a la caminata aleatoria que ocurre dentro de cada una de las comunidades U_k , ponderada por la proporción de tiempo que el caminante está en dicha

comunidad, calculada como

$$p_{\mathcal{O}}^k = \sum_{\alpha \in U_k} p_\alpha + q_{k \rightarrow \cdot}$$

La ecuación (2.9) se conoce como la ecuación del mapa. El objetivo de Infomap es encontrar aquella partición \mathbf{M} que minimice esta entropía $L(\mathbf{M})$. Esta entropía da una noción de la información necesaria para describir el flujo dentro de la red. La partición que pueda describir ese flujo con menor información, será la partición adecuada para la red según el criterio detrás de Infomap.

Datos

Esta parte del trabajo se enfoca en la base de datos utilizada para el estudio comparativo de los dos métodos de caracterización de la agenda social. Aquí se presentará la fuente de estos datos, algunas especificaciones de los mismos, y los criterios tomados para el preprocesamiento de estos con el objetivo de realizar un estudio adecuado.

3.1. Recolección de datos

La base de datos que se utilizará en esta tesis fue construida previamente en el trabajo de [6]. Esta contiene 67336507 tweets y retweets publicados entre los meses de marzo a octubre del año 2019 en Argentina, año de elecciones presidenciales. Los tweets y retweets de la base de datos están agrupadas por su distinto mes de publicación. Los estos tweets y retweets contienen palabras clave relacionadas con la política argentina, como 'Alberto Fernández', 'Macri', 'Cristina', entre otras, así como términos asociados con el proceso electoral o partidos políticos, como 'elecciones', 'PASO', 'UCR', etc. Otros tweets mencionan a usuarios de figuras políticas relevantes del momento, como '@alferdez', '@CFKArgentina' y '@mauriciomacri', entre otros. Todos los tweets y retweets de la base de datos están en castellano. Para cada tweet o retweet de la base de datos, se registra la fecha y hora que fue creado, el texto del tweet o retweet, los hashtags utilizados, el identificador numérico del tweet o retweet y, en el caso de los retweets, la identificación del tweet original.

Antes de implementar los dos métodos de detección de tópicos, es esencial realizar un adecuado preprocesamiento de los datos. Este paso es fundamental para asegurar que el estudio comparativo se realice utilizando una misma base de datos común. El preprocesamiento se aplicó por separado a cada grupo de tweets y retweets de cada mes, e incluyó la preparación de los datos de distintos campos, que implican la selección de tweets con más de dos hashtags, suficientes retweets y estandarización de los datos mediante la limpieza de texto. A continuación se detallará cada criterio, ejemplificado en el mes de agosto, que originalmente tiene 6907089 tweets. Este mes resulta interesante para presentar los resultados ya que fue cuando ocurrieron las elecciones PASO. Estos tweets pasan por un criterio de selección que asegura poder hacer un análisis de ambos métodos sobre la misma base de datos. Con los tweets seleccionados, se toman estos y sus retweets asociados para completar la base de datos final.

3.2. Timestamp

Cada tweet o retweet cuenta con su tiempo de publicación asociado, en formato `mm/dd/aaaa hh:mm:ss`. En principio se le etiquetó a cada tweet o retweet el mes donde se publicó. Una vez agrupados los tweets y retweets por mes de publicación, se le asignó a los mismos, definidos como el conjunto $\{E_i\}$, su marca temporal asociada t_i . Esta marca temporal contiene información del día y la hora en que se publicó el tweet o retweet. Cada tweet y retweet se registra entonces su mes y fecha de publicación con resolución de 1 h.

3.2.1. Detección de tweets irrelevantes

En la extracción de tweets y retweets hecho por Twitter, como mencionado antes, se pueden exigir condiciones que deban cumplir al hacer la extracción. Particularmente, para este trabajo se pidió que estos tweets y retweets contengan algunas de las palabras clave ya mencionadas como 'PASO', o 'elecciones', 'voto', 'votar', etc. Es importante tener en cuenta que a la hora de pedir estas condiciones, Twitter es insensible a algunos tipos de caracteres. Entre ellos, la distinción entre mayúsculas y minúsculas, y las tildes. Estas insensibilidades a la hora de extraer tweets y retweets resulta en tener muchos que no sean de interés para el estudio.

Por ejemplo, muchos tweets y retweets fueron extraídos desde la API de Twitter por tener la palabra 'pasó', por la insensibilidad de las tildes de la API.

Otro subconjunto grande de tweets que pudo extraer la API de Twitter son los que hacen referencia un tiroteo ocurrido en El Paso, Texas, por la insensibilidad de las mayúsculas y minúsculas.

Por otro lado, el uso de la palabra 'voto' extrajo una gran proporción de tweets que hacen referencia a eventos de entretenimiento como los 'Teen Choice Awards' (TCA), donde las personas votan por su artista favorito y compiten estos para ganar premios.

Todos estos tweets y retweets fueron eliminados de los resultados finales de la tesis, filtrandolo por palabras clave y sensibles a tildes y mayúsculas, 'teen choice awards', 'Paso', 'pasó', etc. Esto se decidió ya que, al ser una considerable cantidad de tweets que usan palabras similares a las del contexto político en Argentina, estos son capaces de generar sesgos en los resultados y afectar a la caracterización de la agenda social. Aún así, se puede considerar que estos últimos dos conjuntos de tweets, TCA y El Paso, son tópicos de interés de las personas. El estudio de estos tópicos se pueden encontrar en el apéndice.

Este criterio no necesariamente descartó todos los tweets que hacen referencia a tópicos distintos a los de la política argentina, pero los reduce considerablemente en tamaño.

3.3. Hashtags

En esta etapa de preprocesamiento, se establecen los criterios para seleccionar los tweets que cumplan con los requisitos necesarios para construir una red de co-ocurrencia de hashtags. Además, se realizó una limpieza de texto a los hashtags.

Tal como dice el nombre de la red de co-ocurrencia de hashtags, para construirla se necesita tener tweets donde dos o mas hashtags distintos hayan co-ocurrido. Con esto en mente, para la detección de topicos con la red de co-ocurrencias no se utilizan tweets con menos de dos

hashtags distintos. La cantidad de tweets que cumplen con tener por lo menos 2 hashtags son 308489 y entre tweets y retweets hay 908197 datos.

Una vez con los tweets que cumplan los requisitos necesarios para la construcción de la red de co-hashtags, se realizó una limpieza de texto a los hashtags. La necesidad de procesar los hashtags surge de observar que dos tweets pueden tener hashtags muy similares, pero difieren de algún carácter. Esto, por ejemplo, se puede dar por la sensibilidad a mayúsculas y minúsculas. En el caso de analizar dos hashtags “#Macri” y “#macri”, el método los detectaría como dos objetos distintos.

En principio, los hashtags se estandarizaron quitando todo carácter no alfanumérico y pasando todas las letras a minúsculas. Luego, se notó la ocurrencia usual de hashtags que representan fechas, por ejemplo, fenómenosaug12’ en un tweet del 12 de agosto. Estos hashtags fueron descartados también de la base de datos, ya que no representaban ningún tipo de tópico en la agenda social.

3.4. Rank por retweets

Una vez que los tweets fueron filtrados por criterios de hashtags, se procedió a identificar aquellos que fueron retuiteados de manera significativa. Este análisis se fundamenta en la idea de que estudiar la agenda social implica examinar aquellos tweets que generan un mayor impacto y resuenan en la sociedad, en lugar de centrarse en los que pasan desapercibidos y no son compartidos por otros usuarios.

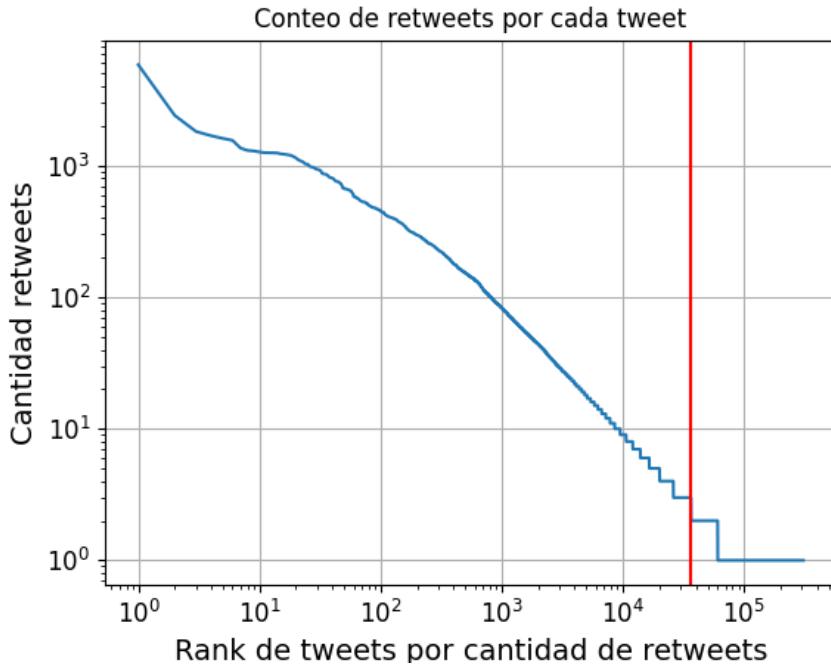


Figura 3.1: Se hace un rank de los tweets por la cantidad de veces que fue retuiteado. La linea roja vertical delimita la cantidad de tweets con 2 o mas retweets asociados (35382). Con esto se determina el nivel de alcance de los tweets en la agenda social. El estudio de la agenda social requiere de analizar información que le haya llamado la atención a las personas. No es de interés estudiar tweets que no hayan sido retuiteados.

Para ello, se realiza un *ranking* de los tweets según el número de retweets que tienen y se seleccionan aquellos que fueron más retuiteados para su posterior estudio y análisis. Este enfoque permite enfocarse en los temas y discursos que están siendo difundidos y debatidos en Twitter, brindándon una visión de la agenda social. En la figura 3.1 se presenta un rank tomando los tweets que fueron retuiteados por lo menos 2 veces. De este filtrado se llegan a tener 35382 tweets para analizar, y 371180 tweets y retweets para analizar el volumen de los tópicos.

3.5. Texto

El proceso de limpieza del texto de los tweets se llevó a cabo una vez seleccionados y guardados los tweets con al menos 2 hashtags únicos y filtrados aquellos tweets que no hayan sido retuiteados lo suficiente para ser considerado relevante. Estos textos pueden tener una variedad de elementos, como enlaces a páginas web (URLs), emojis, hashtags y menciones a otros usuarios de Twitter.

Para la limpieza de los textos se eliminaron los hashtags. Asimismo, se eliminaron todos los URLs incluidos en los textos. Posteriormente, se examinó la frecuencia de las menciones a usuarios de Twitter en los textos, considerando que estas menciones pueden desempeñar un papel relevante en la semántica de los tweets al actuar como sujetos de la oración. Como muestra la figura 3.2, en una muestra de los 35382 tweets ya seleccionados previamente, el

usuario de Macri aparece en $\gtrsim 3000$ de los tweets, Fernández otros ~ 3000 , y así siguiendo. Se optó por reemplazar las menciones a figuras políticas destacadas por sus apellidos. De esta manera se estandarizan las menciones a una misma figura política en todos los textos.

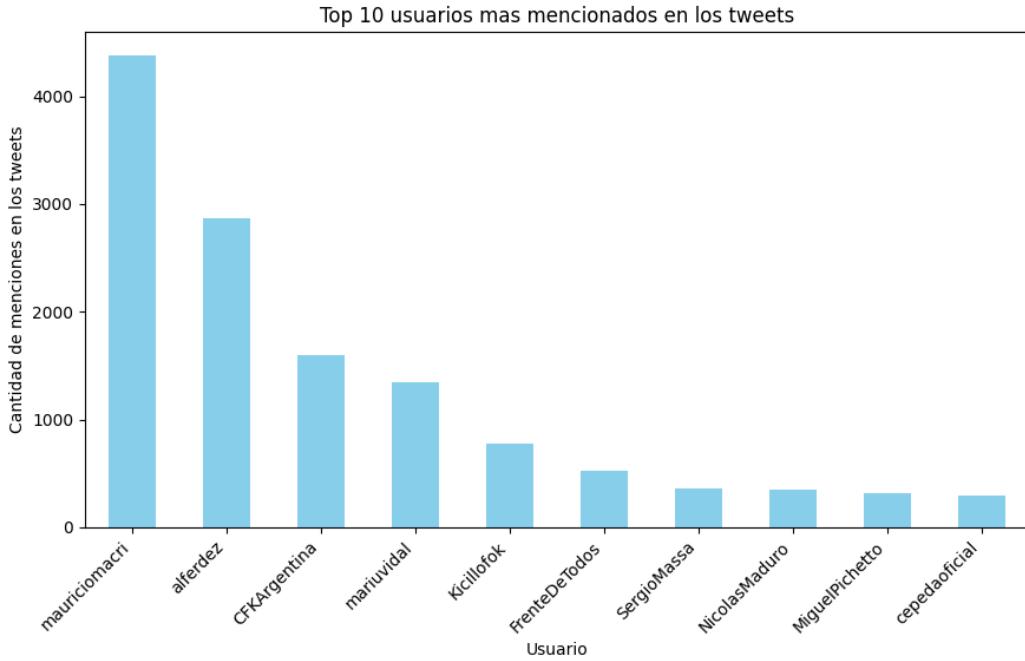


Figura 3.2: Histograma de las menciones a cada figura política en la base de datos. En ~ 35000 tweets, Mauricio Macri fue mencionado $\gtrsim 4000$ veces.

Vale aclarar, que en la figura 3.2 aparece un usuario que no pertenece a la política argentina. Esto surge del buscador de la api de Twitter, como mencionado antes.

Una vez completada esta etapa, se procedió a la eliminación de caracteres no alfanuméricos y la conversión de todas las letras a minúsculas de tanto los hashtags como los textos de los tweets. Por último, se realizó una lematización de las palabras de los tweets usando Stanza, un paquete de Python hecho por Stanford para NLP [33].

Caracterización de tópicos mediante una red de co-ocurrencia de hashtags

Esta parte del trabajo presenta la detección y caracterización de los tópicos subyacentes en los tweets seleccionados en el capítulo 3.1 usando una red de co-ocurrencia de hashtags [5]. La idea detrás de esta metodología es que esta red permite revelar la emergencia de tópicos relacionados a grupos de hashtags que co-ocurren frecuentemente. Modelar estas interacciones en una red de co-ocurrencias y aplicarle un algoritmo de detección de comunas deriva en encontrar grupos de hashtags que tienen una alta interconectividad, es decir, que frecuentan juntos en los tweets. Estos grupos son los que, bajo la hipótesis del trabajo, se considerarían los tópicos de discusión subyacentes de estos tweets estudiados.

En principio, se explicará la forma en la cual se construye la red, que características tiene y como puede ser optimizada para que represente una mejor dinámica entre los hashtags. Una vez construida la red, se le aplicará un algoritmo de detección de comunidades, en particular Louvain, para detectar aquellos grupos de hashtags distintos que se asocian con tópicos subyacentes de los tweets. Con los grupos de hashtags bien definidos, se analizará cada uno de estos por separado para identificar que eje temático representan. Para cada tópico se estudiará el volumen de tweets y retweets asociado y se analizará la serie temporal del mismo para ver el comportamiento a lo largo del mes de agosto, y que eventos dentro del mismo pueden explicar su comportamiento.

4.1. Construcción de la red

Esta sección del trabajo se dedicará a explicar la fase inicial de construcción de la red y el estudio de sus propiedades principales. Posteriormente, se utilizará un modelo nulo para distinguir aquellos enlaces que hayan sido generados por causas aleatorias e irrelevantes de los que son de interés para el estudio. Los enlaces identificados como aleatorios serán eliminados de la red. Finalmente, se estudiará la estructura resultante de la red, analizando su topología y destacando su coherencia temática en los distintos grupos conexos de la red.

a. Desarrollo inicial de la red

Como se mencionó en la subsección 2.2.1 una red de co-ocurrencia representa las relaciones entre agentes que pueden encontrarse juntos en distintos registros. En este caso, los registros a considerar son los tweets que cumplen los requisitos de la sección 3.1 y los agentes son todos

los hashtags que aparezcan en estos tweets. La red de co-ocurrencia que se construye tiene enlaces entre pares hashtags que hayan sido utilizados en un mismo tweet.

Para armar la red de co-ocurrencias de hashtags, se tomó como el conjunto de nodos a todos los hashtags $\{h_i\}$ de los tweets que cumplen todos los criterios mencionados en la sección 3.1. El peso v_{ij} del enlace entre los nodos hashtags h_i y h_j lo determina la cantidad de co-ocurrencias de este par de hashtags en todos los tweets mencionados. En consecuencia, $v_{ij} \in \mathbb{N}$. En este caso, la red construida contiene 22164 hashtags.

b. Selección de enlaces significativos

Una vez establecidos todos los enlaces dado la co-ocurrencia de hashtags se evalúan cuáles de los enlaces son relevantes para la detección de los tópicos de la red. Esto surge de la observación que algunos hashtags pueden tener enlaces muy pesados, solo por el hecho de que ambos hashtags son muy utilizados en Twitter, no por el hecho de que tengan una relación temática fuerte.

Se utilizó el mismo criterio de los trabajos de [8, 34] para determinar que enlaces de la red son considerados relevantes.

La relevancia de un enlace se evalúa calculando la probabilidad de obtener el peso del mismo en un modelo nulo de la red. En otras palabras, asumiendo la hipótesis que la ocurrencia de un hashtag en un tweet es independiente de la ocurrencia de otro hashtag, ¿cuál es la probabilidad de que estos dos hashtags co-ocurran la misma cantidad de veces que el peso del vértice?

Considerando un par de hashtags h_i y h_j , cuyo enlace tiene un peso v_{ij} , calculamos la probabilidad del modelo nulo de este enlace como

$$p_{ij} = \sum_{k \geq v_{ij}} p_{ij}(k). \quad (4.1)$$

En la ecuación (4.1) $p_{ij}(k)$ es la probabilidad de que el peso del enlace entre los nodos n_i y n_j sea exactamente k para el caso del modelo nulo y se define como

$$p_{ij}(k) = \frac{\binom{N}{k, c_i, c_j}}{\binom{N}{c_i} \binom{N}{c_j}} \quad (4.2)$$

donde los valores de c_i y c_j son las ocurrencias de cada hashtag en todos los N tweets usados para el armado de la red, y

$$\binom{N}{k, c_i, c_j} \equiv \frac{N!}{k! c_i! c_j!}.$$

Nótese que las probabilidades de la co-ocurrencia en el modelo nulo se basa únicamente en la ocurrencia total de cada hashtag.

Para eliminar enlaces, se debe definir un p-valor p_0 donde, si la probabilidad $p_{ij} > p_0$, entonces no se puede decir con confianza que el peso v_{ij} haya sido generado por una correlación entre los hashtags h_i y h_j . En tal caso, es preferible para el estudio eliminar ese enlace y no relacionar esos hashtags en la red.

Cuanto menor sea el valor de p_0 , menos enlaces quedarán en la red. Esto implicará que algunos nodos puedan quedar aislados de todo el resto, y se perderán en el estudio de los

tópicos. Por otro lado, valores bajos del p-valor asegurarán que esos enlaces restantes sí existen por una causa temática y no por azar. El p-valor tomado para este estudio es de $p_0 = 10^{-2}$ y con este valor la red contiene 21492 hashtags, aproximadamente un 97% de los hashtags de la red original.

c. Topología resultante de la red

Con la red compleja armada y con los enlaces irrelevantes eliminados, se observa cuán conexa es esta red resultante. La conectividad de la red da una idea de cuán relacionados están los hashtags. Puede ocurrir que la topología de la red varíe dependiendo de la relación subyacente de los hashtags. Si existen distintos grupos de hashtags en la red disjuntos entre ellos, esto podría significar que los tópicos que hacen referencia cada grupo hashtags son completamente independientes entre ellos, hasta podrían representar cada grupo un distinto tópico de discusión. Por otro lado, puede ocurrir que todos los hashtags de la red estén conectados de tal forma que exista siempre un camino de enlaces que conecte un hashtag con otro.

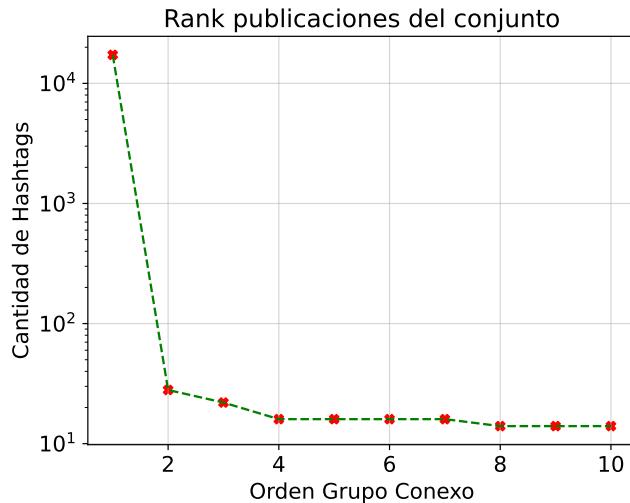


Figura 4.1: Se ordenan los conjuntos conectados de hashtags de mayor a menor volumen de hashtags. De esta manera se da una idea de como es la topología de la red. El gráfico muestra el ranking para los primeros 10 grupos conexos más grandes.

La figura 4.1 muestra el orden de los primeros conjuntos conexos de la red por cantidad de hashtags que lo componen. La red, una vez que se eliminaron los enlaces no relevantes, se partitiona en 1405 grupos de 2 o más hashtags conexos. De la figura 4.1 se puede observar que de estos grupos conexos existe un conjunto de hashtags principal, con la gran mayoría de los hashtags conectados. Este conjunto contiene un 80% de los hashtags. Los conjuntos que le siguen en el rank por tamaño tienen una cantidad significativamente menor de hashtags. En la figura 4.2 se observa un ejemplo de los conjuntos conexos pequeños, después de la componente principal.

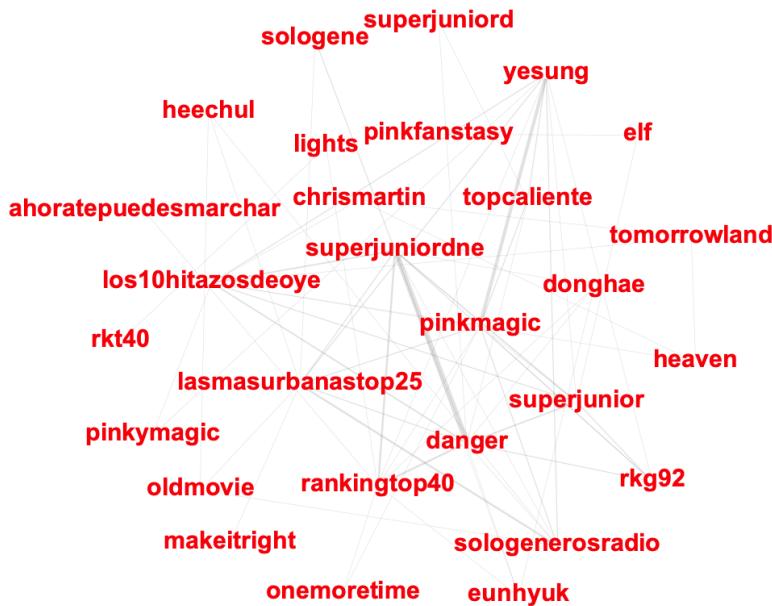


Figura 4.2: Se grafica la tercera componente conexa más grande. En este grupo de hashtags se pueden ver palabras relacionadas por el tópico de la musica pop coreana, tales como 'super junior', una banda de k-pop, o canciones del género como 'pink magic', etc.

Un ejemplo de los grupos conexos siguientes al principal se visualizan en la figura 4.2. La idea que refleja esta figura es que los grupos conexos pequeños de la red pueden representar tópicos ya definidos de la agenda social. Se puede ver que hay una coherencia marcada entre sus hashtags. Si uno investiga los hashtags de este conjunto conexo, puede deducir rápidamente que hace referencia a bandas y canciones del pop coreano.

Aunque estos conjuntos mostrados no hagan referencia a política argentina, se pueden seguir considerando parte de la agenda social. Algún grupo de hashtags puede conformarse por pocos hashtags o incluso pocos tweets, pero puede estar altamente retuiteado por las personas, haciéndolo importante en la agenda social. Por ende, uno podría considerar estos grupos de hashtags para el futuro del estudio.

Por otro lado, la componente conexa más grande contiene la gran mayoría de los hashtags. Estos hashtags pueden conformar varios tópicos distintos. Ante esto, es necesario desglosar este conjunto conexo para entender cuales son los grupos de hashtags que representen tópicos. A la componente más grande se la particionará con el algoritmo de detección de comunas Louvain.

4.2. Detección e identificación de tópicos

La sección anterior se enfocó en construir la red de co-ocurrencia de hashtags, remover aquellos enlaces que se construyeron por azar y estudiar la red resultante con sus componentes conexas. En esta sección del capítulo, se realiza la detección, identificación y caracterización de los tópicos subyacentes en los tweets.

La detección de tópicos se lleva a cabo con un algoritmo de detección de comunas aplicado a la red. Luego de hacer esto, se caracteriza la partición creada por el algoritmo; las comunidades que se detectan y los tamaños de estas en función de los nodos. Luego, se le identifica el tópico asociado a las comunidades. Esta identificación se hace a partir de un análisis cualitativo de los hashtags que la componen, y la ocurrencia de cada uno de los tweets utilizados para el armado de la red.

Finalmente, con los tópicos identificados, se estudiará la presencia de cada tópico en la base de datos estudiada. La presencia se determina a partir del volumen de tweets y retweets que mencionan a estos tópicos. De esta forma se puede ver cuán presente está dicho tópico en la agenda social.

a. Detección de comunidades

A la red resultante de la sección red se le aplica el algoritmo de Louvain para encontrar las comunidades de la red. Louvain partió la componente conexa principal en 67 comunidades. Considerando cada componente conexa siguiente como un propio tópico, se llega a tener en total 1471 grupos de hashtags que podrían, en principio, considerarse como tópicos. En la figura 4.3 se muestran los distintos tamaños en hashtags de las comunas en la red.

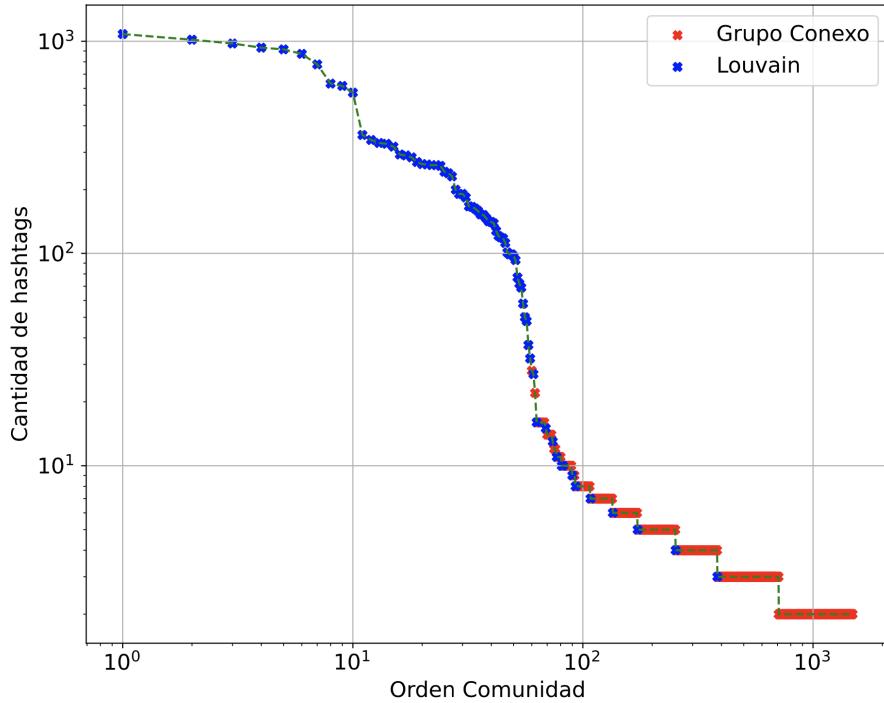


Figura 4.3: Se ordenan las comunidades por la cantidad de hashtags que la componen. Las marcadas en azul son los grupos de hashtags que inicialmente estaban en el grupo conexo principal y fueron detectados como una comunidad por Louvain. Los rojos son aquellos grupos conexos pequeños que fueron considerados como una misma comunidad o tópico.

De la figura 4.3 se puede ver que Louvain partitiona la componente principal en algunas comunidades grandes en nodos y detecta otras considerablemente más pequeñas en relación. Mas aún, las primeras 10 componentes más grandes en hashtags presentan un tamaño del mismo orden, variando desde la más grande con 1081 hashtags a décima con 573.

Tanto estas comunidades detectadas por Louvain como los grupos pequeños de hashtags conexos, se considerarán en teoría como los tópicos que conforman la agenda social.

Uno podría, en principio, tener interés en las comunidades más voluminosas en hashtags. Esto es porque podrían considerarse las más importantes en la red. Por ende, vale la pena estudiar que tópicos conforman. Este paso es una forma de evaluar si la metodología de la red de co-ocurrencia funciona para caracterizar correctamente la agenda de una red social. Si se encuentran tópicos conformados por bolsas de hashtags que muestran una coherencia temática, se tiene un buen indicio de la efectividad de la metodología.

Para no extender el estudio a las 1471 comunas en total, se analizan las 10 comunidades más voluminosas en hashtags. Para esto, se observan los hashtags de cada comunidad y se las graficó en WordClouds. El ejemplo de la comunidad más grande se encuentra en la figura 4.4. Cada hashtag en el WordCloud tiene un tamaño en proporción a la cantidad de ocurrencias que tiene en todos los tweets.



Figura 4.4: WordCloud para el tópico con la mayor cantidad de hashtags. Se grafican los 15 hashtags del tópico con mayor ocurrencias en los tweets. Se puede inferir por los hashtags que el tópico hace referencia al conflicto de Venezuela con los EEUU.

En 4.4 se presenta la primera comunidad más voluminosa en hashtags con 1081 hashtags, donde sus hashtags más predominantes son 'Venezuela', 'Venezuela Heróica', 'No more Trump', 'Venezuela Unida en Batalla', etc. Estos hashtags, particularmente en el contexto de agosto de 2019, referencian a acontecimientos ocurridos en Venezuela, principalmente el bloqueo comercial hecho por los Estados Unidos. Si uno extiende el análisis a los siguientes hashtags, aquellos con mayor cantidad de ocurrencias también se muestran altamente relacionados con temas de Venezuela y su relación con los EEUU. El hashtag 'tuiterosactivos', aparece en tweets que muestran un apoyo hacia Maduro, y rechazo a los Estados Unidos y Donald Trump. Se puede decir con confianza que este tópico hace referencia a la situación política y económica de Venezuela en el mes de agosto de 2019 y su conflicto con los Estados Unidos.

A continuación en la figura 4.5 se presenta los WordClouds de las siguientes 9 comunidades más grandes de la partición.



Figura 4.5: WordClouds de los 9 tópicos siguientes al de Venezuela y EEUU con mayor cantidad de hashtags.

En la figura 4.5 se muestran los WordClouds para las 9 comunidades con mayor cantidad de nodos después de la del tópico asociado con el conflicto de Venezuela y Estados Unidos.

En el WordCloud de la parte a) en 4.5, se presenta un tópico de 1017 hashtags, principalmente asociados a la crítica y el rechazo a Macri. En él se ven hashtags como 'habrán consecuencias', 'Macri es caos', 'riesgo Macri', etc. Entre estos hashtags también aparecen hashtags como 'se van en primera vuelta', haciendo referencia a las elecciones y hasta el libro de Cristina Fernández de Kirchner, 'Sinceramente'. Este tópico se define como 'Rechazo a Macri' y no se define como 'Kirchnerista' o 'A favor de Alberto Fernández' ya que únicamente se presentan hashtags en relación a Macri, con connotación negativa. Por otro lado, la detección de tópicos fue capaz de detectar un tópico distinto que presenta apoyo a Fernández y al Frente de Todos (FdT). Este tópico presenta hashtags frecuentes como 'Alberto presidente', 'Alberto y Cristina', 'Axel gobernador', entre otros.

Por otro lado, la comunidad mostrada en la figura b) contiene 976 hashtags que no hacen referencia a un tópico específico si no a palabras generales del día a día de la política argentina. Estas palabras como 'elecciones', 'Argentina', 'kirchnerismo', 'dolar', 'politica', pueden considerarse como palabras clave de la política y economía general del país. Dentro de los tweets de la base de datos, aquellos que contienen estos hashtags están hablando de política argentina. Algunos de estos, aun así, pueden presentar apoyo a Macri y otros apoyo a Alberto Fernández.

En tercer lugar, el que muestra la figura 4.5 c), tiene hashtags que presentan un claro apoyo a Mauricio Macri y al partido de Juntos Por el Cambio (JxC). Esto se puede ver tanto en 'Yo

voto MM', 'No Vuelven Mas' y otros como '24a vamos todos', '24a yo voy', etc. Estos ultimos hacen referencia a la marcha del 24 de agosto en apoyo a JxC. Este mismo análisis se puede realizar para todos los otros WordClouds de los tópicos. Algunos siguen siendo identificables con mucha claridad, como los tópicos en e) siendo 'Apoyo a Frente de Todos (FdT)', en f) como 'Elecciones PASO' y en g) 'Fútbol'. Otros, pueden requerir de mayor esfuerzo en identificarse. El tópico que se muestra en d), usa hashtags como 'Macri', 'Cristina', es decir figuras políticas, y después surgen mucho las palabras 'Ahora' y 'Animales Sueltos', entre otros. Más aún, observando los tweets que contienen estos hashtags se ve una alta proporción de tweets con citas y con otras referencias a medios de comunicación como C5N. Esto fue considerado suficiente para identificar a este tópico como 'Artículos de Noticias', refiriendo a tweets y retweets con un tono periodístico.

Otro de los tópicos que debió ser analizados en mayor profundidad fue el que se muestra en la subfigura h). Este se puede ver a primera vista que engloba varios países y, analizando los tweets que tienen estos hashtags, suelen hablar de temas de sociopolítica, economía o hasta entretenimiento de los países. Se decidió generalizar este tópico como 'Internacional'.

El ultimo de los tópicos de 4.5 que debió ser observado con atención fue el que aparece en i). Principalmente aparecen palabras como 'cdmx', 'amlo', 'alerta', 'oaxaca' y 'morena'. Estos hacen referencia a ciudades o estados de mexico, al presidente de Mexico Andrés Manuel López Obrador o AMLO, y a diversos acontecimientos que ocurrieron en Mexico como crímenes, elecciones internas del partido PRI, anuncios del presidente, etc. Todo este tópico se catalogó como 'México'.

Con los primeros 10 tópicos identificados, se graficó un muestreo de la red compleja asociada a estos tópicos de discusión, como muestra la figura 4.6.

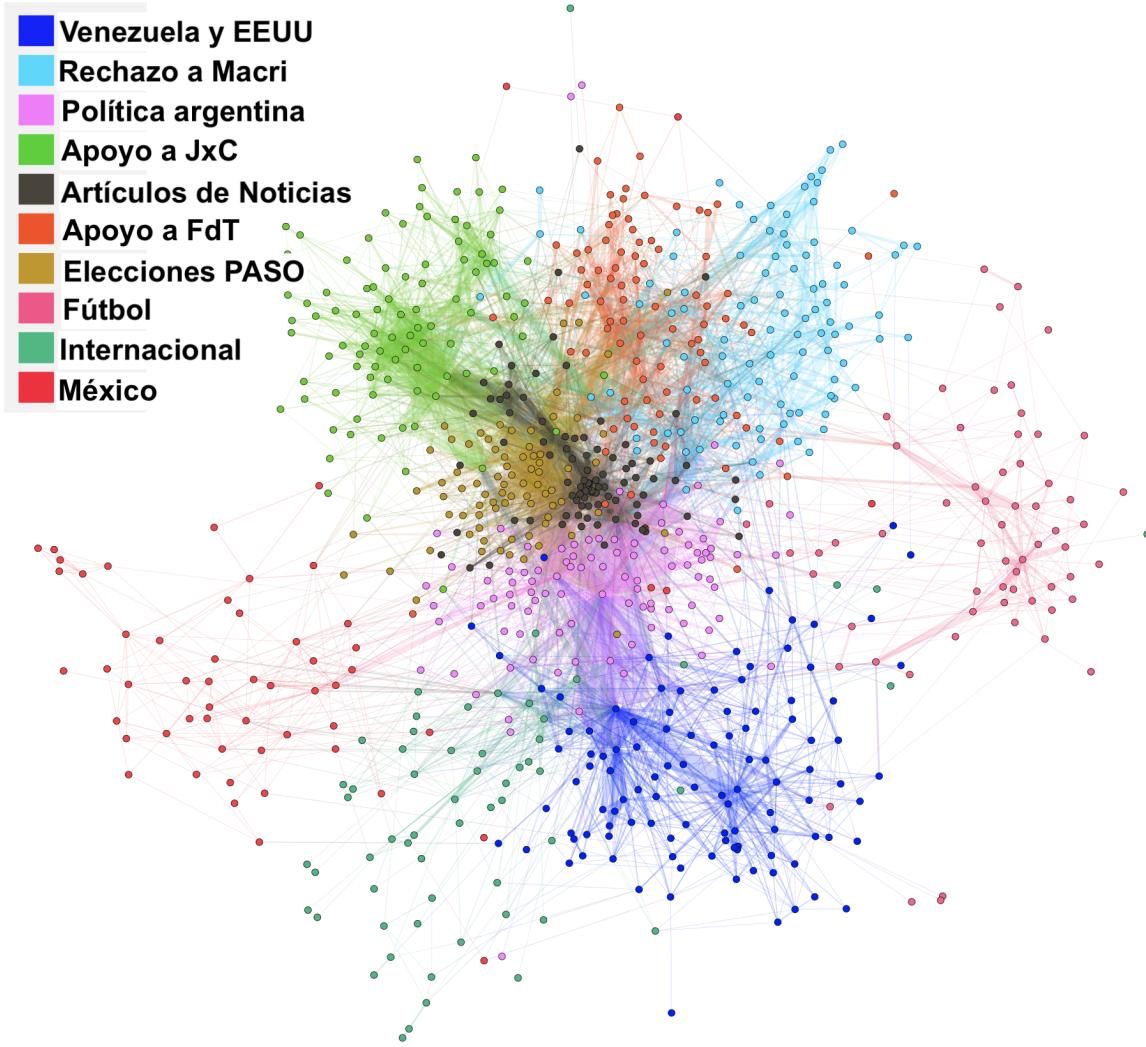


Figura 4.6: Muestreo de la red compleja para los tópicos con mayor cantidad de hashtags. Cada color es un tópico y estan identificados la leyenda ordenados por cantidad de hashtags.

En la figura 4.6 se muestran los tópicos que cuentan con la mayor cantidad de hashtags y como interactúan en la red de co-ocurrencia de hashtags. Se puede ver de la figura 4.6 que aquellos tópicos que se relacionan con temas de política argentina están altamente conectadas entre ellas tales como 'Apoyo a JxC', 'Elecciones PASO', 'Política argentina', 'Rechazo a Macri', 'Artículos de Noticias', etc. a diferencia de otros tópicos menos relacionados como 'Venezuela', 'Fútbol', 'México' e 'Internacional'. En otras palabras, la red de co-ocurrencias refleja la alta relación de los tópicos sobre la situación argentina y el panorama político. Si uno nota, se puede visualizar la alta conexión de enlaces entre los nodos del tópico de 'Rechazo a Macri' con el de 'Apoyo a FdT', dando indicio a una polarización de los partidos políticos JxC y FdT, y sus seguidores.

Como se puede ver, la visualización de la red de co-ocurrencia de hashtags permite ver la dinámica interna de los tópicos. Un par de tópicos de interés que se pueden observar son los del apoyo a JxC y a FdT. Se decidió observar una muestra reducida de la red de co-ocurrencias formada por los hashtags más mencionados de estos dos tópicos. Esto permite entender cuantas

veces se observan la co-ocurrencia de hashtags de estos dos tópicos en un mismo tweet.

En la figura 4.7 se grafica la relación de los tópicos en apoyo a JxC y FdT en la red de co-ocurrencia de hashtags, mostrando los hashtags más ocurrentes en los tweets.

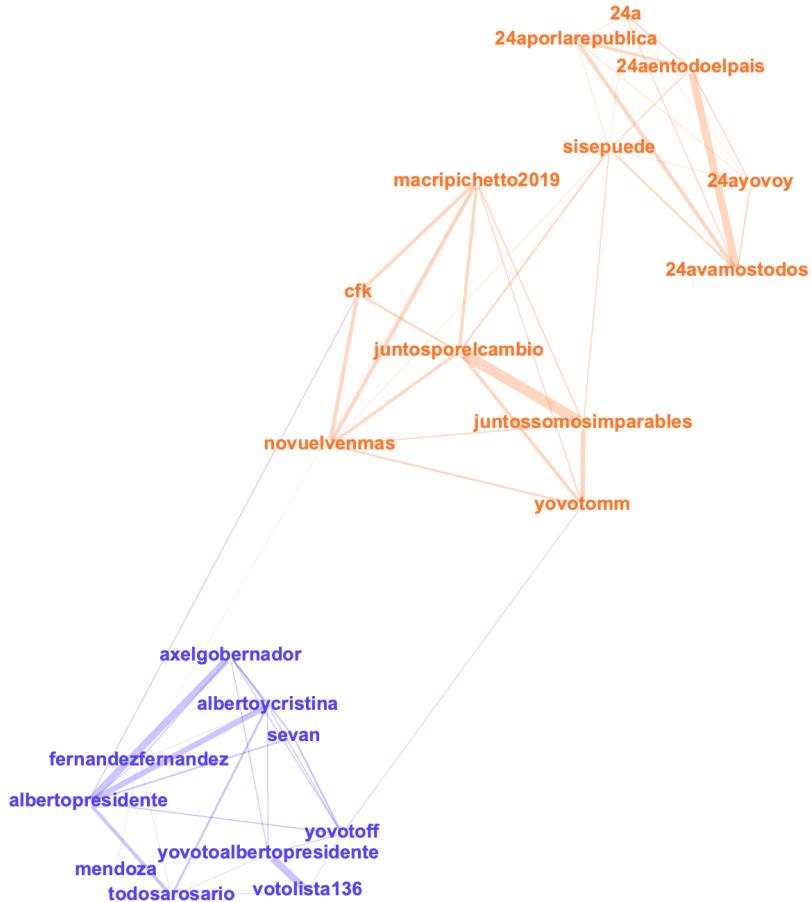


Figura 4.7: Subconjunto de la red de co-ocurrencia que contiene los hashtags más mencionados por los tópicos en apoyo a JxC (rojo) y a FdT (azul).

De la figura 4.7 se puede ver que los hashtags marcados en rojo coinciden en mensajes de apoyo a JxC o en rechazo al partido asociado a Alberto Fernández, FdT. De la misma manera, los hashtags en azul muestran una coherencia en mensajes de apoyo a Alberto Fernández y Cristina Kirchner, candidatos a presidente y vicepresidente de FdT respectivamente, y de rechazo a miembros del partido político opuesto, principalmente hacia Mauricio Macri. Más aún, se puede ver la alta conectividad entre los hashtags de un mismo tópico, en comparación a las conexiones entre hashtags de tópicos diferentes. En algunos casos, los tweets que usan hashtags de ambos tópicos son encuestas para preguntar a otros usuarios a quienes votan en las PASO.

Hasta ahora se dió a conocer la partición de la red, los tópicos más grandes en hashtags y la dinámica de estos tópicos en la red de co-ocurrencia. Todavía queda por determinar adecuadamente cuales de las comunidades de la partición representan los tópicos más importantes de la agenda social. Uno podría argumentar que un tópico voluminoso en hashtags, no significa que sea un tópico predominante en la agenda social. Uno puede estudiar el volumen de los tópicos de otra manera, en la presencia de los tweets y retweets. A continuación se encontrará

la manera de definir que tópicos son predominantes en la agenda social a partir de su volumen medido en ocurrencias en los tweets y retweets.

b. Volumen de los tópicos

Para caracterizar la agenda social, se propone buscar los tópicos que más presentes estén en toda la base de datos. Esto se puede hacer observando el volumen de tweets y retweets que hacen referencia a cada uno de los tópicos.

Con esta pregunta, surge un desafío para definir el volumen de tweets y retweets asociados a un tópico. El desafío reside en que un tweet o retweet puede contener varios hashtags de distintos tópicos. Asociar un tweet a solo un tópico puede resultar en perder información de lo que habla ese tweet. En esta sección se atacará este desafío con una propuesta que evita esta pérdida de información.

Se considera un tweet o retweet que contiene H_i hashtags del tópico i -ésimo, donde $i = 1, 2, 3 \dots I$ siendo I la cantidad de tópicos detectados y $H_i \in \mathbb{N}$. Podemos definir para el tweet o retweet j -ésimo un vector \vec{P}_j , donde sus componentes se definen como

$$(\vec{P}_j)_i = \frac{H_i}{\sum_l H_l}, \quad j = 1, 2 \dots J \quad (4.3)$$

donde J es la cantidad de tweets y retweets de la agenda social. Con estos vectores, se da a entender el porcentaje de cada tópico que reside en un mismo tweet o retweet. Así se llega a una noción de volumen de los tópicos en un tweet o retweet.

Esta definición evita colapsar un tweet entero en un solo tópico. Más aún, se puede contemplar cuánto se menciona a cada tópico por separado en un mismo tweet o retweet. Decir que una publicación puede estar hablando de un solo tópico puede simplificar el estudio perdiendo información valiosa.

Una vez que se dispone de esta definición del volumen de cada tópico por tweet, queda calcular el volumen que ocupa un tópico en todos los tweets y retweets. Esto se puede hacer sumando la contribución de cada vector \vec{P}_i de la forma

$$W_i = \sum_{j=1}^J (\vec{P}_j)_i. \quad (4.4)$$

Ahora hay una forma de evaluar la presencia de cada tópico en todas las publicaciones. Vale notar que $\sum_{i=1}^I W_i = J$.

A partir de ahora, para cada tópico se puede calcular su peso en la agenda social considerando el volumen de tweets y retweets que ocupa. Esto determina un orden de prioridad de un tópico por sobre otro, cuando se buscan los temas de discusión principales de la agenda.

La agenda social se determinó a partir de los primeros tópicos con mayor presencia en la base de datos.

c. Caracterización de la agenda

Con la definición presentada del volúmen de un tópico en tweets y retweets, se puede definir la agenda social a partir de los tópicos más voluminosos. Para esta tesis, se presentarán los 10 tópicos con mayor presencia en la agenda social.

En la figura 4.8, se presenta la red representativa de los tópicos con mayor peso en la agenda social y su tabla de volúmenes en tweets y retweets.

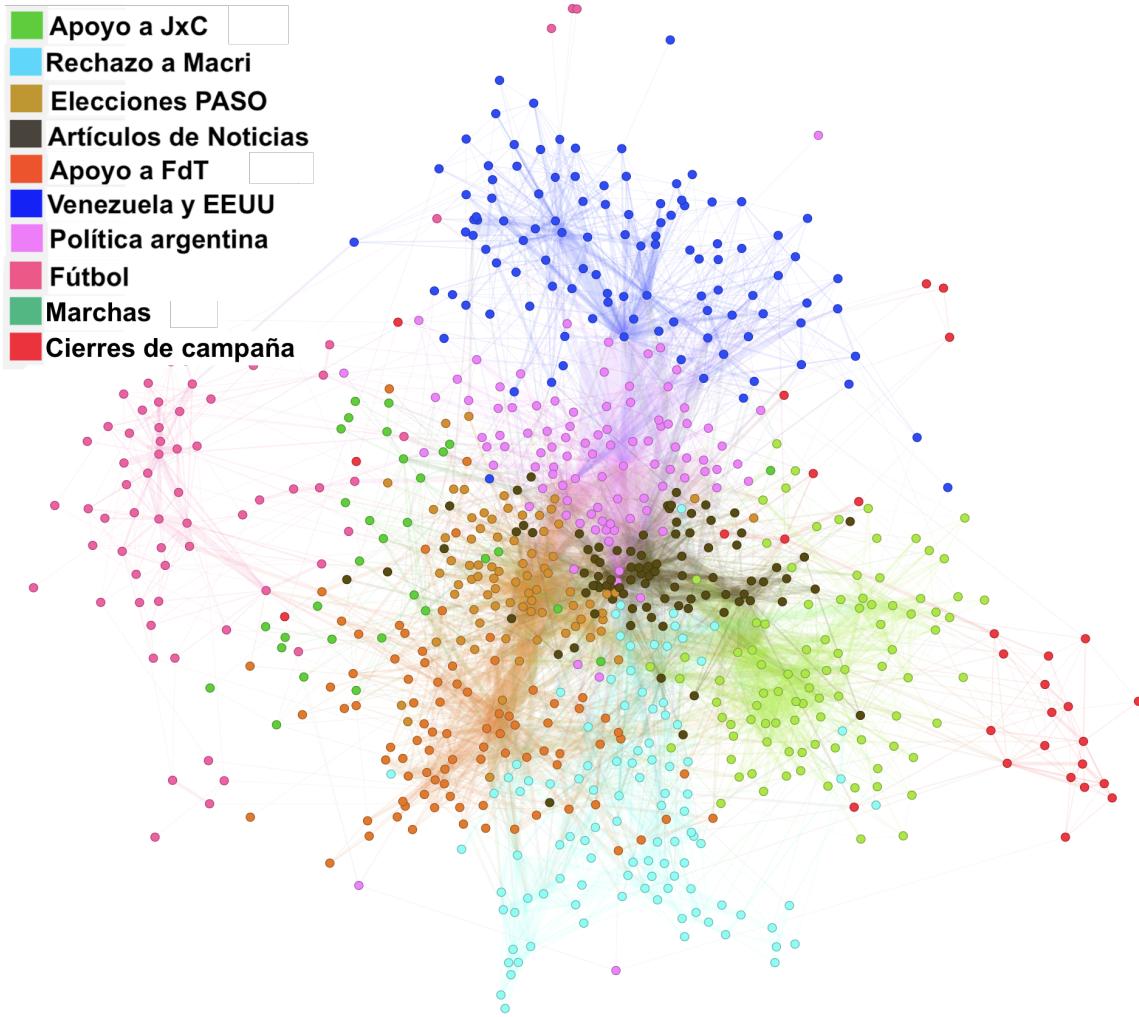


Figura 4.8: Muestreo de la red compleja para los tópicos con mayor presencia en la agenda social. La leyenda ordena a los tópicos por su volumen en tweets y retweets.

En la figura 4.8 se muestran los tópicos más mencionados en la agenda social de los tweets y retweets estudiados. La leyenda de esta figura ordena los tópicos por su volumen de tweets y retweets. Se puede ver como los primeros 5 tópicos de mayor interés sí hacen referencia a temas de política argentina, y recién después aparecen otros tópicos como 'Venezuela y EEUU' o 'Fútbol'. Más aún, todos los tópicos que se relacionan con la política argentina se muestran fuertemente relacionados e interconectados en la red, mientras que los hashtags de tópicos como 'Venezuela y EEUU', o 'Fútbol' se muestra menos conectados con los otros tópicos de la red. Esto resulta interesante ya que la misma metodología de la red de co-ocurrencia de hashtags muestra la alta relación que tienen todos los hashtags de política, pero aún así se pueden seguir segmentando en tópicos más específicos con el uso de un algoritmo de detección de comunidades en la red.

La tabla 4.1 se ve el volumen que tiene cada tópico medido en tweets y retweets. Es interesante notar como, ordenando a los tópicos por su volumen en tweets y retweets, efectivamente se observa una agenda política mayoritariamente politizada con algunos tópicos por fuera de la política argentina. Por otro lado, el tercero, hacer referencia a las PASO, un evento altamente relevante en el país, que es de esperar que tenga una alta presencia en la agenda.

Orden	Tópico	Volumen
1°	Apoyo JxC	48009
2°	Rechazo a Mauricio Macri	46985
3°	Elecciones PASO	42082
4°	Artículos de noticias	31290
5°	Apoyo a FdT	31252
6°	Venezuela y EEUU	28772
7°	Política argentina	26511
8°	Fútbol	8802
9°	Marchas	7463
10°	Cierres de campaña	5699

Cuadro 4.1: Tópicos rankeados por su volumen W_i de tweets y retweets.

En la tabla 4.1 se muestran los tópicos identificados ordenados por el volumen que ocupa cada uno en los tweets y retweets.

Es interesante ver que los tópicos principales de la agenda no necesariamente hacen referencia a temas de política nacional. En la agenda social también puede enfocarse en acontecimientos tanto internacionales de conflictos sociales o hasta de entretenimiento.

Como se mencionó antes, algunos de los tópicos detectados hacen referencia a acontecimientos puntuales, como las elecciones PASO, mientras que otros pueden desarrollarse a lo largo de todo un mes o el año entero, como las opiniones que se tienen hacia Mauricio Macri. Es de esperar que cada uno de estos grupos de tópicos tengan dinámicas distintas a lo largo del tiempo.

4.3. Series temporales

Hasta ahora se hizo una detección e identificación de los tópicos, seguido de una caracterización por el peso que tiene cada tópico en la agenda social basado en el volumen de tweets y retweets que los mencionan.

Otro aspecto que se podría estudiar en cada tópico es su evolución temporal; Cómo se desarrolla a lo largo de un mes, que diferencias pueden presentar distintas series temporales y cuales pueden presentar similitudes.

A los tópicos mostrados en la tabla 4.1 se le realizó un estudio de las series temporales para ver el comportamiento de cada uno a lo largo de un mes. Es de esperar que, por ejemplo, los tweets y retweets que mencionen las elecciones del 11/08 surgan puntualmente alrededor de la fecha de la misma, mientras que otros tópicos más generales como 'Apoyo a JxC' y 'Apoyo a FdT', etc. sean mencionados a lo largo de todo el mes.

Para construir las series temporales, se volvió a utilizar la definición de los vectores \vec{P}_i para el tweet o retweet i -ésimo.

En un principio, se ordenan temporalmente los tweets y retweets por su tiempo de publicación t_i asociados, construido como se mencionó en la subsección 3.2.

La serie temporal para el tópico j se lo define como

$$S_j(T_k) = \sum_{\{t_i \in [T_k; T_k + \Delta T]\}} (\vec{P}_i)_j, \quad T_k = k * \Delta T. \quad (4.5)$$

La resolución de la escala puede variar modificando el parámetro ΔT . Para las series temporales se tomaron dos tiempos característicos distintos. En principio para la evolución total de un mes se tomó $\Delta T = 6h$. Se hizo un proceso de suavización de la serie temporal para eliminar los efectos del ritmo circadiano.

En la figura 4.9 se muestra la evolución temporal de los 10 tópicos con mayor presencia de la agenda social del mes de agosto.

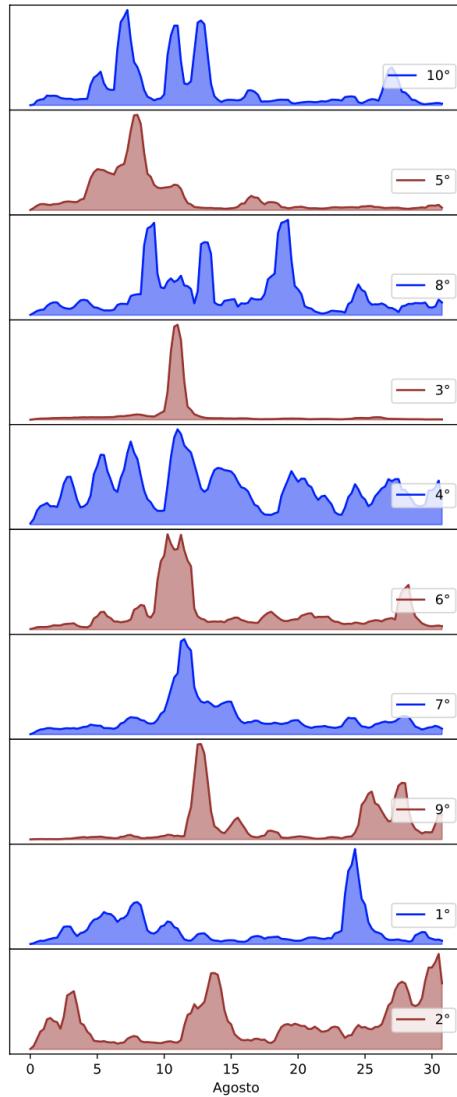


Figura 4.9: Evolucion temporal de los 10 tópicos más mencionados en agosto, en contexto de política argentina. Las series temporales se ordenan segun la fecha de su máximo de tweets y retweets. Cada serie está normalizada a partir de su máximo.

Como se puede ver, los tópicos de las PASO, de Política argentina y de Venezuela, tienen momentos mayor relevancia que en otros momentos del mes de agosto. Es de esperar que esto ocurra ya que las PASO es un evento muy puntual del mes y el conflicto de Venezuela puede manifestarse en mas tweets dentro de una ventana temporal alrededor de las noticias y acontecimientos. El conflicto entre EEUU y Venezuela comenzó a aparecer en los medios de comunicación por la primera mitad de agosto del 2019.

Por otro lado, los tópicos que hablan de fútbol, los artículos de noticias y aquellos tweets y retweets que expresan rechazo a Mauricio Macri tienen una presencia menos localizada en el tiempo, si no que se encuentran presentes en todo el mes.

A continuación, se detallará el estudio de las series temporales asociadas a los tweets y retweets que presentan apoyo a JxC y a FdT, los dos partidos políticos con mayor presencia en las elecciones presidenciales.

4.3.1. Series temporales de JxC y FdT

En esta subsección se buscó comparar la serie temporal del tópico en apoyo a JxC con la serie temporal en apoyo a FdT. Estas son las series temporales asociadas a los partidos políticos más importantes del momento. Resulta interesante que comportamientos tiene cada uno y relacionarlos. Que similitudes y diferencias presenta cada uno.

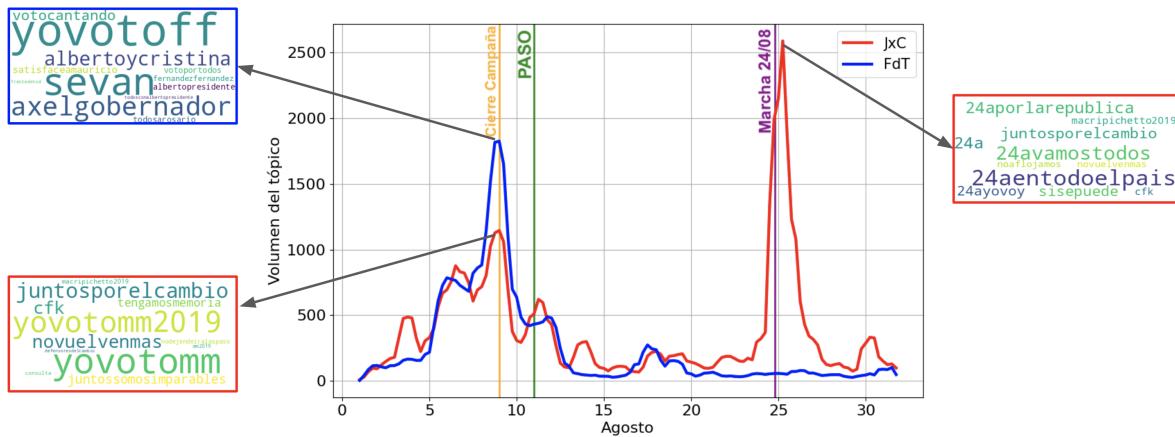


Figura 4.10: Gráfico de la serie temporal del tópico en apoyo a JxC y de la de apoyo a FdT. Las líneas verticales marcan distintas fechas de interés. En verde se marca el cierre de campaña el 09/08, en amarillo las elecciones del 11/08 y en violeta la marcha del 24/08. En los costados se presentan WordClouds de cada tópico de los hashtags ocurriendo en esos momentos.

En la figura 4.10 se muestra la comparación entre las dos series temporales de los tópicos que muestran apoyo a los dos partidos políticos principales de las elecciones del 2019. Se puede ver claramente que ambas están altamente presentes en las fechas alrededor del cierre de la campaña electoral, marcada por la línea vertical amarilla, y de las elecciones PASO, marcada por la línea vertical verde. Estos dos eventos externos son los que amplifican la presencia de ambos tópicos. Más aún, existe un tercer pico en el que coinciden ambas series temporales, alrededor del 6 de agosto. Es interesante notar que estos dos picos pueden tener orígenes distintos para cada serie temporal. Si uno investiga alrededor de la fecha, aparecen 2 acontecimientos relevantes. El primero hace referencia a un discurso de Macri en cierre de campaña, con la famosa frase 'No se inunda más', y el otro hace referencia al rechazo de Alberto Fernández ante el uso de Smartmatic. Esta es una empresa especializada en la automatización de elecciones con el voto electrónico y fue contratada por el gobierno para el escrutinio de las elecciones PASO. Se puede ver entonces que, aunque las series temporales tengan comportamiento similar, algunos picos de interés son por fenómenos que los relacionan como el cierre de campaña o las PASO y otros son por motivos diferentes.

Por otro lado, uno puede encontrar diferencias entre las dos series temporales. La más notoria se presenta en la segunda mitad de agosto particularmente alrededor de las fechas del 24 y 25. La presencia del apoyo a JxC es la mayor del mes, mientras que para la serie temporal de FdT no hay una presencia importante. El máximo de la serie de JxC se debe a la marcha del 24 de agosto en apoyo a Mauricio Macri.

Otro estudio de interés que se realizó en las series temporales fue registrar los hashtags

más utilizados en una ventana temporal reducida de la serie. Para la serie temporal de JxC, se estudiaron dos fechas de interés, el cierre de campaña y la marcha del 24/08. Para la serie de FdT se decidió hacer el mismo análisis, únicamente para el cierre de campaña, donde muestra la línea amarilla. La ventana temporal se centró en los máximos de la cada serie temporal y se le determinó un ancho de 3 días. En esta ventana temporal se registraron todas las ocurrencias de los hashtags de cada uno de los tópicos y se visualizaron en los WordClouds que se muestran en la figura 4.10. Aquellos WordClouds recuadrados en rojo pertenecen a la serie temporal de JxC y el que aparece recuadrado en azul pertenece a la de FdT. Dados estos WordClouds interesante notar dos cosas.

La primera observación es que cerca de las fechas de cierre de campaña, los hashtags que mayor ocurrencia tiene en ambas series temporales hacen referencia principalmente al apoyo de cada partido político en las elecciones. La segunda muestra algo aún más particular. Es interesante ver como, dependiendo de la fecha que uno observa para la serie temporal de JxC, la temática de los hashtags que aparecen en el WordCloud cambia completamente. En el cierre de campaña los mayores hashtags que aparecen son efectivamente sobre las elecciones, mientras que alrededor de la fecha del 24/08 los hashtags hacen referencia a la marcha. Se puede ver entonces que este método de detección de tópicos puede reconocer hashtags que engloban una misma idea general, como el apoyo a JxC, pero que a lo largo del mes se manifiestan en distintas formas. Una de ellas, por ejemplo, es el apoyo en las elecciones, y otra es en la marcha del 24/08.

En este capítulo se llevó a cabo una detección de tópicos subyacentes en los tweets mediante una red de coocurrencia de hashtags. A esta red se le aplicó un algoritmo de detección de comunas resultando en grupos de hashtags que, bajo la hipótesis del trabajo, resultan en los hashtags representativos de cada tópico de interén en la agenda social. Después de la detección de estos tópicos, se hizo una identificación de los mismos basándose en que hashtags de cada tópico son los que más ocurren en los tweets y retweets estudiados. Valiéndose que el hashtag es una compacta unidad de información temática, fue sencillo hacer esta identificación, visualizada en un WordCloud como muestran las figuras 4.5.

Una vez hecha esta identificación de los tópicos, se ordenaron los tópicos privilegiando aquellos que tengan un mayor volumen de tweets y retweets en la base de datos. Para poder realizar este ordenamiento de tópicos, se definió una noción de volumen del tópico en un tweet o retweet, a partir de la expresión (4.3). Con esta noción definida, se pudo calcular el volumen total de un tópico en la expresión (4.4).

En la ultima sección del capítulo se tomaron los tópicos más relevantes del mes y se estudió la evolución temporal de cada uno. En este estudio, se encontró que algunos tópicos aparecen en la agenda social a lo largo de todo el mes, mientras que otros ocurren mayoritariamente en un intervalo acotado, de unos pocos días o hasta horas. Por último, se pudo ver que dentro de un mismo tópico, pueden ocurrir en distintas etapas distintas manifestaciones del mismo. Ejemplificado en el tópico de apoyo a Mauricio Macri, al principio del mes el volumen que ocupaba este tópico era a través de hashtags que hacían referencia a las elecciones. Por otro lado, en fechas cercanas a la marcha del 24/08 en apoyo a Macri, surgieron otros hashtags notorios que hacían referencia específicamente a la marcha.

Caracterización de tópicos mediante una red de similaridad semántica de textos

En el capítulo anterior se desarrolló una técnica de detección de tópicos a partir de modelar las co-ocurrencias de los hashtags con una red compleja. Con esta red, se detectaron grupos de hashtags que representan los distintos tópicos.

En este capítulo se desarrollará una segunda metodología de detección de tópicos que parte del análisis de los textos de los tweets, en vez de los hashtags utilizados. La idea que fundamenta esta metodología es que aquellos tweets cuyos textos tengan semánticas similares deben estar hablando de un tema en común. El objetivo de este capítulo es desarrollar una metodología que encuentre topics que emergen a partir de grupos de tweets cuyos textos sean semanticamente similares.

Para la metodología propuesta en este capítulo, es necesario hacer una representación vectorial de la semántica de los textos de los tweets. Particularmente, en este trabajo, se buscó hacer esta representación con modelos de redes neuronales de tipo transformer encoder. Con estos vectores construidos, se puede cuantificar la similitud de la semántica de los textos asociados, a partir de una métrica de similitud a definir. Estas relaciones de similitud semántica entre los textos es codificada en una red compleja de similitud. En esta red, cada nodo es un tweet y los enlaces codifican la similitud semántica de los textos asociados al par de tweets.

Con la red construida, uno puede realizar una detección de comunidades, encontrando grupos de tweets cuyos textos tengan una alta similitud en su semántica. Nuevamente, la hipótesis de esta metodología es que estos grupos de tweets detectados hacen referencia a distintos tópicos de discusión subyacentes.

Una vez detectados los tópicos, se utilizan distintas técnicas de NLP para la identificación de los mismos. Posteriormente, se realiza un análisis de series temporales y finalmente se realiza una evaluación de cuán efectivo es este método de detección e identificación de los tópicos.

5.1. Representación vectorial de los textos

Para construir una red de similitud a partir de tweets, es fundamental partir de la representación vectorial de los textos. Esta representación permite calcular de manera cuantitativa la similitud entre los tweets, lo que constituye la base de la red. En este caso, se emplea **sentence-transformer**, una adaptación de la arquitectura BERT [18]. Este está diseñado específicamente para extraer representaciones semánticas relevantes de textos, lo

que facilita las comparaciones mediante la similitud coseno. Particularmente, el modelo de `sentence_similarity_spanish_es` permite estudiar textos en español a nivel semántico [35]. La representación vectorial generada por el `sentence-transformer` proyecta cada tweet E_i en un vector \vec{V}_i que lo representa dentro de un espacio de 768 dimensiones. Esta proyección codifica el significado semántico del texto en forma de un vector numérico. Este paso es crucial para poder medir la similitud entre tweets.

El primer paso de la construcción de la red fue representar los 35382 tweets de los tweets en sus vectores asociados, mediante `sentence-transformer`. Como la comparación entre vectores suele definirse mediante la similitud coseno, se consideró que la norma de los vectores no sería relevante para el estudio. Con esto en consideración, se normalizaron todos los vectores para estandarizar el conjunto de vectores.

Una vez que los tweets están representados como vectores, el siguiente paso es calcular la similitud entre ellos. Esto se logra mediante la comparación de los vectores utilizando una métrica de similitud, que se desarrollará con mayor profundidad en la siguiente subsección.

5.2. Construcción de la red de similitud

En esta sección del capítulo, se desarrollarán los pasos tomados para construir una red de similitud entre los textos de los tweets. Este instrumento codifica la similitud semántica de los textos de los tweets, y es fundamental para la detección de aquellos que sean los más similares. Con los textos ya vectorizados en la sección 5.1, se propone una métrica de similitud para comparar pares de vectores de los tweets. Con esta métrica definida, se calcula la similitud de todos los pares de textos posibles y se construye una matriz de adyacencia. Esta matriz, que posteriormente representará una red compleja altamente densa de enlaces, es pasada por un proceso de eliminación de enlaces para reducir el tamaño de la misma, facilitando su estudio a futuro. Una vez eliminados los enlaces no deseados, se consigue la red de similiaridad para estudiar con métodos de detección de comunas.

5.2.1. Métrica de similitud

Una vez que se obtiene el conjunto de vectores representando a los tweets, se busca la métrica de similitud a emplear para la construcción de la red. En la red, cada peso de los enlaces entre tweets se definirá con la similitud entre los pares de tweets. Para ello, es cómodo utilizar una métrica que sea definida positiva. Como se mencionó en la subsección 5.1, la representación vectorial que realiza `sentence-transformers` es construida para hacer comparaciones de tipo coseno. Considerando esto, se decidió hacer una métrica definida positiva, utilizando la similitud coseno. Para ello, se definió una metrica de similitud definida entre 0 y 1 entre un par de vectores \vec{V}_i y \vec{V}_j , que construyen un ángulo θ_{ij} . La métrica se define como

$$\text{simil}(\vec{V}_i, \vec{V}_j) = \frac{1 + \cos(\theta_{ij})}{2}. \quad (5.1)$$

Con esta métrica normalizada entre 0 y 1, se puede cuantificar una similitud normalizada entre pares de vectores.

La definición de 5.1 representa también una similitud semántica entre los textos asociados a los vectores estudiados. La idea del estudio reside en que los pares de textos que hablen de

ideas similares, resulten en valores cercanos a 1 en la definición propuesta de similitud entre vectores.

5.2.2. Construcción de la red y selección de enlaces significativos

a. Construcción de la red

Una vez definida la métrica de similitud adecuada, se construye una matriz de adyacencia \mathbf{A} donde cada elemento

$$A_{ij} = \text{simil}(\vec{V}_i, \vec{V}_j) = A_{ji} \quad (5.2)$$

y en la diagonal $A_{ii} = 1$.

Esta matriz representa una red compleja completa de 35382×35382 . Esto significa que cada nodo comparte enlaces con todos los otros nodos. Cada nodo representa a un tweet y el peso de un enlace entre dos tweets es el valor del elemento A_{ij} , siempre y cuando $i \neq j$.

El hecho de que la red sea completa hace que el estudio de la misma sea computacionalmente costoso. Se buscó reducir el nivel de conectividad de la red manteniendo la información relevante de la similitud entre tweets, comparando los pesos de los enlaces con un *threshold* establecido. Es necesario preguntarse qué *threshold* hay que aplicar y por qué ese valor y no otro.

b. Selección de enlaces

Como la idea es encontrar grupos de tweets que hablen de un mismo tema, los enlaces entre nodos que hay que priorizar son los que mayor peso tiene. De esta manera, los enlaces que contiene la red de similitud son aquellos que relacionan fuertemente los tweets. Más aún, la red resultante de esta selección de enlaces pasa a ser menos costosa computacionalmente de estudiar con un algoritmo de detección de comunidades.

Por otro lado, hay que tener cuidado con la eliminación exigente de enlaces ya que puede tener un impacto en su estructura perdiendo la alta conectividad de la red. Mantener una conectividad adecuada evita la fragmentación de la red, lo que podría llevar a la agrupación de textos similares pero sin un tema común más allá de su contenido literal.

El objetivo no es encontrar oraciones superficialmente similares, si no identificar patrones y grupos de textos que hablan de un tema en común.

Con esto en consideración, se realizó un estudio de la conectividad C de la red en función al umbral u para eliminar los enlaces. Variando el valor del umbral sobre la red, se evaluó la cantidad de nodos que permanecían en el grupo conexo principal en relación a la cantidad de nodos de la red original. Con esto se definió la variable

$$C(u) = \frac{\# \text{ de nodos en la componente conexa principal dado un } u}{\# \text{ de nodos en la red original}} \quad (5.3)$$

Como en la red original todos los nodos permanecían conectados, $C(0) = 1$.

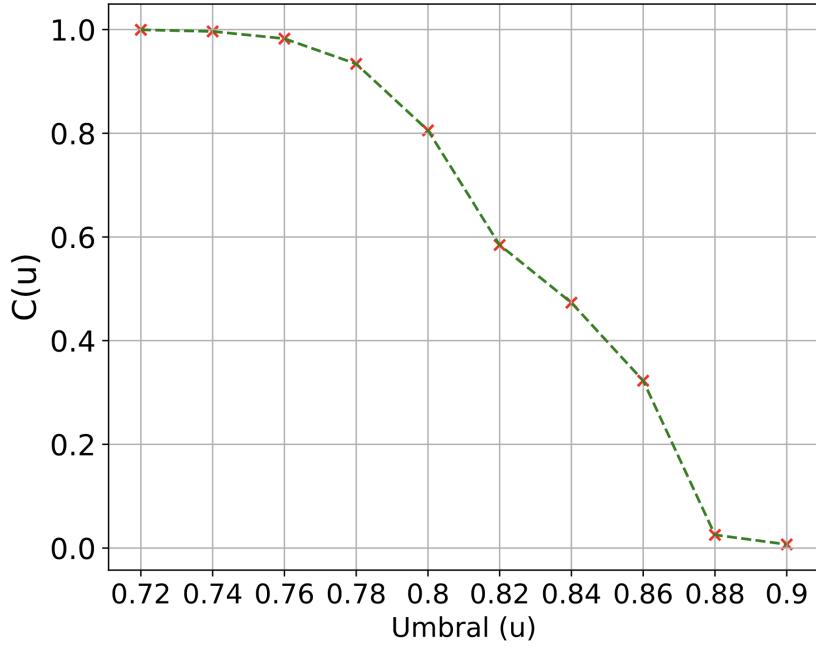


Figura 5.1: Se grafica el porcentaje del volumen de la componente conexa principal C en relación a la red original, en función al umbral aplicado u para la eliminación de enlaces.

Según el gráfico de la figura 5.1, la estructura conexa de la red empieza a verse afectada a partir de eliminar enlaces con pesos similares a 0,72. Esto se observa en que el volumen de la componente conexa principal no coincide con el de la red original, es decir que existe por lo menos otro grupo de nodos que se desconectó de la red.

Con esto en consideración, se eliminaron todos los enlaces cuyo peso sea menor que 0,72. Con este umbral, la componente conexa principal pasó a tener 35366 nodos. Más aún, la red sigue siendo una única componente conexa, sin fragmentarse en subconjuntos pequeños.

Una vez eliminados los enlaces de la red, se le aplicó un algoritmo de detección de comunas para conseguir los grupos de textos que representarían a los tópicos a la red entera resultante.

5.3. Detección de tópicos en la red de similitud

Hasta ahora se tiene una red apta para ser analizada con un algoritmo de detección de comunas. La red contiene a los tweets como nodos y el peso de los enlaces entre nodos se calcula a partir de la similitud (5.1) que tienen las representaciones vectoriales asociadas a los tweets de los nodos. Una vez que se eliminan los enlaces cuyo peso es menor al del umbral, se continua aplicando un algoritmo de detección de comunas a la red. Esta detección de comunas agruparía tweets que, bajo la hipótesis del trabajo, representa a los tópicos subyacentes dentro de la base de datos. Esta detección de comunas se hace nuevamente con Louvain. Consiguiendo una partición de Y comunidades, se la estudia por la cantidad de tweets E_j que contiene cada tópico M_i , $i = 1 \dots Y$. En esta red es importante notar que cada nodo es un tweet, a diferencia de la red de co-ocurrencia de hashtags, donde cada nodo es un hashtag. Esto significa que el tamaño de una comunidad ahora sí afecta directamente en el peso que tiene el tópico en la agenda social.

Se puede ver que, como la red de similitud carece de componentes conexas chicas, los tópicos de la agenda surgen únicamente de la detección de comunas hecha por Louvain. Esto hace que la cantidad de tópicos detectados se reduzca considerablemente. En la metodología anterior se detectaron 1471 tópicos, donde 67 fueron detectados por Louvain y los otros se definieron a partir de los grupos conexos pequeños. Para esta metodología se detectan 30 comunidades. La distribución de los tweets en las comunidades se puede visualizar en la figura 5.2.

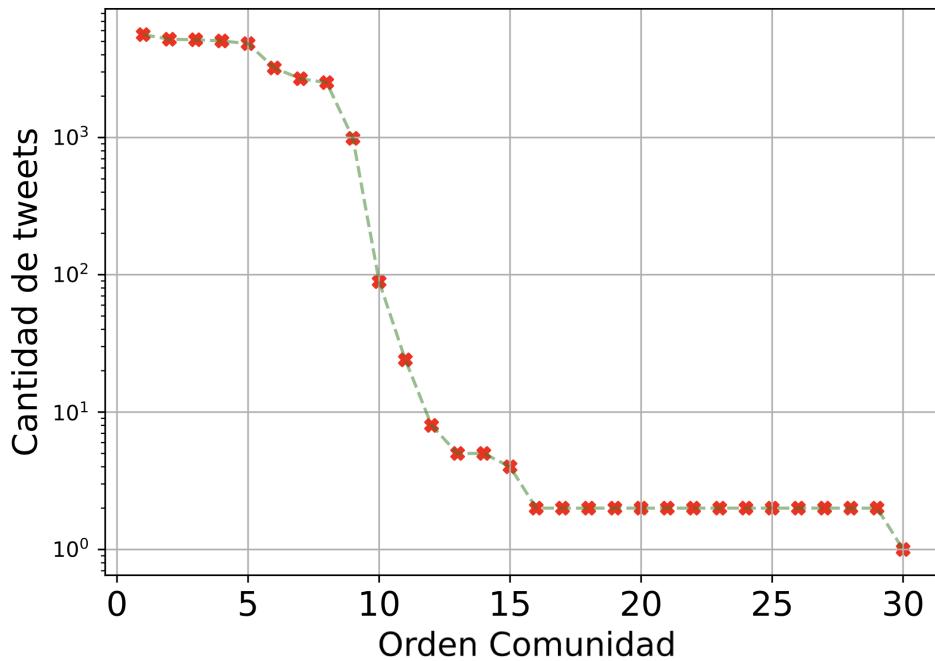


Figura 5.2: Orden de las comunidades detectadas por Louvain de la red de similitud determinado por la cantidad de tweets de contiene cada una.

En la figura 5.2 se puede que los primeros 8 tópicos tienen una cantidad similar de tweets. El tópico número 9 puede considerarse todavía un eje temático importante para estudiar si se compara con los siguientes, que se componen de menos de 100 tweets. Como se mencionó antes, este ranking tiene una influencia mayor en el peso de los tópicos de la agenda, ya que tener más tweets influye directamente en el peso que tiene un tópico en la agenda. Más aún, la diferencia de volúmen de tweets que tienen los primeros 9 tópicos con los siguientes hace considerar que estos últimos son realmente despreciables para el estudio.

Se puede verificar esto observando el peso de cada tópico en la agenda social, a partir de su volúmen de tweets y retweets. Este estudio es más sencillo que en el capítulo 4, ya que el peso de un tópico en esta metodología se calcula sumando todos los tweets asociados al tópico, y los retweets de los mismos.

Con la partición bien caracterizada y los tópicos de interés definidos, el próximo paso es identificar el tópico asociado a cada comunidad.

5.4. Identificación de los tópicos subyacentes

Hasta ahora se detectaron grupos de hashtags que, de acuerdo con las hipótesis propuestas en este capítulo, representan cada uno de ellos un tópico de la agenda social. Aún así, todavía no se le asignó a cada grupo una identidad del tópico al que hacen referencia. Para hacer esto, uno puede extraer las palabras más relevantes de cada grupo de tweets que facilite la identificación de cada tópico. Para esta extracción de palabras se puede hacer con multiples técnicas, desarrolladas en la sección 2.1. Entre ellas, existe el Bag of Words y la co-ocurrencia de términos. A continuación se presenta la aplicación de cada una de estas herramientas para realizar la identificación de los tópicos detectados por Louvain.

5.4.1. Bag of Words

Como se explicó en la sección 2.1, una de las formas adecuadas de estudiar volúmenes de textos es a partir de representaciones vectoriales de tipo VSM. Una de estas es a partir de la técnica de Bag of Words. Esta representación se reduce a contar la cantidad de ocurrencias de cada palabra en un texto o conjunto de textos. De esta manera, se consideran relevantes las palabras que más se usan dentro del conjunto de textos.

Para todos los tweets dentro de la base de datos se les removieron las stopwords. A cada tópico se les agruparon los tweets correspondientes y se realizó un estudio de Bag of Words con el objetivo de ver que palabras son las que más se usan dentro de cada tópico.

A cada tópico se le graficaron las palabras más ocurrentes en un WordCloud.



Figura 5.3: Se presenta cada WordCloud de los 9 tópicos considerados relevantes para la caracterización de la agenda mediática. El orden se determina a partir de la cantidad de tweets y retweets asociados al tópico.

A primera vista se puede ver que la mayoría de los topicos tienen un conjunto de palabras representativas coherentes. Es importante recordar que las palabras que se muestran en estos WordClouds fueron lematizadas para la representación vectorial de los textos. Esto explica las palabras en infinitivo, y también, el pasaje de palabras de femenino a masculino. Este ultimo es muy importante particularmente para aclarar que la palabra 'argentino' en realidad hacer referencia a 'La Argentina'. Principalmente las palabras más importantes en la mayoría de estos tópicos son de política y de las elecciones. Para algunos WordClouds, como los que se muestran en las subfiguras 4°, 7°, 8° e 9°, en [5.3](#) la identificación puede realizarse con suficiente confianza. Por ejemplo, si uno analiza el WordCloud en la subfigura 4°, es probable concluir a partir de las palabras 'voto', 'votar', 'elección', 'domingo', 'candidato', 'paso', etc. que este tópico gira en torno a las PASO.

En el WordCloud de la subfigura 7°, se puede ver que las palabras que lo forman son principalmente figuras políticas del partido de FdT. Aunque también hayan algunas palabras en este WordCloud que muestren una emoción positiva como 'volver' o 'mejor', esto no significa que el tópico en general sea en apoyo al FdT, ya que estas no tienen tanta presencia como 'Fernandez' o 'Kirchner'. Si uno observa los tweets dentro de este tópico, puede encontrarse también con algunos tweets de odio hacia estas figuras políticas.

Por otro lado, en la subfigura 8° se puede ver un WordCloud particular que se construye con palabras como 'ganar', 'victoria', 'paso', 'partido', y después también hay palabras como 'rossi', el jugador de fútbol, 'medalla', 'final', 'equipo', etc. Este tópico parece englobar un mensaje asociado con la victoria, sea tanto en las elecciones como, en este caso, en el mundo de los deportes.

Por último, en el WordCloud 9°) se encuentran palabras asociadas nuevamente con las elecciones PASO, principalmente con 'Domingo' y 'Elección'. Además, aquí se visualiza una asociación con los cierres de campaña dadas las palabras como 'hablar', 'cerrar', 'cierre', 'campana'. Esto se confirma cuando uno observa a los tweets de este tópico. Aún así, entre estos tweets siguen habiendo algunos que no necesariamente hablan de los cierres de campaña, pero si la gran mayoría hace referencia a las elecciones nuevamente.

La mayor desventaja de estos WordClouds que se muestran en [5.3](#), a diferencia de los que se mostraban en [4.5](#), es que aquí se hace una inferencia de los ejes temáticos a partir de palabras individuales, mientras que en la metodología de la red de co-ocurrencia de hashtags, era relativamente sencillo concluir un tópico a partir de los mensajes de los hashtags. Esto era porque cada unidad de información que se analizaba en el capítulo [4](#) tienen compactados mensajes que explican con facilidad una idea. En este caso, al analizar las palabras individuales se corre el riesgo de perder información valiosa, como la connotación detrás de cada palabra.

Otra dificultad que se puede visualizar es que, cuando se comparan los WordClouds, hay muchas palabras que coinciden en los WordClouds. El caso más claro es el de Macri, que aparece como una de las palabras más importantes en los WordClouds de 1°, 2° y 5°. Uno podría guiarse con las otras palabras dentro del WordCloud pero no siempre se puede llegar a una identificación clara.

Para intentar resolver esto, se decide analizar cada tópico mediante el análisis de las palabras que más co-ocurren en cada tópico. De esta manera, se le puede atribuir un significado más profundo a las palabras relevantes. Considerando con qué otras se relacionan, es posible

atribuirle una connotación más clara a una misma palabra.

5.4.2. Co-ocurrencia de palabras

Otra metodología que se usa para la identificación de cada tópico en la agenda es la co-ocurrencia de palabras. Para cada corpus de tweets del mismo tópico se le calculó una matriz de ocurrencias binarias. Considerando un solo corpus de N tweets E_j , $j = 1 \dots N$ con M palabras únicas r_i , se calcula la matriz de ocurrencia binaria como

$$A_{ij} = \begin{cases} 1 & \text{si la palabra } r_i \text{ está en el tweet } E_j \\ 0 & \text{si no.} \end{cases} \quad (5.4)$$

Con esta matriz definida, se calcula una matriz de $M \times M$ como

$$B = A * A^T \quad (5.5)$$

donde el elemento B_{ij} codifica la cantidad de tweets en donde las palabras r_i y r_j co-ocurren.

Para cada tópico se calculó su matriz de co-ocurrencia de palabras asociadas. Cada una de estas se puede interpretar como una matriz de adyacencia, haciendo referencia a la figura 2.3 y se graficaron aquellos enlaces más pesados de la red para visualizar que palabras tienen una relación más grande dentro del corpus.

En la figura 5.4 se visualiza las redes de co-ocurrencias para los tópicos que todavía no fueron identificados en la figura 5.3, es decir las subfiguras 1°, 2°, 3°, 5°, y 6°. En la figura 5.4 se siguen utilizando estos índices para identificar cada tópico.

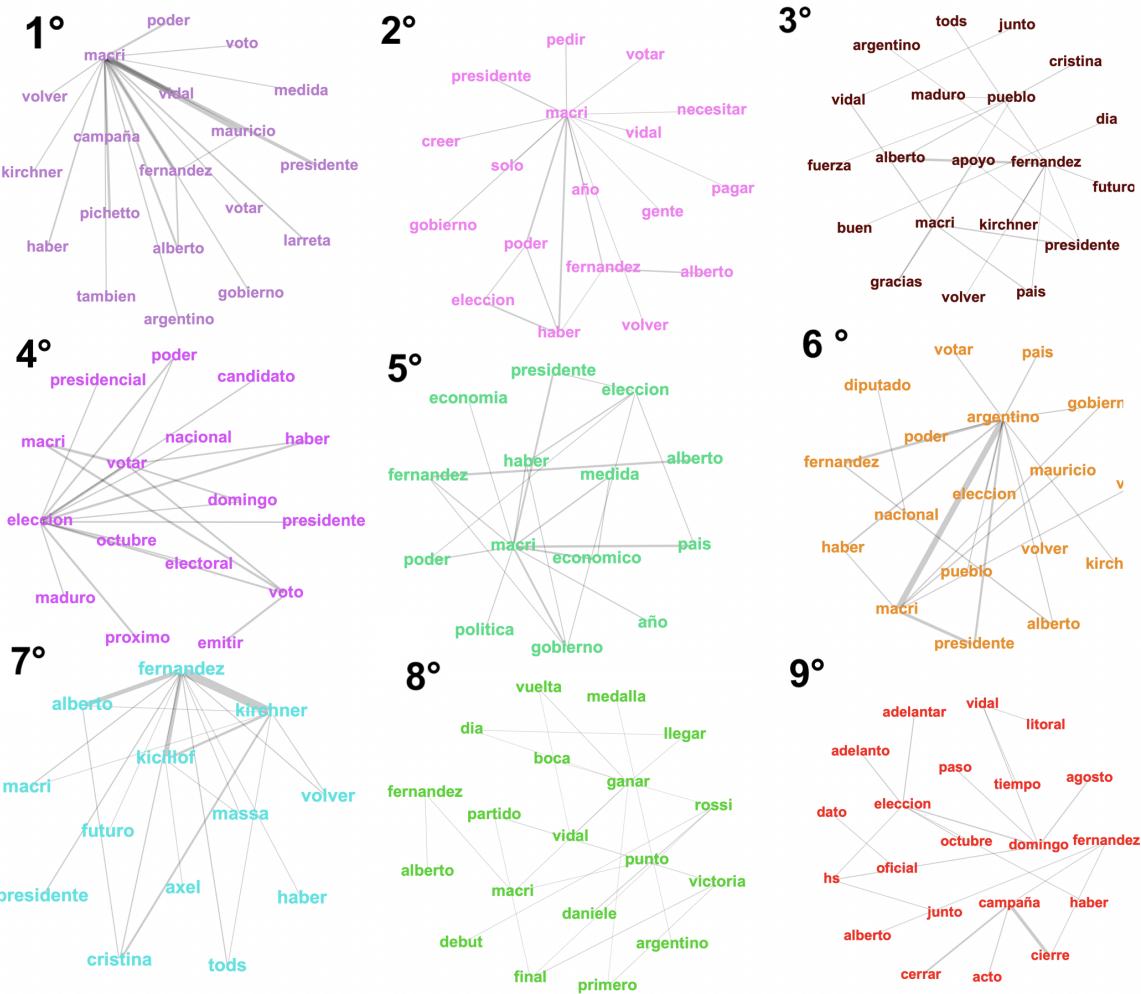


Figura 5.4: Se presentan las redes de co-ocurrencia de los 9 ejes temáticos detectados. Cada red de co-ocurrencia muestra los 20 enlaces más pesados, es decir, los 20 pares de palabras que más co-ocurren en todo el corpus.

Las redes de la figura 5.4 presenta una nueva manera de evaluar cada tópico. En cada red, se puede reinterpretar el rol de cada palabra dentro del tópico, considerando las otras palabras con las que haya co-ocurrido frecuentemente. Hay que remarcar que el hecho de que no haya un enlace entre palabras no significa que no hayan co-ocurrido nunca en el corpus, si no que no es de los enlaces más pesados de la red.

Se puede ver que para los tópicos ya identificados por la técnica de Bag of Words, siguen manteniendo una coherencia muy fuerte con la etiqueta de su tópico. Para el tópico de 'Elecciones' en la subfigura 4°, las palabras más relacionadas siguen siendo las del 'voto', 'elecciones', 'domingo', etc. Lo que no necesariamente agrega información nueva a la identificación de este tópico, pero sí refuerza la idea de que este efectivamente hace referencia a las elecciones PASO. Estas coherencias siguen dandos en las otras subfiguras 7°, 8° y 9°, de los tópicos de FdT, 'Victoria' y de 'Cierre de campaña'.

Por ejemplo, en la red del tópico 1º muestra como Macri es una palabra central en el tópico, esta se conecta fuertemente con todas las otras palabras de la red tales como 'Fernández', 'Voto', 'Votar', 'campaña', 'volver', que son todas palabras relacionadas con el contexto de las

elecciones PASO. Más aún, estas otras palabras no muestran una interacción fuerte entre ellas. Esto puede hacer entender que el tópico está altamente identificado por Macri, particularmente con su relación con la campaña electoral.

La red que se muestra en la subfigura 5° nuevamente muestra a Macri como un sujeto central del tópico, donde es muy común que esta palabra co-ocurra con las otras de la red. Si uno observa estas otras palabras, tales como 'medida', 'economía', 'política', 'país', etc., puede inferir que hablan del mandato presidencial de Macri, y su relación tanto en lo económico, político, elecciones, etc.

Por otro lado, para la red de la subfigura 6°, la palabra central en el tópico parece ser 'Argentino'. Hay que recordar que por producto de la lematización y por el uso de letras en minúscula, 'la argentina' es transformada en 'el argentino'. Con esto en mente, se puede decir con cierta confianza que este tópico tiene como eje principal, 'La Argentina' y todos los aspectos a nivel país.

Hay 2 tópicos que todavía no se pudieron identificar adecuadamente con una etiqueta distintiva. Estas son la 2° y 3°. La 2° presenta la dificultad de que nuevamente surgen palabras individuales muy genéricas para realizar una identificación clara del tópico. Más aún, estas coinciden en su mayoría con las del 1° y 5°. Esto puede deberse a que este tópico es el segundo más grande de la agenda política, y también tiene una cantidad de tweets considerablemente alta, lo cual puede hacer difícil identificar al tópico con un conjunto limitado de palabras. Acá ya se encuentra una limitación fuerte de esta metodología. La identificación de los tópicos requiere de múltiples herramientas de análisis, y a veces que ellas no son suficientes para extraer una etiqueta clara del tópico. Por conveniencia, se puede seguir asociando el tópico 2° a Macri, pero en aspectos generales.

El 3° tópico muestra una particularidad. Si se estudia su Wordcloud ([5.3](#) 3°) y su red de co-ocurrencias, se puede ver que este tópico incluye muchas palabras de apoyo y los nombres de las distintas figuras políticas, tanto Macri como Alberto y Cristina Fernández. Este apoyo no necesariamente es hacia uno de los dos partidos, si no que, hasta ahora, se puede decir que este tópico se puede englobar como la 'Militancia' o 'Apoyo' a los partidos políticos.

Para indagar un poco más sobre esto, se le aplicó una herramienta particular al tópico 3° para verificar que efectivamente este tópico se identifica como tal.

5.4.3. Análisis de sentimiento

El análisis de sentimiento es una herramienta para estudiar si una oración está cargada de emociones positivas o negativas. Este estudio también es parte del NLP, y puede realizarse con modelos de machine learning. Esta herramienta se usa en la tesis para entender el contenido emocional del tópico 3°. A continuación se desarrollarán los pasos que fueron utilizados.

El análisis de sentimiento hecho en esta tesis es con una librería en Python llamada `pysentimiento` [[15](#), [16](#)]. Esta librería no solo estudia el sentimiento positivo, negativo o neutro de una oración si no también puede registrar el sentimiento hacia uno o múltiples sujetos especificados dentro de una misma oración.

Para cada oración y sujeto definido, `pysentimiento` devuelve un vector de 3 componentes. Estas componentes son las probabilidades de que el sentimiento de la oración hacia el sujeto sea positivo, negativo o neutro. Estas componentes suman un total de 1. Al par texto y sujeto

se le asigna una etiqueta de la emoción resultante con mayor probabilidad.

La propuesta para estudiar el contenido emocional del tópico 3° es registrar todos los tweets del mismo que mencionen a distintas figuras políticas, en principio Macri y Fernández y ver cuantos hablan negativamente o positivamente de cada uno.

Para facilitar el estudio, se decidió crear un puntaje de -1 a 1, donde los numeros negativos son las probabilidades de las emociones negativas con un signo invertido, y los valores positivos se refieren a las probabilidades de emociones positivas.

En el estudio se consideraron únicamente los resultados donde la emoción con mayor probabilidad sea positiva o negativa. Y se registro por par de tweet y sujeto, la emoción resultante y su probabilidad asociada a la misma. Para cada sujeto, se vio la distribución de los puntajes a lo largo de los valores de -1 y 1. Esto indica cual es la distribución de emociones negativas y positivas para cada figura política dentro de un tópico.

En la figura 5.5 se ve la distribución de emociones que se manifiestan hacia Mauricio Macri y Alberto Fernandez en el tópico 3°.

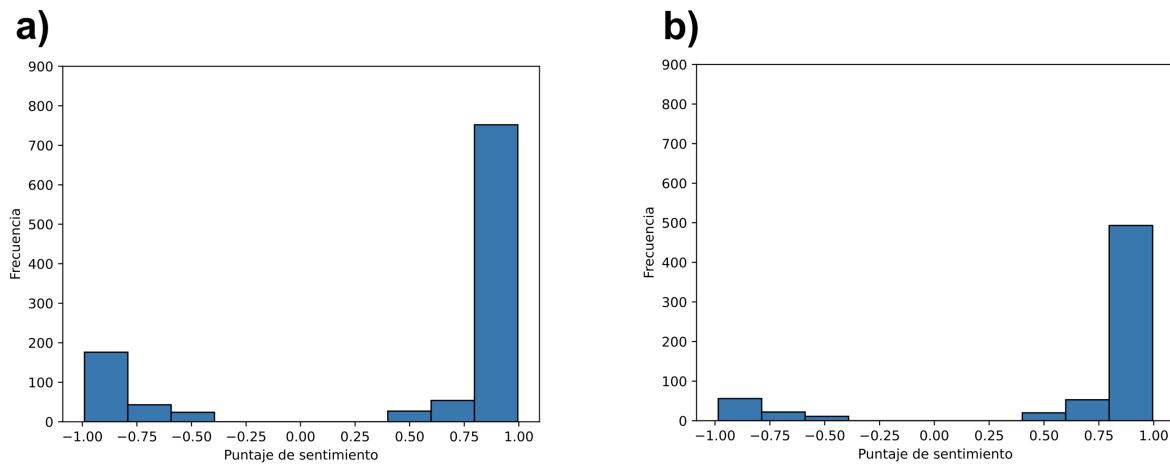


Figura 5.5: En la figura a) se muestra la distribución de puntajes de emoción de los tweets del tópico 3° hacia Macri, mientras que la figura b) muestra la distribución de los puntajes hacia Fernández.

De la figura 5.5 se puede ver que la gran mayoría de los tweet que mencionan a cada uno de las figuras políticas muestran una tendencia a ser de emociones positivas. Tanto para la distribución de puntajes de sentimiento de Macri en la subfigura a), como para Fernandez en b), el resultado global resulta en tweets positivos hacia dichas figuras políticas. Esto se puede deber a la semántica positiva que comparten estos tweets, más allá de quién estén hablando.

Las palabras observadas en el BoW y la red de co-ocurrencias muestra la existencia de estas palabras de apoyo, y esta última figura da a entender que el tópico entero se puede considerar como el apoyo a las distintas figuras políticas.

Con la mayoría de comunidades ya detectadas, se presentan de los tópicos detectados de la agenda caracterizada por la red de similitud en la tabla 5.1.

Orden	Tópico	Volumen
1°	Macri - Elecciones	60495
2°	Macri - General	58810
3°	Apoyo a Partidos	57919
4°	Elecciones	54226
5°	Macri - Presidencia	47625
6°	Argentina	37775
7°	Frente de Todos	21920
8°	"Victoria"	21658
9°	Cierre de campaña	8007

Cuadro 5.1: Tópicos rankeados por su volumen de tweets y retweets.

Se puede ver que las técnicas de identificación dan lugar a tópicos que comparten muchos conceptos, entre ellos a las figuras políticas y las elecciones. Muchos de los tópicos tienen como palabra central a Macri, y estos pueden diferir en otras palabras que surgen del análisis de las BoW como en las co-ocurrencias. Esta similitud en los tópicos puede deberse a que de por sí la gran mayoría de los tweets tratan de un tópico global, la política argentina. Se puede ver que todos los tópicos fueron detectados como temas de política y, más aún, no parecería haber un tópico distintivo de fútbol o de relaciones internacionales, como se llegó a detectar en el capítulo 4.

La visualización de la red de similitud en la figura 5.6 puede explicar en palabras el resultado obtenido en la detección e identificación de los tópicos.



images/BERT/red_similitud.png

Figura 5.6: Distribución de los nodos más relevantes de cada comunidad en la red de similitud. El color de cada nodo se asocia a su comunidad, es decir, a su tópico.

En la red de la figura 5.6 se ve claramente la alta interconectividad de las comunidades. A diferencia de la red mostrada en 2.2.1, esta no tiene una estructura clara que disocia a algunos grupos de nodos (o tweets) de otros, si no que se ve como un único grupo de nodos conectados. Esta estructura de la red puede reflejar los tópicos detectados, que comparten una semántica similar, como muestra la red.

No quita que de cada uno de estos tópicos se puede extraer información relevante. Otra etapa de evaluaciónn de los tópicos sigue siendo el análisis de la serie temporal. Entender el comportamiento de cada uno de los tópicos a lo largo del tiempo puede dar a luz nuevas similitudes o diferencias entre los tópicos detectados.

5.5. Series temporales

Una vez con cada tópico identificado, se estudia las series temporales de cada una, de una manera análoga a la metodología del capítulo 4.

Para cada tópico, se ordenaron los tweets E_j por su tiempo de publicación t_j , y se definió la serie temporal como

$$S_j(T_k) = \sum_{\{t_i \in [T_k; T_k + \Delta T] \wedge E_i \in M_j\}} 1 \quad (5.6)$$

Es decir que ahora, lo que diferencia esta expresión a la presentada en la ecuación (4.5) es que un tweet o retweet no se partitiona en distintos tópicos, si no que se suma una unidad entera por tweet y retweet a la serie temporal.

Con esto definido, se calcularon las series temporales de cada tópico. Estas se muestran en la figura 5.7

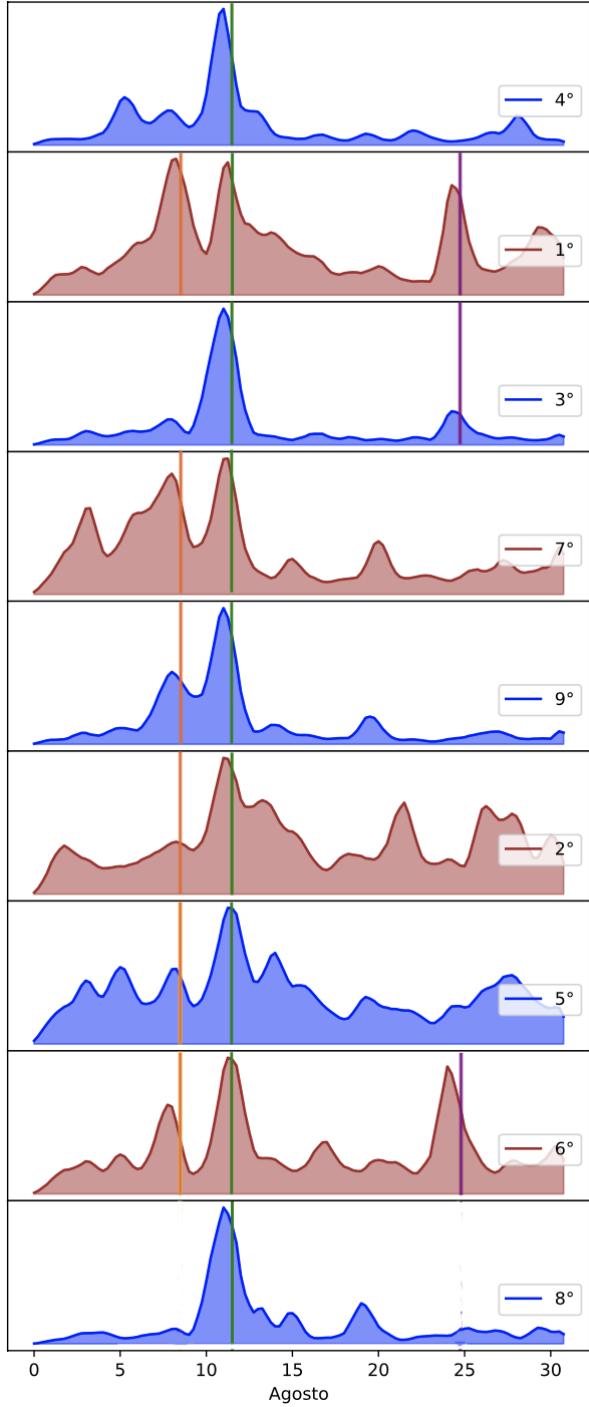


Figura 5.7: Se muestran las 9 series temporales de los tópicos detectados de la agenda social. Se ordenan por la fecha que llegan a su máximo de tweets y retweets. Para las distintas series, se encuentran marcadas 3 de las fechas de interés, el cierre de campaña de los partidos políticos, 9 de agosto, marcado en naranja. La siguiente fecha es la de las elecciones PASO, el 11 de agosto, marcado en verde. Por ultimo, se marca en violeta la fecha del 24 de agosto por la tarde, relacionada con la marcha a la casa Rosada en apoyo a Macri.

Se puede ver que, a diferencia de la figura 4.9, las series temporales de la figura 5.7 tienen

todas el máximo de su serie temporal alrededor del 11 de agosto, en las elecciones PASO. Esto se podría explicar desde el punto de vista que todos los tópicos detectados son relacionados con la política, es de esperar que en la fecha más relevante del mes se encuentre el momento de mayor atención de cada tópico. Aún así, cada serie temporal presenta distintos tópicos aparte del del 11/08. Una pregunta válida de hacerse es si se puede realizar un estudio análogo al de la figura 4.10, donde se observan los subtópicos de cada momento de la serie temporal. Para esto, en los picos de interés de cada serie temporal se agruparon los tweets y retweets publicados en esa ventana temporal y se observaron que pares de palabras consecutivas, o bigramas, suelen aparecer con más frecuencia en todas los tweets y retweets. De esta manera, se puede evaluar si en dos momentos distintos de un mismo tópico, se habla de temas diferentes.

Uno de los ejemplos se muestra en la serie temporal del tópico de Macri-Elecciones. Esta muestra una presencia muy alta en los 3 picos de interés. El objetivo es ver si se puede responder que acontecimientos particulares son de interés en cada uno de estos picos, y si son coherentes con la etiqueta del tópico definido.

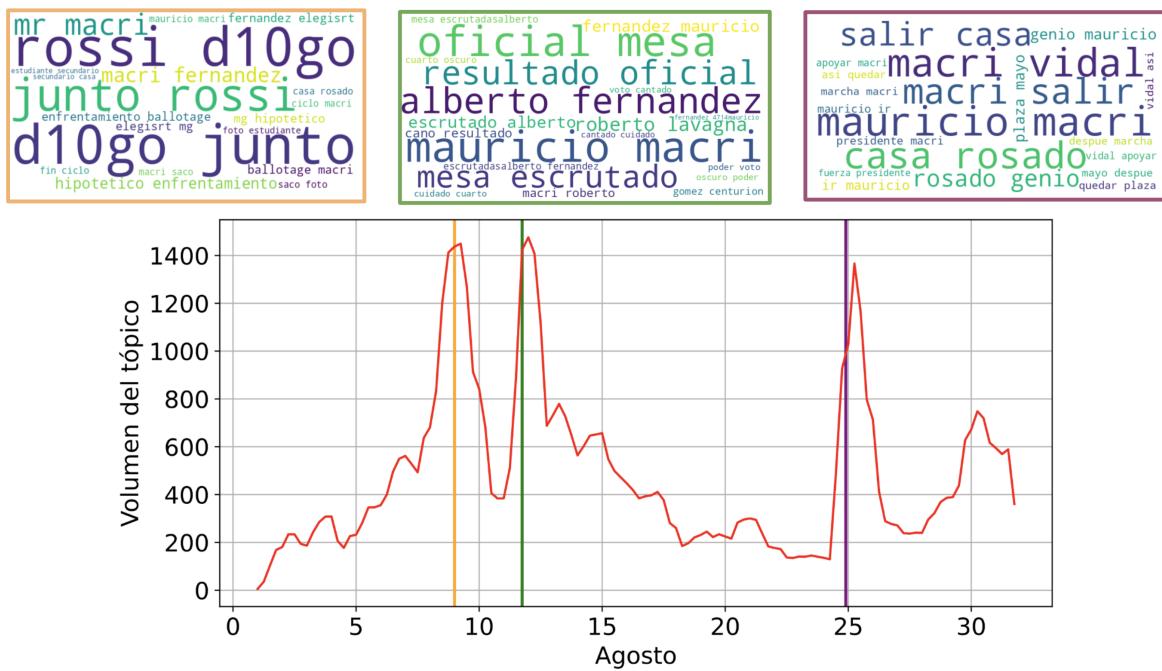


Figura 5.8: Serie temporal asociada al tópico 'Macri-Elecciones'. Se marcan las 3 fechas de interés para estudiar los tweets y retweets de la serie temporal: El cierre de campaña marcado en el 9 de agosto, las elecciones PASO del 11 de agosto, y la marcha en apoyo a Mauricio Macri el 24 de agosto por la tarde. Para cada pico marcado se muestra el WordCloud de bigramas de los tweets y retweets.

De la figura 5.8 se estudia la coherencia dentro de los tweets y retweets del tópico 'Macri-Elecciones' a lo largo del tiempo. De los picos asociados a las paso y a la marcha del 24, se pueden observar palabras del contexto político, que pueden considerarse coherentes con el tópico. Observando los tweets y retweets del 11 de agosto, estos presentan palabras relacionadas a las elecciones como 'mesas escrutado', 'resultado oficial', 'Alberto Fernandez', 'Mauricio Macri'. Para los tweets y retweets del 24 de agosto, se tiene una fuerte presencia de Macri, de

la casa rosada y del apoyo de las personas.

El pico del 9 de agosto muestra que este tópico no está enteramente contruido por tweets de Macri, o si quiera de tweets de política. En este momento de la serie temporal surgen palabras como 'Rossi' con 'D10go' y 'juntos', en referencia a un encuentro entre los dos jugadores de fútbol, Diego Maradona y Daniele De Rossi. Recién los bigramas con menor frecuencia hacen referencia a Macri. No parece haber una relación semántica de estas palabras con las que se detectaron en el estudio de Bag of Words y co-ocurrencias de palabras de la sección 5.4. Uno podría esperar que en esta fecha se haga referencia fuertemente al cierre de campaña de cada partido político, o de Macri, pero no es lo que ocurre.

En este capítulo se buscó caracterizar nuevamente la agenda social dentro de los tweets y retweets de la base de datos de la tesis. Para ello, la idea fue encontrar grupos de textos que compartan una semántica en común, lo que en teoría daría lugar a los distintos temas de la agenda social que hablan las personas en Twitter. Para poder relacionar semánticamente los textos, se utilizó una representación vectorial de **sentence-transformers**, modelo de BERT que funciona para la comparación semántica entre pares de vectores a partir de una similitud coseno. Comparando los 35382 vectores entre ellos mediante una métrica basada en la coseno (ecuación (5.1)), se definió una matriz de adyacencia de 35382×35382 que fue representada en una red compleja completamente conectada, es decir de $35382 * (35382 - 1)$ enlaces.

Para reducir el costo computacional del estudio de esta red, se eliminaron los enlaces menos pesados, en base a un umbral que siga mannteniendo la red totalmente conexa. Con la red modificada, se le aplicó una detección de comunas con Louvain, lo que devolvió 9 grupos relevantes de textos que representarían, en teoría, cada tópico de discusión.

A estos grupos se los buscó idenntificar a partir de tecnicas de VSM como Bag of Words y otras de NLP como co-ocurrencia de términos. Esto permitió encontrar las etiquetas de la mayoría de los tópicos, pudiendo llegar a una agenda social identificada. Estos tópicos compartían palabras relevantes, y no eran completamente independientes una de la otra. Por ejemplo, 4 de los 9 tópicos detectados eran fuertemente identificados por la palabra Macri, y diferían entre ellos por otras palabras que surgían en las BoW y las co-ocurrencias.

A diferencia del capítulo 4, no se detectaron tópicos de entretenimiento como fútbol ni de aspectos internacionales, si no que todos hacían referencia en su mayoría a política argentina.

El estudio de las series temporales remarcó otras características de esta metodología. En principio, todos los tópicos compartían la fecha donde se maximizaba el volumen de tweets y retweets asociados, el 11 de agosto. Esto vuelve a explicar la similitud entre los tópicos, donde todos contienen en su mayoría tweets relacionados con la política argentina. Por otro lado, se pudo realizar un estudio de mayor presición en las series temporales. En ellas se pudo visualizar los temas específicos que se trataban en cada momento, particularmente en fechas de interés como el 9, 11 y 24 de agosto, considerando tanto los tweets como los retweets publicados. En el ejemplo mostrado, se pudo ver que en las fechas del 11 y 24 las palabras más observadas en los tweets y retweets se alineaban fuertemente con los acontecimientos de la fecha, siendo estas las elecciones PASO y la marcha en apoyo a Macri respectivamente. Por otro lado en el día del 9 de agosto, siendo época de cierre de campaña de los distintos partidos políticos, las palabras más utilizadas de los tweets, en un tópico supuestamente de política, eran en particular menciones a temas de fútbol en vez de los cierres de campaña.

Discusiones y conclusiones

En este trabajo se compararon dos metodologías para la detección de ejes temáticos en un conjunto de tweets: una basada en una red de co-ocurrencia de hashtags y otra en una red de similitud semántica de textos. El objetivo fue encontrar similitudes y diferencias, y ventajas y desventajas de cada uno de los métodos para tener en consideración cuál de estas metodologías es más efectiva para capturar los temas subyacentes en las conversaciones de Twitter. A grandes rasgos, se puede ver que la metodología basada en el estudio de la co-ocurrencia de los hashtags resultó en una agenda de tópicos bien definidos, distintivos entre ellos y cada uno con su dinámica temporal característica, como se pudo ver en sus series temporales asociadas. Por otro lado, la metodología asociada a la red de similitud semántica brindó otra agenda social, únicamente compuesta de tópicos de política argentina que compartían muchas propiedades y temáticas entre ellos. Esto se manifestó tanto en la identificación de los tópicos, en la visualización de la red de similitud, como en las series temporales de cada uno de ellos.

La identificación de estos tópicos fue sencilla de realizar apenas con un método de Bag of Words, que dió lugar a los hashtags más importantes por tópico. El hecho de que cada hashtag sea de por sí un elemento útil para definir la temática de un tweet fue ventajoso a la hora de visualizar los WordClouds para cada tópico y poder inferir un eje temático de cada uno. Más aún, con los tópicos bien definidos se pudo ver que la estructura de la red de co-ocurrencias facilitaba en gran parte al algoritmo de detección de comunas a encontrar los tópicos distintivos de la agenda social. Era clara la visualización de aquellos tópicos relacionados con temas de política argentina, en comparación a aquellos que se relacionaban con tópicos de política internacional o entretenimiento. La separación de cada tópico se veía reflejada acordemente en la distribución de los nodos en la red compleja. Por otro lado, la metodología basada en la red de similitud presentó mayores dificultades a la hora de la identificación de cada tópico. Esto no necesariamente se considera como una caracterización errónea de la agenda social, pero sí se puede decir que la segunda metodología presentada no fue capaz de detectar aquellos tópicos específicos que detectó la metodología asociada a la co-ocurrencia de hashtags. Para realizar una identificación adecuada de los tópicos, se necesitó implementar múltiples herramientas tanto como Bag of Words, co-ocurrencia de palabras y hasta análisis de sentimiento para poder caracterizar correctamente los tópicos. Esto fue necesario más que nada porque la identificación de tópicos enteros a través de palabras individuales puede dificultarse por la falta de contexto que tiene cada palabra, a diferencia de los hashtags que en su mayoría no necesitan de un contexto para entenderse. Estas herramientas utilizadas para la identificación de los tópicos del capítulo 5 simplemente se necesitaron para darle un contexto apropiado a

las palabras dentro de cada tópico. La visualización de la red en la figura 5.6 demostró la dificultad de hacer una partición tan detallada como en la red de co-ocurrencias de hashtags. En la red de similitud se visualizaban conjuntos nodos de distintos tópicos altamente conectados, dado por la naturaleza de la construcción de la red. Estos enlaces en primer lugar hacían que la detección de comunidades sea altamente costos en términos de computación, y no resultaba en comunidades distintivos, si no que todos mostraban una similitud fuerte y una asociación con los temas de política argentina.

Las series temporales de cada metodología también puso en evidencia la efectividad de estas en la caracterización de los tópicos. En el capítulo 4, se observó que la serie temporal de cada tópico mostraba un comportamiento que coincidía con las fechas importantes relacionadas con ese tópico, haciendo un contraste con las noticias publicadas a lo largo del mes de agosto de 2019. Aquellos tópicos puntuales mostraban tener un único máximo en la serie temporal, como las PASO o el interés sobre el conflicto de Venezuela con los EEUU, en las fechas adecuadas. Los otros tópicos se veían más distribuidos a lo largo del tiempo, ya que no se relacionaban con un momento puntual del mes. Más aún, se pudo visualizar un resultado muy interesante sobre las series temporales en relación a Juntos por el Cambio y el Frente de Todos. Estas dos series presentaban comportamientos similares en sus series temporales, por motivos compartidos, siendo estos el cierre de campaña y las elecciones PASO. Además, se encontraba una diferencia muy fuerte en la fecha del 24 de agosto, donde el tópico de JxC tenía su máximo, y el tópico de FdT no tomaba mucha presencia. Para la serie temporal de JxC, se pudo ver que este tópico manifestaba a lo largo del tiempo distintas representaciones del apoyo al partido político. Los hashtags predominantes en distintos momentos del mes no eran necesariamente los mismos, pero todos describían eventos diferentes que compartían el mismo eje temático.

Este comportamiento no se visualizó con tanta claridad en las series temporales del capítulo 5. Entre ellas se mostraban similitudes muy marcadas, particularmente el máximo que compartían todas las series en el 11 de agosto. Esto fue explicado justamente por las identidades políticas de cada tópico y como estos necesariamente compartían como fecha relevante las PASO. Al estudiar las ventanas temporales reducidas de las series temporales, se pudo ver que efectivamente esta metodología era capaz de mostrar subtópicos a lo largo del mes, pero, a diferencia de las series temporales del capítulo 4, estos subtópicos no se podían relacionar con una etiqueta tan específica como el caso anterior, si no que simplemente se relacionaban por ser subtópicos de política argentina.

Esta tesis deja inconclusa una comparación cuantitativa de las agendas sociales. Para trabajos futuro se idearon técnicas de similitudes entre las distintas agendas, basadas en representaciones vectoriales de corpus de textos. Por otro lado, sería ideal refinar la metodología presentada en el capítulo 5 y así detectar claramente tópicos con mayor resolución y que comparten menos palabras relevantes entre ellos.

Otra propuesta interesante, basada en trabajos como [6, 8] es unir las dos metodologías para realizar un modelo de machine learning capaz de detectar la orientación política de un tweet. Esto se podría realizar entrenando al modelo con la semántica de los hashtags que tiene cada tweet, y relacionando estos con el texto del mismo. Sería sencillo conocer la semántica de los distintos hashtags sabiendo a qué tópico pertenecen, detectado con la metodología del capítulo 4.

Bibliografía

- [1] Federico Albanese et al. «Analyzing mass media influence using natural language processing and time series analysis». En: *Journal of Physics: Complexity* 1.2 (jul. de 2020), pág. 025005. DOI: [10.1088/2632-072X/ab8784](https://doi.org/10.1088/2632-072X/ab8784). URL: <https://dx.doi.org/10.1088/2632-072X/ab8784>.
- [2] Sebastián Pinto et al. «Quantifying time-dependent Media Agenda and public opinion by topic modeling». En: *Physica A: Statistical Mechanics and its Applications* 524 (2019), págs. 614-624. ISSN: 0378-4371. DOI: <https://doi.org/10.1016/j.physa.2019.04.108>. URL: <https://www.sciencedirect.com/science/article/pii/S0378437119304844>.
- [3] Frank Schweitzer. «Sociophysics». En: *Physics Today* 71 (feb. de 2018), págs. 40-46. DOI: [10.1063/PT.3.3845](https://doi.org/10.1063/PT.3.3845).
- [4] Maxwell Mccombs y Donald Shaw. «The Agenda-Setting function of mass media». En: *The Agenda Setting Journal* 1 (sep. de 2017). DOI: [10.1075/asj.1.2.02mcc](https://doi.org/10.1075/asj.1.2.02mcc).
- [5] Hendrik Schawe et al. «Understanding who talks about what: comparison between the information treatment in traditional media and online discussions». En: *Scientific Reports* 13.1 (7 de mar. de 2023), pág. 3809. DOI: [10.1038/s41598-023-30367-8](https://doi.org/10.1038/s41598-023-30367-8). URL: <https://doi.org/10.1038/s41598-023-30367-8>.
- [6] Zhenkun Zhou et al. «Why polls fail to predict elections». En: *Journal of Big Data* 8.1 (23 de oct. de 2021), pág. 137. DOI: [10.1186/s40537-021-00525-8](https://doi.org/10.1186/s40537-021-00525-8). URL: <https://doi.org/10.1186/s40537-021-00525-8>.
- [7] James Flaminio et al. «Political polarization of news media and influencers on Twitter in the 2016 and 2020 US presidential elections». En: *Nature Human Behaviour* 7.6 (1 de jun. de 2023), págs. 904-916. DOI: [10.1038/s41562-023-01550-8](https://doi.org/10.1038/s41562-023-01550-8). URL: <https://doi.org/10.1038/s41562-023-01550-8>.
- [8] Alexandre Bovet, Flaviano Morone y Hernán A. Makse. «Validation of Twitter opinion trends with national polling aggregates: Hillary Clinton vs Donald Trump». En: *Scientific Reports* 8.1 (6 de jun. de 2018), pág. 8673. DOI: [10.1038/s41598-018-26951-y](https://doi.org/10.1038/s41598-018-26951-y). URL: <https://doi.org/10.1038/s41598-018-26951-y>.
- [9] Gerard Salton. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. USA: Addison-Wesley Longman Publishing Co., Inc., 1989. ISBN: 0201122278.

- [10] Riad Sonbol, Ghaida Rebdawi y Nada Ghneim. «The Use of NLP-Based Text Representation Techniques to Support Requirement Engineering Tasks: A Systematic Mapping Review». En: *IEEE Access* 10 (2022), págs. 62811-62830. ISSN: 2169-3536. DOI: [10.1109/ACCESS.2022.3182372](https://doi.org/10.1109/ACCESS.2022.3182372). URL: <http://dx.doi.org/10.1109/ACCESS.2022.3182372>.
- [11] Tomas Cicchini et al. «News sharing on Twitter reveals emergent fragmentation of media agenda and persistent polarization». En: *EPJ Data Science* 11.1 (19 de ago. de 2022), pág. 48. DOI: [10.1140/epjds/s13688-022-00360-8](https://doi.org/10.1140/epjds/s13688-022-00360-8). URL: <https://doi.org/10.1140/epjds/s13688-022-00360-8>.
- [12] Natalie K. Cygan, • Mentor y Megan Leszczynski. «Sentence-BERT for Interpretable Topic Modeling in Web Browsing Data». En: 2021. URL: <https://api.semanticscholar.org/CorpusID:235343808>.
- [13] Diksha Khurana et al. «Natural language processing: state of the art, current trends and challenges». En: *Multimedia Tools and Applications* 82.3 (1 de ene. de 2023), págs. 3713-3744. DOI: [10.1007/s11042-022-13428-4](https://doi.org/10.1007/s11042-022-13428-4). URL: <https://doi.org/10.1007/s11042-022-13428-4>.
- [14] Divya Khyani y Siddhartha B S. «An Interpretation of Lemmatization and Stemming in Natural Language Processing». En: *Shanghai Ligong Daxue Xuebao/Journal of University of Shanghai for Science and Technology* 22 (ene. de 2021), págs. 350-357.
- [15] Juan Manuel Pérez et al. *pysentimiento: A Python Toolkit for Opinion Mining and Social NLP tasks*. 2023. arXiv: [2106.09462 \[cs.CL\]](https://arxiv.org/abs/2106.09462).
- [16] Tomás Salgueiro et al. «A Spanish dataset for Targeted Sentiment Analysis of political headlines». En: *Memorias de las JAIIO* 8.2 (dic. de 2022), págs. 92-97. URL: <https://publicaciones.sadio.org.ar/index.php/JAIIO/article/view/269>.
- [17] Mamdouh Farouk. «Measuring Sentences Similarity: A Survey». En: *Indian Journal of Science and Technology* 12.25 (jul. de 2019), págs. 1-11. ISSN: 0974-6846. DOI: [10.17485/ijst/2019/v12i25/143977](https://doi.org/10.17485/ijst/2019/v12i25/143977). URL: <http://dx.doi.org/10.17485/ijst/2019/v12i25/143977>.
- [18] Nils Reimers e Iryna Gurevych. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. 2019. arXiv: [1908.10084 \[cs.CL\]](https://arxiv.org/abs/1908.10084).
- [19] Alex Sherstinsky. «Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network». En: *Physica D: Nonlinear Phenomena* 404 (mar. de 2020), pág. 132306. ISSN: 0167-2789. DOI: [10.1016/j.physd.2019.132306](https://doi.org/10.1016/j.physd.2019.132306). URL: <http://dx.doi.org/10.1016/j.physd.2019.132306>.
- [20] Sepp Hochreiter y Jürgen Schmidhuber. «Long Short-term Memory». En: *Neural computation* 9 (dic. de 1997), págs. 1735-80. DOI: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735).
- [21] Jacob Devlin et al. «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». En: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. por Jill Burstein, Christy Doran y Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, jun. de 2019,

- págs. 4171-4186. DOI: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423). URL: <https://aclanthology.org/N19-1423>.
- [22] Daniel Jurafsky y James Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Vol. 2. Feb. de 2008.
- [23] Wei Xu, Xin Liu y Yihong Gong. «Document clustering based on non-negative matrix factorization». En: *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '03. Toronto, Canada: Association for Computing Machinery, 2003, págs. 267-273. ISBN: 1581136463. DOI: [10.1145/860435.860485](https://doi.org/10.1145/860435.860485). URL: <https://doi.org/10.1145/860435.860485>.
- [24] Iqbal H. Sarker. «Machine Learning: Algorithms, Real-World Applications and Research Directions». En: *SN Computer Science* 2.3 (22 de mar. de 2021), pág. 160. DOI: [10.1007/s42979-021-00592-x](https://doi.org/10.1007/s42979-021-00592-x). URL: <https://doi.org/10.1007/s42979-021-00592-x>.
- [25] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: [1706.03762 \[cs.CL\]](https://arxiv.org/abs/1706.03762).
- [26] Dzmitry Bahdanau, Kyunghyun Cho y Yoshua Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. 2016. arXiv: [1409.0473 \[cs.CL\]](https://arxiv.org/abs/1409.0473).
- [27] Ernesto Estrada. *The Structure of Complex Networks: Theory and Applications*. USA: Oxford University Press, Inc., 2011. ISBN: 019959175X.
- [28] Dmitry Zinoviev. *Complex Network Analysis in Python: Recognize - Construct - Visualize - Analyze - Interpret*. 1st. Pragmatic Bookshelf, 2018. ISBN: 1680502697.
- [29] Elad Segev, ed. *Semantic Network Analysis in Social Sciences*. London: Routledge, 2021.
- [30] Santo Fortunato. «Community detection in graphs». En: *Physics Reports* 486.3 (2010), págs. 75-174. ISSN: 0370-1573. DOI: <https://doi.org/10.1016/j.physrep.2009.11.002>. URL: <https://www.sciencedirect.com/science/article/pii/S0370157309002841>.
- [31] Vincent D Blondel et al. «Fast unfolding of communities in large networks». En: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (oct. de 2008), P10008. ISSN: 1742-5468. DOI: [10.1088/1742-5468/2008/10/p10008](https://doi.org/10.1088/1742-5468/2008/10/p10008). URL: <http://dx.doi.org/10.1088/1742-5468/2008/10/P10008>.
- [32] M. Rosvall, D. Axelsson y C. T. Bergstrom. «The map equation». En: *The European Physical Journal Special Topics* 178.1 (1 de nov. de 2009), págs. 13-23. DOI: [10.1140/epjst/e2010-01179-1](https://doi.org/10.1140/epjst/e2010-01179-1). URL: <https://doi.org/10.1140/epjst/e2010-01179-1>.
- [33] Peng Qi et al. «Stanza: A Python Natural Language Processing Toolkit for Many Human Languages». En: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2020. URL: <https://nlp.stanford.edu/pubs/qi2020stanza.pdf>.
- [34] Juan Martinez-Romo et al. «Disentangling categorical relationships through a graph of co-occurrences». En: *Physical review. E, Statistical, nonlinear, and soft matter physics* 84 (oct. de 2011), pág. 046108. DOI: [10.1103/PhysRevE.84.046108](https://doi.org/10.1103/PhysRevE.84.046108).
- [35] José Cañete et al. *Spanish Pre-trained BERT Model and Evaluation Data*. 2023. arXiv: [2308.02976 \[cs.CL\]](https://arxiv.org/abs/2308.02976).