UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingeniería de Sistemas Informáticos

Monitorización de información falsa mediante Análisis de Redes Sociales y Procesamiento del Lenguaje Natural

TESIS DOCTORAL

Presentada para optar al título de Doctor por

Guillermo Villar Rodríguez
Doble Grado en Periodismo y Comunicación Audiovisual

Madrid, 2025

UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingeniería de Sistemas Informáticos

Doctorado en Ciencias y Tecnologías de la Computación para Smart Cities

Monitorización de información falsa mediante Análisis de Redes Sociales y Procesamiento del Lenguaje Natural

TESIS DOCTORAL

Presentada para optar al título de Doctor por:

Guillermo Villar Rodríguez
Doble Grado en Periodismo y Comunicación Audiovisual

Bajo la dirección de:
Dr. David Camacho Fernández
Dr. Alejandro Martín García

Madrid, 2025

Título: Monitorización de información falsa mediante Análisis de Redes Sociales y Procesamiento del Lenguaje Natural

Autor: Guillermo Villar Rodríguez

Programa de doctorado: Ciencias y Tecnologías de la Computación para Smart Cities

Supervisión de la tesis:

Dr. David Camacho Fernández, Catedrático de Universidad, Universidad Politécnica de Madrid (Director)

Dr. Alejandro Martín García, Profesor Contratado Doctor, Universidad Politécnica de Madrid (Director)

Revisores externos:

Comité de defensa de la tesis:

Fecha de defensa de la tesis:

Esta investigación ha sido financiada por el proyecto CIVIC: Caracterización inteligente de la veracidad de la información asociada a la COVID-19, otorgado por las Ayudas a Proyectos de Investigación Científica SARS CoV-2 y COVID-19 Fundación BBVA; por el proyecto PCI2022-134990-2 (MARTINI) del IV programa Cofund 2021 CHISTERA, financiado por MCIN/AEI/10.13039/ 501100011033 y por la European Union NextGenerationEU/PRTR; por el proyecto DisTrack, concedido por la Fundación Mobile World Capital; por el Ministerio de Ciencia e Innovación de España a través de FightDIS (PID2020-117263GB-I00); por MCIN/AEI/10.13039/501100011033/ y la European Union NextGenerationEU/PRTR para XAI-Disinfodemics (PLEC 2021-007681); por la Comisión Europea mediante IBERIFIER - Iberian Digital Media Research and Fact-Checking Hub (2020-EU-IA-0252) e IBERIFIER Plus (DIGITAL-2023-DEPLOY- 04-EDMO-HUBS 101158511), por EMIF, gestionado por la Fundación Calouste Gulbenkian en el proyecto MuseAI, y por la Comunidad de Madrid en el proyecto CIRMA (TEC-2024/COM-404).

Agradecimientos

Gracias a mis tutores de la tesis por valorar los perfiles interdisciplinares y ver mi salto a la inteligencia artificial desde el periodismo y la comunicación como una virtud desde el primer momento. Gracias siempre, David y Álex, por acoger a quienes nos hemos salido de la norma en el ámbito académico para tender estos puentes entre áreas del conocimiento. Por entender, en esencia, que quienes somos *outliers* en la carrera investigadora somos igual de importantes que el resto. Aunque pueda encapsular todas estas palabras en uno de tantos vectores y números que hemos tratado, es muy difícil comprimir por aquí todo el agradecimiento que os tengo. Habéis sido el *Transformer* de mi andadura en el mundo de los campos más técnicos.

Gracias a mis padres, las personas que más quiero, por ser mis dos escoltas de la tesis. Gracias a ti, mamá, que has ido siempre por delante de mí para ponerme los pies en la tierra. Con ellos he dado pasos firmes, gracias a tus cuidados, tu protección, tu comprensión y tu visión pragmática de las cosas. Siempre has sido mi orden en el caos. Este doctorado también es tuyo por ser la persona que más me ha enseñado. Aprender de ti es un orgullo y una lección constante para progresar en la investigación y en la vida. Y, sobre todo, para ser mejor persona, con o sin doctorado.

Gracias a ti, Matías/papá, que has ido siempre detrás de mí para protegerme y apoyarme. Eres el mejor guardaespaldas y el mejor fan, cuando es a ti a quien debo admirar. Las veces que me impulsabas a pedalear sin patines en bicicleta y que me animabas a seguir eran solo un anticipo de lo que vendría después. Me has potenciado durante toda mi vida, levantándome aunque me cayera y celebrando mis logros. Contigo todo ha ido sobre ruedas y no solo has sido el motor de mis éxitos: también les has puesto sus transmisiones, como únicamente tú sabes hacer, para que estos lleguen de la mejor manera posible.

Gracias a mis primos y a mis tíos por darme el privilegio de estudiar arropado por sus lazos, no solo de sangre sino de amor incondicional. He estado dispuesto a estos retos porque con ellos he tenido la suerte de jugar en "modo fácil". Gracias, abuela, por preocuparte en estos años y hacer que cualquier altibajo sea relativo cada vez que me has preguntado por esta travesía. A ellos y a todos los no están pero permanecen en el recuerdo va dedicada esta tesis. Por eso también las gracias van a mi abuelo Máximo, para quien fui "el periodista" hasta sus 102 años, pero que ojalá me hubiera visto como "el investigador". Gracias también a mi tía Pauli por su cariño, por sus rosquillas tanto en España como en el extranjero y por sus enhorabuenas, de la primera a la última cuando empecé esta aventura. Y gracias a mi abuelo Santiago por los triunfos que celebré con él de pequeño, porque esto ha hecho que también los tenga de mayor.

Gracias a mis compañeros elegidos de vida. No hay mejor prueba para medir las amistades como doctorando que el número de veces que me han preguntado "¿y cómo llevas la tesis?". Esto me ha dado la certeza de saber que cuento con grandes personas alrededor: de España a Indonesia y Puerto Rico, pasando por Turín; desde quienes me reclutaron en la adolescencia hasta quienes en mi trabajo en Relevo y fuera de él me han tendido la mano para estar en sus vidas. Sea con palabras de afecto, notas largas de voz o cualquier otro gesto de empatía. Gracias a todos.

Gracias al grupo AIDA por darme todo el sustento necesario para la tesis. No solo han quedado

plasmados en grafos los nodos y aristas de nuestros experimentos, sino todas las conexiones humanas en este largo viaje, de las que he recibido mucho cariño y talento. El afecto y dedicación de Helena, el ambiente de Javi T o los halagos de Sergio son solo algunos ejemplos del baúl de recuerdos de todos y cada uno de los componentes que han pasado o siguen en esta familia investigadora. AIDA no solo lleva detrás el significado de sus siglas, sino la humanidad de todos sus miembros que, sin excepción, han construido esta etapa.

Y, por último, gracias a quienes empezaron y siguen el camino. Esta investigación no habría sido posible sin los avances previos en el campo de la inteligencia artificial y sin la existencia de los *fact-checkers*. Los desmentidos de organizaciones como Maldita, Newtral, Efe Verifica o Verificat en España, acreditados con el sello de la IFCN, han sido materia prima e inspiración para elaborar esta tesis. Espero que estos resultados contribuyan a los siguientes escalones en la lucha contra la desinformación.

Resumen

Esta investigación presenta una metodología innovadora para monitorizar conversaciones sobre desinformación en redes sociales a partir de técnicas computacionales de dos disciplinas: Procesamiento del Lenguaje Natural y Análisis de Redes Sociales. Mediante la aplicación de modelos de lenguaje y grafos, se generan modelos que permiten entender el ecosistema completo de las falsedades, su evolución a lo largo del tiempo y sus desmentidos en estas plataformas sociales.

Mediante las palabras claves extraídas de enunciados de información falsa (*claims*) verificados por *fact-checkers*, este trabajo ha obtenido los *posts* a lo largo del tiempo sobre una falsedad en redes sociales y los ha convertido en *embeddings* semánticos a través de modelos de lenguaje. Sobre estos *posts*, se aplican modelos entrenados en tareas de *Natural Language Inference* con el objetivo de evaluar su nivel de alineamiento con el texto original. Esto permite saber si ambos expresan lo mismo, lo contrario o neutralidad entre si, identificando el papel de cada contenido dentro de la conversación sobre desinformación.

Con estos tres indicadores, unidos a distintos metadatos descargados de la red social, esta tesis plantea una serie de métodos para analizar toda la conversación existente alrededor de una determinada desinformación. Mediante la generación de grafos, estos contenidos sobre la información falsa se estructuran en nodos (los posts en sí) y sus conexiones (republicaciones y respuestas), visualizados en función a su alineación con el *claim* y ordenados cronológicamente. De esta manera, se consigue explorar el ciclo de vida de esta pieza de desinformación en la red social desde su origen, incluyendo los usuarios que la comparten, sus distintas formas o los desmentidos lanzados hacia ella (por ejemplo, por los *fact-checkers*).

Este trabajo intenta realizar un importante progreso en el *fact-checking* semiautomático, donde la integración de modelos de lenguaje y otros métodos de inteligencia artificial, lejos de sustituir la rutina profesional de los *fact-checkers*, trata de ayudarlos. Mediante una combinación con técnicas de Análisis de Redes Sociales y visualizaciones en forma de grafos, el objetivo de esta investigación es proporcionar una herramienta que permita entender mejor los mecanismos bajo los cuales se genera desinformación y se disemina, así como ayudar a detectarla y prevenirla.

Abstract

This research presents an innovative methodology for monitoring conversations that include disinformation in social networks based on computational techniques from two disciplines: Natural Language Processing and Social Network Analysis. Through the application of language processing and graphs, models are generated to monitor the entire ecosystem of falsehoods, their evolution over time and their denials on these social platforms.

By searching through a series of keywords extracted from false information statements (claims) disproved by fact-checkers, this work has obtained the posts over time about a falsehood in social networks and has converted them into semantic embeddings through language models. Models trained in Natural Language Inference tasks are applied on these posts to evaluate their level of alignment with the original text. This allows us to know whether both express the same, the opposite or neutrality to each other, identifying the role of each content within the conversation about misinformation.

With these three indicators, together with different metadata downloaded from the social network, this thesis proposes a series of methods to analyse all the existing conversation around a given piece of disinformation. Through the generation of graphs, these contents about false information are structured in nodes (the posts themselves) and their connections (reposts and responses), visualised according to their alignment with the claim and ordered chronologically. In this way, it is possible to explore the life cycle of this piece of disinformation in the social network from its origin, including the users who share it, its different forms or the disavowals launched towards it (for example, by the fact-checkers).

This work attempts to make important progress in semi-automatic fact-checking, where the integration of language models and other artificial intelligence methods, far from replacing the professional routine of the fact-checkers, tries to help them. Through a combination of Social Network Analysis techniques and graph-based visualisations, the aim of this research is to provide a tool to better understand the mechanisms under which disinformation is generated and disseminated, as well as to help detect and prevent it.

Índice general

Índice de Figuras

Índice de Tablas

A quienes hacen que mi vida no sea un bulo
y permiten que me muestre como soy de verdad

Abreviaturas y acrónimos

ANLI *Adversarial Natural Language Inference* . . . 85
API Interfaz de Programación de Aplicaciones . . . 19
BOW *Bag-of-Words* . . . 24
CBOW *Continuous Bag-of-Words* . . . 26
DL *Deep Learning* . . . 12
DNN *Deep Neural Network* . . . 67
DOSNs *Decentralized Online Social Networks* . . . 114
DSS *Decision Support Systems* . . . 99
FEVER *Fact Extraction and VERification dataset* . . . 85
GRU *Gated Recurrent Units* . . . 23
HDP *Hierarchical Dirichlet Process* . . . 61
IA Inteligencia Artificial . . . 3
IFCN *International Fact-Checking Network* . . . 3
LDA *Latent Dirichlet Allocation* . . . 61
LIA Análisis de la Información del Lenguaje . . . 60
LLMs *Large Language Models* . . . 28
LSTM *Long Short-Term Memory* . . . 23
ML *Machine Learning* . . . 7
MNLI-MT *Machine Translated MultiNLI* . . . 85
MSTSb *Multilingual Semantic Textual Similarity benchmark* . . . 71
MultiNLI *Multi-Genre Natural Language Inference* . . . 30
NER *Name Entity Recognition* . . . 22
NLG *Natural Language Generation* . . . 20
NLI *Natural Language Inference* . . . 5
NLU *Natural Language Understanding* . . . 20
OSNs *Online Social Networks* . . . 2
PCA *Principal Component Analysis* . . . 21
PLN Procesamiento del Lenguaje Natural . . . 3
reLU *Rectified Linear Unit* . . . 67
RNN *Recurrent Neural Networks* . . . 23
RQ *Research Question* . . . 5
RTE *Recognizing Textual Entailment* . . . 30
SI Susceptible-Infectado . . . 40

SICK *Sentences Involving Compositional Knowledge data set* 77
SIR Susceptible-Infectado-Recuperado 40
SIS Susceptible-Infectado-Susceptible 40
SNA *Social Network Analysis* 3
SNLI *Stanford Natural Language Inference Corpus* 30
STS Similitud Textual Semántica 71
STSb *Semantic Textual Similarity benchmark* 70
SVM *Support Vector Machines* 22
tf-idf *term frequency – inverse document frequency* 22
UMAP *Uniform Manifold Approximation and Projection for Dimension Reduction* 73
XNLI *Cross-lingual Natural Language Inference corpus* 30

Capítulo 1

Introducción

Hemos sido víctimas de un bulo.
Nos han atacado en una especie de locura colectiva,
un ataque en el que se hablaba de algo que nunca ocurrió
y de protagonistas que nunca han existido.

— Concha Velasco

El Capítulo 1 inicia con el marco teórico del impacto de información falsa en la sociedad actual (1.1). Esto permite encuadrar el problema de esta tesis pero también los motivos en la investigación para enfrentarlo (1.2). Ya identificados, se presentan las preguntas de investigación (1.3), que se irán resolviendo con los experimentos de las siguientes secciones y que recibirán respuesta en el Capítulo 6. Son el esqueleto de esta tesis, así como los objetivos planteados (1.4) que se perseguirán en esta investigación a la hora de responder a las preguntas en este apartado. A continuación, se explican cuáles son las principales aportaciones fruto del desarrollo de este trabajo (1.5), se describe la estructura que sigue el documento (1.6) y, por último, se indican las publicaciones resultadas del desarrollo de la tesis, incluyendo los artículos de revista y de congreso, que contribuyen de manera principal y secundaria (1.7).

1.1. Introducción al problema: desinformación y su presencia en redes sociales

La desinformación siempre ha estado presente, pero su expansión en las redes sociales y canales de mensajería constituye un problema global de nuestra era. Estas redes han cambiado la forma de relacionarnos con la información y, en consecuencia, de expandir aquella que es falsa. Mientras que antes los contenidos eran de propagación *peer-to-peer*, ahora han pasado a ser de propagación *many-to-many* y esto facilita la diseminación de la desinformación [13], algo ya demostrado en casos concretos [14].

Dentro de este panorama, la desinformación no es siempre accidental, y se puede distinguir entre varios tipos. Para ello, se suelen utilizar los términos ingleses, que en este caso permiten una mayor exactitud (ver Fig. 1.1). La palabra '*misinformation*' engloba toda información falsa en general pero también aquella no intencionada [15, 16]; la palabra '*disinformation*', sin embargo, alude

a la información falsa compartida de forma deliberada [17], una distinción ya señalada mucho antes [18]. En este sentido entra en juego la palabra '*malinformation*', contenido ya diseñado como arma arrojadiza [1, 19].

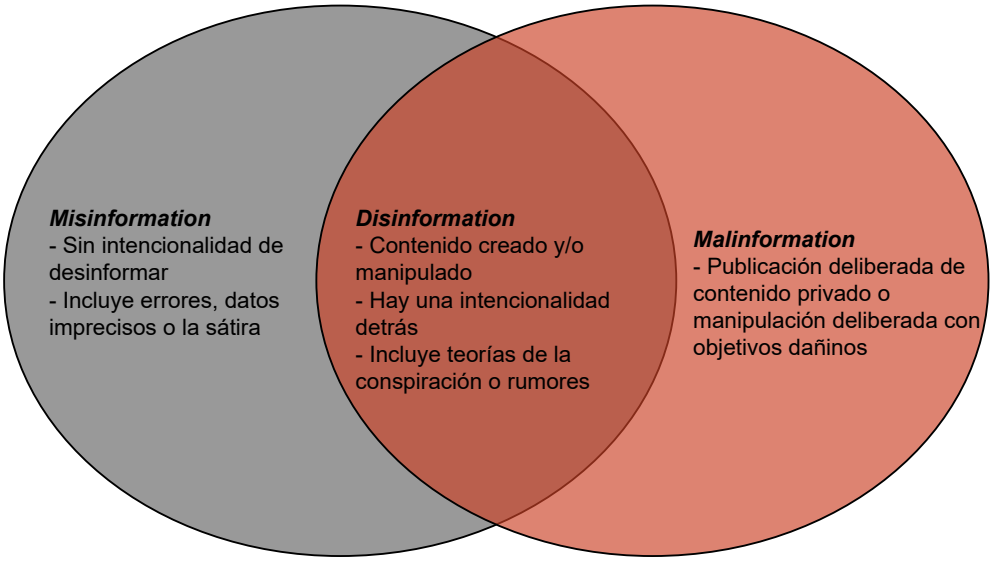


Figura 1.1: Diagrama adaptado a partir del trabajo de Wardle y Derakhshan [1] para comparar los términos '*misinformation*', '*disinformation*' y '*malinformation*'.

Esta irrupción de las falsedades en las *Online Social Networks* (OSNs) tiene como prueba el creciente número de artículos académicos sobre este problema [20, 21, 16, 22]. En este contexto de las redes sociales, no hay consenso de cuándo estalló la expansión de información falsa [13], pero se apunta a 2017, tras las elecciones de 2016 de Estados Unidos, como el momento de un mayor interés por este ámbito de estudio a nivel académico [23], pero no fue lo único que aupó el foco en este tema [24]. Además, la investigación constata cómo las dudas y proclamas contra, por ejemplo, las vacunaciones frente al sarampión en Estados Unidos ya tomaban un significativo impacto en Twitter y Facebook [25], y la investigación de redes sociales en situaciones de crisis ya se había puesto sobre la mesa [26]. Las elecciones a la presidencia y los conflictos internacionales son protagonistas de la desinformación y hacen de su estudio una cuestión de mayor escala.

De acuerdo a una encuesta publicada por el *Digital News Report* del Instituto Reuters en 2024, más de un tercio de la población (36 %) ha recibido información engañosa sobre política, siete puntos porcentuales más que el año anterior. Crece también aquella sobre salud (18 %, 6 puntos más), inmigración (21 %, 7 puntos más), medioambiente y cambio climático (23 %, 6 puntos más). Se mantienen presentes los contenidos espurios sobre el coronavirus (30 %) y la guerra en Ucrania (24 %), y emerge el conflicto Israel-Palestina (27 %) [27].

Más allá de formar parte de la sociedad, las redes sociales destacan como vía para informarse. Además, según estos informes del Instituto Reuters, se accede menos en general a los diarios digitales que en la década anterior, un descenso auspiciado por los jóvenes de 18 a 24 años y de 25 a 34 años frente al consumo más regular de los encuestados que superan estas edades [28, 27]. Se produce así un conflicto de intereses: mientras que el informe refleja una población cada vez más preocupada por la desinformación, hay una tendencia creciente a favor de las redes sociales, masivos propagadores de información falsa. A un 27 % y a un 24 % de quienes usan TikTok

y Twitter (X), respectivamente, les cuesta encontrar información fiable. Todo en un caldo de cultivo con un desinterés cada vez mayor por las noticias durante la última década, pese a la tendencia más estable de 2023 a 2024, con un 39 % de personas encuestadas que directamente las evitan [27].

A nivel periodístico, las organizaciones de *fact-checking* se erigen como la defensa a los ciudadanos frente al auge de la desinformación. A partir de unas pautas definidas por la *International Fact-Checking Network* (IFCN) a través de su 'Código de Principios' para verificar la información y desmentir la que es falsa, los *fact-checkers* combaten las falsedades en todo el mundo. Ya se ha demostrado que las verificaciones de estas organizaciones tienen efecto contra la desinformación sobre las personas en varios países [29].

Pero cuando llegó una crisis internacional como la de la COVID-19, Cristina Tardáguila, la entonces presidenta de la IFCN, apuntó a las enormes dimensiones que los bulos estaban alcanzando en comparación a eventos previos [30]. De la conversación sobre desinformación del coronavirus en inglés, el estudio de la infodemia en 2021 mostró que las elaboraciones de *fact-checkers* y otras verificaciones comprendieron el 16.4 % de su total [31].

1.2. Descripción del problema y motivación

La ola de desinformación que supuso para las organizaciones de *fact-checking* el coronavirus, más una lucha contra ella que no cesa sino que se agrava en crisis como las guerras de Rusia-Ucrania y de Israel-Gaza, apunta a la misma cuestión: el reto de que el volumen de mensajes falsos no sea inabarcable. Contar con herramientas de automatización para una respuesta orquestada frente a las falsedades en redes se convierte así en prioridad dadas las circunstancias.

La Inteligencia Artificial (IA) se posiciona como el paraguas de los métodos automáticos contra la información falsa en crecimiento. Los avances en técnicas de aprendizaje automático han permitido, desde sus inicios, clasificar contenidos y diferenciar unos de otros de acuerdo a los datos que reciben como entrenamiento. Sobre esta base, estos métodos han evolucionado en el área del procesamiento del lenguaje para un etiquetado más refinado de los contenidos según el tipo que sean.

Partir de este punto como motivación contesta a tres preguntas: el **'qué'**, es decir, el problema de la desinformación en aumento; el **'por qué'**, referente a la necesidad de aplacarla de forma automática para que no sobrepase los recursos humanos de las organizaciones de *fact-checking*, y el **'cómo'**, a través de herramientas de IA. Pero esta investigación invita también a la contestación del **'dónde'** y del **'cuándo'**.

El **'dónde'** sería en el ámbito de las redes sociales, por ser el nido de la desinformación por su fácil propagación [13]. Atacar la información falsa en las OSNs es, por tanto, mitigar en la medida de lo posible la cadena de expansión de falsedades para que llegue a menos círculos. La rama de *Social Network Analysis* (SNA), hacia la investigación de los ecosistemas en los *social media* permite abordar esta cuestión.

Con el Procesamiento del Lenguaje Natural (PLN) y el SNA, que, a su vez, concretan más la cuestión del **'cómo'**, se aborda también el **'cuándo'**: no solo mitigar la desinformación *a posteriori*, como un parche, sino desde que se va fraguando en las redes sociales, antes de que crezca tanto que sea muy difícil atajarla. Se busca de esta manera un estudio de la información falsa de principio a fin para entender sus dinámicas y poder mejorar la lucha en el tiempo.

Los modelos actuales de PLN y la generación de grafos dentro del análisis de redes son la base de esta investigación para modelar la propagación de conversaciones sobre desinformación, tomando Twitter (ahora X) como epicentro. Tanto para acceder a fuentes *mainstream* como para explorar otras de índole política, de activismo y/o no hegemónicas, Twitter destaca respecto a otras plataformas por el consumo informativo [28]. Esto produciría el caldo de cultivo hacia los desórdenes de la información y, con ello, la tarea de aplacarlos.

Este seguimiento de conversaciones que incluyen desinformación no consiste en la captura del contenido alrededor de una falsedad en un momento congelado del tiempo, sino en un trazado de toda su trayectoria desde su irrupción en la red. Por un lado, esto permite luchar contra la información falsa con un paso que va más allá de comprobar si un enunciado es verdadero, detectando todo aquel contenido referido a ello; por otro, permite estudiar la trayectoria de todos los mensajes en torno a tal desinformación, los actores que participan en ella y el impacto generado dentro de este ecosistema de posts.

Esta lucha, que pretende abarcar todos los contenidos en una red social con falsedades y sus dinámicas, persigue abandonar las metodologías tradicionales que categorizan como verdaderas o falsas las publicaciones en función a *datasets* de entrenamiento. Por un lado, estos carecen de datos actualizados y confían en los registros de desinformaciones que no tienen por qué ser las se aludan en los contenidos a verificar; por otro, el resultado de tal entrenamiento no etiqueta según el sentido de cada post, sino según patrones comunes no relacionados con la afirmación a examen.

En consecuencia, las metodologías a lo largo de esta investigación no serán completamente automáticas, sino que abrazan un enfoque semiautomático [9, 32], como se explicará después. Así, se partirá de bases del conocimiento en vez de un tratamiento tradicional con *datasets* de entrenamiento. Las organizaciones de *fact-checking* serán el pilar de esta base y la compondrán los enunciados que hayan identificado en sus desmentidos, en vez de los conjuntos masivos de contenidos verdaderos y falsos para entrenar en los métodos tradicionales.

De esta forma, se produce una simbiosis entre el cometido de los *fact-checkers* y los diferentes modelos de lenguaje para convertir los desmentidos en la materia prima de una lucha más tecnológica contra la desinformación, donde con los procedimientos señalados de PLN y de SNA se avance hacia mejores *insights* de la evolución y propagación de los posts sobre información falsa.

1.3. Hipótesis y preguntas de la investigación

Esta investigación parte de la siguiente hipótesis: **"Se puede monitorizar la desinformación en redes sociales"**. Como se ha indicado, el foco aquí no responde a clasificar directamente posts falsos a partir de patrones de un *dataset* ya etiquetado, sino a relacionar estos contenidos con la falsedad ya desmentida a la que se refieren para poder seguir su recorrido en las redes sociales.

La pregunta principal, por tanto, a raíz de la hipótesis planteada y con las disciplinas a abordar para explotar las propiedades textuales y de las redes sociales, se formula así: **"¿Es posible monitorizar las conversaciones sobre informaciones falsas en una red social con PLN y SNA?"**. No solo por la complejidad de esta cuestión la pregunta necesita dividirse en subpreguntas, sino también porque obliga a contestar primero si se pueden identificar estas desinformaciones (mediante PLN) para responder después si se puede mostrar su evolución (mediante SNA). Cada

Research Question (RQ) subyacente abordará estos aspectos.

A la hora de procesar el lenguaje, el primer paso es ver si es factible relacionar un enunciado falso con posts que también comparten esta afirmación. De forma preliminar, se chequeará como punto de partida si las representaciones vectoriales capturan el significado de los contenidos de tal modo que puedan agruparse aquellos parecidos semánticamente, de acuerdo a la similitud semántica calculada entre ellos. Las RQ en este bloque son:

- **RQ 1**. ¿Es posible extraer la cadena de diseminación de una información falsa en una red social?
- **RQ 2**. ¿Se puede generar una representación vectorial donde posts relacionados con la misma desinformación mantengan distancias cercanas?
- **RQ 3**. ¿Se pueden relacionar los posts de distintas falsedades en función a su cercanía semántica?

Después, se incidirá en la rama del *Natural Language Inference* (NLI), para resolver si, más allá de la relación semántica, es posible alinear los enunciados falsos desmentidos ya por las organizaciones de *fact-checking* con los posts que expresen exactamente lo mismo (propagación de esa desinformación) o lo contrario (correcciones de los *fact-checkers* y de otros usuarios). Esta utilidad del NLI se comprobará analizando los mensajes con desinformación y con su desmentido en las OSNs. Las preguntas en este apartado son:

- **RQ 4**. ¿Se pueden separar los posts relacionados con la información falsa en la conversación de aquellos no relacionados con ella?
- **RQ 5**. ¿Se pueden distinguir las publicaciones que propagan una información falsa de aquellas que la contradicen?
- **RQ 6**. ¿Se puede extraer información de las publicaciones que propagan o contradicen la información falsa en función al número de interacciones de los posts?
- **RQ 7**. ¿Difieren las proporciones entre los posts y usuarios que diseminan desinformación frente a aquellos que la contradicen dependiendo del número de interacciones?

Por último, mediante técnicas de SNA se observará si los contenidos a la par y contrarios a una información falsa específica, entre otros, pueden estructurarse dentro de un grafo, en un entramado que permita seguir su recorrido a lo largo del tiempo. Completar esta parte lleva a contestar la siguiente subpregunta, que permitirá a su vez resolver la pregunta principal, producto de completar todo este camino:

- **RQ 8**. ¿Se puede trazar el movimiento de los posts relacionados con una información falsa y los usuarios que la propagan de principio a fin?

1.4. Objetivos

Como objetivos de esta investigación, se plantean los siguientes:

- **Objetivo 1. Monitorización de desinformación mediante PLN y SNA**. Como el título y la hipótesis indican, la primera meta consiste en unir técnicas de SNA y modelos de

PLN para analizar y trazar la diseminación de mensajes falsos. Si bien ambas disciplinas se presentan como los pilares contra las falsedades en el área computacional [32], predomina su uso individual y no combinado. El trazado propuesto en esta investigación para este tipo de conversaciones es, en esencia, el producto de unir las fortalezas individuales de PLN para caracterizar la desinformación [33, 34, 35] y de SNA para seguir las ramificaciones de las publicaciones [36, 7] y desentrañar en grafos estos desórdenes informativos [37, 38] más sus comunidades [39, 40, 41]. En este objetivo, los modelos de PLN posibilitan el filtro para distinguir entre qué es información falsa y qué no, mientras que las técnicas de SNA permiten estructurarlo todo dentro de una red.

- **Objetivo 2. *Fact-checking* semiautomático mediante PLN**. Llegar a este punto obliga también a cambiar las concepciones de estas disciplinas. Dentro de la rama del lenguaje, la segunda meta es explotar más allá del *claim matching* [42, 43] la viabilidad del *fact-checking* semiautomático en el PLN [9]. Se trata de un paso más para apartarse de los enfoques tradicionales totalmente automáticos que clasifican los contenidos a partir de patrones y no porque se hayan desmentido ya de manera oficial. La intención no es solo la de alinear cualquier publicación con el contenido falso al que se refiere, sino hacer esto con toda la conversación sobre este en la red social para visibilizar cómo es la evolución de los textos que lo difunden o contradicen [9]. Así, el *claim monitoring* emergería como la evolución del *claim matching* para un mayor protagonismo dentro del *fact-checking* semiautomático.

- **Objetivo 3. Modelado de la cascada completa de propagación de la desinformación**. Dentro del ámbito específico de las OSNs, el tercer objetivo es conseguir modelar todo el ecosistema de la desinformación y no solo el recorrido de los posts con más impacto. Se habla de las infodemias [30] en la analogía de la desinformación como un virus de rápido y amplio alcance pero esta analogía no se da a la hora de mostrar esta lacra a través de todos los contagios y focos entre cuentas de una red social. Son populares en este aspecto las cascadas de publicaciones surgidas a partir de un mensaje viral, pero en estos casos el SNA y la generación de grafos se siguen concibiendo en cierto modo como *broadcast models*, donde el contagio parte de un único nodo que infecta a los siguientes, aunque estas infecciones posteriores provengan en forma de árbol [7, 36]. Esta tesis, sin embargo, trabajará en la generación de grafos como *viral models* ya desde el inicio, donde la desinformación puede tener varios focos, no partir de una única cascada y contribuir en mayor o en menor medida a otros contagios.

- **Objetivo 4. Visualización del papel de los *fact-checkers* en la diseminación de desinformación**. La cuarta meta, consecuencia de la anterior, es mostrar en el mapa de la desinformación la respuesta del *fact-checking* frente a ella. En el ecosistema de la información falsa al que aspira el objetivo previo, no solamente están en la conversación los posts que difunden las falsedades, sino también aquellos que expresan un significado contrario a ellas, incluyendo los desmentidos por parte de las organizaciones de verificación. Construir los grafos desde cero permite esta imagen más verosímil en la que estos contenidos nocivos confluyen con los posts que los rebaten, no necesariamente dentro de una misma cascada pero sí en el mismo espacio, frente a la concepción de los modelos donde todo se expande a partir de un mismo nodo [7, 36] sin ninguna injerencia.

- **Objetivo 5. Crear una herramienta para un trabajo conjunto efectivo entre *fact-checkers* y modelos de IA**. Las metas anteriores hacen que el fin último de dar una herramienta de monitorización a las organizaciones de *fact-checking* no suponga un

menoscabo a sus fórmulas de rigor. Al igual que los *fact-checkers* buscan cumplir con el 'Código de Principios' de la IFCN para validar su trabajo periodístico [44], esta tesis propone que los enfoques computacionales vayan de la mano con las pautas para verificar y desmentir. Los enfoques totalmente automáticos para la detección directa de la información falsa a partir de patrones van en contra de esta senda, a no ser que lleven consigo metodologías para su explicabilidad contra las cajas negras.

1.5. Aportaciones

En línea con los objetivos marcados, esta investigación supone una evolución respecto a trabajos anteriores mediante las siguientes aportaciones:

- Ofrece un método innovador para monitorizar contenidos desinformativos en redes sociales mediante un enfoque novedoso que integra modelos de lenguaje y la generación de grafos para trazar su recorrido, todo ello en una estructura modular para usar cada una de las aportaciones computacionales de forma independiente en cualquier línea de investigación posible. Además, estos avances se presentan de tal forma que pueden aplicarse a otras plataformas sociales y a otros fenómenos, y son flexibles a mejoras técnicas.

- Antepone el enfoque práctico en la batalla contra estas falsedades al meramente técnico en el PLN. De los métodos avanzados del lenguaje más recientes, también abordados, prioriza las necesidades de los agentes e instituciones contra la desinformación frente a solo la carrera por ofrecer el mejor etiquetado de los textos como verdaderos o falsos en una concepción clásica de esta tarea. Tal camino tradicional taparía la labor necesaria de desmentir un contenido porque realmente sea mentira, ya que solo se basaría en los patrones de los *datasets* de entrenamiento de los algoritmos. Evitar esto supone ampliar el trabajo en el *fact-checking* semiautomático, ya orientado a la coincidencia con los *claims*.

- También busca este pragmatismo en el SNA. El grafo deja de ser la puesta en escena de la cascada de la publicación más compartida y de sus consideraciones teóricas para atender un enfoque realista y práctico, representando todos los mensajes descargados dentro de la conversación, sin excepción. Así, sirve para conocer cómo discurre el contenido de la desinformación a lo largo del tiempo, para saber cuándo hay más repercusión sobre ella y para atajar los posts que sí la difunden gracias a las variables añadidas con PLN.

1.6. Estructura

La estructura de esta tesis está compuesta de seis capítulos:

- **Capítulo 1**: se introduce la desinformación en el contexto actual de las redes sociales, qué problema supone y cómo las técnicas computacionales se postulan para atajarlo. En esta tesis, se profundiza en esto mediante las preguntas de investigación y objetivos planteados para mitigar esta lacra con los métodos de PLN y SNA.

- **Capítulo 2**: se ofrece el marco teórico sobre estos dos campos de la IA contra la desinformación, explicados después para entender el desarrollo de la tesis. Se detalla su uso individual y la unión de ambos, con el *Machine Learning* (ML) como propulsor de estos.

- **Capítulo 3**: se explica el PLN más allá del enfoque tradicional del etiquetado de posts

como verdaderos o falsos a partir de un *dataset*, gracias a las arquitecturas *Transformer*. Se estudia en la práctica a través de la similitud semántica de la representación de los mensajes como *embeddings*, más cerca o lejos entre sí según la desinformación a la que aludan.

- **Capítulo 4**: se presenta la tarea del NLI en el ámbito de la desinformación y su funcionamiento para ver el grado de vinculación de cada post con una falsedad concreta. Se comprueba en la práctica a partir del análisis de publicaciones tras haberlas separado por NLI en función a su alineación con la información falsa.

- **Capítulo 5**: tras describir las formas de transmisión de la información y las consideraciones en torno a ellas, se finaliza como último paso con el SNA a través de los grafos para trazar las conversaciones sobre la desinformación mediante las características del NLI y de la plataforma social. Se muestra esto a partir de tres casos a modo de experimento.

- **Capítulo 6**: se responden las preguntas de la investigación y se referencian los objetivos planteados en el Capítulo 1. A través de ellos, se plantean también los próximos pasos a seguir y se aportan las principales conclusiones de la tesis.

1.7. Publicaciones

En este apartado se exponen las aportaciones principales y extras de la investigación:

1.7.1. Aportaciones principales

Tres papers, dos de primer autor, son las aportaciones que vertebran esta tesis. A continuación, se exponen los detalles de cada uno de ellos y los capítulos que protagonizan, además de contribuir estos tanto a la cuestión teórica de los Capítulos 1 y 2 como a las consideraciones finales y conclusiones del Capítulo 6 con sus resultados.

- **Paper 1**. Martín, A., Huertas-Tato, J., Huertas-García, Á., Villar-Rodríguez, G., Camacho, D. (2022). FacTeR-Check: Semi-automated fact-checking through semantic similarity and natural language inference. Knowledge-based systems, 251, 109265.

  - **Contribución**. Este paper constituye una de las aportaciones principales a los Capítulos 3 y 5 de esta tesis.

  - **Resumen**. Este estudio propone una metodología para el tratamiento de la desinformación en redes, con los pasos de la descarga de posts para obtener toda la conversación sobre la conversación falsa, la similitud semántica para comprobar la cercanía entre el *claim* con cada post y la inferencia del lenguaje para distinguir de estos cuáles sí enuncian o contradicen el contenido falso.

- **Paper 2**. Villar-Rodríguez, G., Souto-Rico, M., Martín, A. (2022). Virality, only the tip of the iceberg: ways of spread and interaction around COVID-19 misinformation in Twitter. Communication & Society, 239-256.

  - **Contribución**. Este paper constituye una de las aportaciones principales al Capítulo 4 de esta tesis.

- **Resumen**. Esta investigación estudia la forma de disgregar a través de inferencia del lenguaje los contenidos sobre desinformación de la COVID-19 entre aquellos que la difunden, los que expresan lo contrario y los que no tienen que ver. Así, se analizan sus propiedades dentro de la red social y cómo se relacionan en función a sus métricas de la plataforma.

- **Paper 3**. Villar-Rodríguez, G., Huertas-García, Á., Martín, A., Huertas-Tato, J., & Camacho, D. (2025). DisTrack: A New Tool For Semi-automatic Misinformation Tracking in Online Social Networks. Cognitive Computation, 17(1), 1-18.

  - **Contribución**. Este paper constituye la principal aportación al Capítulo 5 de esta tesis.

  - **Resumen**. Esta investigación se centra en el análisis y seguimiento de las conversaciones sobre desinformación mediante la unión de los procesos de búsqueda de palabras claves en redes y PLN para la caracterización del contenido descargado con la generación de grafos del SNA. De esta forma, se identifica la trayectoria de los posts sobre desinformación junto a sus autores a lo largo del tiempo.

1.7.2. Aportaciones extras

Los Capítulos 3, 4 y 5, que concentran los pasos de la tesis y las aportaciones principales de las publicaciones ya citadas, se han podido testar a su vez en dos papers y en una investigación expuesta en un congreso, cuyos detalles se exponen a continuación.

- **Paper 1**. Torregrosa, J., D'Antonio-Maceiras, S., Villar-Rodríguez, G., Hussain, A., Cambria, E., & Camacho, D. (2023). A mixed approach for aggressive political discourse analysis on Twitter. Cognitive computation, 15(2), 440-465.

  - **Contribución**. Este paper añade de forma extra los avances en los métodos planteados en el Capítulo 3 de la tesis para el estudio del discurso político en X.

  - **Resumen**. Esta investigación estudia el tono de la campaña política en las elecciones de Madrid de 2021 en redes a través de una metodología mixta con técnicas cuantitativas de PLN como filtro para las técnicas cualitativas en el análisis de contenido.

- **Paper 2**. Vivo, J. M. N., del Mar Grandío, M., Rodríguez, G. V., Martín, A., & Fernández, D. C. (2023). Desinformación y vacunas en redes: Comportamiento de los bulos en Twitter. Revista Latina de Comunicación Social, (81), 3.

  - **Contribución**. Este paper añade de forma extra los avances en los métodos planteados en el Capítulo 4 de la tesis en el estudio de las cuentas de X.

  - **Resumen**. Este estudio explora las características de los difusores de contenidos de desinformación y de los opuestos a ella a través de la descarga de los posts sobre informaciones falsas antivacunas en redes, su filtrado a través de inferencia del lenguaje y las métricas propias de la plataforma social.

- **Congreso**. Villar-Rodríguez, G., Huertas-Tato, J., Martín, A., Camacho, D. (2021). A la desinformación le gusta la compañía: Representación de bulos de Twitter sobre la COVID-19 mediante *embeddings*. XIV Congreso Español de Metaheurísticas, Algoritmos Evolutivos y Bioinspirados (MAEB 2021).

- **Contribución**. Este congreso constituye una de las aportaciones al Capítulo 3 de esta tesis.

- **Resumen**. Este estudio expone una metodología para descargar los posts sobre desinformación de la COVID-19 y modelarlos en función a la similitud semántica, todo ello mediante la distancia de sus *embeddings* en un espacio bidimensional tras comprimirlos. Una red neuronal profunda ayuda a la mayor separación entre publicaciones sobre distintas falsedades.

Capítulo 2

Estado de la cuestión

Tomar alimentos alcalinos
no tiene efectos demostrados
para combatir el coronavirus.

— Verificat

En esta parte, que integra el marco teórico, se profundiza primero en el asunto de la desinformación (2.1), a través de sus tipos, sus temas tratados y las OSNs que la acogen. Después se recogen las disciplinas computacionales de la tesis en la lucha contra esta lacra: por un lado, el PLN (2.2), con la cuestión transversal del aprendizaje automático, su evolución hasta los *Transformers* y el NLI como subdominio beneficioso para este tipo de tareas; por otro, el SNA, abordando primero la representación de grafos más las dinámicas de la difusión de contenidos (2.3), y después las propiedades de la red cambiantes en el tiempo más el papel de la monitorización (2.4). La última parte muestra la unión de PLN y SNA (2.5) en esta batalla.

2.1. El problema de la desinformación

La desinformación es un problema de interés creciente. Prueba de ello es el aumento de publicaciones científicas sobre este tema, con una subida aún más notable en los últimos años [20, 21, 16, 22], como ya se ha avanzado antes. Es una cuestión global en varios aspectos: por un lado, a nivel territorial, ya que Estados Unidos sigue publicando más papers que el resto de países sobre el asunto, pero cada vez es menos dominante en el Top 20 que más publican; por otro, a nivel temático, pues afecta a todas las disciplinas, y así lo demuestra la variedad de áreas entre las revistas con más publicaciones respecto a ello [20, 21]. Por otro lado, cabe destacar el importante papel que juega la IA tanto en la mitigación de este problema como en su amplificación debido, principalmente, a la generación automática de contenidos [45].

Wang et al. analizaron palabras claves de 5.666 papers publicados hasta 2021. Los resultados ponen en valor los conceptos inherentes a la desinformación en investigación: destaca por encima del resto la palabra '*misinformation*', que engloba todas las formas de desinformación, pero en cuarto y quinto lugar aparecen '*fake news*' y '*disinformation*', mostrando las tres expresiones más utilizadas dentro de este fenómeno. Que la segunda expresión más usada sea '*social media*'

pone de manifiesto el foco en la desinformación como problema asociado a las redes sociales [20]. El estudio de KaabOmeir et al., de 21.407 artículos hasta 2022, refuerza más esto: '*social media*' es la *keyword* que más acompaña a términos relacionados con la desinformación.

La explosión del problema de la información falsa en la actualidad se evidencia también a partir de los términos más utilizados: pese a que los estudios de Wang et al. y de KaabOmeir et al., se remontan a las publicaciones desde 2002 y desde 1971, respectivamente, el coronavirus aparece entre los tres términos más destacados: en el caso de Wang et al., 'COVID-19' es la tercera palabra clave; en el de KaabOmeir et al., se sitúa como la segunda [20, 21]. Estas *keywords* arriba en estas listas de las más utilizadas explican el contexto actual: si bien la información falsa no es un problema nuevo y por eso los autores pueden remontarse a principios de siglo o incluso a antes, las OSNs y la pandemia de 2020 han elevado esta cuestión a otro nivel.

La presencia de 'Twitter' en lo alto de estos compendios [20, 21], además de la apariciones de 'Facebook' y 'YouTube' encuadradas en las OSNs, y de 'vacunación', 'salud pública' o 'pandemia' en el marco del coronavirus [20] evidencian más la dinámica actual del fenómeno. Pero la COVID-19 y estas plataformas sociales no son los únicos signos de la repercusión reciente, sino también las técnicas de mitigación abordadas. Los conceptos de '*machine learning*' y '*deep learning*', propios de la IA, también son ampliamente utilizados [20], mostrando la importancia de la rama computacional en la actualidad para la investigación de esta lacra.

Esta distinción entre temas se aprecia en los *clusters* obtenidos por KaabOmeir et al.: uno sobre los conceptos de desinformación (como '*misinformation*' y '*disinformation*'), otro sobre la alfabetización mediática, otro sobre los factores que afectan a la difusión de los contenidos falsos, aquel con todo lo relacionado con la IA y el que aborda toda la parte de las redes sociales [21]. Si bien las *keywords* más utilizadas ya sugieren el carácter transversal de la desinformación [20, 21], estos *clusters* reflejan más cómo la desinformación se puede enfocar en varias direcciones por las áreas a las que afecta.

En el *cluster* de la IA, aparecen cuestiones como la transparencia, la ética y la seguridad, pero también están presentes las diferentes ramas empleadas contra la desinformación [21]. Las expresiones claves en los estudios cienciométricos sobre las técnicas computacionales más abordadas en esta área recalcan esto a través de sus focos: el *Deep Learning* (DL) y sus arquitecturas, la detección de desinformación, la cuestión de las redes sociales y el PLN [45]. '*Natural language processing*', '*social network*' y '*journalism*', tres de las *keywords* más abundantes entre las ya tratadas [20], son, en particular, las que dibujan el marco de esta tesis.

2.1.1. La desinformación según sus tipos

Zhou y Zafarani elaboraron una distinción entre noticias satíricas, desinformación (como *disinformation* y *misinformation*), *cherry-picking*, *clickbait* y rumores mediante las diferencias según su autenticidad (nivel de factualidad, de pertenencia a los hechos), según su intención (engañar, entretener o indefinida) y según estén en formato de noticias o no [46]. Kapantai et al. hacen una revisión de cómo otros autores han categorizado la información falsa: desde aquellos que engloban los términos más generales y que también separan entre *misinformation* y *disinformation* hasta otros que hacen divisiones específicas con más hincapié en la intencionalidad de estas falsedades [47].

Estas clasificaciones [47, 46] pasan de la teoría a la práctica cuando la IA incorpora a sus tareas distinguir entre un tipo de información u otra. Recopilaciones como las de parodia o sátira entre

los tipos de falsedades [48, 49, 1, 50, 51] han sido tratadas por la IA para intentar detectarlas. También ha sido objeto de estos enfoques computacionales el etiquetado de otros desórdenes como el *clickbait* [51] o el discurso de odio [1].

Sin embargo, dentro de las clasificaciones recogidas por Kapantai et al. [47], la tipología de Parikh y Atrey se sale de estas distinciones para estructurar la desinformación según los tipos de elementos del mensaje. Los autores detallan cómo esta puede estar basada en lo visual, en el usuario, en el conocimiento, en la red y en la postura de la afirmación [52]. En su caso, estas clases están definidas de cara a la tarea de detección mediante la IA, pasando de disgregar las falsedades por cuestiones como la intencionalidad a hacerlo por aquellas que tienen que ver con cómo procesar los canales del mensaje, sus características y sus difusores.

Por tanto, la tipología de la desinformación cambia dependiendo del foco. Ejemplo de ello es la revisión realizada por Bondielli y Marcelloni, que también hacen su distinción entre los tipos de falsedad pero después categorizan la información falsa según sus propiedades para su detección computacional. Recogen, por un lado, los enfoques basados en el contenido (léxico, sintaxis y semántica) y, por otro, aquellos guiados por el contexto (basados en el usuario y en la red) [53]. Mridha et al., que hacen una revisión de los contenidos falsos pero ya en base a la detección con DL, ya no se centran en tales taxonomías, pero sí recopilan las características de los *datasets* usados para las tareas de IA contra la información falsa. Es aquí donde las categorías anteriores siguen estando presentes, pero ahora traducidas para que el aprendizaje automático pueda ejercer las tareas contra la desinformación [54].

En esta distinción de *datasets*, las taxonomías se aprecian sobre todo en la diferencia de etiquetas: por un lado, aquellas para la clasificación binaria (fiable frente a no fiable, rumor frente a no rumor, verdadero frente a falso, falso frente a sátira); por otro, aquellas multietiqueta (según el grado de verdad o el grado de acuerdo). Pero también están presentes en la modalidad (separación entre texto, imagen y los datos de las ramificaciones de la propagación) y el tipo de fuente (noticias, datos de redes sociales, desmentidos de *fact-checkers* o declaraciones políticas) [54]. Es un paso más en la teoría llevada a la práctica para combatir la desinformación en sus diferentes formas de manifestarse.

2.1.2. La desinformación según la temática

Como se ha avanzado antes, la cienciometría también revela los temas de la desinformación más abordados en la literatura, dado que tanto Wang et al. como KaabOmeir et al. recogen la importancia de la COVID-19 en este campo dentro de los términos más utilizados, junto a otras cuestiones sanitarias [20, 21]. Es un síntoma del foco mundial por la pandemia en la época de publicación de estas investigaciones.

Referirse a los temas de la desinformación es aludir a las crisis mundiales que han marcado la agenda de nuestro tiempo. Más de diez mil citas tiene el estudio de Allcott y Gentzkow sobre la desinformación en las redes de las elecciones de Estados Unidos en 2016, que concluyó que, si bien estaba por ver que estas plataformas tuvieran un impacto en los votos, ya conformaban una realidad que necesitaba estudiarse más para entender toda su dimensión [55]. Ya antes de la guerra actual de Rusia en Ucrania se evidenciaron los desórdenes de la información en el conflicto para apelar a la emoción y distorsionar la realidad [56].

Tanto los hechos relacionados con la COVID-19 como los conflictos de Palestina-Israel y de Ucrania-Rusia aparecieron en la encuesta del *Digital News Report* en 2024 por ser temas que

han expuesto información engañosa a los interrogados. Pero en esta encuesta se encuentran también, a nivel general, las cuestiones de política, economía, clima/medio ambiente, inmigración y salud [27]. Por eso, a la hora de hablar de los temas de las falsedades, se debe distinguir entre la desinformación según sus crisis, las antes citadas, y aquella según sus ámbitos.

2.1.2.1. Política

La desinformación política trae consigo un problema extra: la creencia en informaciones falsas es también producto de querer apoyar actitudes o afiliaciones políticas concretas [57]. Estos sesgos partidistas llevan a los individuos a aceptar y a compartir falsedades, y a propagar aquellas que ataquen a los grupos ajenos a su identidad política [58].

La irrupción de los contenidos políticos en las OSNs da lugar a tres aspectos claves: una fragmentación de los medios, con muchos espacios a explorar que pueden tener un contenido de menor calidad; una exposición al resto que, con sus beneficios, también puede derivar a la desinformación; y una oleada de incivismo y negatividad en redes que lleve a la polarización. Como en otros tipos de falsedades, los *bots*, los *trolls*, los conspiracionistas, los medios *mainstream* y las webs de información falsa contribuyen a la difusión, pero en este contexto político también participan los políticos y los gobiernos extranjeros en el proceso [58].

2.1.2.2. Clima

La desinformación sobre el cambio climático va de la mano con el escepticismo y el negacionismo. Como en otro tipo de desinformaciones, la adherencia a las informaciones falsas sobre el clima también se debe a los sesgos y creencias de los usuarios y a las cámaras de eco en las OSNs, a través de grupos e individuos filántropos y expandidas por blogs, políticos y medios [59]. De acuerdo a Dunlap y McRight, esta cámara de eco es una máquina organizada de la negación auspiciada por la industria de los combustibles fósiles, los actores conservadores, las asociaciones en defensa de los intereses corporativos, los científicos opositores (tengan o no que ver con la materia) y los movimientos de *astroturfing* [60].

De acuerdo a Lewandowsky, este escepticismo viene de la oposición a añadir políticas de regulación; de la visión contra la "élite" de los medios, judicatura y voces expertas; de las corrientes de opinión pública y, en consecuencia, de la amenaza ideológica y económica que la reacción al cambio climático supone [61]. Estas oposiciones se manifiestan en cinco proclamas: "no es real", "no es por nosotros", "no es malo", "los expertos no son de fiar" y "las soluciones al clima no funcionan", como oposición a las afirmaciones sobre el cambio climático contrarias a ellas [62].

2.1.2.3. Salud

La desinformación sobre salud representa la oposición al consenso científico sobre una cuestión sanitaria. Para estos casos, la afirmación de algo como verdadero o falso evoluciona conforme la investigación científica encuentra las evidencias para apoyar o rebatir un contenido [63]. La vacunación y las enfermedades infecciosas han sido centro de estos contenidos falsos y, ya antes del coronavirus, la desinformación sobre salud deslegitimaba las vacunas, las evidencias sobre los virus del Ébola o el Zika y también contaminaba los contenidos sobre fumar, el cáncer o la nutrición [64].

La desinformación sobre salud provoca: más errores a la hora de interpretar las evidencias y de localizar las fuentes sanitarias; impacto en la salud mental, y más resistencia a la vacunación [65].

Se demuestra, en general, que las falsedades sobre salud prevalecen más que la información verdadera y que persiguen una narrativa del miedo y la ansiedad [64].

2.1.2.4. Inmigración y minorías

Las minorías culturales, étnicas y religiosas son también objeto de la desinformación. El análisis de la desinformación en Europa muestra, entre otros: la abundancia de contenidos sobre personas migrantes o musulmanas y su exposición como una amenaza cultural, económica y para el orden público; alegatos y conspiraciones antisemitas, y el refuerzo a las tensiones, estereotipos y miedos ya visibles en los Estados miembros [66].

Por ejemplo, el análisis de los desmentidos de los *fact-checkers* españoles categorizados como racistas, islamófobos o xenófobos mostró cómo las acusaciones de violencia, obtención de ayudas e inmigración irregular copaban la desinformación hacia las personas migrantes. Imágenes que los mostraban como un colectivo y no de forma individual y pantallazos a modo de manipulación acompañaban esa narrativa de la otredad como violenta, delincuente y beneficiada por el Estado [67].

2.1.3. La desinformación según la red social

Twitter (X), Facebook, Instagram, TikTok, YouTube y WhatsApp son, junto a LinkedIn y el buscador de Google, los espacios que se han abordado en la encuesta del *Digital News Report* para desglosar el porcentaje de individuos con dificultades para captar noticias fiables [27]. Los resultados, desiguales según la plataforma, muestran que el problema de la desinformación tiene sus particularidades dependiendo del ecosistema social en la que prolifera.

2.1.3.1. Twitter

Los estudios sobre desinformación en Twitter (X) se remontan a antes de las elecciones estadounidenses de 2016, momento en el que más importancia social cobró este problema. Hay ejemplos más antiguos en los mismos contextos, como las elecciones surcoreanas de 2012, sobre los que se han explorado campañas de *astroturfing* [68]. Y no fueron los únicos casos que se estudiaron en estos años: cuestiones como la fiebre del Zika ya evidenciaron antes de la COVID-19 el reto de abordar los rumores en las noticias de salud antes de poder confirmarse o descartarse con evidencias científicas, más la necesidad de contar en cualquier tratamiento automático con el apoyo de las fuentes sanitarias de autoridad [69].

En el auge de papers sobre la desinformación durante la COVID-19 [20, 21], se demostró con aprendizaje automático que las afirmaciones totalmente falsas en Twitter discurrían más rápido que los enunciados parcialmente falsos, además de que los usuarios verificados también eran actores de la desinformación en aquel momento [70]. Esta cuestión es una tendencia genérica [36] y ya se comprobó antes con las falsedades del Ébola con los términos 'Ébola', 'prevención' o 'cura' dentro de los posts [71]. Estos estudios ponen de manifiesto dos focos comunes en la red de *microblogging*: cómo trabajar con el texto de sus posts para buscarlos y procesarlos y cómo obtener *insights* de las cadenas de sus interacciones.

La extracción de posts a través de *keywords* es común en Twitter (X) para recoger las muestras de contenidos a analizar. Este paso ha motivado la creación de herramientas para *dashboards* con análisis de los temas, tendencias y sentimiento generado en torno al tema de la desinformación [72]. Por ejemplo, el uso de *keywords* permitió ver en Twitter a Kouzy et al. que en su

muestra el término 'COVID-19' llevaba consigo menor tasa de desinformación y de contenidos inverificables que '#2019_ncov' y 'corona'. En su recopilación, siete de cada diez posts tenían información sanitaria, y el resto política y financiera. Una cuarta parte contenía desinformación, más por parte de cuentas informales individuales y de grupo y más por aquellas sin verificar [14]. Los hipervínculos, que también sirven como filtro para los análisis, permitieron a Singh et al. ver la poca proporción de publicaciones con enlaces a fuentes sanitarias de referencia. Aunque no era tanto el porcentaje de aquellas asociadas a la desinformación, se reposteaban más [73].

2.1.3.2. Facebook

Como en Twitter (X), en Facebook la desinformación se acoge en forma de posts con enlaces, que los usuarios pueden difundir y sobre los que pueden interactuar. Buchanan y Benson ya vieron en esta red el poder de las cuentas reconocidas, ya que los contenidos, incluidos los falsos, tienen más posibilidad de amplificarse si parten de una fuente de confianza [74]. En cuanto a las interacciones, los comentarios de la interfaz sirven de información sobre la reacción de los usuarios y, en este sentido, Barfar se benefició del análisis textual de estos para varios hallazgos: el auge de incivismo y enfado ante la desinformación política; la falta de pensamiento cognitivo en las falsedades a ambos lados del espectro político, y la mayor ansiedad de los usuarios en sus respuestas a la información verdadera [75].

El partidismo afecta en estas desinformaciones, pero también la unión a otros usuarios cercanos en preferencias. Las cámaras de eco pueden alcanzar grandes dimensiones y, cuando la adhesión es a comunidades que consumen fuentes alternativas, los usuarios son más vulnerables a la desinformación incluso como parodia, al encajar con sus narrativas. El poder de la comunidad en esta plataforma es un reto para combatir contra esta lacra: por un lado, los pocos que interactúan con los desmentidos no hacen que su actividad descienda, sino todo lo contrario; por otro, el sentimiento es más negativo cuando las comunidades científica y conspiratoria interactúan, peor a medida que conversan más [76].

Más allá de la política, también otros tipos de falsedades se adhieren a esta red, como la sanitaria. Johnson et al. observaron cómo entre los artículos más populares sobre cáncer, aquellos con falsedades y contenidos perjudiciales conseguían más *engagement* de los usuarios [77]. En relación a la COVID-19, que también impactó en Facebook, Recuero et al. vieron que los grupos sobre este tema asociados con la derecha afín a Bolsonaro, la religión, las fuentes alternativas y la conspiración fueron los que movieron la desinformación, en conjunto con el efecto de medios hiperpartidistas [78]. En línea con la evidencia de que las falsedades se amplifican mejor en las cascadas que las verdades [36], este estudio muestra cómo el *fact-checking* tiene más difícil llegar a ciertas personas y a los grupos de desinformación para poder atajarla [78].

2.1.3.3. Instagram

También son propios de Instagram los estudios sobre sus *likes*. Por ejemplo, se pudo ver que los posts antivacunas del VPH contenían más *likes* pese a que hubiera más posts provacunas [79]. Este estudio, también en analogía a otras redes sociales, clasificó los contenidos por tipos y descubrió teorías de la conspiración y grupos sobre la supuesta ocultación de los efectos nocivos de la vacuna, así como diferencias entre publicaciones con y sin narrativas personales [79].

Pero propia de Instagram es la naturaleza de sus contenidos: en tal estudio, 32 % de la muestra contenía solo imágenes; otro 30 aproximado, composiciones con texto e imagen sin infografía [79]. Esto se aprecia en la desinformación sobre el coronavirus: también los *hashtags* de la pandemia

sirven para recoger los posts de Instagram y estos muestran más proporción de conspiración y, sobre todo, de desconfianza general, pero específicos de la plataforma fueron los *selfies* y memes como parte de estos contenidos, no necesariamente en relación con estos *hashtags* y los pies de foto [80]. También única a esta red es su interfaz para incitar a la desinformación: se evidenció cómo los usuarios percibían como más creíbles los posts con respaldo de usuarios reputados(*trusted endorsement*), visibles con sus *likes*, frente a aquellos que no [81].

También destacó el texto como acompañamiento de la imagen y los pies cortos de foto en un estudio de 400 contenidos de *influencers* políticos en España, más de la mitad de ellos con desinformación y solo uno de ellos señalado como advertencia de desinformación por Instagram. Por tanto, entender los códigos de Instagram supone comprender la investigación sobre información falsa en ella. Tal estudio también subrayó los *reels* como un espacio de proliferación de falsedades al adaptarse estas al contenido audiovisual a la vanguardia también propio de TikTok, y advirtió de las audiencias más jóvenes de Instagram expuestas a estos posts [82].

2.1.3.4. TikTok

El vídeo impera en TikTok para los análisis sobre su desinformación, y esto pone el foco en los estudios de la multimedialidad en las redes sociales [37, 6]. Por ejemplo, en materias de salud, se mostró que los contenidos audiovisuales sobre primeros auxilios para convulsiones omitían recomendaciones refutadas con evidencias en favor de otras ineficaces y perjudiciales [83]. En el marco de Estados Unidos, también se vio cómo se colaban recomendaciones de abortos caseros, y a la audiencia le costaba discernir si eran desinformación o no [84].

En concreto, la desinformación se beneficia del formato de TikTok. De acuerdo a Alonso-López et al., las falsedades en esta red circulan con más facilidad debido a su carácter fresco y visual y a la sencillez a la hora de compartir el contenido, según el análisis en cuatro países distintos. Comprobaron que la información falsa viene sobre todo de individuos sin afiliación política o institucional, seguidos de cuentas *fake*, que pueden ser conectoras de los contenidos de los partidos políticos. Lleva más fácil al engaño el hecho de que parte de los usuarios individuales no solo publica falsedades sino otros contenidos, ganando credibilidad [85].

En esta red, destacan como forma de información falsa los extractos de vídeos descontextualizados de otras plataformas digitales, acompañados con un texto que influye en la percepción de quienes los ven, además de otros elementos como *selfies* y música también contribuyendo a la desinformación. Estos investigadores vieron también que los contenidos falsos no tenían por qué ser los más visibles, pero la plataforma social juega a favor de la información falsa publicada de forma humorística, aunque se refiera a cuestiones políticas. La polarización, como en su caso analizado de España, también está presente [85].

2.1.3.5. YouTube

La desinformación también campa en forma de vídeos en YouTube. Por ejemplo, una cuarta parte de los 75 vídeos más vistos en marzo de 2020 tanto para la búsqueda con el término 'coronavirus' como para aquella con el término 'COVID-19' reproducía contenidos engañosos, frente a la escasez de fuentes de autoridad en salud en la plataforma [86]. El contenido falso siguió durante la vacunación: de los diez vídeos más vistos sobre la vacuna de la COVID-19, alrededor de uno contradecía los contenidos de la OMS y de los centros de control y prevención [87].

Las falsedades contra la vacunación en formato audiovisual no surgieron a raíz de la COVID-19.

Ya en Italia se demostró antes cómo el tono de la mayoría de vídeos al respecto era negativo [88]. Aunque predominen en otros contextos los posicionamientos a favor de las vacunas, llegar a un vídeo antivacunas era la puerta a la desinformación a partir de las recomendaciones de la plataforma [89]. La salud es terreno pantanoso en esta red más allá de este ámbito, a través de un mensaje que ya no es el textual de Twitter o Facebook. Por ejemplo, de los 150 vídeos de cribado y tratamiento del cáncer de próstata, unas tres cuartas partes contenían desinformación o sesgos [90].

El riesgo es doble: por un lado, con las falsedades cobijadas en los vídeos; por otro, con las recomendaciones que dirigen al usuario a más de ellas. Y todo esto en una red social que ya existía antes del auge del DL actual. Por ejemplo, respecto al vídeo, cuando en 2018 las técnicas de procesamiento de imagen para tratar estos contenidos podrían haber sido insuficientes, se emplearon los metadatos de los vídeos como alternativa para entender cómo funcionaba esta desinformación en las conspiraciones, todo ello a través del análisis del *engagement* y del SNA [91]. En cuanto a las recomendaciones, se descubrió el efecto burbuja de los vídeos sugeridos de cuestiones conspiratorias del 11S, de los *chemtrails*, del terraplanismo, de la llegada del hombre a la luna y de las vacunas. No en función al perfil del usuario (género, edad, localización), pero sí en combinación con su historial de visualización [92].

2.1.3.6. WhatsApp y Telegram

Como se ha observado, la desinformación puede estar en cualquier formato, y todos ellos tienen cabida dentro de las redes de mensajería instantánea. Es por ello que la desinformación en los mensajes de WhatsApp a cualquier persona también ha sido tratada. Se ha descubierto que los adolescentes tienden a compartir el contenido en este espacio más en función a sus intereses que a su fiabilidad. Basta con que el asunto sea noticioso para que lo distribuyan, más allá de su naturaleza [93].

Esta falta de juicio ante los mensajes recibidos supone un problema: en el contexto de las elecciones brasileñas, un 13 % de los hipervínculos procedía de espacios de contenidos maliciosos y erróneos (*junk news*), en perjuicio de menos del 3 % de fuentes políticas profesionales, y un 40 % pertenecía a YouTube. Destaca la diseminación a partir de archivos multimedia y, con ella, la no necesidad de asociar imágenes, vídeo o audio a fuentes de autoridad, apartándose de la apariencia y narrativa como noticias. También se recurre al engaño y al odio para alcanzar la viralidad [94].

La falta de regulación de los mensajes supone un punto crítico en Telegram, que se puede extrapolar a los otros espacios de mensajería instantánea. De un análisis de 200.000 publicaciones de esta plataforma en Estados Unidos se descubrió que los hipervínculos de fuentes no fiables se compartían más que aquellos de medios de comunicación profesionales, pero que recibían menos visitas. Los canales que tenían un alto porcentaje de enlaces no fiables eran más activos, pero estos enlaces se distribuían en menos canales que aquellos que dirigían a medios profesionales. Esto mostró que la desinformación no ocupa tanto el espectro de Telegram, pero que puede agitar más los espacios en los que sí se hace un hueco [95].

Esta lucha entre la información verdadera y la falsa es palpable grupo a grupo. De un canal de Singapur sobre la COVID-19 de más de 10.000 integrantes, se encontró de enero a marzo una actitud crítica de los usuarios con la desinformación, también negando y cuestionando las falsedades desmentidas por fuentes oficiales. Este escepticismo por parte de los individuos sirvió, sin embargo, para detectar en los mensajes que también circulaban mentiras que no habían sido señaladas por las autoridades públicas [96].

2.1.4. *Fact-checkers*

En los 2000 aparecieron las primeras acciones de *fact-checking* a las afirmaciones políticas en Estados Unidos. En Europa, el blog anglosajón de Channel 4 News para las elecciones al parlamento fue la semilla [97]. De acuerdo al estudio de Graves y Cherubini en 2016, tres perfiles definen a los *fact-checkers*, de acuerdo a su encuesta realizada, mezclados en la práctica: sobre todo reporteros (*reporters*), en su misión de informar a la población, pero también reformistas (*reformers*) en su objetivo de incitar a una mejor evolución del modelo político y mediático, y/o expertos, por situarse como fuentes independientes de autoridad fuera de los otros dos perfiles [97], entre otros.

También en 2016, estos autores estudiaron cómo los *fact-checkers* pueden usar varias etiquetas para sus desmentidos: una escala de verdadero a falso (la elegida por la mayoría), categorías generales del error detectado sin escalas o incluso no optar por ninguna categorización al uso. En ese momento, también vieron varios hábitos en la recogida de *claims*: del casi consenso de encontrarlos en los políticos, a las diferencias en cuanto a recogerlos también de medios o de otras voces. El estudio también hizo hincapié en otros aspectos como con qué tipo de medios colaboraban para amplificar su misión o su financiación [97].

Estas estructuras de *claims* y de etiquetas en función al tipo de falsedad o error más sus veredictos permiten la conversión del trabajo de los *fact-checkers* a una base de datos. En 2011, el estudio de Cohen et al. sobre periodismo computacional mostró las ventajas de automatizar los enfoques periodísticos, entre ellos el *fact-checking*, a partir de *datasets* que contrapusieran los *claims* a fuentes de información a través de preguntas automatizadas [98]. En 2014, Vlachos y Riedel propusieron la construcción de un *dataset* a partir de los desmentidos, pero aún de forma superficial, lejos de los actuales *embeddings* [99].

Los *fact-checkers* ya han creado *datasets* que pueden después usarse tanto para análisis como para otras tareas de IA. En el contexto de la COVID-19, 35 organizaciones de *fact-checking* se aliaron para formar LatamChequea Coronavirus[1], un espacio con la desinformación descubierta sobre la pandemia, con la base de datos de los desmentidos de estos *fact-checkers* [100]. Sirvió a nivel de análisis para desvelar el carácter político tras los contenidos de salud en las desinformaciones en Latinoamérica [101], y también el amplio espectro de temas de las falsedades en España, sus canales de distribución y su mayor frecuencia tras declararse el estado de alarma en este país ante tal crisis global [102].

Estas recopilaciones también se han realizado en otros contextos. La *European Fact-Checking Standards Network*[2] ha llevado a cabo esta iniciativa para las elecciones europeas de junio de 2024 (Elections24Check) y para la crisis climática (FACTCRICIS). La guerra de Rusia en Ucrania también trajo consigo a partir de la IFCN la base de datos #Ukrainefacts [103]. El *dataset* LIAR se compuso en 2017 por los *claims* descargados a partir de la Interfaz de Programación de Aplicaciones (API) de la organización estadounidense de *fact-checking* Politifact[3] [104], que se presumió en esa época como el más grande hasta la fecha. Este se postuló como adecuado para tareas de ML y de detección automática, en la línea del resto de *datasets* de entrenamiento con informaciones falsas. Aunque en 2014 ya se expusieron conceptos como el *entailment* o *fact-checking* semiautomático [99], son estos *datasets* dirigidos a las tareas de IA los que abren la puerta a que el trabajo de los *fact-checkers* sea también la materia prima de ella.

[1] `https://chequeado.com/latamcoronavirus/`
[2] `https://efcsn.com/projects/`
[3] `https://www.politifact.com`

2.2. PLN y su aplicación en la lucha contra la desinformación

El texto, en mayor o menor medida, es inherente a la comunicación en las OSNs por ser los espacios donde se concentran este tipo de contenidos entre los usuarios que los comparten. Dentro de la IA, el PLN o *Natural Language Processing* es el área computacional que registra y explota las propiedades de estos textos [105]. En el ámbito de la desinformación, el PLN permite detectar y caracterizar los mensajes con falsedades, con el fin de mitigarlos.

Así, el PLN se encargará del tratamiento computacional del discurso humano en cualquiera de sus formas. Esta disciplina cubre tanto las tareas para análisis y entendimiento, lo que lleva al *Natural Language Understanding* (NLU), como aquellas de generación, conformando la rama de *Natural Language Generation* (NLG) [105, 106]. Fonética, morfología, léxico, sintaxis, semántica, discurso y pragmática, todas ellas propiedades de la lingüística, se tienen en cuenta en el PLN dependiendo de su tarea [105, 106]. Aunque fueron ya citadas por Liddy y Feldman a finales de siglo, en esta disciplina encuentran su alcance actual [107, 108].

Chowdhary referencia tres tipos de trabajos con PLN: análisis de la estructura del texto, análisis del significado y herramientas más análisis basados en conocimiento [105]. Patwardhan et al. van más allá y dividen las tareas actuales según su aplicación unimodal (solo de lenguaje) o multimodal (el lenguaje en interacción con otras fuentes) [109]. A las tareas ya conocidas de predecir, clasificar, traducir o producir texto [105, 106], todas unimodales, se unen las multimodales de transformar texto a imagen o viceversa, es decir, pasando de imágenes a descripciones textuales o a respuestas a preguntas sobre este contenido visual [109].

De nuevo, la cienciometría pone de relieve las cuestiones más tratadas, en este caso en el PLN. Las publicaciones analizadas de 2000 a 2019 por López-Martínez y Sierra generaron cinco *clusters*: uno con los temas derivados de esta disciplina (las referencias sobre todo con la IA, el DL, los *embeddings* de palabras y las redes neuronales), otro más en relación con sus tareas (además del ML, las cuestiones más concretas de las minerías de opinión, texto y datos, más la clasificación y el ámbito de las redes sociales) y los otros tres más enfocados en la extracción de los datos y en la generación de conocimiento a partir de ellos [110]. Cuestiones como la similitud semántica, entre estos *clusters*, son también objeto de esta tesis.

Dentro de las OSNs, el estudio cienciométrico de Sandu et al. sobre minería de texto en las redes pone de relieve la cuestión de la desinformación y de las técnicas de procesamiento de texto. El PLN, el ML, el DL y el análisis de sentimiento salen a la luz entre las *keywords* destacadas de las publicaciones en este ámbito [111]. En esta cienciometría se cuelan los términos relacionados con la COVID-19 y la desinformación [111], mostrando el calado del problema actual de la información falsa y las crisis recientes en esta disciplina. Los autores recalcan tres periodos dentro de su análisis de 2010 a 2023: el interés modesto hasta 2015, la subida notable desde ese año en adelante y el gran interés por los investigadores coincidiendo con la pandemia del coronavirus [111].

2.2.1. PLN y aprendizaje automático

En la desinformación, el PLN entra en juego en la detección de propiedades basadas en el contenido, en concreto, en los elementos textuales de las publicaciones [32]. Ya sea diferenciándolos por sus propiedades léxicas o semánticas [53], por los matices del estilo, por la postura adoptada [52] o por las características psicolingüísticas en conjunto con otras propiedades del discurso [22], esta disciplina ha permitido abordar enfoques computacionales contra las falsedades.

El desarrollo de mejores herramientas que involucran el entendimiento del lenguaje humano ha sido posible gracias a los avances en el ML o aprendizaje automático y, de forma más reciente, en el DL o aprendizaje profundo. El aprendizaje automático es el área de la IA que se centra en el desarrollo de modelos entrenados en conjuntos de datos con el objetivo de realizar una tarea concreta, generalizando a ejemplos aún no abordados [112]. En el PLN, gracias al entrenamiento de los algoritmos a través de extensos *datasets*, de únicamente texto o con etiquetas, el DL permite llevar a cabo tareas avanzadas con gran precisión.

Dentro del aprendizaje automático podemos distinguir entre técnicas supervisadas y no supervisadas. En las supervisadas, se entrenan modelos a partir de instancias o ejemplos etiquetados, que forman los *datasets* de entrenamiento. Esta clase de modelos se puede entrenar tanto para tareas de regresión, donde el objetivo es predecir un valor numérico, como de clasificación, donde el objetivo es etiquetar ejemplos [113] entre diferentes categorías. En el caso de tareas relacionadas con el lenguaje humano, podemos, por ejemplo, tener tareas de clasificación en las que el modelo debe categorizar documentos en distintos tipos o filtrar correos no deseados. Para ello, se parte de *datasets* con ejemplos similares sobre los que entrenar, validar y evaluar el modelo obtenido. El objetivo último es aplicar el modelo a nuevos ejemplos no vistos durante el entrenamiento [114].

En las técnicas no supervisadas, donde no hay ejemplos etiquetados sobre los que entrenar, el objetivo principal es extraer información en forma de patrones de un conjunto de datos o realizar tareas de *clustering*, que consisten en generar grupos de ejemplos que comparten características [113]. Otro caso de técnicas no supervisadas son las de reducción de dimensionalidad. El ejemplo más conocido es *Principal Component Analysis* (PCA) o análisis de componentes principales, procedimiento que reordena la información para comprimirla en un espacio de menos dimensiones basado en la extracción de una serie de componentes principales de la distribución de los datos [113].

La desinformación es uno de los muchos problemas donde se ha aplicado con éxito el ML. El enfoque de aplicación más común implica tratarlo como un problema de aprendizaje supervisado, entrenando un modelo de clasificación en un *dataset* etiquetado de ejemplos, los cuales consisten en piezas de información asociadas a una etiqueta binaria como verdaderas o falsas. No obstante, los matices de la desinformación hacen que un gran número de investigaciones aborden esta tarea como una clasificación multietiqueta, donde cada clase corresponde a una escala de falsedad a verdad de un contenido [32].

Menos habitual es encontrar casos de aprendizaje no supervisado para este tipo de tarea, que ofrece como ventaja el hecho de no depender de la calidad de un *dataset* específico de falsedades ni de sus etiquetas. Sin embargo, dada su difícil eficacia en la caracterización de la desinformación, estos métodos suelen quedar relegados como complemento del aprendizaje supervisado o como engranaje de modelos semisupervisados, en una combinación de ambas técnicas [32].

Dentro de las implementaciones de algoritmos de aprendizaje automático, las más famosa y utilizada es la librería Scikit-learn[4], cuya documentación ofrece un diagrama ilustrativo de las técnicas supervisadas y no supervisadas más importantes (ver Fig. 2.1).

Tanto para tareas de aprendizaje supervisado como no supervisado, se ha propuesto un amplio abanico de algoritmos y técnicas de aprendizaje. Así, dentro de las técnicas supervisadas se puede encontrar desde árboles de decisión, máquinas de soporte vectorial, redes bayesianas o

[4] `https://scikit-learn.org/1.4/tutorial/machine_learning_map/index.html`

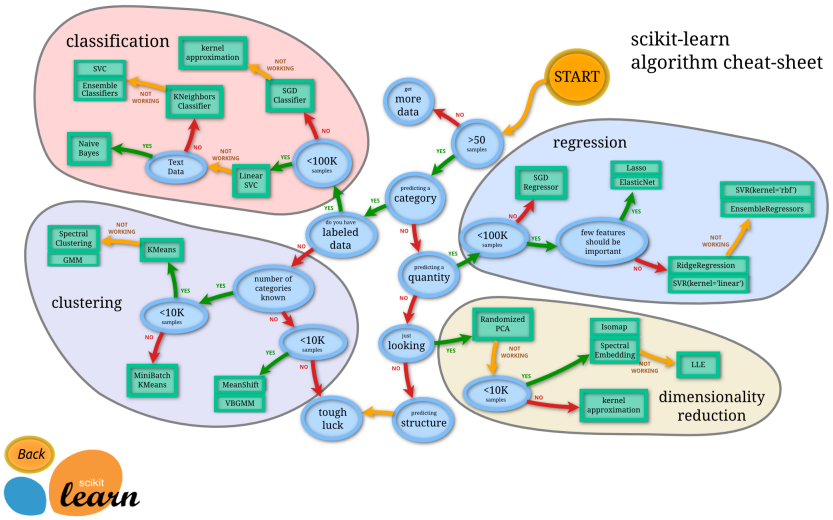


Figura 2.1: Diagrama de la librería Scikit-learn mostrando sus diferentes tipos de algoritmos de aprendizaje automático.

modelos de redes neuronales artificiales, entre muchos otros. Las redes de neuronas artificiales son el paradigma que más investigación y avances ha proporcionado al ML, siendo la semilla del aprendizaje profundo o DL.

2.2.2. Modelos neuronales y no neuronales en tareas de PLN

Dentro de las muchas tareas donde se ha aplicado con éxito el uso de técnicas de aprendizaje automático, se encuentran las relativas al PLN. Numerosas investigaciones han combinado distintas técnicas de ML con texto, tanto modelos clásicos como neuronales, incluyendo tareas relacionadas con la desinformación [115]. Mientras que los primeros enfoques basados en métodos clásicos sentaron las bases de este campo [53], en la actualidad los modelos neuronales predominan en la mayoría de investigaciones.

Al grupo de los no neuronales pertenecen los enfoques tradicionales basados en reglas, estadística, optimización matemática, proximidad y los métodos *ensemble* de estas técnicas. En el campo del PLN, destacan las *Support Vector Machines* (SVM) y *Naïve Bayes* para tareas de clasificación a partir de un *dataset* de entrenamiento para generalizar en casos reales, pero también son importantes aquellos basados en árboles de decisión y regresión logística [115]. Tanto unos como otros se han utilizado extensamente en multitud de investigaciones [116, 113, 114, 53, 117, 118, 112, 119, 2].

Estos modelos clásicos obligan al *feature engineering*, es decir, a un preprocesamiento del texto que permita un mejor tratamiento a través de la segmentación a frases, tokenización de estas frases (conversión a una lista de palabras), eliminación de *stopwords*, supresión de signos de puntuación y otros caracteres, *stemming* (reducción de las palabras a una raíz) y lematización (conversión a la forma base con significado de las palabras) [120]. Tras estas transformaciones, el texto original se transforma en conocimiento interpretable mediante la aplicación de métodos desde el conteo básico de palabras al *term frequency – inverse document frequency* (tf-idf) [121, 122], pasando cuando es necesario por el reconocimiento automático de entidades o *Name Entity Recognition* (NER) dentro de los textos [120].

A los enfoques neuronales [123] pertenecen los modelos bioinspirados en las conexiones entre neuronas. Así, cada neurona se representa de manera computacional como una operación matemática con diferentes entradas a las que se aplican una serie de pesos y una función de activación, y se calcula una salida. Estas neuronas se organizan en capas conectadas entre sí. Su ventaja reside en que todo el conjunto de conexiones construye un modelo no lineal con capacidad de extraer relaciones altamente complejas entre los datos. Estos modelos calculan a través de *back-propagation* un gradiente para ajustar los pesos y los sesgos, que sirve para recalcularlos a través de un algoritmo de optimización hasta minimizar el error [2].

En cuanto a los métodos neuronales aplicados a problemas de PLN, además de las redes compuestas por capas completamente conectadas, también destacan las redes recurrentes o *Recurrent Neural Networks* (RNN), y las redes convolucionales. En la lucha contra la desinformación se han aplicado todos estos enfoques en diferentes estudios [112, 2, 114, 113, 53, 118, 54, 119, 33, 22, 116, 117].

A medida que los modelos neuronales evolucionaban en el ámbito del PLN, las RNN comenzaron a mostrar limitaciones en la captura de dependencias a largo plazo debido al problema del gradiente desaparecido o explosivo [124]. Para mitigar estas deficiencias, se introdujeron variantes como las *Long Short-Term Memory* (LSTM) [125] y las *Gated Recurrent Units* (GRU) [126], que mejoraron el modelado de secuencias largas. Sin embargo, la computación secuencial de estas arquitecturas limitaba su escalabilidad en tareas de gran volumen de datos. Esto llevó a la adopción de modelos más eficientes, como los basados en la atención, culminando en la aparición de los *Transformers* [127], que reemplazaron la recurrencia por mecanismos de autoatención capaces de capturar dependencias a cualquier distancia en la secuencia de entrada.

La llegada de los *embeddings* semánticos hizo posible abandonar en gran medida tareas de *feature engineering*, que implican la pérdida de semántica, adoptando representaciones matemáticas ricas en propiedades contextuales y de significado [2, 128, 129], algo que no es ajeno a la batalla computacional contra la desinformación, tal como muestra la cienciometría en este aspecto [45].

2.2.3. Tratamiento de cadenas de texto mediante modelos de aprendizaje automático

En el ML aplicado al PLN entran en juego los modelos del lenguaje. Dada una secuencia, su misión es predecir los próximos caracteres o palabras. En los modelos clásicos estadísticos, esta predicción se realiza en base a las probabilidades asignadas por la secuencia de palabras, el historial. Formalmente, se computa como la probabilidad de un término x dado un historial h ($P(x|h)$) [2]. En el contexto de la desinformación, ello implicaría resolver, por ejemplo, con qué frases aparece con más frecuencia un término asociado a la información falsa, para comprobar después si estos enunciados necesitan verificación; otro enfoque podría ser el contrario, para ver si una palabra aparece con más probabilidad con frases ya asociadas a falsedades.

La solución más simple a esta tarea es en términos relativos: cuántas veces del historial h aparece ese historial h seguido de x, formalmente $C(hx)/C(h)$, donde hx representa el historial seguido del término. Pero este tratamiento asume que el lenguaje siempre se reproduce de la misma manera. No obstante, una solución más realista, partiendo también de lo básico, sería la regla de la probabilidad en cadena, donde ya no solo entra en juego el término x, sino cada uno de los términos de todo el historial. El término señalado en cuestión será x_n, expresando n la posición, y la probabilidad de que aparezca tras el historial será la probabilidad conjunta de que cada

palabra aparezca en función de las anteriores hasta x_n en cuestión. Matemáticamente esto se traduce a $P(x_1 \dots x_n) = P(x_1)P(x_2 \mid x_1)P(x_3 \mid x_{1:2}) \dots P(x_n \mid x_{1:n-1})$ [2].

Sin embargo, de manera general como en el contexto de los desórdenes de la información, solo sería factible pensar en esta probabilidad simple de $C(hx)/C(h)$ en el caso de los *bots* [3], bajo el supuesto de que serán los únicos usuarios que, de compartir repetidamente una desinformación, lo harán de forma automática con el mismo post exacto. Pero en la práctica, son muchos los actores en el flujo de la información falsa [3], y muchas las formas de expresarla, no a través de la misma secuencia de palabras. Esto también descarta la regla de la cadena de probabilidades, ya que los términos no están sujetos a la misma secuenciación [2]. Tiene más sentido, por tanto, no depender del historial completo, ya sea como un único bloque o en las probabilidades en cadena de sus términos.

Es por ello que en estos métodos tradicionales entran en juego las cadenas de Markov, que asumen que cada elemento depende solo del que precede, y los modelos de n-gramas, que siguen esta asunción de probabilidad y, por tanto, la probabilidad de un término se calcula teniendo en cuenta las n palabras que lo preceden en el historial menos una, la palabra en cuestión. Es decir, los modelos de bigramas computan la probabilidad de un término con el anterior; los trigramas la calcularán en función a los dos anteriores, y así sucesivamente. De manera formal, la probabilidad de una palabra dado el contexto anterior será $P(x_n \mid x_{1:n-1}) \approx P(x_n \mid x_{n-N+1:n-1})$, donde N se refiere a la 'n' de 'n-grama' [2]. En la práctica, esto implica calcular todas las transiciones de un elemento a otro en un texto para que, dado ese elemento, se compruebe cuál es el siguiente con más probabilidad de ocurrir después [129, 130].

Pero en realidad, un elemento no depende solo del anterior. Las redes neuronales recurrentes o RNN permiten un modelado secuencial más avanzado dentro del DL [2], además de lidiar con el problema de las probabilidades a cero y la cuestión de la dimensionalidad [129]. En ellas, no solo depende el elemento anterior de la predicción del contiguo, sino que acumulan toda la información actualizando memoria al deslizarse de un elemento al siguiente. Pero esta actualización hace que la información de los elementos más al principio se desvanezca (desvanecimiento del gradiente). El LSTM resuelve esta cuestión conservando la información más relevante y descartando la que menos en esa actualización [2, 129].

2.2.4. Representaciones semánticas y *embeddings*

El mayor reto en el tratamiento computacional del lenguaje humano textual es conseguir una representación numérica que integre la mayor cantidad de información posible: semántica, lingüística, contextual o pragmática. Las representaciones tradicionales como *Bag-of-Words* (BOW) (ver Tabla 2.1), aunque útiles en ciertos contextos, carecen de estas propiedades, como por ejemplo de la riqueza semántica necesaria para capturar relaciones complejas entre palabras [131]. Este tipo de métodos permite crear representaciones vectoriales de palabras y textos, útiles para entrenar modelos de aprendizaje automático (ver Fig. 2.2). Sin embargo, el rendimiento conseguido es, por lo general, muy limitado.

El principal problema de este tipo de codificaciones reside en que, si un texto contiene una palabra, su sinónimo y su antónimo, cada uno de estos términos formará parte de un recuento diferente. Serán tres sumas distintas en una bolsa de palabras y, por tanto, tres pesos diferentes en relación al resto de términos en el documento en la representación de la bolsa. Más allá de los ajustes que se puedan hacer a estos enfoques para optimizarlos en ausencia de otras técnicas, en

ningún momento de estas operaciones se vincula la cercanía de la palabra con su sinónimo ni la lejanía de estos respecto a su antónimo.

Tabla 2.1: Matriz de la bolsa de palabras [2] de las frases "La mascarilla causa hipoxia" (1), "La comida alcalina cura el coronavirus" (2), "La mascarilla propicia el cáncer" (3), "El bicarbonato mata el coronavirus" (4).

id	alcalina	bicarbonato	causa	comida	coronavirus	cura	cáncer	el	hipoxia	la	mascarilla	mata	propicia
0	0	0	1	0	0	0	0	0	1	1	1	0	0
1	1	0	0	1	1	1	0	1	0	1	0	0	0
2	0	0	0	0	0	0	1	1	0	1	1	0	1
3	0	1	0	0	1	0	0	2	0	0	0	1	0

La concepción de la connotación o significado afectivo para la distribución de palabras en planos diferentes, acuñada por Osgood, conduce hacia este camino donde unas palabras serán distintas a otras respecto a tres dimensiones: la valencia (nivel de agrado del estímulo); la dominancia (control del estímulo) y la activación (intensidad del estímulo) [132]. Se entiende así que una palabra y su sinónimo tendrán estas tres dimensiones con valores iguales o cercanos, mientras que el antónimo diferirá bastante de estos. Si bien las bolsas de palabras suponen la traducción de un texto a los recuentos de sus palabras, los planos de Osgood computarían esta representación vectorial en base a las tres dimensiones sobre las connotaciones [2]. En el contexto de la desinformación, si una palabra y su sinónimo forman parte de textos catalogados como verdaderos y el antónimo se encuentra en aquellos etiquetados como falsos, los clasificadores de ML podrán beber de estas representaciones diferentes para mejores predicciones.

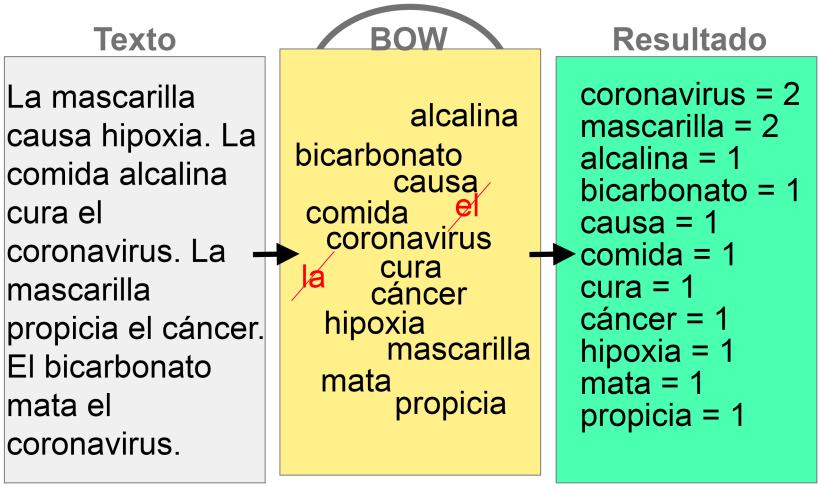


Figura 2.2: Diagrama adaptado a partir de la representación realizada por Jurafsky [2] de la bolsa de palabras.

Esta semántica distribucional se intenta materializar con las matrices término-documento y término-término. En las matrices término-documento, cada documento es un vector y cada uno de sus elementos es el recuento de cada una de las palabras que aparecen. Ya en estas representaciones hay una recreación del espacio vectorial como tal [133] en comparación a los otros documentos y su situación según su cómputo de palabras para que la representación n-dimensional mapee las

relaciones semánticas [2]. Las matrices término-término funcionan igual, pero en este caso cada vector no es un documento sino otra palabra, y los recuentos dentro son el número de veces que cada término aparece en el mismo documento que esa palabra [2].

Tabla 2.2: Matriz tf-idf [2] a partir de las frases "La mascarilla causa hipoxia" (1), "La comida alcalina cura el coronavirus" (2), "La mascarilla propicia el cáncer" (3), "El bicarbonato mata el coronavirus" (4).

id	alcalina	bicarbonato	causa	comida	coronavirus	cura	cáncer	el	hipoxia	la	mascarilla	mata	propicia
0	0.00	0.00	0.57	0.00	0.00	0.00	0.00	0.00	0.57	0.37	0.45	0.00	0.00
1	0.47	0.00	0.00	0.47	0.37	0.47	0.00	0.30	0.00	0.30	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.54	0.34	0.00	0.34	0.43	0.00	0.54
3	0.00	0.49	0.00	0.00	0.38	0.00	0.00	0.62	0.00	0.00	0.00	0.49	0.00

Estos recuentos presentan un problema: las sumas no son en términos relativos. En el ejemplo antes citado de término-documento, una palabra con una frecuencia muy alta a lo largo de los documentos no tendría un verdadero valor que discrimine. Los tf-idf son la solución a esto (ver Tabla 2.2), ya que a estas cuentas del vector (*term frequency*) les contrapone el concepto de *inverse document frequency*, llamado así porque dado el número total de documentos en los que aparece un término (*document frequency*), este término tendrá más valor si aparece en menos documentos, pues es la indicación de que es más distintivo. Matemáticamente, supone la multiplicación de *term frequency* e *inverse document frequency*, con una modificación: por un lado, se aplica el log_{10} a la *term frequency* (además de sumarle 1 para diferenciarla del 0, que sería que no hay coocurrencias), pues se entiende que una palabra con una suma muy elevada no implica una importancia tan grande a tal escala respecto al resto; por otro, también se calcula el log_{10} de la *inverse document frequency* (también para no sobredimensionar ciertos valores), computada como el número total de documentos entre los documentos en los que aparece el término [2].

Para estas frecuencias, el vector va a ser tan grande como el vocabulario y esto hace que sea disperso (con la mayoría de los recuentos de las palabras a cero). Esto cambia con los *embeddings* generados con modelos más complejos, como los modelos de lenguaje, construcciones donde el vector será denso y las dimensiones, muchas menos, no son sobre todo ceros sino valores asignados a los que no se les busca una interpretación, pero que en conjunto representan el significado de la palabra. Esta reducción de dimensiones se traduce en un mejor rendimiento a la hora de clasificar contenidos y de representar el espacio vectorial de los términos como vectores [2]. Este tipo de vectores comenzó con word2vec a partir del entrenamiento de grandes *corpora* [134] y permite la codificación en estas unidades de significado gracias a los métodos de *skip-gram* y la *Continuous Bag-of-Words* (CBOW) [2].

Esta evolución de los vectores no abandona la formalización de la semántica distribucional [121] ni la concepción de Firth [135], sino que la optimiza. Esta vez la codificación numérica de una palabra no depende de los pesos a partir de los recuentos de las palabras, sino de aquellos obtenidos de entrenar un modelo para determinar si el término en cuestión puede aparecer cerca de otra palabra [134, 2]. Esto es posible gracias a la autosupervisión, particular de los modelos del lenguaje porque prescinde del etiquetado a mano para concebir el texto como su señal de supervisión. Es decir, cada término en la secuencia textual guía la tarea de clasificación binaria de si está cerca de la palabra en cuestión [134, 123, 136]. En concreto, el *skip-gram* de word2vec toma los términos afines (los vecinos que están en el mismo contexto de la palabra en cuestión)

y los no afines (aquellos aparte escogidos de forma aleatoria) y a cada uno de ellos les asigna un vector aleatorio, modificado luego para que las palabras en el mismo contexto tengan vectores parecidos entre sí. Un algoritmo devolverá después con la función *sigmoid* un valor cerca de 1 a los términos que crea más afines y cerca de 0 a aquellos que considere menos afines [134, 2].

Con los *word embeddings* se abrió una nueva línea de investigación, con procedimientos similares para nuevos modelos del lenguaje. El surgimiento de fasttext [137] dio solución al vacío de word2vec con las palabras fuera del vocabulario en los *corpora* de entrenamiento y con aquellas que aparecían poco por sus múltiples variaciones en la lengua. Todo ello mediante la descomposición de cada término de n-gramas, para que cada palabra desconocida para el modelo sí tenga representación a través de las partes en las que se haya fragmentado [2]. Por su parte, el nacimiento de GloVe [138] permitió las representaciones basadas en matrices de coocurrencias, por las cuales una palabra en su forma de vector será cercana a otra si sus estadísticas de coocurrencias con otras palabras son parecidas [138, 128, 131].

Gracias a este progreso respecto a los métodos tradicionales de análisis cuantitativo de palabras, las frases similares que se compongan de palabras distintas tendrán una representación parecida. Esto es, en términos matemáticos, muy poca distancia coseno entre ellas y una distribución muy cercana en el espacio latente [139, 2]. Entre otros indicadores, los *embeddings* capturan propiedades sintácticas como temáticas (por ejemplo, nombres ficticios de colegios en películas y series, aunque no sean similares entre sí), coocurrencias de primer nivel (la relación directa con las palabras que suelen estar próximas al término en cuestión) y de segundo nivel (la relación indirecta con las palabras que se dan en vez del término en cuestión en contextos similares), más analogías/relaciones [2].

2.2.4.1. *Transformers* y LLMs

Los *Transformers* [127] suponen un salto en los *word embeddings*: cada palabra tendrá una representación vectorial distinta, en función de cómo le afectan los términos que le rodean. Construcciones como la de word2vec formalizan de una sola manera cada palabra, cuando en verdad esta puede tener un significado u otro dependiendo de la frase en la que esté. En resumen, los *Transformers* evolucionan respecto a los *embeddings* anteriores porque se fijan en el contexto de la palabra en la frase. Mejoran la senda de la semántica distribucional de Harris [121] y, más allá de la traducción matemática de la frase de Firth "*You shall know a word by the company it keeps*" [135], suponen un acercamiento a su noción más cultural del significado de las palabras en base a su colocación, si bien su concepción del contexto de situación va más allá de esto [140].

Los *Transformers* surgen gracias a los mecanismos de autoatención [127], que actualizan los pesos de la palabra para generar el vector en función a la relevancia de las que le rodean en contexto [128, 109, 141]. El punto de partida es el *token*, el término para este tipo de modelos que se convierte en un vector inicial x pero también combina la información de su posición i para tener también en cuenta el lugar de este y del resto de *tokens* en el enunciado. Este vector, formalmente x_i, es el *input* del modelo. La arquitectura devolverá como *output* h_i, siendo h el vector de salida que corresponde a la información semántica con el contexto añadido de x_i. A través de h_i se predice el siguiente término.

En concreto, la transformación de x_i a h_i es posible por el procesamiento en las columnas de bloques, centro de la arquitectura, que contienen: una capa de atención multicabeza para procesar en varias partes las distintas relaciones de la palabra respecto al resto e incluir en la codificación varias formas de dependencias contextuales; redes prealimentadas (*feedforward*) que mejoran esta

formalización del vector procesándolo de forma independiente, y normalización por capas para controlar el aprendizaje. La predicción del siguiente *token* a partir de h_i se hace mediante la cabeza del modelo, con la función *softmax* para obtener las probabilidades de todas las palabras para devolver aquella con la probabilidad más alta [2] (ver Fig. 2.3).

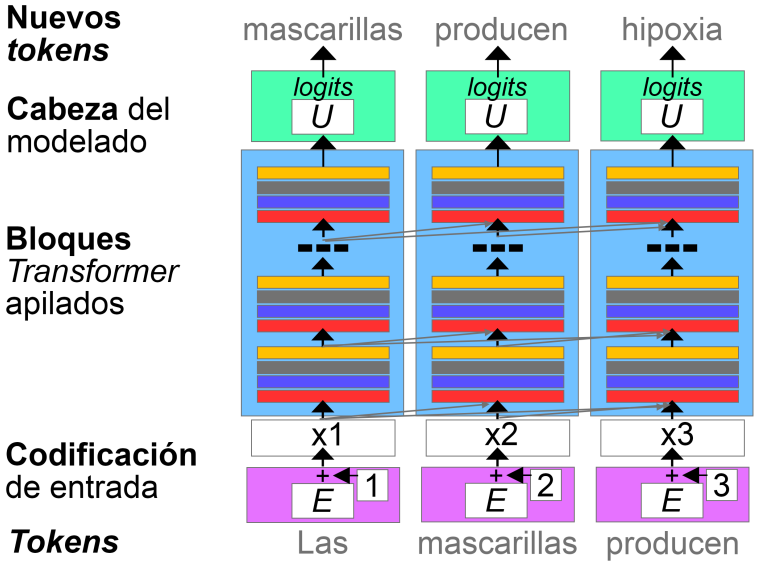


Figura 2.3: Diagrama adaptado a partir de la representación de la arquitectura *Transformer* realizada por Jurafsky [2].

BERT [142] es el modelo característico de esta arquitectura, seguido por otros como ELMo [143], RoBERTa [144], XLM [145] o XLM-RoBERTa [146]. Tareas muy comunes del PLN como la generación de texto [147], la respuesta de preguntas [148] o el análisis del sentimiento [149] explotan su potencial a partir de estos avances, exportados también a otros campos como las tareas relacionadas con imagen, audio, vídeo y aquellas multimodales [150, 151].

Los *Transformers* han demostrado su poder en las tareas contra la información falsa [152, 35]. Tanto es así que han motivado la investigación sobre todos los estudios que, dentro del PLN, utilizan estos procesos para combatir falsedades en las plataformas sociales y, de forma específica, en las redes de *microblogging* [153]. Sus funciones no tienen por qué estar ligadas a la clasificación directa de contenidos en estas redes, ya que sus capacidades permiten también, por ejemplo, la similitud entre enunciados y el cálculo de la relación entre estos [9].

Para este tipo de tareas, los *Transformers* pueden ser preentrenados o ajustados con entrenamiento posterior. Por un lado, con los preentrenados no es necesario ningún *dataset* de entrenamiento, una ventaja cuando no es factible obtener ese conjunto de datos para este cometido. Un proceso más rápido que, sin embargo, va a perder los matices y contexto propio de la jerga de la plataforma social y cuyo enfoque general no está optimizado para la lucha contra la desinformación. Por otro, con los ajustados se gana esa optimización para la tarea, dando mejores resultados, pero ello obliga a altos recursos computacionales, a mucha más especialización para encontrar el mejor ajuste del modelo y a depender del *dataset* diseñado para tal misión [153] (ver Fig. 2.4).

Estos descubrimientos han sido la puerta para los grandes modelos del lenguaje actuales, los *Large Language Models* (LLMs), que, entrenados para adivinar la siguiente palabra, permiten atajar múltiples tareas de PLN. Las actividades relacionadas con el *fact-checking* son también

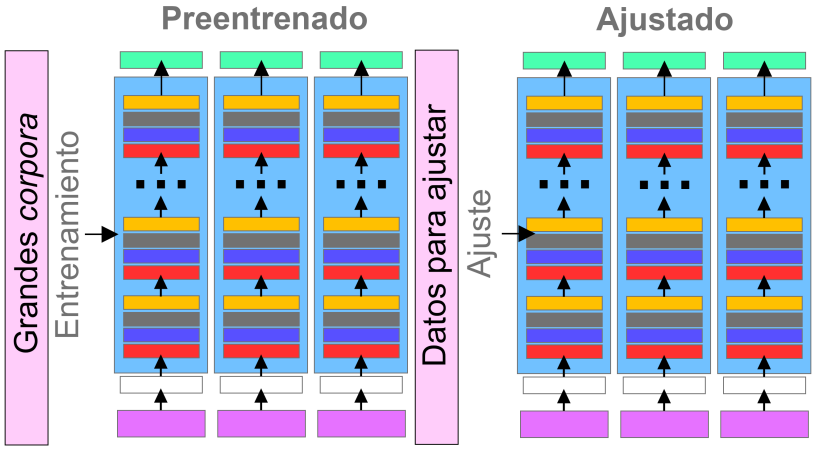


Figura 2.4: Diagrama adaptado a partir de la representación de los modelos del lenguaje preentrenados y después ajustados realizada por Jurafsky [2].

parte de ellas y superan las técnicas tradicionales [33], gracias a ese progreso en la codificación de las características del lenguaje natural como vectores. Los avances a partir de los *Transformers* destacaron en la lucha del PLN contra la desinformación de la crisis del coronavirus [154, 155], también a partir de modelos propios ajustados [156].

2.2.5. *Natural Language Inference*

El NLI, rama del PLN, resuelve la tarea de comprobar si una hipótesis (h) se puede inferir de la premisa (p). Conseguir esto eleva esta área a cuestiones del razonamiento formal y da un paso más dentro del NLU. En otras palabras, esto permite ver qué vínculo hay entre enunciados [11]. Las tres relaciones entre dos frases que se pueden establecer con NLI a partir de los modelos entrenados con los *corpora* para esta tarea son las siguientes: entre A y B se expone el mismo enunciado (*Entailment*); A se refiere a lo opuesto a B (*Contradiction*), y entre A y B no hay relación (*Neutral*) [10, 2, 11] (ver Tabla 2.3).

Tabla 2.3: Ejemplos de *Entailment*, *Contradiction* y *Neutral* dentro de la inferencia del lenguaje [10, 2, 11] para cuatro falsedades, en el contexto de las redes sociales donde se intercalan publicaciones con distinto sentido a partir de las mismas palabras u otras diferentes.

Frase original	***Entailment***	***Contradiction***	***Neutral***
La mascarilla causa hipoxia	Cuidado: el uso de los nasobucos produce hipoxia	Las mascarillas no se relacionan con la aparición de hipoxia	Me quedo sin aire al verte sin mascarilla
La comida alcalina cura el coronavirus	Los alimentos alcalinos eliminan la COVID-19	La COVID-19 no desaparece gracias a los alimentos alcalinos	El cura repartió la comida pese a tener coronavirus
La mascarilla propicia el cáncer	Los barbijos pueden llegar a ser cancerígenos	No es cierto que los tapabocas tengan riesgo de producir cáncer	Los signos cáncer saben ver más allá de las máscaras de la gente
El bicarbonato mata el coronavirus	El bicarbonato es bueno contra el coronavirus	No, el bicarbonato no cura la COVID-19	El bicarbonato mata los olores del desagüe del baño

En la práctica, el NLI puede emplearse para evaluar las respuestas de las interfaces de conversación, eliminar frases que expresen lo mismo para optimizar los resúmenes automáticos, hacer motores de búsqueda basados en el sentido del *input* o comparar traducciones [11]. El éxito en estas actividades avala su uso en la lucha contra la desinformación: también se puede comprobar si dos textos cuentan la misma falsedad.

La tarea del *entailment* textual vino a partir de los desafíos de *Recognizing Textual Entailment* (RTE) [157] para ver si las frases de pares dados estaban vinculadas, a los que años más tarde se añadió la tarea de ver si esa vinculación era de oposición [158]. Una evolución que también trajo muchos más datos de entrenamiento para esta misión. *Datasets* como *Stanford Natural Language Inference Corpus* (SNLI) [159] supusieron un salto respecto a los anteriores datos de referencia por su cantidad de frases emparejadas para ya adivinar las tres relaciones de *Entailment*, *Contradiction* y *Neutral* [158].

Pese al avance en el razonamiento formal, el reto del NLI tiene que ver con el léxico, el significado y la riqueza lingüística de las oraciones a enfrentar [11]. Es, por ejemplo, resolver que dos enunciados están expresando lo mismo pese a que son paráfrasis muy distintas. Los primeros métodos con BOW [121, 122], sensibles a estas limitaciones, y sustituidos por otros más refinados [11], encontraron su evolución en las soluciones neuronales y *embeddings* semánticos [158], como el resto de áreas de PLN.

Por eso, ahora los modelos actuales de NLI se sustentan en el multilingüismo y en el estado del arte del PLN, que incorporan estos últimos avances técnicos de DL y *embeddings*. Destacan modelos como *Multi-Genre Natural Language Inference* (MultiNLI), a modo de mejora de SNLI a partir de descripciones de imágenes [160]; *Cross-lingual Natural Language Inference corpus* (XNLI), a través de un *corpus* de pares en 15 idiomas [161], o *SILT*, en aras de explotar más la arquitectura *Transformer* [34]. El paso más allá de estas propuestas para alinear frases en idiomas diferentes y expresadas de manera muy distinta es el que permite salir de la concepción tradicional de la clasificación de textos dentro de la desinformación para encontrar otros enfoques.

2.2.5.1. *Fact-checking* semiautomático

Frente a los modelos tradicionales, aparece el concepto del *fact-checking* automático, donde también se cuestiona el grado de confianza del ML en función de su aplicación [9]. Se contrapone la imposibilidad de automatizar todo por las competencias humanas que requiere el proceso de verificación frente a la rapidez de la desinformación en comparación a sus desmentidos, precisamente por ese tiempo necesario para la calidad del *fact-checking* [9, 32]. Pero esta aplicación computacional no tiene por qué equivaler a una completa dependencia del uso tradicional del aprendizaje automático. Por esto, las propuestas en investigación sobre *fact-checking* automático avanzan hacia el *fact-checking* apoyado por la IA (en inglés, *AI-supported fact-checking*) [32]

En el *survey* desarrollado por Zeng et al. [42], centran esta disciplina en los *claims*, el término que se refiere a los enunciados susceptibles de ser analizados. La selección de cuál texto es *claim* y cuál no corresponde a los *fact-checkers*, pero en la actualidad los métodos computacionales también se presentan como una ayuda al respecto.

La actividad del *fact-checking* apoyado por la IA en relación a esta clase de enunciados se divide en dos fases: por un lado, un primer paso de *claim detection*, es decir, el filtrado a través de ML de las frases que son susceptibles de someterse al proceso de verificación, a las que además se les puede dar un orden de prioridad para facilitar más el trabajo del *fact-checking*; por otro lado,

un segundo paso de relacionar los *claims*, es decir, si un enunciado corresponde a un contenido ya verificado [42, 9].

Este segundo paso de *claim matching* también se asocia al *claim verification*, es decir, a vincular las afirmaciones con *claims* ya chequeados o con otras bases de datos. La diferencia es solo conceptual, ya que para estas definiciones se consideran *claims* tanto el enunciado a verificar como el ya verificado que se emplea para emparejarlo [43]. Dado que el *claim* es una unidad textual, las tareas de aprendizaje automático para asistir al *fact-checking* van a estar ligadas al PLN [43].

Los enfoques tanto para la detección como para la verificación de *claims* se apoyan por tanto en modelos de IA, operando a la par, sin permitir que se llegue a desmentir directamente un enunciado sin ninguna información previa [42]. Bajo esta perspectiva, el aprendizaje automático se sitúa al principio de la cadena de trabajo para facilitar la detección de aquellas frases que necesitan *fact-checking*. De involucrarse en la verificación como tal, el modelo más fiable es confrontar estas frases contra un contenido ya desmentido, y no desmentir automáticamente desde cero.

La asignación de unas tareas concretas al aprendizaje automático dentro de todo el proceso de verificación abre la puerta a solucionar el que es, según Ruffo et al., un paradigma en este campo periodístico: los costes, precisión y explicabilidad elevados de los *fact-checkers* frente a los bajos de automatizar esta tarea [162]. La solución a esto tiene que ver con la IA *human-in-the-loop*, que deja para los humanos las tareas más difíciles y reserva para el aprendizaje automático otras actividades de procesamiento de datos [162]. Demartini et al. abrazan este concepto y ponen en valor el equilibrio de fuerzas de la computación, de *fact-checkers* y de otros colaboradores para el etiquetado de afirmaciones dependiendo del coste, el tiempo y la dificultad disponibles [163].

En esta concepción de la IA como aliada más que como motor único del proceso, nace el *fact-checking* semiautomático. Aquí los *claims* salen de los desmentidos de los *fact-checkers* y ya a través de ellos sí se pueden encontrar otros enunciados en redes sociales y aplicarles la IA para relacionarlos con estas informaciones falsas ya chequeadas. No solo el *fact-checker* forma parte de unas tareas dentro de la verificación, sino que sigue siendo el corazón de ella para que los procesos de IA partan de esa ventaja de la calidad periodística que el *fact-checking* ofrece [32].

Otra posible aplicación de los métodos de aprendizaje automático es en la recogida de evidencias como información complementaria para determinar o no una falsedad, donde ya se pasa de información verdadera o falsa a contenidos apoyados o refutados por estas pruebas [164]. Esto abre la puerta al estudio de cómo generar las justificaciones de estas decisiones asistidas [164], terreno donde muchas investigaciones se están centrando en los últimos años [165].

Con todo, el aprendizaje automático muestra un amplio abanico de tareas en la lucha contra la desinformación sin la necesidad de acaparar toda la fase humana. Montoro-Montarroso et al. enumeran siete funciones a lo largo de todo el flujo de trabajo, sin absorber enteramente la calidad del *fact-checking* manual de principio a fin: identificar y rastrear cuál es cada *claim*; medir después su relevancia para utilizarlo o no; extraer evidencias para ayudar en la verificación; chequear directamente si el enunciado se ha verificado antes; clasificar este contenido según su nivel de falsedad, y, como pasos finales, diseminar todo el proceso y ganar rapidez documentándolo [32].

2.2.6. Enfoques actuales de PLN para luchar contra la desinformación

En los apartados anteriores se ha descrito la evolución del ML al DL en general y, de forma específica, en la lucha contra la desinformación. Y se revela a su vez el modo tradicional de combatirla: con el avance en los algoritmos para que, dado un enunciado como entrada, cada vez su clasificación como información verdadera o falsa sea mejor. Los enfoques clásicos son, por tanto, una travesía desde los métodos más básicos hasta los últimos modelos que superan los límites de predicción [119, 115, 116, 118, 54]. Esta carrera por la mejor clasificación se evidencia con los *Transformers*, enfrentando los métodos preentrenados que van saliendo con aquellos ajustados y con los ensamblajes de estos [33, 166, 167].

Fuera del enfoque tradicional, la inferencia del lenguaje es la tarea de PLN que mejor representa un proceso de verificación de forma computacional, contrastando un enunciado contra una base de datos verificados, como ya se ha explicado antes. Ya retos como el WSDM 2019 *Fake News Classification challenge* expusieron la necesidad de alinear una información falsa con otro contenido para comprobar si mostraba acuerdo, desacuerdo o ninguna relación con él, un proceso para determinar los tres tipos de vínculos equivalente a NLI [168]. Todo ello ya en el contexto actual de los modelos *Transformer*.

Los trabajos relativos a la aplicación de NLI en este terreno muestran la necesidad de *datasets* de entrenamiento específicos, compuestos de listados de falsedades sobre los que alinear los enunciados en función de las tres categorías mencionadas. Sadeghi et al. hacen una comparación de los métodos tradicionales con los *datasets* de FakeNewsNet y LIAR frente a aquellos con NLI, donde estos registros pasan a ser las hipótesis mientras que los contenidos de información fiable extraída de Politifact son las premisas. Con el enfoque de NLI encontraron mejores resultados para la mayoría de métodos [169].

Shah et al. también rompieron de forma explícita la dependencia en los modelos y *datasets* de ML tradicionales, a partir del sistema VERITAS-NLI. También mediante el conjunto de datos LIAR, tomaron titulares de fuentes fiables (las frases verdaderas) y modificaciones espurias de estos tras alterarlos con modelos del lenguaje (las frases falsas). Así, se enfrentaron estos titulares (la hipótesis) con los resultados de comprobar de forma computacional tales frases en Google (las premisas). El camino del NLI les otorgó a partir de uno de sus modelos sin reentrenar mejor resultado que las soluciones de clasificación automática del *dataset* [170].

En la investigación de Arana-Catania et al., se dio un paso más sobre qué contenidos emparejar con los *claims* para realizar la tarea de NLI. Por un lado, construyeron el *dataset* PANACEA con los enunciados de fuentes de datos curadas y validadas sobre la COVID-19 para etiquetarlos como verdaderos o falsos. Por otro lado, recopilaron los artículos de fuentes especializadas de autoridad y los asociaron a cada enunciado en orden de importancia, calculado por modelos del lenguaje. De los artículos más importantes, se extrajeron sus frases más relevantes a partir de su similitud con la afirmación asociada, en base a la distancia coseno (que también sirvió para evitar enunciados duplicados). Son los emparejamientos con estos extractos los que se testan para comprobar con NLI si las afirmaciones de cada par son verdaderas o falsas, y los resultados se equipararon a los modelos del estado del arte en la tarea de verificación [171].

2.2.6.1. Lecciones aprendidas de estos enfoques de PLN

En retos como el WSDM 2019 *Fake News Classification challenge*, que expusieron en su momento la eficacia de NLI, se han sugerido como trabajos futuros la mejora de modelos y el aumento

de datos [168] para mejorar la clasificación. Pero más allá de estos retos, la propia definición de la tarea revela una desventaja: ya están hechos los pares de enunciados para testar, uno como premisa y otro como hipótesis. Sin embargo, en la práctica esto no sucede así y son necesarias soluciones como la distancia coseno o el *reranking* para formar las parejas de frases. Tampoco supone una puesta en práctica ya en un entorno real.

En el paso más allá de Sadeghi et al. con la creación de un *dataset* de *claims* para encontrar los enunciados a alinear con las falsedades, se dieron cuenta de que no todas ellas tenían ya un desmentido por parte de los verificadores. Como esta contribución con el *dataset* no hace por sí sola el emparejamiento entre premisas e hipótesis, ya plantearon como trabajo futuro el paso de encontrar contenidos similares [169]. Pero de nuevo, no hay un testeo real a partir de los posts descargados de las OSNs, sino que las alineaciones son entre los *claims* descargados de Politifact más los *datasets* de desinformación de referencia.

Aunque estos trabajos se salen de la caja de los métodos tradicionales de clasificación de texto, este conflicto con la puesta en práctica está presente de alguna manera. En el caso de Shah et al., su sistema VERITAS-NLI sí lleva la cuestión a los resultados que devuelve Google, pero sigue recogiendo las afirmaciones de los *datasets* de desinformación [170] en vez de pasar directamente a las informaciones falsas alojadas online. En consecuencia, estas investigaciones plasman los pasos posibles para mitigar la desinformación pero no cuentan con los procedimientos para obtenerla como materia prima si no es a través de conjuntos de datos ya creados.

Los pasos combinados de NLI, colocación de los contenidos por importancia y similitud semántica llevados a cabo en PANACEA suponen un paso adelante: de la maraña de contenidos, se filtran los enunciados para enlazar con las fuentes expertas y después alinearlos [171]. En el contexto de las OSNs, los mensajes pueden ser todos los descargados de la plataforma social escogida para aplicar los mismos procedimientos. En todo caso, los autores advierten de la dificultad de abordar los enunciados falsos cuando son variaciones de otros similares pero con partes de la oración cambiadas (por ejemplo, los nombres propios de las personas que protagonizan la falsedad) [171].

Más allá de las consideraciones propias de cada enfoque de PLN, se aprecian tres retos comunes: primero, la necesidad de asociar las hipótesis con las premisas para saber con qué *claim* comprobar si las afirmaciones son verdaderas o falsas; segundo, el paso a la acción para paliar la desinformación en contextos y crisis específicas, más allá de las aproximaciones con fuentes de datos ya dadas sobre la COVID-19 o similares; por último, librar tales batallas en las OSNs, el ecosistema donde estas falsedades proliferan.

2.3. SNA y su aplicación en la lucha contra la desinformación

El SNA o Análisis de Redes Sociales es la disciplina que tiene como meta el estudio y modelado de las relaciones que se establecen en una red social [172]. Concibe al individuo u objeto a analizar como parte de una estructura y en función a su relación con el resto. Asume las interacciones entre ellos como las conexiones y, en consecuencia, permite plasmar esta concepción estructural como una red [173, 174].

La sociología y la antropología abrieron la puerta al estudio de las relaciones hasta acercarse a la metáfora de la red [173]. Tras los avances llevados a cabo en esta área, el desarrollo de la informática y de las redes sociales ha situado al SNA como instrumento relevante para el análisis

en varios campos. Estas plataformas online han provocado que las unidades a analizar pasen de pequeñas muestras conectadas a millones de ellas [172].

La cienciometría sobre las redes sociales revela que sus datos no son solo cosa de la componente semántica o textual. En el estudio de Esfahani et al. de *big data* y *social media* se aprecian, además de las *keywords* generales, cuestiones específicas del texto (PLN y, dentro de él, análisis de sentimiento o semántica, entre otros) pero también aquellas relacionadas con las métricas de las OSNs en sí (sobre las plataformas, el *social media analytics* y la diseminación de información). Además, no son solo protagonistas en este ámbito las palabras claves relacionadas con la extracción y procesado de datos en la IA, sino también aquellas referentes a su visualización [175]. Esto manifiesta la importancia del SNA, de manera individual y en combinación con el PLN, para la investigación de estos ecosistemas sociales, y pone en valor la representación del flujo de la información entre usuarios en grafos dentro de esa área de visualización estudiada.

2.3.1. Aplicación general del SNA en contextos sociales

Cuando el estudio cienciométrico es sobre el SNA, se evidencia la multitud de aplicaciones en el área. Camacho et al. distinguen la salud (exposición a malas conductas de los usuarios a partir de su comportamiento, comunidades y contenido recibido en OSNs), el marketing (boca a boca, comportamiento de los usuarios sobre una marca y beneficios esperados de esta mediante las OSNs), el turismo/hospedaje (análisis de las valoraciones y momentos compartidos en las estancias) y la ciberseguridad (patrones de delitos, comportamiento de sus perpetradores y formas de contraataque) [37]. Singh et al. también destacan los ámbitos de la educación (descubrimientos de fuentes expertas, de rendimiento escolar, de la actividad de la mujer en la ciencia o de comunidades de interés), deportes (matrices de pases, conexiones de jugadores, rankings de rendimiento), cultura/sociedad (papel de los personajes en literatura y cine, relaciones entre ellos, palabras claves e influencias en medios), gestión de desastres (toma de decisiones y control de efectos en momentos de crisis) o transportes (mitigación de accidentes, congestión o retrasos) [6].

Pero Camacho et al. también desvelan los nichos con cada vez más interés del SNA: política (análisis del discurso de odio, de las opiniones divididas, del rastreo de comunidades y del tipo de información que se comparte), el multimedia (traspaso de tareas más asociadas al texto a imagen, audio y vídeo para su máximo aprovechamiento, más nuevos enfoques) y la detección de desinformación (entramado de los posts o de los autores sobre los que gira el contenido falso) [37]. Multimedia, política y desinformación también vienen incluidos en la investigación de Singh et al., pero estas dos últimas las fusiona en un único dominio [6].

Estos dos estudios aluden a la multidisciplinariedad de perfiles para abordar las aplicaciones del SNA en diferentes dominios [37, 6]. Por eso, en la práctica, estos campos del SNA pueden solaparse. Por ejemplo, los ejemplos citados sobre los contenidos de la COVID-19 para la salud son, a su vez, parte del área de la desinformación. Esto demuestra también cómo las falsedades no tienen por qué circunscribirse solo en la política, clasificación de Singh et al., sino que abarcan más temas. En cuanto al estudio del multimedia que ambas investigaciones destacan [37, 6], el dominio de la imagen y el vídeo en redes como Instagram o TikTok hace que sea cual sea la aplicación del SNA, la exploración multimedia forme ya parte del proceso.

2.3.2. Relevancia del SNA en el estudio de la desinformación

Así como todo tratamiento computacional a los contenidos falsos del texto se relaciona con el área del PLN, todo estudio de las OSNs en las que habitan estas falsedades alude directamente

al SNA, a partir de las métricas de estas plataformas [174]. La red social supone el sitio en el que se amplifica la desinformación y los vínculos entre usuarios construyen el círculo que afecta con los contenidos falsos a más personas dentro de la plataforma. Es papel del SNA diseccionar las propiedades de estos espacios online ante tales falsedades y aquellas de sus actores [12, 3, 46, 36], tomando en cuenta los enfoques orientados al usuario, a la red, al tipo de post y a su propagación además de la parte de PLN [52, 53, 176].

La lucha contra la desinformación no solo consiste en la detección de falsedades, sino en la mitigación de su propagación en las plataformas sociales. Es aquí donde adquiere protagonismo el SNA como disciplina para estudiarlas en este ecosistema. El SNA asume tres condiciones: primero, la estructura de relaciones entre los actores de una red importa más para su estudio que sus características individuales; segundo, el impacto de las conexiones entre estos actores modela su comportamiento, y, tercero, estas uniones son dinámicas y cambiantes [174].

Dentro de la ingeniería, es tal la magnitud de la desinformación que se considera un problema similar al de todo procesamiento de datos actual. Implica superar las tres V de todo conjunto de datos: el volumen de contenido, la velocidad de procesarlo todo y la variedad [177, 174]. El SNA ofrece las formas de atajar el problema, también en paralelo a toda rama en relación a los datos. En un estudio de cinco años de papers en esta disciplina, Camacho et al. identificaron cuatro enfoques para su aplicación: el descubrimiento de patrones (qué se puede hallar), la integración y fusión de la información (qué se puede aglutinar), la escalabilidad (hasta dónde se puede llegar) y la visualización (qué se puede enseñar) [37].

En las OSNs la desinformación parte en una situación de ventaja por naturaleza por ser su nido de amplificación. Ruffo et al. enumeran sus cuatro fortalezas dentro de las redes como caldo de cultivo [162]: la información falsa es más atractiva por aclamar más a la novedad, a la emoción y a la fácil comprensión; puede estar más expuesta en cuanto se repita una y otra vez hasta su amplificación; apela a los sesgos sociales, moviendo a las personas en base la opinión pública y de la reacción de sus grupos y comunidades, y también llama a los sesgos individuales, en base a su predisposición respecto a los hechos enunciados [162].

Esta problemática tanto en las falsedades difundidas como en quienes las extienden pueden estudiarse a partir de los dos tipos de análisis que ofrecen los enfoques del SNA: por un lado, los basados en la estructura (propiedades de las uniones entre usuarios), a partir de la teoría de grafos con métricas para entender la relevancia de determinadas cuentas o de la red en su totalidad, y permiten identificar aspectos como las comunidades de un determinado tema o su viralidad; por otro lado, los basados en el contenido, para convertir la información desestructurada en datos ordenados e inteligibles a partir de un grafo y así generar conocimiento [32].

En consecuencia, el SNA en el ámbito de la desinformación es protagonista en la detección de las propiedades centradas en el contexto de los perfiles de la red social [32, 53]. El tratamiento de los datos de las OSNs permite explotar tres escenarios: el contexto de los usuarios, el contexto de los mensajes y el contexto de la estructura general de la red [32].

2.3.3. El SNA a través de los grafos

El grafo es una conversión matemática de una red y sus conexiones. Los grafos $G(V, E)$ se componen de nodos V y sus enlaces E, que son las aristas que hacen de conexión. Los nodos pueden verse como entidades (personas, animales, países, etc), y determinarán el tipo de enlaces a estudiar [178, 37, 4, 179]. En el caso de la desinformación, estas entidades pueden representar a

las personas a las que afecta o puede afectar, o bien a cada uno de los contenidos vertidos sobre esta falsedad. En el contexto de las OSNs, los nodos pueden ser, por tanto, los usuarios de estas personas o los posts respecto a esta información falsa.

Las aristas se clasifican en: coocurrencias, que unen nodos por algún elemento común (involucramiento en grupos, similitudes, distancias); relaciones sociales, que forman lazos familiares, afectivos y otros niveles de vinculación; interacciones, que enlazan los nodos que realizan una misma acción (actividades, transacciones), y flujos, que trazan los productos de esas interacciones (paso de información, infecciones y otros) [4].

Esto también se puede materializar en el contexto de las OSNs. En este caso, los elementos involucrados en el grafo son las cuentas de la plataforma social y/o los posts sobre los que ellos actúan. La literatura demuestra varias estructuras para representar las relaciones de individuos y contenidos y todas ellas caben en el marco del SNA y, en concreto, de la difusión de la información y sus desórdenes [3]. De manera general, se distinguen entre redes homogéneas y heterogéneas [3] (ver Fig. 2.5).

En su estudio para el control en las OSNs y la desinformación, Shu et al. presentan las redes homogéneas, aquellas con un mismo tipo de nodo y de unión, y heterogéneas, aquellas con varios tipos. Por un lado, en las redes homogéneas, consideran que los usuarios de las plataformas sociales se pueden organizar de manera genérica en grafos de redes de amistad (aristas como unión social) y de difusión (aristas como paso de información a un tiempo y probabilidad determinadas), mientras que los posts pueden plasmarse como grafos de redes de credibilidad (como su nombre indica, aristas según cuán creíbles son los contenidos) [3]. Los grafos de amistad también los mencionan Tabassum et al., y concretan en los enlaces sociales de estas plataformas online poniendo como ejemplos los grafos de seguidores y aquellos de preferencias similares entre usuarios [179].

Por otro, en las redes heterogéneas, Shu et al. explican cómo los grafos pueden ser redes de conocimiento (aristas para marcar los tipos de relaciones entre contenidos), de postura (aristas de los posts a las noticias relacionadas para indicar su posicionamiento al respecto) y de interacción (uniones de los posts a otros usuarios porque los difunden por sus interacciones) [3]. La interacción y la difusión también la detallan Tabassum et al. en sus ejemplos, e incluso otros a los que se refieren como las redes de citación o de coautoría, en el contexto de artículos publicados, también se pueden extrapolar a las OSNs [179].

Dentro del ámbito de la desinformación, se pueden plasmar como redes homogéneas las relaciones sociales entre los miembros de las comunidades que difunden las falsedades; cómo de fácil tienen su expansión entre los usuarios en comparación con las afirmaciones verdaderas, y hasta qué punto estos creen un tipo de contenido u otro cuando lo reciben. También como redes heterogéneas se pueden ilustrar los tipos de relaciones entre los contenidos falsos; el ecosistema de propagadores, falsedades y fuentes de información que tienen una visión positiva o negativa de ellas, y los saltos que estos enunciados falsos realizan de sus creadores a los siguientes usuarios a través de los *likes* o compartidos. En todo caso, estas estructuras básicas [179, 3] muestran que puede haber muchas formas de mapear los fenómenos de las OSNs en general, y los desórdenes de la información en estas plataformas en particular.

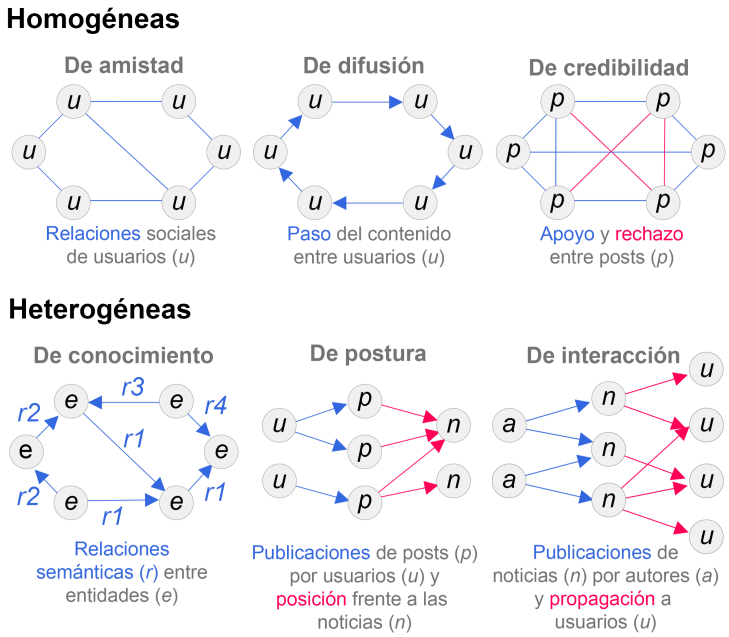


Figura 2.5: Diagrama con la interpretación de las redes homogéneas y heterogéneas a partir del trabajo de Shu et al. [3].

2.3.3.1. Grafos dirigidos y no dirigidos

Los grafos se dividen en dirigidos y no dirigidos. Como el nombre indica, en los grafos dirigidos las conexiones E llevan una dirección y se entienden como flechas de un nodo que desemboca a otro. En los grafos no dirigidos, cada nodo está al mismo nivel y es un vínculo que no concibe esa direccionalidad [4] (ver Fig. 2.6).

Formalmente, un grafo con tres nodos $V = \{A, B, C\}$ será dirigido si las conexiones van de unos pares a otros. Por ejemplo, $E = \{(A, B), (B, C)\}$, yendo de A a B y de B a C. Esto no significa que las conexiones no puedan ser recíprocas, es decir, por ejemplo (B, A) en una conversación en la que A dice algo a B pero también B dice algo a A. No sucede esto en los grafos no dirigidos, donde la conexión es un lazo entre A y B y no una interacción o una consecuencia de A a B [4].

En el campo de la desinformación, en un grafo dirigido el nodo A puede trasladar al nodo B una falsedad, y B contestarle a A con un el desmentido de un *fact-checker*. En un grafo no dirigido, A y B pueden estar conectados si forman parte de la misma comunidad, y aquí el nexo entre ellos representa esa relación social. En el contexto de las OSNs, B y A pueden recibir esa falsedad del contrario a través de sus publicaciones y comentarios (grafo dirigido) y que B y A sean parte de la misma comunidad de seguidores (grafo no dirigido).

Por esto mismo, Shu et al. distinguen dentro de la red de la información, y por tanto también de la desinformación, la dimensión del contenido (el 'qué') y la dimensión social (el 'quién') [3]. Mientras que la primera dimensión expone la información como el amalgama de las noticias de los medios, los posts en redes sociales derivados y los comentarios al respecto, la segunda hace alusión a las relaciones entre los creadores, los difusores y los consumidores [3].

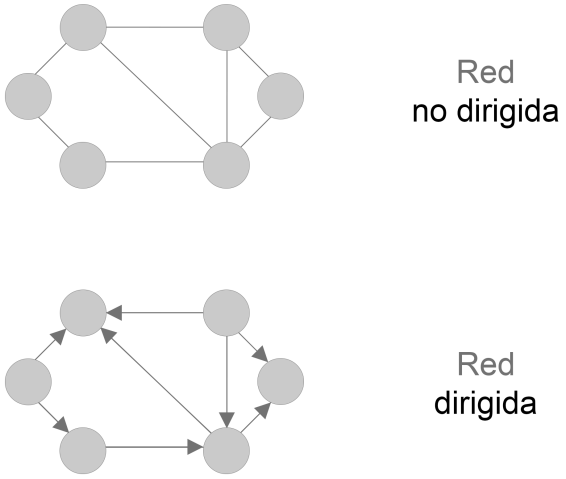


Figura 2.6: Diagrama con la interpretación de las redes no dirigidas y dirigidas a partir del trabajo de Borgatti et al. [4].

2.3.3.2. Grafos no ponderados y ponderados

En la práctica, unas conexiones destacan más que otras en las relaciones. Por ejemplo, de cara al contagio por una información falsa como si fuera un virus, no son iguales todas las uniones entre individuos [5] y cuanto más tiempo pase el no infectado con un infectado, mayor probabilidad tendrá de infección. Aun sin concebir esta analogía, también se puede entender que cuanto más reciba cada involucrado una falsedad, menor tolerancia tendrá a la versión verdadera de los hechos [6]. Los grafos ponderados serán aquellos que representen estos matices a través de los pesos o ponderaciones asignadas para las aristas, a diferencia de los no ponderados [5] (ver Fig. 2.7).

En un escenario real, ya variables como los minutos de conversación entre individuos hacen que unas conexiones tomen más valor que otras y, en consecuencia, que la ponderación de las aristas sea diferente [5]. Tanto el tiempo como cualquier otro factor que afecte la conexión entre actores pueden formar parte de estos pesos, y también esto ocurre en el intercambio de falsedades.

En el contexto de las OSNs, se pueden tomar las republicaciones de contenidos como indicador para ponderar las aristas en vez de como las aristas en sí. Por ejemplo, en una red cuyas conexiones ya sean los seguidores entre usuarios, las veces que un usuario comparte el contenido de otro pueden servir para la ponderación. Lo mismo sucede con cualquier otra medida de *engagement* en la plataforma (comentarios, *likes*, métricas en relación al contenido, etc).

2.3.4. Modelos de difusión usados en SNA

Partiendo de la base de que la información (y la desinformación) es dinámica, el SNA ha abordado varias maneras de estudiar su difusión [180, 181, 12, 6, 5]. Aquí se presentan los modelos más populares para este estudio (ver Fig. 2.8), si bien pueden darse otros en la literatura.

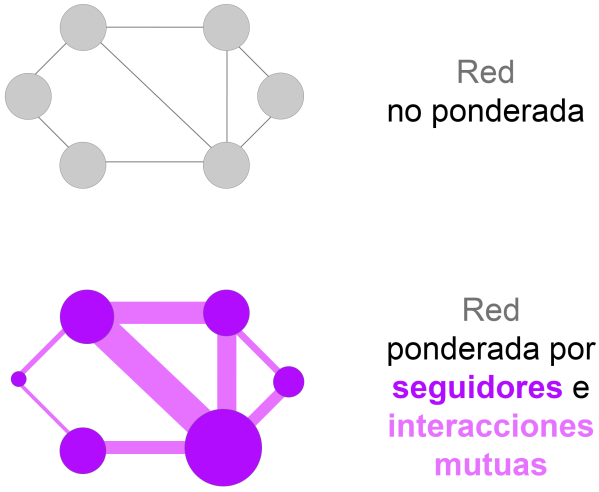


Figura 2.7: Diagrama con la interpretación de las redes ponderadas y no ponderadas a partir del trabajo de Barabasi [5], tomando como hipotéticos valores el número de seguidores y la cantidad de *engagement* en cuanto a interacciones de los usuarios en las redes sociales.

2.3.4.1. Modelos de umbral

Estos modelos constatan el impacto de un objeto a un nodo una vez que pasa un nivel de exposición a este, el umbral. En concreto, se adopta el concepto de umbral lineal porque es la suma lineal de las influencias de los vecinos de la red, bien a modo de recuento o ponderadas, la que se compara con el tope, que también puede ser uniforme o variar en función a un peso asignado para cada nodo [6]. Un mismo nodo podrá activarse todas las veces que se sobrepase ese límite y activar otros en otro momento del tiempo [181].

En el contexto de la desinformación en las OSNs, este objeto serán las informaciones falsas. Sobrepasar el umbral hará que el usuario acepte la falsedad por encima de la verdad. Ello permite varias interpretaciones: habrá cuentas en las OSNs más susceptibles que otras (umbrales más bajos) y se pone en valor el rol del *fact-checking* para que, dada una enunciación falsa, también haya un umbral de consumo de un desmentido y, superado su límite, el individuo afectado pueda recular ante tal mentira.

2.3.4.2. Modelos de cascada

Este modelo asume que cada nodo tiene una probabilidad de impacto sobre sus nodos vecinos [6]. No es la condición del umbral sino la influencia única de cada usuario respecto al siguiente la que posibilita la difusión. Esta propagación nodo a nodo genera una difusión en forma de árbol, la cascada. Es una activación en cadena, donde cada individuo es un eslabón al que se le confiere el poder de activar los siguientes nodos en un nivel de la cascada, pero que no tendrá influencia en otro momento del tiempo ni en los nodos que no sean concurrentes a la cadena [12].

En el contexto de las OSNs, se considera cascada a la cadena de elementos compartidos desde el contenido inicial hasta las últimas ramificaciones que sigue a través de los usuarios sin romperse (cascada de tamaño uno si solo comprende esa publicación inaugural, y de más tamaño conforme se va ampliando la cadena de compartidos) [7, 36, 6]. En el contexto de la desinformación, se

han estudiado cómo son los saltos de la información falsa rama a rama en comparación a la verdadera [36]. Bajo este modelo, se puede entender el *fact-checking* como dique para que la cascada no pase al siguiente nivel.

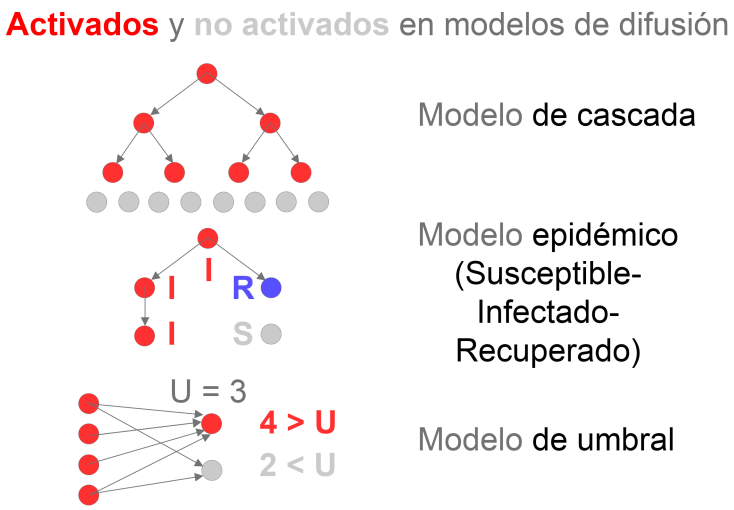


Figura 2.8: Diagrama con la interpretación de los modelos de difusión a partir del trabajo de Singh et al. [6]. A estos se añadirían los modelos de activación, más flexibles porque recogerían los aspectos comunes de estos modelos, y el tratamiento de las redes con la componente temporal.

2.3.4.3. Modelos epidémicos

Otra forma de entender las redes es desde el punto de vista epidémico. Los nodos pueden dividirse en susceptibles (los no afectados por el patógeno), infecciosos (los afectados) y los recuperados (ya recuperados y no infectan a otros). Son estados cambiantes y los susceptibles pueden infectarse y después recuperarse [5, 180, 6]. Los modelos epidémicos principales tienen a su vez variantes, a las cuales se pueden adecuar los diversos escenarios de la propagación de desinformación [12].

Estos tres tipos se entienden bajo las asunciones que toman cada uno de los modelos de infección: en los modelos Susceptible-Infectado (SI), si un susceptible entra en contacto con un infectado, se infecta también en una tasa β (la probabilidad de infección en un tiempo determinado) y no se recuperará; en los modelos Susceptible-Infectado-Susceptible (SIS), el proceso es similar pero el infectado puede recuperarse en una tasa de curación μ, aunque puede volver a infectarse, y en los modelos Susceptible-Infectado-Recuperado (SIR), la recuperación de la infección otorga una inmunidad y los recuperados no pueden infectarse [5].

2.3.4.4. Modelos de activación

Tanto en los modelos de cascada como en de umbral la dinámica es la de nodos que activan a otros para seguir con la cadena. Además, los modelos de umbral acuñan también los términos de contagio propios de los modelos epidémicos (susceptibles, infectados y recuperados), de la misma forma que las cascadas admiten variaciones, por ejemplo, en el tratamiento del tiempo de un paso a otro [180].

Por eso, los modelos de activación resumen las características de los modelos de cascada y de umbral para referirse a las situaciones de los afectados o de los siguientes eslabones en la estructura

de árbol. Proponen un modelo más realista, dado que la propagación de contenidos no tiene por qué encorsetarse a una forma de difusión ni de relación o estado entre nodos [6]. Así, los modelos de activación amplían el estudio de estas dinámicas para abordar toda forma de propagación e influencia entre los actores de la red.

Estos modelos permiten ampliar las miras en la difusión de desinformación: no tienen por qué encajar en los cambios de los modelos epidémicos [5] ni en sus variantes [12]; no tienen por qué llegar a los usuarios a través de cascadas de un solo nodo inicial o propagarse solo de esa manera [36, 7], y existen más factores además del nivel de tolerancia de los individuos a los contenidos [181]. Todo ello en la complejidad de las OSNs, cada una de ellas con sus particularidades, y del comportamiento humano que, incentivado por otros factores [46], cambiará a lo largo del tiempo.

2.3.5. El ciclo de vida de la desinformación

Los estudios sobre desinformación abordan el ciclo de vida de la información falsa, y son distintas las etapas enumeradas de acuerdo a la literatura. Zhou y Zafarani exponen tres: creación (cuando nacen los distintos tipos de enunciados falsos), publicación (cuando surgen en las OSNs) y propagación (cuando se expanden través de sus usuarios). Sin embargo, a la hora de abordar las dimensiones de la desinformación, los autores se refieren a los procesos de creación, propagación e intervención. A esta última no la consideran una fase, sino el tercer cometido de otra triada tras el análisis, la primera meta, y la detección, la segunda [46].

No todas estas etapas coinciden con las presentadas por Nasery: este autor coincide en la primera fase de creación y pasa después a la propagación, pero la tercera fase es la del impacto [182], la consecuencia posterior al ciclo de vida presentado por Zhou y Zafarani [46]. No obstante, de todas las clasificaciones se aprecian puntos en común. Las etapas aquí mostradas buscan tender puentes con todas las perspectivas: la primera comprende así tanto la creación y publicación; la segunda, la propagación; la tercera, la reacción (la que produce en los individuos y la que generan *fact-checkers* y otros actores a modo de oposición).

2.3.5.1. Etapa de creación y publicación

La primera fase comprende el nacimiento de la desinformación. Los estudios sobre esta etapa caracterizan la naturaleza de las falsedades surgidas, pero también de sus actores. Shu et al. distinguen entre varios tipos de participantes en la difusión de la desinformación: los llamados 'persuasores', que extienden la falsedad, los considerados 'clarificadores', que dan puntos de vista opuestos y los denominados 'crédulos', que aceptan el contenido falso (nombrados así porque no son perfiles que confían en una afirmación hasta que se demuestra lo contrario, sino que la aceptan sea fiable o no). En la fase de creación y publicación, todo empieza con los persuasores [3].

Los persuasores en las OSNs buscarán mover la red: en los modelos epidémicos, son el nacimiento de la infección para contagiar a los susceptibles; en los de cascada, el primer eslabón de la cadena para lograr el árbol de propagadores; en los modelos de umbral, los agentes para transmitir tal cantidad de información para que llegue un punto en el que los otros individuos también la acepten. En la concepción genérica de los modelos de activación, son los que encienden a los demás actores de la plataforma social para que ellos continúen con el flujo del contenido [180, 181, 12, 6, 5].

Los persuasores pueden ser actores individuales, pero Shu et al. también aluden a las cuentas

maliciosas, usuarios con la misión de agravar los desórdenes de la información. Se trata de los *bots* (cuentas no humanas para propagar la desinformación), los *trolls* (cuentas humanas que irrumpen en comunidades con el afán de provocar) y los cíborgs (cuentas llevadas por humanos pero con sus funciones en la plataforma automatizadas) como híbrido en favor de la diseminación de falsedades [3]. Estos toman protagonismo en la fase de la creación y publicación y permiten agitar los desórdenes de la desinformación en la siguiente fase de propagación.

2.3.5.2. Etapa de propagación

En esta fase entran en juego los crédulos, los perfiles en los que ha calado la desinformación por obra de los persuasores. Como se ha explicado antes, esta credulidad es diferente a la confianza [3] y será difícil que el usuario cambie de opinión una vez que ha tomado como verdad las afirmaciones falsas. Los modelos de difusión muestran, a su vez, cómo estos crédulos se convierten en persuasores: son los nuevos agitadores de los modelos de activación y, por ende, los nuevos infectados en los modelos epidémicos, los siguientes eslabones de las cascadas y los sujetos en mayor proporción para sobrepasar antes los límites soportados por los no afectados aún en los modelos de umbral [180, 181, 12, 6, 5].

A nivel general, los modelos de difusión siguen dos patrones: por un lado, el boca a boca del contagiado por una información al susceptible, que formará parte del correveidile después, conocido como '*viral model*' (transmisión *peer-to-peer*); por otro, la emisión masiva desde un actor al resto, por ejemplo, en la televisión hacia toda la audiencia, denominado '*broadcast model*' (transmisión *one-to-many*) [7] (ver Fig. 2.9). Las representaciones de los modelos de información muestran esto: 'viralidad' es un término que bebe de los modelos epidémicos y las ramificaciones de los modelos de cascada también ejemplifican esto [180, 181, 12, 6, 5].

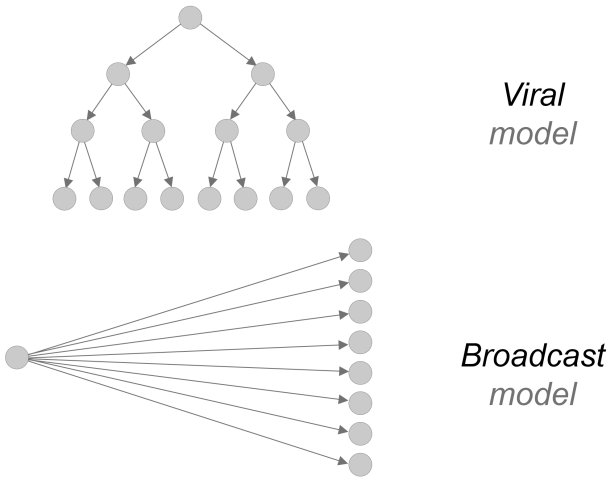


Figura 2.9: Diagrama adaptado a partir de la representación de los *viral models* y *broadcast models* realizada por Goel et al. [7].

Las redes sociales han dado la vuelta al tablero. Mientras que antes era más sustancial el contagio de desinformación a través del nodo padre al resto, las redes sociales hacen que el contagio de la falsedad a través de otros nodos, continuando como una infección [7], suponga una cuestión crítica. Eventos de gran calado como la Superbowl ponen sobre la mesa el poder de los *broadcast*

models, pero ahora las OSNs constituyen en este sentido un riesgo sin necesidad de esta proyección a nivel internacional. El boca a boca de los *viral models* ahora son transmisiones *many-to-many*, donde cada individuo desde una cuenta puede generar y/o republicar cualquier desinformación para expandirlo a otras cuentas [7].

Aquí entra el concepto de viralidad estructural, que se entiende dentro de la representación de los contenidos en una cascada. Este pone de relieve que, más allá de la difusión, existe un patrón. Matemáticamente, $v(T)$ constituye la distancia promedio entre todos los pares de nodos n que forman las ramas del árbol, formalmente $n(n-1)$ porque el nodo no es par de sí mismo. Entran aquí también aquellos nodos que se encuentran al mismo nivel entre sí, no conectados entre ellos pero sí por el mismo nodo padre [7].

En el proceso de difusión viral, el camino de la desinformación está más despejado. Vosoughi et al. descubrieron que la información falsa se expande más y mejor en Twitter que la verdadera tras analizar 11 años de falsedades comprobadas a partir de los desmentidos de seis organizaciones de *fact-checkers*, y no necesariamente por los *bots*, sino por las propias personas [36]. De las cascadas de compartidos, vieron cómo para las falsedades eran mayores el número de saltos a partir de los *reposts* (profundidad), las cuentas involucradas (tamaño) y la amplitud, medida a partir de la viralidad estructural definida antes [7], más aún en el caso de la desinformación política [36].

2.3.5.3. Etapa de reacción

Mientras que creación y propagación son las etapas comunes, antes o después, a los ciclos de la vida estudiados en la literatura, en la última fase se habla tanto de impacto y repercusión como de respuesta [182, 46]. En esencia, todos estos conceptos entrañan una reacción, tanto de forma negativa en la toma de decisiones por culpa de la información falsa, como de forma positiva para evitar su propagación, por ejemplo con el *fact-checking*.

En esta parte entrarían los clarificadores, con la intención de mitigar la desinformación [3]. Los SI ignoran este rol, pero no los modelos SIR ni algunas variantes: mientras que no cambia el contagio de infectados a susceptibles, que a su vez infectarían a más usuarios, el clarificador contribuye a que los infectados pasen a ser recuperados [5, 3]. Ellos serían la oposición a los persuasores de dos maneras: inmunizando a los individuos de la red y expandiendo los contenidos verdaderos a mayor escala [3]. También los clarificadores encuentran su sentido dentro de los otros modelos: son la contraparte a las falsedades que hace más difícil alcanzar el tope de los modelos de umbral para creerlas y las interrupciones de la cascada para que los contenidos falsos pasen en menor medida al siguiente nivel [180, 181, 12, 6, 5].

La respuesta en OSNs a la desinformación permite abordar el ciclo de vida de los contenidos falsos en sí. Se perciben tres comportamientos en su difusión, representados en la regla 80:10:10: alrededor del 80 % sufre una gran caída tras su repunte inicial, sobre el 10 % se mantiene o crece (entre los que se encuentran contenidos en desarrollo o en actualización constante) y el resto desciende en visitas pero luego sube (por otros espacios que redirijan después a tal contenido). Estas fases de mayor atención, descenso en picado y renacer a los meses se aprecia también en los *fact-checks*, donde el descenso de los desmentidos sigue al de las afirmaciones falsas [183].

2.3.6. Roles de los actores en cada etapa

Los nodos de la red son los emisores y receptores de la información y sus desórdenes, cada uno con unas propiedades que afectan a la difusión (ver Tabla 2.4). En el caso de los emisores, entran

en juego el tiempo de reacción (la espera desde que reciben el contenido hasta que lo propagan), la perseverancia (persistencia en el envío de estos contenidos a lo largo del tiempo), el nivel de autoridad (*status* en la red dado por su conocimiento, el cual hace que el impacto del mensaje sea mayor o menor) y la sensibilidad (disposición y emoción del emisor hacia el contenido). En cuanto a los receptores, juegan su papel la actitud (estímulos que modifican el comportamiento en función al contenido), el número de mensajes (la cantidad recibida para instarlo a la difusión, en relación a los modelos de umbral [6]) y la autoridad de la fuente (la impresión del receptor respecto al emisor en función de su *status* en la red) [12].

Tabla 2.4: Propiedades de los emisores, receptores y el mensaje transmitido en la difusión, de acuerdo a Raponi et al. [12].

Nodo	Estados	Inmutable Mutable
	Polaridad	Susceptible Resistente
Emisor	Tiempo de reacción	Inmediata Atrasada Sin reacción
	Perseverancia	A corto plazo A medio plazo A largo plazo
	Nivel de autoridad	Bajo Medio Alto
	Sensibilidad	Positiva Negativa
Receptor (factores de calado)	Actitud	Inmediata Atrasada Sin actualizar
	Mensajes	(Cantidad de información)
	Autoridad de la fuente	Popularidad Especialización

Más allá de los modelos de difusión, las cuestiones comunes a todos ellos son el comportamiento de los nodos cuando reciben el contenido y cómo lo conciben. Esto se traduce en los dos elementos que manejan el éxito o fracaso de la propagación: los estados (paso del nodo de un rol a otro cuando recibe el contenido) y la polaridad (juicio del nodo a la información recibida, que influirá en su estado) [12].

Se puede tomar el modelo epidémico para entender el papel de las cuentas en cada etapa de la difusión de contenidos [5]. Como ya se ha avanzado antes, el patógeno se puede interpretar como la información falsa y su curación como el *fact-checking*, el lugar de contagio serán las OSNs y unos infectados irán contagiando a otros en el ciclo de la desinformación.

La etapa de creación y publicación en la red da lugar al primer infectado. En los modelos epidémicos, pasará este contenido en la plataforma social y quienes lo vean en estas plataformas, los susceptibles, se infectarán a una tasa β [5]. En la etapa de distribución, la asunción de un tipo de modelo u otro define también el papel de los siguientes contagiados: en los modelos SI, aquellos que han recibido la información falsa no aceptarán otra versión de los hechos; en los SIS, los contaminados podrán deshacerse de este contenido falso en pos de la realidad en una tasa de recuperación μ, y esto les dejará expuestos a un nuevo contagio por la misma falsedad después [5]. No obstante, en los modelos SIR, sí se asume una solución [5, 12], dentro de la última fase de reacción.

Este remedio a la infección lo proporcionarían los *fact-checkers* en base a sus desmentidos [5, 12].

Es la interacción con la información verdadera la que da sentido al resto de modelos epidémicos más allá de los SI. Por eso en los modelos SIR la situación es más verosímil y los nodos pueden tener el estado de recuperados y ser inmunes, al contar con la información verdadera o con el desmentido sobre una falsedad [12]. Las extensiones de los modelos SIR plasman los otros escenarios de las etapas segunda y tercera del ciclo de vida de la desinformación (ver Tabla 2.5).

Tabla 2.5: Aplicación de las extensiones de los modelos Susceptible-Infectado-Recuperado destacadas por Raponi et al. [12] al ámbito de la desinformación a partir de su interpretación.

Modelo	Explicación
SIRS	Los recuperados (R) pueden dejar de ser inmunes y volver a ser susceptibles, por lo que necesitan el refuerzo de la información verdadera.
SIHR	Se incluyen a los hibernadores (H), que recibieron la falsedad y no la difundieron, pero que pueden recordarla y expandirla al tratar con los difusores.
SHIR	Se incorporan los *hesitators* (H), los dudosos al tanto de la falsedad pero que no la difunden por el esfuerzo que supone antes contrastarla.
SiRaRu	Los individuos tienen el contenido pero no lo comparten: los que no lo hacen por desinterés (Ra) y los que no lo hacen porque no se lo creen (Ru).
SCIR	Presupone la presencia de los contactados (C) antes de la difusión activa, cuando el individuo ha recibido la desinformación pero aún no ha decidido qué hacer con ella.
SKIR	Aparece el estado *Known* (K), los catalogados como antirumores porque propagan información verdadera frente a la falsedad. Es la difusión opuesta a la de los infectados.
SEIR	Los modelos SEI asumen un estado intermedio como incubador (I) antes de infectarse, porque se entiende que el individuo pasará por un proceso más o menos largo hasta que acepta o niega una falsedad. SEIR es la extensión SIR de esta versión.
SEIZR	Modelo SEIR que incluye la figura de los escépticos (Z), para los que no basta que llegue el contenido falso porque deciden no difundirlo.
ISCAR	Sus siglas corresponden a las entidades en inglés *ignorant*, *carrier*, *spreader*, *advocate* y *removal*, que dan lugar a ocho estados porque el modelo tiene en cuenta los comportamientos ante la información falsa por un lado y ante la información verdadera por otro.

Todas estas extensiones a los modelos epidémicos ponen en valor las transiciones, es decir, el paso de un estado a otro [12]. Gracias a ellas se pueden analizar las variantes en la difusión de desinformación, y se sortea el corsé del modelo epidémico básico SI, donde no hay alternativa al contagio tras el contacto [5]. No obstante, lo que produce la versatilidad de los estados iniciales (susceptible, infectado, recuperado) no son las transiciones por sí solas sino el concepto de la incubación, que introduce el factor de la duda o de la digestión del contenido para estas interpretaciones.

Con este elemento dentro de los modelos epidémicos, no todos los afectados que reciben el contenido son activos y lo difunden directamente, sino que también son pasivos y lo incuban antes de pasar al siguiente comportamiento [12]. Mientras que en los modelos epidémicos tradicionales solo se les da protagonismo a los nodos como emisores para mover la red, en las variantes con incubación los receptores dejan de ser actores secundarios a expensas de su activación y deciden su papel respecto a ella una vez que reciben el contenido.

En estos intercambios son de vital atención los *hubs* o superdifusores, aquellos nodos con una cantidad muy elevada de enlaces. Entendidas estas redes dentro de los modelos de difusión, los *hubs* pueden entenderse como los primeros que contienen el patógeno para expandirlo al resto [5] o como el inicio de una cascada [6, 7], pero también como aquellos con mayor poder de amplificación. Es decir, basta que llegue al *hub* antes o después para ampliar la enfermedad rápidamente [5].

En las comunidades, los grupos de actores operan como una entidad propia mediante la interacción entre ellos y es esto lo que posibilita el intercambio, ya sea de patógenos o, en este caso, de

la información y sus desórdenes [4]. Esto implica que el rol en la difusión no solo se entiende en los estados, polaridad, transiciones e incubaciones entre nodos, sino también en las dinámicas de los actores más influyentes (respaldados por las propiedades antes citadas de tiempo de reacción, perseverancia, nivel de autoridad y sensibilidad) [12] y en las de grupo. La velocidad en el ciclo de las informaciones falsas de la red tiene que ver con esto.

2.3.7. Factores que aceleran o frenan el ciclo

Raponi et al. ahondan en los factores que afectan la propagación de falsedades, clasificándolos en función a los elementos del grafo. Por eso, su distinción es entre factores de la red (tamaño, estructura, vecindad, tráfico, densidad, métricas como grado o densidad), de la comunidad (relación cultural y social de los grupos), del usuario (en cuanto a su bagaje, métricas del perfil en las OSNs y de comportamiento), de la información (su atractivo, veracidad, cantidad, contenido, claridad, finalidad o probabilidad de difusión) y del tiempo (duración de la transmisión, incubación, retención y descarte de las afirmaciones falsas) [12].

De manera más genérica, Li et al. engloban estos factores en influencias individuales y de las comunidades. Por un lado, las individuales aluden al impacto de los líderes de opinión por cómo arrastran al resto de la red. Por otro, las de la comunidad se refieren a las propiedades en común del grupo [181], de manera similar a Raponi et al. [12]. No obstante, Li et al. mencionan en su estudio cómo la descripción de la red y los factores afectivos (contenidos, relaciones, amistades, emociones) son la semilla de estos dos tipos de influencias [181].

De la investigación de Raponi et al. igualmente se extrae que los factores también dependen de estas métricas propias del grafo y de todo el contexto detrás de sus actores [181, 12]. En relación a este contexto, Zhou y Zafarani asocian las teorías de la psicología, la filosofía, las ciencias sociales y la economía a las características de la estructura. Mencionan teorías basadas en el usuario (influencia por sociedad, por él mismo o por la expectativa de que algo en su beneficio ocurrirá), en la propagación (percepción de los hechos según refuercen o contradigan las creencias del usuario) y en el estilo del contenido (características del contenido en función a si es real o ficticio) [46].

En este sentido, Singh et al. separan las propiedades estructurales de la red de las relacionales y las individuales, donde entra en juego esta parte afectiva. Por un lado, las relacionales incluyen la simpatía, la confianza y la reputación, entre otras; por otro, las individuales aglutinan la personalidad, la inteligencia emocional y el análisis del sentimiento, además de la intencionalidad y las experiencias pasadas, también enumeradas [6]. Son, por tanto, múltiples los factores descritos en estudios distintos, pero todos ellos llevan a que las características procesadas por el SNA, sean del tipo que sean, influirán en mayor o menor medida.

Esta combinación de cuestiones afectivas y características de la red muestra que referirse a los factores en la difusión es aludir directamente a las trabas de la red y de sus dinámicas. Bakir y McStay introducen las cuestiones afectivas dentro de los que llama los 'tres frentes' de la desinformación, junto al surgimiento de ciudadanos mal informados y a las cámaras de eco [184]. Esto revela que las acciones por sí solas de los usuarios de la red no son los únicos elementos que potencian la desinformación, sino un conjunto de ingredientes que la hacen más poderosa en las OSNs.

En estos tres frentes, Bakir y McStay incluyen las cámaras de eco porque son el paso siguiente a tener ciudadanos mal informados: el riesgo de mantenerlos siempre así [184]. Shu et al. tampoco

reducen el problema a los actores individuales y a las cuentas maliciosas, y cita las cámaras de eco como forma de aislamiento del individuo dentro de una comunidad, pero también las burbujas de filtro como aislamiento mediante su propio contenido personalizado [3]. Entendido el paso de la desinformación como un modelo epidémico SIR [5], los susceptibles quedan más expuestos a la desinformación mediante estas burbujas y, a su vez, estas los separan de las cuentas difusoras de los desmentidos e información verdadera que pueden revertir su estado de infectado.

En el caso de las cámaras de eco, se forman en las plataformas sociales porque los usuarios se decantan por seguir a cuentas similares y eso hace que consuman el mismo contenido afín. Las consecuencias de esto son dos: por un lado, el usuario tenderá a percibir este contenido como creíble si el resto de esta comunidad similar así lo ve (credibilidad social); por otro, se decantará por las afirmaciones que consuma con más asiduidad, sean verdaderas o falsas (heurística de frecuencia) [3].

En cuanto a las burbujas de filtro, son los sistemas de recomendación de la red social los que ofrecen un tipo de contenido en vez de otro en función a las preferencias de consumo del usuario. Por eso, también recibirá de la plataforma más contenido afín en vez de opuesto, dando lugar a que la percepción de la realidad propia es la válida y a que toda ajena a ella será incorrecta (realismo naíf), y a la preferencia por recibir el contenido que reafirme sus propias ideas (sesgo de confirmación) [3]. Estas burbujas modifican el concepto tradicional de los modelos virales [7] y van un paso más allá de los estudios respecto a los saltos de la información falsa en cascadas [36], ya que el usuario no encuentra la desinformación en la ramificación en forma de árbol ni en interacción con otro individuo en los modelos de difusión [6], sino que esta se salta todo el proceso al ser un contenido recomendado directamente por el algoritmo de la plataforma.

Los momentos de crisis también influyen en las falsedades. En estos contextos, los usuarios tienen más voluntad de reconstruir la verdad y compartirla a través de rumores, convirtiendo a las redes sociales en la sede de afirmaciones sin completar ni confirmar antes de que las fuentes de autoridad puedan emitir informaciones oficiales y verificadas [185, 186, 187, 188, 189]. Mehta et al. recopilaron los factores del contenido falso en estos contextos, como que los usuarios lo acepten siguiendo más su instinto y sus esquemas mentales [190] o porque apele más a la emoción [191], entre otros [186]. De todos los daños recavados en la *survey* de Tran et al., son los daños a la emoción, más aquellos a la credibilidad, a la reputación y a la toma de decisiones los que pueden agravar el flujo de la desinformación en las OSNs en estas circunstancias [192].

Pese al aumento del estudio de la desinformación, Suryana et al. advierten de la necesidad de ahondar más en las dinámicas en escenarios de crisis y esbozan cómo los factores ya mencionados agravan más las falsedades en ellas: afectan, por un lado, las dinámicas psicológicas de inestabilidad emocional y de falta de pensamiento analítico en favor de atajos mentales para llegar a conclusiones, más los sesgos ya advertidos de confirmación, pero también de aquellos de grupo (aceptación de las afirmaciones de la comunidad afín en favor de la cohesión grupal) y la heurística de frecuencia; influyen, por otro, los caldos de cultivo de las redes, con las recomendaciones del algoritmo del contenido polémico y emocional por *engagement*, la presencia de cámaras de eco y la ausencia de mecanismos frente a estos desórdenes, y, condicionan, por último, los impactos de todo esto contra la confianza en las instituciones, a partir de narrativas partidistas, polarización, desobediencia a las recomendaciones, rebelión frente al orden público y rechazo a los medios de comunicación [193].

2.3.8. Visualización de redes de desinformación

El estudio de Camacho et al., que disgrega las cuatro dimensiones del SNA, también realiza un análisis de las herramientas dedicadas a ella. La visualización, última dimensión de las cuatro, es el producto final del análisis de grafos en SNA y la última meta de varias herramientas de este campo [37]. Los criterios usados para su evaluación son la explotación de las Variables Visuales [194] y la interactividad. Por un lado, las Variables Visuales aluden a los elementos gráficos que ayudan a distinguir entre grupos, cambios, secuencias y valores numéricos (color, tamaño, orientación, color, saturación y textura); por otro, la interactividad es la forma alternativa de mostrar la información para no aglutinar demasiadas Variables Visuales (*zoom*, filtro, destacados, agrupamientos y selectores para cambiar de representación de los datos) [37].

En este estudio, a partir de una lista inicial de 70 herramientas de esta disciplina, se destacaron 20 a partir de su tipo de licencia de software, su documentación y su impacto. Cada una de estas 20 recibió una puntuación para cada dimensión en base a la información en tal documentación, en sus webs oficiales y en los trabajos publicados en el área [37]. En el momento de tal investigación, se concluyó que las cinco herramientas subrayadas para este campo eran ORA-LITE/PRO, Grasphistry y Neo4j, seguidas de cerca por Gephi y Cytoscape [37].

El estudio de Rani y Shokeen, dentro de las tareas de visualización, diferencia entre las herramientas con propósito global y aquellas con un fin específico. De manera general, las herramientas de SNA están diseñadas tanto para analizar como para visualizar los grafos. Pusieron de ejemplo Pajek, JUNG, SocNetV, UCINET y Gephi, más los lenguajes de programación MATLAB, R y Python [195]. Sapountzi y Psannis hacen una distinción entre herramientas según su calidad de visualización alta (Gephi), media (NodeXL y Statnet) y baja (Pajek y NetworKit), pero destacando, entre otros, las funciones de NodeXL destinadas a las OSNs [196]. Pero no todas soportan de la misma manera los datos de las redes cuando son muy grandes: el estudio de Akhtar alude a las librerías de Python para las estructuras más ingentes [197], y tanto él como Sapountzi y Psannis asocian softwares como Gephi a redes de menor tamaño [196].

Más allá de sus fortalezas y debilidades, estos autores reflejan el amplio abanico de herramientas de SNA para visualizar el traspaso de la información y de sus desórdenes. Aunque, de acuerdo a la comparativa de Rani y Shokeen, se destaca R para todo propósito en SNA, se documenta en esta abundancia de recursos la importancia de crear flujos de trabajo para el análisis eficaz. Por ejemplo, con Python se pueden obtener los datos a través de llamadas a una API para luego pasar al enfoque más estadístico de R o a la visualización con Gephi [195]. En estas investigaciones se evidencia también el aporte de PLN que puede servir de ayuda en los grafos finales: Sapountzi y Psannis listan las funciones de análisis de sentimiento y las técnicas de procesado de texto en la comparativa de herramientas de SNA [196].

2.3.8.1. Interpretación de patrones visuales: comunidades, *hubs*, patrones de difusión

El *layout* es el primer paso de la visualización del grafo, puesto que define la posición de todos los elementos de la red para que sean interpretables al ojo humano [4, 195]. Así, nodos y aristas se mapean en función de sus características, de su distancia o similitud en algún aspecto concreto con el resto de nodos, y con algoritmos diseñados para esto, contando con estas propiedades de los nodos y permitiendo mejor legibilidad [4]. Dependiendo de la herramienta de SNA, se podrán emplear unos *layouts* u otros, y son varios estudios los que las han analizado en base a los que pueden emplear [197, 195].

Entre otros, se diferencian los siguientes *layouts* (ver Fig. 2.10): aleatorio (posiciones arbitrarias diseñadas para mostrar las métricas), circular (los nodos en anillo, en el centro o alejados de él dependiendo de si tienen muchas o pocas conexiones, respectivamente), en *grid* (disposición rectangular para mostrar la estructura de la red y sus grupos) y dirigida por la fuerza (atracción o repulsión entre nodos en función a las conexiones y cercanía) [8]. Tanto la *layout* circular como la *spring layout* (una de las dirigidas por la fuerza) aparecen en las principales herramientas de SNA analizadas por Akhtar [197].

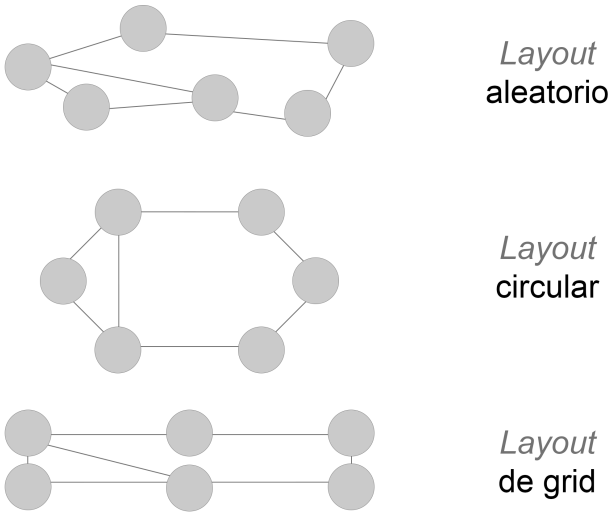


Figura 2.10: Diagrama adaptado a partir de la representación de los tipos de *layout* para los grafos realizada por Singh et al. [8].

Los nodos, representados como puntos, pueden mostrar también las diferencias entre grafos según sus atributos a través del tamaño, la forma y el color [4, 8]. Las formas y colores de los nodos sirven para mostrar las variables categóricas que distinguen entre unos y otros, mientras que la escala de color y el tamaño permiten asignarles en la visualización las variables en números absolutos y relativos. Esto ayuda a la comprensión del fenómeno, mientras no se abuse de la cantidad de propiedades plasmadas [4]. De esta forma quedan mapeadas tanto las métricas como las relaciones entre nodos [4, 8].

La visualización no solo afecta al modo de dibujar los nodos, sino también a las aristas. Si bien la potencia de las uniones se puede considerar una característica de la red [8], también puede concebirse como una forma de visualización. En este sentido, se pueden eliminar los enlaces más débiles o darles un grosor en función a la fuerza calculada de esa unión. Cuando las conexiones no son unidireccionales, esa concepción de aristas como flechas se puede plasmar como tal, además del resto de modificaciones como el grosor u otras [4].

Las etiquetas son también importantes en el grafo porque dan la información explícita sobre los nodos y las aristas, pero suponen a la vez un reto porque pueden hacer confusa la visualización si no se utilizan bien [195]. Esto es crucial en la difusión de desinformación, por ejemplo, para etiquetar los temas que discurren entre usuarios o los *clusters* que forman en función a sus características comunes, los cuales pueden sobrecargar la representación. Por eso, en este aspecto sale también a colación la interactividad, hacia el equilibrio de cuál información mostrar de base y cuál una vez el usuario accede a cada uno de los nodos, aristas o grupos [37].

Aquí entran en valor las formas de encontrar patrones a partir de la interactividad para entender el grafo: el filtrado (eliminando nodos y enlaces en función de las métricas para obtener solo los relevantes), el *clustering* (dividiendo la red en subredes por los patrones de la estructura y sus propiedades para descubrir su anatomía), los agrupamientos (la vista de pájaro para no observar solo las conexiones entre nodos, sino las uniones globales entre los grupos que forman según sus atributos) y la simplificación (transformando el resultado en estructuras más simples comunes a otros grafos) [198]. En resumen, todas ellas permiten ir de lo general a lo particular. Acciones como el filtrado y la simplificación afectan a la modificación de la escala, la tercera dimensión en valor en las visualizaciones además del *layout* y las propiedades de las etiquetas, color, forma y tamaño [8].

2.4. Análisis de redes sociales con evolución temporal

En los apartados anteriores, la dinámica temporal ha sido un elemento transversal: en el rol de los emisores, se mencionó el tiempo de reacción como una de sus características claves, además de ser considerado como uno de los factores que afectan a la propagación en general [12]; por eso, dentro de las estructuras de grafos, aquellas que plasman la difusión no pueden ignorar la probabilidad de los traspasos a través de las aristas pero tampoco el tiempo que necesitan para ello [3]; esto hace que en las redes ponderadas un ejemplo de pesos asignados a las aristas sean estos tiempos de propagación; también que en los modelos epidémicos la infección se vea afectada por este aspecto [5], y que en los modelos de activación se flexibilicen las concepciones clásicas de las cascadas [180].

2.4.1. La componente temporal

La presencia del tiempo muestra, por tanto, que no es una variable más, sino uno de los pilares del SNA. Como ya se ha visto antes, Shu et al. representan en el sistema de la diseminación de información la dimensión del contenido (el qué) y la social (el quién). Pero los autores solo conciben esta estructura con una tercera dimensión, la temporal (el cuándo) (ver Fig. 2.11). Es la que ilustra la evolución de las cuentas reproduciendo los contenidos a lo largo del tiempo, y no solo las uniones entre los usuarios o entre sus publicaciones [3].

Lo mismo sucede en los modelos de difusión: el tiempo es parte de las redes concebidas como infecciones o como cascadas, pero también es protagonista de estas representaciones. Es por ello que, además de los modelos ya vistos, Singh et al. enumeran también los modelos basados en el tiempo. En línea con las anteriores referencias a la parte temporal por el resto de autores, estos modelos tienen en cuenta que la difusión se maximiza en una ventana de tiempo específico [6]. En el contexto de la desinformación, importa también la velocidad de propagación en la publicación y amplificación de los contenidos, en vez de solo el mensaje y los actores que lo transmiten.

La forma tradicional de plasmar nodos y enlaces es la estática, y también la esencial para mostrar las conexiones en una plataforma social, pero esto la despoja de la variable del tiempo. La misma red representada de forma estática en tres momentos de tiempo t diferentes puede ser muy diferente [199]. Esto se debe a que la transmisión de información en las OSNs surge en redes dinámicas: estos contenidos están en movimiento entre cuentas y además estas pueden aparecer, desaparecer, aumentar tanto sus seguidores como su actividad, colaborar en la difusión, bloquearla o quedar en un segundo plano.

Por eso, no es que el concepto del tiempo no exista, es que es una transversal a los modelos antes

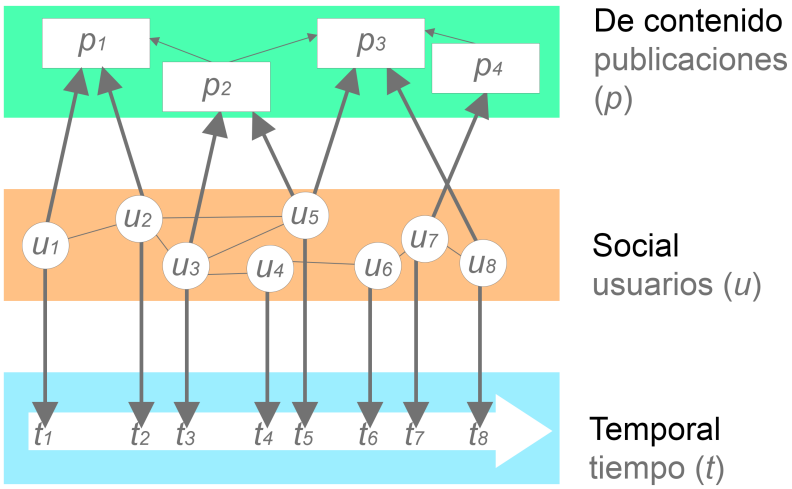


Figura 2.11: Diagrama adaptado a partir de la representación de las dimensiones del análisis de redes realizada por Shu et al. [3].

citados. En la definición formal de un grafo $G = (V, E)$ como red temporal, enlaces y nodos se encuadran en una franja de tiempo $[0, T]$. Un enlace entre los nodos u y v del set de nodos V se concibe como $e = (u, v, t, \delta t)$, porque se incluye el tiempo t dentro de la franja $[0, T]$, en un momento específico donde $0 \leq t \leq T$, y δt es su duración [179]. En el contexto de las OSNs, así como otros, las redes pueden ser grafos de secuencia de contacto donde $\delta t \to 0$ porque el contacto es instantáneo (por ejemplo, una interacción en plataformas sociales) [179].

En los modelos epidémicos, tanto la rapidez como el alcance del patógeno no se pueden entender en su totalidad si se ignora que los patógenos se expanden en redes temporales [5]. Lo mismo sucede con la desinformación: para entender la difusión de estos contenidos, cómo se viralizan en algunos casos y hasta dónde pueden llegar no se puede mirar el fenómeno como una imagen fija. Cuando esta difusión se ha estudiado a modo de cascada, más allá de la concepción tradicional [6, 12], el tiempo también ha permitido ver los momentos de explosión de la propagación [200].

2.4.1.1. Patrones de difusión temporal

Las propiedades de los grafos en las OSNs afectan al tiempo de propagación de la información (y desinformación). Estas redes son de escala libre porque siguen una ley de potencias: mientras que unos nodos concentran muchas de las conexiones, la mayoría de nodos tienen pocas. En términos matemáticos, esto implica que la cantidad de nodos del grado k es proporcional a $k^{-\gamma}$ [4]. Y esto es lo que da lugar a una difusión rápida de los contenidos.

Estas redes de libre escala tendrán una distancia geodésica promedio L muy baja. La distancia geodésica es el número de conexiones para llegar de un nodo a otro y, si de media es poca, la información no necesitará muchos pasos para llegar a los nodos, por su unión a aquellos que concentran gran parte de las conexiones. Para este tipo de estructuras grandes, se presupone que L crecerá proporcionalmente a $logN$ (siendo N el número de nodos). [4]. Es decir, aunque N crezca (por ejemplo, a un millón), L como $logN$ (6 para este caso) crece más lentamente.

Los seguidores en Twitter (X) son un ejemplo de esto, sin necesidad de reciprocidad por parte de

los nodos que reciben la mayoría de *follows*, y bastan estas conexiones para las jerarquías entre ellos. Para este tipo de redes se sigue el efecto Matthew, donde el más tiene, más aún recibirá. Es decir, un nodo con muchas conexiones atraerá también a otros nodos a conectar con él [4].

Las redes sociales no solo contribuyen a la difusión rápida mediante una corta distancia geodésica, sino también mediante un coeficiente de agrupamiento alto por las afinidades que unen a los nodos, es decir, el número de conexiones reales entre los vecinos k de un nodo i dividido entre todos los posibles [4], formalmente $k_i(k_i - 1)/2$ ($k_i - 1$ dado que las conexiones no se producen de un nodo consigo mismo y partido por 2 al estar contando los pares). Por eso, se habla para estos casos también de redes de mundo pequeño, llamadas así por los pocos saltos necesarios para pasar de un nodo a otro en la estructura, en analogía con la popular teoría de los seis grados de separación [4].

La rapidez de la difusión también puede entenderse a partir de los modelos epidémicos. En el modelo epidémico SI se asume el flujo rápido porque los contagiados no vuelven a su estado inicial y los susceptibles no pueden pasar a un estado que les blinde del virus, de tal manera que no hay ninguna injerencia en la cadena de información que corte las conexiones por las que circula [5]. No obstante, el modelo SIR permite una difusión más lenta porque los actores de la red se inmunizan y bloquean el paso de las falsedades a otros susceptibles [5].

La disposición de la variable temporal en el eje x constituye en la literatura un modo de no discriminar la cuestión dinámica de la propagación de información. Vosoughi et al. realizaron el análisis del tiempo de las cascadas de contenidos enfrentándolo a los indicadores propios de estas (tamaño, profundidad, viralidad estructural) en el eje y [36]. A modo de diagrama visualizaron además las interacciones desde el nodo inicial a los siguientes con un *layout* cuyo eje x también representaba el tiempo. Goel et al. también plasmó la relación entre el tamaño de la cascada y las horas por un lado y la formación de la cascada en sí por otro [7]. Por su parte, Tabassum et al., centrados en el modelo temporal, priorizan la estructura de la red pero añaden como anotaciones el orden de los momentos en los que los nodos han interactuado entre sí [179].

Sea de una forma o de otra, estos gráficos permiten visualizar la velocidad con la que puede propagarse una falsedad. Sacar la variable del tiempo de la ecuación imposibilitaría esto, así como el orden en que los distintos actores toman partido en la difusión y hasta qué punto la hacen más rápida [36, 7, 179]. Por esto mismo, los modelos epidémicos no pueden ignorar este factor en la analogía de la desinformación como un virus: sus visualizaciones necesitan mostrar en el eje x el paso del tiempo para reflejar la cantidad de población infectada dependiendo del momento [5, 201] y, con ello, evaluar el aumento o reducción del contagio en cada etapa. Estas evoluciones de la desinformación de la mano del tiempo también pueden hacerse en comparación con el *fact-checking* [202].

2.4.2. Aplicación de técnicas de SNA para monitorizar la desinformación

Cada nodo de la estructura tiene una importancia que también puede medirse con las matemáticas. Se le llama centralidad a la posición del nodo en la red en función a su contribución en ella. Un nodo con mucha centralidad romperá la red en gran medida, pero si apenas la tiene no afectará [4]. Esta centralidad se puede medir de varias maneras en función al objetivo del estudio.

2.4.2.1. Centralidad de grado

El grado es la forma más sencilla de abordar la centralidad en cualquier red no dirigida y sin pesos ya establecidos. La centralidad de grado mide cuántos enlaces tiene con otros [4, 8]. Hay dos tipos de grados: *in-degree* en favor del número de aristas que van al nodo y *out-degree* en cuanto a las que salen de él [8]. Esto hace que esta medida pueda ser tanto de contribución (*in-degree*) como de exposición (*out-degree*) [4, 8].

En el contexto de la desinformación en las OSNs, esta centralidad de grado puede entenderse de ambas maneras. Por un lado, tendrán mayor centralidad los mayores amplificadores de las falsedades, porque llegan a más usuarios (*out-degree*); pero también la tendrán aquellas cuentas más vulnerables a las falsedades porque reciban, en total, muchos enlaces con los propagadores (*in-degree*). Ambas cuestiones son fundamentales en el estudio de estos desórdenes.

2.4.2.2. Centralidad de cercanía

La cercanía como medida de centralidad se define como la proximidad de un nodo con el resto de la red [8]. Formalmente, supone la distancia geodésica $d(u, v)$ más corta entre un nodo u y un nodo v. Por tanto, el grado de cercanía de cada nodo dentro de la red se medirá por la suma de aristas necesarias para llegar a todos los nodos. Como en este caso un mayor valor determinará una mayor distancia, se divide la distancia mínima geodésica posible por la distancia geodésica real calculada. El mayor valor aquí será 1 (la distancia geodésica real es la mínima posible), la posición más central. Este mínimo será solo una conexión por nodo, y por tanto se puede representar con el total n de nodos menos el nodo analizado en cuestión $(n - 1)$ [4, 8].

Adaptada esta métrica al área de la desinformación, un difusor de falsedades o un contenido falso tendrá mayor centralidad cuantas menos conexiones necesite para llegar al resto de actores de una comunidad. En la comparación con la información verdadera en general y con los desmentidos en particular, si bien el grado solo describía su total de conexiones, esta otra medida de centralidad permite ver si es más difícil para estas verdades y sus amplificadores llegar al resto de usuarios.

2.4.2.3. Centralidad de intermediación

La centralidad de intermediación calcula la relevancia de cada actor a través de los caminos más cortos entre pares [8]. También toma la distancia geodésica, pero en este caso no se mide si un nodo tiene las conexiones más cortas con el resto, sino si este hace mejor de puente de otros nodos, es decir, si permite el camino más corto entre ellos. Así, la intermediación de un nodo v es la suma del total de caminos más cortos de cada uno de los pares de nodos que pasan por v divididos entre todos los caminos posibles de esos pares de nodos, pasen o no por v. El nodo que obtenga un valor más alto será el que mayor intermediación tenga [4].

En el área de la desinformación, el nodo con la máxima intermediación puede reflejar al usuario que permite a una falsedad concreta dar menos pasos para que llegue al resto de cuentas a las que está este usuario conectado. Si bien la cercanía mostraba cómo un difusor es central porque su contenido llega antes a los nodos, la intermediación revela que es central cuando se convierte en el mejor nexo para que el contenido generado por otros llegue de la forma más corta al resto.

2.4.2.4. Centralidad de vector propio

Son más los tipos de centralidades además de las de grado, cercanía e intermediación. La literatura también destaca la centralidad *eigenvector*, una variación de la centralidad de grado porque no

solo se mide el número de conexiones del nodo sino cuánto de importantes son, es decir, en ponderación con su influencia, representados en una matriz de adyacencia [8]. Formalmente, e_i es la centralidad del nodo i, e_j es la del nodo vecino j, λ es el *eigenvalue* y A_{ij} es el valor en la matriz de adyacencia, (1 cuando hay conexión, 0 cuando no la hay).

En el contexto de la información falsa, si una cuenta en la red social tiene pocos contactos pero estos son justamente los que tienen más impacto en la red a la hora de mover desinformación, mientras que su centralidad de grado sería baja, la centralidad de vector propio sería alta precisamente por esta ponderación en relación a la matriz de adyacencia con todas las conexiones.

2.4.3. Análisis de cohesión y densidad de la red

Además de las medidas para conocer la importancia del nodo dentro del conjunto, la red también puede evaluarse de manera global. Es aquí donde entra el concepto de la cohesión, cómo de enredados están los nodos de la estructura [4]. El número de aristas ya permite dar valores de guía de la densidad de la red. Esta se puede comprender en función al total de pares enlazados posibles (el número de nodos n multiplicado por todas las conexiones a los nodos menos la conexión a él mismo, $n-1$) divididos entre dos ($(n-1)/2$) si se entiende el grafo como no dirigido (la conexión entre A y B es la misma que entre B y A) [4]. No obstante, son varias las concepciones de la densidad, y también puede entenderse en función de la suma del total de conexiones entre dos nodos (bidireccionales para cada nodo en el par en las redes dirigidas y unidireccionales en las no dirigidas, formalmente como $2M$ y M, respectivamente) [6].

Como esta métrica juega en favor de las redes pequeñas porque es más fácil la conexión entre pocos nodos que la conexión entre una cantidad ingente de ellos, la densidad también se puede medir con el grado de los nodos en promedio. Estas medidas también pueden aplicarse únicamente a los nodos del mismo tipo para saber cómo es la densidad solo entre ellos, sin contar al resto, obteniendo un valor de cohesión por grupo para entender cada clase de actores en la difusión. No obstante, esto pone de manifiesto otra cuestión: la medida global de la conexión no permite diferenciar entre las zonas muy enredadas de aquellas más sueltas, apenas cohesionadas en el grafo [4].

Esto abre la puerta a medir la cohesión a partir del número de componentes del grafo. Se entiende por componente el grupo de nodos unidos por las conexiones directas (un nodo conectado con todos los de su grupo) o indirectas [4] (por ejemplo, nodo A y nodo C conectados porque A y C se conectan a B). Normalizando esta medida de los componentes, siendo así el ratio de componentes, se puede comparar cómo de conectadas están unas redes respecto a otras. Consiste en dividir $c-1$ entre $n-1$, donde c es el número de componentes y n es el número de nodos [4]. Si todos están conectados, c será 1 y, por tanto, el numerador será 0, dando lugar a la máxima conexión. De lo contrario, todos los nodos aislados darían lugar a tantos componentes como nodos y, por ello, restar 1 en el denominador permite llegar al 1 como resultado de la mínima cohesión.

Sin embargo, esta medida no define, dentro de los componentes, si sus nodos están muy conectados o no. Por eso se proponen medidas como la conectividad [203], la proporción de nodos conectados en la estructura más allá de la cantidad de componentes [4]. Supone uno de los conceptos claves para entender la estructura del grafo, junto a las subredes (los grupos de nodos) y los caminos (la secuencia de nodos hasta llegar a uno en concreto), aparte del binomio nodo-arista [8]. Esta métrica implica sumar todos los pares de nodos que están en un mismo componente (r_{ij}, siendo i y j los nodos del par, nunca iguales entre sí, y r igual a 1 cuando estén en el mismo componente)

y dividir entre el número total de pares del nodo menos la unión consigo mismo ($n(n-1)$). Una alternativa sería la compacidad, que no cuenta las uniones entre dos nodos sino la longitud geodésica inversa para que la suma sea de cada par ponderado en función a esta distancia ($1/d_{ij}$ en vez de r_{ij}, siendo d la distancia geodésica) [4].

En este tipo de características también tiene cabida la robustez, el grado de dificultad de romper la red al eliminar nodos o enlaces. Se consideran puntos de corte y puentes los nodos y aristas, respectivamente, que cuando desaparecen fragmentan la red, y son sets de corte los conjuntos de nodos y aristas que la rompen. Si son muchos los que tienen que deshacerse, se entenderá la red como robusta y, por ende, cohesionada. Para esta propiedad, se tienen en cuenta como métricas el mínimo de nodos y de enlaces necesarios para desunir la red [4]. Esto concuerda con otra de las concepciones de cohesión como tal, entendida también como el mínimo de nodos cuya desaparición rompe la red o aumenta la cantidad de componentes [8].

2.4.4. Detección de anomalías en SNA

Singh describe junto a más autores las verticales de estudio dentro del SNA: la diseminación de información (los modelos de difusión ya tratados), la predicción de enlaces (cálculo de nuevos enlaces de la red en base a la similitud de sus características), la maximización de la influencia (detección de los nodos con más impacto) y la detección de comunidades (localización de los grupos que aúnan ciertas dinámicas y/o características) [6, 8]. De estas verticales, mientras que la predicción de enlaces apunta a un escenario futuro, dado que busca adivinar aquellos que todavía no existen, tanto la maximización de la influencia como la detección de comunidades se centran en el presente. En el campo de la desinformación, estas dos últimas ramas permiten saber de antemano los grupos en los que se generan las falsedades y los usuarios con más calado en ellas.

El objetivo de la maximización de la influencia es hallar los nodos iniciales del impacto, los llamados 'semillas' [6, 181]. Esta meta se enmarca en el análisis de la influencia individual y de la comunidad, pues el fin puede ser encontrar a los actores más influyentes por sí solos o de manera colectiva. Genéricamente, la maximización de la influencia comprende una fase de selección de las semillas (de una sola vez o en un proceso iterativo) y una fase de acción para ver cómo se extiende su impacto. En este segundo paso ya entran en juego las teorías basadas en los métodos de difusión como las activaciones por cascada o por superación de umbral [181].

Dentro de las técnicas de maximización de influencia en el contexto de las OSNs, los métodos de aproximación son menos eficientes en las redes grandes por su coste computacional al usar métodos voraces, por lo que las reglas prácticas de los heurísticos son una alternativa que sacrifica las soluciones más óptimas en favor de otras más simples para estas cuestiones a gran escala. Por su parte, los métodos de detección de comunidades permiten encontrar esas semillas a través de los *clusters* de los grupos. Las metaheurísticas se erigen como recurso actual en su equilibrio entre las soluciones avanzadas de los enfoques costosos y la simpleza de los heurísticos [6].

La detección de comunidades es posible a través de la partición de grafos en comunidades, de los *clustering* tanto espectrales (patrones en las matrices de las conexiones) como jerárquicos (dendrogramas para ver las uniones y separaciones en los nodos a cada nivel), de los métodos dinámicos (métricas obtenidas de los caminos aleatorios entre nodos) y de los algoritmos basados en densidad (para detectar grupos muy unidos en contraste al resto de conexiones) [6]. Para esta tarea se puede partir de dos asunciones: que cada nodo sea parte de una sola agrupación (con

métodos que trabajan mejor en redes pequeñas) o parte de varias, con un valor de adherencia a cada comunidad, donde se agrupa la mayor parte del esfuerzo en esta área [8].

Son variadas, por tanto, las formas de detectar comunidades, pero eso no quiere decir que todas valgan como soluciones para el mismo problema. Es conocido, por ejemplo, el algoritmo de Girvan-Newman para este aspecto. Calcula la centralidad de intermediación de las aristas y se eliminan aquellas con mayor valor para dividir el grafo en subcomunidades y obtener la mayor modularidad Q de la red. En concreto, el valor óptimo de Q resulta de un proceso iterativo para eliminar la(s) arista(s) con la máxima centralidad y seguir con las siguientes en el ranking hasta encontrar la Q más alta. Las subcomunidades que hayan producido tal Q óptima tras eliminarse una cantidad específica de aristas son el resultado de este algoritmo. Sin embargo, rinde peor en redes grandes por el coste computacional de todos los cálculos y por la dificultad de detectar pequeñas comunidades en ellas [4].

Aunque el algoritmo de Girvan-Newman pueda funcionar bien para muestras ya filtradas y, por tanto, no tan grandes, algoritmos como el de Louvain vienen mejor para la puesta en práctica en las plataformas sociales, con estructuras no tan sencillas. Este algoritmo depende de la modularidad Q, pero a menor coste y sin tener que calcular la centralidad de intermediación de las aristas [4]. Son tres pasos: primero, la etapa voraz toma la comunidad de conexiones de cada nodo, se mueve luego a la comunidad vecina (las conexiones de las conexiones del nodo inicial) y esto forma un supernodo que, con los otros supernodos da un valor de modularidad Q; segundo, mira las conexiones de los supernodos para rehacer el primer paso y calcular Q de los nuevos supernodos; tercero, repite el proceso hasta llegar al Q máximo [4]. Pero su límite de la resolución le da dificultades para encontrar también grupos pequeños [8].

Pero, más allá de esto, las redes dinámicas como las OSNs presentan una dificultad: sus nodos y aristas evolucionarán [8]. Es aquí donde entra la detección de cascadas, donde el objetivo ya no es captar los grupos, sino los patrones de propagación. En el ámbito de la desinformación, se parte de la idea de que las cascadas de contenidos falsos tendrán un comportamiento distinto a las de los verdaderos [46], algo que va en la línea de cómo las falsedades discurren mejor por estas ramificaciones que las verdades [36]. Por tanto, estas características de la propagación sirven para diferenciar unas cascadas de otras.

Si se logra discernir entre los caminos por los cuales discurren las informaciones falsas y aquellos por los que no, se llega también al punto de descubrir dónde circulan estas falsedades, lo que lleva a detectar la desinformación como tal. Por una parte, está la detección de información falsa basada en cascadas, a partir de los valores de similitud entre cascadas que entrenan los algoritmos para separar las ramificaciones con y sin falsedades, o a partir de la representación con redes neuronales de las propiedades de las cascadas, que también diferenciarán entre ambos tipos. Por otra parte, se encuentra la detección basada en la estructura, llamada así porque se distingue la información verdadera de la falsa dentro de los diversos esqueletos de un grafo [3] gracias a las representaciones del conocimiento de sus elementos dentro (por ejemplo, sus usuarios y sus contenidos) [46].

2.4.5. Monitorización continua y alerta temprana

Como se ha podido ver en los anteriores apartados, los modelos utilizados para representar grafos muestran que el intercambio de información (y de desinformación) no es estático. En los modelos virales, se aprecia la diferencia entre un contenido que no ha ido más allá de su origen y entre aquel que ya ha conseguido extenderse hasta las últimas ramificaciones de la

cascada [36, 7]; en los modelos epidémicos, los estados de la infección en los usuarios también son cambiantes en función a si les llega la desinformación o no [5], y los modelos temporales son los que precisamente muestran cómo el paso del tiempo modifica el grafo [6]. En resumen, luchar contra la desinformación no es capturar una foto fija en las OSNs, sino tener en cuenta su evolución para, de acuerdo a estos modelos, parar las cascadas, evitar más infectados y hacer que el tiempo no vaya en perjuicio de la información verdadera, en la medida de lo posible.

La literatura aborda varias formas para monitorizar las redes. Shu et al., además de cubrir la detección de la información falsa, también describen los modos de mitigarla, paso posterior en la lucha: primero, identificando el origen de la desinformación a partir del grado y la cercanía de los nodos para detectar los principales focos a falta del difusor inicial; segundo, detectando a los líderes al precisar con qué conjunto de usuarios llegan más mensajes falsos; tercero, estimando el tamaño de toda la red afectada a partir de los contaminados de muestras independientes entre sí (por ejemplo, una de Twitter y de Facebook); cuarto, haciendo campañas con contenidos verdaderos a las cuentas que reciben información falsa para contrarrestarla; quinto, minimizando el impacto al bloquear la parte del grafo que más favorezca el flujo de falsedades [3].

Dentro de las verticales del estudio del SNA mencionadas por Singh et al., aunque la difusión de la información o la maximización de la influencia apuntan a la detección de los patrones en la red ya creados, en la monitorización temprana entra en juego la predicción de enlaces citada antes. Esta detección de futuras uniones entre nodos se adivina a partir de técnicas de similitud entre los miembros de cada par, de la probabilidad condicional de que aparezca un nuevo enlace en base a los rasgos de la red o de la reducción de dimensiones a partir de *embeddings* de sus propiedades para comprobar los actores vecinos en el espacio latente [6]. No obstante, Li et al. incluyen también la predicción de la difusión de la información dentro del estudio sobre la maximización de influencia [181], donde el foco ya no es la unión entre nodos sino el salto del contenido que discurre, a través de compartir los mismos posts o de crear otros nuevos.

Esta reducción del impacto bebe de las formas de difusión en cascada y epidémica para prestar atención a las probabilidades para un nodo de contagiarse dentro de las ramificaciones de la propagación [3]. Eliminada de la ecuación cualquier intervención propia de las plataformas sociales, pasar a estas estrategias dentro de la monitorización implica incorporar también el rastreo de los desmentidos de los *fact-checkers*.

Shao et al. estudiaron esta contraposición entre falsedades y *fact-checking*, donde cobra importancia la respuesta rápida de los agentes contra la desinformación [204]. En tal estudio, se toma en cuenta el tiempo de reacción entre los *posts* con afirmaciones falsas y aquellos con sus *fact-checks*. Dentro de los modelos epidémicos, conciben este problema como un análisis de supervivencia, comparando la duración para desarrollar la vacuna (el desmentido) y los afectados por el virus (la desinformación), del cual se evidencia el tiempo necesario hasta terminar la verificación para la muestra analizada [204]. Este tipo de enfoque también se planteó en otros escenarios, como las falsedades que pasan el filtro en Wikipedia hasta que son señaladas [205].

2.4.6. Enfoques actuales de SNA para luchar contra la desinformación

El uso de técnicas de SNA ha servido para analizar varios fenómenos de la desinformación y ver similitudes y diferencias entre estos. Uno de ellos fue a través del uso del *hashtag* '#FilmYourHospital' [40] (en relación a la información falsa de que los hospitales estaban vacíos y no saturados en el punto álgido de la pandemia de la COVID-19). Se descubrieron tres grupos mayoritarios

como altavoces de falsedades y un grupo aislado más minoritario. Aquí predominaron los enlaces a YouTube y la presencia de *bots* se consideró reducida [40].

En otra investigación relacionada, el *hashtag* '#COVID19', pero también el término genérico 'coronavirus' sirvieron para el análisis en Twitter (X) de varios escenarios de desinformación. Si bien estas formas de búsqueda aluden a las cuestiones más globales de la pandemia, Pascual-Ferrá et al. recopilaron los posts de tres eventos a analizar (declaraciones de la COVID-19 como emergencia de salud pública, como pandemia y como amenaza global en activo después) para controlar toda la conversación alrededor de esos momentos claves. Sus visualizaciones mapearon los nodos con mayor centralidad de grado y detallaron los usuarios con más menciones, más publicaciones y más *reposts*. Descubrieron que las cuentas de salud pública más repetidas perdían fuerza conforme avanzaba la pandemia en favor de figuras políticas y concluyeron, a partir de los grafos, que estos debates estaban muy divididos, descentralizados y con conexiones débiles [206].

Los trabajos de SNA también se centraron en las campañas de desinformación una vez que empezaron las narrativas sobre la vacuna de la COVID-19. En este sentido, Durmaz y Hengirmen compararon las conversaciones antes y después de la pandemia en franjas similares de tiempo mediante el *hashtag* '#aşı' ('vacunación' en turco, el idioma elegido para este estudio) a partir de todas las interacciones en Twitter. El color y el tamaño se diseñaron en función a las centralidades de grado en unos grafos, y a las centralidades de intermediación en otros, así como también el color sirvió para distinguir en otras visualizaciones entre antivacunas, provacunas y el resto. Comprobaron la gran influencia del ministro de salud, seguido por los medios convencionales según estas métricas, y el aumento de los usuarios en la conversación. En concreto, los antivacunas fueron 22 veces más que los provacunas y pasaron de casi un 2 % a más de una cuarta parte del grafo [207].

Pero la desinformación no es solo Twitter ni es solo la COVID y su vacuna. Para Instagram, Masey et al. también se valieron de los *hashstags* como la clave en el SNA para mapear las falsedades sobre la vacuna del VPH ('#HPV', '#HPVVaccine' y '#Gardasil' en este caso). También enfocaron la centralidad según el grado de los nodos y definieron su tamaño por la cantidad de *likes*. Los colores diferenciaban, entre otros, antivacunas y provacunas (relleno del nodo) y narrativas personales frente a información y fuentes (borde del nodo), mientras que las formas geométricas categorizaban los tipos de actores de estos contenidos. Entre los descubrimientos, se comprobó que los posts antivacunas venían más de cuentas de individuos no vinculados a la salud, no de colectivos, e incluían más narrativas basadas en su experiencia personal. Como con la vacuna de la COVID, las conspiraciones y falsedades sobre las enfermedades que provocan formaban parte de estas cadenas de desinformación [79].

Estas metodologías encuentran elementos comunes: la descarga de publicaciones, la unión de cuentas o mensajes en función a las interacciones, las medidas de centralidad y la separación tanto de comunidades como de tipos de usuarios. Trabajos como el de Duzen et al. toman estos pasos como módulos para que el objetivo final, más allá del análisis de la campaña específica, sea la creación de una herramienta capaz de desglosar el fenómeno, sea el que sea. Entre sus módulos, destacan el uso de Iffy.news para etiquetar los posts no fiables, el empleo conjunto de varias centralidades y la implantación de algoritmos para detección de comunidades, probados después en un caso real para demostrar la utilidad de su creación en un escenario concreto [208].

2.4.6.1. Lecciones aprendidas de estos enfoques de SNA

Ahmed et al. advierten de que dentro de los posts con sus *hashtags* analizados sobre la COVID, aparecían contenidos sin relación y aquellos que contradecían las falsedades en vez de seguir su senda. Aunque el foco de su estudio es la conspiración detrás, los autores ven necesario distinguir entre el porcentaje de posts falsos frente a otros [40]. En su otro trabajo sobre la diseminación de la supuesta ligazón entre el coronavirus y el 4G sí hicieron un análisis de contenido sobre las posiciones respecto a esta afirmación [41]. No obstante, los académicos realizan todo este proceso manualmente.

Esta cuestión manifiesta cómo el SNA pone el foco en las conexiones pero, a su vez, lo pierde en el análisis de los contenidos del flujo, los cuales pueden producir nuevos *insights*, cambiar el sentido de esas uniones y/o discernir entre otros grupos. Es el caso de los resultados de Durmaz y Hengirmen, que también indicaron la ausencia de análisis de los mensajes en su investigación y la dependencia a los *hashtags*, ya que aquellos asociados a proclamas antivacunas pueden contener, sin embargo, enunciados de provacunas y viceversa [207].

Aunque Massey et al. sí tuvieron en cuenta los contenidos a favor y en contra de las vacunas para el análisis guiado por los *hashtags*, también la búsqueda es dependiente a ellos y tampoco se describen métodos computacionales para separar unos mensajes de otros. Estos autores exponen además la desventaja de no ofrecer un análisis temporal de los posts, en este caso de Instagram, para ver tendencias de la dinámica en las publicaciones. Así, aunque sus visualizaciones sí distinguen entren las falsedades dentro de los nodos, no se aprecia cómo evolucionan a lo largo del tiempo [79].

Entre las barreras de Pascual et al., se señalaron aquellas relacionadas con la API por un lado, por no obtener más de un número determinado de posts cada cierto tiempo, y con los usuarios de la red por otro, porque que no interactúen con estos mensajes no significa que los ignoren, y pueden reproducirlos después [206]. Ambos problemas sugieren una cuestión de fondo sobre los posts de descarga: por una parte, en cuanto a no poder capturar el total de textos por las restricciones de las herramientas; por otra, en cuanto a discriminar otros contenidos no recogidos que, sin necesidad de interactuar, también formen parte del flujo de la información y sus desórdenes.

Estas cuestiones son también cruciales para caracterizar y mitigar la desinformación, también en las herramientas creadas al respecto. Entre los pasos de la solución sugerida por Duzen et al., se presenta la recogida de datos pero sin ir más allá del uso de *keywords* y también se especifica la anotación de estos, pero está basada en encontrar las fuentes no fiables listadas en Iffy.news y no en una comprobación real de si un contenido es verdadero o falso [208]. Esto indica que cada una de estas fases de la metodología se puede modificar para culminar en un grafo que refleje de una forma más realista el flujo de desinformación.

2.5. Integración de PLN y SNA en la lucha contra la desinformación

Como se ha visto en los anteriores apartados, el contenido apela el uso de PLN para el tratamiento del lenguaje natural de las redes sociales y el contexto reivindica el empleo del SNA para las conexiones dentro de estas plataformas [32, 53]. Por ello, cuando se trata la desinformación es necesario hablar tanto de PLN como de SNA.

El PLN y el SNA son las dos disciplinas sobre las que sustenta la lucha contra la desinformación desde la parte computacional [32] y, como ya se ha explicado previamente, esta investigación tiene como objetivo unirlas para que su aplicación no sea individual sino conjunta. A través de un enfoque semiautomático, se expondrán los modelos del lenguaje actuales en el campo del PLN por un lado, y la explotación de métricas de redes sociales con la generación de grafos en el ámbito del SNA por otro.

Sin embargo, mientras que Bondielli y Marceloni aluden estrictamente a las características del texto para el contenido y a las internas de la red social para el contexto [53], Montoro-Montarroso et al. se refieren al contexto para todo el mensaje en sí, información a la que podría llegar el PLN actual. En concreto, dentro de los elementos contextuales mencionados (usuarios, mensajes, estructura), precisamente hacen referencia al procesamiento textual [32]. Con el carácter multimodal de las publicaciones, el texto no solo es la fuente de contenido sino que también construye todo lo que lo rodea. Ponen de ejemplo el trabajo de Zhang et al. sobre la construcción de la 'escena visual' de los posts (lugar, estación del año, meteorología) para ver diferencias estadísticas en la información verdadera y falsa [209]. Esto es, en esencia, una extracción de los *insights* con una metodología para abordar las cuestiones más contextuales.

Así, contenido y contexto son dos caras de la misma moneda. Esto se puede ver con las propiedades de los *embeddings*, que recogen todas las cuestiones semánticas inherentes al texto, y en concreto las de los *Transformers*, que también se refieren al contexto, en este caso para dar un significado distinto a la palabra en función al resto del enunciado [127]. Más allá de los matices a los que puede llevar esta explicación, este estado de la cuestión sugiere que el contexto no tiene por qué ser solo territorio del SNA.

Los tipos de análisis dentro del SNA [37] también se ofrecen a la unión con el PLN: por un lado, los análisis de estructura sobre las uniones y composición del grafo pueden definirse por las relaciones en el texto; por otro, los análisis del contenido para generar conocimiento pueden contar también con las propiedades textuales para conformarlo. Los grafos, como producto del SNA para ordenar el *big data*, asumen explotar todo lo que hay dentro de él: textos, audios y vídeos, además de posts, para encontrar patrones. Esto no es nuevo y bebe de la aplicación del SNA en la multimodalidad de los documentos en la archivística [174].

Así, los cuatro enfoques de Camacho et al. [37] también pueden plantearse con el PLN como ingrediente del SNA: descubrir (qué hallar cuando se combinan ambas disciplinas), integrar (qué incluir entre las propiedades del texto al grafo), escalar (hasta qué punto se pueden trasladar estas propiedades) y visualizar (qué se puede mostrar en esa combinación). De esta forma, el flujo general de las ramas de datos tiene sentido en el objetivo particular de abordar la desinformación: primero procesando las falsedades de una red social; luego extrayendo sus propiedades intrínsecas de estas plataformas más las propias del lenguaje natural; después explotando esto al máximo para recrear toda la conversación sobre el contenido falso más allá de la detección de este, y por último visualizando este ecosistema alrededor de la desinformación en grafos junto a las variables obtenidas en este proceso.

Son varios los estudios de SNA donde la parte de PLN aparece intrínseca. Singh et al. aluden al Análisis de la Información del Lenguaje (LIA) como forma para procesar a través del texto la dimensión psicológica, social y emocional entre conexiones, ya que su meta es inferir a partir de todo este contexto las relaciones y posiciones de los usuarios en las comunidades y en la red [8]. Los tipos de análisis en este aspecto, como los presentados por Szurawitzki, tienen en cuenta cuestiones como la cantidad de palabras, el idioma, el estilo, la semiótica o el discurso [210].

Sapountzi y Psannis desglosaron en otro estudio los retos en la rama de SNA, entre los que se encontraban los relacionados con el PLN: análisis de sentimiento, extracción de la opinión, análisis de tendencias, captación de temas y, en general, minería de texto. Fue esta minería de texto una de las funciones que pormenorizaron después para cada uno de los softwares de SNA, junto a las de otras métricas propias de esta disciplina y cuestiones como la detección de comunidades. Se demuestra así cómo estas herramientas tampoco se apartan de la parte textual, ya existente en el momento de tal investigación a través de paquetes de Python (en NetworKit) o de R (en Statnet), de una API (en Gephi) o de actividades propias de análisis de sentimiento (en NodeXL). Los autores extienden el abanico a herramientas de monitorización de contenidos, también con enfoques de PLN [196].

2.5.1. Enfoques actuales de PLN y SNA en la lucha contra la desinformación

Varios estudios actuales apuestan tanto por el PLN como por el SNA para aportar nuevos enfoques contra la desinformación, pero menos son los que consiguen combinarlos. Por ejemplo, Bahja y Safdar presentan una metodología que depende de ambos campos para sus exploraciones: por un lado, la extracción de los temas de los posts sobre la COVID-19 en Twitter (X) con *Latent Dirichlet Allocation* (LDA) y análisis del sentimiento; por otro, la creación de grafos con los bigramas resultantes de las publicaciones. Sin embargo, estos enfoques no se implementan de manera combinada, y no se enriquece el grafo con los resultados del LDA y los *insights* del sentimiento, si bien se da un paso más allá con el empleo de la parte textual para las uniones de los nodos [211]. Lo mismo ocurre con el trabajo de Kalantari et al., que separan detección de temas, SNA y análisis de sentimiento para los posts de Twitter, en vez de enriquecer las conexiones del grafo con las otras propiedades [212]. Li et al. profundizaron en los errores de su clasificador de contenido falso en Weibo a través del análisis de temas, pero también apartaron estos métodos de su aplicación del SNA para entender la difusión de mensajes en esta red en conjunto [213].

Esta unión sí la realizan Paraschiv et al. de un *dataset* de posts de Twitter (X) extraídos a partir de su API, en este caso de las elecciones de 2016, con los que recrearon las cascadas de contenidos a partir de los usuarios registrados y sus interacciones con los posts. Como la API solo documenta la publicación y usuario de origen pero no la cadena de cuentas por las que pasa el post entre medias, los autores simularon una cascada a partir de los seguidores de cada usuario también involucrados en la difusión y capturaron con *embeddings* las propiedades semánticas entre ellos. Es aquí cuando el PLN entra en juego con la detección de los temas de cada una de las cascadas a partir de *Hierarchical Dirichlet Process* (HDP) y el análisis de sentimiento con *Transformers* para cada uno de los posts iniciales que comienzan cada estructura en árbol. El producto final es un metagrafo que entrelaza las cascadas en función a los perfiles comunes, temas tratados y sentimiento [214].

Dentro de las combinaciones de PLN y SNA, la literatura incluye formas de crear estructuras en base a las propiedades textuales. Sivasankari y Vadivu parten del hecho de que las conexiones entre contenidos no tienen por qué estar ligadas a las interacciones o mensajes originales, sino que es el parecido entre ellos el que define cómo una publicación se desplaza desde su origen hasta otros puntos de la conversación. En concreto, ellos utilizan los posts del *dataset* LIAR y los une si entre ellos hay al menos una similitud de 0.7 o más en el coeficiente de Jaccard [215]. No obstante, esta asunción excluye el análisis del trazado real de los posts en las OSNs como punto clave del SNA y no usa el PLN como una capa añadida para enriquecer el grafo de la plataforma social, sino que define el propio grafo en sí.

2.5.2. Herramientas y claves para monitorizar y analizar la desinformación en redes sociales

Los apartados de la sección del marco teórico sirven de guía para la elaboración de metodologías para trazar el recorrido de la información falsa y analizarla. Estas explicaciones demuestran que la aplicación conjunta del PLN y SNA, más allá de ser ramas aliadas contra la desinformación [32], pueden pasar de solo complementos a herramientas a la vanguardia en esta batalla gracias a las técnicas actuales descritas. Esto se aprecia con las carencias de cada uno de los campos de forma separada mientras el otro de ellos aporta un valor añadido.

Por un lado, se ha comprobado cómo el PLN ha bebido del SNA en su foco por las OSNs, en cuestiones como la descarga de los posts de una API para su análisis masivo, y el rastreo de todas las publicaciones que contengan *keywords* o *hashtags*, elementos destacados de estos ecosistemas virtuales. Pero la evolución desde los modelos de n-gramas y recuentos tradicionales hasta los *Transformers* [127] revela que la minería del texto en las OSNs para desgranar sus dinámicas ya puede ser más avanzada, desde la forma de extraer estos mensajes hasta la curación de contenidos después para no depender solamente de palabras claves como filtro.

Por otro, el SNA también se ha alimentado del PLN, tal como se ha visto en los estudios de esta disciplina que valoran los análisis del lenguaje, incluso con la integración de técnicas de procesado de texto en sus softwares para la creación y análisis de grafos. Pero este acercamiento a las vías para procesar el texto no suele traspasar los enfoques más tradicionales, en perjuicio de los *Transformers* [127]. Las visualizaciones realizadas de los grafos, dimensión final del SNA [37], no tienen en cuenta por lo general los LLMs para enriquecer el resultado.

Como se ha apreciado en estos apartados, el *fact-checking* semiautomático es una transversal que daría un enfoque actual al tratamiento de las OSNs dentro del PLN y que a su vez otorgaría nuevos *insights* al resultado de las visualizaciones de SNA. Las herramientas adquieren así un nuevo escenario donde ya no solo se exploran las métricas básicas, el sentimiento, la opinión o los temas del texto sino directamente la alineación con los enunciados falsos para que el grafo muestre qué es desinformación y qué no, para entender mejor sus dinámicas. Los modelos de NLI también evocan el ML tradicional, pero han evolucionado junto a los *Transformers* [127] para tener cabida dentro del estado del arte del PLN, cuyo conocimiento se puede traspasar al SNA.

Capítulo 3

Cribado de la información mediante similitud semántica

Es falso que un grupo de inmigrantes saqueara un restaurante en Tenerife.

— EFE Verifica

El marco teórico ya abordado revela los avances de los *Transformers*, de las áreas del PLN como el NLI y del *fact-checking* semiautomático, entre otros. Esto abre el camino a abordar nuevas propuestas frente al uso tradicional de modelos de clasificación, previamente entrenados en *datasets* cerrados que contienen conjuntos de enunciados clasificados como verdaderos y falsos, entre otras categorías. En esta sección se presenta como primer paso de esta alternativa la representación de *embeddings* semánticos, tomando como punto de partida los desmentidos de los *fact-checkers*.

La metodología para lograr estas representaciones del conocimiento abre esta sección (3.1), con la obtención de los textos, la visualización en un espacio bidimensional y el análisis adicional de las falsedades a lo largo del tiempo. Después, se abordan las formas para lograr mejores resultados en este tipo de tarea (3.2), tanto a nivel de modelos como de técnicas de reducción de las dimensiones, y se expone un caso de uso más allá de la cuestión de la desinformación (3.3).

3.1. Representación de la desinformación mediante *embeddings*

La clasificación tradicional de contenidos de forma binaria, como verdaderos o falsos, es una manera de combatir la desinformación a través del aprendizaje automático, pero no la única. Este flujo de trabajo presenta una debilidad: se depende de un *dataset* cuyo etiquetado se produce de forma estática, limitando así el *fact-checking* a solo los textos usados en el entrenamiento, al conjunto de dos o más etiquetas utilizado [33, 216, 32] y a un intervalo temporal concreto.

Esto motiva una corriente alternativa donde las fuentes fiables son el epicentro para mejorar las decisiones asistidas de forma computacional [32]. Si el *dataset* deja de ser la base para el entrenamiento y se convierte en la base para el conocimiento [217], ya la tarea no consistirá en

la clasificación binaria o multicategoría. Así, el *dataset* lo compondrán los *claims*, enunciados de cada información falsa, y podrá ser actualizado con cada uno de los siguientes hechos señalados oficialmente como falsedad. Aquí no sería necesario aumentar de manera masiva los datos de un set de entrenamiento para mejorar la predicción, sino contar con un *claim* por cada información falsa nueva. En consonancia, el cometido en este caso será directamente comprobar si un mensaje coincide con alguno de estos *claims* del *dataset* y pondrá así a los *fact-checkers* en el centro de la ecuación en un proceso no automático pero sí semiautomático [9].

Para llevar a cabo este proceso es necesario contar con modelos que puedan evaluar el grado de alineamiento entre un enunciado candidato (la nueva información a contrastar) y cada uno de los hechos señalados como falsos por los verificadores en una base de datos (ver Fig. 3.1). Esto se realiza en esta investigación a través de las tareas de inferencia del lenguaje natural o NLI. Un modelo entrenado para esta tarea confronta normalmente dos textos y devuelve tres probabilidades, determinando entre ellos el grado de alineamiento (*Entailment*), de contradicción (*Contradiction*) y de ausencia de relación (*Neutral*).

Pero antes, la aplicación del NLI plantea un reto fundamental: ver con qué afirmación dentro de la base de datos se debe emparejar cada nuevo enunciado a contrastar, un trabajo muy lento y costoso. Por ello, en esta investigación se ha diseñado un proceso de dos pasos. El primero, descrito en esta parte de la tesis, implica un cribado mediante el cálculo de distancias semánticas entre textos. De este modo, se consigue reducir el conjunto de hechos a aquellos relevantes en términos semánticos. En segundo lugar, se aplica el modelo de NLI, esta vez únicamente sobre el subconjunto de los hechos más relevantes. Este segundo módulo se explica en el Capítulo 4.

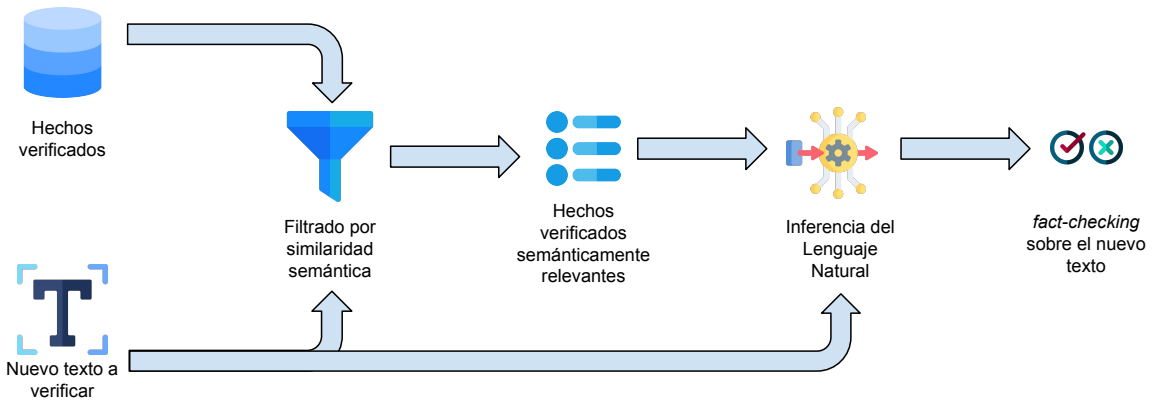


Figura 3.1: Diagrama general del enfoque propuesto para la realización de *fact-checking* semiautomático.

La similitud entre textos es una tarea ya estudiada dentro del PLN. Esto incluye una diversidad de métodos, como aquellos basados en el cálculo de distancias entre vectores generados con BOW y tf-idf o métodos basados en grafos que relacionan de manera predeterminada unas palabras con otras [218]. Pero, ante solo estos métodos, la literatura topa con la necesidad de formas más precisas de calcular la similitud semántica [219]. Los *embeddings* generados mediante arquitecturas *Transformer* [127] consiguen dar un valor semántico real a cada término *per se* en base a su significado y en relación al resto del texto y permiten innovar respecto a las tareas tradicionales de clasificación.

Una vez generada una buena representación contextual y semántica [219, 218], el cálculo de la

similitud tradicionalmente se realiza mediante la distancia coseno [139], que para dos vectores u y v viene dada por la ecuación:

$$CosSim(u,v) = \frac{\sum_{i=1}^{N} u_i v_i}{\sqrt{\sum_{i=1}^{N} u_i^2}\sqrt{\sum_{i=1}^{N} v_i^2}} = \frac{\langle u,v\rangle}{\|u\|\ \|v\|} \tag{3.1}$$

donde N representa el número de dimensiones del vector o *embedding* u y v, $\langle u,v\rangle$ es el producto escalar entre dos vectores, y $\|.\|$ es la norma L2.

Esta combinación ha sido ya evaluada con éxito en la lucha contra la desinformación [152, 35]. Permite también lidiar con contenidos con lenguaje muy variado, como el de WhatsApp [220], con el que es más difícil crear una base de datos y, por tanto, utilizarla como entrenamiento para una clasificación al uso. Esta corriente dentro de la IA que mide la cercanía matemática entre mensajes se ha desarrollado para contextos como el coronavirus, potenciando la mejora de las bases de conocimiento a partir de *claims* para casos concretos [221].

Una vez obtenido el conjunto de hechos ya desmentidos sobre los que construir esa base del conocimiento, es posible precalcular sus *embeddings*, evitando así calcularlos repetidamente en toda comprobación futura. Por contra, de cada nuevo enunciado a verificar sí se tendrá que computar su representación en tiempo real.

3.1.1. Generación de *embeddings* semánticos de informaciones falsas mediante modelos de lenguaje

Como se ha indicado previamente, en esta tesis se indagará en la similitud semántica gracias a los modelos *Transformer* actuales, en línea con previos estudios [152, 35]. El uso de estas representaciones vectoriales permitirá desenmarañar las conversaciones sobre desinformación de cara a comprender su funcionamiento, haciendo uso de técnicas actuales de PLN.

La semántica es el nivel del lenguaje que se ocupa del significado de una frase centrándose en las interacciones a nivel de palabra. El objetivo de esta investigación es inferir y comprender el contenido de los textos para hacer frente a la desinformación mediante la comparación de vectores o *embeddings* extraídos a partir de frases que condensan el nivel semántico del lenguaje. Mediante las características semánticas y contextuales de estos vectores, se evaluará el grado de similitud entre un nuevo enunciado y una base de datos de afirmaciones ya desmentidas. El resultado será un subconjunto de textos contrastados que garanticen un grado mínimo de similitud.

Para evaluar el enfoque propuesto en el dominio de esta investigación, se han elegido 20 bulos descubiertos por LatamChequea Coronavirus [100], principal repositorio de *claims* falsos en español sobre la COVID-19, de acuerdo a la IFCN. Lo forman 35 organizaciones de *fact-checking* hispanoparlantes bajo la coordinación de Chequeado. Cada uno de estos enunciados falsos se ha almacenado y, a partir de ellos, se han realizado búsquedas de posts relacionados en Twitter mediante su API. Para ello, se han construido cadenas de búsqueda o *queries*, compuestas por combinaciones de palabras claves de cada falsedad y operadores lógicos. Para este experimento, las *queries* se han creado de forma manual, incluyendo expresiones con el mismo sentido y sinónimos hacia una búsqueda más optimizada de todas las publicaciones alrededor de la desinformación entre los usuarios de habla hispana. Estos 20 bulos se encuentran en la Tabla 3.1.

ID	Falsedad	*Fact-checker*
0	Las vacunas de ARN mensajero puedan hacer que seamos un ser transgénico	Maldita.es
1	Las vacunas contra la COVID-19 causan convulsiones	Newtral.es
2	Estados Unidos admitió que solo un 6 % de las muertes informadas fueron realmente por coronavirus	Chequeado
3	Las mascarillas causan enfermedades neurodegenerativas	Maldita.es
4	Una imagen de una patente en Países Bajos de un método para «testear el COVID-19» desde 2015	Newtral.es
5	La vacuna contra el coronavirus te puede dejar estéril	Chequeado
6	Hay un plan diseñado para el COVID-19 desde 2017 en documentos del Banco Mundial	Mala Espina Check
7	Se ha descubierto que la vacuna contra la COVID-19 destruye nuestro sistema inmunológico de forma permanente	Maldita.es
8	Beber mucha agua y hacer gárgaras con agua caliente y sal elimina el coronavirus	AFP
9	Se recomienda mantener el cuerpo en un estado alcalino	Ecuador Chequea
10	El eucalipto previene o elimina el nuevo coronavirus	AFP
11	La hoja del árbol de guayaba puede prevenir o revertir los efectos de COVID-19	Maldita.es
12	La Nasa catalogó al dióxido de cloro como antídoto universal en 1988	Animal Político
13	Tomar vino puede ser beneficioso frente al COVID-19	EFE Verifica
14	El uso de la mascarilla provoca muertes por neumonía bacteriana	Maldita.es
15	La vitamina C previene el virus	Maldita.es
16	Christine Lagarde dijo: Los ancianos viven demasiado y eso es un riesgo para la economía global	Chequeado
17	Existe una relación entre el laboratorio biológico chino de Wuhan, las compañías farmacéuticas Glaxo y Pfizer y personas como George Soros y Bill Gates entre otros	Maldita.es
18	El coronavirus muere a los 27^{o}C	Convoca
19	El científico Charles Libier fue detenido por crear el coronavirus COVID-19	Animal Político

Tabla 3.1: Lista de enunciados falsos en español a partir de distintos *fact-checkers*.

Por cada *query* a partir de cada *claim* falso, se han descargado posts pertenecientes, en mayor o menor medida, a las palabras claves de la conversación sobre tal desinformación. Estos posts tienen un ID para identificarlos con la falsedad a la que se refieren y distinguirlos de otros descargados a partir de otra *query*, y con todos ellos se ha creado una base de datos sobre desinformación relacionada con el coronavirus en Twitter.

La selección de las desinformaciones escogidas para representar la base de datos de posts extraídos de las *queries* se ha realizado bajo las siguientes condiciones: 1) Que la *query* del *claim* descargue una cantidad suficiente de posts a analizar a lo largo del tiempo; 2) Que estos posts descargados estén relacionados con el *claim*, es decir, que no solo coincidan las palabras claves sino también el tema tratado; 3) Que estas descargas no se refieran exclusivamente a los desmentidos por parte de los *fact-checkers*, que sí pertenecen a tal conversación sobre esa desinformación pero no son posts propagando la falsedad en sí.

El primer paso para este experimento, antes de aplicar los algoritmos, consiste en preprocesar el texto de las publicaciones descargadas. Esto implica quitar los símbolos de *hashtags* ('#'), saltos de línea y guiones bajos, más toda mención ('@'). Es después de esto cuando se ha realizado la conversión de texto a *embeddings* semánticos que guardan su significado en forma de *features* codificados numéricamente en vectores. En concreto, se ha aplicado el modelo *Transformer* DistilBERT [222], multilingüe y preentrenado. Dado que la tarea es representar las propiedades de los contenidos para comprobar la relación entre ellos, es necesaria la extracción del *token* CLS de la última capa de la red neuronal porque es el que contiene el vector del texto completo y, con él, las 768 *features* a contraponer con las de los otros contenidos para ver su proximidad o usarlas para tareas de clasificación.

Después, se ha realizado un PCA [223]. Esto permite simplificar las *features* de 768 a 650,

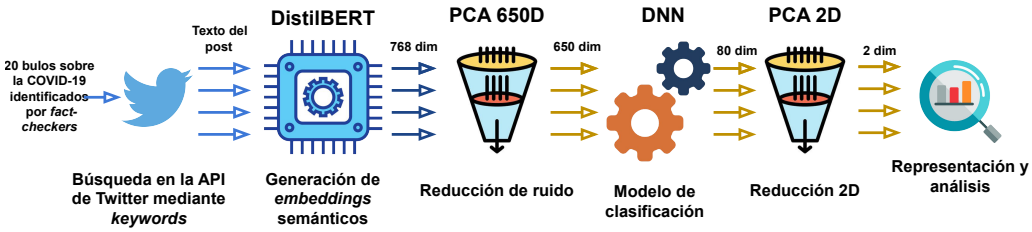


Figura 3.2: Pasos desde la obtención de posts hasta su representación en 2D, tras la conversión a *embeddings* y la transformación con redes neuronales.

eliminando el ruido con el cúmulo de todas las propiedades iniciales y, de esta manera, el modelo no estará sobreajustado y podrá generalizar en esta tarea en el set de testeo. Se ha tomado el set de entrenamiento, compuesto por la mitad de los posts, para tal PCA.

A continuación, se ha optado por una segunda transformación de los vectores a partir de una *Deep Neural Network* (DNN) para separar más entre los posts de una falsa información (identificados con un ID) de aquellos pertenecientes a otra (identificados con otro ID) para que las falsedades similares se distingan de las no relacionadas. Esta red neuronal consiste en una capa de 80 neuronas, *Rectified Linear Unit* (reLU) como función de activación [224], capa de *batch normalization*, capa de *dropout* (con 50 % de probabilidad) y capa de salida con *softmax* como activación [225], con *Adam* de optimizador [226] a la hora de entrenarla con el 50 % de los *embeddings*.

3.1.2. Visualización de *embeddings* de informaciones falsas

La salida de estas transformaciones son vectores resultantes de la capa oculta de la DNN, de 80 *features*. Dado que lo que se busca es la representación final bidimensional y no de 80 dimensiones, de nuevo se ha realizado PCA en los vectores para dejarlos en dos dimensiones. Este resultado permite la visualización de los posts de informaciones falsas separados de otras para comprender las similitudes y diferencias entre ellos a nivel semántico, como consecución del proceso (ver Fig. 3.2).

La visualización a partir de los *embeddings* en Fig. 3.3 como último paso de esta metodología se lee de la siguiente manera: cada post es un punto; el color del punto representa su ID (ver Tabla 3.1), es decir, la información falsa de la que proceden, y los ejes x e y son las dos dimensiones finales a las que se han reducido las 768 *features* iniciales y que sitúan a cada punto / post en un espacio de más proximidad o lejanía respecto a otro en función a su similitud semántica. En esta bidimensionalidad también se ha representado el *claim* del *fact-checker* indicado en la tabla 3.1 como punto de referencia destacado sobre el resto.

Las desinformaciones muestran similitud entre sí frente a las pocas que, por su lejanía, se diferencian del resto. En concreto, las informaciones falsas número 2, 5 y 16 se separan de las demás frente al solapamiento de las otras. Estas tres falsedades, cada una con un ID, son más distintas en cuanto a significado, de acuerdo a las *features* reducidas a dos dimensiones.

En cuanto a los posts relacionados con el bulo con el ID 2, la desinformación hace alusión a cómo, falsamente, en Estados Unidos la proporción que moría por COVID-19 era pequeña cuando irrumpió la pandemia. Si bien todos los IDs pertenecen a informaciones falsas sobre el coronavirus, se podría entender la separación de los posts con el ID 2 del resto porque la mayoría

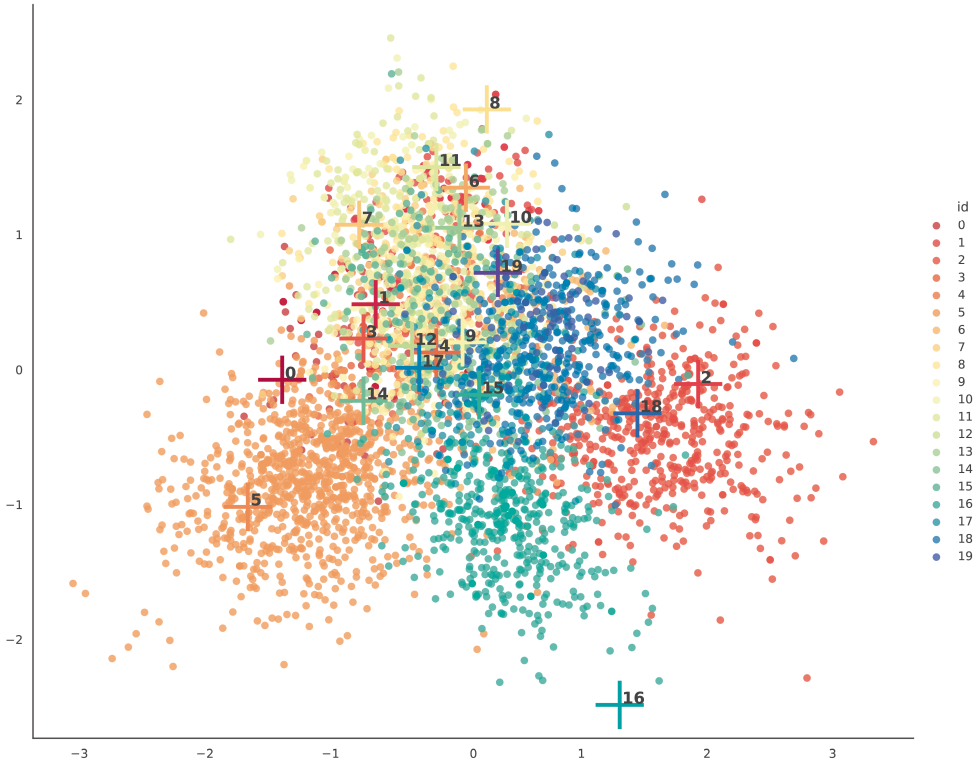


Figura 3.3: Visualización de los *embeddings* de la capa oculta tras el análisis de componentes principales para comprimirlos a 2D. Cada punto es un post; cada color, una información falsa.

de bulos son sobre curas erróneas o sobre conspiraciones de la procedencia de la pandemia. De haber en otra selección más desinformaciones que, como el ID 2, sean de negación de la catástrofe derivada de la pandemia, estas, por su significado, podrían estar más cercanas entre sí.

Respecto a las publicaciones con el ID 16, estas se refieren a la falsedad de que Christine Lagarde, presidenta del Banco Central Europeo, expresa que las personas ancianas "viven demasiado", enmarcada en el contexto de los mayores que morían por COVID-19. Dependiendo de si este aspecto del coronavirus aparece o no reflejado en el contenido dentro del post, se podría entender la separación de estos *embeddings* respecto a los restantes. En línea con la desinformación con el ID 5, no hace referencia a la tónica común de remedios o cuestionamientos de la pandemia.

Pero en cuanto a los contenidos con el ID 5, esta desinformación se asemeja a aquellas de las publicaciones con los IDs 0, 1 y 7 sobre los peligros de las vacunas. En concreto, difunde el riesgo a quedar estéril tras vacunarse y queda más próximo en el espacio bidimensional con la falsedad número 0, pero no tanto con aquellas pertenecientes al 1 y al 7. Dadas estas posiciones de los *embeddings*, a pesar de que los cuatro tipos de desinformación están bajo el paraguas común de los falsos problemas de las vacunas, la carga semántica del tema de la esterilidad podría afectar su posición respecto a los posts con otros IDs.

Por su parte, la cercanía entre las desinformaciones con los IDs 4, 12 y 17 se comprende por el nexo de ser cuestionamientos a instituciones, en estos casos científicas (creación de una patente en el marco neerlandés, declaración sobre el dióxido de cloro por la NASA y el vínculo de Wuhan con Pfizer y Glaxo, respectivamente). Coincide también la alusión a hechos pasados (en 2015 la

falsa patente, en 1988 la falsa afirmación y antes de la pandemia el vínculo entre compañías), impacte esto o no a la posición de los vectores. Pero más allá de las relaciones específicas, el tema común de la ciencia parece afectar. Por esto, el contenido con el ID 9, con términos científicos (los falsos remedios en base a un cuerpo alcalino) estaría también cerca, aunque no ocurre esto con aquellos con los IDs 6 y 9 dentro de este ámbito.

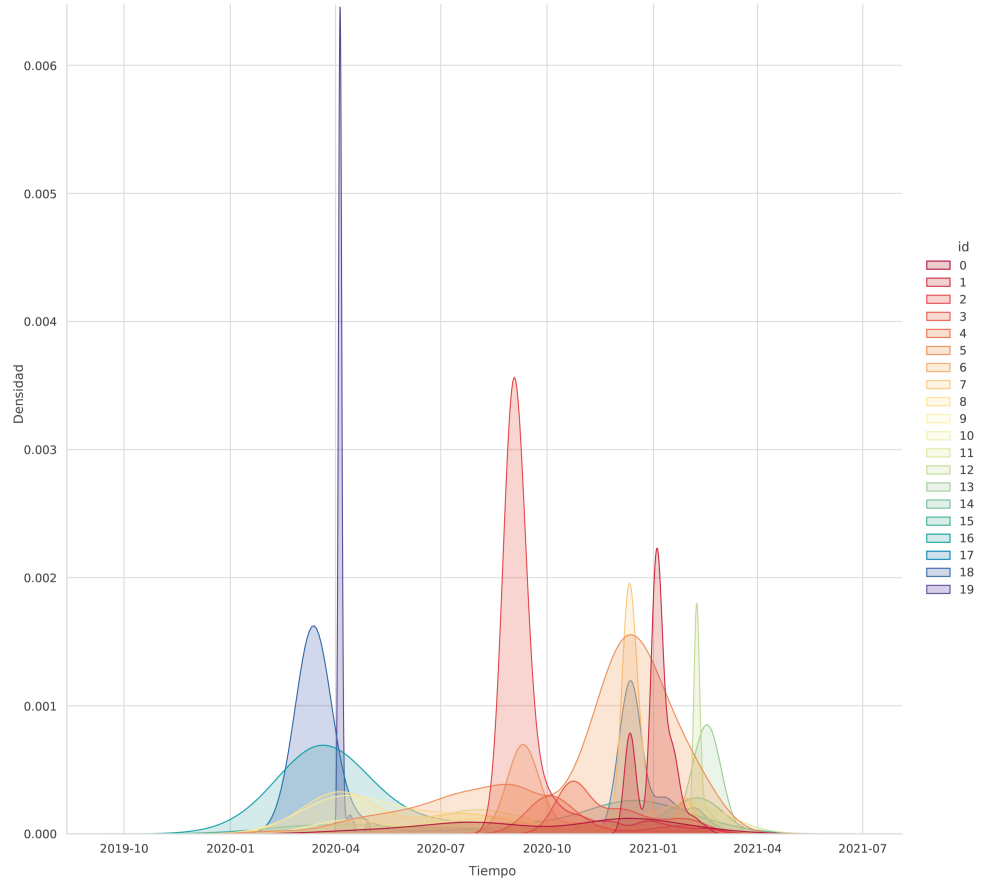


Figura 3.4: Evolución del número de posts observado por cada uno de los *claims* falsos (identificados con cada color) e indicados en la tabla 3.1.

3.1.3. Análisis temporal del número de informaciones falsas generadas

El potencial de estos *embeddings* abarca muchas otras posibilidades. Por ejemplo, es posible analizar el impacto temporal de la desinformación, observando cómo se produce un efecto ola (ver Fig. 3.4). Los contenidos con el ID 5 (falsa esterilidad por la vacuna), cuya segunda ola fue mayor en cantidad de posts, son un ejemplo de ello, y contrastan con posts como los del ID 19 (científico falsamente arrestado por crear la COVID-19), con un gran pico pero muy breve. Ninguna de las falsedades desapareció por completo durante 2020, año donde tuvieron protagonismo por la pandemia.

Las diferentes curvas de la infodemia en estos ejemplos muestran la necesidad de entender las distinciones por temáticas y entrever así las relaciones de la desinformación. Se aprecia cómo los posts con los IDs 0, 1, 5 y 7 sobre vacunas aumentaron en 2021 y al final de 2020, coincidiendo con el comienzo de las campañas de vacunación frente al coronavirus, y cómo la ola de los contenidos numerados como 0 es algo más llana que con aquellos pertenecientes al 5. Como

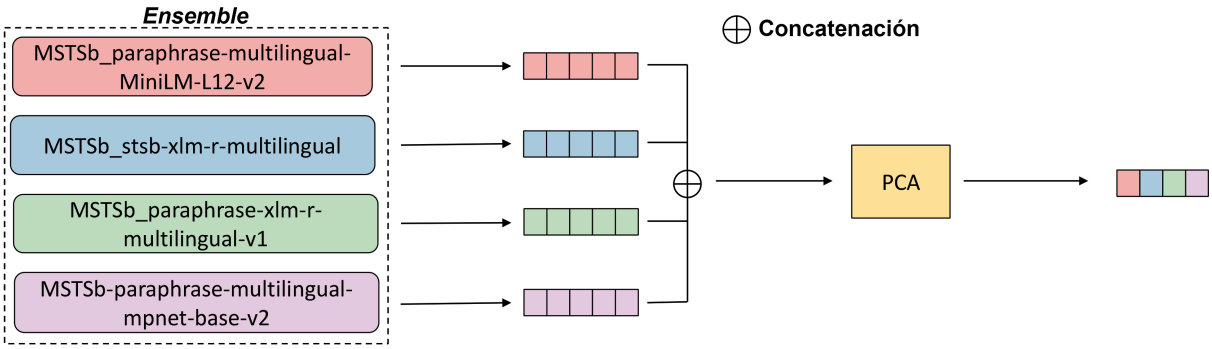


Figura 3.5: Enfoque de *ensemble* propuesto en FacTeR-Check [9], que incluye la aplicación de análisis de componentes principales para reducir las dimensiones de los *embeddings* concatenados procedentes de cuatro modelos multilingües de Sentence Transformers.

se puede comprobar, el uso de estas representaciones vectoriales es de gran utilidad para manejar computacionalmente la desinformación, desarrollando soluciones que permitan estudiarla, comprenderla y/o mitigarla.

3.2. Combinación de modelos para una mejor representación

Como extensión del trabajo mostrado en la anterior sección, el trabajo conjunto de FacTeR-Check [9] siguió explorando el uso de *embeddings* para incrementar su calidad. Así, se optó por construir un *ensemble* de modelos, un enfoque usado en la literatura para mejorar los resultados en múltiples problemas [227, 228]. Mediante una combinación de *embeddings* con diferentes características se consiguen representaciones más robustas y un mejor rendimiento que mediante los métodos basados en un único vector. Además, otra ventaja de los métodos *ensemble* es la ampliación de la cobertura del vocabulario.

3.2.1. *Ensemble* de modelos de similitud semántica

Entre los diversos modelos ya preentrenados para similitud semántica en FacTeR-Check [9], se investigaron los siguientes:

- `paraphrase-xlm-r-multilingual-v1`: una versión destilada de RoBERTa [229], entrenada en un gran *dataset* de datos de parafraseo usando XLM-R [230] como un *student model*.
- `stsb-xlm-r-multilingual`: BERT destilado [142] entrenado en NLI [160] y *Semantic Textual Similarity benchmark* (STSb) [231] usando XLM-R como un *student model*.
- `paraphrase-multilingual-MiniLM-L12-v2`: versión multilingüe de la versión MiniLM de Microsoft [232] entrenada en un gran *dataset* de parafraseo.
- `paraphrase-multilingual-mpnet-base-v2`: versión destilada del modelo MPNet de Microsoft [233] y entrenada en un gran *dataset* de datos de parafraseo usando XLM-R [230] como un *student model*.

En el *ensemble* propuesto en FacTeR-Check [9] (ver Fig. 3.5), el *output* se calcula como la concatenación de la salida de los cuatro modelos arriba indicados, todos ellos ajustados mediante

fine-tuning sobre el *dataset* de *Multilingual Semantic Textual Similarity benchmark* (MSTSb)[1], una versión multilingüe ampliada del STSb [231]. Normalmente, las tareas de Similitud Textual Semántica (STS) incluyen ejemplos compuestos por un par de frases y una puntuación que oscila entre 0 y 5 según el grado de similitud. STSb[2] comprende una selección de los conjuntos de datos en inglés utilizados en las tareas STS entre 2012 y 2017 de SemEval. Con el fin de trabajar en un escenario multilingüe, el STSb se amplió en tal trabajo a 15 idiomas diferentes utilizando la API de traducción de Google.

Estos modelos preentrenados se ajustan en MSTSb utilizando la función de pérdida por la similitud coseno de Sentence Transformers [234]. Para obtener los mejores resultados y evitar el sobreajuste, en esta investigación se organizaron los siguientes hiperparámetros utilizando el método de *grid search*: *learning rate*, *epochs*, *batch size*, *scheduler* y *weight decay*. Los valores de los hiperparámetros seleccionados y el modelo resultante están publicados en HuggingFace[3].

3.2.2. Aplicación de técnicas de reducción de dimensionalidad

Aunque el enfoque propuesto consigue mejorar el resultado de modelos individuales, sufre sin embargo de una complejidad elevada, tanto en memoria como en tiempo. Para reducir la carga durante el cálculo de distancias y la necesidad de espacio en memoria, en FacTeR-Check [9] se aplica PCA para disminuir el número de dimensiones de los vectores concatenados, tal y como se muestra en Fig. 3.5.

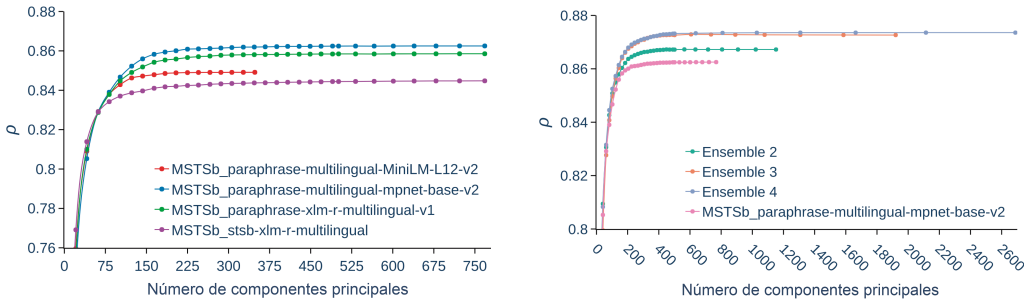
Para llevar a cabo una evaluación de este enfoque, se utilizó una porción para *test* del STSb multilingüe (generado con Google Translator). Los resultados generales en los conjuntos de prueba se muestran en la Tabla 3.2. Mientras que la columna 'EN-EN' se refiere al conjunto de datos original de STSb, 'EN-ES' y 'ES-ES' se calculan utilizando la versión traducida. Estos resultados revelan que el mejor rendimiento se obtiene con el modelo ajustado `MSTSb-paraphrase-multilingual-mpnet-base-v2`.

También se presentan los resultados obtenidos con diferentes combinaciones de los modelos en FacTeR-Check [9]. El mejor resultado se consigue con solo dos modelos: `MSTSb-paraphrase-xlm-r-multilingual-v1` y `MSTSb-paraphrase-multilingual-MiniLM-L12-v2`; el *ensemble* 3 agrega el modelo `MSTSb-paraphrase-multilingual-mpnet-base-v2`, mientras que el *ensemble* 4 incluye todos los modelos alcanzando un máximo de 2.688 dimensiones. Sorprendentemente, solo el *ensemble* 3 supera el mejor modelo individual, lo que implica incorporar más del doble de dimensiones (ver Tabla 3.3).

Como era de esperar, el uso de enfoques basados en *ensemble* aumenta masivamente el número de dimensiones. Así, y al igual que en el enfoque propuesto en la sección anterior, tal investigación conjunta aplica PCA para reducir la dimensionalidad. Se utilizan 90 mil oraciones paralelas que representan 15 idiomas[4], extraídas de tres recursos conocidos (TED2020[5], WikiMatrix [235] y OPUS-NewsCommentary [236]) para ajustar el PCA para cada modelo. La relación entre el rendimiento obtenido y el tamaño de reducción se muestra en Fig. 3.6. Como se puede ver tanto en el caso de modelos individuales afinados como en arquitecturas *ensemble*, el rendimiento converge con menos de 200 componentes principales, lo que proporciona una reducción sustancial

[1] `https://github.com/Huertas97/Multilingual-STSB`
[2] `http://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark`
[3] Modelos disponibles en `https://huggingface.co/AIDA-UPM`
[4] Los idiomas utilizados en este escenario son: *ar*, *cs*, *de*, *en*, *es*, *fr*, *hi*, *it*, *ja*, *nl*, *pl*, *pt*, *ru*, *tr*, *zh*
[5] `https://www.ted.com/participate/translate`

del espacio. El mejor espacio de PCA se selecciona de acuerdo con el rendimiento promedio en el conjunto de testeo de MSTSb en varios idiomas.



(a) Modelos individuales. **(b)** Modelos *ensemble*.

Figura 3.6: Selección del número de componentes en el conjunto de desarrollo MSTSb en FacTeR-Check [9]. Coeficiente medio de correlación de Spearman de los modelos individuales con *fine-tuning* (a) y modelos *ensemble* (b) utilizando la distancia coseno para 15 idiomas en función del número de componentes del conjunto de testeo ampliado de STSb. La media de los coeficientes de correlación se calcula transformando cada coeficiente de correlación en un valor z de Fisher, promediándolos y volviéndolos a transformar en un coeficiente de correlación.

El uso de *embeddings* semánticos a partir de los modelos *Transformer*, primero entrenados con una capa de aprendizaje profundo y después construidos a partir de *ensembles*, ha servido para mostrar la distancia entre los significados de los contenidos de varias desinformaciones, una vez que han sido reducidos a dos dimensiones para su visualización. Las publicaciones con el mismo ID, es decir, pertenecientes a la misma falsedad, aparecen por lo general juntas entre sí. La proximidad entre posts con IDs diferentes y la lejanía a su vez de otros sugieren que, dentro de este ecosistema, discurren textos falsos con temáticas en común y otros pertenecientes a otras cuestiones.

La observación de los posts en la visualización, en base a la desinformación a la que aluden, ha intentado desgranar estas relaciones de cercanía o lejanía entre contenidos. Los enunciados de las informaciones falsas se han utilizado para barajar posibles vínculos, ya que las propiedades de las dos dimensiones mediante la compresión de todas las *features* no descubren las causas entre distancias como tal. En todo caso, este análisis pone sobre la mesa cómo relaciones como la conspiración, las vacunas o los remedios son temáticas generales que pueden afectar el espacio bidimensional, por lo que se podrían agrupar los mensajes en estos grupos más genéricos en experimentos y gráficos análogos, en la línea de estudios previos [237]. LatamChequea Coronavirus [100] tenía etiquetas con estos macrotemas pero no eran uniformes para todo el *dataset*.

En resumen, este experimento introduce con éxito la alternativa de relacionar los posts con una desinformación sin acudir a los enfoques tradicionales. Gracias a los modelos *Transformer* y a cómo concentran sus *embeddings* las propiedades semánticas, se puede ver qué mensajes en una conversación pertenecen a una información falsa, y ello abre las puertas a una línea de investigación fuera del marco común de la detección clásica de desinformación con aprendizaje supervisado.

Modelo	Dim	EN-EN		EN-ES		ES-ES		Avg	
		r	ρ	r	ρ	r	ρ	r	ρ
MSTSb_paraphrase-mltl-MiniLM-L12-v2	348	85.26	86.17	81.45	81.49	83.30	83.68	81.38	81.47
MSTSb_stsb-xlm-r-mltl	768	84.21	85.10	82.65	83.04	83.20	83.83	81.75	82.09
MSTSb_paraphrase-xlm-r-mltl-v1	768	84.80	85.59	82.90	83.19	83.41	83.71	82.39	82.60
MSTSb-paraphrase-mltl-mpnet-base-v2	768	86.80	**87.40**	84.42	84.45	85.19	**85.52**	83.48	83.59
***Ensemble* 2**	1152	85.90	86.72	83.68	83.87	84.39	84.67	83.25	83.41
***Ensemble* 3**	1920	86.34	87.13	84.18	84.34	84.86	85.14	83.67	**83.84**
***Ensemble* 4**	2688	85.73	86.59	84.16	**84.53**	84.67	85.25	83.33	83.62

Tabla 3.2: Coeficiente de correlación de Spearman ρ y Pearson r entre la representación de las oraciones de modelos multilingües y las etiquetas para el conjunto de prueba de STSb.

Modelo + PCA	Dim	EN-EN		EN-ES		ES-ES		Avg	
		r	ρ	r	ρ	r	ρ	r	ρ
MSTSb_paraphrase-mltl-MiniLM-L12-v2	184	84.92	85.71	81.04	81.04	83.08	83.28	81.03	81.02
MSTSb_stsb-xlm-r-mltl	408	84.35	85.11	82.84	83.17	83.39	83.89	81.85	82.08
MSTSb_paraphrase-xlm-r-mltl-v1	286	84.79	85.50	82.73	82.97	83.38	83.58	82.23	82.39
MSTSb-paraphrase-mltl-mpnet-base-v2	306	86.69	87.27	84.21	84.28	84.93	85.19	83.20	83.28
***Ensemble* 2**	347	85.91	86.72	83.49	83.69	84.42	84.68	83.12	83.28
***Ensemble* 3**	367	86.64	87.55	84.50	84.80	85.24	85.72	83.85	84.21
***Ensemble* 4**	429	86.77	**87.78**	85.00	**85.52**	85.56	**86.20**	84.24	**84.71**

Tabla 3.3: Coeficiente de correlación de Spearman ρ y Pearson r entre la representación de las oraciones de modelos multilingües con reducción de dimensionalidad mediante análisis de componentes principales y las etiquetas para el conjunto de prueba de STSb.

3.3. Caso de uso: análisis del discurso político agresivo

La similitud semántica de enunciados dentro de una misma conversación en redes sociales en un espacio bidimensional se ha probado, gracias a este primer experimento de la tesis, en otros contextos fuera de la desinformación. Uno de ellos es el ámbito de la política en las OSNs, donde la distancia coseno a través de *embeddings* ha demostrado su validez en el trabajo conjunto con Torregrosa et al. para analizar el discurso agresivo político en Twitter (X) [238].

En concreto, en el ámbito de las elecciones de Madrid en mayo de 2021, se exploró el tono del discurso de los posts con técnicas computacionales cuantitativas, que hacían de filtro para el estudio cualitativo después (conformando así una metodología mixta) [238]. En total, se analizaron 252.881 posts en X (a partir de 61.926 originales) de 589 candidatos políticos pertenecientes a los seis partidos destacados para estas elecciones. Dentro de la parte cuantitativa con PLN se encuentran el análisis de sentimiento con VADER [239]; el análisis de n-gramas de los términos positivos y negativos resultantes, y, por último, el análisis semántico. Es en este último donde entran en juego las representaciones semánticas en el espacio latente, previas a la parte cualitativa.

Esta investigación muestra el éxito de la representación de la similitud semántica para cuestiones más acotadas. Mientras que los anteriores apartados han abordado estos métodos para tareas de desinformación a nivel genérico, empleando todos los posts descargados, para este caso se seleccionaron solo las publicaciones con los términos 'libertad' y 'libertades', palabras destacadas por la ambivalencia de su uso en la campaña. Tras el filtro, tales mensajes se convirtieron en *embeddings* con el modelo *Transformer* preentrenado `twitter-xlm-roberta-base` [240], reducidos mediante *Uniform Manifold Approximation and Projection for Dimension Reduction* (UMAP) [241] para

conservar la estructura de sus propiedades en el modelado final.

Dentro de la dispersión final, que refleja la variedad semántica de la conversación, el estudio demuestra mediante los *embeddings* la distinción del significado de las palabras 'libertad' y 'libertades' otorgado por el partido político que más las ha usado en la campaña frente al resto. Esto se ve en la representación en dos dimensiones, con estos vectores separados de aquellos que encapsulan el significado diferente de los otros partidos políticos. Incluso sin entrenamiento posterior con redes neuronales, la visualización final ilustra los diversos registros semánticos de palabras concretas a lo largo del discurso [238].

Capítulo 4

Hacia el *fact-checking* semiautomático mediante NLI

No, el nombre de Volodímir Zelenski no se traduce al español como "el maligno es dueño del mundo".

— Maldita

En esta sección se explica el segundo proceso necesario para implementar el enfoque de *fact-checking* semiautomático que se propone en esta investigación. Este segundo avance se conoce como inferencia del lenguaje, y supone ir un paso más allá a la hora de evaluar si dos enunciados son equivalentes. Mientras que dos frases pueden estar semánticamente muy relacionadas, no tienen porque estar alineadas. De este modo, una afirmación y su opuesta estarían altamente relacionadas en términos semánticos, pero serían contrarias entre sí.

La primera parte de esta sección comprende la explicación al detalle de esta subdisciplina del PLN en el ámbito de la desinformación a través de los *claims* y los enunciados de los posts, más el trabajo a nivel de evaluación, recopilación y ejecución de esta tarea a través de la monitorización de olas de informaciones falsas como caso de uso (4.1). La segunda parte presenta la aplicación del NLI para un análisis exhaustivo de posts sobre desinformación, y muestra así cómo se distribuye el contenido falso y aquel que lo contradice dependiendo del tipo y la cantidad de interacciones en la red (4.2).

4.1. NLI para contrastar desinformación

Como se explica en el Capítulo 3, las arquitecturas *Transformer* se sitúan como base dentro del PLN actual para el estudio de la desinformación, abandonando los enfoques tradicionales de clasificación [35, 34, 9]. Sin embargo, el *fact-checking* semiautomático no se reduce a contrastar una base de conocimiento a partir de *claims* y la similitud semántica entre contenidos. El experimento en la sección previa mostraba cercanía y lejanía de los posts entre ellos en función a las desinformaciones a las que más se asemejaban en significado, pero no desvelaba el papel de estos contenidos en esta propagación de la información falsa.

Como su nombre indica, la inferencia en lenguaje natural se utiliza para inferir la relación entre un enunciado (hipótesis) y una serie de afirmaciones verificadas (premisa) [217, 11]. Esta relación puede ser de alineamiento o congruencia, (del inglés *Entailment*), de contradicción (*Contradiction*) o de ninguna de ellas (*Neutral*). Mientras que la similitud semántica es incapaz de detectar estos matices más sutiles, un modelo de NLI consigue captar una relación de alineamiento o contradicción dado un par de enunciados. Si se logra comprobar que una afirmación implica una falsedad, se puede asumir con seguridad que tal enunciado lo apoya y, por tanto, que contiene información errónea. No obstante, es importante mencionar que en este proceso no se considera la intencionalidad que hay detrás de un enunciado, una cuestión que no se aborda en esta investigación.

Para describir mejor la tarea de NLI, dejemos que $\langle p, h \rangle$ sea un par de frases de un bulo y de una afirmación a comprobar, respectivamente. Utilizando NLI se puede inferir las probabilidades de *Entailment*, *Contradiction* y *Neutral*. De esta manera, se comprueba si el enunciado es una falsedad h_f (*Entailment*) o si no se puede determinar la naturaleza de este (h_u). Formalmente se quiere aproximar la Ec. 4.1.

$$f(p, h) \approx P(p|h_f) \tag{4.1}$$

donde p es la afirmación falsa desmentida por *fact-checkers* y se tiene la certeza de que implica tal falsedad, h es la afirmación a verificar encontrada por similitud semántica y h_f es el evento en el que la afirmación contiene información falsa. Por lo tanto, el propósito es encontrar una función f adecuada que sea capaz de aproximar esta probabilidad. Encontrar $P(p|h_f)$ es equivalente a hallar la probabilidad de alineamiento de $\langle p, h \rangle$. Por otro lado se puede decir con certeza que $1 - P(p|h_f) = P(p|h_u)$ ya que la contradicción y neutralidad de $\langle p, h \rangle$ no da una explicación significativa para h.

4.1.1. Alineamiento entre *claims* y hechos verificados

La aplicación del NLI en el campo de la desinformación necesita por tanto de los *claims* de la información falsa detectada por los *fact-checkers* y de las publicaciones en OSNs susceptibles de llevar estas falsedades (apoyándolas o desmintiéndolas) para convertirlos en las premisas e hipótesis, respectivamente.

Pero esta cuestión también requiere los modelos preparados con este fin: desde aquellos entrenados en el *dataset* SNLI, 570.000 pares de frases usadas de forma habitual como referencia en este campo [159], hasta MultiNLI con textos enriquecidos [242], o XNLI, interlingual para el uso de varios idiomas [34]. Del mismo modo que los *Transformers* suponen un paso adelante en la asociación semántica, la unión de estos con la inferencia del lenguaje permite comparar mejor entre enunciados en distintos idiomas.

Gracias a este uso de NLI, el análisis no trata solo de las publicaciones más compartidas, sino de todas aquellas que, con independencia de su impacto, contribuyen a la conversación en la difusión de la desinformación o en su contradicción. Aprovechando así arquitecturas *Transformer*, precisas en el cálculo de la similitud semántica [35] y ahora también en la inferencia [34], esta investigación se aparta de los usos tradicionales de clasificación de un texto como verdadero o falso. Por eso, cuando ya se combinan además con métricas propias de la red social, los resultados de NLI permiten una mayor comprensión del comportamiento alrededor de las falsedades surgidas en

estas plataformas online.

Como ya se ha apuntado en el estado de la cuestión, esta tarea de NLI se encuadra dentro del *fact-checking* automático en la medida que antepone la alineación de los contenidos con un *claim* (*claim matching*). Esto implica que un post pertenece a una afirmación que ya ha sido desmentida por los *fact-checkers* y, por tanto, no necesita pasar por el mismo proceso de verificación de nuevo para constatar que está expresando lo mismo que una falsedad o lo contrario [42].

Para construir este módulo de NLI, se optó por el modelo *Transformer* `XLM-RoBERTa-large` [230]. Las arquitecturas *Transformer* para NLI experimentan problemas cuando se transfieren a dominios no vistos [243], por lo que se ha puesto especial atención al proceso de *fine-tuning*. Para entrenar esta red se utilizan dos conjuntos de datos, XNLI [161] y *Sentences Involving Compositional Knowledge data set* (SICK) [244].

El modelo XLM-R se ajusta primero en XNLI partiendo de los pesos disponibles en el repositorio de Huggingface[1]. Tras este paso, se añade una cabeza de clasificación al modelo que incluye: 1) un *global average pooling* del último estado oculto del modelo del *Transformer*; 2) una capa lineal con 768 neuronas y una activación *tanh*; 3) un *dropout* de 10 % para el entrenamiento, y 4) una capa lineal clasificadora con una *softmax*. Esta cabeza de clasificación se entrena en SICK, congelando los pesos de XLM-R para preservar el preentrenamiento previo. Se optimiza mediante *Adam* [245] con una tasa de aprendizaje de 0.001. Los mejores pesos se deciden en el subconjunto de validación de SICK.

4.1.2. Evaluación

El módulo NLI se encarga de determinar la relación entre dos afirmaciones (una afirmación verificada) y una nueva afirmación de entrada. Como ya se ha avanzado antes, esta relación, ya sea *Entailment*, *Contradiction* o *Neutral*, se basa en diferentes probabilidades. Por tanto, hay que definir un umbral para asignar la etiqueta final. Así, una vez establecido este umbral y ya recibida la nueva afirmación a verificar, se contrastará mediante el módulo NLI con aquellas afirmaciones verificadas que existan en la base de datos por encima de un determinado grado de similitud semántica. Como resultado, si se encuentra un grado suficiente de vinculación, la nueva afirmación de entrada se etiquetará de acuerdo con la afirmación verificada encontrada.

Para evaluar el enfoque propuesto, se utiliza el conjunto de testeo de SICK. Los resultados se presentan en la Tabla 4.1. Para comparar, se incluyen los resultados de dos métodos de referencia: GenSen [246] e InferSent [247]. En el caso de GenSen, alcanza una precisión del 87.8 %, mientras que InferSent llega a 86.3 %. El enfoque propuesto alcanza un 87.7 % de precisión manteniendo las capacidades multilingües de XLM-RoBERTa, lo que resulta útil para contrastar información de falsedades en idiomas distintos o culturalmente separadas. Esto se representa en las secciones en español e interlingual de la Tabla 4.1, donde se calculan las mismas métricas. Se observa un ligero descenso de la calidad, debido sobre todo a que SICK es monolingüe, aunque los resultados en español e interlingual son bastante sólidos por sí solos, con un 82.9 % y un 85.3 % de precisión respectivamente. Es de destacar la alta precisión alcanzada por el módulo al mezclar idiomas, lo que posibilita un análisis de la desinformación sin que el lenguaje suponga una barrera.

[1] `https://huggingface.co/joeddav/xlm-roberta-large-xnli`

Idioma			*Precision*	*Recall*	*F1-score*	*Support*
Inglés (enfoque propio)	**Etiqueta**	***Contradiction***	0.9158	0.7486	0.8238	712
		Entailment	0.8475	0.8946	0.8704	1404
		Neutral	0.8856	0.9022	0.8938	2790
	Resumen	***Macro Avg.***	0.8830	0.8484	0.8627	4906
		Weighted Avg.	0.8791	0.8777	0.8770	4906
	-	***Accuracy***	0.8777	-	-	4906
Español (enfoque propio)	**Etiqueta**	***Contradiction***	0.8511	0.7388	0.7910	712
		Entailment	0.7446	0.9031	0.8162	1404
		Neutral	0.8797	0.8451	0.8461	2790
	Resumen	***Macro Avg.***	0.8251	0.8190	0.8178	4906
		Weighted Avg.	0.8369	0.8292	0.8296	4906
	-	***Accuracy***	0.8292	-	-	4906
Inter (enfoque propio)	**Etiqueta**	***Contradiction***	0.8825	0.8737	0.8072	1424
		Entailment	0.7925	0.8989	0.8423	2808
		Neutral	0.8828	0.8586	0.8705	5580
	Resumen	***Macro Avg.***	0.8526	0.8337	0.84	9812
		Weighted Avg.	0.8569	0.8534	0.8533	9812
	-	***Accuracy***	0.8534	-	-	9812
GenSen (solo inglés)	**Resumen**	***Accuracy***	0.878	-	-	-
InferSent (solo inglés)	**Resumen**	***Accuracy***	0.863	-	-	-

Tabla 4.1: Resultados del conjunto de pruebas de SICK. Los resultados en español se extraen de las traducciones automáticas del conjunto de pruebas del *dataset*. Los resultados interlinguales se obtienen emparejando indistintamente las instrucciones en español e inglés.

4.1.3. NLI19-SP: un *dataset* de NLI en español compuesto por bulos y hechos verificados por *fact-checkers*

Gracias a la operación conjunta de similitud semántica e inferencia del lenguaje, durante esta investigación se construyó el *dataset* NLI19-SP. Se trata de un conjunto de datos en español de desinformación acerca de la COVID-19, que incluye falsedades detectadas por *fact-checkers* así como posts de la red social X que las apoyan o desdicen. Para construir dicho conjunto de datos, se ha seguido un proceso de cuatro pasos:

1. **Recopilación de bulos:** se ha obtenido un conjunto de 61 bulos identificados por organizaciones de *fact-checking*.

2. **Generación de consultas:** es necesario construir consultas (*queries*) representativas con palabras claves para recuperar los posts de los bulos de la API de Twitter.

3. **Recuperación de posts:** utilizando FacTeR-CheckKey [9], se construye una consulta de búsqueda para cada uno de los bulos con el fin de descargar los posts relacionados con ellos de la API de Twitter.

4. **Filtrado por similitud semántica:** se implementa el módulo de similitud semántica para filtrar los posts relacionados semánticamente con cada bulo.

5. **Etiquetado por inferencia de lenguaje natural:** se aplica el módulo NLI para etiquetar los posts según su relación con el bulo original, detectando aquellos que apoyan o contradicen la afirmación falsa.

El resultado de aplicar este proceso es un conjunto de posts semánticamente similares para

cada falsedad etiquetada como *Entailment*, *Contradiction* o *Neutral*. Para la extracción de las afirmaciones falsas ya identificadas por los *fact-checkers* se ha utilizado LatamChequea Coronavirus [100]. De entre todos los indicadores de esta base de datos, la variable utilizada para este propósito será el título de cada post falso registrado. Dado que los módulos de NLI y de similitud semántica exigen que la afirmación falsa se exprese de la forma más clara posible, se descartan palabras redundantes como 'bulo' o 'mensaje' que hagan referencia a la falsedad en sí.

El segundo paso consiste en generar cadenas de búsqueda (*queries*) para cada información falsa. Ya se ha visto en los experimentos anteriores de esta tesis que son construcciones de distintas palabras claves relacionadas con la falsedad y combinadas mediante operadores lógicos 'AND' y 'OR'. Estas cadenas se utilizan a continuación a través de la API de Twitter para encontrar mensajes que compartan ese tipo de desinformación. Cada cadena generada se ha mejorado después manualmente para recuperar el máximo número de posts que difunden esa información falsa.

La consulta resultante se compone de las palabras claves potenciales de esa falsedad enlazadas y del uso de paréntesis para mejorar la recopilación (en el Capítulo 5 se muestra cómo actualizar esta parte). Posteriormente se ha optimizado cada cadena con sinónimos y expresiones similares para captar distintas formas de expresar una misma información falsa, ya que esta tiene por qué propagarse con las mismas palabras en la red social. Esto permite recoger variantes del mismo mensaje en diferentes zonas geográficas hispanas y evita la realización de una búsqueda sesgada de posts de un único país hispano.

El tercer paso define la búsqueda automatizada en la API de Twitter utilizando las consultas de búsqueda generadas. Esta búsqueda se limita al periodo comprendido entre el 1 de enero de 2020 y el 14 de marzo de 2021. Además, no se han excluido los posts de respuestas que coinciden con la consulta, ya que también pueden contener información errónea. El resultado de este proceso comprende 61 consultas seleccionadas para la búsqueda automatizada a partir de *claims* desmentidos y posts recogidos a través de ellos gracias a esta API. En las Tablas 4.2 y 4.3 se muestra el conjunto de enunciados falsos empleado.

En el siguiente paso, el conjunto de datos ha sido refinado con el módulo de similitud semántica para filtrar los posts que realmente presentan cercanía con la falsedad identificada por los *fact-checkers*. Por último, se aplica el componente de NLI para etiquetar cada post como *Entailment*, *Contradiction* o *Neutral* según la relación con tal información falsa. De acuerdo con la normativa de Twitter y para garantizar la privacidad de los individuos en la plataforma, no se han publicado los usuarios ni los textos.

4.1.4. Caso de uso: NLI para monitorizar olas de desinformación

La combinación de un filtro por similitud semántica junto con un proceso de NLI construye una herramienta de gran utilidad con multitud de posibilidades. Así, más allá del propio *fact-checking*, la comprobación automática de *claims* permite hacer una evaluación exhaustiva de cómo se distribuye la desinformación. En este caso de uso se presenta la aplicación de ambas técnicas para explorar cómo se propagó la información falsa en la red social X (Twitter) durante la pandemia de la COVID-19.

Para ello, se utiliza el conjunto de datos NLI19-SP presentado en el apartado anterior. Cada post del conjunto de datos recibe una etiqueta (*Entailment*, *Contradiction* o *Neutral*) según su relación con la falsedad más similar. Además, también se han identificado los posts de las cuentas

Id	Bulo (en español)	Bulo (en inglés)	*Fact-checkers*
1	La PCR no distingue entre coronavirus y gripe	*PCR tests do not distinguish between coronavirus and the flu*	Newtral.es
2	Las vacunas de ARN-m contra el coronavirus nos transforman en seres transgénicos	*mRNA vaccines against coronavirus transform us into transgenic beings*	Animal Político, Maldita.es, Newtral.es
3	La vacuna contra la COVID-19 se crea con células de fetos abortados	*COVID-19 vaccines are made of cells from aborted fetuses*	Agencia Ocote, Agência Lupa, Chequeado, ColombiaCheck, Maldita.es, Newtral.es
4	Merck asocia las vacunas contra la COVID-19 con un genocidio	*Merck associates COVID-19 vaccines with a genocide*	Ecuador Chequea, Newtral.es
5	Una imagen relaciona la prueba PCR con la destrucción de la glándula pineal en el Antiguo Egipto	*An image links PCR tests to the destruction of the pineal gland*	Maldita.es
6	La vacuna contra la COVID-19 produce parálisis facial	*COVID-19 vaccines produce facial paralysis*	Chequeado, Newtral.es
7	La primera ministra de Australia finge ponerse la vacuna contra la COVID-19	*Australia first minister pretends to get the COVID-19 vaccine*	Agência Lupa, La Silla Vacía
8	La vacuna contra la COVID-19 produce convulsiones	*COVID-19 vaccines produce seizures*	Maldita.es, Newtral.es
9	Mueren 53 personas en Gibraltar tras ponerse la vacuna contra la COVID-19	*53 people dead after being vaccinated against COVID-19 in Gibraltar*	Maldita.es, Newtral.es
10	Detienen en un Lidl de Gijón a 11 personas con COVID-19	*11 people with COVID-19 arrested in Lidl supermarket in Gijón*	Maldita.es, Newtral.es
11	Ya no existen las enfermedades respiratorias que no son COVID-19	*Respiratory diseases that are not COVID-19 do not exist anymore*	Newtral.es
12	La PCR da positivo por nuestros genes endógenos, no por coronavirus	*PCR tests positive due to our endogenous genes, not due to coronavirus*	Newtral.es
13	La ciudad de Rosario (Argentina) para la vacunación por los efectos adversos de la vacuna	*The city of Rosario (Argentina) stops vaccination because of the adverse effects of the vaccine*	Chequeado, Maldita.es
14	La OMS dice que llevar a los niños al colegio sirve como consentimiento para su vacunación	*The WHO says that taking our children to school gives consent for their vaccination*	Maldita.es
15	La definición de pandemia cambió en 2009 por la OMS	*The definition of pandemic was changed in 2009 by the WHO*	Newtral.es
16	Muere una enfermera de Tennessee (Estados Unidos) tras vacunarse contra la COVID-19	*A nurse from Tennessee (United States) died after being vaccinated against COVID-19*	La Silla Vacía, Maldita.es, Newtral.es
17	Solo el 6 % de las muertes por coronavirus en Estados Unidos fueron realmente por esta causa	*Only 6 % of coronavirus deaths in United States were actually due to this cause*	AFP, Agência Lupa, Animal Político, Chequeado, La Silla Vacía
18	La PCR da positivo por los exosomas, no por coronavirus	*PCR tests positive due to exosomes, not due to coronavirus*	Newtral.es
19	La mascarilla produce enfermedades neurodegenerativas	*Masks produce neurodegenerative diseases*	Maldita.es, Newtral.es
20	En Países Bajos existe desde 2015 una patente de test de COVID-19	*A patent of COVID-19 test exists in the Netherlands since 2015*	Maldita.es, Newtral.es
21	La vacuna contra la COVID-19 causa esterilidad	*Pfizer vaccines cause sterility*	Animal Político, Chequeado, ColombiaCheck, La Silla Vacía, Maldita.es, Newtral.es
22	Un estudio de 2008 financiado por la Comisión Europea ya incluía la COVID-19	*A study funded by the European Commission in 2008 already included COVID-19*	Newtral.es
23	Varios vacunados con la vacuna UQ-CSL contra la COVID-19 contraen el VIH	*Several COVID-19 vaccinated people with UQ-CSL contracted HIV*	Newtral.es
24	La vacuna contra la COVID-19 es aún experimental porque está en fase 4	*Vaccines against COVID-19 are still experimental because they are in phase 4*	Animal Político, Maldita.es
25	El Banco Mundial tenía planes para la COVID-19 desde 2017	*The World Bank had plans for COVID-19 since 2017*	Animal Político, Aos Fatos, Mala Espina Check
26	La vacuna contra la COVID-19 destruye nuestro sistema inmunológico	*Vaccines against COVID-19 destroy our immune system*	Maldita.es, Newtral.es
27	Pirbright Institute patentó la COVID-19 en 2018	*Pirbright Institute patented COVID-19 in 2018*	Maldita.es
28	Las gárgaras con agua y sal previenen o curan el coronavirus	*Gargling with water and salt prevents or cures coronavirus*	#NoComaCuento (La Nación), AFP, Chequeado, ColombiaCheck, Ecuador Chequea, Efecto Cocuyo, El Surtidor, La Silla Vacía, Maldita.es, Spondeo Media, Verificador (La República)
29	La dieta alcalina previene o cura el coronavirus	*Alcaline diets prevent or cure coronavirus*	Agência Lupa, Animal Político, Bolivia Verifica, Chequeado, ColombiaCheck, Cotejo.info, EFE Verifica, Ecuador Chequea, Efecto Cocuyo, #NoComaCuento (La Nación), La Silla Vacía, Mala Espina Check, Maldita.es, Newtral.es
30	El coronavirus fue fabricado en un laboratorio chino	*Coronavirus was made in a Chinese lab*	Chequeado, Ecuador Chequea, Estadão verifica

Tabla 4.2: Relación de bulos incluidos en el dataset NLI19-SP - Parte 1

Id	Bulo (en español)	Bulo (en inglés)	Fact-checkers
31	La mascarilla causa hipoxia	*Masks cause hypoxia*	Agencia Ocote, Agência Lupa, Animal Político, Aos Fatos, Bolivia Verifica, Chequeado, ColombiaCheck, Cotejo.info, EFE Verifica, Ecuador Chequea, Efecto Cocuyo, La Silla Vacía, Maldita.es, Newtral.es, Verificado, Verificador (La República)
32	El eucalipto previene o cura el coronavirus	*Eucalyptus prevents or cures coronavirus*	AFP
33	El matico cura el coronavirus	*Matico cures coronavirus*	Bolivia Verifica
34	El biomagnetismo mata el coronavirus	*Biomagnetism kills coronavirus*	Bolivia Verifica, Maldita.es
35	La hoja de guayaba previene o cura el coronavirus	*Guava leaf prevents or cures coronavirus*	Animal Político, Bolivia Verifica, Maldita.es, Newtral.es
36	La NASA catalogó el dióxido de cloro como antídoto universal en 1988	*NASA catalogued chlorine dioxide as a universal antidote in 1988*	Animal Político
37	El vino previene o cura el coronavirus	*Wine prevents or cures coronavirus*	Chequeado, EFE Verifica, Maldita.es, Newtral.es
38	La mascarilla causa la muerte por neumonía bacteriana	*Masks cause death due to bacterial pneumonia*	Maldita.es
39	La vitamina C previene o cura el coronavirus	*Vitamin C prevents or cures coronavirus*	AFP, Chequeado, EFE Verifica, Agência Lupa, Maldita.es, Verificado
40	La prueba de antígenos no sirve para la COVID-19 porque da positivo con Coca-Cola	*Antigen tests are useless for COVID-19 because they test positive with Coca-cola*	Maldita.es, Newtral.es
41	La homeopatía previene o cura el coronavirus	*Homeopathy prevents or cures coronavirus*	Chequeado, Mala Espina Check, Maldita.es, Periodismo de barrio / El Toque
42	La COVID-19, el MERS y el H1N1 coinciden con la instalación del 5G, 4G y 3G, respectivamente	*COVID-19, MERS and H1N1 coincide with the installation of 3G, 4G and 5G, respectively*	Poligrafo
43	Los indígenas protegen a los niños con hierbas y árboles frente a la COVID-19	*Indigenous groups protect their children from COVID-19 with herbs and trees*	Ecuador Chequea
44	Los mosquitos transmiten el coronavirus de contagiados	*Mosquitoes transfer coronavirus from infected people*	Maldita.es
45	Beber agua o sorbos previene o cura el coronavirus	*Drinking or sipping water prevents or cures coronavirus*	#NoComaCuento (La Nación), AFP, Bolivia Verifica, ColombiaCheck, La Silla Vacía, Maldita.es, OjoPúblico
46	Mueren 55 personas en Estados Unidos tras vacunarse contra la COVID-19	*55 people dead after being vaccinated against COVID-19 in the United States*	EFE Verifica
47	Las mascarillas producen pleuresía y neumonía	*Masks produce pneumonia and pleurisy*	AFP
48	Las personas sanas llevan la mascarilla con la parte blanca hacia fuera para no contagiarse de COVID-19	*Healthy people wear their masks with the white part on the outside not to get COVID-19*	Newtral.es
49	El SARS-COV-2 no cumple los postulados de Koch, Rivers e Inglis para considerarlo enfermedad y coronavirus	*SARS-COV-2 does not fulfill Koch, Rivers and Inglis' postulates in order to be considered as coronavirus and as a disease*	EFE Verifica
50	Christine Lagarde dijo que los ancianos viven demasiado	*Christine Lagarde said that the elderly live too long*	Chequeado, ColombiaCheck, Ecuador Chequea, Maldita.es
51	La COVID-19 es una bacteria	*COVID-19 is a bacteria*	Animal Político, Chequeado, ColombiaCheck, La Silla Vacía, Maldita.es, Verificador (La República)
52	Galicia aprueba una ley para aislar a los positivos COVID-19 en campos de concentración	*Galicia approves a law to aisle COVID-19 positives in concentration camps*	Maldita.es
53	Las ondas electromagnéticas del 5G propagan el coronavirus	*5G electromagnetic waves spread coronavirus*	Chequeado, Ecuador Chequea
54	La OMS rcomienda un test pulmonar para identificar el coronavirus	*The WHO recommends a pulmonary test to detect coronavirus*	EFE Verifica
55	Las pandemias tienen lugar cada 100 años	*Pandemics take place every 100 years*	AFP, Animal Político, ColombiaCheck, Verificador (La República)
56	El laboratorio de Wuhan tiene relación con Glaxo y Pfizer	*Wuhan lab is related to Glaxo and Pfizer*	Animal Político, Chequeado, La Silla Vacía, Maldita.es, Newtral.es
57	El coronavirus desaparece a los 27 grados	*Coronavirus disappears at 27 degrees*	Bolivia Verifica, Convoca, Agência Lupa
58	Hubo 17000 y 26000 muertes más en 2019 y 2018 respectivamente que en 2020	*There were 17000 and 26000 more deaths in 2019 and 2018 respectively than in 2020*	Maldita.es, Newtral.es
59	El polisorbato 80 de la vacuna contra la gripe causa coronavirus	*Polysorbate 80 in the flu vaccines cause coronavirus*	EFE Verifica, Maldita.es
60	Detienen a Charles Lieber por crear y vender el coronavirus	*Charles Lieber arrested for creating and selling coronavirus*	#NoComaCuento (La Nación), AFP, Animal Político, Efecto Cocuyo, Agência Lupa, Mala Espina Check, Maldita.es, Newtral.es
61	En Israel no mueren por coronavirus gracias a una receta de limón y bicarbonato	*No deaths in Israel due to coronavirus thanks to a recipe with lemon and bicarbonate*	Newtral.es, Verificado

Tabla 4.3: Relación de bulos incluidos en el dataset NLI19-SP - Parte 2

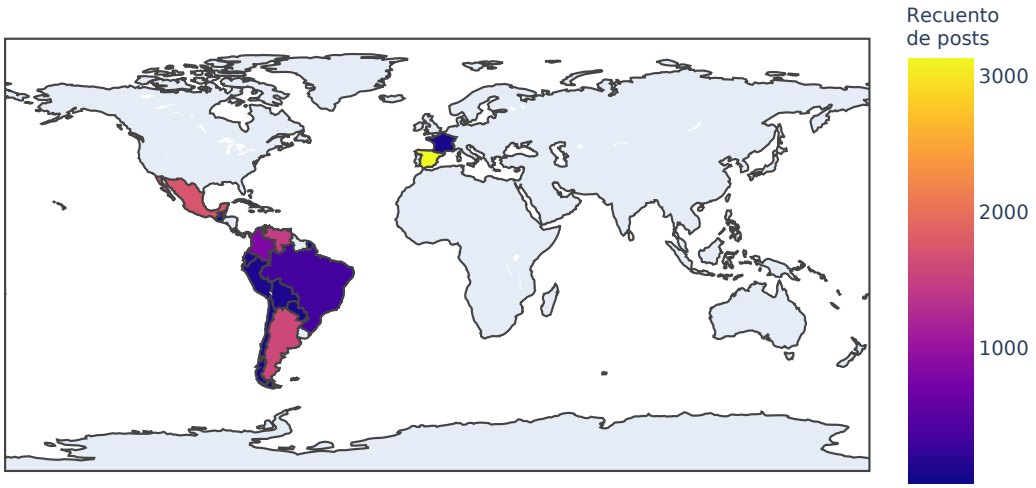


Figura 4.1: Mapa que muestra el número de posts que apoyan una falsedad según la nacionalidad del *fact-checker* que lo ha identificado. Aparece también Francia, aunque no es un país hispanohablante, por las informaciones falsas en español recogidas por la organización de *fact-checking* Factual AFP, de origen francés.

de Twitter de los *fact-checkers*. Toda esta información permite inferir patrones y características relevantes de las afirmaciones sobre desinformación difundidas durante la pandemia del coronavirus. Para acotar el análisis, el foco son los mensajes escritos en español. Fig. 4.1 muestra la distribución de los posts encontrados según la nacionalidad de la organización de *fact-checking* utilizada para identificar el bulo. Aunque hay un número importante de posts recogidos de enunciados falsos señalados por las organizaciones españolas, no se encontraron grandes diferencias entre los países de habla hispana.

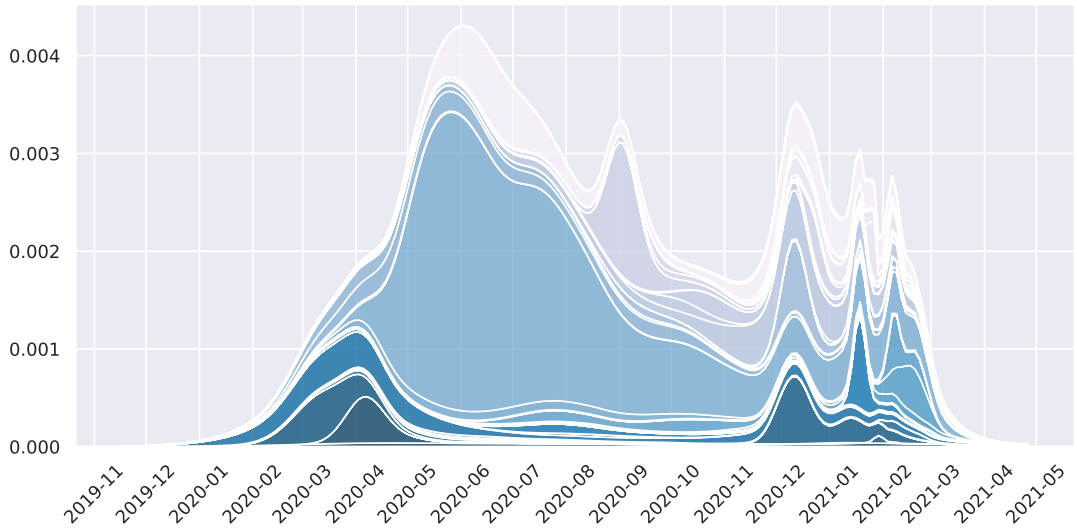


Figura 4.2: Distribución temporal de los posts que apoyan los 61 bulos identificados, lo que evidencia patrones comunes con múltiples picos de actividad compartidos.

Fig. 4.2 muestra el gráfico de distribución acumulativa para una visión general de los posts

que apoyan las diversas falsedades. Se aprecia como conclusión la tendencia común hacia las oleadas de desinformación. Este comportamiento refleja que la información falsa se alimenta de sí misma y que los difusores pueden operar de forma coordinada para la evolución de estas olas. De acuerdo a las definiciones vistas al inicio, en estas oleadas puede haber tanto *disinformation*, es decir, la desinformación deliberada para desinformar, como *misinformation*, aquella sin una intencionalidad concreta.

También merece la pena tener en cuenta este fenómeno a la hora de tomar medidas para contrarrestar la propagación de la desinformación. Además, la gran representación de falsedades específicas es también un elemento importante a estudiar. Así, uno de los bulos más difundidos (número 31 en la Tabla 4.3) es el de que "las mascarillas causan hipoxia". El gran número de posts apoyando esta falsa afirmación es la razón de la gran oleada centrada en junio de 2020. Del mismo modo, el pico situado en abril de 2020 se debe principalmente al bulo 50 ("Christine Lagarde dijo que los ancianos viven demasiado").

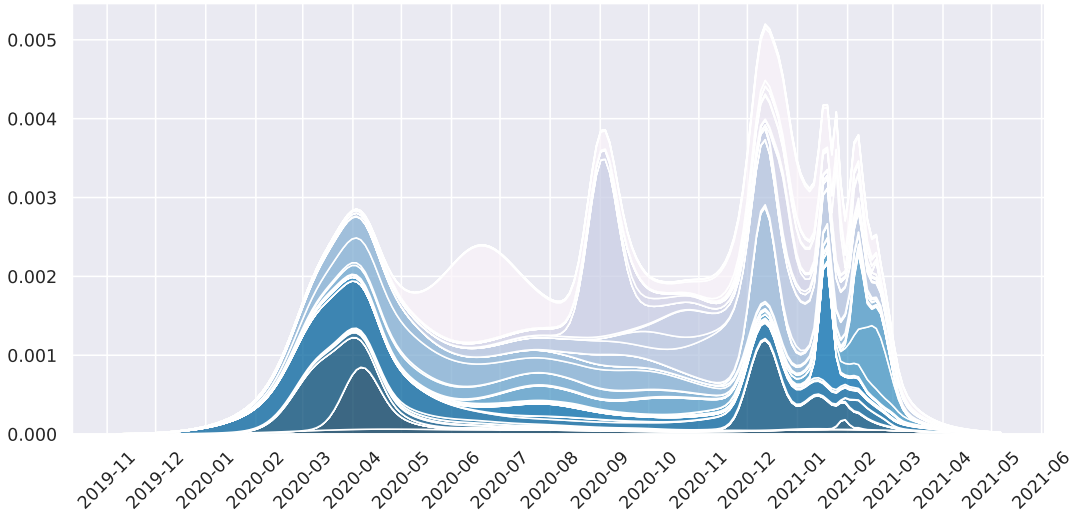


Figura 4.3: Distribución temporal de los posts que apoyan los bulos identificados sin representar el bulo con el ID 31, relacionado con la falsa afirmación "las mascarillas causan hipoxia".

Para visualizar mejor la distribución de posts que apoyan la información falsa, en Fig. 4.3 se muestra el mismo gráfico sin incluir el bulo 31, que concentra gran parte de las publicaciones. Aunque la gran onda desaparece en este gráfico, reflejando que fue causada por el bulo eliminado, se puede ver cómo las curvas siguen siendo visibles, evidenciando los patrones de comportamiento comunes de cómo circula la desinformación.

Para un análisis más profundo de la desinformación que circuló durante la pandemia de la COVID-19, Fig. 4.4 muestra la distribución temporal de los posts que apoyan una selección de bulos y los posts opuestos a ellos como *fact-checking*. En los casos de los bulos 28, 37, 50 y 60, la campaña lanzada a través de los desmentidos causó más posts contrarrestando tales falsedades que posts apoyándolas. Para el resto de bulos analizados, la respuesta contraria a la falsedad fue más leve. En el caso del bulo 15 ("La definición de pandemia fue cambiada en 2009 por la OMS") no se ha detectado una reacción a través de desmentidos. Estas visualizaciones revelan lo complejos que son estos escenarios.

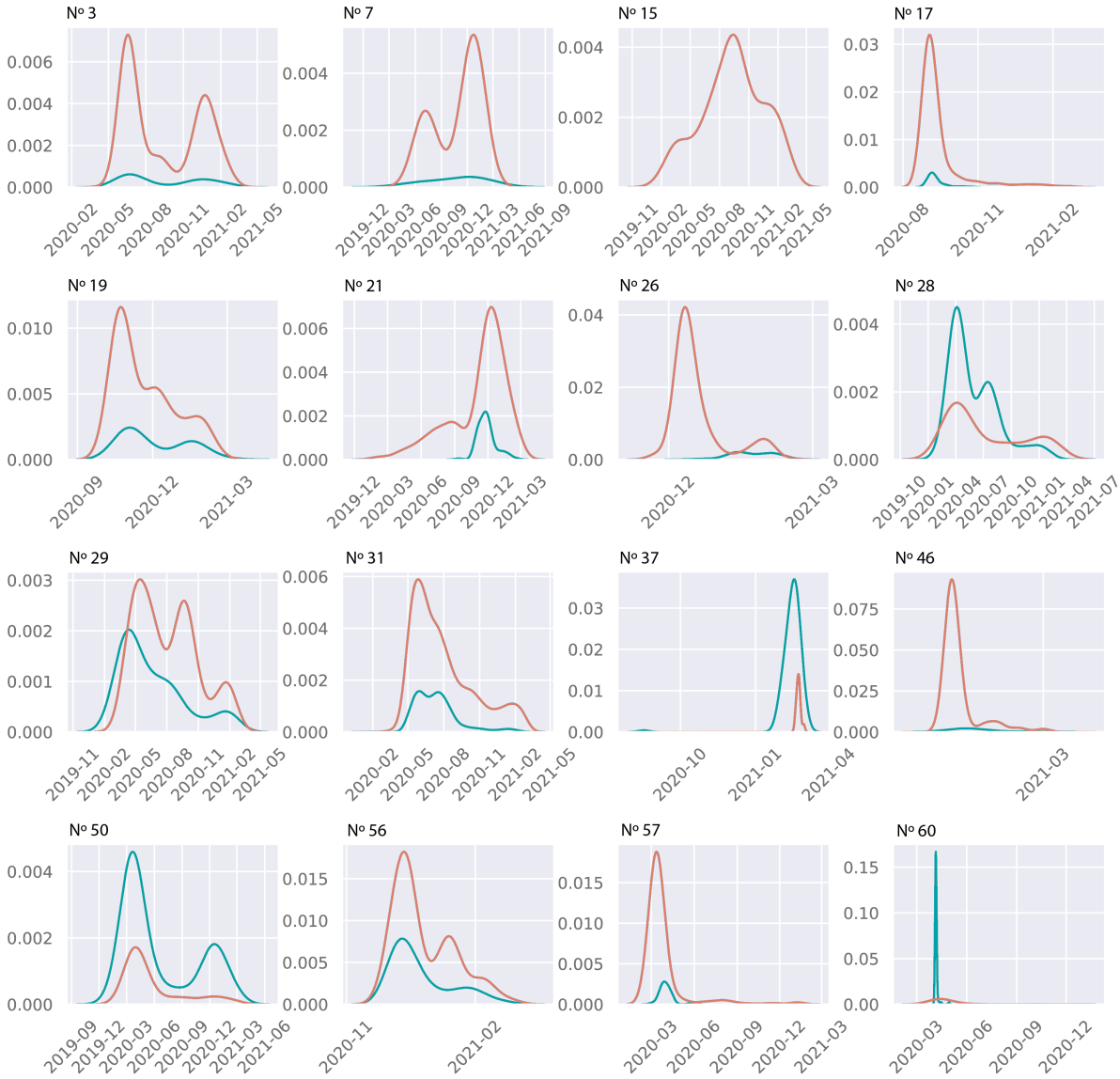


Figura 4.4: Comparativa entre la distribución de posts que apoyan cada falsedad enumerada (en naranja) y la de aquellos que la desmienten (en azul).

4.2. Analizando los mecanismos de difusión de desinformación en X

Para desarrollar el módulo de NLI en esta investigación, se ha ajustado un modelo siguiendo el enfoque presentado en la sección anterior. Este ajuste es de `XLM-RoBERTa-large` [146] con dos *datasets* de entrenamiento [9]: uno a partir de los datos del modelo *Machine Translated MultiNLI* (MNLI-MT) [242] para el procesamiento interlingual, y otro también con XNLI [161], además de *Adversarial Natural Language Inference* (ANLI) [248], SNLI [159] y *Fact Extraction and VERification dataset* (FEVER) [249] para textos en inglés. Todo ello con un tamaño del *batch* de 1.024, una tasa de aprendizaje de $2e$-5 y el optimizador *Adam* [226] como hiperparámetros[2].

No obstante, aunque el proceso descrito es un mecanismo de *fact-checking* automático, se trata en verdad de *fact-checking semiautomático*, porque a partir de esta tarea con el *claim*, la correspondiente a su detección ya queda cubierta por los desmentidos realizados por los *fact-checkers* de manera manual, y de este proceso humano dependen estas afirmaciones para alinear después con los posts [42]. De acuerdo al estado de la cuestión antes detallado, mientras que el *claim detection* automático tiene su ventaja en encontrar enunciados susceptibles de verificar, el *claim matching* encuentra su fuerte *a posteriori* con los *fact-checks* ya realizados para no replicar esfuerzos [9], en vez de escoger otras bases de conocimiento. Esta tesis se centra en esta fase posterior, no como un proceso aparte del manual sino potenciador de este para aprovechar el ejercicio profesional ya realizado por las organizaciones de *fact-checking*.

4.2.1. Análisis de posts con NLI

Para este otro experimento, también se han tenido en cuenta los posts en base a su alineación falsa, tengan éxito o no dentro de Twitter. De nuevo, el proceso pasa por obtener *claims* de varias desinformaciones ya desmentidas por las organizaciones de *fact-checking*, también sobre el coronavirus; por generar *queries* para la búsqueda automática mediante la API de la plataforma, y por etiquetar los posts mediante su NLI para indicar el sentido del *claim* respecto al enunciado dado.

También se detalla una serie de condiciones para que estos *claims* funcionen en este análisis: 1) como ya se ha avanzado, estar ligados a la COVID-19, como afirmaciones sobre las medidas para afrontar la pandemia en su momento o que, directamente, cuestionan la existencia del virus; 2) haber sido desmentidos por las entidades reconocidas por el sello de la IFCN [44]; 3) no limitarse a un idioma pese a que las *queries* sí contengan en este caso los términos en inglés para obtener los posts en esta lengua; 4) no referirse a falsedades dadas por la falta de contexto, más complejas de encontrar mediante *queries* y que pueden sufrir ambigüedades; 5) no repetirse, ya que cada *claim* debe aludir a una falsedad diferente dentro de la base del conocimiento.

El *dataset* de partida para este experimento lo componen 26 falsedades sobre la COVID-19 en pandemia tratadas por 13 organizaciones de *fact-checking*. Se dividen en 11 mentiras antivacunas, cinco frente a la existencia del virus (más una de ambos temas), seis sobre las máscaras, dos sobre los remedios alternativos y una sobre la gestión de la pandemia. Seis de ellas apuntan contra instituciones o personalidades públicas. Con estos *claims* como unidades de las distintas informaciones falsas, se pueden formar las *queries* con sus *keywords*, variaciones, sinónimos y/o jerga específica para descargar todos los contenidos posibles.

[2]Este modelo está publicado en `https://huggingface.co/AIDA-UPM/xlm-roberta-large-snli_mnli_xnli_fever_r1_r2_r3`

La muestra final a partir de la API la componen 17.570 posts, de los que 2.837 entre enero de 2020 y diciembre de 2021 para este experimento aluden directamente a la información falsa para difundirla (1.737 posts con el 99 % de *Entailment*) o para rebatirla (1.102 con *Contradiction*). Estas publicaciones se han guardado asociadas a las *queries* y los enunciados que han posibilitado su descarga. Al centrarse el análisis en los indicadores de los posts originales, no se han tenido en cuenta aquellos reposteados por otras cuentas, para los que solo se trata el número de *reposts*.

De nuevo, cada post tendrá consigo una etiqueta con su tipo de alineación con la desinformación (*Entailment*, *Contradiction* o *Neutral*). De estos, se usarán los considerados como *Entailment* y *Contradiction* para analizar las publicaciones en función de las métricas de la red social. Los 'me gusta', *reposts*, respuestas y citados se obtuvieron de la API de Twitter, antes de su cambio de dueño y de nombre. Con todo ello, el análisis permite saber para cada variable la proporción de posts con *Entailment* y *Contradiction* desde dos perspectivas: teniendo en cuenta solo el peso de las creaciones originales y considerando también todas las interacciones.

A estos indicadores se suman las repeticiones (contenidos exactamente iguales publicados desde cuentas distintas). La repetición es un recurso usado dentro de la propagación de la información falsa. Por ejemplo, se ha demostrado que una cuarta parte de las falsedades en el discurso político se repite [250]. En este contexto entran en juego los posts de desinformaciones ya expuestas que siguen circulando de forma manual o automática mediante *bots* [251, 36].

4.2.2. *Reposts*

En cuanto a los posts originales, de todos aquellos en relación con la desinformación (*Entailment* más *Contradiction*), más del 75 % no tiene *reposts*. Menos del 20 % restante tiene de 1 a 10 *reposts*, alrededor de un 3 % los sobrepasa en un rango de hasta 100 *reposts* y solo 1 de cada 200 los supera (ver Fig. 4.5).

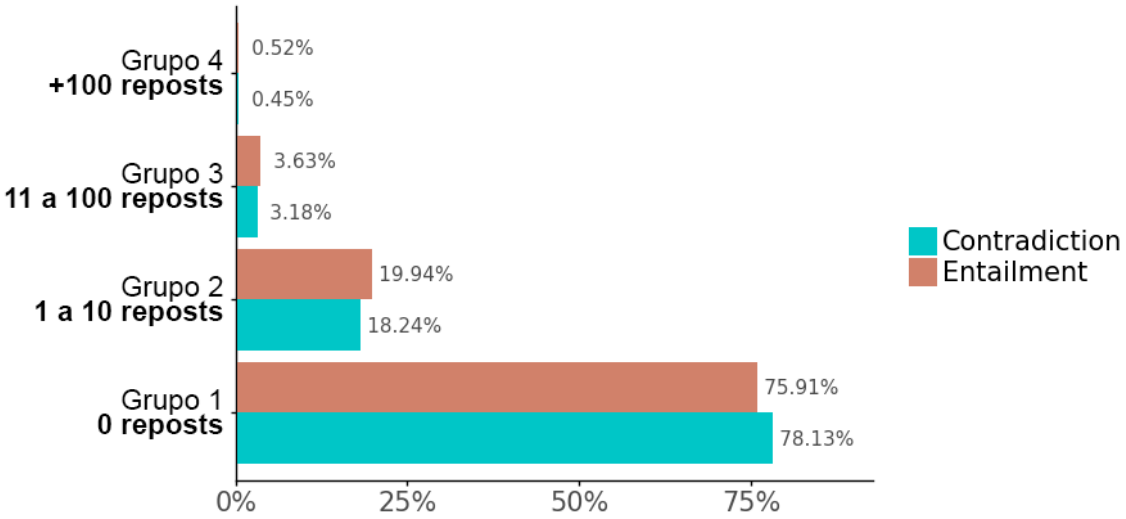


Figura 4.5: Proporción de posts con *Entailment* y *Contradiction*, por número de *reposts*.

Las publicaciones con *Contradiction* sin *reposts* (78.13 %) superan ligeramente a aquellas con *Entailment* (75.91 %), dando a entender que los posts cuyo enunciado es opuesto a la información reciben algo menos de interés, pero no mucho. Esto hace que el porcentaje de contenidos con uno o más *reposts* sea mayor para la etiqueta de *Entailment*, pero en términos de máxima viralidad,

ambas clases están prácticamente a la par (0.45 % para *Contradiction* y 0.52 % para *Entailment*).

Respecto al peso real de los posts en la red social, contando todas las veces que un contenido se ha distribuido tanto de forma original como con *reposts*, la situación es diversa (ver Fig. 4.6). Bajo esta perspectiva, el peso de las publicaciones sin *reposts* constituye menos del 20 %, así como el de aquellas reposteadas 10 veces o menos. Bajo este análisis, los posts con más 10 *reposts* y sobre todo los de más de 100 adquieren más peso.

Teniendo en cuenta el peso de todos los *reposts*, también cambian las estadísticas de los porcentajes de *Contradiction* y *Entailment*. Hay menos peso de los contenidos sin repostear opuestos al enunciado de la desinformación (17.61 %) frente a los que sí la exponen como tal (19.51 %), debido a la cantidad más alta de publicaciones con más de 100 *reposts* con *Contradiction* (45.41 %) frente a aquella con *Entailment* (32.06 %). Por tanto, la suma del resto de grupos con menos *reposts* dará una mayor proporción para los posts que exponen la información falsa (67.94 %) que para los que la contradicen (54.59 %).

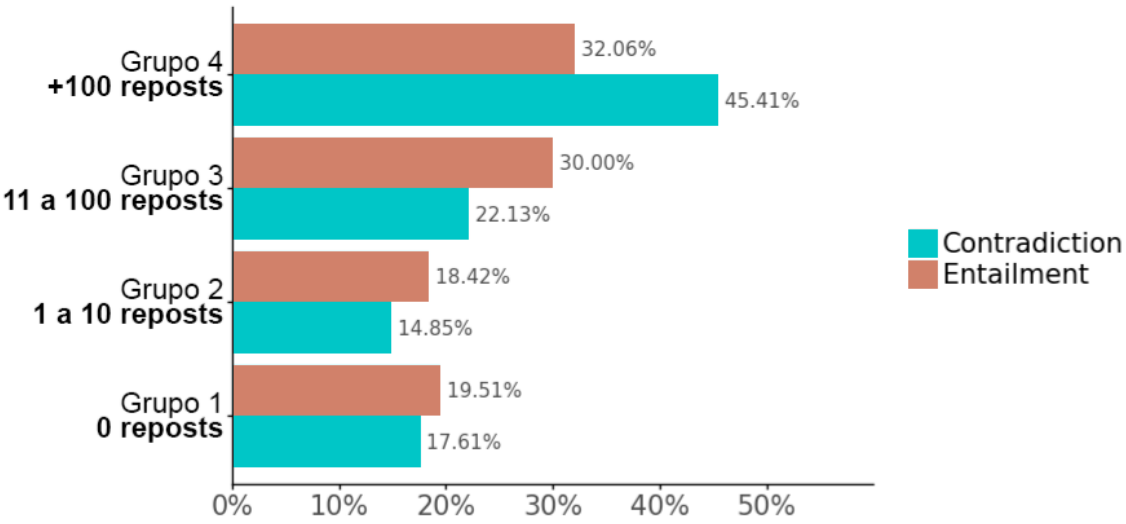


Figura 4.6: Proporción de posts con *Entailment* y *Contradiction* más la suma de sus *reposts*, por número de *reposts*.

4.2.3. *Likes*

Hay más implicación por parte de los usuarios hacia los contenidos sobre información falsa con los 'me gusta' que con los *reposts* (ver Fig. 4.7): mientras que los contenidos sin *reposts* eran más del 75 %, en el caso de aquellos sin *likes* apenas sobrepasan el 50 %. Las publicaciones con más de 100 *likes* constituyen menos del 2 % de todos los posts, en su mayoría con 0 a 10 'me gusta'.

Hay más posts sin *likes* que replican la información falsa (57.06 %) que los que la contradicen (53.90 %), debido sobre todo a que las publicaciones de 1 a 10 'me gusta' comprenden más porcentaje en el segmento de *Contradiction* (37.30 %) que en el de *Entailment* (34.81 %). En los grupos restantes, no destacan las diferencias de estas dos categorías.

Cuando el recuento es según el número de usuarios que dan *likes* y no solo del número de los posts que los reciben, se aprecia aún más la implicación en la red social en forma de 'me gusta' frente al *repost* (ver Fig. 4.8). Esto hace que la pequeña proporción de publicaciones con más de 100

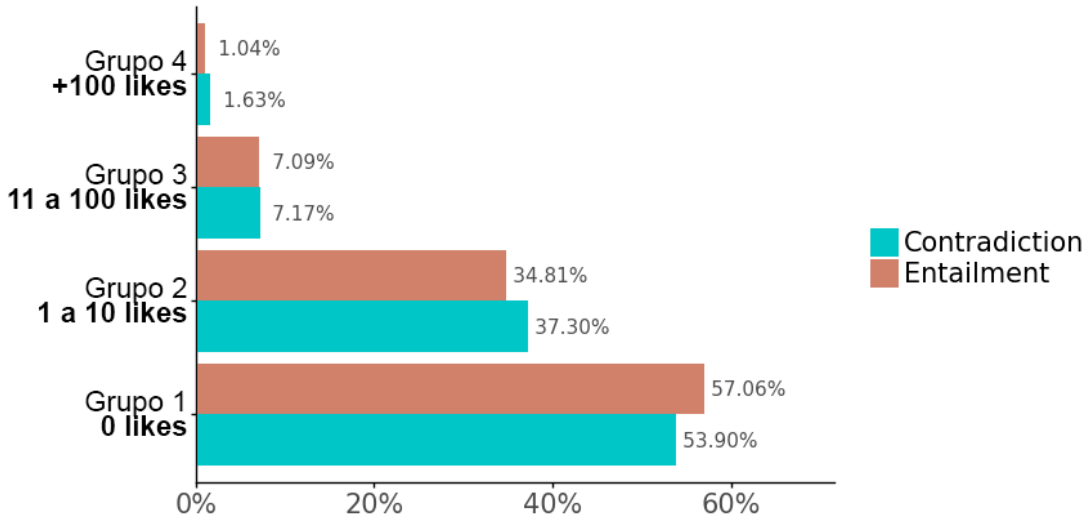


Figura 4.7: Proporción de posts con *Entailment* y *Contradiction*, por número de *likes*.

likes sea, de lo contrario, la predominante respecto a los posts sin *likes* cuando las proporciones se toman en función al peso real de cada una de las interacciones de este tipo.

Este contraste respecto al peso de los posts por sí solos se aprecia más en el desglose por categorías. Mientras que las publicaciones etiquetadas como *Contradiction* con más de 100 'me gusta' son el 1.63 % del total, el peso ponderado a partir de estos *likes* corresponde al 73.05 %. Este cambio de tornas minimiza las proporciones del resto de grupos, algo que sucede en menor medida en el segmento de *Entailment*.

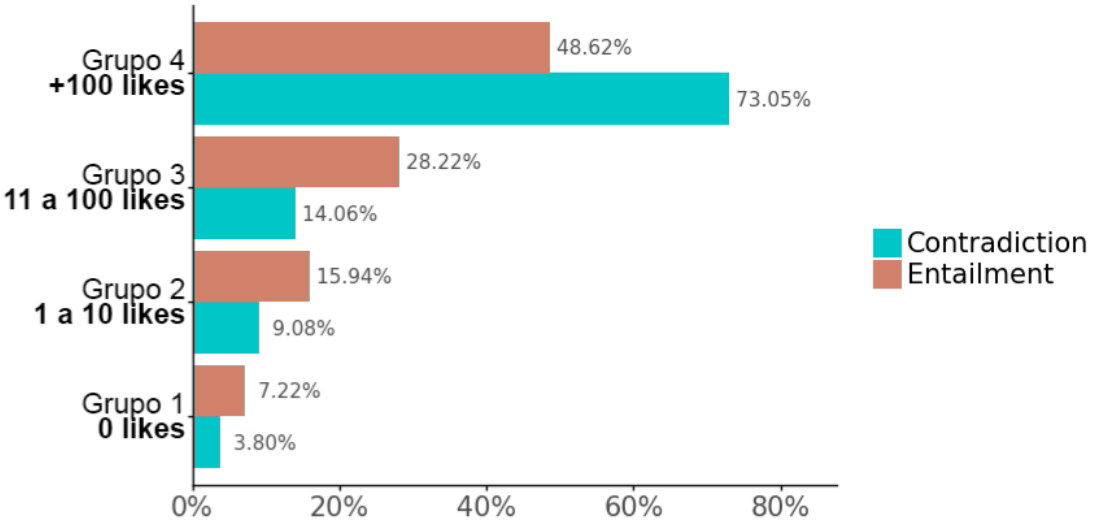


Figura 4.8: Proporción de posts con *Entailment* y *Contradiction* más la suma de sus *likes*, por número de *likes*.

4.2.4. Respuestas

En el análisis de las respuestas, los posts con ausencia de ellas destacan sobre el resto (proporción superior al 60 %), en la misma tendencia que los indicadores anteriores (ver Fig. 4.9). Más del

30 % tienen de 1 a 10 respuestas, por lo que, de igual modo que los 'me gusta', se marca la diferencia respecto a las proporciones más pequeñas de los *reposts*. No obstante, a diferencia de los *likes* y los *reposts*, es más difícil encontrar posts superando las 10 respuestas.

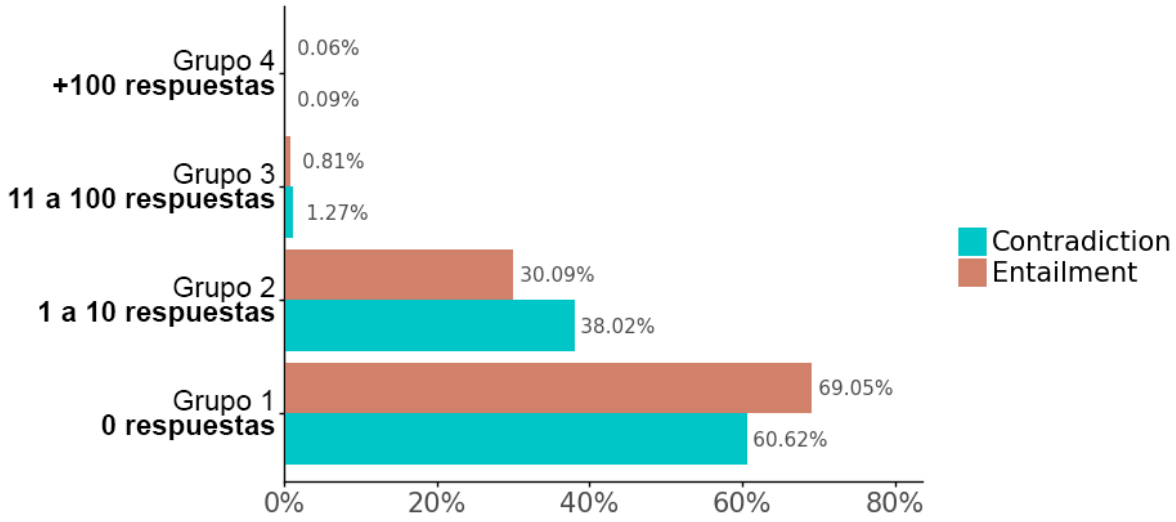


Figura 4.9: Proporción de posts con *Entailment* y *Contradiction*, por número de respuestas.

En esta métrica, son más evidentes las diferencias entre los posts que enuncian la información falsa y los que apuntan lo contrario a ella. Esto se ve con el 69.05 % de aquellos sin respuesta con *Entailment* frente al 60.02 % con *Contradiction*. En consecuencia, cuando las publicaciones contienen de 1 a 10 respuestas, el 30.09 % de *Entailment* es inferior al 38.02 % de *Contradiction*. En función a estas cifras, los usuarios se inclinarían por responder más contenidos que expresan una información contraria a la falsa.

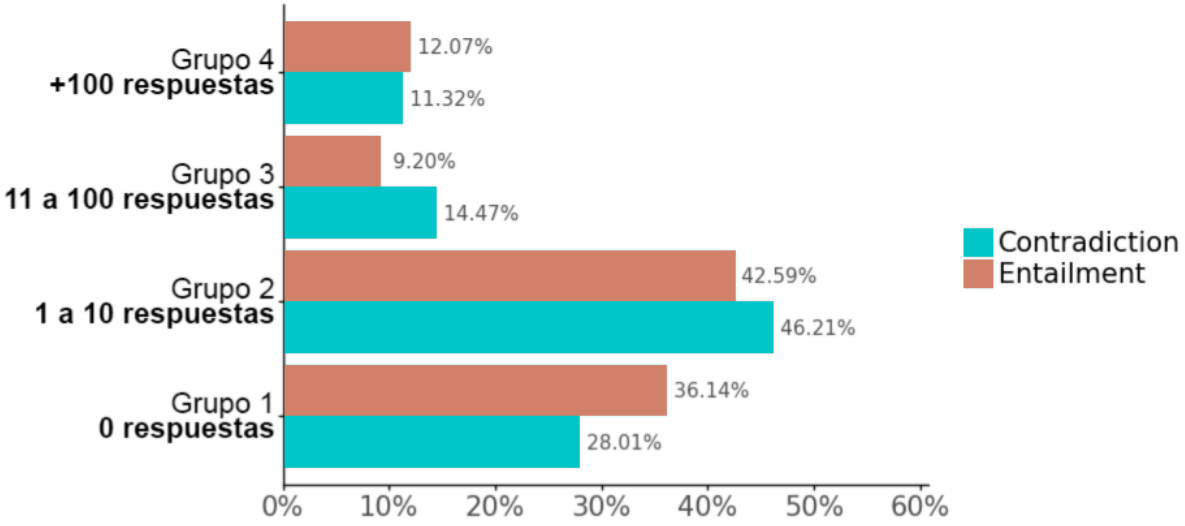


Figura 4.10: Proporción de posts con *Entailment* y *Contradiction* más la suma de sus respuestas, por número de respuestas.

Cuando se suma el número de respuestas al de los posts en sí para determinar el peso ponderado de las publicaciones con estas participaciones de los usuarios, de nuevo el grupo de más de 100

interacciones destaca más que antes (ver Fig. 4.10), en concordancia con las métricas anteriores, pero con un matiz: en este caso los grupos de posts sin respuestas y de 1 a 10 siguen superando el porcentaje del grupo con más de 100.

Estos pesos sugieren de nuevo la tendencia de contestar más a los posts con el sentido opuesto a la información falsa (46.21 % y 14.47 % en los grupos 2 y 3, respectivamente) que con tal desinformación en sí (42.59 % y 9.20 %), menos el grupo 4, de más de 100 respuestas. En este caso, el 12.07 % de *Entailment* apenas supera el 11.32 % de *Contradiction*.

4.2.5. Citados

De acuerdo a este análisis, los usuarios no citan la desinformación ni los contenidos con un significado contrario porque más del 90 % de los posts carecen de esta interacción (ver Fig. 4.11).

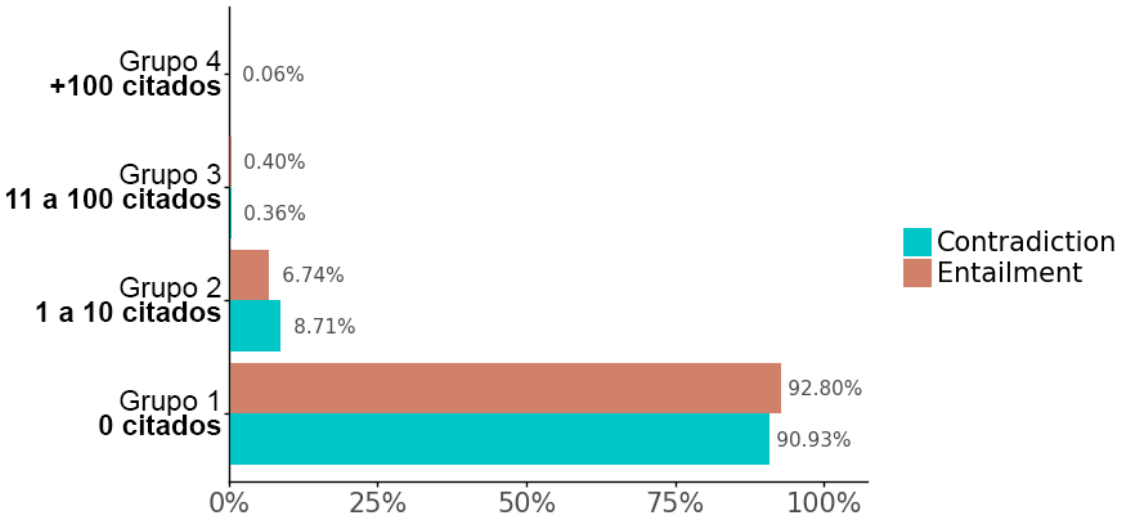


Figura 4.11: Proporción de posts con *Entailment* y *Contradiction*, por número de citados.

Esta falta de participación más activa concuerda con las cifras de *reposts* (más del 75 % de las publicaciones no los contenía), a excepción de pocos contenidos que sí se llevaban todo este tipo de reacciones de los usuarios. Al grupo sin citar le sigue en proporción los posts de 1 a 10 citados y son mínimos los casos con más de este número de interacciones.

En esta variable no se aprecia mucha diferencia entre las cifras de los contenidos con el *claim* en sí o que dicen lo contrario a él, aunque, como en las respuestas, se reacciona a más posts con *Contradiction* (8.71 % del grupo de 1 a 10 citas) que a aquellos con *Entailment* (6.74 %), ambas pequeñas proporciones en comparación a las publicaciones sin citados (92.80 % y 90.93 % con *Entailment* y *Contradiction*, respectivamente).

En este caso, los pesos ponderados con todas las cuentas que citan incluidas en la suma no cambian mucho respecto a los iniciales, dado que siguen destacando en proporción los posts sin citados (ver Fig. 4.12). Los grupos 3 y 4 aumentan su porcentaje debido a su número de reacciones de este tipo. La distinción entre posts con información falsa y con el significado contrario tampoco rompe respecto a lo anterior. Mientras que el 19.49 % de peso ponderado de los contenidos con *Contradiction* supera al 14.78 % de *Entailment* en el grupo 2, el 4.66 % de *Entailment* del grupo 4 equilibra la balanza entre ambas categorías.

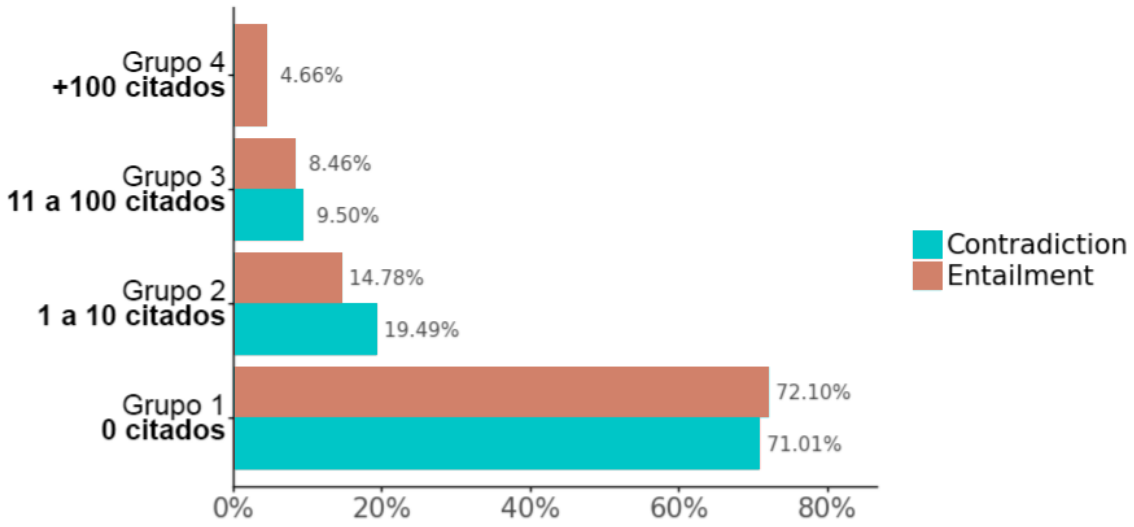


Figura 4.12: Proporción de posts con *Entailment* y *Contradiction* más la suma de sus citados, por número de citados.

4.2.6. Repeticiones

Si bien no es una métrica propia de X ni del anterior Twitter, el contenido permite también ver qué posts son repeticiones de otros (ver Fig. 4.13). Es decir, qué publicaciones, *a priori*, originales, son calcadas a otra que ya existe. Según esta muestra analizada, los posts usados para repeticiones no predominan (menos del 5 %) en el campo de la desinformación. Pero se ha observado también cómo los repetidos, cuando ocurren, no solo se dan una vez, sino hasta más de 10 veces, de acuerdo al desglose por grupos que también se ha hecho para este indicador.

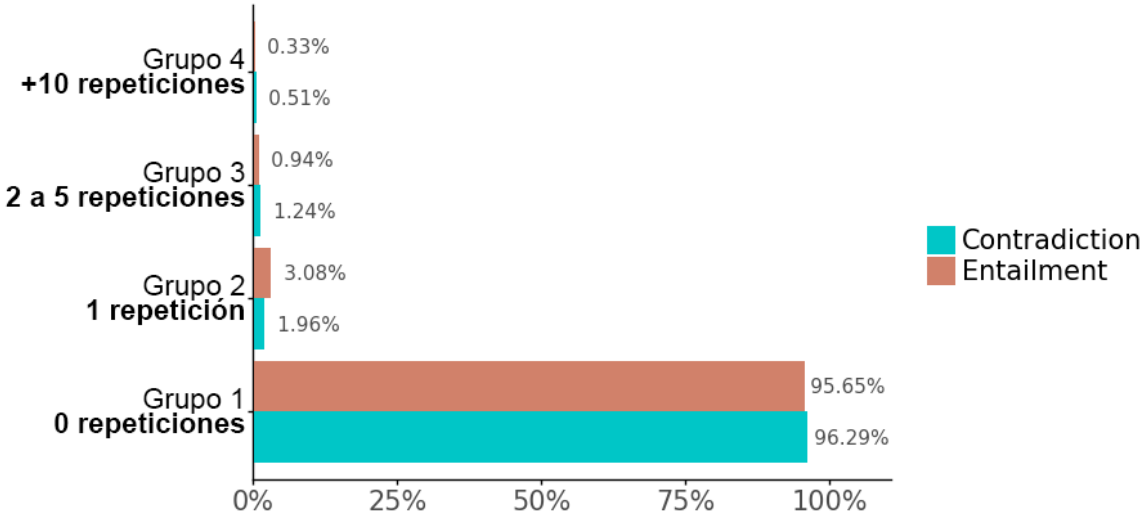


Figura 4.13: Proporción de posts con *Entailment* y *Contradiction*, por número de repeticiones.

En general, las cifras de *Entailment* no difieren de forma significativa respecto a las de *Contradiction* en el análisis de esta métrica. Dentro de la mayoría de posts sin repetir, la desinformación en sí y las publicaciones que apuntan lo opuesto se sitúan casi a la par (95.65 % y 96.29 %, respectivamente). Hay algo más de *Entailment* que de *Contradiction* entre los posts clonados solo

una vez (3.08 % y 1.96 %, respectivamente).

Sin embargo, más del 15 % de los contenidos alrededor de la información falsa son repeticiones que amplifican un mensaje con exactamente el mismo contenido (ver Fig. 4.14), de acuerdo al análisis de los pesos ponderados (no solo el recuento de los posts a repetir sino también de todas las repeticiones ya hechas). Además, la mitad de ellos son posts de 10 o más calcos, y esto permite apreciar más distinción entre *Entailment* y *Contradiction*.

En porcentaje, hay más contenidos sin clones en la categoría de *Contradiction* (84.85 %) que en *Entailment* (82.42 %). Cuando se trata de expandir la información falsa, más del 9 % pertenece al grupo de más de 10 repeticiones (9.11 %). Pero esta particularidad también ocurre en el caso de *Contradiction* (7.35 %) y muestra que el fenómeno es común también para los textos que vierten lo contrario a una desinformación.

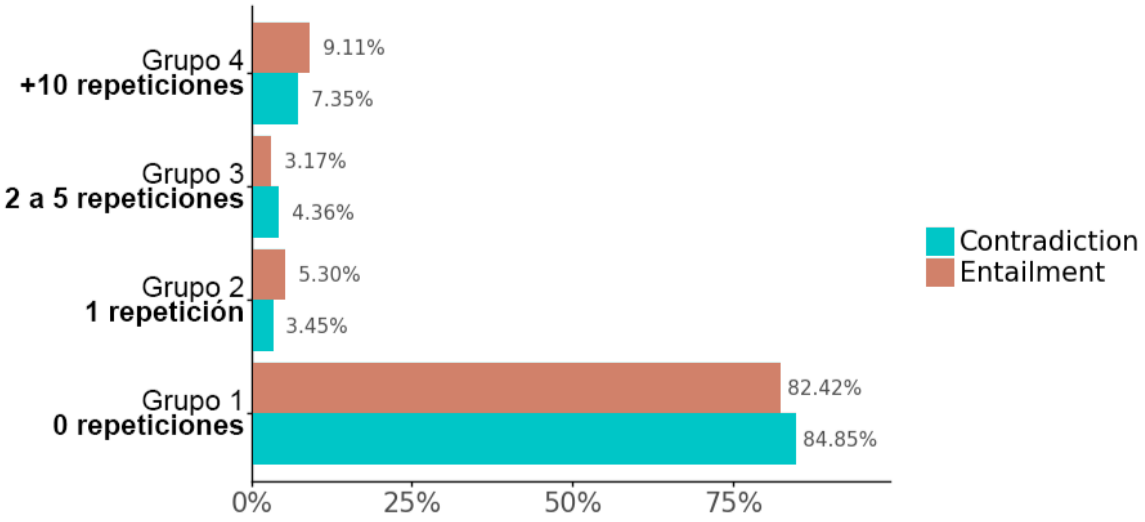


Figura 4.14: Proporción de posts con *Entailment* y *Contradiction* más la suma de los repetidos, por número de repeticiones.

4.2.7. Conclusiones

La cantidad de posts sin interacción, como *reposts* y/o 'me gusta', en comparación al resto muestra que la conversación sobre la desinformación no solo parte de publicaciones virales, sino lo contrario, ya que la mayoría de información que se mueve procede de contenidos sin tal repercusión. Pero es el elevado número de usuarios que reacciona a una mínima parte de mensajes virales el que hace que estos destaquen más, si se mira el peso ponderado de estos posts junto a todas las interacciones. Esta dualidad entre publicaciones sin reacciones y otras que copan todas ellas son las dos caras de una misma moneda en el ecosistema de la desinformación en redes.

No obstante, cuando se trata de un movimiento más activo por parte del usuario en la red social, es decir, contestar, citar o repetir con las mismas palabras un post, el impacto no es tan grande como para que el peso ponderado con este tipo de interacciones cambie mucho respecto al recuento de las publicaciones sin más. De esto se extrae que el individuo no se involucra tanto de forma proactiva mediante las acciones de la plataforma que implican un esfuerzo extra, sino de forma más reactiva en base a *reposts* y 'me gusta'.

La reescritura de posts calcando el contenido como una de las acciones con mayor esfuerzo

suma con todas las repeticiones más del 15 % de las publicaciones sobre información falsa, pero el número de contenidos originales en los que se fijan estos clones es mínimo. En el caso de que estas réplicas correspondan a *bots*, estas cifras muestran que no suponen una práctica tan abundante en contraposición a estudios anteriores sobre el coronavirus [252], el tema de este experimento. Respaldan así que la mayor parte del ecosistema de la desinformación prolifera por obra humana [36]. Más allá de la naturaleza de los posts clonados, no solo aparecen estos cuando se desinforma, sino también cuando se contradice esta desinformación.

La métrica de las respuestas muestra la distinción más relevante entre las categorías de *Contradiction* y *Entailment*, en la medida que se responde más a los mensajes opuestos a la desinformación y no a la desinformación en sí, que en teoría podría haber sido más susceptible de recibir contestaciones por la falsedad que expone. No son tan importantes el resto de diferencias, aunque llama la atención cómo los porcentajes de *Entailment* superan a los de *Contradiction* en los *reposts*, pero ocurre lo contrario con los *likes*. Dicho de otro modo, de este análisis se deduce que las desinformaciones son más propensas a recibir *reposts* que aquello que diga lo contrario a ellas pero, en porcentaje, menos de ellas en comparación encuentran el apoyo más silencioso en forma de 'me gusta'.

Las proporciones ponderadas de los posts, a partir del recuento de cada uno de los tipos de interacciones, amplían la perspectiva de este ecosistema en torno a la desinformación. Entre otras características, se aprecia que tanto *reposts* como 'me gusta' ponen por encima en porcentaje a los contenidos con *Contradiction* frente a aquellos con *Entailment*. Por tanto, las cuentas enfrentadas a los *claims* desinformativos también juegan un papel en la propagación. Sin embargo, toda diferencia en este aspecto es sensible porque muchas de las reacciones en las proporciones finales vienen de los pequeños porcentajes de los posts con más impacto observados en los análisis sin ponderar. Sí existen paralelismos con los pesos añadidos para variables como las respuestas, pero esto no sucede en todos los parámetros propios de la plataforma de X.

Los resultados invitan a replicar este proceso del NLI en tres direcciones. En primer lugar, de cara a conversaciones en las OSNs sobre otras desinformaciones o sobre otras cuestiones, con *queries* manuales para la descarga de posts pero también automáticas [9]. En segundo lugar, para conocer cómo fluye la información falsa con y sin interacciones en relación a su temática para seguir la senda de los estudios que comprueban que, por ejemplo, las falsedades discurren mejor [36]. En tercer lugar, con el fin de cruzar las métricas propias de la red social, como los seguidores de los usuarios, para desentrañar más detalles sobre el fenómeno de la viralidad.

4.2.7.1. Resultados adicionales: métricas basadas en los autores del post

La caracterización de la desinformación obtenida no solo ofrece importantes conclusiones con las métricas propias de los posts de estas OSNs, como se ha podido comprobar, sino también con los indicadores de estas plataformas en relación a sus autores [253]. Esto hace que el estudio dentro del área computacional haya derivado también a investigaciones sobre estos perfiles [254, 255] ante estas falsedades. Por eso, los métodos de este apartado de la tesis sirvieron después en el trabajo conjunto con Noguera-Vivo et al. para estudios sobre información falsa con este fin [253].

De manera específica en el campo de la desinformación sobre la vacunación, se exploraron las dinámicas de los difusores del discurso antivacunas en el contexto de la pandemia de la COVID-19. De una descarga inicial con *queries* de 200.246 posts en X a partir de desmentidos de la organización de *fact-checking* Maldita, se separaron con NLI 36.292 que sí tenían que ver con la desinformación, y se distinguieron entre aquellos que la enuncian y aquellos que la contradicen.

De esta manera, la explotación de las métricas de sus diseminadores en la red social, también descargadas, permitió ver las diferencias entre los autores de las publicaciones con *Entailment* frente a los de los posts con *Contradiction* [253].

De estos usuarios, se hicieron grupos según sus características en este ecosistema social: el número de seguidores; su *status* de verificado y de no verificado entonces; su número de posts publicados; su antigüedad en la plataforma; su permanencia en listas públicas, y su ratio entre seguidores y seguidos. Esto permitió hacer el análisis de la proporción de publicaciones con *Entailment* y *Contradiction* para los grupos, pero también comprobar en cada uno de ellos qué parte ocupaban de este porcentaje sus creaciones originales, sus *reposts*, sus respuestas y sus citas a otros mensajes [253].

Gracias a aplicar estos avances de la tesis, el estudio reveló que los usuarios con más *followers* y aquellos entonces verificados compartían más desinformación creada directamente por ellos (contenido original no difundido por *reposts*); que la información falsa discurría más mediante perfiles nacidos en la red entre 2013 y 2020, y que la contradecían más aquellos con mayor reconocimiento, medido por la permanencia en listas y por el ratio entre seguidores y seguidos. Esto ratifica cómo el NLI no solo desvela patrones en las métricas de los posts, sino también en las de las cuentas a las que pertenecen [253].

Capítulo 5

Trazado de la desinformación mediante generación de grafos

La Aemet no ha bajado los umbrales de temperatura para poder declarar "más alertas por calor extremo".

— Newtral

En las secciones anteriores se han presentado dos elementos claves para poder abordar tareas de *fact-checking* guiadas por IA: en primer lugar, los modelos de similitud ayudan a filtrar los contenidos relevantes en términos semánticos; en segundo lugar, los modelos de NLI permiten valorar si dos frases están alineadas o si se contradicen. Entre las distintas aplicaciones que tienen ambas misiones, una de las más relevantes en esta investigación es la de analizar los mensajes en redes sociales y monitorizar cómo determinadas desinformaciones circulan y se distribuyen, así como el papel que juegan las cuentas que las diseminan y las que contrarrestan (por ejemplo, *fact-checkers*). Es ahora cuando se introduce el uso de grafos para entender cómo determinados contenidos falsos se originan y se difunden, y cómo ciertos usuarios toman un papel fundamental en esta distribución.

Para terminar en los grafos en torno a estas falsedades, se repite el proceso con una versión optimizada de la descarga de información de la API (5.1) y, a partir de aquí se aborda la aplicación del SNA como tal en esta tesis en el campo de la desinformación (5.2). Las anteriores secciones, a través de los avances presentados de PLN, permiten la representación con los nodos y aristas con la aportación del NLI incorporada (5.3), la meta tras aplicar todos los módulos de esta investigación. Los grafos se evalúan a partir del análisis de tres casos de estudio (5.4), para llegar después a los resultados de este apartado (5.5), presentados también en las redacciones.

5.1. Obtención de información

El primer paso de cara a construir una representación visual de la propagación de una pieza de desinformación es la obtención de un conjunto de datos suficientemente amplio de la red social a analizar. Las restricciones que imponen las API de cada plataforma influyen directamente en este aspecto. Así, espacios como Facebook conllevan importantes restricciones.

En esta investigación, el foco se puso en la red social X, previamente Twitter[1]. El primer paso consiste en la descarga de todos los posts relacionados con el contenido a analizar. Mediante la API de Twitter para académicos (*for Academic Research*) se pudo recuperar información mediante llamadas que reciben como entrada la *query*, formada con los términos representativos y palabras claves del bulo, como ya se realizó en las partes anteriores de la tesis.

En el caso de X, las búsquedas no tienen en cuenta la parte semántica y se restringen, a nivel general, a los términos introducidos. Esto limita los posts que se pueden obtener a aquellos con las palabras claves incluidas en la cadena de búsqueda. Para dinamizar este comportamiento, en los experimentos previos se propusieron cadenas de búsqueda más complejas, consistentes en la unión de palabras mediante operadores lógicos como 'AND' u 'OR', permitiendo así indicar varias alternativas para una misma palabra o variaciones. Por eso, la creación de una *query* más elaborada constituye *a priori* un ejercicio manual de ensayo y error a la hora de encontrar publicaciones sobre un tema.

5.1.1. Generación de cadenas de búsqueda

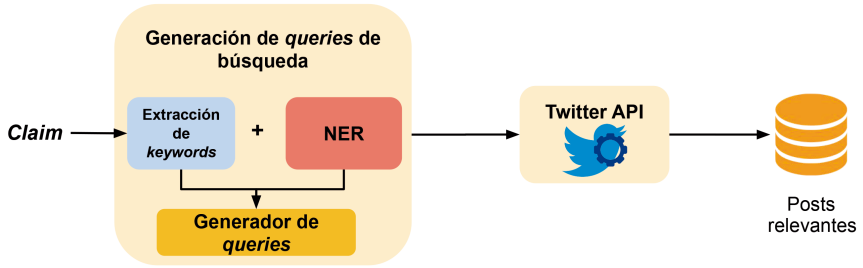


Figura 5.1: Flujo de generación de cadenas de búsqueda para la API de Twitter.

Este estudio, en línea con las metodologías semiautomáticas ya descritas contra la desinformación, se beneficia también de la automatización para generar *queries* como primer paso para obtener toda la conversación sobre las falsedades. Fig. 5.1 muestra el flujo seguido para la generación de estas *queries* y su envío a la API de Twitter. Se apuesta en esta investigación por los modelos *Transformer* multilingües para captar el significado de un *claim* y extraer sus palabras más importantes, que serán los términos a concatenar mediante 'AND' como el operador lógico de unión, de acuerdo a la investigación en esta línea [9]. Para los casos en español, se ha tenido en cuenta también el reconocimiento de entidades nombradas o NER. Modelos como KeyBERT [256] suponen un precedente para este tipo de actividades.

Sin embargo, la gran variedad de léxico del lenguaje implica que una misma afirmación se pueda escribir con un vocabulario y expresiones muy diferentes. Es por ello que construir estas cadenas de búsqueda no es un asunto trivial. Los términos más importantes extraídos dependen ya de esa subjetividad inicial de una frase que podría haber sido escrita de cualquier otra forma y, en

[1] A lo largo del desarrollo de esta tesis doctoral, la API académica de Twitter dejó de estar operativa, por lo que algunas aplicaciones previstas en esta investigación conllevarían determinados gastos, al tener que usar la API comercial. No obstante, la utilidad y vigencia de los métodos y herramientas no ha cambiado. Además, las técnicas planteadas son fácilmente extrapolables a otras redes similares como Bluesky o plataformas como Telegram.

consecuencia, la recogida de posts puede omitir una parte de la conversación sobre desinformación en la recopilación, con publicaciones que contengan algunas de estas palabras pero no todas.

Las mejoras en este sentido, bajo el modelo `MSTSb-paraphrase-multilingual-mpnet-base-v2` [233], derivado del STSb [257], y `bert-spanish-cased-finetuned-ner` para el NER, en base a investigación previa [9], incluyen: por un lado, la creación de todas las secuencias posibles alternando los términos con los operadores lógicos 'AND' y 'OR'; por otro, un parámetro para nivelar la cantidad de términos escogida para cada una de estas concatenaciones posibles, y, por último, el cambio de números a palabras para captar posts con cifras expresadas de cualquier manera. Este trabajo previo tuvo en cuenta el etiquetado multilingüe de Flair [258] y los modelos de Spacy [259] para eliminar conjunciones, adverbios y preposiciones, entre otras *stopwords* [9].

5.2. Aplicación de técnicas de SNA al análisis de la desinformación

El experimento del Capítulo 4 mostraba cómo, a través de NLI, las arquitecturas *Transformer* no solo permiten calcular con precisión distancias semánticas sino que propulsan la distinción de los posts para separar la información falsa en sí de sus desmentidos y del resto de contenidos no relacionados. Sin embargo, estos cálculos de distancias semánticas e inferencias del lenguaje natural caracterizan la desinformación pero no sus dinámicas a largo del tiempo ni sus formas de entrelazarse. Por eso, es necesario recurrir a técnicas de SNA para satisfacer esta tarea y construir la evolución de las publicaciones en torno a la información falsa en la red social.

Más allá de todo análisis mediante técnicas de PLN tras la descarga de las publicaciones, los grafos son la forma de representar estos flujos entre ellas. Como ya se ha abordado en el marco teórico, un grafo está formado por nodos y las aristas que los unen para identificar sus relaciones [178, 37, 4, 179]. Tal como se avanzó entonces, en el ámbito de las OSNs los nodos pueden ser los usuarios o las publicaciones asociadas a ellos y las aristas son las que indican, a partir de las métricas propias de la red social, las interacciones y vínculos entre ellos. A partir de aquí, toda relación puede establecerse a partir de las propiedades encontradas dentro de los posts descargados.

Por eso, las técnicas de SNA no solo comprenden, como su nombre indica, la explotación de las variables propias de estas plataformas, y es el grafo la forma de unir todas ellas para entender todos los nexos entre usuarios [178, 37, 4, 179]. Diferentes algoritmos entran en juego en esta disciplina, entrenando, por ejemplo, modelos a partir de los *likes* [260], para separar unas publicaciones de otras. El problema de la propagación actual de falsedades en estos ecosistemas ha postulado al SNA como una disciplina clave en el estudio de la desinformación, entre otras áreas como la política, cuando antes se reservaba más a turismo, marketing, ciberseguridad o salud [37].

El origen de los posts, las respuestas de los usuarios y la información sobre estos contenidos son, entre otros, los factores que se tienen cuenta para esta tarea [176], al demostrarse en el Capítulo 4 que también importan los indicadores propios de la plataforma que las acogen. Los matices no textuales, como aquellos sobre el comportamiento de las cuentas de la red, además de otros [52], también son relevantes.

Que se demostrara en 2018 cómo la información falsa llegaba más lejos en Twitter que la verdadera (todo ello a partir de las variables de la profundidad de la cascada generada, los usuarios que promovían tal contenido y cuánto duraba la difusión) [36] anima a no menospreciar el SNA. Con

la pandemia se descubrió que esta propagación mayor dependía también de la amplitud de las cuentas difusoras, ya que los usuarios que expandían estos posts nocivos publicaban más que los que no los propulsaban [261], comprobando también la importancia de no ignorar el análisis de las redes sociales en esta ecuación.

Aquí entra en juego la cuestión de la viralidad. Ya antes de la COVID-19, mediante técnicas de análisis de redes se demostró que los círculos en las conversaciones sobre vacunación incidían en la decisión final de vacunarse o no [39], y, por tanto, si el círculo es antivacunas, existe el riesgo de rechazar esta prevención. En esencia, no solo es agente desinformador el creador de una falsedad, sino todos los que en ese círculo contribuyen a ella, contagiando al resto. Estos contagios aluden a los *viral models* [7], además de a las concepciones de los modelos epidémicos y de cascada [5, 6], pero con un matiz: el contagio no tiene por qué darse de una única cascada, pues esto sería no salir del todo de la concepción de los *broadcast models*, precisamente opuestos a los *viral models*, porque se aceptaría que la desinformación solo va también de un emisor al resto de receptores [7] (aunque con intermediarios) y no como resultado de distintos focos ajenos entre sí.

Los propagadores principales de la información falsa (*super-spreaders*) y el paso de contenidos de unos círculos a otros por obra de estos han formado parte de las investigaciones en forma de grafos [38]. Antes de la etapa de X con Elon Musk, se demostró que los perfiles entonces verificados eran 50 veces más potentes para expandir cuestiones sobre vacunas en relación a las cuentas sin verificar [262]. En la época de la COVID-19, una tercera parte de la desinformación procedía de las acciones de famosos e instituciones en esta red [70], y, en otras situaciones, las cuentas de políticos también se han encargado de la expansión de las falsedades [263].

Pero en estos contextos, no solo contribuyen estos perfiles, sino también aquellos anónimos [263] y, además, las investigaciones también referencian a los *bots* dentro de la expansión [252]. Se forma así una combinación de cuentas influyentes y desconocidas en la difusión de desinformación, donde, en el caso de la irrupción del coronavirus, los *hashtags* de la cuestión [40, 41] también amplificaron la información falsa. En estos ámbitos, las *keywords* se convierten en objeto para las metodologías que siguen las redes y sus comunidades [117].

5.3. Reconstrucción de la cascada de desinformación mediante NLI y grafos

Dado el avance del SNA en los estudios de desinformación, el análisis de las métricas de las redes y sus grafos puede ir de la mano con el análisis textual de los contenidos para comprender más las dinámicas de las falsedades que discurren en X. A través del apartado anterior y también mediante el marco teórico de esta tesis, queda claro que detrás de cada mensaje de la plataforma está toda la comunidad de cuentas que interactúan y que, por tanto, también influyen en el movimiento de la información falsa. El uso aislado de PLN ignoraría todo ese flujo de las OSNs, que otorga también conocimiento en esta problemática y permite la tarea final de descubrir y trazar su impacto.

La fusión de PLN y SNA eleva la investigación hacia una radiografía más realista de las conversaciones que giran en torno a la desinformación. Dentro del contagio de un nodo a otro en un grafo, el PLN ha demostrado en el Capítulo 4 que la desinformación no se trata de un mismo mensaje expandido en una única cascada sino de una propagación descentralizada de varios contenidos diferentes e independientes entre sí. Como ya se ha abordado antes, abrazar el concepto de los *viral models* y partir, en consecuencia, de un propagador inicial para el contagio al resto

de individuos en forma de árbol no dista mucho de los *broadcast models* [7], con un único emisor masivo y varios receptores.

De esta manera, PLN y SNA configuran una visión alternativa de la información falsa donde se tiene en cuenta el máximo número de publicaciones, a partir de la búsqueda con *queries* dentro de una plataforma, y no solo las que están conectadas al post de mayor alcance. Así, se ha apreciado cómo una parte de conversación sobre la desinformación surge a raíz de la viralización de posts, pero otra la conforman contenidos de todo tipo de más o menos interacciones, o incluso ninguna [264]. Los usuarios que participan en ella también varían de mayor a menor popularidad [253].

Los descubrimientos del Capítulo 4, sin embargo, no van más allá de la tipología de las publicaciones en función a sus interacciones y a su relación dentro del NLI (*Entailment* o *Contradiction*). Al excluirse la parte de generación de grafos, propia del SNA, desaparece la posibilidad de visualizar ese ecosistema alternativo de la información falsa. Así, se produce una brecha entre la parte teórica a partir de tal experimento, la cual ha revelado que la desinformación no fluye solo en forma de cascada jerárquica desde el post con más interacciones, y la parte práctica, la cual representaría cómo fluye en realidad a través de los nodos y sus conexiones, y permitiría ver el recorrido de principio a fin de las falsedades en la plataforma en una franja de tiempo seleccionada.

Esta última parte de la tesis consiste, por tanto, en plasmar del trazado de los contenidos en una conversación sobre desinformación mediante grafos para entender sus dinámicas y evolución a lo largo del tiempo. La representación gráfica mostrará las publicaciones y las cuentas que han formado parte de estas difusiones, independientemente de sus interacciones, para llegar a la visualización más realista posible de la información falsa y de sus desmentidos.

Aquí entra en juego el área de los sistemas de computación cognitiva: ya no se trata de la mera aplicación de los algoritmos, sino de enfocarlos para recrear y facilitar la resolución humana de problemas [265]. En este caso, aquellos vinculados a los profesionales del *fact-checking* para descubrir la información falsa y seguir su progreso. El PLN se erige como uno de los campos que vela por estos sistemas cognitivos y su análisis dentro de las plataformas sociales. Asimismo, el SNA se enmarca también en los *Decision Support Systems* (DSS), estructuras de computación cognitiva que aúnan los resultados de combinar la algoritmia y el tratamiento elaborado de datos [265, 266].

A nivel técnico, esta metodología supone la unión de los tres módulos para trazar la información falsa en X (ver Fig. 5.2): la descarga de posts, la aplicación de *Transformers* para NLI y el uso de SNA para la parte de grafos, además de para la explotación de las métricas propias de la red social. Sobre ellos, se apuntan los siguientes matices:

- **Descarga de posts**: el Capítulo 3 de esta tesis ha mostrado la posibilidad de capturar todos los posts relevantes de una conversación en redes sociales sobre una información falsa. Partiendo de una serie de *keywords*, se construye una cadena de búsqueda o *query* que se lanza a la API de la red social para descargar estas publicaciones, convertidas posteriormente en vectores semánticos o *embeddings*.

- **Filtrado semántico e inferencia del lenguaje**: el Capítulo 3 ha plasmado el potencial de la aplicación de modelos de lenguaje para determinar los posts semánticamente relacionados con el *claim* original. Después, el Capítulo 4 ha revelado cómo el uso de modelos de NLI

permite etiquetar cada texto en función de si apoya o refuta una falsedad. Ahora, el Capítulo 5 tomará también el fin último de estos métodos de categorizar los posts como *Entailment* (la desinformación en sí), *Contradiction* (lo contrario a ella) y *Neutral* (nada en relación).

- **Generación de grafos**: es en esta parte de la tesis cuando se culminará con las creaciones de grafos que trazarán la conversación de una información falsa. Cada post descargado constituirá un nodo; sus interacciones, las conexiones entre ellos. En esta unión de PLN y SNA, los datos obtenidos con NLI se reflejarán en los nodos, configurando la visualización final.

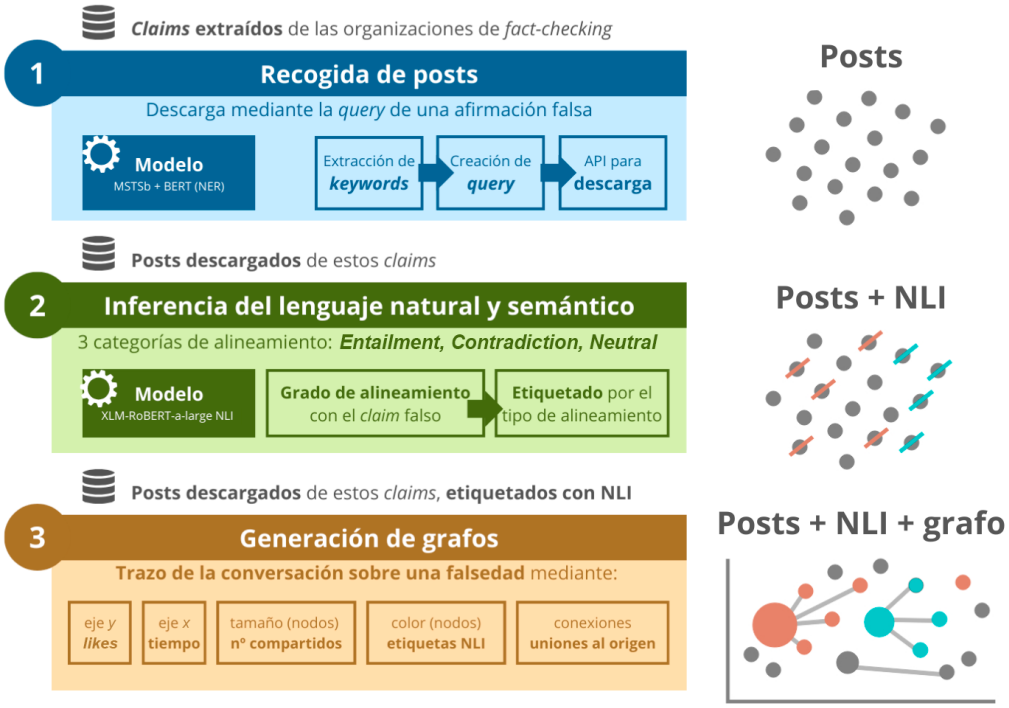


Figura 5.2: Los tres pasos para seguir la conversación sobre desinformación a partir de todas las secciones en conjunto.

El grafo final como resultado de los tres módulos constituye, a nivel práctico, un centro de monitorización para cada información falsa que emerge en X. Permite controlar la diseminación de los posts relacionados con cada *claim*, sus propagadores con más impacto, su evolución a partir del resto de cuentas y las franjas de tiempo en las que la desinformación en sí acumula más publicaciones dentro de la conversación. En este grafo, quedan plasmadas las conexiones: si un post, representado con un nodo, es la respuesta, citado o *repost* de otra publicación, este post precedente será también otro nodo, el nodo padre, y estarán conectados por una arista. Esto se debe a que la información descargada también ha tenido en cuenta el post y el autor del que deriva cada publicación.

Llegar a este punto es otorgar valor a los grafos como paso final dentro del SNA para el estudio de la desinformación, pero también dar utilidad a los módulos anteriores. De nuevo, de forma independiente se establecen las bases para el análisis de los contenidos dependiendo de si son *Entailment* o *Contradiction* y de las características propias de la red social. Son estos módulos

previos al culmen con el grafo los que vuelven a poner sobre la mesa el ecosistema de posts con muchas interacciones junto a aquellos con pocas, cuestionando otra vez la atención solo a lo más viral [264].

Gracias al uso de los metadatos descargados de cada post, el grafo pasa a tener una aplicación práctica como arquitectura de monitorización de los contenidos y autores asociados a cada conversación de la desinformación a lo largo del tiempo. De esta forma, el modelado con NLI más el SNA a modo de grafos es la fórmula para una herramienta de visualización para el trazado de la desinformación desde su posible origen. Fig. 5.3 muestra un ejemplo del tipo de representación buscada.

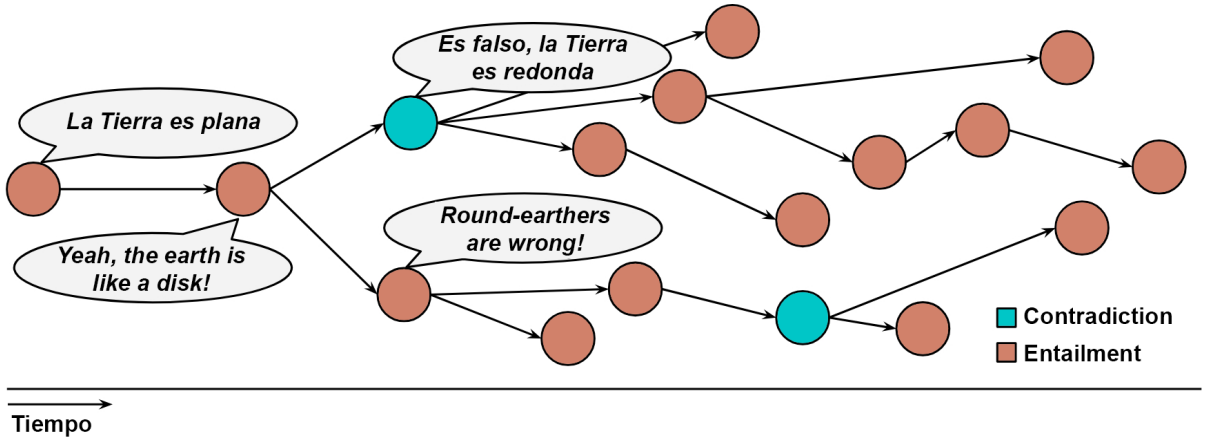


Figura 5.3: Explicación visual del grafo con la conversación sobre una información falsa trazada.

El orden cronológico representa el eje x. Difiere de la variable del tiempo en sí, pues la intención no es una representación cronológica lineal sino una representación discreta del orden. Hay dos motivos detrás de esto: primero, no solapar los contenidos en los momentos de mayor afluencia de posts, cuando la información falsa y su conversación tienen su pico en la red; segundo, no dejar espacios vacíos correspondientes a los momentos en los que tal desinformación no ha estado presente en redes (casos de, por ejemplo, inactividad sobre una falsedad concreta que luego resurja). Así, si hay muchas publicaciones en un breve plazo de tiempo, esta decisión dará a tales posts la misma importancia que al resto. Si los siguientes posts ocurren, de lo contrario, muy después, en este orden discreto aparecerán a continuación, sin un espacio en el eje x que haga el gráfico menos legible.

El eje y define la influencia de las cuentas de la conversación sobre desinformación a partir de su número de seguidores, en escala logarítmica para una visualización más clara de aquellas cuentas con mayor impacto por sí mismas frente a aquellas que tienen menos, pero que también participan. De esta forma, dos posts que estén como nodos situados uno a continuación del otro en el orden cronológico (eje x) tendrán una posición distinta en el eje y dependiendo del número de seguidores de la cuenta que haya originado cada uno de ellos.

Pero el impacto del post en sí se define por el número de *likes*, además de por el número de citados, *reposts* y respuestas, ya incluidos en el grafo en las uniones con aristas. Por eso, el tamaño de los nodos lo establece la cantidad de 'me gusta' que han recibido, dando más peso en la visualización a aquellos con más interacciones en este sentido frente a los que menos. Como resultado, esta evolución no solo muestra el recorrido en el tiempo de la conversación sobre una

información falsa, sino que también, en esa aproximación realista hacia la representación de este ecosistema, busca plasmar todos los planos del impacto de las cuentas y sus posts a través de las métricas de X.

Por último, el aporte diferencial del NLI en el grafo lo dará el color dependiendo del tipo de alineación del post con el *claim*. Para este experimento, los nodos con *Entailment* serán naranjas; aquellos con *Contradiction*, azules; los que no tengan relación con el *claim*, grises. De esta forma, se integran tanto NLI y SNA y la visualización ya no consiste solo en la unión de todos los posts descargados a partir de unas *keywords* sobre una falsedad, sino en una monitorización de los posts con desinformación y aquellos que la contradicen, dentro de toda la maraña de mensajes en la conversación generada.

5.4. Casos de estudio

Se han escogido tres informaciones falsas de tres temas distintos para comprobar el funcionamiento de todos los pasos de esta tesis en conjunto. En línea con los anteriores módulos, se han descargado todos los posts alrededor de la conversación de las tres desinformaciones elegidas y se ha realizado un análisis exploratorio que atiende a las métricas desglosadas en el Capítulo 4, también a aquellas en relación a los autores de tales publicaciones [264, 253]. El último módulo será el que concluya con la generación de los grafos finales con las propiedades de estos pasos anteriores para recorrer la trayectoria del contenido sobre las falsedades seleccionadas.

5.4.1. Análisis exploratorio

Los tres temas diferentes escogidos para esta metodología son: por un lado, el descrédito institucional y la islamofobia, por otro, las proclamas antivacunas de la era del coronavirus, y, por último, los comentarios vertidos de la guerra entre Ucrania y Rusia:

- **Caso 1: "El 80 % de los musulmanes viviendo en Europa viven de las ayudas sociales y rechazan trabajar"**. Afirmaciones como esta pueden mermar el apoyo a las instituciones y dirigir el odio hacia segmentos de población como, en este caso, la comunidad musulmana. La muestra contiene 32 posts originales, más de ellos con *Entailment* que con *Contradiction*. Con los *reposts* en el recuento, la suma de publicaciones en relación a esta desinformación llega a 84.

- **Caso 2: "Las vacunas de ARN contra el coronavirus incluyen óxido de grafeno"**. Frases como esta son desinformaciones antivacunas en el contexto de la COVID-19. Como el Caso 1, también son 32 los posts originales, aunque en este ejemplo hay un equilibrio entre los contenidos con *Entailment* y *Contradiction*. Contando los *reposts*, el conjunto a analizar constituye 128 publicaciones sobre esta falsedad.

- **Caso 3: "Zelenski vendió 17 millones de hectáreas de tierra a Monsanto, Dupont y Cargill"**. Este *claim* tiene lugar entre toda la desinformación sobre la guerra Rusia-Ucrania. Lo componen 26 posts originales y prima el *Entailment* sobre *Contradiction*. Se aprecia el grueso de los usuarios que han compartido el contenido original, pues en total para este ejemplo son 916 posts vinculados a esta falsedad, contando los *reposts*.

En los Casos 1 y 3 (ver Fig. 5.4), hay más publicaciones que difunden la información falsa que aquellas que la contradicen, contando el peso de los posts sobre *fact-checking*. El Caso 2, en el

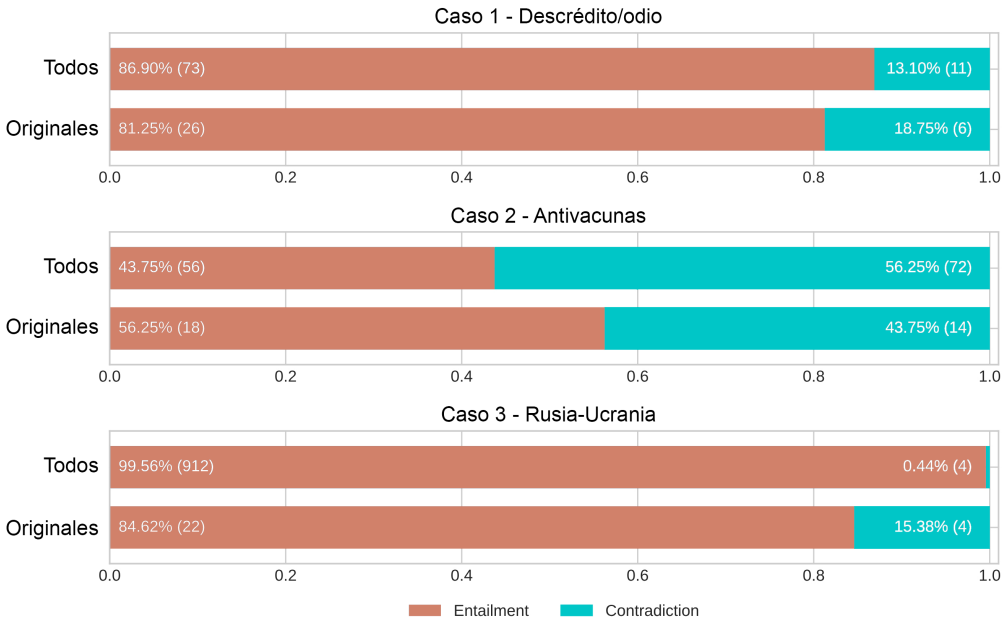


Figura 5.4: Proporción de posts con *Entailment* y *Contradiction*, en total y solo originales, para cada caso.

contexto de la pandemia, muestra sin embargo cómo los usuarios también se dedican a expresar lo contrario a tal falsedad, equilibrando la balanza.

Los *reposts* y *likes* (Fig. 5.5) son reveladores en el estudio de la desinformación para entender el impacto de los mensajes sobre ella. Predominan los posts de uno a diez *reposts*, seguidos de los de cero. En cuanto a *likes*, destacan las publicaciones que no han recibido ningún 'me gusta', pues las que van de uno a diez tienen una proporción más baja. Cuando van de 11 a 100 *likes*, únicamente alcanza el Caso 2 un porcentaje considerable con la desinformación antivacunas. Solo una publicación en el Caso 3 sobre el conflicto Rusia-Ucrania logra sobrepasar los 100 'me gusta'.

En línea con el Capítulo 4, obedecen a la tendencia estudiada en la plataforma social [264]: son más abundantes los posts originales con menos repercusión que aquellos mas virales en términos de *reposts*. También destacan más en cantidad aquellas publicaciones originales sin ningún *like* en comparación a otras que sí lo han recibido. Se vuelve a mostrar así un ecosistema de la información falsa donde no solo repercuten los mensajes con más impacto en interacciones, sino aquellos que discurren con un número pequeño o inexistente de ellas.

El análisis adicional de los *followers* (Fig. 5.6) ayuda también a entender la dinámica de estas desinformaciones, y corresponde a la línea ya explorada en la aportación extra del Capítulo 4 aplicada a otros estudios [253]. En los tres casos, la mayoría de las creaciones originales corresponde a cuentas con al menos un centenar de seguidores. En el Caso 1, en relación al descrédito y al odio, estos contenidos propios provienen más de cuentas entre 1.001 y 10.000 *followers*; en los Casos 2 y 3 el grueso pertenece a aquellos entre 101 a 1.000. En los tres ejemplos, no son tantos los usuarios con más de 10.000 seguidores.

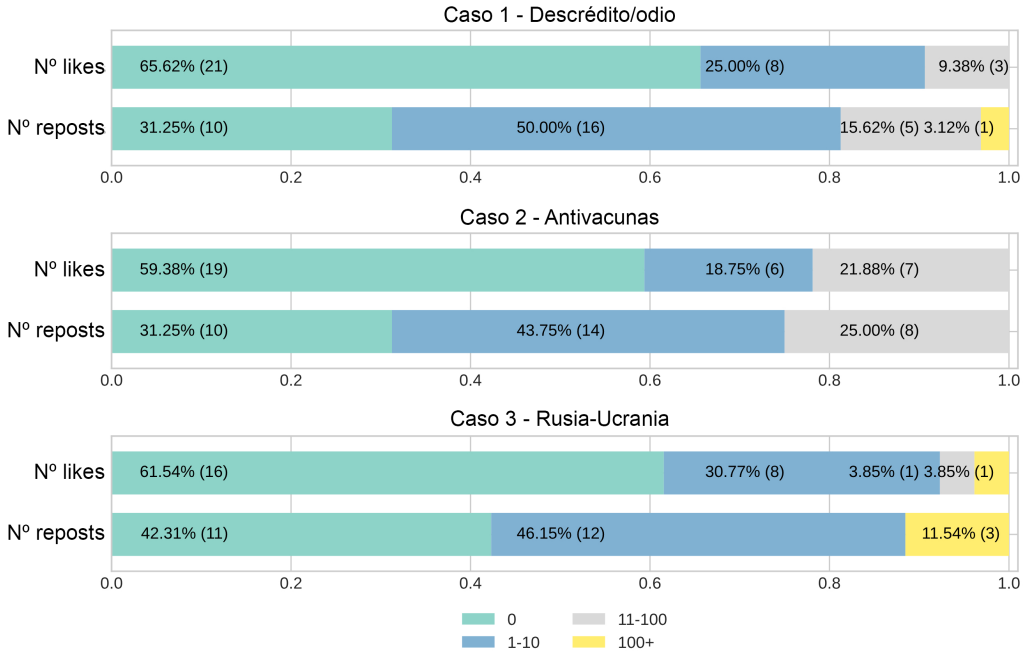


Figura 5.5: Proporción de posts por grupos según el número de *likes* y *reposts*, para cada caso.

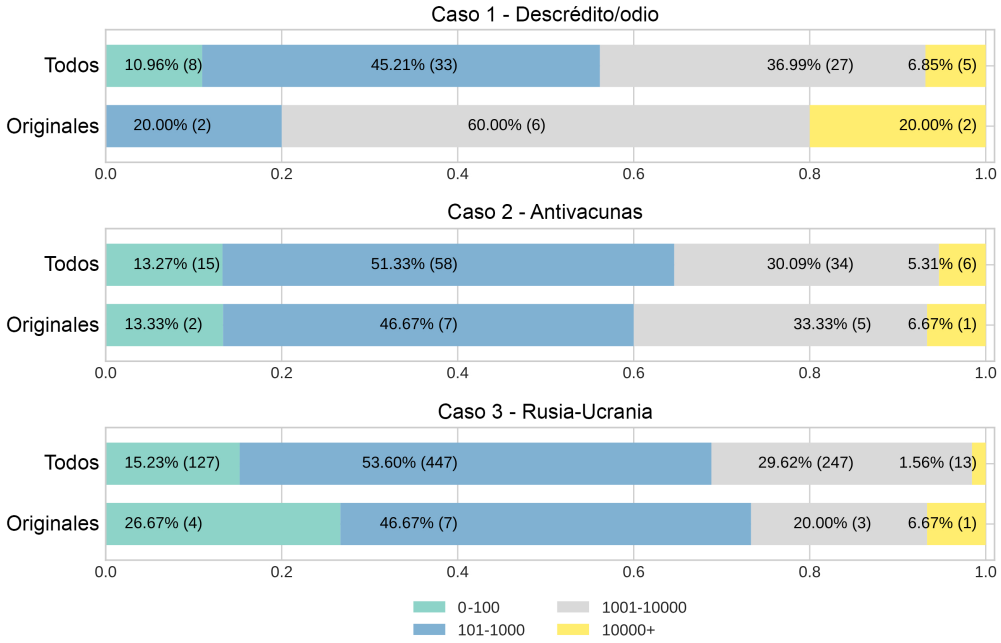


Figura 5.6: Proporción de posts por grupos, en total y solo originales, según el número de *followers* de quienes los comparten, para cada caso.

5.4.2. Visualización de los grafos

El módulo 5 culmina en un grafo que traza la conversación de la desinformación para cada uno de los tres casos de estudio, a partir de los nodos y uniones que lo componen. Cada publicación se representa en forma de nodos circulares, más grandes o más pequeños en función de su número

de 'me gusta', mientras que los rombos aluden a los *reposts*. La monitorización es posible con la representación de las publicaciones en orden cronológico en el eje x, mientras que la cantidad de seguidores de los partícipes de cada mensaje se representa en el eje y (en escala logarítmica).

5.4.2.1. Caso 1

El primer ejemplo (Fig. 5.7) de estudio concierne la conversación sobre la falsa información en torno al *claim* "El 80 por ciento de los musulmanes viviendo en Europa viven de las ayudas sociales y renuncian a trabajar". Esta desinformación, reproducida a través del grafo generado mediante los posts con *Entailment*, se repite a lo largo de las publicaciones, a las que se añaden alusiones a otras cuentas, menciones a supuestos investigadores para vestir de credibilidad los posts falsos y comentarios a modo personal en favor de tal falsedad.

Autor	**Nº. seguidores**	**Nº. interacciones**	**Máx. Nº. *reposts***	**Máx. Nº. *likes***	**Nº. posts**
0	25.464	0	0	0	3
1	12.537	7	7	5	4
2	8.795	4	1	2	1
3	3.881	1	1	0	1
4	3.856	0	0	1	1
5	2.669	0	0	1	1
6	2.641	27	22	72	2
7	1.141	1	1	2	1
8	417	0	0	0	1
9	146	38	32	60	5

Tabla 5.1: Ranking de las cuentas activas en el Caso 1, ordenadas por número de seguidores.

En cuanto a los *fact-checks*, detectados en el grafo como posts etiquetados como *Contradiction*, son enunciados del *claim* en la forma negativa para decir lo opuesto. Son frases que, además, vienen acompañadas de los usuarios oficiales de *fact-checking* en X y/o los hipervínculos de los desmentidos de tal información que han difundido en su medio. Estos *fact-checks* se distribuyen en ocasiones muy alejados entre sí pero son de aspecto similar.

Diez meses antes de la publicación más reposteada difundiendo la falsa información el 2 de abril de 2022 ya circulaban posts con *Entailment*, según el grafo. En ese momento, el 8 de junio de 2021, un post con esta desinformación recibió solo un *repost* y fue ya cinco meses antes de la publicación más reposteada cuando otro post, el segundo con más *reposts* (11 en total), tuvo más repercusión activa entre los usuarios. Los mensajes que reproducen el *claim* aparecen hasta el último tramo del año, de acuerdo a la visualización.

Dentro de los posts con *Entailment*, dos usuarios con más de 10.000 seguidores que crearon sus propias publicaciones con información falsa lo hacen, además, varias veces a lo largo del tiempo (ver Tabla 5.1): uno (con 25.464 seguidores en el momento de la captura), tres veces; el otro (con 12.537), cuatro. A partir del *repost*, otras cuentas con más de 10.000 *followers* también difunden la desinformación. Dentro de las publicaciones contrarias a la falsedad, Maldito Bulo (una de las cuentas de la organización de *fact-checking* Maldita) contribuye a combatir la desinformación en la red social con el *repost* de usuarios que citan sus desmentidos en este caso.

Este grafo muestra cómo toda la falsedad no nace de una cascada a partir del post más viral, en este caso, en abril de 2022. Antes de esa viralidad, la publicación con solo un *repost* en junio de

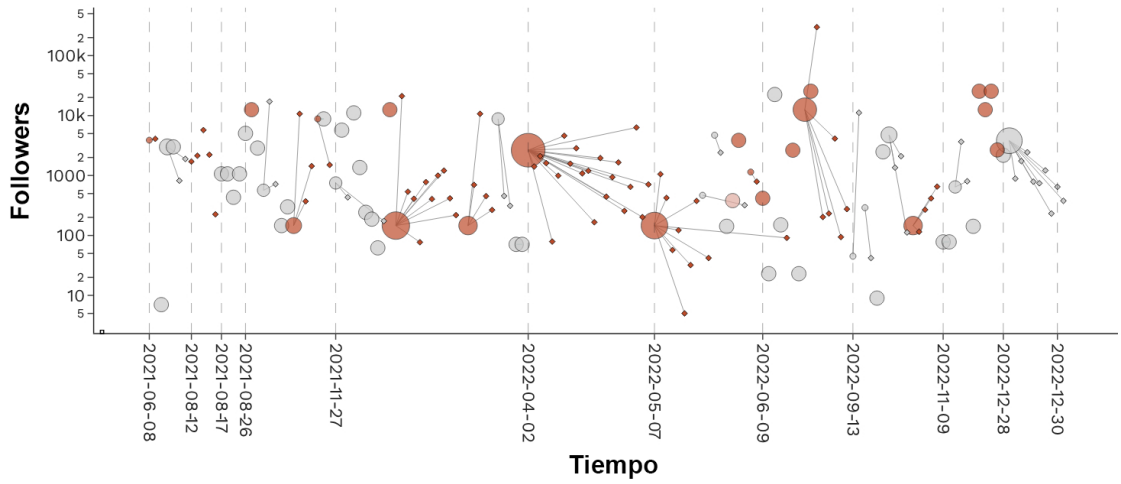


Figura 5.7: Visualización del grafo perteneciente al *claim* del Caso 1.

2021 muestra cómo ya la información falsa circulaba a través de cuentas con menos interacciones. La senda de la desinformación tampoco acaba en este caso tras el mensaje más viral de abril u otros cercanos, sino que se extiende a través de distintas cuentas y enunciados hasta el último trimestre de 2022.

5.4.2.2. Caso 2

El segundo caso de estudio, mostrado en Fig. 5.8, engloba la conversación sobre desinformación a través del *claim* "las vacunas de ARN mensajero contra la COVID-19 contienen óxido de grafeno". Mientras que los ejemplos con *Entailment* del caso anterior estaban muy encorsetados a la composición del enunciado del *claim*, los de este caso destacan por su parafraseo como recurso para expandir la desinformación de distintas formas. En este grafo se suceden publicaciones más enunciativas y otras más emocionales.

Autor	**Nº. seguidores**	**Nº. interacciones**	**Máx. Nº. *reposts***	**Máx. Nº. *likes***	**Nº. posts**
0	9.547	26	23	23	1
1	1.903	6	6	13	1
2	1.854	8	1	4	3
3	632	8	6	10	1
4	445	9	9	11	1
5	248	2	1	1	1
6	158	11	9	11	1
7	116	2	1	2	3
8	81	0	0	0	1
9	1	0	0	1	1

Tabla 5.2: Ranking de las cuentas activas en el Caso 2, ordenadas por número de seguidores.

Dentro de los ejemplos con *Contradiction*, no varían tanto los *fact-checks*, en la línea del caso de estudio anterior. Pero esta vez, a diferencia de antes, el primer enunciado opuesto al significado del *claim* en el grafo no viene de un desmentido, sino de otra cuenta cuya frase también expresa lo contrario a la información falsa, algo que el sistema de NLI ha detectado.

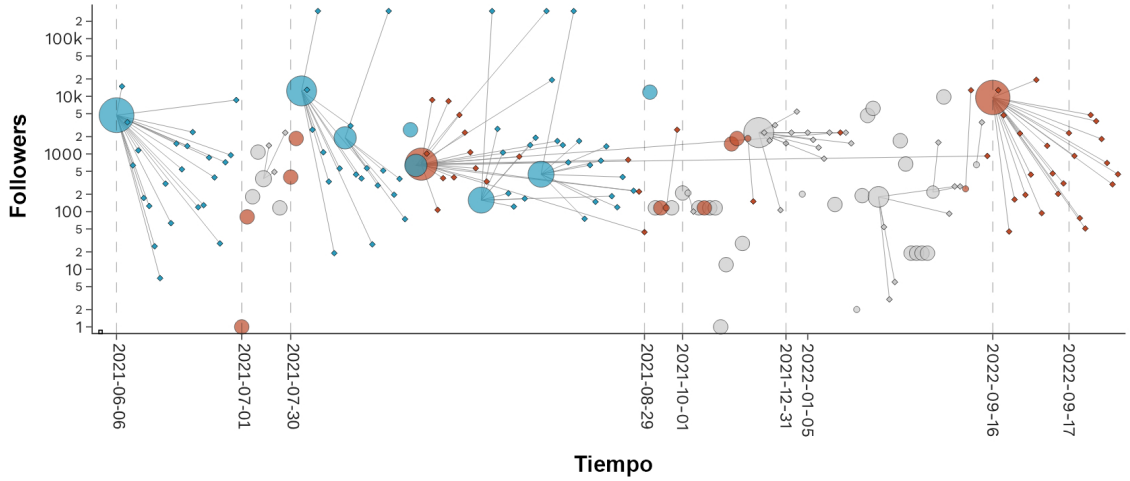


Figura 5.8: Visualización del grafo perteneciente al *claim* del Caso 2.

Esta publicación con *Contradiction*, la más reposteada (24 *reposts*), se dio el 6 de junio de 2021 y da inicio al grafo. La falsedad en sí aparece casi un mes después y se replica después varias veces hasta septiembre de 2022, fecha en la que llega el texto más reposteado con la etiqueta de *Entailment* (23 *reposts*), tras varios meses con contenidos de la categoría *Neutral*. Como se aprecia también en el caso anterior, tal mensaje no es el eje de la difusión de información falsa, sino que antes y después ya la propagan otros posts y *reposts*.

En este caso, los usuarios con más de 10.000 seguidores sí han creado publicaciones propias con *Contradiction* (ver Tabla 5.2) pero no con *Entailment* (se le acerca un perfil con 9.547 seguidores en el momento de la recogida de datos). De nuevo, la cuenta oficial Maldito Bulo, de la organización de *fact-checking* Maldita, expande con *reposts* los enunciados de otros usuarios que citan sus desmentidos, además de otro usuario con 19.269 seguidores.

El grafo indica cómo, otra vez, las publicaciones más virales son solo una mínima porción de toda la información falsa que circula, con diferentes paráfrasis. La conversación sobre la desinformación se amplía en esta visualización a través de cuentas que la propagan o la enfrentan con enunciados contrarios, como se puede ver en el último post con *Entailment* en septiembre de 2022, posterior a los desmentidos.

5.4.2.3. Caso 3

El tercer ejemplo (Fig. 5.9) de estudio atañe la conversación sobre la información falsa del *claim* "Zelenski vendió 17 millones de hectáreas de tierra a Monsanto, Dupont y Cargill". La abundancia de nodos (usuarios) destaca en este caso respecto a los otros, pero en cuanto a forma y contenido se asemeja a los anteriores por los distintos recursos para enunciar una misma desinformación, si bien en este caso el grueso de la falsedad lo comprenden *reposts* de un mismo contenido.

Esta publicación (ver Tabla 5.3), el 19 de septiembre de 2022 (663 *reposts*, 1.000 *likes*) es protagonista no solo en su irrupción sino también después, ya que sus *reposts* se extienden hasta tres meses después, el 30 de diciembre. En el grafo se aprecian tanto el momento de mayor impacto con el estallido de este mensaje más viral como el de menor impacto con estos *reposts* de meses posteriores, gracias a la distribución a lo largo del eje x, que no mapea los nodos en función a

cuándo ocurrieron, pero sí los ordena cronológicamente en una secuencia con la misma separación entre unos y otros, con independencia de si surgen más o menos cercanos en el tiempo.

Autor	Nº. seguidores	Nº. interacciones	Máx. Nº. *reposts*	Máx. Nº. *likes*	Nº. posts
0	43.502	774	663	1.000	1
1	2.933	8	7	11	1
2	2.211	0	0	1	1
3	1.925	3	1	1	1
4	843	2	2	4	1
5	248	2	1	4	1
6	223	0	0	1	1
7	147	0	0	0	1
8	113	1	0	1	1
9	8	0	0	0	1

Tabla 5.3: Ranking de las cuentas activas en el Caso 3, ordenadas por número de seguidores.

Se pueden hacer más apreciaciones a partir del grafo. Por ejemplo, se muestra cómo un mismo usuario republica el mismo post viral, propagándose en momentos muy espaciados en el tiempo. También se observa cómo el grafo lo inicia otra serie de *reposts* antes del mensaje más viral. Estos *reposts* no están unidos con ninguna arista, por lo que se entiende, en relación a la API de Twitter en ese momento, que el post original (con 1.597 *reposts*, de acuerdo a la información de los nodos que le suceden) se suspendió o eliminó.

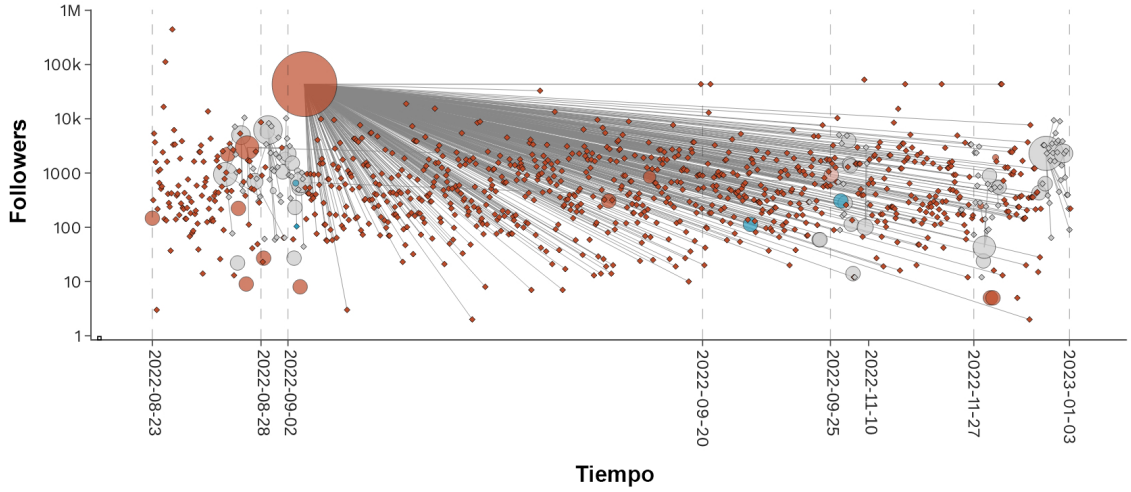


Figura 5.9: Visualización del grafo perteneciente al *claim* del Caso 3.

Si bien en este ejemplo una cascada sí capturaría la mayoría de los posts en torno a la información falsa, al surgir como *reposts* de la publicación más viral, la estructura de este grafo permite, a lo largo de los ejes x e y, observar las dinámicas a lo largo del tiempo, como la presencia de *reposts* de la publicación viral muy alejados temporalmente, y la tipología de los usuarios claves. De esta forma, el mapa de la desinformación no es solo una sucesión de nodos a partir del nodo principal con más *reposts*, sino un entendimiento de cómo ha fluido la conversación de la información falsa antes, durante y después de este mensaje.

5.5. Resultados experimentales

Los tres casos de estudio muestran el éxito de la aplicación de los *Transformers* [35, 35] para construir grafos de la conversación en torno a la información falsa. La inferencia del lenguaje ha permitido detectar los enunciados de la desinformación como tal y aquellos que expresan el sentido contrario en sus distintas formas, desde aquellas más cercanas al *claim* original (Caso 1) a otras más lejanas a partir de parafraseos (Caso 2 y Caso 3).

En los grafos de estos tres casos, se intercalan en la propagación de la información falsa cuentas con muchos seguidores, los ya denominados 'superpropagadores' en los estudios de desinformación [38, 262], con aquellas con menos *followers*, mostrando que no solo se expanden los mensajes a través de la viralidad, sino mediante toda la conversación. Esto permite y anima la investigación exhaustiva del comportamiento de las falsedades en las redes sociales ante las grandes crisis y de fenómenos como *bots* durante ellas [252].

La necesidad de los anteriores módulos se hace factible ya previamente en el análisis exploratorio. Sus resultados revelan la heterogeneidad de contenidos y usuarios, de mayor a menor influencia [264, 253] en función a su número de interacciones y *followers*, respectivamente, para la aproximación realista del mapeo de la desinformación. Este proceso muestra la importancia de estudios más enfocados en la extracción de mensajes de otras redes y en su aprovechamiento mediante NLI.

Lo que distingue la generación de grafos como paso final es el conocimiento obtenido a partir del rastreo de una falsedad, máxima expresión de la combinación del PLN y del SNA como áreas claves para luchar contra la desinformación [32]. Los tres casos de estudio indican que una información falsa no tiene por qué desaparecer poco a poco, sino que puede resurgir, y que esta tampoco tiene por qué originarse en el momento de la publicación viral, sino ya antes. A través de ellos, se observa también cómo los posts con *Contradiction*, de *fact-checkers* y otros, no solo nacen a raíz del momento de mayor afluencia, sino también en otras franjas de tiempo.

Estas visualizaciones de la desinformación pero también del contenido que la contradice y la desmiente a partir del *fact-checking* dan un paso más allá respecto a los experimentos de la sección anterior. Sin la parte de grafos, *Entailment* y *Contradiction* solo serían las categorías de los posts sobre las cuales se podrían extraer estadísticas [264, 253], pero las representaciones gráficas de los tres casos de estudio demuestran su utilidad para elaborar sistemas de supervisión en esta lucha, a través de las falsedades en sí y de los posts que dicen lo contrario. Esto permitiría a *fact-checkers* y agentes contra la información falsa una ayuda en la respuesta coordinada contra este problema.

5.5.1. Resultados adicionales: puesta en práctica en redacciones

Los enfoques de PLN a partir de la experimentación con *embeddings* abren un camino de metodologías prácticas fuera de la clasificación. Cuestiones como la distancia coseno de los posts convertidos a vectores [152, 35], abordados en el Capítulo 3 de esta tesis, se erigen como alternativa por las organizaciones para mitigar la desinformación. Una aplicación que no tiene por qué ser hacia una plataforma en concreto, sino hacia múltiples redes y contextos.

Esta línea de investigación de similitud semántica entre publicaciones ya se ha implementado en las redacciones [267], en situaciones como la vivida con la desinformación de la COVID-19. Así, los *fact-checkers* pueden almacenar toda falsedad desmentida en forma de *claims* para generar así

mejores bases del conocimiento para cuestiones concretas [221] que no terminan en el momento de la negación de tal información falsa, sino que continúa con la semiautomatización de la tarea.

Si estos métodos ya forman parte de las aplicaciones reales contra la desinformación en las organizaciones de *fact-checking*, también puede serlo todo este camino de la tesis. La unión de estos pasos fue la idea original de la que partió el proyecto DisTrack, uno de los ganadores para financiación del *Call for solutions: Tech against Disinformation*, en colaboración con medios y organizaciones de *fact-checking*[2].

Para ello, se utilizaron los módulos de descarga de posts, de empleo de NLI y de generación de grafos de este apartado para ofrecer una herramienta para los *fact-checkers* a través de las informaciones falsas detectadas por ellos. Así, los resultados en forma de visualizaciones de los nodos y sus conexiones para rastrear las conversaciones de cada falsedad se presentaron como una aplicación interactiva a modo de ayuda para estos profesionales.

[2]La convocatoria y la descripción del piloto están publicados en `https://digitalfuturesociety.com/disinformation-fakenews/`

Capítulo 6

Respuestas a las preguntas de la investigación, discusión y conclusiones

Es falso que la Antártida haya ganado hielo
y que esto demuestra que el cambio climático no existe:
se trata de una tergiversación de un estudio científico.

— Chequeado

Tras todos los pasos seguidos, esta sección trae de vuelta las preguntas y subpreguntas del Capítulo 1 para darles una respuesta tras resolver los módulos propuestos de la investigación (6.1). Con los resultados de todos los experimentos y el marco teórico que los motiva, se hace balance de los métodos planteados y se plantean futuras líneas de trabajo (6.2). Tras las respuestas y la reflexión por ellas, se cierra la tesis con las conclusiones (6.3).

6.1. Respuesta a las preguntas de la investigación

El trabajo realizado a lo largo de los capítulos y sus experimentos permite responder las RQ de la tesis:

- **RQ 1: ¿Es posible extraer la cadena de diseminación de una información falsa en una red social?** A lo largo de este trabajo se ha mostrado cómo el uso de cadenas de búsqueda o *queries*, compuestas por palabras claves extraídas de un determinado *claim* permiten obtener todos los posts posibles relacionados. En los Capítulos 3 y 4, la creación de *queries* ha sido manual a partir de los *claims* desmentidos por los *fact-checkers*. Para el Capítulo 5, la generación ha sido automática a partir de estos enunciados. Las conclusiones de los experimentos indican que sí se puede extraer todo un conjunto de posts de la red social que representa su cadena de diseminación.

- **RQ 2: ¿Se puede generar una representación vectorial donde posts relacionados con la misma desinformación mantengan distancias cercanas?** De la conversación

extraída se muestra en el módulo del Capítulo 3 que se puede obtener conocimiento de las relaciones semánticas entre afirmaciones. Mediante el uso de *embeddings* semánticos y proyecciones a dos dimensiones, usando técnicas de reducción de dimensionalidad, es posible crear representaciones visuales donde se refleja la cercanía entre pares de *claims* relacionados semánticamente.

- **RQ 3: ¿Se pueden relacionar los posts de distintas falsedades en función a su cercanía semántica?** Es más difícil determinar las causas de por qué los *embeddings* se acercan o se alejan en la representación. Se han observado ciertos vínculos de cercanía entre diferentes posts de diversas desinformaciones en el Capítulo 3, pero estos vectores no desvelan las propiedades por las que estas aparecen más juntas. Responder a la anterior pregunta es más fácil porque todas las publicaciones de un *claim* aparecen dibujadas en grupos y se entiende que es por la relación a una misma desinformación, pero no se pueden entrever las dinámicas entre publicaciones de falsedades distintas, si bien se han valorado de forma experimental ciertos vínculos también en el Capítulo 3.

- **RQ 4: ¿Se pueden separar los posts relacionados con la información falsa en la conversación de aquellos no relacionados con ella?** La respuesta a las anteriores preguntas da luz verde a los siguientes experimentos con NLI por el potencial de los *embeddings* para relacionar textos entre sí. El segundo módulo en el Capítulo 4 pone a prueba esta capacidad en el contexto de las redes sociales para caracterizar la desinformación. Gracias al uso de modelos entrenados para inferencia del lenguaje, se ha comprobado que se puede separar la conversación referida a cada falsedad de aquella que no está vinculada a ella.

- **RQ 5: ¿Se pueden distinguir las publicaciones que propagan una información falsa de aquellas que la contradicen?** Los modelos de NLI no solo consiguen separar la conversación sobre desinformación de los posts no relacionados, sino que también logran discernir entre las publicaciones que reproducen la desinformación y aquellas que la contradicen, además de apartarlas del resto de descargas que no tienen que ver con la falsedad gracias al etiquetado de posts como *Entailment*, *Contradiction* y *Neutral* en base a las probabilidades generadas por estos modelos.

- **RQ 6. ¿Se puede extraer información de las publicaciones que propagan o contradicen la información falsa en función al número de interacciones de los posts?** Gracias al NLI, se ha descubierto que la mayoría de posts en la conversación sobre una desinformación dada no tienen reacciones (entendidas dentro del estudio de X como *reposts*, *likes*, respuestas y citas, a las que se añaden las repeticiones de exactamente el mismo contenido). Fundamenta la necesidad de esta disciplina para entender las conversaciones sobre la información falsa, para proseguir después con la generación de grafos, y da una respuesta afirmativa a esta pregunta porque se han podido analizar los posts cruzando los *insights* del NLI con los de las métricas de la red social.

  En este aspecto, se subrayan tres afirmaciones: por un lado, tanto las desinformaciones como sus contradicciones (entre las que están sus desmentidos) no solo se mueven en la viralidad, sino en la ausencia de ella a través de los círculos de seguidores donde se expresan sin tanto impacto; en segundo lugar, por esta mayoría de posts no virales, se entiende que los que sí lo son no lo deben solo a expresar o refutar tal falsedad sino a otras características en torno a la publicación en sí y a la autoría; finalmente, los contenidos no son necesariamente producto de interacciones con posts anteriores que sembrasen la falsedad como ramificaciones de una

gran cascada, abriendo la vía a que se ha consumido la desinformación de otra manera, bien por parte de otras publicaciones con o sin interacciones, o bien fuera de la red social.

- **RQ 7. ¿Difieren las proporciones entre los posts y usuarios que diseminan desinformación frente a aquellos que la contradicen dependiendo del número de interacciones?** No se observan notables diferencias en las dinámicas de los textos falsos en las plataformas sociales frente a los posts con enunciados opuestos. Predominan en el recuento del experimento del módulo de NLI las publicaciones con información falsa frente a la contraria a ella en significado, pero no se puede responder de forma positiva a esta pregunta porque estas publicaciones opuestas a tales falsedades también surgen de pequeña a gran escala, desde aquellas con las métricas a cero hasta las más virales. No obstante, sí es de destacar el número desigual de respuestas a los posts con *Entailment* respecto a aquellos con *Contradiction*.

- **RQ 8. ¿Se puede trazar el movimiento de los posts relacionados con una información falsa y los usuarios que la propagan de principio a fin?** Caracterizar el contenido en una red social en función a cómo se alinea con el *claim* y explicar sus rasgos en función a las interacciones demuestra el potencial del NLI y de las métricas en la plataforma para explotar el SNA junto a este filtro de inferencia. Los tres casos de estudio planteados en el Capítulo 5, cada uno en relación a un tipo distinto de información falsa, han dado lugar a tres grafos que aúnan todas las propiedades extraídas en los pasos anteriores, con los nodos representando los posts y las aristas encarnando las uniones para aquellos que sean derivados de otros anteriores. De esta forma, se ha conseguido desdeñar la evolución de la información falsa.

  Los grafos han representado en los tres casos cómo han discurrido tres desinformaciones en el tiempo y su diversidad de publicaciones y usuarios, desde los posts con ninguna interacción hasta aquellos con más impacto, y desde los usuarios con menos seguidores hasta aquellos con más comunidad. En cada ejemplo, el contenido falso ha circulado con y sin viralidad. Por tanto, a la octava pregunta la respuesta es 'sí'. En el Capítulo 5, esta tesis ha impreso la imagen del recorrido de las desinformaciones desde su inicio, antes de que tuvieran su publicación más viral y también después de esta, todo ello también en forma de otros posts con pocas interacciones o ninguna. Queda así configurado el ecosistema de la información falsa y de sus propiedades.

"¿Es posible monitorizar las conversaciones sobre informaciones falsas en una red social con PLN y SNA?" es la pregunta principal de esta investigación, de la cual han partido todas las subpreguntas. Con el camino completado, el flujo de trabajo propuesto con la combinación del PLN y el SNA ha resultado exitoso, y se puede contestar de forma afirmativa a esta cuestión. Se puede trazar la desinformación en las OSNs si se siguen los pasos de estas secciones: primero, la descarga de la conversación con *queries*, que puede ser automática mediante la asistencia del PLN; después, la división entre los posts que contienen informaciones falsas, su significado contrario y nada en relación a estas, y, finalmente, la generación de grafos para el mapeo de estos mensajes y de las cuentas de donde proceden. Se construye así una herramienta de monitorización al servicio de las organizaciones de *fact-checking* y de futuros avances en esta línea de investigación académica.

6.2. Discusión y trabajo futuro

Frente a los trabajos que exploran los avances individuales del SNA, el PLN y el ML, tanto de forma genérica como en la lucha contra la desinformación, esta tesis ha ido un paso más allá a partir de una nueva metodología de tres bloques: el filtrado semántico, la aplicación de NLI y la generación de grafos. Esta aportación permite el empleo flexible de los módulos, juntos o por separado, pero además la posibilidad de innovar en cada uno de ellos. Toda evolución es bienvenida para cada una de las partes en escenarios futuros sin necesidad de cambiar esta estructura para el seguimiento de las conversaciones sobre información falsa.

Respecto al módulo de descarga, aunque se ha diseñado para Twitter, se puede adaptar a otras redes sociales para captar toda la conversación sobre desinformación. Más allá de X, el informe del Instituto Reuters ya exponía a YouTube, Facebook, TikTok e Instagram y a las plataformas de mensajería WhatsApp y Telegram como espacios de consumo de noticias [268] antes de sus estudios más recientes [28, 27]. Aunque cada una de ellas tiene un formato distinto y, en consecuencia, una forma diversa de acceder a sus datos, los pasos de esta tesis sientan las bases para configurar este bloque en otras OSNs y desentrañar los otros ecosistemas de la información falsa.

Si bien se ha usado X para probar los módulos de esta investigación, otras redes de la misma naturaleza están ahora en el punto de mira. Aunque estas plataformas se transformen a lo largo del tiempo, los usuarios se organizarán en espacios en línea para seguir consumiendo contenidos y, por ello, la desinformación que también habita en estos. Ya se han abordado estudios de los traslados de usuarios de X a Mastodon, Bluesky y Threads [269], y, aunque plataformas como ya el antiguo Twitter muten en el futuro, se revela cómo los individuos tendrán alternativas para compartir y recibir información. Para estos ecosistemas, los bloques de esta tesis también tienen cabida.

Respecto a Mastodon, es una de las redes del llamado 'fediverso': alude al ecosistema de las *Decentralized Online Social Networks* (DOSNs) [270], plataformas descentralizadas cada una operativa con un servidor, propiedad y reglas diferentes, pero unidas por las interacciones entre usuarios de distintos servidores, como alternativa a las redes conocidas centralizadas donde el individuo ha perdido el poder y el control de unos mecanismos que provocan, entre otros aspectos, la exposición a la desinformación [271]. No obstante, pese a estas ventajas, la investigación ya ha desvelado flaquezas que invitan a mecanismos más fuertes de moderación [272]. Ello anima a adaptar y reinterpretar los módulos de NLI y SNA de este trabajo para los problemas de las informaciones falsas que también surjan en un futuro en estos espacios.

En el PLN, el constante progreso de los nuevos modelos del lenguaje puede servir a dos niveles: por una parte, dentro del módulo de descarga, los nuevos sistemas para la captura de temas y para la obtención de *keywords* de un enunciado permitirán optimizar las *queries* o toda forma de búsqueda refinada con el texto como *input*; por otra, dentro del filtrado semántico y de caracterización de la información falsa, ambos tipos de procesos podrán también mejorarse para etiquetar cada publicación en función a su alineación con el *claim*. Esto hará aún más verosímil el mapa de la desinformación a partir del grafo final, tanto por los mensajes descargados como por la mejor clasificación de estos.

No obstante, estas mejoras del PLN pueden no ser suficientes para los términos camuflados. Los *Transformers* capturan los significados del texto de los posts y el lenguaje natural de las OSNs, pero pueden propagarse entre círculos de forma encubierta, sustituyendo parte de las letras del mensaje por números u otros caracteres para que sean entendibles por el ojo humano pero no por

los buscadores de texto y, en última instancia, por estos modelos del lenguaje. Ya se muestran avances para generar y detectar estos términos [273] y para blindar los *Transformers* frente a ellos [274], y esto permitiría el trabajo futuro en las dos direcciones ya mencionadas: la generación automática de *queries* y los modelos de NLI.

Además, el texto a analizar puede aparecer dentro de las imágenes compartidas en las redes sociales. Se han mostrado avances en el tratamiento de texto dentro de los memes en las OSNs, superpuesto para estos casos en las imágenes con intención cómica o irónica, en campos como la detección de contenido misógino [275]. La conversión del texto reconocido a *embeddings* para combinarlo con la información de las imágenes procesadas puede ayudar en contextos donde la desinformación también se exprese con estos elementos textuales plasmados en una imagen. Esta línea de investigación invita, por tanto, a que el NLI no se limite al texto descargado sino a las otras fuentes de información visuales.

En esta combinación de información, por tanto, no todo gira alrededor del texto. Otra de las cuestiones a atajar en futuras investigaciones es el carácter multimodal de las publicaciones en OSNs [32]. Redes como TikTok, con más usuarios que no distinguen qué información es fiable [27], tienen la imagen y el vídeo como materias primas de difusión. Si bien la investigación ha puesto en el foco en el lenguaje natural, el tratamiento de la imagen, del vídeo y de los recursos en general que conviven con la fuente textual forma parte del estado de la cuestión en los métodos computacionales que unen todos estos elementos del mensaje [276, 277].

En esta línea, las investigaciones ya abordan las formas de procesamiento de todos los canales de información, los modos de fusionarlos y los *datasets* utilizados [277], pero la clasificación del contenido multimodal se enfrenta a los mismos problemas que el etiquetado de texto. Estos son la dependencia a los datos de entrenamiento para poder captar la desinformación y la hegemonía de enfoques totalmente automáticos para separar el contenido verdadero del falso [9, 42, 32]. Tal como se ha hecho con el NLI para el texto, esta tesis invita a repensar los procesos para acercar el procesado multimodal también al *fact-checking* semiautomático propuesto en estas secciones, donde también, por ejemplo, se tome una unidad de referencia como el *claim* en el lenguaje natural para evaluar también la coincidencia entre contenidos visuales y audiovisuales con la desinformación.

No solo se puede innovar con la multimodalidad en las partes acaparadas por el PLN, sino también en las del SNA. Como ya se ha mencionado, la información de las OSNs, como todo conjunto de datos, comparte las tres V [177] y, precisamente, estas no se conciben solo como texto. La V de variedad alude a los contenidos de estas plataformas como un amalgama de imagen, audio y texto en múltiples formatos. Esto afecta a las otras V (velocidad y volumen) y anima a reconfigurar la forma de procesar toda esta información en conjunto [174]. Este carácter multimodal encaja también en las cuatro dimensiones comunes del SNA como disciplina de datos: qué hallar, qué visualizar, hasta dónde abarcar y, sobre todo, qué información mezclar, la cual explícitamente hace referencia a la unión de todos los canales [37].

Esta tesis ha aportado en la lucha contra la desinformación, pero sus módulos son transversales a otras tareas. Como ya se ha apuntado, el análisis de sentimiento dentro del PLN ha constatado que las OSNs son un termómetro de la emoción en periodo de elecciones gubernamentales [238] tras obtener el grado más positivo o negativo del texto con la metodología VADER [239]. Este descubrimiento alude, para otro dominio, a los dos primeros módulos de descarga de posts y de variables adicionales de PLN en esta tesis, pero no al de generación de grafos. Esta investigación impulsa la necesidad de también combinar PLN y SNA con los tres módulos para monitorizar,

por ejemplo, la polarización política, cambiando el cometido de caracterizar los posts sobre desinformación por el de analizar el sentimiento.

Por eso, este trabajo es la entrada a la monitorización de todo fenómeno que pueda ser procesado de forma computacional en las OSNs y al NLI como disciplina para enriquecer los grafos, tanto en el caso antes mencionado como para cualquier otro. Otros escenarios donde estos módulos tienen cabida son: todo enfoque general con *Transformers* para detectar el discurso de odio [278, 279, 280] y seguir su recorrido; propuestas en torno al perfilado de autores, no solo para captar con los últimos avances en PLN los contenidos con un mismo creador [281], sino para añadir esa información a grafos, o el modelado de temas (*topic modeling*), aupado por *Transformers* como BERTopic [282], donde estas variables agregadas por el procesado del lenguaje se incorporen a los nodos y sus conexiones.

6.3. Conclusiones

Este trabajo cumple las metas indicadas al principio de esta tesis. Se pueden enumerar los siguientes objetivos alcanzados:

- La investigación muestra que **SNA y PLN son necesarios en la lucha contra la desinformación**. Los módulos de este trabajo han revelado cómo combinar ambas disciplinas, primera de las metas planteadas. Esto supone un avance en esta batalla y ha sido posible readaptando los puntos fuertes de ambos campos: por una parte, esta tesis ha replanteado las tareas de clasificación del PLN para ver la relación de los posts con enunciados falsos y detectar si su sentido es el mismo, el contrario o uno ajeno; por otra, ha redirigido el camino del SNA para que muestre la estructura completa de todos los mensajes y sus conexiones y no solo aquella de posts destacados.

- Esta investigación evidencia también **el potencial del filtrado semántico y la inferencia del lenguaje natural en el *fact-checking* semiautomático**. El replanteamiento de las tareas del PLN ha sido también una ruptura con los métodos computacionales tradicionales que se saltan el proceso de verificar y etiquetan contenidos como verdaderos o falsos solo en base a los patrones de un conjunto de desinformaciones distintas. Los experimentos expuestos han transformado el *claim matching* en *claim monitoring* porque, gracias al NLI, no solamente se ha alineado de forma aislada cualquier post con el *claim*, sino que se ha buscado esa relación del enunciado falso con el sentido de todos los contenidos descargados para caracterizar así toda la conversación de la desinformación y estructurarla después con el SNA.

- Gracias a los pasos realizados, se ha conseguido **trazar con SNA todo el ecosistema de usuarios y publicaciones en torno a una conversación sobre desinformación**, tercera meta de la tesis. Se demuestra así un cambio en la forma de concebir la difusión de falsedades en OSNs: la información falsa se mueve por una proporción considerable de posts con ninguna o pocas interacciones, y no solo por aquellas más virales en la red, que no representarían el fenómeno en su totalidad. Con independencia de su impacto, el mapa de la desinformación dibujado con NLI y SNA también tiene esta serie de posts, cada uno de ellos en un círculo de seguidores que, aunque no republiquen el contenido mediante el mensaje original, sí pueden haberlo consumido.

  Así, estos resultados formulan una alternativa a los trabajos que exploran solo la trayecto-

ria tradicional en cascada a partir del contenido más compartido en una plataforma [36, 7]. De tomar esa senda, la cascada excluiría la proporción considerable de posts poco o nada compartidos porque no forma parte de esas ramificaciones como producto de republicaciones. Pero esta tesis evidencia que las olas de desinformación también se componen de estos posts sin interacciones. En analogía con las pandemias, la infodemia no va a tener un único foco, sino varios con mayor o menor alcance y, aunque su origen fuera el mismo, dentro o fuera de una red social, no dibujan un mismo trazado y no tienen por qué manifestarse como una concatenación de una publicación anterior.

- Este trabajo también ha logrado **construir representaciones que incluyen la respuesta frente a la desinformación**, cuarto objetivo, y no solo las falsedades en sí dentro de la conversación. Por un lado, el NLI ha permitido etiquetar los contenidos contrarios al sentido del *claim* además de aquellos que expresaban lo mismo que este. Por otro, el SNA ha posibilitado, a través de la creación de grafos, que se aprecie cómo los posts con mensajes falsos y aquellos opuestos a ellos se intercambian a lo largo del tiempo. Ante la cantidad ingente de información falsa contra la que luchan los *fact-checkers* [22, 30], esta investigación permite optimizar la respuesta contra esta lacra, ya que el rastreo será también de los desmentidos que rebaten tales publicaciones nocivas.

- Esta investigación ha servido para **crear una herramienta de monitorización** que va de la mano de la práctica profesional de las organizaciones de *fact-checking*, y este era el último objetivo de la tesis. Ya se ha visto en el primer módulo que, aunque se aprecien las relaciones de cercanía semántica entre posts con temas en teoría similares, los *embeddings* no dan las causas de las relaciones textuales entre contenidos, en un esfuerzo por la IA explicativa de revertir este obstáculo [165]. El NLI no cambia esto pero sí restringe la tarea a vincular el *claim* con los posts que expresen lo mismo, en vez de con todos con los que haya relación semántica de cualquier tipo. Así, la decisión computacional queda subordinada al enunciado falso obtenido mediante los desmentidos del *fact-checking* y, por ello, al rigor periodístico para llegar a ellos, ligado al 'Código de Principios' de la IFCN para los *fact-checkers* miembros [30]. Toda revisión humana después de este proceso de IA ya velaría por el lazo con el sentido del *claim* y no por descifrar por qué unos patrones desconocidos etiquetan una información como verdadera o falsa a partir de los datos de otras falsedades que no tienen que ver.

Esta tesis ha demostrado que es posible innovar en las técnicas computacionales en la lucha contra la desinformación. No por sus métodos planteados en PLN con el NLI o en SNA con la generación de grafos, que ya existían, sino por cuestionar los que ya había para replantear cuál debe ser su rol de forma combinada. Este trabajo se ha salido de la rueda de los procedimientos automáticos en la IA gracias al *fact-checking* semiautomático, reorganizando el PLN y el SNA sin minar las fortalezas de cada uno de estos campos. No entra en conflicto con las futuras investigaciones enfocadas en el estudio individual de estas dos áreas por separado, sino todo lo contrario: los avances técnicos de los grandes modelos de PLN aún por venir y también los del SNA se podrán integrar en la estructura de módulos planteada en estas secciones y contribuir a esta acción coordinada.

Así, los esfuerzos de la IA contra la información falsa ya no tienen por qué ir separados con los del *fact-checking*. Las falsedades vertidas no empiezan en los *datasets* de entrenamiento con otras desinformaciones previas, sino en el día a día, en las OSNs. Como se ha planteado en esta tesis, la lucha contra esta lacra no tiene que basarse únicamente en los *outputs* de los modelos

computacionales, sino en reenfocar tanto el uso de los algoritmos en esta disciplina como sus resultados. Etiquetar contenidos como verdaderos o falsos automáticamente en entornos reales puede ser un enfrentamiento de la computación contra la desinformación, pero estos pasos de la tesis han intentado que la batalla también sea del *fact-checking* en sí, y no solo de la rama de la informática como un elemento separado.

El *fact-checker* ha sido el centro a la hora de configurar los módulos. Sus desmentidos son el resultado de toda su práctica profesional, y son los que sirven como escudo para todo el contenido falso recién surgido en las plataformas sociales. Estos desmentidos no se han tratado de manera secundaria a través de los modelos automáticos de predicción, sino que han sido el punto de partida de la tesis. El NLI en estos módulos los ha potenciado hasta ver qué relación tienen las falsedades que señalan con todos los posts de la conversación en una red social, y el SNA los ha estructurado para permitir su seguimiento de principio a fin. Una línea de investigación con dos disciplinas computacionales de la mano, pero también con la fusión de dos aliadas, el *fact-checking* y la IA, que no tienen por qué ir separadas contra este problema.

Bibliografía

[1] C. Wardle, H. Derakhshan, Information disorder: Toward an interdisciplinary framework for research and policymaking, volume 27, Council of Europe Strasbourg, 2017.

[2] D. Jurafsky, J. H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, third edition draft ed., 2024. URL: https://web.stanford.edu/~jurafsky/slp3/.

[3] K. Shu, H. R. Bernard, H. Liu, Studying fake news via network analysis: detection and mitigation, Emerging research challenges and opportunities in computational social network analysis and mining (2019) 43–65.

[4] S. P. Borgatti, F. Agneessens, J. C. Johnson, M. G. Everett, Analyzing social networks, SAGE publications Ltd, 2024.

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

[6] S. S. Singh, S. Muhuri, S. Mishra, D. Srivastava, H. K. Shakya, N. Kumar, Social network analysis: A survey on process, tools, and application, ACM Computing Surveys 56 (2024) 1–39.

[7] S. Goel, A. Anderson, J. Hofman, D. J. Watts, The structural virality of online diffusion, Management Science 62 (2016) 180–196.

[8] S. S. Singh, V. Srivastava, A. Kumar, S. Tiwari, D. Singh, H.-N. Lee, Social network analysis: a survey on measure, structure, language information analysis, privacy, and applications, ACM Transactions on Asian and Low-Resource Language Information Processing 22 (2023) 1–47.

[9] A. Martín, J. Huertas-Tato, Á. Huertas-García, G. Villar-Rodríguez, D. Camacho, Facter-check: Semi-automated fact-checking through semantic similarity and natural language inference, Knowledge-Based Systems (2022) 109265.

[10] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. R. Bowman, N. A. Smith, Annotation artifacts in natural language inference data, arXiv preprint arXiv:1803.02324 (2018).

[11] B. MacCartney, Natural language inference, Stanford University, 2009.

[12] S. Raponi, Z. Khalifa, G. Oligeri, R. Di Pietro, Fake news propagation: A review of epidemic models, datasets, and insights, ACM Transactions on the Web (TWEB) 16 (2022) 1–34.

[13] J. Posetti, A. Matthews, A short guide to the history of fake news and disinformation, International Center for Journalists 7 (2018) 2018–07.

[14] R. Kouzy, J. Abi Jaoude, A. Kraitem, M. B. El Alam, B. Karam, E. Adib, J. Zarka, C. Traboulsi, E. W. Akl, K. Baddour, Coronavirus goes viral: quantifying the covid-19 misinformation epidemic on twitter, Cureus 12 (2020).

[15] R. Salaverría, N. Buslón, F. López-Pan, B. León, I. López-Goñi, M.-C. Erviti, Desinformación en tiempos de pandemia: tipología de los bulos sobre la covid-19, Profesional de la Información 29 (2020).

[16] E. M. Said-Hung, M. A. Merino-Arribas, J. Martínez-Torres, Evolución del debate académico en la web of science y scopus sobre unfaking news (2014-2019), Estudios Sobre el Mensaje Periodístico 27 (2021) 961.

[17] A. M. Guess, B. A. Lyons, Misinformation, disinformation, and online propaganda, Social media and democracy: The state of the field, prospects for reform 10 (2020).

[18] N. A. Karlova, K. E. Fisher, A social diffusion model of misinformation and disinformation for understanding human information behaviour, Ïnformation Research"(2013).

[19] C. Ireton, J. Posetti, Journalism, fake news & disinformation: handbook for journalism education and training, Unesco Publishing, 2018.

[20] S. Wang, F. Su, L. Ye, Y. Jing, Disinformation: A bibliometric review, International journal of environmental research and public health 19 (2022) 16849.

[21] F. KaabOmeir, S. Khademizadeh, R. Seifadini, S. O. Balani, M. Khazaneha, Overview of misinformation and disinformation research from 1971 to 2022, Journal of Scientometric Research 13 (2024) 430–447.

[22] M. Choraś, K. Demestichas, A. Giełczyk, Á. Herrero, P. Ksieniewicz, K. Remoundou, D. Urda, M. Woźniak, Advanced machine learning techniques for fake news (online disinformation) detection: A systematic mapping study, Applied Soft Computing 101 (2021) 107050.

[23] D. Freelon, C. Wells, Disinformation as political communication, 2020.

[24] S. Altay, M. Berriche, H. Heuer, J. Farkas, S. Rathje, A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field, Harvard Kennedy School Misinformation Review 4 (2023) 1–34.

[25] M. S. Deiner, C. Fathy, J. Kim, K. Niemeyer, D. Ramirez, S. F. Ackley, F. Liu, T. M. Lietman, T. C. Porco, Facebook and twitter vaccine sentiment in response to measles outbreaks, Health informatics journal 25 (2019) 1116–1132.

[26] M. Imran, C. Castillo, F. Diaz, S. Vieweg, Processing social media messages in mass emergency: A survey, ACM Computing Surveys (CSUR) 47 (2015) 1–38.

[27] N. Newman, R. Fletcher, C. T. Robertson, A. R. Arguedas, R. K. Nielsen, Digital News Report 2024, Technical Report, RISJ: Reuters Institute for the Study of Journalism, 2024.

[28] N. Newman, R. Fletcher, K. Eddy, C. T. Robertson, R. K. Nielsen, Digital News Report 2023, Technical Report, RISJ: Reuters Institute for the Study of Journalism, 2023.

[29] E. Porter, T. J. Wood, The global effectiveness of fact-checking: Evidence from simultaneous experiments in argentina, nigeria, south africa, and the united kingdom, Proceedings of the National Academy of Sciences 118 (2021) e2104235118.

[30] J. S. Brennen, F. M. Simon, P. N. Howard, R. K. Nielsen, Types, sources, and claims of COVID-19 misinformation, Ph.D. thesis, University of Oxford, 2020.

[31] S. Evanega, M. Lynas, J. Adams, K. Smolenyak, C. G. Insights, Coronavirus misinformation: quantifying sources and themes in the covid-19 infodemic, JMIR Preprints 19 (2020) 2020.

[32] A. Montoro-Montarroso, J. Cantón-Correa, P. Rosso, B. Chulvi, Á. Panizo-Lledot, J. Huertas-Tato, B. Calvo-Figueras, M. J. Rementeria, J. Gómez-Romero, Fighting disinformation with artificial intelligence: fundamentals, advances and challenges, Profesional de la información 32 (2023).

[33] A. Tretiakov, A. Martín, D. Camacho, Detection of false information in spanish using machine learning techniques, in: International Conference on Intelligent Data Engineering and Automated Learning, Springer, 2022, pp. 42–53.

[34] J. Huertas-Tato, A. Martín, D. Camacho, Silt: Efficient transformer training for inter-lingual inference, Expert Systems with Applications 200 (2022) 116923.

[35] A. Huertas-García, J. Huertas-Tato, A. Martín, D. Camacho, Countering misinformation through semantic-aware multilingual models, in: International conference on intelligent data engineering and automated learning, Springer, 2021, pp. 312–323.

[36] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, science 359 (2018) 1146–1151.

[37] D. Camacho, Á. Panizo Lledot, G. Bello Orgaz, A. Gonzalez Pardo, E. Cambria, The four dimensions of social network analysis: An overview of research methods, applications, and software tools, Information Fusion 63 (2020) 88–120.

[38] A. Bodaghi, J. Oliveira, The theater of fake news spreading, who plays which role? a study on real graphs of spreading on twitter, Expert Systems with Applications 189 (2022) 116110.

[39] G. Bello-Orgaz, J. Hernandez-Castro, D. Camacho, Detecting discussion communities on vaccination in twitter, Future Generation Computer Systems 66 (2017) 125–136.

[40] W. Ahmed, F. L. Seguí, J. Vidal-Alaball, M. S. Katz, et al., Covid-19 and the "film your hospital" conspiracy theory: social network analysis of twitter data, Journal of medical Internet research 22 (2020) e22374.

[41] W. Ahmed, J. Vidal-Alaball, J. Downing, F. L. Seguí, et al., Covid-19 and the 5g conspiracy theory: social network analysis of twitter data, Journal of medical internet research 22 (2020) e19458.

[42] X. Zeng, A. S. Abumansour, A. Zubiaga, Automated fact-checking: A survey, Language and Linguistics Compass 15 (2021) e12438.

[43] E. Lazarski, M. Al-Khassaweneh, C. Howard, Using nlp for fact checking: A survey, Designs 5 (2021) 42.

[44] L. Graves, M. A. Amazeen, Fact-checking as idea and practice in journalism, 2019. URL: https://oxfordre.com/communication/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-808. doi:doi: 10.1093/acrefore/9780190228613.013.808.

[45] P. Dhiman, A. Kaur, C. Iwendi, S. K. Mohan, A scientometric analysis of deep learning approaches for detecting fake news, Electronics 12 (2023) 948.

[46] X. Zhou, R. Zafarani, A survey of fake news: Fundamental theories, detection methods, and opportunities, ACM Computing Surveys (CSUR) 53 (2020) 1–40.

[47] E. Kapantai, A. Christopoulou, C. Berberidis, V. Peristeras, A systematic literature review on disinformation: Toward a unified taxonomical framework, New media & society 23 (2021) 1301–1326.

[48] J. Pamment, H. Nothhaft, A. Fjällhed, Countering information influence activities: The state of the art, Technical Report, MSB, 2018.

[49] J. Valant, Online consumer reviews: The case of misleading or fake reviews. european parliamentary research service, 2015.

[50] D. Tambini, Fake news: public policy responses, Media Policy Brief 20 (2017).

[51] S. Zannettou, M. Sirivianos, J. Blackburn, N. Kourtellis, The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans, Journal of Data and Information Quality (JDIQ) 11 (2019) 1–37.

[52] S. B. Parikh, P. K. Atrey, Media-rich fake news detection: A survey, in: 2018 IEEE conference on multimedia information processing and retrieval (MIPR), IEEE, 2018, pp. 436–441.

[53] A. Bondielli, F. Marcelloni, A survey on fake news and rumour detection techniques, Information sciences 497 (2019) 38–55.

[54] M. F. Mridha, A. J. Keya, M. A. Hamid, M. M. Monowar, M. S. Rahman, A comprehensive review on fake news detection with deep learning, IEEE access 9 (2021) 156151–156170.

[55] H. Allcott, M. Gentzkow, Social media and fake news in the 2016 election, Journal of economic perspectives 31 (2017) 211–236.

[56] I. Khaldarova, M. Pantti, Fake news: The narrative battle over the ukrainian conflict, in: The Future of Journalism: Risks, Threats and Opportunities, Routledge, 2020, pp. 228–238.

[57] J. Jerit, Y. Zhao, Political misinformation, Annual Review of Political Science 23 (2020) 77–94.

[58] J. A. Tucker, A. Guess, P. Barberá, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, B. Nyhan, Social media, political polarization, and political disinformation: A review of the scientific literature, Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018) (2018).

[59] K. M. d. Treen, H. T. Williams, S. J. O'Neill, Online misinformation about climate change, Wiley Interdisciplinary Reviews: Climate Change 11 (2020) e665.

[60] R. E. Dunlap, Organized Climate Change Denial, Oxford Handbook of Climate Change and Society/Oxford University Press, 2011.

[61] S. Lewandowsky, Climate change disinformation and how to combat it, Annual Review of Public Health 42 (2021) 1–21.

[62] J. Cook, Understanding and countering misinformation about climate change, Research anthology on environmental and societal impacts of climate change (2022) 1633–1658.

[63] B. Swire-Thompson, D. Lazer, et al., Public health and online misinformation: challenges and recommendations, Annu Rev Public Health 41 (2020) 433–451.

[64] Y. Wang, M. McKee, A. Torbica, D. Stuckler, Systematic literature review on the spread of health-related misinformation on social media, Social science & medicine 240 (2019) 112552.

[65] I. J. B. Do Nascimento, A. B. Pizarro, J. M. Almeida, N. Azzopardi-Muscat, M. A. Gonçalves, M. Björklund, D. Novillo-Ortiz, Infodemics and health misinformation: a systematic review of reviews, Bulletin of the World Health Organization 100 (2022) 544.

[66] J. Szakacs, E. Bognar, The impact of disinformation campaigns about migrants and minority groups in the eu, Policy Department for External Relations Directorate General for External Policies of the Union. https://www. europarl. europa. eu/meetdocs/2014_2019/plmrep/COMMITTEES/INGE/DV/2021/07-12/IDADisinformation_migrant_minorities_EN. pdf (2021).

[67] J. Gamir-Ríos, R. Tarullo, M. Ibáñez-Cuquerella, et al., Multimodal disinformation about otherness on the internet. the spread of racist, xenophobic and islamophobic fake news in 2020, Anàlisi (2021) 49–64.

[68] F. B. Keller, D. Schoch, S. Stier, J. Yang, Political astroturfing on twitter: How to coordinate a disinformation campaign, Political communication 37 (2020) 256–280.

[69] A. Ghenai, Y. Mejova, Catching zika fever: Application of crowdsourcing and machine learning for tracking health misinformation on twitter, in: 2017 IEEE International Conference on Healthcare Informatics (ICHI), IEEE Computer Society, 2017, pp. 518–518.

[70] G. K. Shahi, A. Dirkson, T. A. Majchrzak, An exploratory study of covid-19 misinformation on twitter, Online social networks and media 22 (2021) 100104.

[71] S. O. Oyeyemi, E. Gabarron, R. Wynn, Ebola, twitter, and misinformation: a dangerous combination?, Bmj 349 (2014).

[72] K. Sharma, S. Seo, C. Meng, S. Rambhatla, Y. Liu, Covid-19 on social media: Analyzing misinformation in twitter conversations, arXiv preprint arXiv:2003.12309 (2020).

[73] L. Singh, S. Bansal, L. Bode, C. Budak, G. Chi, K. Kawintiranon, C. Padden, R. Vanarsdall, E. Vraga, Y. Wang, A first look at covid-19 information and misinformation sharing on twitter, arXiv preprint arXiv:2003.13907 (2020).

[74] T. Buchanan, V. Benson, Spreading disinformation on facebook: Do trust in message source, risk propensity, or personality affect the organic reach of "fake news"?, Social media+ society 5 (2019) 2056305119888654.

[75] A. Barfar, Cognitive and affective responses to political disinformation in facebook, Computers in Human Behavior 101 (2019) 173–179.

[76] F. Zollo, W. Quattrociocchi, Misinformation spreading on facebook, Complex spreading phenomena in social systems: Influence and contagion in real-world social networks (2018) 177–196.

[77] S. B. Johnson, M. Parsons, T. Dorff, M. S. Moran, J. H. Ward, S. A. Cohen, W. Akerley, J. Bauman, J. Hubbard, D. E. Spratt, et al., Cancer misinformation and harmful information on facebook and other social media: a brief report, JNCI: Journal of the National Cancer Institute 114 (2022) 1036–1039.

[78] R. Recuero, F. B. Soares, O. Vinhas, T. Volcan, L. R. G. Hüttner, V. Silva, Bolsonaro and the far right: How disinformation about covid-19 circulates on facebook in brazil, International Journal of Communication 16 (2022) 24.

[79] P. M. Massey, M. D. Kearney, M. K. Hauer, P. Selvan, E. Koku, A. E. Leader, Dimensions of misinformation about the hpv vaccine on instagram: Content and network analysis of social media characteristics, Journal of medical Internet research 22 (2020) e21451.

[80] E. K. Quinn, S. S. Fazel, C. E. Peters, The instagram infodemic: cobranding of conspiracy theories, coronavirus disease 2019 and authority-questioning beliefs, Cyberpsychology, Behavior, and Social Networking 24 (2021) 573–577.

[81] P. Mena, D. Barbe, S. Chan-Olmsted, Misinformation on instagram: The impact of trusted endorsements on message credibility, Social Media+ Society 6 (2020) 2056305120935102.

[82] M. Palacios López, F. Bonete, R. Gelado Marcos, et al., New agents of mass disinformation. analysis of the publications produced by spanish political influencers on instagram, Estudios sobre el Mensaje Periodístico (2023).

[83] A. A. Birkun, Misinformation on first aid for seizures communicated through the fastest growing social media platform: A cross-sectional study of tiktok content, Epilepsy & Behavior 161 (2024) 110116.

[84] F. Sharevski, J. V. Loop, P. Jachim, A. Devine, E. Pieroni, Abortion misinformation on tiktok: Rampant content, lax moderation, and vivid user experiences, arXiv preprint arXiv:2301.05128 (2023).

[85] N. Alonso-López, P. Sidorenko-Bautista, F. Giacomelli, et al., Beyond challenges and viral dance moves: Tiktok as a vehicle for disinformation and fact-checking in spain, portugal, brazil, and the usa, Anàlisi (2021) 65–84.

[86] H. O.-Y. Li, A. Bailey, D. Huynh, J. Chan, Youtube as a source of information on covid-19: a pandemic of misinformation?, BMJ global health 5 (2020) e002604.

[87] H. O.-Y. Li, E. Pastukhova, O. Brandts-Longtin, M. G. Tan, M. G. Kirchhof, Youtube as a source of misinformation on covid-19 vaccination: a systematic analysis, BMJ global health 7 (2022) e008334.

[88] G. Donzelli, G. Palomba, I. Federigi, F. Aquino, L. Cioni, M. Verani, A. Carducci, P. Lopalco, Misinformation on vaccination: A quantitative analysis of youtube videos, Human vaccines & immunotherapeutics 14 (2018) 1654–1659.

[89] L. Tang, K. Fujimoto, M. Amith, R. Cunningham, R. A. Costantini, F. York, G. Xiong, J. A. Boom, C. Tao, "down the rabbit hole" of vaccine misinformation on youtube: Network exposure study, Journal of Medical Internet Research 23 (2021) e23262.

[90] S. Loeb, S. Sengupta, M. Butaney, J. N. Macaluso Jr, S. W. Czarniecki, R. Robbins, R. S. Braithwaite, L. Gao, N. Byrne, D. Walter, et al., Dissemination of misinformative and biased information about prostate cancer on youtube, European urology 75 (2019) 564–567.

[91] M. N. Hussain, S. Tokdemir, N. Agarwal, S. Al-Khateeb, Analyzing disinformation and crowd manipulation tactics on youtube, in: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2018, pp. 1092–1095.

[92] E. Hussein, P. Juneja, T. Mitra, Measuring misinformation in video search platforms: An audit study on youtube, Proceedings of the ACM on Human-Computer Interaction 4 (2020) 1–27.

[93] P. Herrero-Diz, J. Conde-Jiménez, S. Reyes de Cózar, Teens' motivations to spread fake news on whatsapp, Social Media+ Society 6 (2020) 2056305120942879.

[94] C. Machado, B. Kira, V. Narayanan, B. Kollanyi, P. Howard, A study of misinformation in whatsapp groups with a focus on the brazilian presidential elections., in: Companion proceedings of the 2019 World Wide Web conference, 2019, pp. 1013–1019.

[95] A. Herasimenka, J. Bright, A. Knuutila, P. N. Howard, Misinformation and professional news on largely unmoderated platforms: the case of telegram, Journal of Information Technology & Politics 20 (2023) 198–212.

[96] L. H. X. Ng, J. Y. Loke, Analyzing public opinion and misinformation in a covid-19 telegram group chat, IEEE Internet Computing 25 (2020) 84–91.

[97] L. Graves, The rise of fact-checking sites in europe, Reuters Institute for the Study of Journalism (2016).

[98] S. Cohen, C. Li, J. Yang, C. Yu, Computational journalism: A call to arms to database researchers, in: 5th Biennial Conference on Innovative Data Systems Research, CIDR, 2011, pp. 148–151.

[99] A. Vlachos, S. Riedel, Fact checking: Task definition and dataset construction, in: Proceedings of the ACL 2014 workshop on language technologies and computational social science, 2014, pp. 18–22.

[100] LatamChequea, Información chequeada sobre el coronavirus, 2020. URL: `https://chequeado.com/latamcoronavirus/`.

[101] L. Quintana Pujalte, M. F. Pannunzio, Fact-checking en latinoamérica. tipología de contenidos virales desmentidos durante la pandemia del coronavirus, Revista de Ciencias de la Comunicación e Información 26 (2021) 27–46.

[102] J. M. Sánchez-Duarte, R. M. Rosa, Infodemia y covid-19. evolución y viralización de informaciones falsas en españa, Revista española de comunicación en salud (2020) 31–41.

[103] M. Bulo, # ukrainefacts: una base de datos mundial y colaborativa para luchar contra la desinformación, Maldita. es 26 (2022).

[104] W. Y. Wang, liar, liar pants on fire: A new benchmark dataset for fake news detection, arXiv preprint arXiv:1705.00648 (2017).

[105] K. Chowdhary, K. Chowdhary, Natural language processing, Fundamentals of artificial intelligence (2020) 603–649.

[106] D. Khurana, A. Koli, K. Khatter, S. Singh, Natural language processing: state of the art, current trends and challenges, Multimedia tools and applications 82 (2023) 3713–3744.

[107] E. D. Liddy, Enhanced text retrieval using natural language processing, Bulletin of the American Society for Information Science and Technology 24 (1998) 14–16.

[108] S. Feldman, Nlp meets the jabberwocky: Natural language processing in information retrieval, ONLINE-WESTON THEN WILTON- 23 (1999) 62–73.

[109] N. Patwardhan, S. Marrone, C. Sansone, Transformers in the real world: A survey on nlp applications, Information 14 (2023) 242.

[110] R. E. Lopez-Martinez, G. Sierra, Natural language processing, 2000-2019—a bibliometric study, Journal of Scientometric Research 9 (2020) 310–318.

[111] A. Sandu, L.-A. Cotfas, A. Stănescu, C. Delcea, A bibliometric analysis of text mining: Exploring the use of natural language processing in social media research, Applied Sciences 14 (2024) 3144.

[112] T. P. Nagarhalli, V. Vaze, N. Rana, Impact of machine learning in natural language processing: A review, in: 2021 third international conference on intelligent communication technologies and virtual mobile networks (ICICV), IEEE, 2021, pp. 1529–1534.

[113] M. A. El Mrabet, K. El Makkaoui, A. Faize, Supervised machine learning: a survey, in: 2021 4th International conference on advanced communication technologies and networking (CommNet), IEEE, 2021, pp. 1–10.

[114] A. Singh, N. Thakur, A. Sharma, A review of supervised machine learning algorithms, in: 2016 3rd international conference on computing for sustainable global development (INDIACom), Ieee, 2016, pp. 1310–1315.

[115] R. Oshikawa, J. Qian, W. Y. Wang, A survey on natural language processing for fake news detection, arXiv preprint arXiv:1811.00770 (2018).

[116] D. Rohera, H. Shethna, K. Patel, U. Thakker, S. Tanwar, R. Gupta, W.-C. Hong, R. Sharma, A taxonomy of fake news classification techniques: Survey and implementation aspects, IEEE Access 10 (2022) 30367–30394.

[117] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, R. Procter, Detection and resolution of rumours in social media: A survey, ACM Computing Surveys (CSUR) 51 (2018) 1–36.

[118] J. Alghamdi, S. Luo, Y. Lin, A comprehensive survey on machine learning approaches for fake news detection, Multimedia Tools and Applications 83 (2024) 51009–51067.

[119] H. F. Villela, F. Corrêa, J. S. d. A. N. Ribeiro, A. Rabelo, D. B. F. Carvalho, Fake news detection: a systematic literature review of machine learning algorithms and datasets, Journal on Interactive Systems 14 (2023) 47–58.

[120] A. Tabassum, R. R. Patil, A survey on text pre-processing & feature extraction techniques in natural language processing, International Research Journal of Engineering and Technology (IRJET) 7 (2020) 4864–4867.

[121] Z. Harris, Distributional structure, 1954.

[122] K. Sparck Jones, A statistical interpretation of term specificity and its application in retrieval, Journal of documentation 28 (1972) 11–21.

[123] Y. Bengio, P. Vincent, J.-F. Paiement, O. Delalleau, M. Ouimet, N. Le Roux, Spectral clustering and kernel PCA are learning eigenfunctions, volume 1239, Citeseer, 2003.

[124] Y. Bengio, P. Simard, P. Frasconi, Learning long-term dependencies with gradient descent is difficult, IEEE transactions on neural networks 5 (1994) 157–166.

[125] S. Hochreiter, Long short-term memory, Neural Computation MIT-Press (1997).

[126] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder–decoder for statistical machine translation, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1724–1734. URL: https://aclanthology.org/D14-1179/. doi:doi: 10.3115/v1/D14-1179.

[127] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[128] S. Wang, W. Zhou, C. Jiang, A survey of word embeddings based on deep learning, Computing 102 (2020) 717–740.

[129] K. Jing, J. Xu, A survey on neural network language models, arXiv preprint arXiv:1906.03591 (2019).

[130] S. Bickel, P. Haider, T. Scheffer, Predicting sentences using n-gram language models, in: Proceedings of human language technology conference and conference on empirical methods in natural language processing, 2005, pp. 193–200.

[131] D. Jurafsky, J. H. Martin, Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition, Pearson Prentice Hall, Upper Saddle River, N.J., 2009. URL: http://www.amazon.com/Speech-Language-Processing-2nd-Edition/dp/0131873210/ref=pd_bxgy_b_img_y.

[132] C. Osgood, The measurement of meaning, Urban III University of Illinois (1957).

[133] S. Dierk, The smart retrieval system: Experiments in automatic document processing—gerard salton, ed.(englewood cliffs, nj: Prentice-hall, 1971, 556 pp.,), IEEE Transactions on Professional Communication (1972) 17–17.

[134] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems 26 (2013).

[135] J. Firth, A synopsis of linguistic theory 1930-1955, Studies in Linguistic Analysis, Special Volume/Blackwell (1957).

[136] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, Journal of machine learning research 12 (2011) 2493–2537.

[137] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the association for computational linguistics 5 (2017) 135–146.

[138] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[139] J. Wang, Y. Dong, Measurement of text similarity: a survey, Information 11 (2020) 421.

[140] M. Brunila, J. LaViolette, What company do words keep? revisiting the distributional semantics of jr firth & zellig harris, arXiv preprint arXiv:2205.07750 (2022).

[141] G. Tucudean, M. Bucos, B. Dragulescu, C. D. Caleanu, Natural language processing with transformers: a review, PeerJ Computer Science 10 (2024) e2222.

[142] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[143] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, 2018. URL: `https://arxiv.org/abs/1802.05365`. `arXiv:1802.05365`.

[144] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[145] G. Lample, A. Conneau, Cross-lingual language model pretraining, arXiv preprint arXiv:1901.07291 (2019).

[146] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2019).

[147] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, arXiv preprint arXiv:1904.09675 (2019).

[148] W. Yang, Y. Xie, A. Lin, X. Li, L. Tan, K. Xiong, M. Li, J. Lin, End-to-end open-domain question answering with bertserini, arXiv preprint arXiv:1902.01718 (2019).

[149] U. Naseem, I. Razzak, K. Musial, M. Imran, Transformer based deep intelligent contextual embedding for twitter sentiment analysis, Future Generation Computer Systems 113 (2020) 58–69.

[150] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, M. Shah, Transformers in vision: A survey, ACM computing surveys (CSUR) 54 (2022) 1–41.

[151] T. Lin, Y. Wang, X. Liu, X. Qiu, A survey of transformers, AI open 3 (2022) 111–132.

[152] A. Huertas-García, J. Huertas-Tato, A. Martín, D. Camacho, Civic-upm at checkthat! 2021: Integration of transformers in misinformation detection and topic classification., in: CLEF (Working Notes), 2021, pp. 520–530.

[153] R. Anggrainingsih, G. M. Hassan, A. Datta, Transformer-based models for combating rumours on microblogging platforms: a review, Artificial Intelligence Review 57 (2024) 1–69.

[154] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, S. Yang, P. W. Eklund, T. Huynh-The, T. T. Nguyen, Q.-V. Pham, I. Razzak, E. B. Hsu, Artificial intelligence in the battle against coronavirus (covid-19): a survey and future research directions, arXiv preprint arXiv:2008.07343 (2020).

[155] C. Shorten, T. M. Khoshgoftaar, B. Furht, Deep learning applications for covid-19, Journal of big Data 8 (2021) 1–54.

[156] A. Wani, I. Joshi, S. Khandve, V. Wagh, R. Joshi, Evaluating deep learning approaches for covid19 fake news detection, in: Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1, Springer, 2021, pp. 153–163.

[157] I. Dagan, O. Glickman, B. Magnini, The pascal recognising textual entailment challenge, in: Machine learning challenges workshop, Springer, 2005, pp. 177–190.

[158] S. Storks, Q. Gao, J. Y. Chai, Recent advances in natural language inference: A survey of benchmarks, resources, and approaches, arXiv preprint arXiv:1904.01172 (2019).

[159] S. R. Bowman, G. Angeli, C. Potts, C. D. Manning, A large annotated corpus for learning natural language inference, arXiv preprint arXiv:1508.05326 (2015).

[160] A. Williams, N. Nangia, S. Bowman, A broad-coverage challenge corpus for sentence understanding through inference, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1112–1122. doi:doi: 10.18653/v1/N18-1101.

[161] A. Conneau, G. Lample, R. Rinott, A. Williams, S. R. Bowman, H. Schwenk, V. Stoyanov, Xnli: Evaluating cross-lingual sentence representations, arXiv preprint arXiv:1809.05053 (2018).

[162] G. Ruffo, A. Semeraro, A. Giachanou, P. Rosso, Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language, Computer science review 47 (2023) 100531.

[163] G. Demartini, S. Mizzaro, D. Spina, Human-in-the-loop artificial intelligence for fighting online misinformation: Challenges and opportunities., IEEE Data Eng. Bull. 43 (2020) 65–74.

[164] Z. Guo, M. Schlichtkrull, A. Vlachos, A survey on automated fact-checking, Transactions of the Association for Computational Linguistics 10 (2022) 178–206.

[165] N. Kotonya, F. Toni, Explainable automated fact-checking: A survey, arXiv preprint arXiv:2011.03870 (2020).

[166] S. Raza, C. Ding, Fake news detection based on news content and social contexts: a transformer-based approach, International Journal of Data Science and Analytics 13 (2022) 335–362.

[167] S. Gundapu, R. Mamidi, Transformer based automatic covid-19 fake news detection system, arXiv preprint arXiv:2101.00180 (2021).

[168] K.-C. Yang, T. Niven, H.-Y. Kao, Fake news detection as natural language inference, arXiv preprint arXiv:1907.07347 (2019).

[169] F. Sadeghi, A. J. Bidgoly, H. Amirkhani, Fake news detection on social media using a natural language inference approach, Multimedia Tools and Applications 81 (2022) 33801–33821.

[170] A. Shah, H. Shah, V. Bafna, C. Khandor, S. Nair, Veritas-nli: Validation and extraction of reliable information through automated scraping and natural language inference, arXiv preprint arXiv:2410.09455 (2024).

[171] M. Arana-Catania, E. Kochkina, A. Zubiaga, M. Liakata, R. Procter, Y. He, Natural language inference with self-attention for veracity assessment of pandemic claims, arXiv preprint arXiv:2205.02596 (2022).

[172] U. Can, B. Alatas, A new direction in social network analysis: Online social network analysis problems and applications, Physica A: Statistical Mechanics and its Applications 535 (2019) 122372.

[173] J. Scott, What is social network analysis?, Bloomsbury Academic, 2012.

[174] D. Knoke, S. Yang, Social network analysis, SAGE publications, 2019.

[175] H. Esfahani, K. Tavasoli, A. Jabbarzadeh, Big data and social media: A scientometrics analysis, International Journal of Data and Network Science 3 (2019) 145–164.

[176] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, Y. Liu, Combating fake news: A survey on identification and mitigation techniques, ACM Transactions on Intelligent Systems and Technology (TIST) 10 (2019) 1–42.

[177] D. Laney, et al., 3d data management: Controlling data volume, velocity and variety, META group research note 6 (2001) 1.

[178] A. Panizo LLedot, J. Torregrosa, G. Bello Orgaz, J. Thorburn, D. Camacho, Describing alt-right communities and their discourse on twitter during the 2018 us mid-term elections, in: International conference on complex networks and their applications, Springer, 2019, pp. 427–439.

[179] S. Tabassum, F. S. Pereira, S. Fernandes, J. Gama, Social network analysis: An overview, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8 (2018) e1256.

[180] Z.-K. Zhang, C. Liu, X.-X. Zhan, X. Lu, C.-X. Zhang, Y.-C. Zhang, Dynamics of information diffusion and its applications on complex networks, Physics Reports 651 (2016) 1–34.

[181] M. Li, X. Wang, K. Gao, S. Zhang, A survey on information diffusion in online social networks: Models and methods, Information 8 (2017) 118.

[182] M. Nasery, Fake News on Social Media: From Fake News Lifecycle to Fake News Combat Cycle, Ph.D. thesis, "McMaster University", 2024.

[183] C. Castillo, M. El-Haddad, J. Pfeffer, M. Stempeck, Characterizing the life cycle of online news stories using social media reactions, in: Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing, 2014, pp. 211–223.

[184] V. Bakir, A. McStay, Fake news and the economy of emotions: Problems, causes, solutions, Digital journalism 6 (2018) 154–175.

[185] Y. L. Huang, K. Starbird, M. Orand, S. A. Stanek, H. T. Pedersen, Connected through crisis: Emotional proximity and the spread of misinformation online, in: Proceedings of the 18th ACM conference on computer supported cooperative work & social computing, 2015, pp. 969–980.

[186] A. M. Mehta, B. F. Liu, E. Tyquin, L. Tam, A process view of crisis misinformation: How public relations professionals detect, manage, and evaluate crisis misinformation, Public relations review 47 (2021) 102040.

[187] T. Tran, P. Rad, R. Valecha, H. R. Rao, Misinformation harms during crises: When the human and machine loops interact, in: 2019 IEEE International Conference on Big Data (Big Data), IEEE, 2019, pp. 4644–4646.

[188] T. Tran, R. Valecha, P. Rad, H. R. Rao, An investigation of misinformation harms related to social media during humanitarian crises, in: Secure Knowledge Management In Artificial Intelligence Era: 8th International Conference, SKM 2019, Goa, India, December 21–22, 2019, Proceedings 8, Springer, 2020, pp. 167–181.

[189] T. G. Van der Meer, Y. Jin, Seeking formula for misinformation treatment in public health crises: The effects of corrective information type and source, Health communication 35 (2020) 560–575.

[190] X. Lu, Y. Jin, Information vetting as a key component in social-mediated crisis communication: An exploratory study to examine the initial conceptualization, Public relations review 46 (2020) 101891.

[191] K. Na, R. K. Garrett, M. D. Slater, Rumor acceptance during public health crises: Testing the emotional congruence hypothesis, Journal of Health Communication 23 (2018) 791–799.

[192] T. Tran, R. Valecha, P. Rad, H. R. Rao, An investigation of misinformation harms related to social media during humanitarian crises, in: S. K. Sahay, N. Goel, V. Patil, M. Jadliwala (Eds.), Secure Knowledge Management In Artificial Intelligence Era, Springer Singapore, Singapore, 2020, pp. 167–181.

[193] C. Suryana, B. Budiandru, K. E. T. Naibaho, Y. Setianti, H. Purwosusanto, The impact of fake news on public opinion during crisis situations, The Journal of Academic Science 1 (2024) 395–407.

[194] M. Monmonier, Semiology of graphics: Diagrams, networks, maps., 1985.

[195] P. Rani, J. Shokeen, A survey of tools for social network analysis, International Journal of Web Engineering and Technology 16 (2021) 189–216.

[196] A. Sapountzi, K. E. Psannis, Social networking data analysis tools & challenges, Future Generation Computer Systems 86 (2018) 893–913.

[197] N. Akhtar, Social network analysis tools, in: 2014 fourth international conference on communication systems and network technologies, IEEE, 2014, pp. 388–392.

[198] B. Shneiderman, C. Dunne, Interactive network exploration to derive insights: Filtering, clustering, grouping, and simplification, in: Graph Drawing: 20th International Symposium, GD 2012, Redmond, WA, USA, September 19-21, 2012, Revised Selected Papers 20, Springer, 2013, pp. 2–18.

[199] T. Y. Berger-Wolf, J. Saia, A framework for analysis of dynamic social networks, in: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006, pp. 523–528.

[200] J. Jasser, Dynamics of misinformation cascades, in: Companion Proceedings of The 2019 World Wide Web Conference, 2019, pp. 33–36.

[201] D. R. Farine, When to choose dynamic vs. static social network analysis, Journal of animal ecology 87 (2018) 128–138.

[202] J. Kim, B. Tabibian, A. Oh, B. Schölkopf, M. Gomez-Rodriguez, Leveraging the crowd to detect and reduce the spread of fake news and misinformation, in: Proceedings of the eleventh ACM international conference on web search and data mining, 2018, pp. 324–332.

[203] D. Krackhardt, Graph theoretical dimensions of informal organizations, in: Computational organization theory, Psychology Press, 2014, pp. 107–130.

[204] C. Shao, P.-M. Hui, P. Cui, X. Jiang, Y. Peng, Tracking and characterizing the competition of fact checking and misinformation: case studies, IEEE access 6 (2018) 75327–75341.

[205] S. Kumar, R. West, J. Leskovec, Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes, in: Proceedings of the 25th international conference on World Wide Web, 2016, pp. 591–602.

[206] P. Pascual-Ferrá, N. Alperstein, D. J. Barnett, Social network analysis of covid-19 public discourse on twitter: implications for risk communication, Disaster medicine and public health preparedness 16 (2022) 561–569.

[207] N. Durmaz, E. Hengirmen, The dramatic increase in anti-vaccine discourses during the covid-19 pandemic: a social network analysis of twitter, Human vaccines & immunotherapeutics 18 (2022) 2025008.

[208] Z. Duzen, M. Riveni, M. S. Aktas, Analyzing the spread of misinformation on social networks: A process and software architecture for detection and analysis, Computers 12 (2023) 232.

[209] G. Zhang, A. Giachanou, P. Rosso, Scenefnd: Multimodal fake news detection by modelling scene context information, Journal of Information Science 50 (2024) 355–367.

[210] M. Szurawitzki, Analyzing the language of social networking sites-an analysis model, in: Proceedings of 24th Scandinavian Conference of Linguistics, 2012, pp. 355–363.

[211] M. Bahja, G. A. Safdar, Unlink the link between covid-19 and 5g networks: an nlp and sna based approach, Ieee Access 8 (2020) 209127–209137.

[212] N. Kalantari, D. Liao, V. G. Motti, Characterizing the online discourse in twitter: Users' reaction to misinformation around covid-19 in twitter, in: 2021 IEEE International Conference on Big Data (Big Data), IEEE, 2021, pp. 4371–4380.

[213] Y. Li, D. Wang, X. Li, Y. Zhai, C. Hon, Misinformation features detection in weibo: Unsupervised learning, latent dirichlet allocation, and network structure, IEEE Access (2024).

[214] M. Paraschiv, N. Salamanos, C. Iordanou, N. Laoutaris, M. Sirivianos, A unified graph-based approach to disinformation detection using contextual and semantic relations, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 16, 2022, pp. 747–758.

[215] S. Sivasankari, G. Vadivu, Tracing the fake news propagation path using social network analysis, Soft Computing 26 (2022) 12883–12891.

[216] H. Jwa, D. Oh, K. Park, J. M. Kang, H. Lim, exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert), Applied Sciences 9 (2019) 4062.

[217] R. Vijjali, P. Potluri, S. Kumar, S. Teki, Two stage transformer model for covid-19 fake news detection and fact checking, arXiv preprint arXiv:2011.13253 (2020).

[218] W. H. Gomaa, A. A. Fahmy, et al., A survey of text similarity approaches, international journal of Computer Applications 68 (2013) 13–18.

[219] R. Mihalcea, C. Corley, C. Strapparava, et al., Corpus-based and knowledge-based measures of text semantic similarity, in: Aaai, volume 6, 2006, pp. 775–780.

[220] J. Gaglani, Y. Gandhi, S. Gogate, A. Halbe, Unsupervised whatsapp fake news detection using semantic search, in: 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE, 2020, pp. 285–289.

[221] X. Guo, H. Mirzaalian, E. Sabir, A. Jaiswal, W. Abd-Almageed, Cord19sts: Covid-19 semantic textual similarity dataset, arXiv preprint arXiv:2007.02461 (2020).

[222] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv (2019).

[223] H. Abdi, L. J. Williams, Principal component analysis, Wiley interdisciplinary reviews: computational statistics 2 (2010) 433–459.

[224] A. Agarap, Deep learning using rectified linear units (relu), arXiv preprint arXiv:1803.08375 (2018).

[225] J. S. Bridle, Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition, in: Neurocomputing: Algorithms, architectures and applications, Springer, 1990, pp. 227–236.

[226] D. P. Kingma, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[227] R. Speer, J. Chin, An ensemble method to produce high-quality word embeddings (2016), 2019. arXiv:1604.01692.

[228] W. Yin, H. Schütze, Learning meta-embeddings by using ensembles of embedding sets, 2015. arXiv:1508.04257.

[229] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.

[230] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzman, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, arXiv (2019).

[231] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, L. Specia, SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1–14. URL: https://www.aclweb.org/anthology/S17-2001. doi:doi: 10.18653/v1/S17-2001.

[232] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020. arXiv:2002.10957.

[233] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, Mpnet: Masked and permuted pre-training for language understanding, Advances in neural information processing systems 33 (2020) 16857–16867.

[234] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).

[235] H. Schwenk, V. Chaudhary, S. Sun, H. Gong, F. Guzmán, Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia, arXiv preprint arXiv:1907.05791 (2019).

[236] J. Tiedemann, Parallel data, tools and interfaces in opus., in: Lrec, volume 2012, Citeseer, 2012, pp. 2214–2218.

[237] L. H. X. Ng, K. M. Carley, The coronavirus is a bioweapon: Analysing coronavirus fact-checked stories, arXiv preprint arXiv:2104.01215 (2021).

[238] J. Torregrosa, S. D'Antonio-Maceiras, G. Villar-Rodríguez, A. Hussain, E. Cambria, D. Camacho, A mixed approach for aggressive political discourse analysis on twitter, Cognitive computation 15 (2023) 440–465.

[239] C. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: Proceedings of the international AAAI conference on web and social media, volume 8, 2014, pp. 216–225.

[240] F. Barbieri, L. Espinosa Anke, J. Camacho-Collados, XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 258–266. URL: https://aclanthology.org/2022.lrec-1.27.

[241] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426 (2018).

[242] A. Williams, N. Nangia, S. R. Bowman, A broad-coverage challenge corpus for sentence understanding through inference, arXiv preprint arXiv:1704.05426 (2017).

[243] A. Talman, S. Chatzikyriakidis, Testing the generalization power of neural network models across nli benchmarks, arXiv:1810.09774 [cs] (2019). ArXiv: 1810.09774.

[244] M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, R. Zamparelli, A sick cure for the evaluation of compositional distributional semantic models, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), 2014, p. 216–223. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/363_Paper.pdf.

[245] D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, arXiv (2014).

[246] S. Subramanian, A. Trischler, Y. Bengio, C. J. Pal, Learning general purpose distributed sentence representations via large scale multi-task learning, arXiv preprint arXiv:1804.00079 (2018).

[247] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, A. Bordes, Supervised learning of universal sentence representations from natural language inference data, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 670–680.

[248] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, D. Kiela, Adversarial nli: A new benchmark for natural language understanding, arXiv preprint arXiv:1910.14599 (2019).

[249] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, Fever: a large-scale dataset for fact extraction and verification, arXiv preprint arXiv:1803.05355 (2018).

[250] I. Larraz, R. Salaverría-Aliaga, J. Serrano-Puche, Combating repeated lies: The impact of fact-checking on persistent falsehoods by politicians, Media and Communication (2024).

[251] E. C. Tandoc Jr, The facts of fake news: A research review, Sociology Compass 13 (2019) e12724.

[252] M. Himelein-Wachowiak, S. Giorgi, A. Devoto, M. Rahman, L. Ungar, H. A. Schwartz, D. H. Epstein, L. Leggio, B. Curtis, Bots and misinformation spread on social media: Implications for covid-19, Journal of medical Internet research 23 (2021) e26933.

[253] J. M. Noguera-Vivo, M. del Mar Grandío-Pérez, G. Villar-Rodríguez, A. Martín, D. Camacho, Disinformation and vaccines on social networks: Behavior of hoaxes on twitter, Revista Latina de Comunicación Social (2023) 44–62.

[254] K. Shu, X. Zhou, S. Wang, R. Zafarani, H. Liu, The role of user profiles for fake news detection, in: Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining, 2019, pp. 436–439.

[255] K. Shu, S. Wang, H. Liu, Understanding user profiles on social media for fake news detection, in: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), IEEE, 2018, pp. 430–435.

[256] M. Grootendorst, Keybert: Minimal keyword extraction with bert., 2020. URL: https://doi.org/10.5281/zenodo.4461265. doi:doi: 10.5281/zenodo.4461265.

[257] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, L. Specia, Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation, arXiv preprint arXiv:1708.00055 (2017).

[258] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, R. Vollgraf, Flair: An easy-to-use framework for state-of-the-art nlp, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations), 2019, pp. 54–59.

[259] M. Honnibal, I. Montani, spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, 2017. To appear.

[260] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, L. De Alfaro, Some like it hoax: Automated fake news detection in social networks, arXiv preprint arXiv:1704.07506 (2017).

[261] D. Saby, O. Philippe, N. Buslón, J. del Valle, O. Puig, R. Salaverría, M. J. Rementeria, Twitter analysis of covid-19 misinformation in spain, in: Computational Data and Social Networks: 10th International Conference, CSoNet 2021, Virtual Event, November 15–17, 2021, Proceedings 10, Springer, 2021, pp. 267–278.

[262] R. Carrasco Polaino, M. Á. Martín Cárdaba, E. Villar Cirujano, Participación ciudadana en twitter. polémicas anti-vacunas en tiempos de covid-19, Comunicar: Revista científica iberoamericana de comunicación y educación.(Ejemplar dedicado a: Participación ciudadana en la esfera digital) 29 (2021) 21–31.

[263] A. Almansa-Martínez, M. J. Fernández-Torres, L. Rodríguez-Fernández, Desinformación en españa un año después de la covid-19. análisis de las verificaciones de newtral y maldita, Revista Latina de Comunicación Social (2022) 183–200.

[264] G. Villar-Rodríguez, M. Souto-Rico, A. Martín, Virality, only the tip of the iceberg: ways of spread and interaction around covid-19 misinformation in twitter, Communication & Society (2022) 239–256.

[265] I. Hasan, S. Rizvi, Review of ai techniques and cognitive computing framework for intelligent decision support, in: 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom), IEEE, 2021, pp. 891–898.

[266] H. J. Watson, Preparing for the cognitive generation of decision support., MIS Quarterly Executive 16 (2017).

[267] I. Larraz, F. Sallicati, et al., Semantic similarity models for automated fact-checking: Claimcheck as a claim matching tool, Profesional de la información 32 (2023).

[268] S. Hölig, J. Behre, W. Schulz, Reuters Institute Digital News Report 2022: Ergebnisse für Deutschland, Technical Report, Verlag Hans-Bredow-Institut, 2022.

[269] U. Jeong, A. Nirmal, K. Jha, S. X. Tang, H. R. Bernard, H. Liu, User migration across multiple social media platforms, in: Proceedings of the 2024 SIAM International Conference on Data Mining (SDM), SIAM, 2024, pp. 436–444.

[270] A. Datta, S. Buchegger, L.-H. Vu, T. Strufe, K. Rzadca, Decentralized online social networks, Handbook of social network technologies and applications (2010) 349–378.

[271] R. Salaverría, M.-P. Martínez-Costa, C. González Tosat, Decentralised networks as a tool for fighting disinformation and censorship: The fediverse and free, collaborative and open networks, in: Journalism, Digital Media and the Fourth Industrial Revolution, Springer, 2024, pp. 15–25.

[272] C. A. Bono, L. La Cava, L. Luceri, F. Pierri, An exploration of decentralized moderation on mastodon, in: Proceedings of the 16th ACM Web Science Conference, 2024, pp. 53–58.

[273] Á. Huertas-García, A. Martín, J. Huertas-Tato, D. Camacho, Countering malicious content moderation evasion in online social networks: Simulation and detection of word camouflage, Applied Soft Computing 145 (2023) 110552.

[274] Á. Huertas-García, A. Martín, J. Huertas-Tato, D. Camacho, Camouflage is all you need: Evaluating and enhancing language model robustness against camouflage adversarial attacks, arXiv preprint arXiv:2402.09874 (2024).

[275] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, Semeval-2022 task 5: Multimedia automatic misogyny identification, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), 2022, pp. 533–549.

[276] A. Giachanou, G. Zhang, P. Rosso, Multimodal multi-image fake news detection, in: 2020 IEEE 7th international conference on data science and advanced analytics (DSAA), IEEE, 2020, pp. 647–654.

[277] S. Hangloo, B. Arora, Combating multimodal fake news on social media: methods, datasets, and future perspective, Multimedia systems 28 (2022) 2391–2422.

[278] R. T. Mutanga, N. Naicker, O. O. Olugbara, Hate speech detection in twitter using transformer methods, International Journal of Advanced Computer Science and Applications 11 (2020).

[279] S. G. Roy, U. Narayan, T. Raha, Z. Abid, V. Varma, Leveraging multilingual transformers for hate speech detection, arXiv preprint arXiv:2101.03207 (2021).

[280] A. Huertas-García, J. Huertas-Tato, A. Martín, D. Camacho, Profiling hate speech spreaders on twitter: Transformers and mixed pooling., CLEF (Working Notes) 2021 (2021).

[281] J. Huertas-Tato, A. Martín, D. Camacho, Understanding writing style in social media with a supervised contrastively pre-trained transformer, Knowledge-Based Systems 296 (2024) 111867.

[282] M. Grootendorst, Bertopic: Neural topic modeling with a class-based tf-idf procedure, arXiv preprint arXiv:2203.05794 (2022).