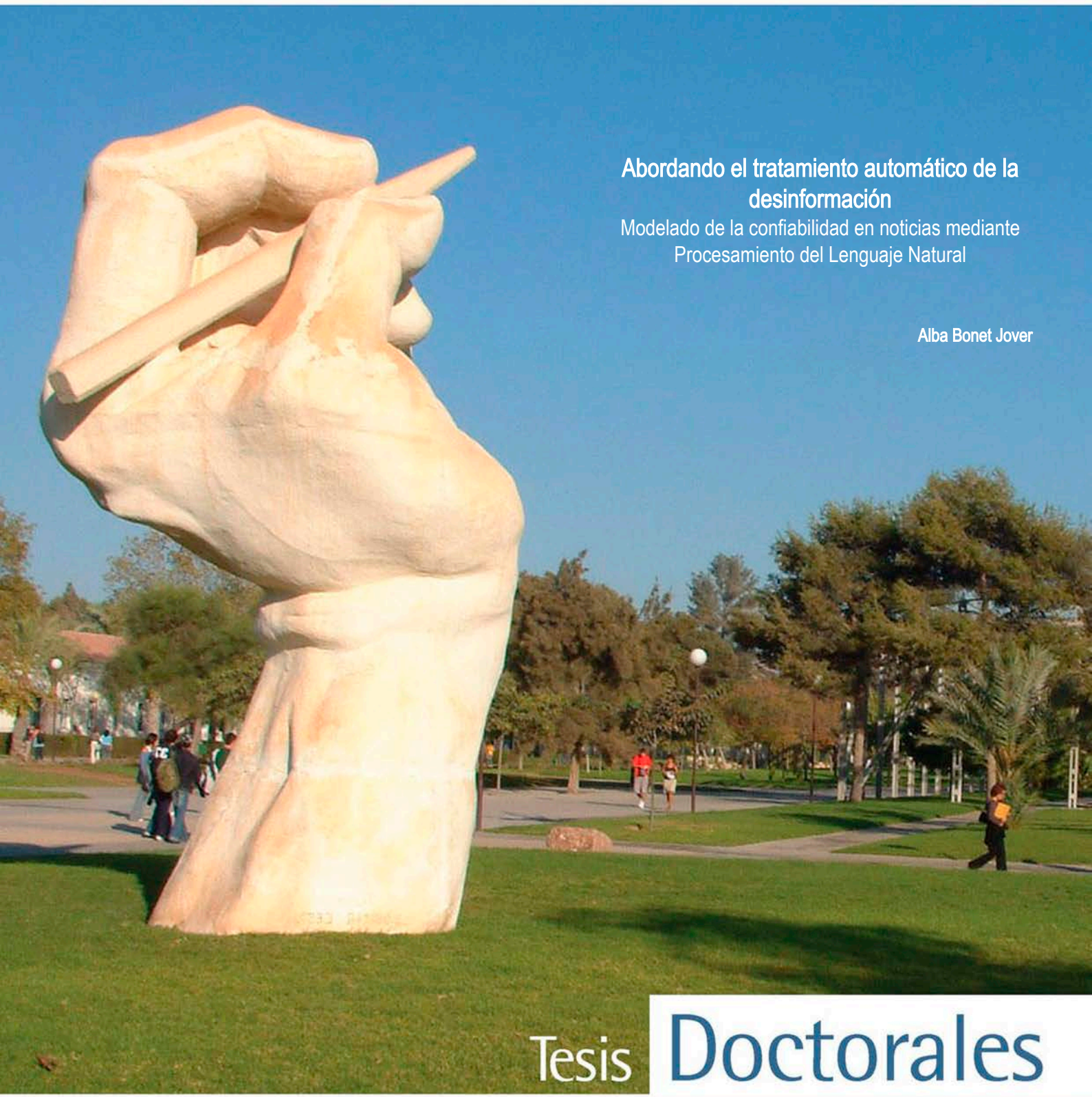
Universitat d'Alacant
Universidad de Alicante



Abordando el tratamiento automático de la desinformación

Modelado de la confiabilidad en noticias mediante Procesamiento del Lenguaje Natural

Alba Bonet Jover

Tesis Doctorales
UNIVERSIDAD de ALICANTE

Unitat de Digitalització UA
Unidad de Digitalización UA

Universitat d'Alacant
Universidad de Alicante

Departamento de Lenguajes y Sistemas Informáticos
Escuela Politécnica Superior

Abordando el tratamiento automático de la desinformación

Modelado de la confiabilidad en noticias mediante Procesamiento del Lenguaje Natural

Alba Bonet Jover

Tesis presentada para aspirar al grado de

DOCTORA POR LA UNIVERSIDAD DE ALICANTE

DOCTORADO EN INFORMÁTICA

Dirigida por

Dra. Estela Saquete Boró
Dr. Patricio Martínez Barco

Tesis financiada por la Generalitat Valenciana a través del Programa para la promoción de la investigación científica, el desarrollo tecnológico y la innovación en la Comunitat Valenciana (ACIF/2020/177)

Agradecimientos

Cuando empecé esta aventura, no imaginaba lo que en realidad me iba a deparar. De hecho, nunca imaginé que llegaría hasta aquí. La vida a veces te regala oportunidades que ni habías contemplado pero que aparecen para ayudarte a crecer. Eso ha supuesto precisamente para mí esta tesis, un gran crecimiento a nivel personal y académico que me ha permitido volcar mis conocimientos y aprender en un mundo que era totalmente desconocido para mí. Ha sido una carrera de fondo, intensa y continua. Una carrera que no hubiese podido recorrer yo sola. Este trabajo no hubiese sido posible sin la ayuda y el apoyo de todos los que han sido testigos de alguna forma de los esfuerzos y los logros de esta investigación, a los cuales les estoy enormemente agradecida.

A mis padres, que son mis grandes pilares en todos los aspectos de mi vida, mi apoyo incondicional. Ellos son la razón de mis éxitos y a ellos les debo todo lo que he conseguido, pues soy reflejo de toda la maravillosa educación que me han dado y del gran amor que me han transmitido. Ellos son los que me impulsan a perseguir mis sueños cueste lo que cueste, los que me animan a seguir mi estrella. A mi padre, por tener siempre unas palabras reconfortantes, por sus valiosos consejos, que me ayudan en todas las decisiones que tomo, por su ejemplo de lucha, por su pasión por todo lo que hago, por los momentos juntos, trabajando o disfrutando, y por su cariño. A mi madre, por su eterna dulzura, por apoyarme y cuidarme cada día, por sus abrazos revitalizantes y su positividad, por los dulces que me regala para animarme y por las risas compartidas.

A Javier, por su paciencia estos tres años, por sacarme miles de sonrisas, por aguantar mi estrés y cansancio, por abrazarme en mis peores días y celebrar mis mejores momentos. Por animarme cada día a dar un pasito más en este reto y por ser el otro gran pilar de mi vida.

A mi familia, por estar siempre a mi lado y regalarme momentos de paz. A Syra, mi labradora, por todas las horas de trabajo tumbada a mi lado haciéndome compañía.

A mis amigos, por entender mi desconexión y alegrarme los días que nos juntábamos. A Luis, por seguir recorriendo juntos el mismo camino, aunque por sendas diferentes, por seguir apoyándonos y ayudándonos, por las cenas hasta las seis de la mañana hablando de traducción, derecho comparado o fake news.

A Robiert, porque sin él no hubiese podido sacar la tesis a flote, por su eterna paciencia, sus clases y sus horas de revisión, por toda la experimentación que ha

fundamentado mi trabajo, por haber sido un gran apoyo y un buen amigo.

A María, por ser la otra lingüista perdida en un mundo de tecnologías, por compartir tantas risas y momentos de alegría, por estar siempre dispuesta a ayudarme.

A Mario, por facilitarme el trabajo técnico y ayudarme a avanzar en la investigación, por estar dispuesto a ayudarme a cualquier hora.

A Paul, por estar siempre dispuesto a ayudarme y a perfeccionar mis textos en inglés, aún en la distancia.

A Bea, Javi, Sergio, Iván, Fabio y José, por los ánimos, las pausas-café, los almuerzos improvisados y por ser un equipo maravilloso. A Tania por estar siempre dispuesta a ayudar en los artículos y por las excelentes correcciones que han mejorado el trabajo. A Ale y Suilan, por su colaboración y ayuda en la etapa inicial de mi tesis. A Yoan, por el apoyo constante y los ánimos para seguir avanzando en esta carrera. A todo el grupo GPLSI por hacerme sentir una más del equipo y por esta oportunidad que me ha ayudado a crecer tanto personalmente como profesionalmente.

A Estela y Patricio por ser los mejores directores que podrían haberme guiado en este camino, por ayudarme en cualquier momento y por el gran esfuerzo de revisión, corrección y mejora de este trabajo. A Estela, por su apoyo desde el inicio, por orientarme y ayudarme a crecer, por los buenos momentos compartidos y también por su empuje constante, que ha permitido que cumplamos los plazos y saquemos la tesis adelante en menos tiempo de lo planeado. A Patricio, por estar siempre atento a los pasos que debía dar, por sus consejos, por disipar todas mis dudas y por facilitarme el camino en un mundo nuevo para mí.

Todos ellos son la verdadera esencia de esta tesis, pues me han dado la energía suficiente para llegar a la meta.

Gracias de corazón.

La investigación desarrollada en esta tesis ha sido financiada por el Gobierno de España a través de los proyectos "Modelang: Modeling the behavior of digital entities by Human Language Technologies" (RTI2018-094653-B-C22) y "LIVING-LANG: Living Digital Entities by Human Language Technologies" (RTI2018-094653-B-C21/C22); por el Ministerio de Ciencia e Innovación y el Fondo Europeo de Desarrollo Regional (FEDER) a través del proyecto "TRIVIAL: Technological Resources for Intelligent VIral AnaLysis through NLP" (PID2021-122263OB-C22); así como financiado por la Generalitat Valenciana a través del proyecto "SIIA: Tecnologías del lenguaje humano para una sociedad inclusiva, igualitaria y accesible" (PROMETEU/2018/089), el proyecto "NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation" (CIPROM/2021/21) y la subvención del "Programa para la promoción de la investigación científica, el desarrollo tecnológico y la innovación en la Comunitat Valenciana", para la contratación de personal investigador de carácter predoctoral (ACIF/2020/177).

Resumen

La llegada de Internet y de las nuevas tecnologías dio lugar al nacimiento de la era de la información, una era que ha conectado a la sociedad de forma global y le ha permitido acceder libremente a la información digital. Con esta facilidad de acceso, cualquier persona, aún sin ser experta en la materia, puede publicar y acceder a la información sin ningún coste, lo que ha ocasionado un exceso de información no contrastada que muchas veces oculta intenciones como el engaño, la manipulación o los fines económicos. De esa forma, la era de la información se ha transformado en la era de la desinformación. La incesante necesidad de estar informados ha motivado que el consumo de la información se convierta en una rutina, ya sea siguiendo las últimas noticias en portales digitales o leyendo a diario publicaciones de personas afines.

Antes, la información viajaba en forma de sonido a través de la radio o en forma de tinta a través de los periódicos, pero ahora una desmedida cantidad de información se propaga a través de algoritmos. Las tecnologías han propiciado la sobreabundancia de información, así como la propagación de noticias falsas y bulos, hasta tal punto que resulta imposible contrastar y procesar manualmente tales volúmenes de desinformación en tiempo real. No obstante, lo que se considera un problema puede convertirse en una solución, pues igual que los algoritmos y el entorno digital son los causantes de la viralización de la información falsa, estos pueden ser a su vez los detectores de la desinformación.

Es aquí donde el Procesamiento del Lenguaje Natural desempeña un papel clave en la relación humano-máquina, modelando el lenguaje humano a través de la comprensión y generación automática del lenguaje, y entrenando modelos a través de la retroalimentación del experto. El trabajo coordinado entre la ingeniería computacional y la lingüística es decisivo a la hora de frenar el fenómeno de la desinformación. Son necesarias las dos perspectivas para abordar la detección automática de la forma más completa y precisa posible, pues el análisis lingüístico permite detectar y estudiar patrones textuales que hacen que la información de una noticia sea o no sea confiable, mientras que el entorno tecnológico se encarga de automatizar la detección de los patrones anotados mediante el entrenamiento de algoritmos de aprendizaje automático.

Específicamente para esta tarea, donde la noticia es el objeto de estudio, el análisis a nivel periodístico también es fundamental. La noticia suele presentar una estructura determinada, técnica conocida como la Pirámide Invertida, don-

de la información sigue un orden de relevancia concreto con el fin de captar la atención del lector. Además, suele organizar el contenido de forma precisa y completa respondiendo a seis preguntas clave, conocidas como las 5W1H. Estas dos técnicas periodísticas permiten construir una noticia siguiendo unos estándares de calidad y son la base de la anotación de la presente investigación.

Para contribuir a la tarea de la detección de desinformación, la presente investigación presenta dos guías de anotación de grano fino diseñadas para anotar tanto la veracidad (guía FNDeepML) como la confiabilidad (guía RUN-AS) de las noticias. Además, se presentan los dos corpus obtenidos y anotados con las guías de anotación, uno de ellos compuesto por 200 noticias verdaderas y falsas (corpus FNDeep) y otro que incluye 170 noticias confiables y no confiables (corpus RUN), ambos en español. Un extenso marco de evaluación se lleva a cabo para validar tanto la calidad de la anotación como la de los recursos, obteniendo resultados prometedores que muestran que el entrenamiento con las características de la anotación mejoran notablemente los modelos de predicción. Asimismo, otras dos aportaciones de la tesis relacionadas más bien con el proceso de anotación y de detección son, por un lado, la propuesta de una metodología semiautomática de anotación que agiliza la tarea del experto anotador y, por otro lado, una arquitectura para la detección de desinformación basada en una capa de estructura y otra de predicción. Las aportaciones de este trabajo permiten abordar una parte del problema de la detección de la desinformación aplicando técnicas de Procesamiento del Lenguaje Natural, pero desde un enfoque lingüístico, lo que permite profundizar en el estudio del problema desde su raíz. El conocimiento profundo del lenguaje de las noticias, y específicamente el modelado de un lenguaje propio de la desinformación, permite no solo dar un paso más en su detección, sino además justificar la confiabilidad de la noticia.

Abstract

Internet and new technologies gave birth to the information age, a period that has connected society globally and given it free access to digital information. With this accessibility, anyone, including non-experts, can publish and access information at no cost. This has led to an excess of unverified information that often hides ulterior motives such as deception, manipulation or economic purposes. Thus, the age of information has turned into the age of disinformation. The incessant need to be informed has turned information consumption into a routine, whether it is following the latest news on digital portals or reading daily posts by like-minded people.

In the past, information travelled in the form of sound through radio or in the form of ink through traditional newspapers. Nowadays, however, an inordinate amount of information is propagated through algorithms. Technologies have led to a glut of information combined with the spreading of fake news and hoaxes, to such an extent that it is impossible to manually check and process such volumes of disinformation in real time. However, the cause of the problem may also be the solution: just as algorithms and the digital environment are responsible for the spreading of disinformation, they can also be used to detect it.

Natural Language Processing plays a key role in the human-machine relationship, by modelling human language through automatic language understanding and generation, and training models through expert feedback. The coordinated work between computer engineering and linguistics is decisive in curbing the disinformation phenomenon. Both perspectives are necessary to render automatic detection as complete and precise as possible, as the linguistic analysis allows the detection and study of textual patterns that determine whether the information in a news item is reliable, while the technological environment is responsible for automating the detection of the annotated patterns by training machine-learning algorithms.

Specifically for this task, where the news item is the object of study, analysis at journalistic level is also fundamental. News usually presents a specific structure by following the technique known as the Inverted Pyramid, where the information is set down in a specific order of relevance in order to catch the reader's attention. In addition, content is usually organised in a precise and complete way by answering six key questions, known as the 5W1H. These two journalistic techniques allow the construction of a news story according to quality standards

and are the basis for the annotation of this research.

To contribute to the task of detecting disinformation, the present research presents two fine-grained annotation schemes designed to annotate both the veracity (FNDeepML scheme) and the reliability (RUN-AS scheme) of news. In addition, the two datasets obtained and annotated with the annotation schemes are presented, one consisting of 200 true and false news items (FNDeep dataset) and the other including 170 reliable and unreliable news items (RUN dataset), both in Spanish. An extensive evaluation framework is established to validate both the quality of the annotation and the resources, obtaining promising results that show that training with the annotation features significantly improves the prediction models. Two other contributions of the thesis related to the annotation and detection processes are, on the one hand, the design of a semi-automatic annotation methodology that accelerates the task of the expert annotator and, on the other hand, an architecture for disinformation detection based on a structure layer and a prediction layer. The contributions of this work address part of the problem of disinformation detection by applying Natural Language Processing techniques, but from a linguistic approach, which allows us to delve into the root causes of the problem. In-depth knowledge of the language of news, in particular the modelling of a specific disinformation language, not only allows us to take a step further towards detecting it, but also to demonstrate the reliability of the news item.

Índice general

Índice de figuras

Índice de tablas

Acrónimos

AL	Aprendizaje Activo —*Active Learning*—
API	*Application Programming Interface*
AdaBoost	*Adaptive Boosting*
BoW	*Bag-of-Words*
BERT	*Bidirectional Encoder Representations from Transformers*
BO	*Boosting*
CRF	*Conditional Random Fields*
CTF	*COVID-19 Twitter Fake News dataset*
DL	Aprendizaje Profundo —*Deep Learning*—
DT	*Decision Tree*
FAQ	preguntas frecuentes —*Frequently Asked Questions*—
GaussianNB	*Gaussian Naive Bayes*
HITL	*Human-in-the-loop*

IA	Inteligencia Artificial
IAA	acuerdo entre anotadores —*Inter-Annotator Agreement*—
IFCN	Red Internacional de Verificación de Datos —*International Fact Checking Network*—
LIWC	*Linguistic Inquiry and Word Count*
LR	Regresión Logística —*Logistic Regression*—
LSTM	*Long Short-Term Memory*
ML	Aprendizaje Automático —*Machine Learning*—
MLP	*Multilayer Perceptron*
MNB	*Multinomial Naive Bayes*
POS	*Part-Of-Speech*
PLN	Procesamiento del Lenguaje Natural
QA	*Question Answering*
RF	*Random Forest*
SEPLN	Sociedad Española para el Procesamiento del Lenguaje Natural
SQuAD	*Stanford Question Answering Dataset*
SVM	Máquinas de Soporte Vectorial —*Support Vector Machine*—

SOTA	estado del arte —*State of the Art*—
TF-IDF	*Term Frequency-Inverse Document Frequency*
TLH	Tecnologías del Lenguaje Humano

CAPÍTULO 1

Introducción

1.1 Motivación

Vivimos en la era global de la desinformación, una era dominada por el conocido fenómeno de la infodemia (*infodemic* en inglés), neologismo utilizado para hacer referencia a la sobreabundancia de información sobre un tema[1]. Además, este término es utilizado por la Organización Mundial de la Salud para hacer referencia al exceso de información (incluida la difusión deliberada de información engañosa) durante un brote de enfermedad, que perjudica la salud física y mental de las personas e incrementa la estigmatización, el discurso de odio y la polarización[2]. A lo largo de la historia, la desinformación ha sido empleada como una técnica de engaño para manipular la verdad. Un ejemplo de ello es la propaganda política utilizada en entornos bélicos para condicionar las opiniones del enemigo o llevarle a tomar decisiones con información distorsionada (Zabala, 2021), como fue el caso de la propaganda nazi de Joseph Goebbels en la Segunda Guerra Mundial o la desinformación acerca de la guerra Ucrania-Rusia. La viralización de noticias falsas afectó a la campaña electoral de los Estados Unidos de 2016 influyendo en la victoria de Donald Trump, así como al referéndum del Brexit en Reino Unido. Además del contexto político, otro escenario susceptible a la manipulación de la verdad es el de la salud y esto se ha podido ver recientemente con la pandemia del covid-19, situación de emergencia en la que la propagación de bulos y noticias falsas se acentúa.

El poder de los textos escritos para influir en el pensamiento y la opinión sobre alguien o algo siempre ha existido en la vida social (Nycyk, 2015), pues las palabras tienen un gran poder a la hora de perfilar las opiniones y creencias de

[1] https://www.fundeu.es/recomendacion/infodemia/

[2] https://www.who.int/es/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation

las personas (Rashkin, Choi, Jang, Volkova, y Choi, 2017). Ese es precisamente el mayor peligro de la desinformación: el poder de manipular la opinión pública, de jugar con las emociones de las personas, de generar miedo y preocupación, de crear prejuicios, todo ello con el fin de obtener generalmente un beneficio económico o conseguir un fin ideológico. La desinformación es muy peligrosa, pues además de causar daños emocionales (esperanza, miedo), ideológicos (religión, política) y económicos, puede incluso llegar a causar daños físicos cuando afecta a temas tan importantes como el de la salud pública, al conducir a la gente a probar "curas milagrosas", remedios caseros o técnicas no aprobadas ni recomendadas por especialistas. En este capítulo se contextualiza el fenómeno de la desinformación, poniendo el foco en las noticias falsas, y se presentan las necesidades, hipótesis y objetivos de la presente tesis.

1.1.1 Desinformación

La desinformación se ha convertido en un problema social a nivel global. En inglés, existen principalmente dos términos que hacen referencia a toda aquella información engañosa, incorrecta o imprecisa: *misinformation* y *disinformation*. La principal diferencia entre ambos términos reside en la intención (Rubin, 2019), en si la información inexacta resulta de un intento deliberado de engañar o confundir, en cuyo caso hablaríamos de *disinformation*, o si resulta de un error honesto (Hernon, 1995), es decir, que no se creó con la intención de hacer daño (Wardle y Derakhshan, 2017), donde hablaríamos de *misinformation*. En el caso de *misinformation*, la información inexacta puede deberse a un error, negligencia o sesgo inconsciente, mientras que con el término *disinformation* estamos aludiendo a la información creada por una persona que está decidida a engañar (Fallis, 2014).

Aunque la intención no sea la misma, la relación entre ambos conceptos es estrecha, pues una persona puede compartir con buena intención (*misinformation*) una noticia creada por otra persona con la voluntad de engañar (*disinformation*), sin saber que la información que está compartiendo es falsa. La intencionalidad marca la diferencia entre ambos conceptos, lo que influye en la manera de crear las noticias, como podría verse en rasgos como la subjetividad, el intento de convencer o de influir en aspectos emocionales o el discurso de odio. No obstante, en otros muchos casos esa diferencia no se aprecia tan fácilmente. En español se utiliza el término **desinformación** para designar ambos conceptos y durante el presente trabajo se utilizará ese término para aludir a toda aquella información falsa difundida con una aparente finalidad, centrándonos específicamente en la información no confiable.

A nivel social, el problema de la desinformación afecta a las relaciones y a la confianza entre las personas, pues, tal y como afirma Marc Amorós, "una sociedad con mala salud informativa vive condenada a la ceguera y si llegamos al punto en que no podemos confiar en las noticias, solo nos creeremos las que reafirmen nuestro pensamiento" (García, 2018). La emoción es una de las principa-

les tácticas que utiliza la desinformación para llegar al lector, pues "las fake news tienen el objetivo de provocar que nuestra emoción ante una revelación nos nuble la razón y nos haga ser impulsivos y no reflexivos" (García, 2018). Cuando una noticia nos hace sentir miedo, esperanza o rabia, no podemos leerla y olvidarla simplemente, sino que surge la necesidad de compartirla y de mostrar nuestro punto de vista sobre un determinado tema. Con ello, queremos reforzar nuestra forma de ser, crearnos nuestra burbuja informativa compartiendo contenido con gente que piensa igual para así sentirnos aceptados y sentir que tenemos la razón: "es más fácil creernos una información que nos da la razón que una que nos la quita" (García, 2018). Esta forma de (des)informarse y compartir información es la que intensifica la polarización de la sociedad.

A nivel comunicativo, la desinformación se nutre de tres elementos: contenido engañoso, lectores susceptibles a creerse dicho contenido y un medio potente de viralización. (Rubin, 2019) presenta una analogía entre un concepto epidemiológico y el fenómeno de la desinformación: el triángulo de la desinformación (*the disinformation and misinformation triangle*). Este modelo conceptual se basa en el modelo de George McNew que refleja la interacción entre los tres factores que causan una enfermedad: patógeno, huésped y medio. Rubin traslada este concepto al de la desinformación para mostrar que, al igual que para tener una enfermedad es necesaria la presencia de los tres factores mencionados, para que se viralice una información falsa también es necesaria la presencia simultánea de tres factores: la noticia falsa (patógeno), el lector (huésped) y la plataforma digital (medio). Por ello, para mitigar el problema, propone como posibles soluciones la detección automática mediante Procesamiento del Lenguaje Natural (PLN), la educación de los lectores en desinformación y la regulación legislativa del fenómeno en el medio informativo. La presente investigación se centra en el primer factor (patógeno), que es la desinformación, y en la primera propuesta de solución, la de la detección automática mediante PLN.

1.1.2 Fake News

Las noticias falsas, *fake news* en inglés, es uno de los fenómenos más conocidos y extendidos de desinformación. Tal y como define (X. Zhou y Zafarani, 2020), las fake news son noticias intencionadamente falsas publicadas por un medio de comunicación, que pueden ser creadas por periodistas y no periodistas, y que abarcan artículos, afirmaciones, declaraciones, discursos, publicaciones, entre otros tipos de información relacionada con personajes públicos y organizaciones.

En esta tesis se utilizará tanto el término acuñado como el anglicismo que, a pesar de ser un extranjerismo no adaptado, se utilizará sin cursiva debido a su repetida aparición a lo largo de la presente investigación. Las noticias falsas siempre han existido y han circulado a través de distintos medios, pero la gran diferencia en la actualidad es que existen medios de difusión más potentes que el papel, la radio o el discurso oral: las redes sociales e Internet. El desarrollo de

las nuevas tecnologías y la llegada de Internet han propiciado la necesidad de estar informándose continuamente y, aunque el progreso tecnológico ha dado pie a un mundo más conectado, existe el riesgo de que esa mayor conectividad se utilice incorrectamente (Shu, Bhattacharjee, y cols., 2020), creando así un caldo de cultivo para la desinformación. La fácil accesibilidad a estos medios, tanto a la hora de crear como de compartir una determinada información, ha influido en la calidad periodística y el criterio de la verdad, dando lugar a un mayor volumen de noticias no contrastadas por profesionales y sin fundamento científico.

El fenómeno de las fake news ha sido clasificado por diversos investigadores. En su artículo de First Draft[3], Claire Wardle presenta siete tipos de desinformación: sátira o parodia (que pueden convertirse en *misinformation* si se malinterpreta el mensaje), conexión falsa, contenido engañoso, contexto falso, contenido impostor, contenido manipulado y contenido fabricado (Wardle y Derakhshan, 2017). Las define de la siguiente manera:

- Sátira o parodia: no tienen el objetivo de causar daño, pero tienen potencial para engañar.
- Conexión falsa: cuando los titulares, imágenes o leyendas no apoyan el contenido.
- Contenido engañoso: uso engañoso de la información para incriminar a alguien.
- Contexto falso: cuando se comparten contenidos auténticos con información contextual falsa.
- Contenido impostor: cuando se suplanta la identidad de fuentes auténticas.
- Contenido manipulado: cuando se manipula información o imágenes auténticas para engañar.
- Contenido fabricado: el nuevo contenido es completamente falso, creado para engañar y causar daño.

En la presente investigación, se analizan varios tipos de fake news, como el de contenido engañoso, contexto falso, contenido manipulado y contenido fabricado. (Tandoc Jr, Lim, y Ling, 2018) presenta una clasificación similar: noticia satírica, parodia, fabricación, manipulación, publicidad y propaganda, siendo la fabricación y manipulación (al igual que contenido el fabricado y manipulado) los dos tipos que encajan especialmente con nuestro análisis. Por otro lado, en lugar de utilizar el término desinformación o fake news, (Salaverría y cols., 2020) recurre al término bulo para referirse a aquellos contenidos que, adoptando una apariencia verdadera, tienen la finalidad deliberada de engañar.

[3] https://firstdraftnews.org/articles/fake-news-complicated/

Respecto a la difusión de la información falsa, esta se da generalmente en periódicos, revistas, medios de comunicación o redes sociales (Sadiku, Eze, y Musa, 2018), siendo estas últimas el entorno principal de difusión (89,1 %), muy por delante de los medios periodísticos (4 %) y otras plataformas diversas (6,9 %) (Salaverría y cols., 2020). Aunque cada uno de los tipos de fake news mencionados tiene un fin distinto, todos comparten el propósito de influir de una forma u otra en las emociones y en la forma de pensar del lector.

En la mayoría de los casos, las noticias falsas no son totalmente falsas, sino una versión distorsionada de algo que realmente ocurrió o una manipulación de hechos reales (Juez y Mackenzie, 2019). Las fake news se caracterizan por mezclar información falsa y verídica, lo que dificulta su detección. Para estudiarlas, se va a analizar cómo influyen la estructura y el contenido periodísticos en su detección. Aunque parte de la desinformación suele difundirse en redes sociales como Facebook, Twitter o cadenas de WhatsApp, una parte importante también se difunde a través de medios tradicionales digitales, medios a los que la sociedad recurre con frecuencia para informarse y ponerse al día, por lo que la presente tesis se centrará en la desinformación que se difunde en periódicos digitales y que sigue el formato tradicional de artículo periodístico.

Las noticias suelen presentar una estructura específica para captar la atención de los lectores y para proporcionarles la información de forma atrayente y organizada. Las noticias son normalmente más extensas que las publicaciones de redes sociales, por ello, para facilitar la legibilidad del texto y atraer al lector, las noticias suelen seguir dos conocidas técnicas periodísticas: la estructura de la Pirámide Invertida y la técnica de las 5W1H. Ambas técnicas son la base de la guía de anotación propuesta en la presente tesis.

La Pirámide Invertida es una técnica muy utilizada en periodismo para reflejar la objetividad de una noticia que, junto al principio de la neutralidad, es un rasgo distintivo de las noticias bien construidas y fidedignas (Thomson, White, y Kitley, 2008). En la hipótesis de la Pirámide Invertida, algunas partes de la noticia contienen diferentes niveles de información útil (Khan, Islam, Aleem, Iqbal, y Ahmed, 2018) y dichos niveles sitúan la información más importante al principio de la noticia y la menos relevante al final (H. Zhang y Liu, 2016). Esta estructura permite que los usuarios adquieran rápidamente los puntos clave de la historia y facilita el procesamiento de la información (DeAngelo y Yegiyan, 2019). Las noticias bien construidas suelen presentar cinco partes estructurales: el titular, el subtítulo, la entradilla, el cuerpo de la noticia y la conclusión. De estas partes, el titular, la entradilla y el cuerpo de la noticia son partes esenciales que constituyen el esqueleto de la noticia, mientras que el subtítulo y la conclusión no están siempre presentes.

Además de la estructura, los puristas del periodismo sostienen que una noticia no está completa hasta que se presenta todo el contenido esencial respondiendo a seis preguntas: qué (*what*), quién (*who*), dónde (*where*), cuándo (*when*), por qué (*why*) y cómo (*how*). Este método se conoce como las 5W1H y representa los constituyentes semánticos de una oración que son más sencillos de

entender e identificar (Chakma, Swamy, Das, y Debbarma, 2020). Además, las 5W1H ayudan a presentar claramente la información clave de una noticia de forma explícita (H. Zhang, Chen, y Ma, 2019) y permiten extraer la información semántica, por lo que son esenciales para entender toda la historia (W. Wang, Zhao, Zou, Wang, y Zheng, 2010). Si una noticia responde a todas estas preguntas y, por lo tanto, la información se presenta de forma completa, esta tendrá un mayor grado de confiabilidad que una noticia que no comunique la información con tanta precisión. Por otro lado, las 5W1H describen el evento principal de la noticia. Un evento es una forma natural de explicar las relaciones complicadas entre personas, lugares, acciones y objetos (Hordofa, 2020), pero también es la forma natural de describir una noticia, la forma en la que los consumidores entienden lo que ocurre en el mundo (Hou y cols., 2015).

La descripción detallada de cada elemento de estructura y contenido se proporciona en el Capítulo 3.

Las fake news llegan más lejos, se difunden más rápido y calan significativamente más que la verdad en todas las clases de información (Vosoughi, Roy, y Aral, 2018). Una vez se viralizan y penetran en la sociedad, son difíciles de refutar, pues "cuanto más viral se vuelve la noticia falsa, más problemas tendremos en el futuro para recordar que no era verdad" (García, 2018).

1.1.3 Confiabilidad

La confiabilidad es una métrica esencial a considerar a la hora de evaluar la calidad de la información de las noticias. Los conceptos de confiabilidad y veracidad están muy relacionados, pero por lo que se puede observar en la literatura, el término veracidad se suele utilizar en tareas en las que se contrasta y verifica la información (Vosoughi y cols., 2018), mientras que el concepto de confiabilidad se utiliza más en métodos en los que se investiga la credibilidad de la fuente de la que proviene la noticia, como es el caso del método basado en la fuente propuesto por (X. Zhou y Zafarani, 2020).

Las fake news son noticias que incluyen tanto información confiable como no confiable, lo que hace que ambos términos estén muy vinculados. Sin embargo, se puede apreciar un sutil matiz en sus definiciones de los diccionarios Oxford[4] y Collins[5]. Respecto al concepto de veracidad, ambos diccionarios lo definen como la cualidad de ser verdadero. El criterio de veracidad se utiliza para categorizar esencialmente si la noticia es verdadera o falsa. Aquí influyen diversos factores, como el perfil del lector, el contexto, el estilo del periodista y, sobre todo, el conocimiento del mundo. Únicamente con el texto de una noticia no es posible decidir ese valor de veracidad, se necesita del conocimiento externo y de la verificación de datos para poder contrastar la información y determinar la veracidad de las noticias. La presente tesis no se centra en la verificación de hechos, sino en un análisis puramente lingüístico y estructural de las noticias. Por

[4] https://www.oxfordlearnersdictionaries.com/
[5] https://www.collinsdictionary.com/

ello, el criterio de veracidad no encaja con nuestro objetivo, aunque sí lo apoya, pues nuestra meta es ser capaces de detectar automáticamente la confiabilidad de forma que se pueda utilizar dicha confiabilidad predecida como soporte a la detección de la veracidad de la noticia.

Una de nuestras hipótesis defiende que la forma en la que se comunica una noticia, el estilo, el lenguaje utilizado, la estructura, la neutralidad o la precisión del contenido son características esenciales que pueden marcar la diferencia entre una noticia confiable y otra que no lo es. El diccionario Oxford describe el concepto de confiabilidad como la cualidad de ser probablemente correcto o verdadero. El diccionario Collins también afirma que la información fiable o que procede de una fuente fiable es muy probable que sea correcta. Por lo tanto, el término confiabilidad está asociado a la probabilidad de ser creíble, a diferencia del concepto de veracidad, que alude a la certeza de serlo.

La confiabilidad permite extraer conclusiones y determinar si existe suficiente evidencia en la forma en la que se presenta una noticia como para considerarla confiable o no. Nuestro estudio defiende que la detección de indicadores lingüísticos en una noticia ayuda a medir la confiabilidad de una noticia, pues proporciona información útil únicamente en un primer nivel de anotación textual y permite conocer qué características hacen que una noticia sea confiable o no. De esa forma, se puede detectar información sospechosa y generar un informe de confiabilidad para analizar rápidamente una noticia antes de aplicar técnicas de verificación de datos.

Este primer análisis textual, además de servir de apoyo a los lectores, puede ser una herramienta de ayuda para las agencias de *fact-checking* (organizaciones que verifican la información y los datos de una noticia *a posteriori*) o incluso una herramienta de ayuda para escritores y periodistas, como guía para seguir unos estándares periodísticos que les permita comunicar una noticia de forma precisa y completa. Tanto el criterio de veracidad como el de confiabilidad permiten dar un paso más en la detección de fake news, pero con focos distintos: la veracidad se centra en el fondo de la noticia (en su verificación), mientras que la confiabilidad se centra en su forma y apariencia (análisis textual). Aunque en un inicio se experimentó con el fondo de la noticia mediante la anotación a partir de información externa, la presente tesis se centrará especialmente en la forma, pues pretende detectar cuándo una noticia tiene la apariencia de no ser confiable y así generar duda en el lector sobre su veracidad antes de que este tome la decisión de creerse o compartir la noticia.

El lenguaje desempeña un papel muy importante y existen indicadores lingüísticos clave a la hora de definir una noticia. Algunos de los indicadores que influyen en la confiabilidad de una noticia son: la ambigüedad de la información, la no confirmación de los datos, la falta de fidelidad en relación con las fuentes, la intencionalidad al ocultar información que no se quiere dar a conocer, el deseo de transmitir y dar a conocer una idea propia del comunicador disfrazada de creencia popular (Tarrés, 2000), la representatividad u opacidad del titular, las citas externas de expertos, las citas de estudios y organizaciones, el tono, las

expresiones exageradas y con carga emocional, por ejemplo de desprecio o ira, (A. X. Zhang y cols., 2018), la extensión, la puntuación o el uso de mayúsculas (Horne y Adali, 2017), los signos de puntuación en titulares (especialmente los exclamativos), los puntos suspensivos, la falta de datos y fuentes (Mottola, 2020), el léxico utilizado para exagerar (como léxico subjetivo o superlativos), las cifras o las marcas personales (como el uso de la primera o segunda persona) (Rashkin y cols., 2017).

A la hora de evaluar la confiabilidad de la noticia, el componente emocional y la burbuja informativa tienen mucho peso, como en el caso del fenómeno conocido como *bandwagon effect*, el cual alude al hecho de seguir la opinión pública, apoyar lo que hace o piensa la mayoría. El concepto metafórico de este fenómeno se remonta a finales del siglo XIX y alude al vagón que transporta a la banda durante un desfile y que atrae a una gran multitud de seguidores que los siguen para disfrutar de la música (Schmitt-Beck, 2015). Aplicado a la desinformación, alude al hecho de creer y compartir una noticia por el mero hecho de que la mayoría lo hace. La característica que define este fenómeno es que la gente lo sigue por razones superficiales y no por evidencia científica (Rikkers, 2002), siendo considerado como un factor importante en la construcción de la opinión mayoritaria (Lee, Ha, Lee, y Kim, 2018), lo que incrementa la polaridad y afecta al pensamiento crítico de la sociedad. La detección de patrones lingüísticos que reflejen la confiabilidad de una noticia permite facilitarle pistas al lector para así fundamentar objetiva y racionalmente su decisión a la hora de creerse o no la noticia, dejando a un lado el factor emocional.

1.2 Necesidades de la investigación

Internet ha alimentado la necesidad de estar continuamente informado y esa sed de información se traduce en una difusión más rápida de noticias no verificadas, ya que cualquier persona puede compartir y acceder a la información fácilmente. Por ello, el fenómeno de la desinformación se ha convertido en un reto para muchos investigadores de diferentes áreas de investigación. En PLN se aborda el problema desde varios enfoques, como el *fact-checking* automático, el análisis de sentimientos, la detección del engaño y la postura, la detección de contradicciones, la credibilidad, etc. Aunque estas líneas aborden la problemática desde distintas perspectivas, todas ellas son complementarias porque comparten objetivos y necesidades comunes (Saquete, Tomás, Moreda, Martínez-Barco, y Palomar, 2020). El problema de la desinformación no se puede resolver desde un solo enfoque, es necesario aplicar distintas soluciones para abordarlo, ya sea desde el plano comunicativo, tecnológico o lingüístico.

Una compleja mezcla de sesgos cognitivos, sociales y algorítmicos nos hace más vulnerables a creernos la desinformación en línea y a ser manipulados (Shao, Ciampaglia, Varol, Flammini, y Menczer, 2017), de ahí la necesidad de luchar contra la desinformación en el mismo ámbito en el que esta se genera: el

mundo digital. La multiplicidad de fuentes y su posible anonimato, la ausencia de estándares de calidad de la información, la facilidad para manipular y alterar la información o la falta de claridad del contexto son algunos de los factores que hacen necesaria la creación de sistemas que ayuden a detectar la confiabilidad (Viviani y Pasi, 2017). Además, el gran volumen de desinformación, así como su rápida viralización, hace imposible el procesamiento y tratamiento de datos de forma manual en el tiempo requerido. Este problema hace necesaria la automatización de tareas y el desarrollo de modelos computacionales mediante PLN, los cuales necesitan corpus de referencia anotados por expertos para aprender y entrenar, pues estos son un prerrequisito para la evaluación y entrenamiento de herramientas de vanguardia para muchas tareas de PLN (Stenetorp y cols., 2012).

La intervención de expertos en tareas de PLN, especialmente en la anotación de recursos, y el uso de algoritmos e Inteligencia Artificial (IA) son factores clave a la hora de abordar el problema de la desinformación. Por un lado, la intervención de expertos es esencial para aportar el conocimiento y los ejemplos anotados de los que va a aprender el modelo entrenado, así como para supervisar las decisiones tomadas por dicho modelo. No obstante, esta tarea es costosa y exige mucho tiempo, por lo que se vuelve igualmente necesaria la automatización o semiautomatización de aquellas tareas que un humano no puede realizar manualmente, debido al volumen, el tiempo o el coste, como es el caso de la tarea de anotación. Así pues, las necesidades de la presente investigación se resumen en:

- El modelado lingüístico de la desinformación y más concretamente de la confiabilidad en noticias.
- La generación de recursos para la tarea de detección de desinformación.
- El desarrollo de modelos computacionales usando PLN y recursos de calidad anotados.

1.2.1 Necesidad de recursos

Tal y como se refleja en las necesidades de la investigación, el uso de recursos anotados es primordial para entrenar modelos y que estos aprendan del conocimiento del experto para automatizar tareas. No obstante, uno de los desafíos en PLN es el de la escasez de corpus de entrenamiento, especialmente en tareas como la detección del engaño y en idiomas que no sean el inglés, como es el caso del español. Esto se debe a que la anotación de corpus es una de las tareas más costosas en PLN, tanto en tiempo como en dinero (Stenetorp y cols., 2012). Además de esos dos factores, se necesita conocimiento y especialización para construir un corpus eficiente, pues las anotaciones de corpus pueden presentar distintos grados de dificultad. Así pues, nos enfrentamos principalmente a dos desafíos en nuestra investigación: la escasez de corpus de calidad anotados en

español para entrenar modelos de detección de desinformación y la dificultad que requiere obtener dichos corpus.

Estas dos problemáticas han impulsado la necesidad de crear un corpus en español para la tarea teniendo en cuenta técnicas de IA para agilizar el proceso. De esta forma, se prueba la anotación diseñada en un corpus *ad hoc*, en una tarea concreta, que es la detección de desinformación en noticias, y con un objetivo específico, el de establecer patrones lingüísticos de información confiable y no confiable.

1.2.2 Soluciones tecnológicas

Actualmente, vivimos en un mundo sobrecargado de información. El Procesamiento del Lenguaje Natural es una rama que, gracias a la interacción entre la Lingüística y la Inteligencia Artificial, trabaja a fondo en el filtrado, la interpretación, el análisis y la detección de ese exceso de información. Los algoritmos son los responsables de la difusión de la desinformación, pero también de su mitigación, por lo que la causa del problema puede ser a su vez su solución.

En el terreno de las noticias falsas, la solución tecnológica propuesta en la presente tesis es la de un sistema entrenado capaz de detectar indicadores lingüísticos de desinformación que alerten a los usuarios sobre aquellos elementos que hacen que una noticia no sea confiable. El objetivo no es que el sistema decida si algo es completamente verdadero o falso, sino que oriente a los usuarios a construir su propia opinión, pero es el humano el que tiene la decisión final. Este sistema no pretende ser una herramienta categórica, sino una herramienta de apoyo, pues no se pretende reemplazar al humano ni anular su sentido crítico, sino ayudarlo en la verificación de la noticia para que, con la información adecuada, el usuario sea capaz de razonar y tomar una decisión fundamentada. Uno de los problemas actuales con el consumo de información en línea es que los usuarios no la verifican ni se documentan por falta de tiempo, ganas o recursos, siguiendo así sus convicciones, las cuales solo hacen que reafirmar las opiniones sobre un tema.

Por otro lado, dentro de la tarea de entrenamiento del modelo para la detección automática de desinformación, se pretende trabajar en una subtarea para agilizar el proceso de anotación aplicando técnicas de IA, como son el Aprendizaje Automático —*Machine Learning*— (ML) y el Aprendizaje Profundo —*Deep Learning*— (DL), las cuales se definirán en el Capítulo 2. La construcción de tecnología con IA, junto a la intervención humana, permite que las tareas manuales sean asistidas por ML. Además, esta tecnología suele depender de la retroalimentación del experto y, en este sentido, hay que destacar el vanguardista concepto de *Human-in-the-loop* (HITL), el cual permite combinar la inteligencia humana y la de la máquina con el fin de aumentar la eficiencia de la tarea, asistir al humano en tareas complejas e incrementar a su vez la precisión del modelo entrenado con más rapidez (Monarch, 2021). Las estrategias de HITL no tienen como objetivo reemplazar al experto, sino facilitar su trabajo y, tal y como

explica (Wu y cols., 2022), combinar la inteligencia humana y la computacional con el fin de entrenar un modelo de predicción preciso con un coste mínimo integrando el conocimiento y la experiencia del experto humano.

1.3 Objetivos de la investigación

El principal objetivo de la presente tesis se centra en establecer un marco de evaluación de la desinformación en textos, aplicando PLN e IA, para su futura detección automática. Para ello, la tarea principal se enfoca en el diseño de una anotación basada en los criterios de veracidad y de confiabilidad que permite anotar indicadores lingüísticos y estructurales. Otras tareas que ayudan a conseguir el propósito de nuestra investigación son:

- Adopción de un criterio de confiabilidad que permita evaluar características como el estilo, la estructura, la precisión, la objetividad, la evidencia o el componente emocional a través del análisis lingüístico.
- Diseño de una anotación lingüística y semántica para la clasificación de la desinformación en noticias, basada en dos técnicas periodísticas (Pirámide Invertida y 5W1H) y en tres niveles de anotación (Estructura, Contenido y Elementos de Interés).
- Construcción de recursos para la tarea de detección de desinformación en noticias en español.
- Anotación tanto individual (elementos estructurales y semánticos) como global (asignación del valor de veracidad/confiabilidad a la noticia en su totalidad) de los corpus generados.
- Diseño e implementación de una metodología semiautomática de anotación aplicando estrategias de Aprendizaje Activo —*Active Learning*— (AL), ML, DL y HITL para asistir al experto en la tarea de anotación e incrementar el corpus reduciendo tiempo y coste.
- Diseño de una arquitectura para la detección de la desinformación.
- Evaluación de la calidad de la anotación y de los recursos generados teniendo en cuenta métricas como la relación tiempo-esfuerzo, precisión o acuerdo entre anotadores.
- Evaluación del rendimiento de diferentes modelos entrenados con los recursos generados para detectar tanto la confiabilidad como la veracidad.

Nuestra propuesta realiza una triple contribución a la tarea de la detección de la desinformación: (i) en primer lugar, se propone una clasificación de veracidad, pero también otra de confiabilidad, siendo esta última una forma novedosa de clasificar las noticias teniendo únicamente en cuenta características

lingüísticas; (ii) en segundo lugar, según la literatura, los corpus publicados suelen anotar las noticias con un único valor global de veracidad, mientras que esta propuesta pretende anotar todas las partes estructurales y los elementos semánticos de una noticia con mayor precisión; (iii) y finalmente, esta anotación de grano fino permite crear un recurso de calidad en español teniendo en cuenta el lenguaje y la estructura de la noticia, sin recurrir al conocimiento externo, lo que serviría como paso previo y herramienta de ayuda antes del proceso de verificación de datos.

1.4 Hipótesis

A continuación, se plantean las cuatro hipótesis que han motivado esta investigación, describiendo brevemente la problemática que da pie a cada una de ellas.

1. **¿Es posible detectar el engaño a partir del lenguaje?** El lenguaje es el armazón de una noticia y la expresión de la realidad o de la falsedad tiene que materializarse en palabras para llegar al usuario. La forma de comunicar un hecho o de expresar una idea se ve reflejada a través del lenguaje y, por esa razón, nuestra primera hipótesis sostiene que un análisis puramente lingüístico, sin necesidad de recurrir a la verificación de datos, da pistas sobre qué información es sospechosa y cuál es confiable. ¿Pero puede el lenguaje proporcionar suficientes pistas para diferenciar una noticia confiable de otra que no lo es? ¿Existen patrones específicos que reflejen el lenguaje engañoso?

2. **¿La confiabilidad global de una noticia depende de la confiabilidad de cada uno de sus elementos?** La desinformación se caracteriza por mezclar datos verídicos y falsos, y tanto los elementos confiables como los que no lo son pueden encontrarse en cualquier parte de la noticia (en el titular o en la conclusión, por ejemplo) y en cualquier elemento lingüístico (como en estructuras sintácticas o en el léxico). Nuestra segunda hipótesis defiende que la anotación individual de cada uno de los elementos de contenido y estructura de una noticia por separado y la clasificación de dichos elementos en confiable o no confiable permiten conocer el valor de confiabilidad de la noticia completa, pues cada uno de esos valores influyen en su clasificación global. ¿Pero es posible detectar qué partes o elementos son los que hacen que la noticia no tenga credibilidad?

3. **¿Es posible construir recursos de calidad a partir de una anotación lingüística basada en el criterio de confiabilidad?** Aunque al inicio de la presente investigación las noticias se clasificaban según los valores de veracidad de verdadero y falso, finalmente se adoptó el criterio de confiabilidad que anota las noticias con los valores de confiable y no confiable. Este

segundo enfoque permite centrarse únicamente en el análisis lingüístico sin depender del conocimiento externo para saber si una noticia contiene información sospechosa o no. Pero, ¿este criterio permite detectar indicadores de desinformación decisivos a la hora de evaluar una noticia? ¿La anotación de la confiabilidad es útil en la tarea de detección de desinformación a pesar de no contrastar la información con técnicas de verificación de datos?

4. **¿Un sistema de anotación asistida puede detectar el lenguaje del engaño?** La presente tesis gira en torno a la aplicación de una guía de anotación en noticias, por lo que este trabajo exige anotar un corpus de noticias para evaluar la anotación. Teniendo en cuenta que las noticias tienen una extensión considerable y un estilo propio dependiendo de la fuente, autor o tema, y considerando la complejidad de la anotación, la cual se basa en varios niveles de anotación a nivel lingüístico y estructural, ¿puede un sistema semiautomático detectar esos patrones lingüísticos a pesar de la subjetividad del lenguaje y de la complejidad de la anotación? ¿Esos patrones extraídos son suficientes para considerar que el corpus es un recurso fiable y útil?

Las preguntas planteadas se irán respondiendo a lo largo de la presente tesis y en las conclusiones del Capítulo 6, tras llevar a cabo los estudios y experimentos pertinentes y evaluar los resultados obtenidos.

1.5 Estructura de la tesis

La presente tesis contiene los siguientes capítulos:

- **Capítulo. 1 Introducción**: contextualización del problema de la desinformación, las noticias falsas y la confiabilidad en noticias y presentación de las necesidades, objetivos e hipótesis de la investigación.
- **Capítulo 2. Estado del arte**: contextualización del estado del arte —*State of the Art*— (SOTA) en materia de desinformación en el ámbito comunicativo, lingüístico y tecnológico.
- **Capítulo 3. Modelado de la desinformación**: presentación de las dos guías de anotación, una en el plano de las noticias falsas y la otra en el de la confiabilidad, así como el recurso generado únicamente de forma manual.
- **Capítulo 4. Propuesta de anotación asistida**: presentación de la metodología semiautomática construida a partir de técnicas de *Human-in-the-loop* y de extracción de resúmenes, así como el recurso generado a partir de esta metodología y la evaluación en otro corpus ya existente.

- **Capítulo 5. Marco de evaluación para la detección de la desinformación**: evaluación de los dos esquemas de anotación en los dos recursos generados, arquitectura diseñada para la detección de desinformación, experimentación y resultados.
- **Capítulo 6. Conclusiones y trabajo futuro**: resumen de las conclusiones globales de la presente investigación, principales aportaciones, validación de las hipótesis, trabajo futuro y publicaciones.

Además, el presente trabajo cuenta con apartado de Agradecimientos, Resumen, Referencias bibliográficas y Apéndices.

CAPÍTULO 2

Estado del arte

El presente capítulo enmarca las principales investigaciones que se han tomado de referencia y de estudio para el diseño de nuestra propuesta, de forma que se tuviesen en cuenta los avances realizados hasta el momento y se pudiese presentar un proyecto novedoso que ayude a la comunidad científica a avanzar en el campo de la detección de la desinformación. Se contextualizará el problema y se abordará a través de dos planos: el comunicativo y el tecnológico. A través del plano comunicativo se presenta la estructura típica de una noticia tradicional, las técnicas periodísticas estándar que suelen seguir o las características lingüísticas que determinan su confiabilidad, todo ello desde un enfoque más bien periodístico. En cuanto al tecnológico, se introducen las distintas metodologías y técnicas adoptadas en PLN para abordar el fenómeno de la desinformación, como son las ramas de ML y DL o la construcción de recursos anotados.

2.1 Contextualización del problema

El fenómeno de la desinformación se ha convertido en un problema social que afecta nuestro día a día y que, de manera silenciosa, manipula nuestras opiniones, sentimientos y pensamiento crítico. Numerosas investigaciones en PLN se centran en el estudio de la detección automática de la desinformación y, para ello, se han creado corpus que hacen posible el entrenamiento con ejemplos reales anotados por humanos. Teniendo en cuenta la premisa de que las noticias falsas son noticias poco fiables que suelen mezclar información verdadera y falsa, determinar la confiabilidad de las partes esenciales por separado puede ayudar no sólo a determinar la confiabilidad global de la noticia, sino también a conocer qué partes o elementos textuales influyen a la hora de determinar la confiabilidad de una noticia.

Según (Habgood-Coote, 2019), hay tres elementos que suelen estar presentes en las definiciones de las fake news: el formato de la noticia (información fal-

sa disfrazada de noticia), el grado de falsedad (información parcial o totalmente falsa) y la intención que hay detrás (engañar a los lectores y usuarios con fines políticos o económicos). No obstante, son diversos los nombres que adopta este concepto. Hay investigadores que consideran que el término fake news es vulnerable a ser politizado y, en su lugar, recomiendan utilizar el término desinformación (Ireton y Posetti, 2018). Las razones por las que se prefiere no utilizar este anglicismo es, por un lado, porque no llega a describir el complejo fenómeno de la contaminación informativa y, por otro lado, tal y como se acaba de mencionar, porque ha sido empleado por personalidades políticas para aludir a toda aquella información con la que no estaban de acuerdo (Wardle y Derakhshan, 2017).

Así pues, (Wardle y Derakhshan, 2017) presentan un concepto más genérico, que es el del "desorden informativo", también conocido como "contaminación informativa", el cual permite clasificar los mensajes verdaderos y falsos teniendo en cuenta la intención con la que se crea o se distribuye. Estos autores clasifican el "desorden informativo" en tres nociones: *disinformation*, que es información deliberadamente falsa difundida con algún fin concreto; *misinformation*, información falsa pero transmitida con el convencimiento de su verdad; y *malinformation*, que es información verdadera de ámbito privado que se saca a la luz con la intención de hacer daño y que va en contra de la ética periodística (Salaverría y cols., 2020). Esta triple distinción de algunos textos académicos en inglés es bastante precisa, pero en español estos términos pierden sus matices, utilizando el término desinformación para cualquier tipo de desorden informativo o, si se quiere ser más concreto, empleando los términos en inglés.

Otros investigadores se centran en el término bulo (*hoax* en inglés) y lo definen como "todo contenido intencionadamente falso y de apariencia verdadera, concebido con el fin de engañar a la ciudadanía y difundido públicamente por cualquier plataforma o medio de comunicación social" (Salaverría y cols., 2020). Podríamos decir que esta definición coincide con la de las noticias falsas y la desinformación, por lo que muchas veces el concepto es el mismo, pero adopta distintos nombres en función del matiz que se le quiera dar. Otros trabajos, como el de (Habgood-Coote, 2019), deciden utilizar tanto el término fake news como el de desinformación, siendo este último más preciso a la hora de hablar de información falsa o engañosa, sin aludir al formato de las noticias.

Del mismo modo, la presente tesis utilizará principalmente el término **desinformación**, aunque en numerosos casos, para aludir con mayor precisión al formato o a las características de las noticias, se utilizará también el término **noticias falsas** o los anglicismos **fake news** y **true news** (este último, para referirse a las noticias verdaderas). Ambos anglicismos se utilizarán sin cursiva debido a su constante aparición en la tesis.

Una vez detectado el problema de la difusión de desinformación a través de noticias necesitamos trabajar en la forma de detectarla. Para ello, la presente tesis se centra en el concepto de confiabilidad, que se explica con más detalle en el apartado 2.2.4. A continuación se presenta el fenómeno desde la perspectiva co-

municativa y se introducen las características de las noticias y algunas técnicas periodísticas.

2.2 Ámbito comunicativo

El desarrollo de sistemas automáticos para la detección de noticias falsas en el contexto de esta propuesta requiere el análisis de las principales características de los artículos periodísticos, como la forma en la que están estructurados o cómo se presenta el contenido. Es importante centrarse en todo aquello que pueda servir como elemento diferenciador entre las noticias verdaderas y las falsas.

2.2.1 Estructura y contenido de las noticias

Las noticias suelen seguir una estructura propia y presentan el contenido de forma atrayente y organizada. Aunque existen diferentes formas de redactar una noticia, hay dos principios clave en los que debe basarse toda noticia bien construida: la neutralidad y la estructura de la Pirámide Invertida (Thomson y cols., 2008). Así, la objetividad de una noticia dependerá de estos dos factores, por lo que la detección de patrones que no sigan estos estándares periodísticos podría proporcionar una pista a la hora de detectar noticias falsas. En la hipótesis de la Pirámide Invertida, la información más importante se sitúa al principio, mientras que la menos relevante aparece al final (H. Zhang y Liu, 2016). Las tres partes comunes y más importantes que estructuran una noticia son el titular, la entradilla y el cuerpo de la noticia. Otros elementos estructurales importantes pero secundarios son el subtítulo o la conclusión, pues, aunque suelen aparecer en los artículos periodísticos bien construidos, no siempre están presentes.

En un artículo bien construido, las partes deben aparecer en el siguiente orden (Figura 2.1):

- Titular: el titular de la noticia proporciona la idea principal de la historia. Normalmente resume en una frase la información básica y esencial de la noticia. Su principal objetivo es captar la atención del lector.

- Subtítulo: se trata de un segundo título que presenta con más detalle, pero también de forma resumida, la información mencionada en el titular. A veces completa la información dada en el titular y otras aporta detalles no mencionados antes.

- Entradilla: es el párrafo (o párrafos) que desarrolla la información principal. Además, presenta los elementos relevantes de la noticia y simultáneamente funciona como principio de la historia (Bednarek y Caple, 2012). Toda la información principal de la noticia debe presentarse claramente en esta sección. La entradilla y el titular se consideran a veces como una unidad porque la entradilla suele repetir la idea dada por el titular, pero

con más detalle y precisión (Thomson y cols., 2008). Su objetivo es mantener la atención del lector y animarlo a seguir leyendo la noticia.

- Cuerpo: toda la información desarrollada se encuentra en esta parte de la noticia. El cuerpo de la noticia presenta con detalle todos los antecedentes, hechos, elementos y argumentos del artículo. Como mencionan (Thomson y cols., 2008), el cuerpo del texto no desarrolla nuevos significados, sino que remite al titular/entradilla mediante una serie de especificaciones.

- Conclusión: la idea principal de la historia puede resumirse en una frase o en un párrafo, pero, aunque la conclusión forme parte de una noticia bien construida, esta no siempre aparece. No presenta información novedosa, ya que es sólo un resumen.

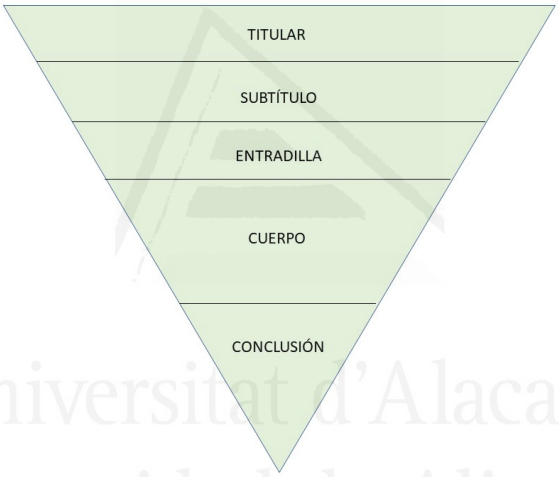


Figura 2.1: Hipótesis de la Pirámide Invertida.

Además de la estructura, el contenido de una noticia suele incluir todos aquellos elementos esenciales que permiten presentar la información de forma completa. Para ello, se suele utilizar la técnica conocida como las cinco W y la H (5W1H) (Chakma y Das, 2018; Kim, Son, y Baik, 2012; W. Wang y cols., 2010), la cual consiste en contestar a seis preguntas clave: qué (*what*), quién (*who*), dónde (*where*), cuándo (*when*), por qué (*why*) y cómo (*how*). Estas preguntas se suelen responder tanto en la entradilla, de forma introductoria, como en el cuerpo de la noticia, de forma más detallada.

- QUÉ: las circunstancias, el acontecimiento, los hechos.

- QUIÉN: las personas implicadas en los hechos.

- DÓNDE: el lugar de los hechos.
- CUÁNDO: el momento del evento.
- POR QUÉ: la razón o causa del acontecimiento.
- CÓMO: la forma en la que se desarrollan los hechos.

El método de las 5W1H es importante en la construcción de la entradilla (Chagas, 2019), la cual es una parte esencial de la noticia ya que presenta los elementos principales de un artículo: hechos, actores, lugares, tiempo, motivos y modo, respondiendo así a las seis preguntas que son clave para comunicar la información con precisión. Sin embargo, todas estas preguntas no siempre se responden en la entradilla. Basta con que sólo se respondan las dos o tres preguntas más importantes y el resto se responda con detalle en el cuerpo de la noticia.

2.2.2 Corpus basados en técnicas periodísticas

Tal y como se muestra en el apartado anterior, nuestra anotación está basada en dos conceptos periodísticos reconocidos: la Pirámide Invertida y las 5W1H. Este apartado se centra en presentar brevemente algunos corpus que también utilizan estas técnicas. Muchos de estos estudios se centran en tareas de extracción de eventos o de etiquetado de roles semánticos.

(Norambuena, Horning, y Mitra, 2020) proponen el método *Inverted Pyramid Scoring* (Clasificación de la Pirámide Invertida) para evaluar en qué medida un artículo periodístico sigue la estructura de la Pirámide Invertida mediante el uso de extracción de descriptores de eventos principales (5W1H) y resumen de noticias. Su propuesta, evaluada en un dataset formado por 65 535 artículos de *Associated Press News* (AP News), muestra que el método adoptado ayuda a distinguir las diferencias estructurales entre las noticias de última hora y las que no lo son, llegando a la conclusión de que las noticias de última hora tienden a seguir más la estructura de la Pirámide Invertida.

Otro trabajo interesante relacionado con el concepto periodístico 5W1H es el de (Chakma y Das, 2018), en el que se describe un enfoque de anotación para asignar roles semánticos. Esta propuesta no se aplica a noticias, sino a un corpus de 3000 tuits muestreados aleatoriamente y relacionados con las elecciones estadounidenses de 2016. Para anotar y extraer las respuestas a las 5W1H se utilizó un enfoque de *Question Answering* (QA), sistema que extrae respuestas a partir de grandes recolecciones de texto, clasificando la pregunta tipo y usando palabras clave o patrones asociados a las preguntas para identificar posibles respuestas (Narayanan y Harabagiu, 2004). Por otro lado, (Khodra, 2015) introduce un nuevo corpus enfocado en las 5W1H de noticias indonesias para entrenar la extracción de eventos. El corpus, que consta de 90 noticias obtenidas de sitios web de noticias populares, fue anotado manualmente por tres anotadores

humanos siguiendo el concepto de las 5W1H y extrayendo la información del evento de la noticia.

En la presente investigación se utilizan estas técnicas periodísticas para detectar partes estructurales, eventos semánticos e indicadores lingüísticos que puedan ayudar a clasificar la confiabilidad de las noticias. Los eventos semánticos quedarán identificados a través de las preguntas 5W1H (quién, qué, cuándo, dónde, por qué y cómo). En los trabajos analizados hasta el momento, la técnica de las 5W1H se ha utilizado para detectar únicamente el evento principal de una noticia, que suele encontrarse al principio de la misma, en el titular o en la entradilla. Sin embargo, en esta tesis se pretende explotar la técnica no solo para los párrafos iniciales, sino para todas las ideas presentes en la noticia. La desinformación puede encontrarse en cualquier parte de una noticia y en cualquier frase o idea de la misma, no sólo en el evento principal. El estudio de las 5W1H de todas las partes de una noticia (desde el titular hasta la conclusión) permite un análisis más profundo de todo el artículo.

2.2.3 Características lingüísticas para la detección de desinformación

En este apartado se quieren destacar algunas investigaciones relevantes en el estudio de las características lingüísticas que influyen a la hora de determinar la confiabilidad de una noticia. Este aspecto es clave para nuestra investigación porque nuestro análisis pretende encontrar indicios de desinformación a través de características lingüísticas. (A. X. Zhang y cols., 2018) presentan un corpus de 40 artículos anotados con un conjunto de indicadores, tanto de contenido como de contexto, para la detección de la credibilidad. Respecto a los indicadores de contenido, que son en los que se centra nuestro estudio, los autores introducen algunos indicadores que se pueden determinar analizando el titular y el texto del artículo sin consultar fuentes externas o metadatos. Estos indicadores son: representatividad del titular, titular de tipo *clickbait*, citas de expertos externos, citas de organizaciones y estudios, calibración de la confianza (lenguaje mediante el cual el autor muestra la confianza o incertidumbre en sus afirmaciones, como lenguaje evasivo o vago), falacias lógicas (argumentos engañosos, pobres pero tentativos, que muchas veces intentan manipular), tono e inferencia.

Por otro lado, (Horne y Adali, 2017) afirman que el estilo y el lenguaje son rasgos que permiten diferenciar artículos falsos de verdaderos. Este estudio se realiza en tres corpus distintos (con noticias verdaderas, falsas y de sátira) y con base en tres categorías de características de contenido: estilística, de complejidad y psicológica. Al estudiar las similitudes entre las noticias, muestran que hay una notable diferencia en los titulares y el contenido entre las noticias falsas y las verdaderas en cuanto a longitud, puntuación, citas, rasgos léxicos o palabras en mayúsculas. Para las características estilísticas y psicológicas, los autores utilizaron *Linguistic Inquiry and Word Count* (LIWC) (Pennebaker, Boyd, Jordan, y Blackburn, 2015), que es un programa de análisis de texto que cuenta las pala-

bras en categorías con significado psicológico y que está disponible en diferentes idiomas.

Otro estudio que demuestra que las características lingüísticas pueden ayudar a determinar la veracidad del texto es el de (Rashkin y cols., 2017). Este trabajo compara el lenguaje utilizado en noticias verdaderas con el utilizado en noticias de sátira, bulos y propaganda. Para analizar los patrones lingüísticos, seleccionaron una muestra de artículos de noticias fiables estándar del corpus *English Gigaword* y recopilaron mediante *crawling* artículos de siete páginas de noticias no fiables diferentes.

(Mottola, 2020) también lleva a cabo un estudio comparativo entre el italiano y el español con el fin de identificar características textuales comunes de la desinformación digital. Para ello, introduce un corpus formado por noticias falsas publicadas en plataformas digitales tanto por usuarios italianos como españoles y reconocidas como falsas por dos conocidas agencias de *fact-checking*: Bufale un Tanto Al Chilo[1] y Maldita[2]. A través de este análisis lingüístico se demuestra que existen diversas características que comparten las noticias falsas relacionadas con los titulares, la puntuación, las mayúsculas, la falta de datos o la carga emocional.

2.2.4 Confiabilidad desde el punto de vista periodístico

En este apartado se quiere proporcionar un enfoque general desde el punto de vista periodístico y comunicativo, de forma que se tenga una referencia de qué criterios suelen seguir los profesionales del periodismo a la hora de considerar la confiabilidad de una noticia.

Aunque no existe una única forma de ejercer la práctica periodística, existen unos principios básicos que ayudan a crear contenido de calidad. (Ireton y Posetti, 2018) afirman que el papel distintivo del periodismo reside hoy en su capacidad para aportar claridad y generar confianza en torno a contenidos verificados. Cita siete principios básicos: precisión, independencia, imparcialidad, confidencialidad, humanidad, responsabilidad y transparencia.

Como comentábamos en el apartado 1.1.3 y teniendo en cuenta la literatura consultada, el concepto de veracidad, el más utilizado en la clasificación de noticias, se relaciona con el proceso de verificación de información. De hecho, (Vosoughi y cols., 2018) argumentan que la veracidad es una característica clave en la difusión de noticias y plantean que la única forma de estudiar de manera sólida las noticias verdaderas y falsas es mediante la verificación de las mismas a través de múltiples organizaciones independientes de verificación de hechos. Por otro lado, el concepto de confiabilidad suele estar más vinculado a la credibilidad de las fuentes de donde se obtienen las noticias. No obstante, estos mismos autores opinan que depender de fuentes confiables no es un proceso

[1] https://www.butac.it/
[2] https://maldita.es/

objetivo y supone un problema debido a la politización del significado y clasificación de fuentes confiables, no hay acuerdo sobre qué fuentes son fiables. Una métrica sólida como la credibilidad de las fuentes se ha visto afectada por la polarización de la sociedad, causando que las fuentes tiendan más a apoyar una postura u otra.

(Appelman y Sundar, 2016) sugirieron que hay diez indicadores que permiten detectar con más probabilidad la confiabilidad de un mensaje y estos se dan cuando el mensaje es: completo, conciso, coherente, bien presentado, objetivo y representativo; cuando no contiene giros, utiliza fuentes expertas, se percibe que tiene un impacto y es profesional. Por otro lado, (Hinsley y Holton, 2021) mencionan que la exhaustividad, la precisión o las medidas de credibilidad de una fuente (como autoría y confiabilidad) son factores que pueden influir en la confianza del usuario al tomar la decisión sobre si una noticia es falsa o no.

Así pues, el concepto de confiabilidad se puede enfocar desde el factor de la fuente, que es como suele utilizarse en muchas investigaciones, aunque también se puede enfocar desde otros factores más relacionados con la práctica periodística como son la objetividad, la precisión o el uso de fuentes expertas. En nuestro caso, enfocaremos el concepto de confiabilidad en estos últimos rasgos, en las características lingüísticas que hacen que una noticia sea comunicada de forma completa y precisa.

2.3 Ámbito tecnológico

En este apartado se aborda el problema de la desinformación desde el enfoque tecnológico, el cual es esencial para llevar a cabo la automatización de la detección de este problema. No obstante, son varios los enfoques que se pueden tener en cuenta a la hora de detectar las fake news. (X. Zhou y Zafarani, 2020) plantean la detección de las noticias falsas desde cuatro perspectivas: (i) Métodos basados en el conocimiento *(Knowledge-based)*, que detectan las noticias falsas verificando si el conocimiento del contenido de la noticia (texto) es consistente con los hechos; (ii) Métodos basados en el estilo *(Style-based)*, relacionados con la forma en la que están escritas las noticias falsas (por ejemplo, si contienen una alta carga emocional); (iii) Métodos basados en la propagación *(Propagation-based)*, en los que se detectan las noticias falsas en función de cómo se propagan en línea; y (iv) Métodos basados en la fuente *(Source-based)*, que detectan las noticias falsas investigando la credibilidad de las fuentes de las noticias en varias etapas (creación, publicación en línea y difusión en las redes sociales). Esta tesis se centra en el segundo enfoque, el método basado en el estilo, pues se estudia la forma en la que se redactan y estructuran las noticias desde el punto de vista lingüístico.

Es importante entender asimismo los cuatro conceptos clave que se tratan en esta sección y en los que se enmarca la presente tesis:

- **Inteligencia Artificial**: es la capacidad de las máquinas para usar algorit-

mos, aprender de los datos y utilizar lo aprendido en la toma de decisiones tal y como lo haría un ser humano, pero procesando grandes volúmenes de información a la vez (Rouhiainen, 2018).

- **Procesamiento del Lenguaje Natural**: rama de la Inteligencia Artificial que consiste en la utilización de un lenguaje natural para la comunicación con la computadora, facilitando el desarrollo de programas que realicen tareas relacionadas con el lenguaje o bien desarrollando modelos que ayuden a comprender los mecanismos humanos relacionados con el lenguaje (Vásquez, Quispe, Huayna, y cols., 2009).

- ***Machine Learning***: es una rama evolutiva de los algoritmos computacionales diseñada para emular la inteligencia humana aprendiendo del entorno que la rodea mediante un proceso computacional que utiliza datos de entrada para realizar una tarea deseada sin estar literalmente programado (codificado) (El Naqa y Murphy, 2015). Esta rama es necesaria para que los ordenadores realicen sofisticadamente la tarea sin intervención de los seres humanos mediante el aprendizaje y la experiencia en constante aumento para comprender la complejidad del problema y la necesidad de adaptabilidad (Alzubi, Nayyar, y Kumar, 2018).

- ***Deep Learning***: es una categoría de métodos pertenecientes al campo del ML, que se basan principalmente en el uso de tipos específicos de redes neuronales artificiales, a veces con un número muy elevado de capas y nodos (Visvikis, Cheze Le Rest, Jaouen, y Hatt, 2019). El DL permite a los modelos computacionales compuestos de múltiples capas de procesamiento aprender representaciones de datos con múltiples niveles de abstracción (LeCun, Bengio, y Hinton, 2015).

2.3.1 Fake news utilizando Procesamiento del Lenguaje Natural

Teniendo en cuenta que la información digital se difunde de forma exponencial, los enfoques de PLN y ML desempeñan un papel fundamental en la detección de noticias falsas (Dale, 2017). Dado que evaluar la veracidad de una noticia es una tarea compleja técnicamente, la comunidad científica la está abordando desde diferentes perspectivas (Saquete y cols., 2020). Las investigaciones actuales de detección de noticias falsas se han enfocado en el tratamiento de las noticias siguiendo distintas aproximaciones: (i) las aproximaciones lingüísticas o basadas en contenido, donde se analiza el contenido léxico, sintáctico y semántico de la noticia en su conjunto; (ii) las aproximaciones de redes o basadas en contexto, orientadas a analizar el contexto digital, como redes de conocimiento, identificación de fuentes (Conroy, Rubin, y Chen, 2015), usuarios o difusión; (iii) las aproximaciones de verificación o basadas en conocimiento, que utilizan fuentes externas para verificar si una noticia es verdadera o falsa (Seddari y cols., 2022); o (iv) las aproximaciones híbridas, que pueden combinar los enfoques anteriores.

Aproximaciones basadas en contenido

Las aproximaciones basadas en contenido (*content-based approaches*) aplicadas a la detección de noticias falsas utilizan diversas técnicas de ML, redes neuronales artificiales y PLN para analizar el contenido de la noticia y clasificar el artículo como falso o verdadero (Bani-Hani, Adedugbe, Benkhelifa, Majdalawieh, y Al-Obeidat, 2020). La detección de noticias falsas se centra actualmente en el estudio de aspectos lingüísticos de la falsedad mediante la identificación de diferentes tipos de características textuales de las noticias falsas. (L. Zhou y Zhang, 2008) propusieron un sistema con las clases de características, como cantidad de información, complejidad del lenguaje, expresividad, contenido del mensaje, como n-gramas o afecto (emociones positivas o negativas), etc. (Pérez-Rosas, Kleinberg, Lefevre, y Mihalcea, 2017) describieron un conjunto similar de características, agrupadas por categorías generales, como n-gramas, puntuación, características psicolingüísticas, legibilidad y sintaxis. En este tipo de estudios es muy común utilizar las características de LIWC (Newman, Pennebaker, Berry, y Richards, 2003), como en el caso de (Almela, Valencia-García, y Cantos, 2013), que presentan un clasificador automático para la detección del engaño en textos escritos en español, o como en la investigación de (Mihalcea y Strapparava, 2009), la cual se centra en la identificación de lenguaje engañoso en textos escritos mediante la detección de patrones de palabras destacadas que permiten diferenciar los textos engañosos.

Por otro lado, la estilometría es la aplicación del estudio del estilo lingüístico generalmente al lenguaje escrito. En cuanto a la detección automática de fake news, (Potthast, Kiesel, Reinartz, Bevendorff, y Stein, 2018) utilizaron la estilometría, combinando características de estilo de redacción (como n-gramas, *stop words* y partes del discurso) y otras características específicas del dominio de las noticias mediante el uso de los diccionarios *General Inquirer* (Stone, Dunphy, y Smith, 1966), un conjunto de procedimientos que permite identificar patrones recurrentes dentro de la variedad de comunicaciones escritas y habladas.

(Afroz, Brennan, y Greenstadt, 2012) también utilizaron la estilometría para detectar el engaño estilístico en documentos escritos. Seleccionaron más de 700 características (léxicas, sintácticas, de contenido específico, de complejidad gramatical y de vocabulario, etc.) y utilizaron tres conjuntos de características para identificar el engaño estilístico: (i) conjunto de características léxicas, sintácticas y de contenido propias del estilo del autor; (ii) conjunto de características eficaces en la detección de la mentira (como la complejidad del vocabulario o gramatical, la especificidad o la expresividad); (iii) conjunto de características de atribución de autoría, entre ellas recuento de frases, índices de legibilidad o media de sílabas por palabra. La principal conclusión que se extrajo fue que, aunque no se pueda detectar al autor de un documento cuya autoría ha sido plagiada, existen rasgos lingüísticos que cambian cuando las personas ocultan su estilo de escritura y que, al identificar dichos rasgos, se puede reconocer el engaño estilístico.

Aproximaciones basadas en contexto y en conocimiento

En cuanto a las aproximaciones basadas en contexto (*context-based approaches*), estas se centran en la información contextual de las noticias, como son los usuarios, las interacciones sociales y mensajes generados por estos o las redes (Bani-Hani y cols., 2020). (Shu, Wang, y Liu, 2019) propusieron una técnica que explota las relaciones entre editores, noticias y usuarios para predecir las noticias falsas mediante un clasificador lineal y asignaron a cada usuario una puntuación de credibilidad basada en su comportamiento en línea, donde una puntuación de credibilidad baja se correlacionó con las noticias falsas. En el proceso de viralización de una noticia en medios sociales, se contempla también la información contextual relativa a la relación entre editores, noticias y usuarios, la cual es útil a la hora de predecir las noticias falsas (Shu y cols., 2019).

El contexto es a su vez crucial en las tareas de verificación de datos, pues los sistemas necesitan localizar las fuentes necesarias para predecir la veracidad de la etiqueta (Vlachos y Riedel, 2014). En estas tareas se suele acudir igualmente a las aproximaciones basadas en conocimiento (*knowledge-based approaches*), las cuales utilizan fuentes externas para verificar la información y clasificar su veracidad. Tal y como afirman (Shu, Sliva, Wang, Tang, y Liu, 2017), las características textuales no suelen ser suficientes para detectar noticias falsas, por lo que es preciso recurrir a una base de conocimiento mediante este tipo de aproximaciones.

Aproximaciones híbridas

Finalmente, las aproximaciones híbridas (*hybrid approaches*) permiten combinar varios de los enfoques anteriores, especialmente el contenido con el contexto o conocimiento externo, para poder mejorar el modelo de detección de noticias falsas. En ese sentido, (Volkova, Shaffer, Jang, y Hodas, 2017) consideran que la incorporación de características lingüísticas y de red aumentan el rendimiento del modelo, por lo que presentan una técnica que clasifica publicaciones sospechosas mediante modelos que aprenden del contenido y de las interacciones de las redes sociales (contexto). Siguiendo la misma línea, (Ruchansky, Seo, y Liu, 2017) sostienen que no solo es importante considerar el texto de la noticia, sino también la respuesta que recibe un artículo o los usuarios de los que procede y presenta un modelo que combina las tres características.

Asimismo, (Seddari y cols., 2022) afirman que el uso de una sola técnica para detectar contenidos falsos no puede alcanzar el nivel de eficacia requerido, por lo que propone un sistema híbrido de detección de noticias falsas que combina tanto características lingüísticas de contenido (titular, número de palabras, legibilidad, diversidad léxica y sentimiento) como características basadas en conocimiento (reputación, comprobación de hechos y cobertura).

Técnicas utilizadas

Tanto las aproximaciones de contenido como las de contexto aplican diferentes técnicas de ML y DL. La mayoría de los sistemas citados anteriormente utilizan ML como sistema de detección, y concretamente Máquinas de Soporte Vectorial —*Support Vector Machine*— (SVM), en la mayoría de los casos. En los últimos años también se han incorporado sistemas basados en DL en general, los cuales utilizan sistemas abiertos como *Bidirectional Encoder Representations from Transformers* (BERT) (Devlin, Chang, Lee, y Toutanova, 2018). Del análisis de la literatura se concluye que la combinación de rasgos lingüísticos con enfoques de ML o DL obtiene algunos resultados interesantes, pero parecen alcanzar el límite en términos de rendimiento. Esto sugiere que las metodologías híbridas que combinan estos enfoques de contenido con información de contexto o de conocimiento podrían proporcionar una estrategia para mejorar el rendimiento. Además, las aproximaciones de ML o DL se comportan a menudo como cajas negras, lo que dificulta la explicabilidad de los modelos generados. El uso de distintos modelos de aprendizaje (*ensemble learning*) puede incrementar el rendimiento, especialmente cuando se agregan varios modelos de bajo rendimiento, o modelos con diferentes espacios de hipótesis. Por ejemplo, un modelo basado en características lingüísticas y otro basado en conocimiento externo.

2.3.2 Corpus anotados para la detección de desinformación

El objetivo de esta sección es presentar la metodología que siguen actualmente los corpus en PLN y la estructura que suelen presentar los corpus más relevantes utilizados por los sistemas de detección de noticias falsas. Dado que nuestro objetivo es estudiar qué tipo de anotación se está utilizando actualmente, hemos analizado los corpus presentados en la literatura, aunque su idioma sea principalmente el inglés.

Metodologías para la construcción de corpus

De acuerdo con la literatura consultada, la construcción de corpus en PLN puede abordarse desde distintas metodologías. Aunque hay casos en los que tanto la tarea de compilación como la de anotación están completamente automatizadas (Abacha, Dinh, y Mrabet, 2015) o se llevan a cabo de manera manual (Evrard, Uro, Hervé, y Mazoyer, 2020), muchos corpus creados para esta tarea utilizan una metodología semiautomática para la compilación de los datos del corpus, pero no para la anotación. En este enfoque, la recopilación de datos se suele realizar de forma automática mediante *Application Programming Interfaces* (APIs) de medios sociales o páginas de verificación de datos, o mediante *web crawling* y *web scraping*, dos procesos de navegación web que permite recoger contenido de forma automática de páginas web. Tras la obtención automática de los datos, la anotación suelen realizarla expertos de forma manual, como es el caso de los corpus publicados por (Shahi y Nandini, 2020; W. Y. Wang, 2017).

Aunque la anotación manual permite obtener ejemplos de calidad, creados y verificados por expertos, se trata de un proceso arduo, por lo que los corpus obtenidos suelen ser de tamaño más reducido y, por tanto, se necesita más tiempo para conseguir el objetivo deseado.

Una alternativa a la metodología semiautomática es la práctica de *crowdsourcing* (Mitra y Gilbert, 2015; Färber, Burkard, Jatowt, y Lim, 2020; Pérez-Rosas y Mihalcea, 2015), la cual permite la externalización de trabajo y la subcontratación masiva de múltiples tareas de etiquetado, normalmente con un bajo coste global y una rápida realización (Hsueh, Melville, y Sindhwani, 2009). Esta práctica permite acelerar la producción de extensos corpus de entrenamiento, pero la calidad no suele ser la misma que presentan aquellos corpus desarrollados específicamente por grupos de expertos que trabajan en el mismo campo y que cooperan en el mismo grupo de investigación. Además de los corpus compilados semiautomáticamente y de *crowdsourcing* como técnicas para acelerar la producción de corpus, cabe mencionar los corpus construidos con aprendizaje supervisado, que es una de las herramientas más eficaces a la hora de automatizar tareas cognitivas complejas (Cañizares-Díaz y cols., 2021). Esta metodología permite crear recursos de calidad supervisados por expertos humanos, de forma que se obtenga un corpus considerable a partir de la automatización al mismo tiempo que se conserva la calidad del proceso humano. En este caso, el sistema toma decisiones de forma automática pero bajo la supervisión del experto que corrige, valida o rechaza dichas decisiones. (Feller y cols., 2018) aceleran el proceso de anotación manual mediante la aplicación del aprendizaje semisupervisado, pues afirman que el excesivo esfuerzo humano que se requiere para compilar un corpus anotado de tamaño considerable supone una barrera para el uso de las redes neuronales profundas.

Corpus anotados para la detección de desinformación

Respecto a la anotación de corpus para la tarea de detección de desinformación, (Vlachos y Riedel, 2014) son los primeros en publicar un corpus de detección de noticias falsas y comprobación de hechos que incluye 221 declaraciones recogidas de PolitiFact[3] y Channel 4[4]. Las declaraciones se anotan utilizando una escala de cinco valores: verdadero, mayormente verdadero, medio verdadero, mayormente falso y falso (*true, mostlytrue, halftrue, mostlyfalse and false*). Otros dos corpus a destacar que se centran en la detección del engaño son el corpus LIAR, que contiene 12 836 declaraciones breves del mundo real (W. Y. Wang, 2017), y el corpus EMERGENT (Ferreira y Vlachos, 2016), formado por 300 declaraciones y 2595 artículos de noticias asociados. Respecto a la anotación, el corpus LIAR presenta una escala de seis etiquetas de grano fino (*pants-fire, false, barely-true, half-true, mostly-true and true*), mientras que el corpus EMERGENT clasifica las noticias en tres valores de veracidad (*true, false and unveri-*

[3] http://www.politifact.com/
[4] http://blogs.channel4.com/factcheck/

fied) y asigna una etiqueta de postura al titular con respecto a la afirmación (*for, against and observing*).

(Pérez-Rosas y cols., 2017) se centraron en identificar automáticamente contenidos falsos en noticias en línea e introdujeron dos nuevos corpus de noticias que cubren siete dominios, uno obtenido a través de *crowdsourcing* (240 noticias legítimas y 240 noticias falsas) y otro obtenido a partir de fuentes web (100 noticias legítimas y 100 noticias falsas). Además, analizaron las diferencias lingüísticas entre los artículos de noticias legítimas y falsas. Para evaluar la capacidad humana de detectar noticias falsas y la precisión de su sistema, crearon una interfaz de anotación y pidieron a los anotadores que etiquetaran los corpus desarrollados eligiendo entre *fake* o *legitimate*, según sus percepciones. Su sistema funcionó bien, pues incluso superó el trabajo humano.

BuzzFeedNews[5] es un corpus compuesto por una muestra de noticias publicadas en Facebook de 9 agencias de noticias durante una semana próxima a las elecciones estadounidenses de 2016. Cabe destacar también el corpus de noticias falsas de Kaggle (*Kaggle Fake News dataset*), proporcionado por la competición Kaggle[6], que es una plataforma popular con excelentes recursos para aquellos que quieren aprender ML e incluso ciencia de datos. El corpus de Kaggle contiene artículos de noticias falsas y verdaderas en inglés de 2015 a 2018, así como texto y metadatos de 244 sitios web, que representa 12 999 publicaciones en total. También podemos destacar el *CLEF-2021 CheckThat! Lab: Task 3 on Fake News Detection* (Shahi, Struß, y Mandl, 2021), competición enfocada en evaluar la tecnología que permite, especialmente en esta tarea y en particular en la subtarea 3A, la detección automática de la veracidad de la noticia. Teniendo en cuenta el texto y el titular de una noticia, el objetivo era predecir si la afirmación principal del artículo era verdadera, parcialmente verdadera, falsa u otra. A los participantes se les proporcionó un corpus de entrenamiento compuesto por 900 artículos de noticias, dejando 354 artículos para el conjunto de entrenamiento.

Por otro lado, la pandemia ha causado un aumento en la viralización de la desinformación, lo que ha propiciado nuevas líneas de investigación y corpus centrados en el covid-19. Existen dos corpus recientes que abordan este tema: un corpus de noticias falsas que consta de 10 700 noticias falsas y verdaderas anotadas como *real* o *fake* (Patwa y cols., 2021) y un extenso corpus de noticias falsas de covid-19 en Twitter, *COVID-19 Twitter Fake News dataset* (CTF), de (Paka, Bansal, Kaushik, Sengupta, y Chakraborty, 2021), que trabaja con tuits anotados y no anotados con dos tipos de etiquetas: *fake* y *genuine*.

[5] https://github.com/BuzzFeedNews/2016-10-facebook-fact-check/tree/master/data

[6] https://www.kaggle.com/

Corpus creados para otros idiomas distintos al inglés

En cuanto a los corpus en otros idiomas, los recursos en español son escasos, lo que acentúa la necesidad de entrenar en este idioma. (Posadas-Durán, Gómez-Adorno, Sidorov, y Escobar, 2019) proponen un corpus para el estudio de la detección automática de noticias falsas en español, que consta de 491 noticias verdaderas y 480 noticias falsas anotadas con dos etiquetas: *real* y *fake*. Este corpus se ha utilizado para algunos experimentos de la presente tesis con nuestra guía de anotación. (Almela y cols., 2013) presentaron un corpus de textos escritos en español sobre tres temas diferentes: opiniones sobre la adopción homosexual, opiniones sobre la tauromaquia y sentimientos hacia un mejor amigo. Recopilaron 100 declaraciones verdaderas y 100 falsas para cada tema.

En cuanto a los corpus de desinformación en otros idiomas, (Silva, Santos, Almeida, y Pardo, 2020) presentaron un corpus (Fake.Br) de noticias verdaderas y falsas anotadas en portugués, compuesto por 7200 noticias (3600 falsas y 3600 legítimas). Durante la construcción del corpus, para cada noticia falsa se obtuvo la noticia verdadera correspondiente, relacionada temáticamente, obteniendo así un corpus de noticias verdaderas y falsas alineadas. Por otro lado, (Assaf y Saheb, 2021) presentaron un corpus novedoso de noticias falsas en árabe que contiene 323 artículos (100 noticias clasificadas como *reliable* y 223 como *unreliable*) y que se centra en características lingüísticas tradicionales. Las noticias fueron recopiladas manualmente por periodistas y anotadas por dos expertos humanos, cuyo acuerdo se midió mediante el Coeficiente kappa de Cohen.

Además, para estudiar las diferencias entre los artículos de noticias de fuentes fiables y no fiables, (Gruppi, Horne, y Adali, 2018) construyeron dos corpus de artículos de noticias de política pertenecientes tanto a fuentes estadounidenses (2841 noticias de 16 fuentes) como a fuentes brasileñas (5511 noticias de 19 fuentes). Para cada artículo, calcularon cada característica del titular y del cuerpo del texto por separado a partir de un conjunto de características significativas en ambos idiomas y asignaron una clase *reliable (R), unreliable (U), satire (S)* en función de la fuente de la que se recopiló el artículo. Después, crearon un conjunto de características aproximadamente equivalentes en ambas lenguas y las clasificaron en 4 categorías (complejidad, estilo, lingüística y psicológica). Este estudio les permitió demostrar que existen diferencias entre los artículos de noticias procedentes de fuentes fiables y no fiables. Lo que diferencia a estos dos corpus del resto es su clasificación en confiable y no confiable, una anotación que se acerca más a nuestro enfoque y que apenas se utiliza.

Resumen de las características generales de los corpus analizados

Una gran parte de los corpus actuales creados para esta tarea utilizan una clasificación binaria basada en los valores de Falso (*fake*) y Verdadero (*true*) (Salem, Al Feel, Elbassuoni, Jaber, y Farah, 2019; Silva y cols., 2020). Otros, como aquellos que se centran en tareas de verificación de datos, utilizan escalas que cubren va-

rios grados de veracidad (W. Y. Wang, 2017; Vlachos y Riedel, 2014). No obstante, aunque algunos corpus están anotados con una clasificación basada en matices graduados de veracidad, en todos los casos la anotación es de la unidad textual completa y no de las partes que la componen. La mayoría clasifica y anota las noticias con un único valor de veracidad global sin tener en cuenta la veracidad o confiabilidad de las distintas partes del documento, tanto a nivel estructural como lingüístico.

Esta única clasificación global de las noticias, ya sea con valores binarios o múltiples, depende de conocimiento externo, como el de los desmentidos de las agencias de *fact-checking*. Pocos corpus usan una clasificación de confiabilidad y esta normalmente se establece con base en la credibilidad de la fuente y utilizando plataformas de *fact-checking* (Dhoju, Main Uddin Rony, Ashad Kabir, y Hassan, 2019; Khalil, Jarrah, Aldwairi, y Jararweh, 2021). Esta credibilidad se mide en función de criterios de contenido, como la manipulación, fabricación, falso contexto, etc. (Assaf y Saheb, 2021), o únicamente en función del contenido textual y según un conjunto de criterios predefinidos de credibilidad como son información incompleta o incorrecta, manipulación de fotos, titulares engañosos, identidad de la fuente o relación con preguntas como quién, dónde, cuándo, por qué y cómo (Hammad y Hemayed, 2013). Así pues, aunque ya se emplean enfoques de contenido, semánticos o textuales a la hora de analizar la confiabilidad de las noticias, estos se emplean por separado o utilizando plataformas de verificación de datos. Lo que propone nuestra investigación es la combinación de características semánticas, estructurales y lingüísticas para la detección de la confiabilidad, todas al mismo tiempo y sin utilizar la verificación externa.

Finalmente, los corpus enfocados en la tarea de desinformación en español (Posadas-Durán y cols., 2019) son escasos, pues normalmente se publican en inglés. Por ello, la presente tesis se enfoca en la creación de dos recursos en español para entrenar en la tarea de detección automática de desinformación. Estos recursos presentan dos ventajas: la clasificación individual y global de la noticia tanto de veracidad como de confiabilidad, siendo la anotación de confiabilidad la más novedosa por basarse únicamente en un análisis lingüístico sin recurrir a conocimiento externo; y la anotación multinivel, tanto a nivel estructural como de contenido semántico. La complejidad del análisis lingüístico incorporado en nuestro enfoque permite un análisis profundo de todos los elementos de una noticia, así como la detección de patrones de información sospechosa que permiten aconsejar al usuario antes de que este tome su decisión final.

2.4 Conclusiones

El fenómeno de la desinformación se ha convertido en un problema social y en un desafío para los investigadores de PLN. Son diversos los términos que se utilizan para aludir a este fenómeno, desde “desorden informativo” para abordar el problema de forma más genérica, a “bulos” o “noticias falsas” cuando

se le quiere dar un matiz más específico. En este capítulo se da una amplia visión del fenómeno a través de dos planos: el comunicativo y el tecnológico. El enfoque comunicativo y periodístico desempeña un papel clave en la presente tesis, pues el objeto de estudio es la noticia, lo que requiere una investigación en profundidad de su estructura y contenido semántico, tanto a nivel periodístico como lingüístico. Por un lado, es importante llevar a cabo un análisis periodístico para contemplar todos los elementos que forman parte de una noticia, las técnicas periodísticas más comunes y los principios básicos de veracidad y confiabilidad y, por otro lado, un análisis lingüístico que permita analizar aquellos indicadores propios del lenguaje de las noticias, tanto de las verídicas como de las falsas.

En este sentido, la investigación se ha centrado en un primer lugar en estudiar el esqueleto de las noticias mediante dos técnicas periodísticas reconocidas que permiten analizar tanto la estructura (técnica de la Pirámide Invertida) como el contenido semántico (técnica de las 5W1H). Como se ha podido observar en este capítulo, numerosos estudios se han centrado en estas técnicas periodísticas aplicadas a tareas de verificación de datos, etiquetado de roles semánticos o tareas de clasificación.

Respecto a la parte lingüística de la investigación, este plano es sumamente importante a la hora de modelar un lenguaje de la mentira y definir indicadores lingüísticos que influyan en la clasificación de la veracidad o confiabilidad de las noticias. Como se muestra en este capítulo, varias investigaciones han demostrado que existe una relación entre el lenguaje y la falsedad en las noticias, destacando indicadores lingüísticos que pueden ser clave a la hora de detectar el engaño, como son la construcción del titular, la puntuación, la carga emocional o la evidencia científica.

Finalmente, el otro plano esencial en la investigación es el del enfoque tecnológico, pues es el que permite abordar la tarea de detección automática de la desinformación. No existe una única manera de abordar el problema de la desinformación, pues se puede contemplar teniendo en cuenta distintos enfoques: de contenido (características textuales), de contexto (credibilidad de la fuente de la que procede o forma de propagarse), de conocimiento (verificación de la información) o híbrido (combinación de varios de los enfoques mencionados). Independientemente del enfoque desde el que se aborde, la investigación en PLN y el uso de técnicas de ML desempeñan un papel esencial en esta tarea, al igual que la generación de corpus para entrenar, que suelen anotarse generalmente de forma automática o manual.

Según la literatura consultada, los recursos en español generados para esta tarea son escasos, siendo el inglés el idioma en el que más recursos se generan. Otra observación es que la mayoría de las investigaciones construyen corpus anotados con un valor de veracidad global de toda la noticia, sin la anotación detallada por partes. Respecto a la anotación de dichos corpus, la falta de un consenso en la anotación de la veracidad o la confiabilidad de la noticia hace necesaria la creación de un estándar de anotación.

Teniendo en cuenta que el formato de una noticia es extenso, contiene mucha información textual, sigue una estructura particular y comunica la información siguiendo unos criterios concretos, nuestra propuesta aborda la anotación de la noticia de una forma más detallada y profunda, proporcionando no solo el valor global de la misma, sino también el individual por partes, y apoyándose en las dos técnicas periodísticas mencionadas para fundamentar si una noticia está bien construida y sigue los criterios periodísticos. A continuación, el Capítulo 3 se centrará en presentar las dos guías de anotación creadas para la detección de desinformación, así como los dos recursos en español diseñados *ad hoc* para la presente investigación.

CAPÍTULO 3

Modelado de la desinformación

En tiempos turbulentos, la desinformación se convierte en un gran enemigo. (Shu, Wang, Lee, y Liu, 2020) definen la desinformación como información falsa o inexacta que se difunde intencionalmente para confundir y/o engañar. Cuando se trata de temas políticos, sociales y de salud, factores como el desorden, el miedo y los intereses económicos o ideológicos aumentan el volumen de desinformación. Este problema social forma parte de nuestras vidas hasta tal punto que se crean términos específicos para referirse a él, como en el caso de infodemia (definido en la sección 1.1 como el exceso de información falsa durante un brote de enfermedad) o posverdad, definida por la Real Academia Española[1] como distorsión deliberada de una realidad, que manipula creencias y emociones con el fin de influir en la opinión pública y en actitudes sociales.

Con el fin de dar un paso más en la tarea de la detección de desinformación, el presente trabajo aborda el problema mediante la creación de una anotación de grano fino capaz de considerar la calidad de la información distribuida por los medios. El modelado de la desinformación en este trabajo comenzó siendo un modelado enfocado en las fake news y en el criterio de veracidad, del que surgió la idea inicial de anotación así como un recurso siguiendo la guía diseñada. En un paso posterior, esta guía inicial evolucionó y se orientó a la anotación de indicadores de confiabilidad a nivel lingüístico, de forma que contemplase el concepto de confiabilidad explicado en el punto 1.1.3. De esta segunda versión se generó a su vez otro recurso en español.

3.1 Modelado de fake news

Al comienzo de la presente tesis, el fenómeno de las fake news estaba en su apogeo: las noticias se difundían con facilidad y apenas se filtraban o verificaban. Sin embargo, con el tiempo, el control de la desinformación se fue hacien-

[1] https://dle.rae.es/posverdad

do cada vez más estricto mediante agencias de verificación de datos[2], secciones de verificación propuestas por diarios digitales[3] e incluso cursos de formación para concienciar a la población de la importancia de educar sobre la desinformación[4]. La evolución de este fenómeno ha afectado a nuestra investigación, cambiando nuestros objetivos y metodología. En este primer apartado, el del modelado de fake news, hablamos de nuestro enfoque inicial de anotación, el cual se basa en la comparación de la información entre las noticias recopiladas y sus respectivos desmentidos para asignarle un valor de veracidad con base en la información externa.

3.1.1 Esquema de anotación: FNDeepML

Para la presente investigación, se empezó a diseñar un esquema *ad hoc* enfocado en la anotación de noticias y denominado FNDeepML *(Fake News Deep Markup Language)*, el cual fue diseñado y aplicado por una anotadora con formación lingüística en traducción e interpretación, que es la propia autora de la presente investigación.

La guía FNDeepML (Apéndice A) permite la anotación de la estructura y del contenido de cada noticia siguiendo las técnicas periodísticas de la Pirámide Invertida y las 5W1H. Todos los elementos se anotaron como Verdadero (*True*) o Falso (*False*) según la información proporcionada por el desmentido, o como Desconocido (*Unknown*) si el desmentido no contenía información que contrastara los datos. La etiqueta Desconocido no significa que la información sea Verdadera o Falsa, sino simplemente que no puede verificarse. Además, es importante destacar que, para esta primera guía, las noticias se anotaron según la información contrastada por el desmentido, no se realizó ninguna otra verificación en fuentes externas. Todos los desmentidos pertenecían a reconocidas agencias españolas de *fact-checking* pertenecientes a la Red Internacional de Verificación de Datos —*International Fact Checking Network*— (IFCN)[5].

A continuación se describen los dos niveles de anotación de la guía FNDeepML (Estructura y Contenido), con sus respectivas etiquetas y atributos, y se muestra un ejemplo de la anotación con este esquema (Figura 3.2).

Nivel 1: Anotación de la Estructura

Este nivel divide una noticia en diferentes partes siguiendo la técnica periodística de la Pirámide Invertida. Esta técnica es una de las más utilizadas por los periodistas para reflejar la objetividad en una noticia (Thomson y cols., 2008), pues presenta la información en orden de relevancia, situando la información más importante al principio de la noticia y la menos relevante al final. Además,

[2] https://maldita.es/
[3] https://www.rtve.es/noticias/verificartve/
[4] https://escuela.elpais.com/talleres/verificacion-fuentes-e-investigacion/
[5] https://www.poynter.org/ifcn/

esta estructura permite que los usuarios adquieran rápidamente los puntos clave de la historia y facilita a su vez el procesamiento de la información (DeAngelo y Yegiyan, 2019).

Las cinco etiquetas del nivel de Estructura son titular (TITLE), subtítulo (SUBTITLE), entradilla (LEAD), cuerpo de la noticia (BODY) y conclusión (CONCLUSION). En la guía de anotación se utilizan las etiquetas en inglés y en mayúsculas, mientras que los atributos se utilizan en minúsculas para poder marcar la diferencia. La anotación de estas partes permite proporcionar información acerca de si la noticia sigue o no la estructura estándar periodística. Aunque se asume que cada periodista tiene su estilo personal a la hora de redactar o presentar una noticia y, por lo tanto, no todas las partes tienen por qué estar presentes (como el subtítulo o la conclusión), la falta de algunas partes esenciales (como el titular, la entradilla o el cuerpo) sugiere claramente que una noticia está mal estructurada. A continuación se muestra un ejemplo de texto anotado con las etiquetas estructurales:

<TITLE>Un vaso de agua caliente con limón puede salvarte la vida</TITLE>

<SUBTITLE>El limón, además de ser un componente ideal para nuestras comidas, puede salvarnos la vida, ya que previene y cura el cáncer.</SUBTITLE>

<LEAD>Son muchas las propiedades que tiene el limón, pero seguro que no sabías que desde hacía millones de años expertos médicos lo han utilizado para curar el cáncer, pues tomar un vaso de agua caliente con trozos de este cítrico todos los días mata las células cancerígenas de nuestro cuerpo y crea un escudo protector que previene futuros tumores.</LEAD>

<BODY>Existen muchos estudios que a lo largo de los años han demostrado que el limón tiene propiedades milagrosas para nuestra salud. Se ha llegado a demostrar que hasta es 100 veces más efectivo que la quimioterapia. No obstante, hay que saber cómo prepararlo para que este cítrico tenga los efectos deseados en nuestro cuerpo. En primer lugar, se debe utilizar agua caliente con trocitos de limón y tomarlo en ayunas todos los días [...]</BODY>

<CONCLUSION>En resumen, el limón es un alimento anticáncer que puede salvar tu vida gracias a sus propiedades anticancerígenas. Tomar una infusión de agua caliente con una rodaja de limón te ayudará a prevenir y matar esta dura enfermedad, así que no dudes en difundir esta noticia a todo el mundo.</CONCLUSION>

En esta primera versión de la guía de anotación, además de las etiquetas de la Pirámide Invertida, como parte de la estructura se incluyó también una etiqueta que permite marcar la presencia de citas o reproducir ideas de terceros en la noticia: la etiqueta QUOTE. Aunque forma parte del nivel de estructura, se trata de un elemento que se diferencia de las etiquetas básicas de la estructura piramidal porque se utiliza para enmarcar un conjunto de información externa, como son las citas de terceros que aparecen dentro de cada etiqueta estructural.

Un ejemplo de la etiqueta QUOTE es:

<QUOTE>"Es solo cuestión de tiempo"</QUOTE>, declaró el experto.

Nivel 2: Anotación del Contenido

La otra técnica utilizada por los periodistas para redactar una noticia de forma precisa y completa es la de las 5W1H que, tal y como se ha presentado en la sección 2.2.1, se centra en los elementos esenciales necesarios para comunicar con precisión una noticia, que consiste en responder a seis preguntas clave: qué (WHAT), quién (WHO), dónde (WHERE), cuándo (WHEN), por qué (WHY) y cómo (HOW). Estas preguntas describen el acontecimiento principal (Hamborg, Breitinger, Schubotz, Lachnit, y Gipp, 2018) y suelen encontrarse al principio de la noticia, como en el titular o la entradilla.

A nivel de contenido, la anotación marca los eventos de la noticia según los elementos semánticos relacionados con las preguntas 5W1H. Tal y como define (Hordofa, 2020), un evento es una forma natural de explicar relaciones complicadas entre personas, lugares, acciones y objetos, pero también es la forma natural de describir una noticia, la forma en la que los consumidores entienden lo que ha ocurrido en el mundo (Hou y cols., 2015). Como afirman (Chakma y cols., 2020), las 5W1H representan los constituyentes semánticos de una frase que son comparativamente más sencillos de entender e identificar. Si una noticia responde a todas estas preguntas, significará que la información se comunica de forma completa y, por lo tanto, la noticia presentará un mayor grado de confiabilidad que una noticia que no comunique la información de forma tan completa, lo que contribuye a la verificabilidad de la noticia, es decir, que tiene más posibilidad de ser verificada.

Finalmente, cabe resaltar que en esta primera versión de la anotación no se anotó con las 5W1H la parte del cuerpo de la noticia (etiqueta BODY) por su extensión, debido a que ralentizaba la anotación, por lo que se decidió dejarla sin anotar hasta que se comprobara que la anotación funcionaba correctamente con el resto de las partes. A continuación se presenta un ejemplo de las etiquetas 5W1H:

```
<WHO>Un científico italiano</WHO>
<WHAT>fue detenido</WHAT>
<HOW>mediante el uso de la fuerza</HOW>
<WHEN>ayer</WHEN>
<WHERE>en Milán</WHERE>
<WHY>por vender una vacuna no autorizada</WHY>
```

Atributos de las etiquetas

Además de anotar la Estructura y Contenido, tal y como se puede observar en la Figura 3.1, las etiquetas incluyen atributos con valores específicos que proporcionan información adicional a la anotación:

- ***type* (tipo)**: indica la veracidad de las etiquetas y presenta tres valores, que son *True* (Verdadero), *False* (Falso) y *Unknown* (Desconocido). De este modo, se pueden detectar elementos falsos y verdaderos en una misma

Figura 3.1: Esquema de anotación FNDeepML.

noticia. Este atributo se utiliza con todas las etiquetas, tanto con las de Estructura como con las de Contenido.

`<TITLE type ::= False>`Un vaso de agua caliente con limón puede salvarte la vida`</TITLE>`

`<WHAT type ::= True>`Se ha detectado una variante del virus`</WHAT>`

- ***author_stance*** **(postura del autor)**: se utiliza únicamente con la etiqueta QUOTE y sirve para anotar la postura del autor con respecto al texto citado. A diferencia del resto de etiquetas, la etiqueta QUOTE no presenta el atributo *type*, únicamente el *author_stance*, el cual se representa a partir de los siguientes valores: *Disagree* (en Desacuerdo), si el autor no respalda la cita; *Agree* (de Acuerdo), si el autor apoya lo que se dice en la cita; y *Unknown* (Desconocido), si la postura del autor no está clara y utiliza la cita únicamente para informar sin mostrar su postura.

  Según el experto, `<QUOTE author_stance ::= Unknown>`"Solo es cuestión de tiempo"`</QUOTE>`

<TITLE type ::= False> <WHO type ::= False>El doctor Chen</WHO> <WHAT type ::= False>afirmó que el cáncer se cura</WHAT> <HOW type ::= False>infusionando agua con una rodaja de limón</HOW> <WHEN type ::= False>cada día</WHEN></TITLE>

<LEAD type ::= False> <WHAT type ::= True>El limón tiene numerosas propiedades</WHAT>, pero <WHO type ::= Unknown>expertos médicos</WHO> <WHAT type ::= False>lo han utilizado</WHAT> <WHERE type ::= False>en Asia</WHERE> <WHEN type ::= False>durante millones de años</WHEN> <WHY type ::= False>porque cura el cáncer</WHY>. Se sabe que <WHAT type ::= True>el limón tiene beneficios para la salud</WHAT>, pero <WHO type ::= Unknown>oncólogos reconocidos</WHO> <WHEN type ::= Unknown>ahora</WHEN> <WHAT type ::= False>afirman que es posible matar las células cancerígenas</WHAT> <HOW type ::= False>consumiendo agua caliente con zumo de limón</HOW> <WHEN type ::= False>cada día</WHEN> <WHY type ::= False>debido a que sus vitaminas son cien veces más efectivas que la quimioterapia.</WHY></LEAD>

Figura 3.2: Ejemplo de la anotación de una noticia utilizando el esquema FN-DeepML.

3.1.2 Corpus FNDeep

El primer corpus que se creó para la presente investigación fue el corpus FNDeep, compuesto por 200 noticias verdaderas y falsas en español obtenidas de portales de salud y nutrición, así como de periódicos digitales generalistas. Las noticias pertenecen al ámbito de la salud, más concretamente a dos temas: covid-19 (100 noticias), por un lado, y problemas de salud y enfermedades generales (100 noticias), por otro lado. Entre ellas, 105 se clasificaron como true news y 95 como fake news, anotadas manualmente siguiendo la guía de anotación FNDeepML definida en la sección 3.1.1.

Se recopilaron manualmente noticias verdaderas, noticias falsas y desmentidos de periódicos digitales para poder contrastar la información a la hora de asignar el valor de veracidad. Inicialmente, la anotación empezó a realizarse en un fichero de texto, pero esta metodología ralentizaba y dificultaba la tarea, pues los elementos anotados no se visualizaban correctamente, lo que llevaba al anotador a cometer muchos errores. Por este motivo, se cambió el formato de anotación y se comenzó a utilizar la aplicación GATE Developer 8.6.1[6], que se configuró según nuestra guía de anotación.

Para anotar la veracidad de cada noticia, así como de cada uno de sus elementos de estructura y contenido, se llevó a cabo un procedimiento manual de verificación de información mediante el contraste de datos verificados por agencias oficiales pertenecientes a la IFCN, como Newtral[7], Salud sin Bulos[8],

[6] https://gate.ac.uk/download/
[7] https://www.newtral.es/
[8] https://saludsinbulos.com/

Maldita[9], Chequeado[10] o AFP Factual[11]. Estas agencias contrastan la información en diferentes fuentes y con diferentes técnicas, publicando desmentidos que ponen a disposición del público para dar a conocer la información verídica y falsa de una noticia determinada. Por otro lado, además de estas agencias de desmentidos, para comprobar la veracidad de la información también se utilizó la aplicación en línea denominada *Google Fact Check Explorer*[12], herramienta que lleva a cabo una búsqueda rápida de desmentidos a partir de palabras clave o temas concretos.

La categoría de veracidad de cada elemento 5W1H depende a su vez del contexto en el que se incluya la afirmación. Es posible tener diferentes valores de veracidad de una misma etiqueta, por ejemplo de un WHO. Si se observan los dos ejemplos que se presentan a continuación, donde la etiqueta WHO (Donald Trump) aparece en distintos artículos de noticias, tras el procedimiento manual de verificación, una etiqueta se consideró Verdadera y la otra Falsa. Aunque lingüística y semánticamente se trate del mismo término, dependiendo del contexto en el que se enmarque se considerará de una forma u otra y esa diferencia es la que se marca con ayuda de los desmentidos.

`<WHO type ::= True>`Donald Trump`</WHO>` es el nuevo candidato para las elecciones estadounidenses de 2020.

`<WHO type ::= False>`Donald Trump`</WHO>` descubre la vacuna contra el covid-19.

A continuación se muestra una serie de tablas descriptivas del corpus FNDeep. En la Tabla 3.1 se presentan cifras más concretas relativas a las tres partes esenciales de una noticia (titular, entradilla y cuerpo) del corpus construido.

Tipo noticia	Núm. docs	Núm. tokens	Media tokens por doc	Media tokens TITLE	Media tokens LEAD	Media tokens BODY
True news	105	75 951	723	12	77	562
Fake news	95	58 581	617	12	63	494
Total	200	134 532	670	12	70	530

Tabla 3.1: Descripción cuantitativa de las tres partes esenciales de las noticias del corpus: título, entradilla y cuerpo de la noticia.

Por otro lado, la Tabla 3.2 presenta una descripción cuantitativa de todas las partes estructurales de la noticia, anotadas manualmente como *True*, *False* y *Unknown* siguiendo el esquema de anotación FNDeepML.

La Tabla 3.3 presenta una descripción cuantitativa de los elementos 5W1H anotados como *True*, *False* y *Unknown* del corpus completo.

Finalmente, la Tabla 3.4 presenta los porcentajes de los distintos valores de

[9] https://maldita.es/
[10] https://chequeado.com/
[11] https://factual.afp.com/
[12] https://toolbox.google.com/factcheck/explorer

Tipo	TITLE	SUBTITLE	LEAD	BODY	CONCLUSION
True	50,75%	52,22%	46,45%	53%	50,40%
False	45,27%	28,89%	33,88%	47%	33,60%
Unknown	3,98%	18,89%	19,67%	0%	16,00%
Total	200	90	183	200	125

Tabla 3.2: Descripción cuantitativa de etiquetas de estructura (Pirámide Invertida) clasificadas como *True*, *False* y *Unknown* de todo el corpus.

Tipo	WHAT	WHO	WHEN	WHERE	WHY	HOW
True	41,64%	0,13%	30,41%	39,67%	32,26%	42,72%
False	35,43%	0,26%	25,77%	19,33%	45,16%	38,35%
Unknown	22,93%	99,61%	43,81%	41,00%	22,58%	18,93%
Total	1112	766	194	300	62	206

Tabla 3.3: Descripción cuantitativa de etiquetas de contenido (5W1H) clasificadas como *True*, *False* y *Unknown* de todo el corpus.

veracidad obtenidos tanto para las etiquetas de la estructura como para las de contenido. Esta tabla sólo incluye las cifras extraídas de las fake news del corpus, excluyendo las true news en las que todos los elementos son verdaderos.

Etiqueta	**False (%)**	**True (%)**	**Unknown (%)**	**Total etiquetas**
TITLE	**95,79**	0	4,21	95
SUBTITLE	68,42	10,53	21,05	38
LEAD	75,31	6,17	18,52	81
BODY	**100**	0	0	95
CONCLUSION	80,38	5,88	13,73	51
WHAT	**68,70**	6,11	25,18	409
WHERE	24,79	22,31	**52,89**	121
WHEN	43,02	9,30	**47,67**	86
WHO	0,59	0	**99,41**	340
WHY	**61,54**	15,38	23,08	39
HOW	**60,82**	16,49	22,68	97

Tabla 3.4: Distribución de las etiquetas *True*, *False* y *Unknown* de las fake news del corpus, excluyendo las true news.

Tras un análisis manual del corpus y de las cifras presentadas en la Tabla 3.4, se extrajeron algunas conclusiones preliminares sobre las noticias falsas del corpus:

- **Estructura de las noticias**: el titular es casi siempre falso en las noticias clasificadas como falsas. Por otro lado, el titular y el cuerpo de la noticia aparecen en todas las noticias, pero la entradilla no siempre está presente en la estructura de las noticias falsas.
- **Contenido 5W1H**: la etiqueta WHAT es la etiqueta que contiene más infor-

mación falsa, aunque también presenta un alto grado de información no definida. La información falsa proporcionada en el WHY y el HOW también es elevada y cercana a los valores del WHAT. En el caso de las etiquetas WHO, WHEN y WHERE, existe un alto grado de vaguedad, especialmente en la etiqueta WHO. Las noticias objetivas proporcionan datos precisos y concretos, por lo que detectar estas imprecisiones nos permite determinar si una noticia es fiable. Algunos ejemplos de etiquetas WHO imprecisas son los términos genéricos que no especifican la fuente o los autores (como "los expertos" o "investigadores"), debido a que las noticias falsas suelen evitar revelar fuentes concretas que restarían credibilidad a la noticia. En cuanto a la etiqueta WHERE, algunos ejemplos imprecisos son "en algunas ciudades" o "en otros países". Estos ejemplos no aluden a un lugar concreto, lo que hace que la información sea inexacta. Finalmente, "hace unos meses" o "en los próximos años" serían ejemplos de una etiqueta WHEN imprecisa. Al igual que los lugares, el tiempo aquí también es ambiguo, por lo que la información sigue sin ser fiable.

3.1.3 Acuerdo entre anotadores

Una vez realizada la primera anotación, se analizó la calidad del esquema y del recurso generado en función del acuerdo de anotación. Para ello, se llevó a cabo un acuerdo entre el anotador experto, con perfil lingüístico, y otros dos anotadores con perfil periodístico, mediante el coeficiente kappa de Cohen (Cohen, 1960).

El acuerdo analizó los elementos en los que ambos anotadores coincidían y, posteriormente, se analizaron las etiquetas que habían sido anotadas de forma diferente con el fin de llegar a un consenso e introducir las modificaciones necesarias, tanto en el esquema como en el corpus. Del acuerdo se obtuvo k=0,737 en el nivel de la Pirámide Invertida y k=0,851 en el nivel de las 5W1H, resultado que validó la anotación.

3.2 Modelado de confiabilidad

En este apartado, el del modelado de la confiabilidad, se cambia el enfoque de la anotación de la veracidad a la anotación de la confiabilidad, obteniendo así la última versión de la guía de anotación creada: RUN-AS (*Reliable and Unreliable News Annotation Scheme*). El cambio de enfoque se produjo por las siguientes razones: (i) verificar las noticias requiere un conocimiento del mundo que no siempre se puede alcanzar; (ii) a menudo el documento carece de contexto suficiente para poder entender completamente la afirmación y, por tanto, poder verificarla; (iii), la falsedad o verdad en una noticia es en muchas ocasiones difícil de cuantificar y depende del punto de vista subjetivo del verificador. Para eliminar dicha subjetividad en la emisión de juicio decidimos centrarnos en el criterio de confiabilidad.

Ante este cambio de enfoque se tuvo que reorientar la metodología sin cambiar el objetivo fijado inicialmente. Era imposible obtener un desmentido para cada noticia recopilada (porque a menudo el tema es tan reciente o con tan poca repercusión que no interesa verificarlo, o simplemente porque la difusión de la desinformación es más rápida que la de la información verificada). Si a esto se le añade que el lenguaje sin contexto no aporta información suficiente para hacer una clasificación y que necesita del conocimiento del mundo para contrastar la información, se puede concluir que la clasificación en Verdadero o Falso necesita apoyarse de la tarea de *fact-checking* o conocimiento externo para asignar un valor de veracidad. Sin embargo, esta tarea no forma parte de los objetivos de la tesis, por lo que se reorientó la anotación y se cambió el enfoque que se estaba siguiendo.

La propuesta de anotación actualizada (Apéndice C) se centra en clasificar las noticias en Confiable (*Reliable*) o No Confiable (*Unreliable*), desde una perspectiva lingüística y estructural, sin conocimiento externo. Además, la guía sigue basándose en las dos técnicas periodísticas de la Pirámide Invertida y las 5W1H, centradas en estructurar una noticia de forma clara y en ofrecer toda la información de forma completa y concisa. Este nuevo enfoque permite detectar rápidamente patrones de información sospechosos antes de contrastar el contenido con fuentes oficiales.

Nuestra hipótesis de partida se basa en la idea de que la confiabilidad de todo el texto depende de la confiabilidad de cada una de sus partes individuales, razón por la que se utilizan ambas técnicas periodísticas como base de nuestra anotación. A partir de esta hipótesis inicial, se plantean las siguientes subhipótesis: (i) las fake news son noticias no confiables que suelen mezclar información verídica con información falsa, por lo que el análisis individual de las partes y los elementos de contenido pueden determinar la confiabilidad global de la noticia; (ii) existen características lingüísticas y estructurales que permiten diferenciar las noticias no confiables de las noticias confiables, sin utilizar el conocimiento del mundo; (iii) una clasificación de confiabilidad puede proporcionar información útil para predecir la veracidad de una noticia, siendo una herramienta de apoyo para usuarios, agencias de verificación de datos y periodistas.

3.2.1 Esquema de anotación: RUN-AS

La propuesta de anotación RUN-AS es una evolución de la guía FNDeepML que permite detectar las partes esenciales de una noticia, que son la estructura y el contenido, junto con la confiabilidad de sus elementos semánticos, así como otros elementos lingüísticos de interés que permiten encontrar patrones de desinformación. El objetivo de esta propuesta de anotación deja de lado el enfoque de la veracidad y de la comparación de desmentidos para centrarse en un análisis puramente estructural y lingüístico de la noticia con el fin de averiguar si la forma en la que está estructurada o redactada una noticia influye en su confiabilidad. La clasificación en Confiable o No confiable puede ayudar a generar

un informe inicial que justifique esa decisión para que, en una fase posterior, pueda ser verificada con técnicas de verificación de datos.

Para saber si una noticia presenta información objetiva y sigue los estándares periodísticos, esta propuesta plantea una anotación basada en dos técnicas periodísticas y tres niveles de anotación lingüística: Estructura (Pirámide Invertida), Contenido (5W1H) y Elementos de Interés (EoI). Como los dos primeros niveles se han descrito en la sección 3.1.1, a continuación se presenta el nuevo nivel incorporado, así como los nuevos atributos de cada etiqueta.

Nivel 3: Anotación de los Elementos de Interés

Este nivel permite anotar información textual adicional que puede ser de interés a la hora de marcar la diferencia entre una noticia no confiable y una que sí lo es, pues, a diferencia de las etiquetas de estructura y contenido, estas permiten detectar la presencia de marcas subjetivas, exageradas, sesgadas o con carga emocional.

KEY_EXPRESSION (expresiones clave): fraseología que insta a los lectores a compartir la información o que expresa emociones como miedo, desprecio, alarma o fines económicos. Permiten encontrar indicadores con carga emocional que reflejen la opinión o intención del autor.

<KEY_EXPRESSION>Vamos a salvar vidas compartiendo esta gran información</KEY_EXPRESSION>

FIGURE (cifras): esta etiqueta permite marcar cifras, las cuales podrían verificarse fácilmente con técnicas de *fact-checking*. El mero hecho de que se trate de una característica verificable proporciona confiabilidad sin necesidad de que la verificación se haga efectiva.

<FIGURE>45</FIGURE> pacientes han dado positivo.

QUOTE (citas): como se ha definido en la sección 3.1.1, esta etiqueta marca la presencia de citas en la noticia (palabras o frases que se citan textualmente o reproducen una idea externa, no del propio periodista) y ya formaba parte del esquema inicial de anotación FNDeepML. No obstante, esta etiqueta se ha trasladado ahora al nivel de Elementos de Interés, pues más que una etiqueta de estructura como las de la Pirámide Invertida, se trata de una etiqueta que aporta información textual adicional, entre ellas, marcas de objetividad o subjetividad. La aparición de citas proporciona objetividad al texto siempre y cuando no se muestre ningún sesgo en el texto citado. Cuando las citas provienen de expertos externos, el grado de confiabilidad de la noticia aumenta.

<QUOTE>"Los casos de covid-19 han descendido notablemente"</QUOTE>, afirmó el ministro de Sanidad.

ORTHOTYPOGRAPHY (ortotipografía): etiqueta que permite marcar errores gramaticales, ortográficos o de formato, como frases enteras en mayúsculas, puntos suspensivos en medio del texto o incompletos, espacios dobles, muchos signos exclamativos, errores gramaticales u ortográficos, falta de cohesión, etc. La presencia de varios de estos elementos reflejan una redacción de baja calidad y, por lo tanto, la confiabilidad de la noticia disminuye.

`<ORTHOTYPOGRAPHY>`Camviará totalmente tu VIDA!!!`</ORTHOTYPOGRAPHY>`.

Atributos de las etiquetas

Además de anotar la Estructura, Contenido y Elementos de Interés, algunas de las etiquetas incluyen atributos nuevos con valores específicos que proporcionan información adicional en la anotación. Uno de los principales cambios realizados en el nivel de la Estructura (Pirámide Invertida) con respecto a la guía FNDeepML es la eliminación del atributo de veracidad. Debido a la cantidad de información que contiene cada parte (que puede combinar texto falso y texto verídico), en lugar de etiquetar cada parte estructural con un único valor de veracidad/confiabilidad, es el contenido semántico dentro de cada una de esas partes el que influye en la evaluación global. En este caso, se pretende únicamente marcar la presencia de las partes de la estructura, pues una noticia bien redactada debería seguir la técnica de la Pirámide Invertida. Como excepción, solo el titular presenta atributos, el resto de etiquetas no contiene ningún tipo de atributo.

Así pues, los dos atributos de la etiqueta TITLE son:

- ***style* (estilo)**: permite marcar dos valores, que son *Objective* (Objetivo) y *Subjective* (Subjetivo). Un titular objetivo presenta la información de forma precisa e informativa, mientras que un titular subjetivo tiende a ser alarmista, connotativo y emocional, llegando a veces a comportarse como un titular con gancho (*clickbait title*), que son titulares que pueden resultar exagerados para captar la atención del usuario, que este haga clic y permanezca en la página el mayor tiempo posible. Los titulares utilizados por esta estrategia no responden a los criterios periodísticos tradicionales, cuyo objetivo es informar a los usuarios, sino que el titular con gancho tendría como objetivo principal la comercialización o difusión de la información (Orosa, Santorum, y García, 2017).

- ***title_stance* (postura del titular)**: este atributo permite marcar si la información presentada en el cuerpo de la noticia es consistente con la información del titular. Esta consistencia se representa a partir de los siguientes valores: *Agree* (de Acuerdo), cuando la información es consistente en ambas partes; *Disagree* (en Desacuerdo), cuando la información no es consistente en una de las partes; y *Unrelated* (Sin relación) cuando la

información del titular no tiene ninguna relación con el resto de la noticia.

<TITLE style ::= Subjective title_stance::= Agree>Un vaso de agua caliente con limón puede salvarte la vida</TITLE>.

Respecto al nivel de Contenido (5W1H), se adoptan los atributos siguientes:

- ***reliability* (confiabilidad)**: es el atributo principal de esta anotación y permite clasificar cada elemento, así como la noticia global, con dos valores: *Reliable* (Confiable) y *Unreliable* (No confiable), dependiendo del nivel de precisión y neutralidad, así como de los criterios lingüísticos descritos en el apartado 3.2.2. El atributo *type* de la primera versión del esquema queda reemplazado por el de *reliability* para adaptar la guía al nuevo enfoque de confiabilidad. Este atributo se utiliza con todas las etiquetas 5W1H.

  <WHAT reliability ::= Unreliable>Un vaso de agua caliente con limón puede salvarte la vida</WHAT>.

- ***lack_of_information* (falta de evidencia)**: se utiliza con las etiquetas 5W1H para indicar si faltan datos importantes o evidencia científica. Este atributo se marca con un único valor (*Yes*), el cual indica que falta dicha evidencia. Cuando no es el caso, simplemente no se utiliza el atributo. Este atributo se utiliza con todas las etiquetas 5W1H, en caso necesario.

  <WHAT lack_of_information ::= Yes>Existen muchos estudios que han demostrado que el limón tiene propiedades milagrosas</WHAT>.

- ***main_event* (evento principal)**: este atributo se utiliza únicamente con la etiqueta WHAT, que designa el hecho o evento principal de la noticia y ayuda a diferenciarlo de otros eventos secundarios.

  <WHAT main_event>Un vaso de agua caliente con limón puede salvarte la vida</WHAT>.

- ***role* (rol)**: solo se utiliza con la etiqueta WHO para indicar el papel que desempeña el sujeto del evento. Este atributo se puede indicar con uno de los siguientes valores: *Subject* (Sujeto), *Tarjet* (Objeto) o *Both* (Ambos).

  <WHO role::= Target>Un científico italiano</WHO> fue detenido ayer en Milán.

La imagen 3.3 muestra un resumen de las diferentes etiquetas del esquema de anotación RUN-AS con sus atributos específicos y los posibles valores de cada atributo.

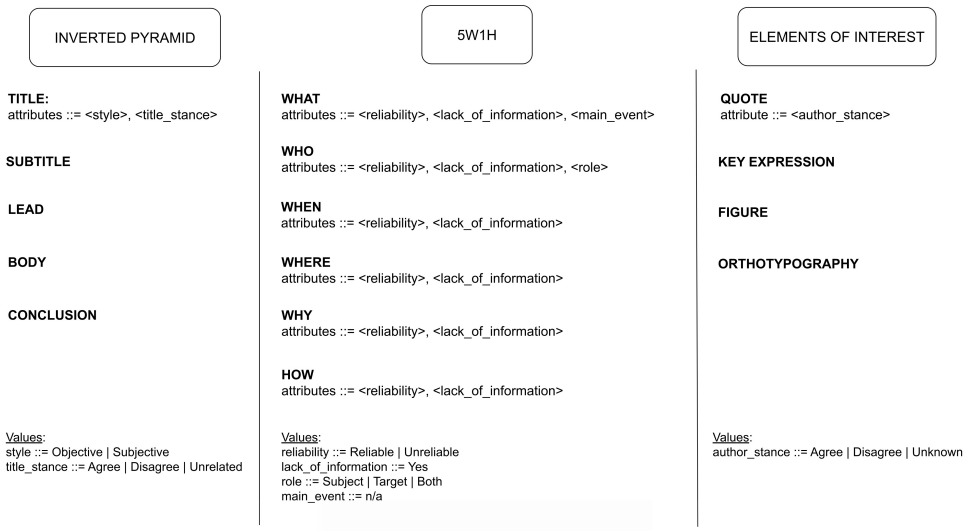


Figura 3.3: Esquema de anotación RUN-AS.

3.2.2 Criterios de confiabilidad

Este trabajo se centra en asignar un valor de confiabilidad a las etiquetas de contenido esencial (5W1H) descritas en el esquema de anotación. La complejidad de la tarea se debe a la detección de indicadores de desinformación sin corroborar la información con fuentes externas y teniendo en cuenta únicamente las características lingüísticas. Esto hace que la anotación sea más subjetiva, ya que el análisis depende de factores como el estilo de redacción y el propósito del autor, la carga emocional presente en el lenguaje o el nivel de candidez del lector. A pesar de esta subjetividad, existen rasgos lingüísticos que permiten detectar la confiabilidad de una noticia y determinar todos los elementos confiables y no confiables de cada parte de la noticia, lo que permite evaluar la confiabilidad global de la noticia.

Para anotar la confiabilidad de las etiquetas se tiene en cuenta la precisión y la neutralidad de la información proporcionada, así como la presencia de marcas personales, lenguaje despectivo, expresiones con carga emocional, falta de evidencia científica o lenguaje que influya negativa o positivamente en la noticia y que suela tener una intención concreta, como expresiones persuasivas o exhortativas.

Los criterios de confiabilidad presentados en esta propuesta se agrupan en dos principios clave: precisión y neutralidad. Para explicar los criterios establecidos por la guía RUN-AS a la hora de clasificar la confiabilidad, nos basamos en algunas investigaciones cuyas características son similares a las utilizadas en este trabajo.

Precisión

Para que la información facilitada sea confiable es importante que los datos sean exactos y no dejen margen para vaguedades o ambigüedades. Es uno de los factores clave para determinar la confiabilidad de la información. En nuestro modelado de la confiabilidad hemos considerado los siguientes indicadores:

- **Vaguedad y ambigüedad.** Las expresiones evasivas o vagas indican que se está ocultando información o que esta no se puede justificar, lo que hace que la información facilitada no sea fiable. Es más fiable dar una fecha exacta o detalles precisos sobre un científico (nombre, institución, título) que generalizar o dar datos inexactos. Por ejemplo, una etiqueta WHEN con valor Confiable sería "el viernes 19 de marzo", mientras que "hace mucho tiempo" carecería de precisión. Por el contrario, la presencia de cifras indica información precisa que se puede verificar fácilmente con fuentes externas, lo que denota confiabilidad, por ejemplo "se han administrado 60 000 dosis de vacunas".

  `<WHO reliability ::= Unreliable>`Los expertos`</WHO>`.

  `<WHO reliability ::= Reliable>`Investigadores del Grupo de Investigación en Biomecánica e Ingeniería de Rehabilitación`</WHO>`.

- **Falta de evidencia.** La ausencia de datos importantes en el texto (como la causa de un suceso, el sujeto de la acción, etc.) o la falta de evidencia como estudios científicos o datos oficiales y verificados denota una información poco fiable. A veces, el autor afirma que la información se basa en estudios científicos sin especificar cuáles, lo que aporta poca confiabilidad. La falta de datos y fuentes es otra característica típica de la desinformación, que convierte las noticias en relatos carentes de contenido informativo (Mottola, 2020).

  `<WHY lack_of_evidence ::= Yes>`Gracias a recientes estudios científicos`</WHY>`.

  `<WHY>`Gracias al estudio publicado en la revista científica *Science*`</WHY>`.

- **Ortotipografía.** Los errores ortográficos, un estilo de redacción pobre o descuidado, una puntuación inadecuada o el uso constante de mayúsculas restarán confiabilidad a una noticia de calidad. Algunos ejemplos de ortotipografía son: frases enteras en mayúsculas, puntos suspensivos en medio del texto o incompletos, dobles espacios, muchos signos de exclamación, errores gramaticales, faltas de ortografía, falta de cohesión, etc.

  `<ORTHOTYPOGRAPHY>`Akí en nuestro Pais`</ORTHOTYPOGRAPHY>`.

  `<ORTHOTYPOGRAPHY>`El Peregil Sana y es Natural!!!`</ORTHOTYPOGRAPHY>`.

Neutralidad

La neutralidad es un componente clave en las noticias. Una noticia tiene más probabilidad de ser confiable cuando la información se presenta de forma objetiva, es decir, cuando no influye ni positiva ni negativamente en el lector y no muestra la postura del autor. Algunos indicadores sobre la neutralidad del texto (o la falta de ella), considerados en el esquema RUN-AS, son los siguientes:

- **Observaciones personales y mensajes emotivos.** Cuando el autor habla en primera persona, cuenta su experiencia personal o la de alguien que conoce es un signo de poca confiabilidad, ya que el autor está tratando de asustar, persuadir o hacer que el lector se sienta más cercano a la historia y, por lo tanto, empatice con ella. Este tipo de comentarios personales hacen que la historia sea más subjetiva y que el lector sea más vulnerable a creer la noticia. De hecho, coincidimos con (Rashkin y cols., 2017), ya que sus resultados muestran que los pronombres en primera y segunda persona se utilizan más en noticias menos fiables o más engañosas.

  <KEY_EXPRESSION>En nuestra opinión</KEY_EXPRESSION>.

  <KEY_EXPRESSION>Evite que sus amigos y conocidos se enfermen, yo lo hago y me funciona</KEY_EXPRESSION>.

  Por otra parte, los mensajes ofensivos, esperanzadores, alarmistas o exhortativos son una clara señal de la falta de confiabilidad porque el autor intenta manipular al lector y jugar con sus emociones. (A. X. Zhang y cols., 2018) afirman que los lectores pueden ser engañados por el tono emocional de los artículos y este tono puede encontrarse en afirmaciones exageradas o textos cargados de emoción, como expresiones de desprecio, indignación o rencor.

  <KEY_EXPRESSION>Esta noticia podría salvar el mundo</KEY_EXPRESSION>.

  <KEY_EXPRESSION>Esta gentuza miserable</KEY_EXPRESSION>.

- **Citas y postura del autor.** La presencia de citas añade neutralidad a la noticia, ya que indica que la información procede de expertos u organizaciones externos y estudios (A. X. Zhang y cols., 2018). Sin embargo, la postura del autor de una cita (indicada con el atributo *author_stance* y los valores *Agree*, *Disagree* o *Unknown*) también puede influir en la falta de confiabilidad de una noticia. En ese caso, si el texto muestra que el autor apoya (*Agree*) o refuta (*Disagree*) una idea, se estará introduciendo un claro matiz de subjetividad, ya que el autor estará dando su opinión. Sin embargo, una cita etiquetada con el valor *Unknown* (Desconocido) es señal de neutralidad, ya que el autor sólo estará reproduciendo las palabras de un tercero para informar y no para influir en el lector.

  La IARC califica como <QUOTE author_stance ::= Unknown>"probable carcinógeno humano"</QUOTE> la acrilamina.

<QUOTE author_stance ::= Disagree>“Nunca se necesitaron los ventiladores, ni la unidad de cuidados intensivos”</QUOTE>, declaró un experto.

- **Estilo y postura del titular.** Los titulares de las noticias suelen dar pistas importantes sobre la confiabilidad de su contenido. Las noticias poco fiables suelen contener titulares alarmistas, subjetivos y llamativos. En nuestra propuesta de anotación, se marca esta característica con el atributo *style*, que puede ser Objetivo o Subjetivo. Asimismo, los titulares engañosos u opacos sobre un tema pueden ser un claro ejemplo de los titulares con gancho (A. X. Zhang y cols., 2018). Incluso ciertos rasgos morfosintácticos como la excesiva longitud de un titular, el uso de más palabras en mayúsculas (Horne y Adali, 2017) y de signos de puntuación (sobre todo de exclamación) y elipsis pueden conducir a una falta de neutralidad (Mottola, 2020). Además, cuando la información del titular (o postura del titular, *title_stance*) no coincide con el contenido del resto de la noticia puede haber información sospechosa.

  <TITLE style ::= Subjective title_stance ::= Agree>Atención: El uso prolongado de la mascarilla produce hipoxia</TITLE>.

  <TITLE style ::= Subjective title_stance ::= Unrelated>Gran escándalo! La EMA ve vínculos entre AstraZeneca y los coágulos y trombosis</TITLE>.

3.3 Conclusiones

En este capítulo se han presentado las guías de anotación diseñadas para la presente investigación y se han descrito detalladamente cada uno de los niveles de anotación, así como las etiquetas, atributos y valores. Además, en aras de una mejor comprensión, se muestran ejemplos de cada elemento de anotación. Son dos los esquemas de anotación que se han creado para este estudio, ambos se basan en las técnicas periodísticas de la Pirámide Invertida y las 5W1H y presentan una anotación detallada multinivel. No obstante, cada uno ha sido creado con un propósito distinto.

La primera guía diseñada, FNDeepML, se enfoca en el criterio de veracidad y atribuye a las noticias un valor de Verdadero, Falso o Desconocido dependiendo del conocimiento externo proporcionado por los desmentidos recopilados. Esta anotación tiene dos niveles (Estructura y Contenido) y su principal atributo es el de *type*, el cual permite asignar el valor de veracidad a la noticia. De esta guía se obtiene el corpus FNDeep, construido y anotado íntegramente de forma manual y que está formado por 200 noticias falsas y verdaderas en español.

Posteriormente, se reorientó la tesis y, por lo tanto, también la anotación. Dejamos de enfocarnos en asignar un criterio de veracidad basándonos en conocimiento externo y verificación de datos y orientamos la clasificación de las

noticias hacia un criterio de confiabilidad sin recurrir a los desmentidos, únicamente con un análisis puramente lingüístico. Como resultado de este cambio obtuvimos nuestra segunda guía de anotación: RUN-AS. El esquema RUN-AS se basa en los dos mismos niveles de anotación de Estructura y Contenido, pero además incorpora un tercer nivel (Elementos de Interés) que permite detectar otros indicadores de información a nivel textual, como la subjetividad, la emoción o la redacción de baja calidad.

Asimismo, se añaden nuevos atributos que proporcionan más información del texto y, principalmente, se sustituye el atributo que marca la veracidad (atributo *type*) por el atributo de confiabilidad (atributo *reliability*). Por otro lado, al no contrastar la información con conocimiento externo, se definen unos criterios de confiabilidad basados fundamentalmente en los principios de precisión y neutralidad que permiten detectar indicadores de información sospechosa de la noticia. A diferencia del corpus FNDeep, el recurso generado con la guía RUN-AS, denominado RUN, ha sido anotado y construido de forma semiautomática, por lo que el Capítulo 4 se enfocará en el proceso de construcción y anotación de este corpus.

CAPÍTULO 4

Propuesta de anotación asistida

4.1 Introducción

La viralización de grandes volúmenes de desinformación hace necesaria la automatización de su detección, ya que se precisa procesar la mayor cantidad de información posible para extraer datos concluyentes y eso solo se puede conseguir automatizando el proceso. Para automatizar la detección, los algoritmos existentes necesitan de la retroalimentación del experto humano, la cual se consigue a partir de ejemplos que conforman un corpus. Los corpus anotados son herramientas indispensables para entrenar modelos computacionales en PLN. Cuando un problema se aborda desde la perspectiva de la IA, ya sea con técnicas de ML o DL, se requieren millones de instancias de retroalimentación humana para obtener los corpus que se utilizarán para entrenar y evaluar los sistemas que se encargarán de resolver el problema (Stenetorp y cols., 2012).

Uno de los principales retos de nuestra investigación, y del campo de PLN en general, es la obtención de corpus de entrenamiento, pues la anotación es una tarea costosa, ardua y que requiere mucho tiempo, lo que se traduce en una escasez de recursos para entrenar algoritmos de ML y DL. Así pues, teniendo en cuenta el añadido de que nuestra anotación es además compleja semánticamente, una tarea propuesta en la presente tesis es la de presentar una metodología de anotación semiautomática basada en diferentes estrategias de *Human-in-the-loop*. Estas estrategias se han utilizado para construir un corpus de confiabilidad de noticias en español con el fin de combatir la desinformación.

La tarea propuesta permite obtener un recurso de calidad mediante la implementación de una metodología de anotación asistida que aumenta la eficacia y la velocidad del anotador en la tarea, optimizando así el rendimiento del modelo sin la necesidad de obtener tantos ejemplos. La metodología consta de tres fases incrementales basadas en estrategias de HITL que dan como resultado la construcción del corpus RUN. La validez de la metodología de anotación asisti-

da se evaluó mediante el cálculo del tiempo, el error de preanotación, el acuerdo entre anotadores y el rendimiento del modelo tras entrenar con el corpus semiautomático RUN, métricas de evaluación que se presentarán en el apartado 4.8.

Construir corpus eficientes es una tarea compleja, ya que las anotaciones de los corpus pueden tener distintos grados de dificultad. Esto puede implicar no sólo un coste de tiempo, sino también la necesidad de un alto nivel de conocimiento en una anotación determinada. Un corpus eficiente sería aquel que se construye de la forma más rápida y económica posible, y que además incluye los ejemplos anotados más apropiados para ayudar en el aprendizaje de la resolución del problema. Por lo tanto, mediante la tarea de anotación asistida propuesta se pretende crear un corpus de calidad hecho a medida y seleccionado según criterios específicos que aumente, o al menos mantenga, la precisión al tiempo que ahorra tiempo y esfuerzo al anotador. El principal objetivo es implementar un procedimiento de anotación semiautomático que permita obtener un recurso de calidad para la detección de desinformación combinando la anotación automática y la manual. Esta tarea contribuye en las siguientes aportaciones al área de investigación:

- El diseño e implementación de una metodología de anotación semiautomática aplicando diferentes estrategias HITL que asista a los anotadores y optimice el rendimiento del recurso. Una de esas estrategias es la del Aprendizaje Activo —*Active Learning*— (AL), la cual elige los ejemplos de entrenamiento de los que quiere aprender. De esta forma se puede mejorar el rendimiento del modelo con menos instancias de entrenamiento y se minimizan los ejemplos a anotar.

- La creación de un corpus anotado de forma eficiente mediante la aplicación de la metodología propuesta, que es un requisito fundamental para las tareas de IA y PLN. La evaluación de la calidad y de los beneficios de aplicar este tipo de metodología semiautomática se evidencian con la disminución del tiempo-esfuerzo en la construcción del corpus necesario, la mejora en el control de la calidad de su anotación y la optimización en la precisión del modelo aprendido.

- La disponibilidad para la comunidad investigadora del corpus semiautomático generado [1], una vez corroborada la validez del recurso.

4.2 Inteligencia Artificial y Human-in-the-loop

En la propuesta de anotación asistida nos centramos en el vanguardista concepto *Human-in-the-loop* (HITL), un conjunto de estrategias que combinan la inteligencia humana y la de las máquinas en aplicaciones que utilizan IA. Estas comprenden un amplio campo de investigación que abarca la intersección

[1] Disponible en `https://github.com/livinglang/NewsReliabilityAnnotation`

de la informática, la ciencia cognitiva y la psicología, y se está aplicando en el área de PLN. Construir tecnología de IA con intervención humana permite que las tareas humanas sean asistidas por ML para aumentar la eficiencia (Monarch, 2021). Dado que nuestro reto es entrenar con una anotación compleja semánticamente, hemos optado por aplicar técnicas HITL para ayudar en el proceso de anotación seleccionando los mejores ejemplos para entrenar y facilitando la tarea de anotación mediante una preanotación automática, lo que permite reducir el tiempo y los ejemplos a anotar. El concepto HITL se basa en que la máquina ayude al humano a ayudar a la máquina, pues esta estrategia no tiene como finalidad reemplazar al anotador, sino crear modelos que asistan e interactúen con el experto humano, de forma que faciliten su trabajo.

La mayoría de los sistemas de IA actuales no pueden aprender por sí mismos y necesitan retroalimentación y evaluación por parte de expertos humanos. Aproximadamente el 90 % de las aplicaciones actuales basadas en ML utilizan aprendizaje supervisado, es decir, aprenden a partir de ejemplos creados por humanos. Tal y como establece (Okoro, Abara, Umagba, Ajonye, y Isa, 2018), es necesaria una solución de modelo híbrido que combine tanto los esfuerzos humanos como los de las máquinas y, para ello, es esencial la creación de recursos para entrenar. El diseño, creación y anotación de un corpus es una tarea clave en el desarrollo de herramientas y corpus en PLN. No obstante, el número de corpus anotados para entrenar es escaso y la recopilación de datos es uno de los desafíos en la tarea de detección del engaño (Saquete y cols., 2020). Esta escasez se debe al tiempo y coste que requiere la tarea de anotación, pues anotar y compilar un corpus requiere esfuerzo, tiempo, consistencia y conocimiento experto. Este problema está a la vanguardia de las investigaciones en PLN y especialmente en las investigaciones de detección de información, pues el desarrollo de nuevos recursos, como es el caso de los corpus anotados, pueden ayudar a aumentar el rendimiento de los métodos automáticos que se centran en detectar este tipo de noticias (Posadas-Durán y cols., 2019).

Debido a la complejidad de nuestra propuesta de anotación semántica y dado que los investigadores pasan más tiempo generando datos que construyendo modelos de ML, nos hemos centrado en crear una metodología HITL para aumentar la eficacia de nuestro trabajo. Los sistemas de IA enfocados en HITL mejoran continuamente gracias a la aportación humana, abordando las limitaciones de las soluciones de IA previas y tendiendo un puente entre la máquina y el ser humano. Estos sistemas pretenden aprovechar la capacidad de la IA para escalar el procesamiento a cantidades muy grandes de datos, al tiempo que se apoyan en la inteligencia humana para realizar tareas muy complejas, como en el caso de la comprensión del lenguaje natural (Demartini, Mizzaro, y Spina, 2020). La metodología HITL se está utilizando en varios estudios para aumentar la eficiencia en la recopilación de datos, como en los casos de (Fanton, Bonaldi, Tekiroglu, y Guerini, 2021; Cañizares-Díaz y cols., 2021), ya que “el bucle ejecutivo continuo desarrolla una relación humano-IA más fiable hasta cierto punto, contribuyendo a una mayor precisión y una mayor robustez del sistema de PLN”

(Wu y cols., 2022).

Uno de los principios del HITL es ayudar en las tareas humanas con el aprendizaje automático para aumentar la eficiencia. De acuerdo con esto, el presente trabajo construye un corpus anotado de forma semiautomática, utilizando HITL en muchas partes del circuito de ML, desde el muestreo de datos no etiquetados hasta la actualización del modelo.

4.2.1 Active Learning

Una estrategia HITL muy extendida es la del ***Active Learning***. De hecho, la anotación y el AL son las piedras angulares del HITL (Monarch, 2021). El AL se utiliza cuando la obtención de datos etiquetados exige una gran cantidad de tiempo o dinero, pues esta estrategia tiene como objetivo seleccionar ejemplos con alta utilidad para el modelo (Tomanek, Wermter, y Hahn, 2007) y aumentar así el rendimiento del modelo de aprendizaje al tiempo que reduce la cantidad de datos anotados necesarios (Kholghi, Sitbon, Zuccon, y Nguyen, 2016). Para ciertas tareas, como la clasificación de documentos o la extracción de información, la obtención de instancias anotadas puede ser repetitiva y costosa, tanto en términos de tiempo como de dinero, y para que un sistema de aprendizaje supervisado funcione bien a menudo debe entrenarse con cientos o miles de ejemplos anotados. La idea clave del AL se basa en que si un algoritmo de ML puede elegir los datos a partir de los que quiere aprender, obtendrá mejores resultados con menos entrenamiento, porque elegirá aquellos ejemplos que le permitan incrementar su rendimiento (Settles, 2009).

El AL se aplica a varias tareas de ML, como la detección de objetos, la segmentación semántica, el etiquetado de secuencias o la generación del lenguaje (Monarch, 2021). En el presente trabajo, el AL se utiliza en la construcción semiautomática de un corpus para la detección de desinformación y permite optimizar el rendimiento con menos noticias, pero mejor elegidas, para incorporar al conjunto de entrenamiento. Un aspecto importante del AL es el proceso iterativo, ya que permite el reentrenamiento haciendo uso de la retroalimentación humana, que a su vez permite al sistema mejorar la precisión.

Existen muchas estrategias de AL y para evaluar las instancias no etiquetadas se pueden seleccionar las muestras a partir de una distribución determinada. Dos amplias estrategias de muestreo de AL son (Monarch, 2021):

- Muestreo de incertidumbre: se trata del conjunto de estrategias para identificar elementos no etiquetados que están cerca de un límite de decisión en su modelo de aprendizaje automático actual.

- Muestreo de diversidad: es el conjunto de estrategias para identificar elementos no etiquetados que están infrarrepresentados o son desconocidos para el modelo de aprendizaje automático en su estado actual. El objetivo de este muestreo es seleccionar elementos nuevos, inusuales e infrarre-

presentados para su anotación, con el fin de proporcionar al modelo una imagen más completa del problema.

Además del AL, las estrategias HITL incluyen dos objetivos distintos que normalmente se combinan: mejorar la precisión del modelo de ML a través de la retroalimentación humana y facilitar la tarea humana con la ayuda del ML. Este último es nuestro objetivo principal, pero el primero (el de aumentar la precisión del modelo) mejorará a medida que se obtenga un corpus más amplio.

El concepto HITL se ha aplicado con éxito en varias áreas como en la gubernamental (Benedikt y cols., 2020), la médica (Budd, Robinson, y Kainz, 2021), y la energética (Jung y Jazizadeh, 2019). En cuanto a la aplicación del HITL en la detección de desinformación, algunos trabajos son clave. (Demartini y cols., 2020) presentaron los desafíos y oportunidades de combinar enfoques de verificación de hechos, tanto automáticos como manuales, para la desinformación, desarrollando un marco humano-IA. Este trabajo se centra más en la comprobación de hechos y no en la confiabilidad. Por otro lado, (Daniel, 2021) pretendía determinar qué conjunto de técnicas era la más adecuada para la detección de la desinformación y cuál era la mejor forma de utilizar cada técnica para este fin. Sin embargo, ninguno de estos trabajos aprovecha el proceso de anotación de un corpus de desinformación.

A continuación se presentan dos propuestas de metodología semiautomática: en la primera se integran diferentes estrategias HITL al proceso de construcción de corpus, con el fin de aumentar la cantidad de datos anotados y alcanzar así la precisión deseada con mayor rapidez y facilidad (sección 4.3); mientras que en la segunda se mantienen las estrategias HITL, pero además se incorpora la técnica de extracción de resúmenes con el fin de anotar únicamente la información relevante resaltada y así agilizar aún más el proceso de anotación (sección 4.7).

4.3 Metodología semiautomática de construcción del corpus RUN

Con el fin de simplificar las tareas de compilación y anotación, se propone una estrategia HITL para automatizar gradualmente las tareas de construcción de corpus. De este modo, se minimiza el esfuerzo del anotador humano y se crea un corpus más extenso y de forma menos costosa.

La metodología aquí propuesta se desarrolla en tres fases que consisten en automatizar gradualmente el proceso de anotación y observar los cambios en comparación con la anotación totalmente manual. El objetivo es comprobar si la anotación asistida crea un recurso de calidad de forma más rápida y aumenta la eficacia del anotador humano. Las noticias se anotaron en pequeños lotes y siguiendo diferentes estrategias, pero manteniendo el mismo número de noticias para cada lote y respetando el esquema de anotación en todas las etapas.

La Figura 4.1 muestra un diagrama que representa el proceso de anotación llevado a cabo siguiendo estrategias HITL. Como se ha indicado anteriormente, el proceso de anotación se dividió en las tres fases siguientes, representadas en la Figura 4.1: Fase 1. Compilación y anotación manual; Fase 2. Compilación automatizada y anotación manual; y Fase 3. Compilación automatizada y anotación semiautomática.

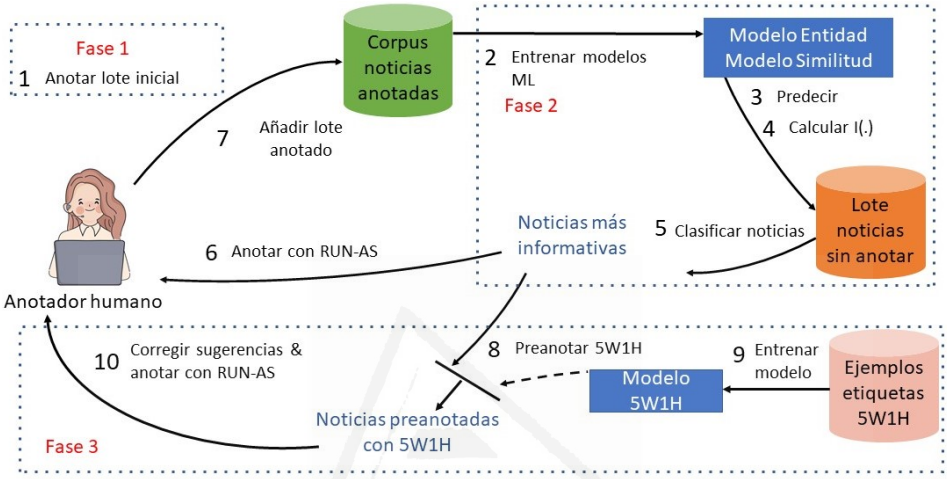


Figura 4.1: Proceso de anotación semiautomática utilizando estrategias HITL.

A continuación se describe la metodología adoptada en cada fase de la construcción del corpus, mientras que las medidas de tiempo y error de cada fase se presentan en el apartado 4.8.

4.3.1 Fase 0: Recopilación de datos

Esta fase es una etapa de preparación centrada en la recopilación de un amplio conjunto de noticias procedentes de diversas fuentes. La recopilación de datos se realiza antes de la tarea de anotación. En esta etapa se lleva a cabo una recopilación manual de datos, así como mediante *webcrawler*, según las necesidades de las fases posteriores.

4.3.2 Fase 1: compilación y anotación manual

En la Fase 1, las noticias que se recopilaron y anotaron de forma totalmente manual en la Fase 0 se utilizan como entrada, como se ilustra en el paso 1 de la Figura 4.1. En cuanto a la tarea de compilación, se obtuvo un total de 40 noticias en español de 9 fuentes. Las noticias se anotaron manualmente siguiendo la guía RUN-AS, es decir, los niveles de la Pirámide Invertida, las 5W1H y los Elementos de Interés. Cada nivel de anotación y la noticia en su conjunto se clasificaron como Confiable o No Confiable en función de varios criterios lingüísticos

y semánticos como la neutralidad, la precisión, la evidencia o la fraseología con carga emocional, entre otros. Esta primera fase fue un proceso arduo y lento, ya que buscar las noticias una a una y anotarlas desde cero fueron tareas que requirieron mucho tiempo.

El resultado de esta fase es la primera versión del corpus (el cilindro verde en la Figura 4.1), en la que se obtuvieron las noticias anotadas que sirvieron de entrada para la Fase 2.

4.3.3 Fase 2: compilación automática y anotación manual

La Fase 2 toma como entrada las noticias recopiladas mediante *web crawler* e introduce el uso de la estrategia de AL para aumentar la productividad del proceso de anotación. En esta fase, el modelo de AL se utiliza únicamente para guiar el proceso de anotación sugiriendo qué noticias nutren más al modelo y ayudando así a equilibrar el corpus automáticamente.

El proceso se lleva a cabo de la siguiente manera. Partiendo de un pequeño lote de noticias anotadas en la Fase 1, se entrena un modelo supervisado que predice sobre un lote mayor de noticias sin anotar (cilindro naranja de la Figura 4.1). Para cada noticia, se calcula automáticamente una métrica informativa basada en un equilibrio entre la incertidumbre del modelo y la diversidad del contenido. Todas las noticias sin anotar se ordenan según esta puntuación y se crea una lista provisional de sugerencias, intercalando noticias de distintas fuentes. Así, de esta lista de sugerencias, en la que aparecen primero las noticias más informativas teniendo en cuenta la diversidad de contenidos, un anotador experto filtra las que no sigan los criterios del idioma, formato, extensión u otras características semánticas deseadas. La lista final se compone de las noticias más informativas que se ajustan a todos los criterios deseados, evaluados antes de la anotación, y equilibradas en cuanto a las fuentes originales. Por último, este lote de noticias se anota y se añade al conjunto de entrenamiento (paso 6 y paso 7 en la Figura 4.1) y se repite todo el circuito de AL. En la Figura 4.1, los pasos 3, 4 y 5 muestran el paso en el que el modelo propone las noticias más apropiadas para ser anotadas a partir del conjunto de datos sin anotar.

En concreto, en esta fase se anotaron un total de 4 lotes y se seleccionaron un total de 10 noticias por lote. Así, una vez finalizada la Fase 2, se añaden al corpus 40 noticias nuevas que han sido anotadas manualmente por un anotador experto, pero seleccionadas con base en la estrategia de AL, lo que contribuye a garantizar un nivel mínimo de diversidad y consistencia difícil de alcanzar con una selección puramente manual.

4.3.4 Fase 3: compilación automática y anotación semiautomática

Por último, la Fase 3 es una evolución de la Fase 2 con el objetivo de mejorar significativamente los tiempos de anotación. Como entrada también se utilizaron las noticias recopiladas mediante *web crawler*. La novedad aquí es la preano-

tación automática de los elementos 5W1H realizada por el sistema para ayudar al experto. El objetivo es que el anotador no etiquete desde cero, sino que se limite a revisar y completar la preanotación realizada por el sistema. En esta fase, el sistema solo preanota las noticias con el segundo nivel de anotación (etiquetas 5W1H) definido en el esquema RUN-AS por ser el nivel de anotación más complejo y que requiere más tiempo.

Como se observa en la Figura 4.1, la intervención humana sigue siendo importante ya que es necesario verificar que la selección automática de noticias y la preanotación propuesta por el sistema semiautomático cumplan los criterios del corpus. Con respecto a la preanotación, se amplía el circuito presentado en la Fase 2 (pasos 8, 9 y 10). Esta fase utiliza un modelo 5W1H, previamente entrenado (paso 9) con ejemplos de las etiquetas 5W1H (cilindro rosa de la Figura 4.1), para preanotar etiquetas 5W1H (paso 8) en las noticias seleccionadas. En el último paso de esta fase (paso 10), el anotador edita los elementos preanotados por el modelo 5W1H y añade el resto etiquetas de la anotación RUN-AS. Por último, se añade un nuevo lote anotado al corpus (paso 7) para concluir el circuito.

En esta fase se anotaron inicialmente 40 noticias para mantener el mismo número de noticias anotadas que en las fases anteriores. Sin embargo, tras validar que la anotación semiautomática aceleraba el proceso, se anotaron otras 50 noticias (90 en total en la Fase 3), con el fin de aumentar el corpus. En total se anotaron 9 lotes.

4.4 Implementación de la metodología

Teniendo en cuenta la definición conceptual de la metodología, se creó un sistema de anotación asistida para automatizar las Fases 2 y 3. En la Fase 2 se anotaron 4 lotes y en la Fase 3 se anotaron 9 lotes. Cada lote está formado por 10 noticias.

4.4.1 Fase 2 (estrategias de Active Learning)

El modelo de AL utilizado en esta fase es una implementación basada en la propuesta de (Cañizares-Díaz y cols., 2021) para la anotación de entidades y relaciones. El modelo original consta de dos clasificadores diferentes, uno para el reconocimiento de entidades y otro para la extracción de relaciones. Sin embargo, dado que nuestro esquema de anotación no contiene relaciones, sólo se utiliza el clasificador de entidades.

El modelo se basa en un clasificador de Regresión Logística —*Logistic Regression*— (LR) entrenado con etiquetas de entidades a nivel de token. Así pues, se realiza un preprocesamiento del texto anotado que transforma la anotación basada en Brat, que se define a nivel de tramo de texto, en una secuencia de tokens anotados. El modelo de LR se alimenta de características sintácticas y

semánticas a nivel de token (extraídas con Spacy) y características contextuales (es decir, las características combinadas de una pequeña ventana de tokens circundantes). La configuración de los parámetros se detalla en el Apéndice C.

4.4.2 Fase 3 (preanotación de las 5W1H)

Para llevar a cabo la Fase 3, se requirió un modelo que anotase automáticamente las etiquetas 5W1H. Para llevar a cabo esta tarea, se propuso utilizar un modelo de QA disponible en el enlace Hugging Face[2]. Este modelo se construyó con una versión afinada del modelo BETO (Canete, Chaperon, Fuentes, y Pérez, 2020) del corpus *Stanford Question Answering Dataset* (SQuAD)-es-v2.0 (Rajpurkar, Zhang, Lopyrev, y Liang, 2016) para adaptarse a la tarea de QA.

Con el objetivo de adaptar este modelo a nuestro corpus, lo que se conoce como *fine-tuning* o ajuste fino, los ejemplos 5W1H de las Fases 1 y 2 se dividieron en tres conjuntos: *training* (entrenamiento), *development* (desarrollo) y *test* (prueba). El ajuste fino se llevó a cabo mediante el entrenamiento y la evaluación con los conjuntos de *training* y *development*. Este proceso se llevó a cabo utilizando la biblioteca Simple Transformers[3]. Como entrada para entrenar el modelo se utilizan las preguntas 5W1H, su contexto y su respectiva respuesta. El modelo devuelve una respuesta así como una puntuación que representa la probabilidad de certeza asociada a la respuesta. La configuración de los parámetros se detalla en el Apéndice C.

Implementación y *fine-tuning* del modelo de QA

El rendimiento del modelo de QA en las 5W1H, implementado en la Fase 3, se comparó con y sin *fine-tuning*, cuyos resultados se muestran en la Tabla 4.1. Para llevar a cabo la comparativa entre ambos modelos, se tuvieron en cuenta las siguientes métricas:

- *Exact Match* (EM): el número de coincidencias exactas entre la respuesta predicha y la respuesta manual. Por ejemplo, **científicos** anotado como WHO por el modelo de QA y como WHO por el anotador de forma manual. Es una coincidencia exacta porque ambas anotaciones han coincidido en la etiqueta elegida.

- *Similar Match* (SM): el número de coincidencias parciales entre la respuesta predicha y la respuesta manual. Por ejemplo, **España** anotado como WHERE por el modelo de QA y como WHO por el anotador. Aunque la elección de la etiqueta por parte del modelo es correcta, pues este sustantivo denota un lugar, en ciertos contextos (p. ej. **España ha decretado el estado de alarma**) puede haber una personificación del nombre. En ese

[2] https://bit.ly/3zfnisx
[3] https://simpletransformers.ai/

caso, el país funcionaría como el sujeto de la oración y, por lo tanto, como una etiqueta WHO.

- *Incorrect Match* (IM): el número de respuestas predichas que no coinciden con las anotaciones manuales. Por ejemplo, **hace dos años** anotado como WHERE por el modelo de QA y como WHEN por el anotador. La respuesta del modelo sería incorrecta, ya que el ejemplo mencionado denota tiempo y no lugar.

- *Total Exact Match* (Total EM): porcentaje de las coincidencias exactas sobre el número de los ejemplos predichos.

- F_1: la medida F_1 es la media de las métricas *precision* y *recall* (Grandini, Bagli, y Visani, 2020). El número de palabras compartidas entre la predicción y la verdad es la base de la puntuación F_1 (Rajpurkar y cols., 2016), donde el F1 alcanza su mejor valor en 1 y la peor puntuación en 0 (Hernández Rubio y cols., 2022).

$$F_1 = \frac{2 * Precision * Recall}{Precison + Recall} \tag{4.1}$$

Modelo	**EM**	**SM**	**IM**	**Total EM**	F_1
QA sin fine-tuning en 5W1H	30	**396**	**141**	5,2	19,1
QA con fine-tuning en 5W1H tras Fase 2	236	178	153	41,6	61,3
QA con fine-tuning en 5W1H tras lote 6 de Fase 3	**263**	152	152	**46,3**	**64,1**

Tabla 4.1: Comparación entre los modelos de QA con y sin *fine-tuning*.

Como se muestra en esta tabla, el modelo de QA con *fine-tuning* obtiene mejores resultados en las principales métricas. Se ha considerado que la métrica más importante en esta tarea de preanotación es el *Total Exact Match*, porque maximizar esta métrica puede ayudar a los expertos anotadores a reducir, descartar y modificar ejemplos. Estos resultados confirman que el *fine-tuning* es beneficioso para la tarea de preanotación de las etiquetas 5W1H, pues aunque la métrica de *Incorrect Match* se mantenga prácticamente igual, los *Exact Match* mejoran notablemente, lo que permite que el anotador no se demore en buscar los elementos semánticos y anote desde cero tanta etiquetas. Por ejemplo, es más costoso añadir 233 etiquetas desde cero, que corregir 11 etiquetas anotadas incorrectamente, pues en este último caso es solo modificar la etiqueta o eliminarla.

Tras anotar el lote 6 de la Fase 3, se volvió a entrenar el modelo con 60 noticias más. El nuevo modelo de QA (con *fine-tuning* tras el lote 6) obtuvo los mejores resultados en términos de *Exact Match*, *Total Exact Match* y F_1 (Tabla

4.1). Estos resultados confirman que con un mayor número de ejemplos de las etiquetas 5W1H se puede obtener un modelo de alta precisión que reduce los tiempos de anotación y asiste al anotador humano en esta tarea compleja.

El objetivo de la preanotación con el modelo de QA es anotar el mayor número posible de elementos 5W1H con gran precisión para que el anotador tenga que descartar muy pocos ejemplos como incorrectos, ya que cuanto mayor sea el número de correcciones, mayor será el retraso en el proceso de anotación. Para reducir la tasa de error en el número de ejemplos preanotados por el modelo, se anotaron automáticamente las etiquetas 5W1H en 10 noticias (lote 1 de la Fase 3) y se clasificaron manualmente como correctas o incorrectas. En la Figura 4.2, se representa mediante un gráfico el proceso mediante el cual se definió un umbral que separa la respuesta incorrecta (punto naranja) de la correcta o de las similares (punto azul). En este caso, el umbral seleccionado fue de 0,11 porque fue la puntuación más cercana en todo momento a la línea de tendencia de regresión, separando así los tipos de etiquetas clasificadas manualmente. Este umbral se configuró en el sistema de anotación semiautomática con el mejor modelo de QA obtenido para iniciar el proceso de anotación en la Fase 3 (QA de 5W1H con *fine-tuning* tras la Fase 2). La configuración de los parámetros se detalla en el Apéndice C.

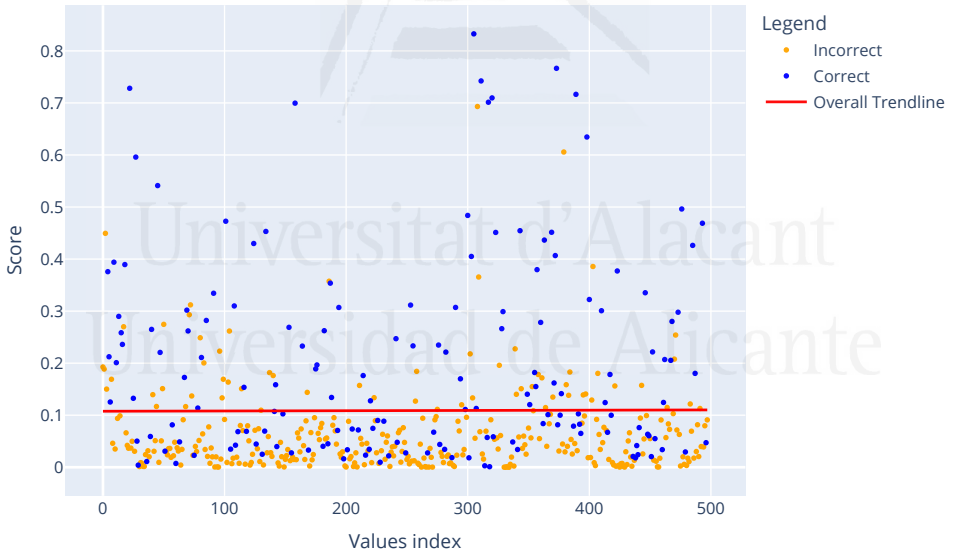


Figura 4.2: Representación de las etiquetas 5W1H mediante puntuaciones de predicción del modelo de QA, índice de cada etiqueta y clasificación manual por un anotador experto de las etiquetas correctas y similares (puntos azules) e incorrectas (puntos naranjas).

Prototipo computacional

En la Fase 3 el modelo propone las noticias que deben anotarse mediante AL. Esto permite al anotador seleccionar, descartar y preanotar noticias en la misma interfaz (Figura 4.3)[4].

La interfaz de usuario es la herramienta Brat[5] junto con el sistema asistido implementado que permite al anotador descartar, aceptar o modificar las propuestas de anotación (Figura 4.4). Brat es una herramienta de anotación intuitiva basada en la web que tiene como objetivo mejorar la productividad del anotador integrando estrechamente la tecnología PLN en el proceso de anotación (Stenetorp y cols., 2012). El uso de esta interfaz permite realizar anotaciones de forma rápida, precisa y sencilla y, al estar alojada en un servidor, guarda automáticamente el trabajo realizado. Además, Brat presenta una interfaz visual y cómoda, pues muestra las etiquetas con colores y símbolos sin necesidad de situar el cursor encima para identificar cada etiqueta, lo que supone una ventaja para una anotación tan compleja como la nuestra que tiene tantos tipos de etiquetas y atributos.

Además, en esta fase tanto las tareas de compilación como las de anotación se realizan en la misma interfaz, sin tener que cambiar de pantalla y buscar en sitios externos. El sistema de anotación asistida se basa en una interfaz predictiva que, como señala (Monarch, 2021), consiste en elementos que han sido preanotados por un modelo de ML. Este tipo de interfaz permite a los anotadores editar los elementos y reajustar el modelo con los errores detectados y corregidos. Gracias a esta navegabilidad, el sistema integra la recomendación de noticias, que no sólo ahorra tiempo, sino que también tiene en cuenta la selección del anotador y, sobre esa base, reentrena el modelo mediante AL.

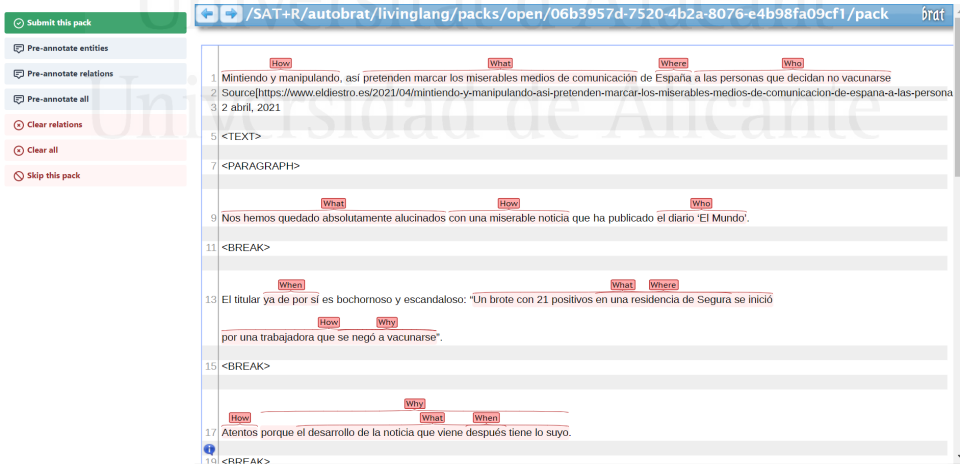


Figura 4.3: Preanotación en Brat.

[4]Fuente de la noticia utilizada en la Figura: `https://bit.ly/3PZJHk4`
[5]`https://brat.nlplab.org/`

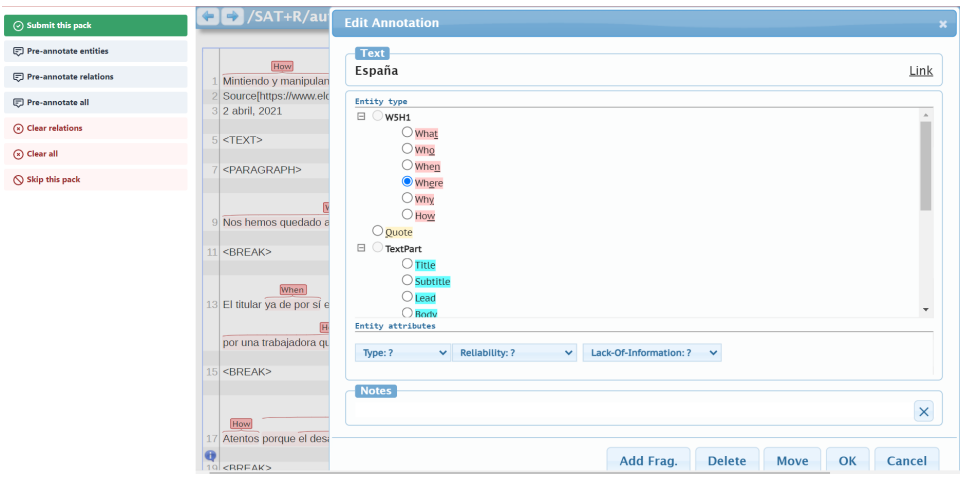


Figura 4.4: Modificación y selección de etiquetas y atributos en Brat.

4.5 Corpus RUN

Tras entrenar con la metodología semiautomática aplicando HITL y anotar con el nuevo enfoque de confiabilidad siguiendo la guía de anotación RUN-AS (distinto al enfoque inicial adoptado para el corpus FNDeep), se ha obtenido como resultado el corpus RUN (*Reliable and Unreliable News*). Este corpus está compuesto por noticias Confiables y No confiables (tanto de España como de Latinoamérica). El idioma elegido es el español porque, a pesar de ser la tercera lengua más hablada del mundo y la segunda en número de hablantes nativos[6], existen pocos corpus construidos en español para avanzar en la detección automática de la desinformación. Aunque el dominio de los contenidos recopilados es principalmente el de la salud, uno de los temas más susceptibles de recibir el impacto de la desinformación, hay otros dominios cubiertos, como el de la política.

El corpus RUN contiene 170 noticias recopiladas de medios digitales tradicionales, de las cuales una parte ha sido anotada de manera manual (40 noticias) y la otra (130 noticias) con anotación asistida (esta última siguiendo la metodología explicada en el apartado 4.3). Con el fin de analizar la estructura periodística tradicional se descartaron publicaciones de redes sociales, blogs o noticias con formato de preguntas frecuentes —*Frequently Asked Questions*— (FAQ). El corpus está equilibrado en cuanto a la clasificación de Confiable (85) y No confiable (85). La descripción cuantitativa de las etiquetas de contenido (5W1H) se presentan en la Tabla 4.2. Tanto la anotación de los elementos internos (Estructura y Contenido) como la confiabilidad global de la noticia se anotan según dos valores: Confiable (*Reliable*) y No confiable (*Unreliable*), en función de los criterios de confiabilidad descritos en la sección 3.2.2. Tanto el corpus como el

[6] https://cvc.cervantes.es/lengua/anuario/anuario_20/informes_ic/p01.htm

esquema de anotación están disponibles en el repositorio de Github[7].

5W1H	**Unreliable**	**Reliable**	**Total**
WHAT	687	1600	2296
WHEN	117	573	690
WHERE	685	58	747
WHO	326	1525	1856
WHY	142	241	384
HOW	165	358	529
TOTAL corpus	2122	4355	6502

Tabla 4.2: Descripción cuantitativa de las 5W1H en el corpus RUN.

Para compilar el corpus se han seguido tres criterios principales: i) dominio (salud y política), ii) idioma (español) y iii) estructura tradicional del contenido de las noticias (Pirámide Invertida). Respecto a los pasos que se siguieron para la creación del corpus en su etapa inicial (antes de implementar la metodología semiautomática), en primer lugar se definió el corpus y se delimitó en función de los tres criterios mencionados arriba (dominio, idioma y estructura tradicional). En segundo lugar, las noticias se recopilaron manualmente y mediante un *web crawler*. En tercer lugar, la estructura y el contenido semántico de las noticias fueron anotados manualmente por el anotador experto (la autora de la tesis) siguiendo los niveles de la Pirámide Invertida, las 5W1H y los Elementos de Interés presentados en la Figura 3.3, y se asignó un valor de confiabilidad a cada etiqueta 5W1H. En cuarto lugar, la confiabilidad global de cada noticia fue asignada por dos anotadores no expertos con conocimientos en PLN, teniendo en cuenta únicamente el texto plano, sin las etiquetas anotadas por el experto. Por último, se calcularon dos acuerdos entre anotadores.

4.6 Acuerdo entre anotadores

Tras llevar a cabo la anotación del corpus RUN, se analizó la calidad del esquema y del corpus mediante la métrica del acuerdo entre anotadores —*Inter-Annotator Agreement*— (IAA), que se calculó entre dos anotadores expertos en traducción y lingüística siguiendo la fórmula utilizada por (Névéol, Doğan, y Lu, 2011): **IAA = número de coincidencias/(número de coincidencias + número de no coincidencias).**

Al comparar las anotaciones, se consideraron coincidencia o *match* todas las etiquetas 5W1H en las que los dos anotadores (A y B) coincidieron en asignar la misma categoría. Por ejemplo, **científicos** anotado como WHO por el anotador A y como WHO por el anotador B.

[7] https://github.com/marionieto51/NewsReliabilityAnnotation

También se consideraron como *match* aquellos casos en los que había una ligera diferencia de longitud en cuanto a la extensión de los elementos a anotar, pero la etiqueta 5W1H asignada era la misma. Por ejemplo **científicos** anotado como WHO por el anotador A y **científicos especializados en biofísica** anotado como WHO por el anotador B.

Se marcaba como no coincidencia o *non-match* cuando los anotadores no estaban de acuerdo en utilizar la misma etiqueta 5W1H para un intervalo de elementos. Por ejemplo: **científicos** anotado como WHO por el anotador A y anotado como WHERE por el anotador B.

En esta investigación, el IAA tiene como objetivo medir el acierto o el error entre los anotadores al utilizar la guía de anotación. Los resultados más actualizados se presentan en la Tabla 4.3.

Nivel anotación	**Corpus RUN**
Pirámide Invertida	0,80
5W1H	0,70
Elementos de Interés	0,63
Anotación completa	0,70

Tabla 4.3: IAA por nivel de anotación del corpus RUN.

A pesar de que para ambos esquemas de anotación se ha probado a calcular el coeficiente kappa de Cohen (Cohen, 1960), se ha llegado a la conclusión de que no es la medida más apropiada para tareas de anotación a nivel de token (en las que gran parte de los tokens no están anotados en el texto). Por lo tanto, para la evaluación del corpus RUN se utilizó el IAA como medida de acuerdo entre anotadores. Para calcular este acuerdo, sólo se tuvo en cuenta la presencia de las etiquetas 5W1H.

Como puede deducirse de los resultados del IAA, el porcentaje de acuerdo en todos los niveles es bastante alto. Se obtuvo un buen acuerdo entre anotadores en las 5W1H, que es el nivel de anotación más complejo de RUN-AS y el único que se ha automatizado. El nivel de la Pirámide Invertida obtuvo el mayor acuerdo debido a su naturaleza estructural y a la falta de carga semántica. Por último, en cuanto al nivel de Elementos de Interés, el acuerdo obtenido también refleja, al igual que el nivel 5W1H, la complejidad de la semántica intrínseca de nuestra propuesta de anotación.

RUN-AS es una anotación multinivel bastante densa, con múltiples etiquetas y atributos que se aplican a artículos de noticias enteros. Se ha comprobado que la naturaleza semántica y la complejidad de la propuesta de anotación, así como la extensión de las noticias (a diferencia de otros medios como tuits o publicaciones) pueden dar lugar a una elección subjetiva del texto a anotar por parte del experto humano, lo que dificulta el acuerdo entre anotadores. Debido a la extensión, la anotación se vuelve flexible y subjetiva, por lo que para intentar reducir esta subjetividad se ha experimentado con el uso de extracción de

resúmenes para limitar el texto a anotar.

4.7 Metodología semiautomática del corpus RUN-AS-SFN

El uso de técnicas de resúmenes de texto ha tenido un impacto positivo en diferentes áreas, como la educación —donde los resúmenes se utilizan para apoyar las tareas de comprensión de lectura (Brown, 2018; Engelen, Camp, van de Pol, y de Bruin, 2018; Lin, Jhang, y Dong, 2018; Barreiro, 2019)—, de negocios —al producir, por ejemplo, un resumen automático de registros de eventos para ayudar a los analistas (Dijkman y Wilbik, 2017)—, o de salud, independientemente de si los resúmenes se crearon manualmente (Petkovic y cols., 2016; Hartling, Gates, Pillay, Nuspl, y Newton, 2018) o automáticamente (Liu, Song, y Chen, 2019). Las ventajas de los resúmenes se deben en parte a su capacidad para identificar la información más relevante de un documento y condensarla en un nuevo texto, lo que ayuda a reducir tiempo y recursos a la hora de gestionar grandes cantidades de datos. Estos métodos han demostrado su eficacia cuando se integran como componente intermedio de sistemas más complejos.

Los enfoques para la obtención automática de resúmenes de textos pueden clasificarse en extractivos y abstractivos (Lloret y Palomar, 2012). Los enfoques extractivos se centran en detectar la información más relevante de un texto, que luego se copia textualmente en el resumen final. Por el contrario, los enfoques abstractivos detectan la información relevante y, a continuación, se lleva a cabo un proceso más elaborado mediante el cual la información se fusiona, comprime o incluso se infiere (Jani, Patel, Yadav, Suthar, y Patel, 2022). Además de estos dos, se han desarrollado enfoques híbridos, que combinan métodos extractivos y abstractivos (Kirmani, Manzoor Hakak, Mohd, y Mohd, 2019).

Teniendo en cuenta las técnicas de extracción de información relevante a través de resúmenes presentadas y con la finalidad de simplificar aún más el proceso de anotación, se presenta una evolución de la metodología HITL descrita previamente, que permite una mayor simplificación en el tiempo de anotación sin perder la información realmente relevante para la tarea de detección de desinformación. En la metodología previa se anotaba la noticia completa, pero no siempre es necesario. Con la extracción de resúmenes, únicamente las frases consideradas relevantes serán anotadas. En este caso, el enfoque de resúmenes utilizado será el extractivo, que consigue mejores resultados en la generación de un texto comprensible.

La metodología de anotación semiautomática en este caso consta de dos niveles: (i) selección de la información más relevante de cada noticia utilizando técnicas de resúmenes y (ii) aplicación de estrategia HITL para la preanotación automática. La información relevante se preanota con la estructura de la noticia y las 5W1H, mostrando esta preanotación al anotador humano para que la corrija y la complete. En el proceso de HITL, esta retroalimentación humana se utiliza para volver a entrenar el modelo de preanotación. La Figura 4.5 mues-

tra la metodología y los pasos seguidos para anotar el corpus RUN-AS-SFN de forma semiautomática, corpus obtenido a partir de la anotación de un subconjunto del corpus *The Spanish Fake News Corpus* de (Posadas-Durán y cols., 2019) que se presenta en la sección 4.7.3.

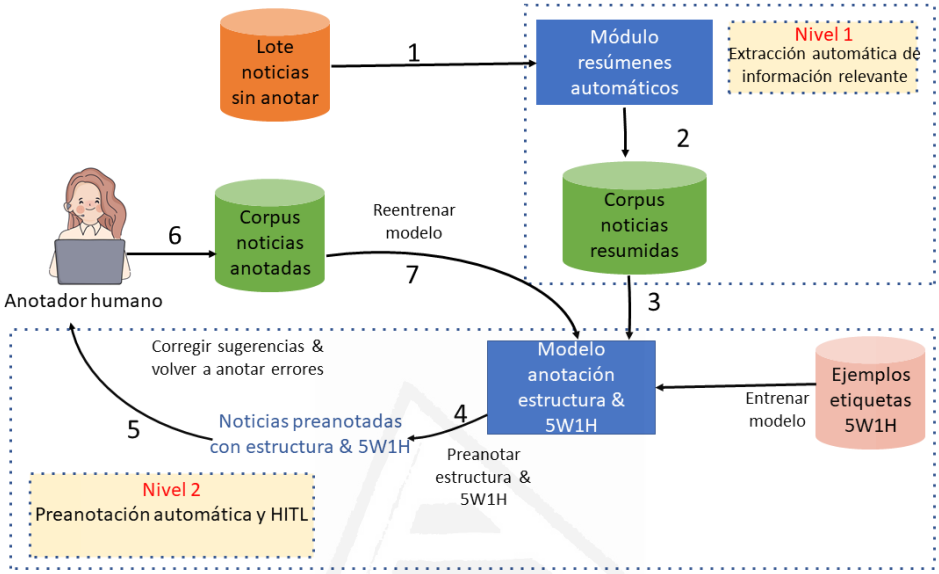


Figura 4.5: Diseño de la metodología de anotación semiautomática para el corpus RUN-AS-SFN.

4.7.1 Nivel 1: extracción automática de información relevante

Debido a la complejidad semántica de nuestro esquema de anotación, en lugar de anotar todo el documento, se aplican técnicas de resúmenes para extraer la información relevante de la noticia (paso 1 de la Figura 4.5). Las noticias resumidas se almacenan (paso 2) y se utilizan como entrada para el siguiente nivel (paso 3). Para el resumen, se definió un número de diez frases por noticia y se utilizó un desarrollo propio del algoritmo de resumen extractivo TextRank (Mihalcea y Tarau, 2004), dado su buen rendimiento, tiempo de ejecución y disponibilidad de implementación[8].

4.7.2 Nivel 2: preanotación automática

Una vez obtenida la información relevante, en este nivel el sistema realiza una preanotación automática de la estructura y de las etiquetas 5W1H para ayudar al experto (paso 4). La ventaja de esto es que el anotador no necesita etiquetar desde cero, sino que simplemente revisa y completa la preanotación reali-

[8] https://pypi.org/project/sumy/

zada por el sistema (paso 5). La estructura de las noticias ha sido anotada por un sistema basado en reglas que se desarrolló siguiendo la teoría de la Pirámide Invertida. En el caso de las etiquetas 5W1H, se utilizó un modelo de DL previamente entrenado con ejemplos de etiquetas 5W1H (cilindro rosa, representado en la Figura 4.5).

El anotador humano comprueba que la preanotación propuesta por el sistema semiautomático cumple los criterios del corpus, la edita siguiendo el esquema RUN-AS y, por último, el nuevo lote anotado se añade al corpus (paso 6). Este corpus anotado se utiliza no sólo para entrenar modelos de detección de confiabilidad, sino también para volver a entrenar el modelo 5W1H, cerrando así el circuito humano-máquina (paso 7).

Implementación y ajuste del modelo de QA

Al igual que para el modelo utilizado para preanotar las 5W1H en la Fase 3 de la metodología semiautomática del corpus RUN (ver apartado 4.4.2), para aplicar la metodología con HITL y extracción de resúmenes se utilizó también el modelo de QA preentrenado para detectar las etiquetas 5W1H, llevando a cabo aquí también el proceso de *fine-tuning* (se preentenó y evaluó dos veces, comparando manualmente la anotación entre el modelo previo y el preentrenado, hasta que se comprobó que el modelo de HITL mejoraba). En este caso se utilizó como conjunto de ejemplo inicial de entrenamiento el corpus RUN con las 170 noticias previamente anotadas con las 5W1H. El corpus RUN se dividió en los conjuntos de *training*, *validation* y *test*. Los conjuntos *training* y *validation* se actualizaron con los nuevos ejemplos 5W1H anotados durante el proceso HITL, proceso que permite mejorar el modelo 5W1H con más lotes de noticias anotadas. Para reentrenar el modelo, se utilizaron las siguientes entradas: preguntas 5W1H, contexto de las preguntas y sus respectivas respuestas.

En este proceso de HITL, en primer lugar se obtuvo el modelo **M1** tras llevar a cabo *fine-tuning* en las noticias anotadas inicialmente (corpus RUN). Tras este primer entrenamiento, se volvió a entrenar el modelo 5W1H con 250 noticias más obtenidas con la metodología propuesta (150 del conjunto de *training* y 100 del conjunto de *test*), obteniendo el segundo modelo 5W1H (**M2**).

La Tabla 4.4 muestra los resultados de los modelos de las 5W1H utilizados para predecir el conjunto de *test*. Los resultados de predicción del modelo de QA preentrenado de las 5W1H sin *fine-tuning* también figuran.

Modelo	**EM**	**SM**	**IM**	**Total EM**	F_1
QA preentrenado	30	396	141	5,29	19,15
M1 con *fine-tuning*	263	152	152	46,38	64,14
M2 con *fine-tuning*	**272**	**162**	**133**	**47,97**	**66,32**

Tabla 4.4: Comparación entre el modelo de QA con y sin *fine-tuning*.

El **M1** muestra un rendimiento destacado frente al modelo preentrenado.

El **M2** obtuvo los mejores resultados en términos de *Exact Match*, *Total Exact Match* y F_1 (Tabla 4.4). Estos resultados confirman que la obtención de un mayor número de ejemplos de etiquetas 5W1H mejora el modelo y reduce el tiempo de la tarea de anotación.

La Tabla 4.11 muestra la media de las categorías de *Exact Match*, *Incorrect Match* y *Similar Match* y de las etiquetas 5W1H del recuento manual llevado a cabo por los modelos **M1** y **M2** tras el reentrenamiento en bucle.

Prototipo computacional

En el Nivel 2, el modelo preanota las noticias en la misma interfaz de anotación utilizada para la sección 4.4.2: la herramienta de Brat. Esta herramienta permite anotar de forma rápida y precisa, así como corregir la preanotación propuesta por el modelo. Además de la anotación asistida, en esta fase el modelo extrae los resúmenes de las noticias, lo que permite anotar las frases que forman parte del resumen, marcadas como HL (Figura 4.6). El modelo marca los resúmenes, pero también muestra la noticia completa para que el anotador pueda contextualizar la información.

Figura 4.6: Preanotación en Brat con resúmenes marcados como HL.

4.7.3 Corpus RUN-AS-SFN

Además de probar la metodología propuesta que combina resúmenes con preanotación, se quiso evaluar si la detección de la confiabilidad podría ser aplicada en la tarea de detección de veracidad. Para ello, se utilizó un corpus publicado en español por (Posadas-Durán y cols., 2019) que presenta la veracidad

anotada de las noticias: *The Spanish Fake News Corpus* (SFN).

El corpus SFN contiene un total de 971 noticias (491 verdaderas y 480 falsas). En lugar de utilizar el corpus RUN, se quiso probar la nueva metodología con este corpus por una razón: el corpus SFN está anotado con un valor de veracidad global, por lo que además de probar la metodología, se puede comprobar si nuestra clasificación de confiabilidad y la anotación con los indicadores lingüísticos definidos sirven para detectar la veracidad anotada en el corpus SFN, es decir, si nuestro criterio sobre si una noticia contiene información no confiable coincide con el valor de veracidad asignado por este corpus.

Se seleccionó un subconjunto de 400 noticias (alrededor del 50 % del corpus). Conservando la anotación inicial del corpus, se aplicó la anotación semiautomática siguiendo el esquema RUN-AS y obteniendo así el nuevo corpus RUN-AS-SFN (un subconjunto del corpus SFN como resultado de aplicar la anotación RUN-AS semiautomática con resúmenes y HITL). De este subconjunto, se utilizaron 300 noticias para el entrenamiento y 100 noticias para las pruebas. El subconjunto anotado RUN-AS-SFN abarca los siguientes temas: economía, deportes, ciencia, educación, salud, sociedad, entretenimiento, política y seguridad. Las cifras relativas al corpus tras la anotación semiautomática con RUN-AS se presentan en la Tabla 4.5.

	Veracidad		**Confiabilidad**	
	True	False	Reliable	Unreliable
TRAINING	143	157	160	140
TEST	48	52	62	38
TOTAL	191	209	222	178

Tabla 4.5: Descripción cuantitativa del corpus anotado RUN-AS-SFN.

Como se explica en la metodología, la anotación de las distintas partes de la noticia se realizó únicamente en las frases extraídas de los resúmenes. Sin embargo, a cada noticia se le asigna un valor de confiabilidad global teniendo en cuenta todo el contenido de la noticia. Como se muestra en la Tabla 4.5, el subconjunto seleccionado estaba equilibrado en noticias Falsas y Verdaderas, tanto para el conjunto de *training* como para el de *test*. Tras anotar la confiabilidad, el número de noticias Confiables frente a las No confiables estaba bastante equilibrado, con sólo un número ligeramente mayor de noticias Confiables que de No confiables.

Las noticias mezclan información verídica con información no contrastada, lo que dificulta la tarea de detección de desinformación. Por eso es importante que las distintas partes y el contenido esencial de una noticia tengan valores de confiabilidad específicos, que influyan en el valor de confiabilidad global de una noticia. En la Tabla 4.6 se presenta una descripción cuantitativa sobre las etiquetas 5W1H anotadas como Confiables y No confiables en el corpus RUN-AS-SFN.

5W1H	Unreliable	Reliable	Total
WHAT	1465	2670	4135
WHEN	133	1200	1333
WHERE	103	1543	1646
WHO	521	3588	4109
WHY	324	512	836
HOW	194	568	762
TOTAL	2740	10 081	12 821

Tabla 4.6: Descripción cuantitativa de las etiquetas 5W1H en el corpus RUN-AS-SFN.

Como puede observarse en la Tabla 4.6, el número de 5W1H Confiables es mucho mayor que el de No confiables, a pesar de que las noticias Confiables y No confiables están bastante equilibradas. Esto se debe a que se disfraza a las noticias falsas incorporando más información fiable que no fiable, lo que dificulta aún más la detección de la veracidad de una noticia.

4.8 Marco de evaluación de la metodología semiautomática

En este apartado se presentan las dos métricas principales utilizadas para evaluar la metodología semiautomática descrita a lo largo del presente capítulo. Esta evaluación se lleva a cabo con base en el cálculo del tiempo y el cálculo del error de preanotación, tanto para la anotación del corpus RUN como para la del corpus RUN-AS-SFN.

4.8.1 Evaluación del corpus RUN

Cálculo del tiempo

Una de las métricas esenciales a la hora de evaluar nuestra anotación es la del tiempo, pues como bien se ha ido indicando a lo largo de la tesis, es imprescindible reducir el esfuerzo, tiempo y coste de la tarea de anotación y construcción del corpus.

Para medir la evolución de la construcción del corpus con el sistema semiautomático, se midió el tiempo empleado en cada fase, tanto en las tareas de compilación como en las de anotación. El tiempo empleado en la tarea de compilación se calculó en función del tiempo que se tardó en encontrar, leer y guardar cada noticia, mientras que la tarea de anotación se calculó en función de las palabras por segundo anotadas en cada noticia (Tabla 4.7). Para cada fase se proporciona una media del tiempo de compilación por noticia y de anotación por palabra.

Fase	Compilación	Anotación
1	15 min/noticia	1,9642 s/palabras
2	3 min/noticia	2,2704 s/palabra
3	1,3 min/noticia	1,5461 s/palabra

Tabla 4.7: Comparación de fases medidas en tiempo.

Cada parámetro se midió por fase para comparar la evolución en tiempo de la construcción del corpus:

En la Fase 1, el proceso de compilación llevó una media de 15 minutos/noticia, ya que el anotador tuvo que buscar las fuentes, leer las noticias para asegurarse de que eran adecuadas para el corpus, guardarlas en el ordenador y añadirlas a la interfaz. Todo ello se procesó manualmente. En cuanto al proceso de anotación, las noticias se anotaron manualmente y, en esta primera fase, el tiempo medio empleado en la tarea de anotación fue de 1,9642 segundos/palabra.

En la Fase 2, tras evaluar el tiempo empleado en la tarea de compilación y compararlo con la Fase 1, se demostró que el sistema asistido ahorra tiempo, ya que sugiere noticias automáticamente y el experto sólo tiene que leerlas y comprobar si encajan con el corpus. El experto no tiene que dedicar tiempo a buscar fuentes y noticias ni a descargarlas, simplemente tiene que aceptarlas o rechazarlas. Además, no se tienen en cuenta las noticias repetidas o ya descartadas, lo que a su vez también reduce el tiempo. En esta fase, el tiempo medio empleado en el proceso de compilación fue de 3 minutos/noticia. Por lo tanto, se ve una clara diferencia en la tarea de compilación entre la Fase 1 y la Fase 2. Lo que se tardaba dos horas y media en recopilar 10 noticias, se tarda treinta minutos con la automatización de esta tarea. En cuanto a la tarea de anotación, como se siguió realizando manualmente, casi no hubo diferencia con respecto a la Fase 1. El tiempo empleado en la anotación de la Fase 2 fue de 2,2704 segundos/palabra.

En la Fase 3, el proceso de compilación llevó una media de 1,3 minutos/noticia. Como el sistema muestra directamente la noticia propuesta en la interfaz de anotación, el experto sólo tiene que leerla y decidir si es válida o no. La navegabilidad del sistema facilita ambas tareas (compilación y anotación). En cuanto a la tarea de anotación, la preanotación redujo el tiempo medio de anotación a 1,5461 segundos/palabra.

La reducción de tiempo en cada fase durante este proceso de anotación se puede observar en la Figura 4.7, teniendo en cuenta que cada fase comprende 4 lotes, es decir, 40 noticias por fase (10 noticias cada lote) con un número medio de 20 000 palabras por fase.

Para concluir, en el proceso de anotación se produjo una reducción de tiempo del 42,17 % cuando la anotación se realizó en la Fase 2 en comparación con la anotación totalmente manual (992,84 min/40 noticias en la Fase 1 frente a 574,08 min/40 noticias en la Fase 2) , y una reducción final del 63,61 % tras llevar

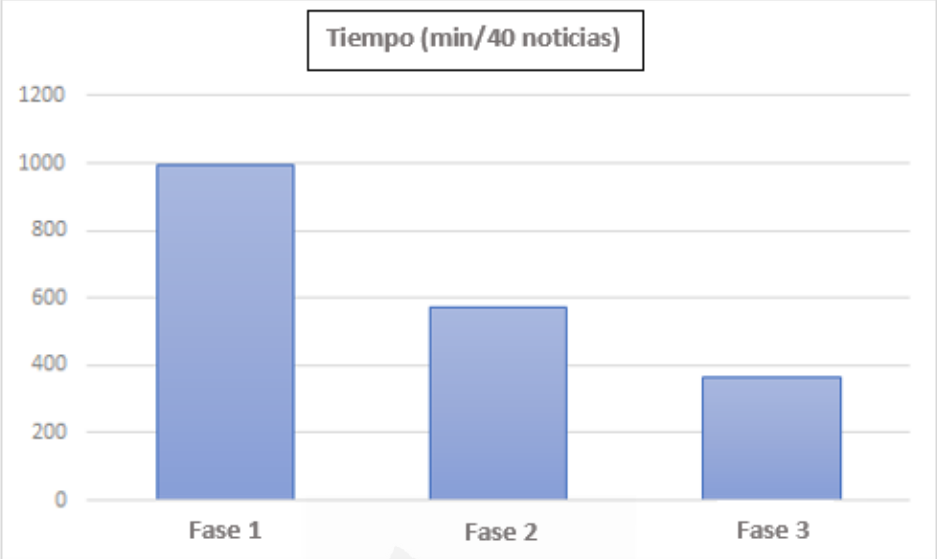


Figura 4.7: Reducción de tiempo durante el proceso de anotación para cada fase, considerando 4 lotes de 10 noticias con una media de 20 000 palabras por fase.

a cabo el proceso de anotación en la Fase 3 en comparación con la anotación totalmente manual (992,84 min/40 noticias en la Fase 1 frente a 361,25 min/40 noticias en la Fase 3). Por lo tanto, los anotadores también son más eficientes cuando tareas como la compilación automática o la preanotación los asisten durante el proceso de anotación.

Cálculo del error de preanotación

Antes de aplicar el modelo de QA, se llevó a cabo un análisis de los errores cometidos por el modelo de preanotación, de forma que se pudiese analizar qué etiquetas deben ser preanotadas automáticamente y cuáles no. Si el anotador tiene que corregir un porcentaje muy alto de un tipo de etiqueta, puede ser más conveniente que ese tipo de etiqueta no se preanote.

Para el análisis se seleccionó el lote 6 de la Fase 3 porque se obtuvo un mejor modelo de QA (de las 5W1H con *fine-tuning*) después de entrenar con las noticias de este lote. El lote se compone de 10 noticias que se anotan previamente con el modelo de QA correspondiente (QA de las 5W1H con *fine-tuning* después de la Fase 2 –modelo uno o **M1**– y el modelo de QA de las 5W1H con *fine-tuning* después del lote 6 en la Fase 3 –modelo dos o **M2**–) y se presentan al anotador experto para que verifique las etiquetas. La tasa de error de todas las etiquetas 5W1H se calculó manualmente sobre la base de las métricas mencionadas en el apartado 4.4.2: *Exact Match*, *Similar Match* o *Incorrect Match*. La Tabla 4.8 muestra la media de estas tres categorías con respecto a las etiquetas 5W1H del

recuento manual realizado con ambos modelos.

Etiquetas 5W1H	**EM**		**SM**		**IM**	
	M1	M2	M1	M2	M1	M2
WHAT	74	**88**	**15**	9	11	**4**
WHEN	28	**70**	7	**12**	40	**22**
WHERE	38	**62**	**7**	1	54	**30**
WHO	57	**84**	**16**	3	27	**12**
WHY	16	**45**	7	**10**	77	**45**
HOW	24	**46**	**13**	8	63	**46**

Tabla 4.8: Media de *Exact Match*, *Similar Match*, *Incorrect Match* en las etiquetas 5W1H del recuento manual en los lotes 6 y 7.

Cuando se analizaron los resultados del modelo **M1** en el lote 6, la media de los resultados marcados como *Exact Match* era muy baja, excepto en el caso de las etiquetas WHAT (0,74) y WHO (0,57). En el caso de las etiquetas marcadas como *Similar Match*, el problema es aún más evidente. Teniendo en cuenta las etiquetas marcadas como *Incorrect Match*, las etiquetas WHY (0,77) y HOW (0,63) son las etiquetas peor predichas, donde la mejor predicha es la etiqueta WHAT con solo 0,11 de media incorrecta. En el caso de la media de valores incorrectos, a diferencia de las otras dos categorías, hay que tener en cuenta que los mejores valores son los más pequeños porque se quieren minimizar los errores en la anotación.

Una vez finalizado el lote con estos resultados, se cambió el modelo de QA por el obtenido tras reentrenar el modelo con los seis primeros lotes de la Fase 3 (QA de las 5W1H con *fine-tuning* tras el lote 6 de la Fase 3), lo que mejoró los resultados en el conjunto *test* utilizado (Tabla 4.1). Además, se aumentó experimentalmente el umbral de 0,11 a 0,14 para comprobar si el modelo era capaz de reducir la media de etiquetas marcadas incorrectamente, reduciendo así los tiempos de anotación.

Volviendo a los resultados de la Tabla 4.8, se puede observar que tras el recuento manual de las etiquetas marcadas como *Exact Match*, *Similar Match* o *Incorrect Match*, realizado con el modelo **M2** en el lote 6, los resultados mejoran significativamente. En el caso de las marcadas como *Exact Match*, todas mejoran, siendo 0,42 el mayor incremento en el caso de la etiqueta WHEN. Por otro lado, para las etiquetas marcadas como *Similar Match* no hubo una mejora significativa en general, produciéndose una mejora sólo en el caso de las etiquetas WHEN y WHY. Esto se debe a que muchos ejemplos que se marcaron como *Similar Match* en el modelo **M1**, mejoraron la precisión y pasaron a marcarse como *Exact Match* en el modelo de predicción **M2**.

Por último, la media de las etiquetas marcadas como *Incorrect Match* en el modelo de predicción **M2** en el lote 6 disminuyó en general, pero de forma pronunciada en el caso de las etiquetas WHY, WHERE y HOW. En resumen, las eti-

quetas peor predichas son WHY y HOW, pero se obtuvo una mejora significativa en su predicción tras volver a entrenar el modelo QA con más ejemplos anotados (QA de las 5W1H con *fine-tuning* tras el lote 6 de la Fase 3). Por otro lado, aumentar el umbral a 0,14 también mejoró la precisión de la anotación. Tras este análisis, llegamos a la conclusión de que el modelo de QA ayuda al anotador y es factible preanotar todas las etiquetas 5W1H de este modo.

4.8.2 Evaluación de la metodología del corpus RUN-AS-SFN

Cálculo del tiempo

Para demostrar la eficacia de la aplicación de la metodología semiautomática en el proceso de anotación, se comparó el tiempo de anotación entre: i) la anotación manual, ii) la anotación semiautomática en noticias completas y iii) la anotación semiautomática en noticias resumidas. El tiempo medio de anotación se calculó en lotes de diez noticias para cada uno de los tipos de anotación mencionados. Los tiempos medios para cada tipo de procedimiento de anotación se muestran en la Tabla 4.9.

MÉTODO	$\overline{\textit{Tiempo}}$ **(min.)**	$\overline{\textit{Palabras}}$
Anotación manual	16,71	510,57
Semiautomática (noticias completas)	12,32	482,22
Semiautomática (resúmenes)	8,07	383,11

Tabla 4.9: Tiempo medio de anotación por noticia según cada método de anotación.

Considerando una longitud media de palabras similar, que oscila entre 350 y 550 palabras, se observa una reducción del tiempo medio de anotación por noticia de 16,71 min/noticia (anotación manual) a 12,32 min/noticia gracias a la semiautomatización de la anotación para la noticia completa. El uso de resúmenes para la anotación de texto redujo aún más el tiempo de anotación a 8,07 min/noticia. El procedimiento consigue una reducción de tiempo del 50 % desde una anotación totalmente manual a este proceso de anotación semiautomático con resúmenes.

Además hay otro factor importante de este corpus que influye en el tiempo: el tema. Como este factor influyó en la anotación se llevó a cabo un análisis más detallado de las noticias seleccionadas según el tema. Los datos específicos sobre el tiempo medio de anotación por tema se presentan en la Tabla 4.10. Por su complejidad, los temas que más dificultaron la tarea de anotación fueron: economía (9,7 min/noticia), salud (9 min/noticia), ciencia (8,75 min/noticia) y política (7 min/noticia). Esto se debe a que estos temas contienen datos numéricos y terminología específica, así como un estilo de redacción más denso para presentar la información de forma más objetiva. Por el contrario, los temas de sociedad (7,2 min/noticia), entretenimiento (6 min/noticia) y educación (5

min/noticia) se anotan a un ritmo diferente porque tienden a presentar la información en un estilo más informal, además de que estos temas no requieren tantos conocimientos previos o especializados por parte del lector.

Para anotar la confiabilidad global es necesario leer la noticia entera, lo que influye en el tiempo de anotación. En la Tabla 4.10, además de la media de tiempo, se presenta la media de palabras por tema. Como era de prever, se ha observado que los temas que contienen más palabras por noticia requieren más tiempo de anotación que los más breves, como economía (444 palabras/noticia), ciencia (401,87 palabras/noticia), salud (431,83 palabras/noticia) y política (473,60 palabras/noticia) frente a entretenimiento (298,60 palabras/noticia).

Las noticias deportivas suelen incluir muchas citas y el hilo conductor es más difícil de seguir que aquellas en las que la información se presenta en un orden más concreto. En el caso del tema de entretenimiento, las noticias buscan distraer al lector e informar de forma más sencilla. Otro factor influyente en el análisis es la lengua del corpus. La mayoría de las noticias del corpus elegido para la anotación están redactadas en español latinoamericano, lo que puede provocar cierta dificultad de comprensión para el anotador (con un perfil lingüístico español de España) y retrasar la anotación de ciertos temas con información cultural relacionada con deportistas, políticos, famosos o sociedad. Sin embargo, aunque esto puede limitar la comprensión y ralentizar la tarea de anotación, también permite que el anotador, al tener un menor conocimiento de esa cultura, no se vea influido por el contexto o el conocimiento del mundo ya adquirido, lo que le permite ser más objetivo a la hora de clasificar la confiabilidad de los datos.

TEMA	$\overline{\textit{Tiempo}}$ **(min)**	$\overline{\textit{Palabras}}$
Economía	9,7	444,00
Deportes	8,5	326,70
Ciencia	8,75	401,87
Salud	9	431,83
Sociedad	7,2	365,80
Entretenimiento	6	298,60
Política	7	473,60
Seguridad	7	314,50
Educación	5	285,33

Tabla 4.10: Tiempo medio de anotación y media de palabras de noticia por tema.

Cálculo del error de preanotación

Al igual que en el apartado 4.8.1, se midió la tasa de error de la preanotación dos veces. Para evaluar estas medidas, se seleccionaron ocho noticias y se

preanotaron con el modelo **M1**, teniendo en cuenta las tres categorías mencionadas anteriormente (*Exact Match*, *Similar Match*, *Incorrect Match*). Tras anotar 150 noticias para el conjunto de *training* y 100 noticias para el conjunto de *test*, estas se utilizaron para volver a entrenar el modelo 5W1H sin las ocho noticias seleccionadas para medir la tasa de error. Tras eso, se obtuvo el modelo **M2** y se inició el segundo entrenamiento de HITL. Por último, se midió la misma preanotación utilizando el modelo **M2**.

Etiquetas 5W1H	**EM**		**SM**		**IM**	
	M1	M2	M1	M2	M1	M2
WHAT	87,27	**89,39**	**12,72**	9,09	**0**	1,51
WHEN	**84,31**	81,48	**3,92**	3,70	**11,76**	14,81
WHERE	65,57	**75,80**	**8,19**	8,06	26,22	**16,21**
WHO	87,5	**95,71**	**6,94**	2,85	5,55	**1,42**
WHY	32,43	**39,47**	**2,70**	2,63	64,86	**57,89**
HOW	42,85	**48,97**	**11,42**	8,16	45,71	**42,85**

Tabla 4.11: Media de etiquetas 5W1H clasificadas en *Exact Match*, *Similar Match* o *Incorrect Match* del recuento manual del rendimiento de los modelos M1 y M2.

Según los resultados presentados en la Tabla 4.11, utilizando el modelo **M2**, la precisión de preanotación mejora significativamente, como indica el recuento manual de las etiquetas marcadas como *Exact Match*, *Similar Match* o *Incorrect Match*. Para las marcadas como *Exact Match*, todas las etiquetas mejoran excepto la etiqueta WHEN (que disminuye en 2,83 puntos porcentuales), obteniendo el mayor incremento en puntos porcentuales (10,23) la etiqueta WHERE. En el caso de las etiquetas marcadas como *Similar Match* no se produjo ninguna mejora, por la misma razón comentada en 4.8.1. Por último, la media de las etiquetas marcadas como *Incorrect Match* en el modelo de predicción **M2** disminuyó en el caso de las etiquetas WHERE, WHO, WHY y HOW. Aunque hay un pequeño incremento en el error de las etiquetas WHAT y WHEN, el hecho de que el sistema sea capaz de detectar correctamente un número significativamente mayor de etiquetas (es decir, más etiquetas clasificadas como *Exact Match*) tiene relevancia, ya que es más fácil eliminar una etiqueta errónea de la anotación que tener que anotar un elemento desde cero. Además, tal y como se presenta en la Tabla 4.4, el total de coincidencias exactas y el $F1$ del modelo **M2** superaron al modelo **M1**.

4.9 Conclusiones

La viralización de grandes volúmenes de desinformación hace necesaria la automatización de su detección. Para ello, el modelo de detección necesita de la retroalimentación del experto humano, que se consigue a partir de ejemplos anotados extraídos de un corpus, y del entrenamiento de algoritmos de ML y

DL. No obstante, la tarea de anotación es una tarea costosa y lenta, por lo que la escasez de corpus en la tarea de detección de desinformación, especialmente en español, es uno de los mayores retos para los investigadores de PLN.

A esto hay que sumarle la complejidad semántica de las guías de anotación presentadas en esta investigación. Por ello, este capítulo se ha enfocado en presentar una metodología de anotación semiautomática basada en diferentes estrategias *Human-in-the-loop* que permite asistir al anotador y facilitar la ardua tarea de construir y anotar un corpus. Esta metodología ha propiciado la creación de forma semiautomática de un corpus de noticias en español anotadas con el criterio de confiabilidad y ha reducido el tiempo del anotador en la tarea.

El principal objetivo en este capítulo es implementar un procedimiento de anotación semiautomático que permita obtener un recurso de calidad para la detección de desinformación combinando la anotación automática y la manual según criterios específicos que aumente, o al menos mantenga, la precisión al tiempo que ahorra tiempo y esfuerzo al anotador. Esto se consigue especialmente con el proceso de HITL, el cual permite un entrenamiento en bucle entre el experto humano y la máquina, de forma que la máquina entrene cada vez que el experto anota y, por lo tanto, pueda ir aprendiendo de las correcciones o ejemplos del anotador humano. Esto se consigue especialmente gracias a la estrategia de *Active Learning*, que permite que el modelo elija los ejemplos que considera más útiles y le proponga al anotador los siguientes ejemplos a anotar.

Este capítulo se centra especialmente en la construcción del corpus RUN, el recurso generado con la guía de anotación RUN-AS y la metodología semiautomática propuesta. A diferencia del primer corpus generado para esta investigación (corpus FNDeep), construido manualmente, el corpus RUN se obtiene tras implementar una metodología semiautomática basada en tres fases que van integrando gradualmente la automatización de las tareas de construcción de corpus. Por otro lado, se evalúa también la guía RUN-AS en un corpus de noticias anotadas con el criterio de veracidad, con el fin de saber si la detección de indicadores de confiabilidad a nivel lingüístico son útiles en la tarea de detección de noticias falsas. Para ello, se introduce la metodología empleada para la anotación con RUN-AS de un subconjunto de este corpus público (al cual le hemos denominado RUN-AS-SFN), empleando además la técnica de extracción de resúmenes para agilizar el proceso de anotación.

Finalmente, se evalúa la metodología semiautomática de ambos corpus, tanto del RUN como del RUN-AS-SFN, siguiendo dos métricas: (i) el cálculo del tiempo empleado en la compilación y anotación de los corpus, en el que se observa una disminución del tiempo gracias a la preanotación y la extracción de resúmenes, y (ii) el cálculo del error de preanotación del modelo de QA, lo que permite ajustarlo de forma que se anote correctamente el mayor número de etiquetas para que el anotador no pierda más tiempo corrigiendo que anotando. Para cerrar la tesis, los esquemas de anotación y los recursos presentados a lo largo de este trabajo se evalúan a continuación en el Capítulo 5, el cual detalla toda la experimentación de la presente investigación.

CAPÍTULO 5

Marco de evaluación para la detección de la desinformación

5.1 Introducción

En este último capítulo se presenta el marco de evaluación, tanto para la primera versión de nuestra guía de anotación (FNDeepML) como para la segunda (RUN-AS), y se explica la arquitectura diseñada para el modelo de detección de desinformación. Además, se presenta toda la experimentación llevada a cabo en los tres recursos utilizados en esta investigación: (i) el corpus FNDeep, (ii) el corpus RUN (ambos diseñados y anotados *ad hoc*) y (iii) el corpus publicado por (Posadas-Durán y cols., 2019), en el que se ha aplicado nuestro enfoque de anotación y probado su eficacia en la detección de noticias falsas (corpus RUN-AS-SFN Dataset). Toda la experimentación y evaluación desarrollada en este capítulo demuestra la viabilidad de nuestra propuesta de anotación en la tarea de detección de la desinformación.

Para evaluar los corpus, se tienen en cuenta cuatro métricas habituales en PLN, que son *precision* (precisión), *recall* (exhaustividad), *F-measure* (media F1) y *accuracy* (exactitud).

- ***Precision***: *Precision* (P) es el ratio entre las observaciones positivas predichas correctamente y el total de observaciones positivas predichas.

$$P = \frac{\#VerdaderoPositivo}{\#VerdaderoPositivo + \#FalsoPositivo} \tag{5.1}$$

- ***Recall***: *Recall* (R) es el ratio entre las observaciones positivas predichas correctamente y todas las observaciones que son realmente positivas.

$$R = \frac{\#VerdaderoPositivo}{\#VerdaderoPositivo + \#FalsoNegativo} \tag{5.2}$$

- ***F1-score***: *F1-Score* (F_1) es la media ponderada de *Precision* y *Recall*.

$$F_1 = \frac{2 * Precision * Recall}{Precison + Recall} \tag{5.3}$$

- ***Accuracy***: *Accuracy* (Acc) es la medida de rendimiento más intuitiva y es el ratio entre las observaciones predichas correctamente y el total de observaciones.

$$Acc = \frac{\#VerdaderoP + \#VerdaderoN}{\#VerdaderoP + \#FalsoP + \#VerdaderoN + \#FalsoN} \tag{5.4}$$

Además, cuando es necesario, se indican las medias macro y micro de cada medida. La media macro es la media de cada una de las medidas, mientras que la media micro es una media ponderada por el valor de soporte (que es el número de instancias verdaderas para cada etiqueta). El uso de estas medidas también es importante porque la media macro será baja si alguna clase es pequeña, pero la media micro penalizará menos en clases con muy pocos elementos. La diferencia entre macro y micro indica cuánto afecta el desequilibrio del corpus al modelo.

5.2 Evaluación del esquema FNDeepML

Al inicio de la presente investigación, tal y como se presenta en la sección 3.1.1, nos enfocamos en el diseño de una guía de anotación centrada en asignar un valor de veracidad en función de la información externa proporcionada por desmentidos, por lo que el recurso anotado con esta anotación está formado por un conjunto de noticias verdaderas y falsas. Aunque el enfoque inicial se reorientó (Sección 3.2), se llevó a cabo una amplia experimentación con el recurso anotado con la versión inicial FNDeepML y se construyó una arquitectura basada en varias fases con la finalidad de probar la predicción de la veracidad de las noticias en función de la anotación aplicada.

5.2.1 Diseño e implementación de la arquitectura del modelo de detección

Para evaluar la guía de anotación FNDeepML diseñada para la clasificación de la veracidad de noticias verdaderas y falsas, se diseñó un sistema de detección con una arquitectura de dos capas basada en un *pipeline*, el cual permite segmentar un proceso de cálculo en varios subprocesos que son ejecutados por unidades autónomas (Ramamoorthy y Li, 1977). La arquitectura consta de cinco fases diferentes, estructuradas en dos capas, y se representa gráficamente en la Figura 5.1.

Las dos capas y sus correspondientes fases de la arquitectura son las siguientes:

- **Capa de Estructura**: Esta capa se encarga de estructurar el texto según los dos niveles de representación de la información. En primer lugar, se divide

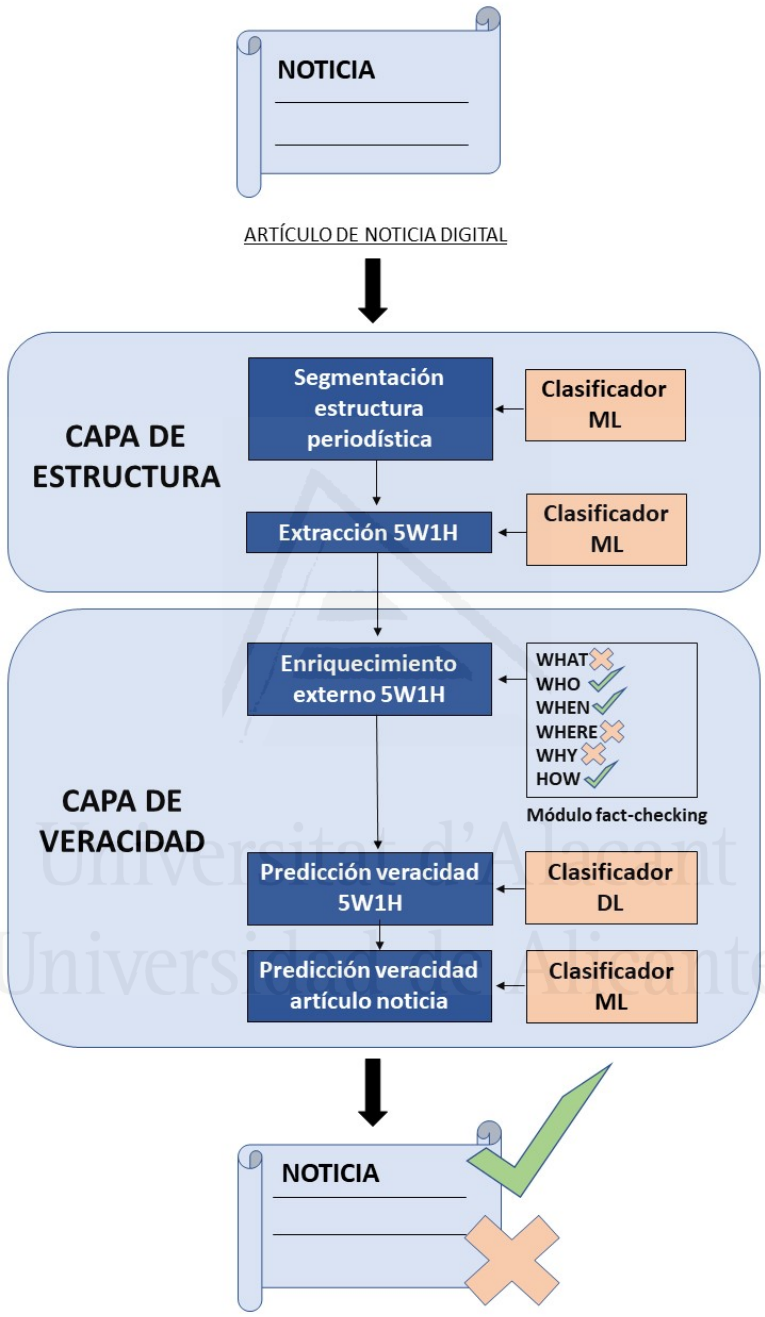


Figura 5.1: Arquitectura del sistema de detección de noticias falsas.

la noticia según la estructura periodística y, a continuación, se determinan los elementos 5W1H de cada parte de la estructura.

  - *Fase 1. Segmentación de la estructura periodística*: teniendo como entrada una noticia de un medio digital tradicional, este primer módulo se encarga de dividir la noticia en las partes de la estructura definida en la guía FNDeepML. Por lo tanto, la salida de este módulo es la noticia dividida en titular, subtítulo, entradilla, cuerpo de la noticia y conclusión.
  - *Fase 2. Extracción del contenido esencial (5W1H)*: teniendo como entrada la noticia dividida en partes, este módulo extrae los elementos 5W1H de cada parte de la noticia.

- **Capa de Veracidad**: En esta capa el objetivo es determinar la veracidad de cada una de las partes previamente detectadas, así como predecir la veracidad de la noticia utilizando la veracidad de los diferentes elementos.

  - *Fase 3. Enriquecimiento externo del contenido esencial (5W1H)*: teniendo los elementos 5W1H de la noticia, este módulo se encarga de enriquecer la información de cada elemento utilizando conocimiento externo mediante la verificación de datos.
  - *Fase 4. Predicción de la veracidad del contenido esencial (5W1H)*: este módulo, tras utilizar la anotación de todas las características posibles (características textuales y conocimiento externo) de los elementos 5W1H, clasifica cada elemento con un valor de veracidad.
  - *Fase 5. Predicción de la veracidad del artículo de la noticia*: el último módulo, tras utilizar la clasificación de veracidad de cada elemento, se encarga de predecir la veracidad de la noticia completa, que es la salida final de la arquitectura propuesta.

En los siguientes apartados se explica con más detalle el desarrollo de cada una de las fases mencionadas.

Fase 1: Segmentación de la estructura periodística

En esta fase se estructura la noticia de acuerdo con la estructura periodística presentada anteriormente en la sección 2.2.1. La noticia original se extrae a través de un *web crawler* que permite obtener un texto con el titular y el subtítulo. El texto restante se divide en entradilla, cuerpo y conclusión para después, utilizando la biblioteca Spacy[1], realizar una tokenización de la noticia y obtener un conjunto de características para cada token (Tabla 5.1). Las características se definen en la documentación de la biblioteca Spacy.

[1] `https://spacy.io/`

Característica	Descripción
`text`	Texto original del token.
`lemma`	Versión lematizada del token.
`pos`	Etiqueta de la parte del discurso, ej. VERBO, NOMBRE, etc.
`tags`	Etiquetas precisas de la parte del discurso, como persona, número, tiempo, etc.
`dep`	Etiqueta del token en el árbol de dependencia.
`shape`	Representación sintáctica de la forma del token.
`ent_type`	Etiqueta de entidad de uso general, ej. PERSONA, ORG, etc.
`is_alpha`	Valor booleano que indica si el token es alfanumérico.
`is_stop`	Valor booleano que indica si el token es un stopword.
`index`	Índice relativo del token en el documento, entre 0 (primer token) y 1 (último token).

Tabla 5.1: Características a nivel de token extraídas con Spacy.

La segmentación a nivel de token se lleva a cabo mediante un modelo de *Conditional Random Fields* (CRF) (Sutton, McCallum, y cols., 2012) entrenado con las características de los tokens descritas en la Tabla 5.1. Para introducir contexto, cada conjunto de rasgos de los tokens se complementa con los rasgos de los tokens circundantes (tanto antes como después) en una pequeña ventana de tamaño 0 a 3. El modelo CRF se entrena utilizando *sklearn-crfsuite*[2]. El problema de segmentación se modela como un problema de etiquetado de secuencias, en el que a cada token se le asigna una de estas etiquetas: LEAD, BODY y CONCLUSION.

Tras este proceso, se obtiene como salida de este módulo la noticia segmentada, como se muestra en el siguiente ejemplo:

(1)

Token	Características	Parte Estructura
token1	| ... |	=> LEAD
token2	| ... |	=> LEAD
token3	| ... |	=> BODY
token4	| ... |	=> BODY
...		
tokenN	| ... |	=> CONCLUSION

Fase 2: Extracción del contenido esencial (5W1H)

Utilizando todas las características por token obtenidas anteriormente, se utiliza un segundo modelo CRF para clasificar cada token de cada parte en uno de los elementos 5W1H, o en ninguno (NONE). Como se observa en la Tabla 3.3 de la Sección 3.1.2, existe un gran desequilibrio en la distribución de las etiquetas del corpus FNDeep, lo que provoca un rendimiento pobre de los modelos entrenados para predecir todas las clases a la vez. Por este motivo, se lleva a cabo una clasificación jerárquica en dos niveles, en la que las etiquetas se dividen en dos conjuntos: el primer nivel está formado por las etiquetas más comunes (NONE y WHAT), mientras que las etiquetas menos comunes se agrupan en una

[2] https://sklearn-crfsuite.readthedocs.io/en/latest/

clase especial, denominada restante (REST); y el segundo nivel comprende sólo las clases menos comunes (HOW, WHEN, WHERE, WHY y WHO). Esto permite entrenar dos modelos distintos que pueden tratar mejor la distribución desbalanceada de las etiquetas, permitiendo que cada modelo se centre únicamente en un conjunto más pequeño de clases para las que sus números relativos son similares.

El hecho de que una de las características obtenidas por Spacy sean las entidades nombradas es muy útil en este módulo, ya que están relacionadas con algunas de las preguntas de las 5W1H, como la ubicación para la etiqueta WHERE, la persona/organización para la etiqueta WHO o el tiempo para la etiqueta WHEN. Además, se utilizan las mismas características que se muestran en la Tabla 5.1 para representar cada token. Asimismo, se puede ajustar el tamaño de la ventana para incluir más contexto.

En el caso de clases con un conjunto menor de ejemplos en el corpus, además de las características utilizadas en el primer nivel, también se utilizarán los roles semánticos del texto en el segundo nivel del modelo jerárquico. Según (Moreda, Llorens, Saquete, y Palomar, 2011), el uso de roles semánticos puede mejorar la detección de respuestas de las 5W1H, especialmente cuando se trata de preguntas cuya respuesta no es en sí misma una entidad nombrada. Por ejemplo, en esta frase en la que se anotan los roles semánticos, el rol `AM_LOC` es la respuesta a la pregunta WHERE (dónde):

(2) ¿Dónde nació Pitágoras? Pitágoras de Samos nació [`AM-LOC` en la isla de Samos].

Para anotar los roles semánticos se utilizó Freeling (Padró y Stanilovsky, 2012), pues esta herramienta también anota roles semánticos en español. A continuación se presenta un ejemplo de la salida obtenida por este módulo para cada noticia. Como se puede observar, cada 5W1H puede abarcar varios tokens, como en el caso de la etiqueta WHAT, que abarca los tokens 3, 4 y 5:

(3)

Token	**Características**	**Parte Estructura**	**Etiquetas 5W1H**
token1	| ... |	LEAD	=> NONE
token2	| ... |	LEAD	=> WHAT
token3	| ... |	BODY	=> WHAT
token4	| ... |	BODY	=> WHAT
token5	| ... |	BODY	=> WHAT
token6	| ... |	BODY	=> WHO
...			
tokenN	| ... |	CONCLUSION	=> WHERE

Fase 3: Enriquecimiento externo del contenido esencial (5W1H)

Este módulo se encarga de enriquecer cada elemento 5W1H utilizando conocimiento externo mediante *fact-checking*. Como el objetivo es buscar sólo el contenido esencial, es decir, el tratamiento de cada elemento 5W1H, el proceso se lleva a cabo utilizando esos elementos y no el texto plano. Hay que destacar que desarrollar un módulo de *fact-checking* no es una tarea trivial y requiere de una investigación en profundidad en sí misma, lo cual queda fuera del alcance

de este trabajo. No obstante, con el fin de añadir conocimiento externo a la arquitectura propuesta, se ha implementado un módulo sencillo de *fact-checking* capaz de detectar si los elementos 5W1H de una noticia forman parte de algún desmentido. Evidentemente, esto implica que el hecho expuesto en la noticia ha tenido que ser refutado previamente. El objetivo de este módulo no es determinar la veracidad de cada 5W1H, sino extraer información externa que, sumada al contenido textual, ayude en la predicción de la veracidad de cada elemento realizada en la Fase 4.

Este módulo utiliza la API de la herramienta Google Fact Check[3] que se basa en el esquema ClaimReview[4]. Un ejemplo de un desmentido en español se muestra en la Figura 5.2.

Figura 5.2: Ejemplo de un desmentido en español.

El contenido esencial de la noticia (5W1H) se busca del siguiente modo. Para cada parte del documento (titular, subtítulo...), los elementos 5W1H se envían por separado a la API para comprobar su veracidad. Si se encuentra un valor, la etiqueta se actualiza con ese valor. Si alguno de los elementos 5W1H no recibe un valor de veracidad o recibe valores contradictorios, se realizará una segunda comprobación con todos los elementos 5W1H de esa parte para añadir información de contexto. Para ello, se concatenarán todos los elementos y se enviarán de nuevo para comprobar su veracidad. En este caso, el valor obtenido servirá para actualizar el valor de veracidad de cada elemento. La calificación textual de la API se asigna a una de nuestras categorías: Verdadero, Falso, Desconocido.

Fase 4: Predicción de la veracidad del contenido esencial (5W1H)

Esta fase se encarga de predecir el valor de veracidad de cada elemento 5W1H de las noticias, basándose en todas las evidencias recogidas en la Fase 3 más el contenido textual de cada elemento. Debido a la complejidad de la tarea, este problema se aborda utilizando DL, ya que la resolución del problema en esta fase requiere no sólo tratar con características textuales de los elementos sino también con características de alto nivel obtenidas a partir de conocimiento externo que enriquece los elementos (*fact-checking* en este caso). Para predecir

[3] https://developers.google.com/fact-check/tools/api/
[4] https://schema.org/ClaimReview

la veracidad de cada elemento, el módulo utiliza un modelo secuencial de tipo *Long Short-Term Memory* (LSTM) y otro convolucional. La configuración de los parámetros se detalla en el Apéndice C.

El modelo de DL se entrena independientemente en cada secuencia continua de tokens que pertenece a la misma parte 5W1H para predecir su valor de veracidad. Al final, utilizando todas las características extraídas previamente en el *pipeline*, el módulo predice la veracidad de cada elemento.

Un ejemplo de la salida de este módulo es:

(4)

Token	**Características**	**Parte Estructura**	**5W1H**	**Veracidad**
token1	| ... |	LEAD	=> NONE	-
token2	| ... |	LEAD	=> WHAT	T
token3	| ... |	BODY	=> WHAT	T
token4	| ... |	BODY	=> WHAT	T
token5	| ... |	BODY	=> WHAT	T
token6	| ... |	BODY	=> WHO	F
...				
tokenN	| ... |	CONCLUSION	=> WHERE	T

En este ejemplo, a la etiqueta WHAT de la entradilla se le asigna un valor de veracidad Verdadero, al WHAT del cuerpo se le asigna un valor Verdadero y a la etiqueta WHO un valor Falso. Esto significa que el acontecimiento presentado en el cuerpo de la noticia tiene lugar, pero la persona implicada no era la indicada en la noticia.

Fase 5: Predicción de la veracidad del artículo de la noticia

Finalmente, la última fase se encarga de dar la predicción final de la noticia, utilizando uno de los varios modelos clásicos de ML: Regresión Logística —*Logistic Regression*— (LR), *Decision Tree* (DT), Máquinas de Soporte Vectorial —*Support Vector Machine*— (SVM), *Multinomial Naive Bayes* (MNB) o un modelo de referencia aleatorio. La configuración de los parámetros se detalla en el Apéndice C.

En este módulo, para representar los documentos, para cada parte de la estructura del documento se agregan sus elementos 5W1H según su valor de veracidad y se cuenta el número de cada elemento dentro de cada valor de veracidad. Así, considerando que hay 5 partes en la estructura (titular, subtítulo, entradilla, cuerpo y conclusión) y 6 posibles tipos de elementos 5W1H (qué, quién, cuándo, dónde, por qué y cómo) dentro de cada parte, y que cada uno de estos elementos 5W1H puede tener uno de los tres valores de veracidad (Verdadero, Falso, Desconocido), el número final de características numéricas generadas es de 90.

Si tenemos en cuenta la anotación del titular y de la entradilla de una noticia, un ejemplo de un subconjunto de las características numéricas extraídas serían las siguientes:

```
{
    TITLE_WHAT_TRUE: 0,
    TITLE_WHAT_FALSE: 1,
```

```
    TITLE_WHAT_UNKNOWN: 0,
    TITLE_WHO_TRUE: 0,
    TITLE_WHO_FALSE: 1,
    TITLE_WHO_UNKNOWN: 0,
    TITLE_WHEN_TRUE: 0,
    TITLE_WHEN_FALSE: 1,
    TITLE_WHEN_UNKNOWN: 0,
    # ...
    LEAD_WHAT_TRUE: 2,
    LEAD_WHAT_FALSE: 2,
    LEAD_WHAT_UNKNOWN: 0,
    LEAD_WHO_TRUE: 0,
    LEAD_WHO_FALSE: 1,
    LEAD_WHO_UNKNOWN: 1,
    LEAD_WHEN_TRUE: 0,
    LEAD_WHEN_FALSE: 3,
    LEAD_WHEN_UNKNOWN: 0,
    # ...
}
```

El mismo tipo de características se generará a partir de las demás partes de la estructura del documento. Cada característica indica el número de elementos 5W1H con una etiqueta y veracidad específicas que aparecen en cada parte de la noticia. Por ejemplo, `LEAD_WHAT_TRUE: 2`, indica que el LEAD contiene dos elementos WHAT anotados con un valor de veracidad True. El modelo se entrena para predecir la etiqueta de veracidad global del documento basándose en estas características numéricas.

5.2.2 Experimentación con el corpus FNDeep

El principal objetivo de la experimentación propuesta en esta sección es demostrar la hipótesis de que, dado que las noticias falsas son una combinación de información falsa y verdadera cuyo objetivo es crear confusión entre los lectores, una aproximación adecuada al problema de la detección automática de noticias falsas sería una arquitectura de dos capas. Para cada una de las capas, se llevaron a cabo los siguientes experimentos:

- **Rendimiento de la Capa de Estructura**: se ha llevado a cabo una serie de experimentos relacionados con las dos primeras fases para evaluar posibles áreas de mejora que aumenten la eficacia.
  - *Rendimiento de la Fase 1. Segmentación de la estructura periodística*: se mide el rendimiento del módulo que realiza la segmentación en entradilla, cuerpo y conclusión del texto.
  - *Rendimiento de la Fase 2. Extracción del contenido esencial (5W1H)*: en este experimento se mide el rendimiento en la detección de los distintos segmentos que corresponden a las respuestas de las 5W1H.

- **Rendimiento de la Capa de Veracidad**: se ha implementado un conjunto de experimentos para medir las dos fases que determinan tanto la veracidad de los elementos como la veracidad de la noticia. Además, un experimento final permite determinar la validez de la hipótesis de este trabajo midiendo la capa de veracidad en su conjunto.
  - *Rendimiento de la Fase 3. Enriquecimiento externo de las 5W1H:* la Fase 3 no tiene un experimento individual ya que es una fase de enriquecimiento y su validez viene dada por los resultados de la Fase 4, que se han medido utilizando tanto la información de la Fase 3 como sin utilizarla para determinar sus beneficios.
  - *Rendimiento de la Fase 4. Predicción de la veracidad del contenido esencial (5W1H):* este experimento mide el rendimiento del módulo que predice el valor de veracidad de cada elemento de la noticia. Para probarlo y determinar la validez de este módulo de forma aislada, se han utilizado las etiquetas 5W1H del corpus *gold standard* y se ha medido el rendimiento del módulo utilizando diferentes configuraciones: i) utilizando sólo las características textuales del contenido de los elementos 5W1H; ii) utilizando sólo las características de *fact-checking* y iii) utilizando la combinación de ambas.
  - *Rendimiento de la Fase 5. Predicción de la veracidad del artículo de noticia:* para medir la precisión de esta fase en este experimento, la fase se mide de forma aislada, utilizando como entrenamiento las noticias *gold standard* anotadas manualmente con las distintas partes de la estructura y los elementos 5W1H con su valor de veracidad. De este modo, se evitan los errores de las fases anteriores y se mide la validez de este módulo por separado. Este es uno de los experimentos más importantes, ya que demuestra la validez de la propuesta.
  - *Rendimiento de las Fases 3+4+5. Capa de Veracidad.* Este experimento pretende determinar la eficacia de la Capa de Veracidad pero evitando los errores de segmentación producidos por la capa de estructura. En concreto, utilizando la segmentación *gold standard* del texto, se realizan y miden conjuntamente las Fases 3, 4 y 5.

Finalmente, se mide el rendimiento de todo el *pipeline* y se realiza una validación entre dominios para explorar la aplicabilidad de nuestra propuesta en otros dominios.

5.2.3 Resultados y discusión de la evaluación del corpus FNDeep

En esta sección se presentan los resultados obtenidos en cada uno de los experimentos descritos en la sección 5.2.2 y una discusión de dichos resultados.

Rendimiento de la Fase 1. Segmentación de la estructura periodística

La Tabla 5.2 presenta el rendimiento a nivel de token del módulo de segmentación de la estructura que corresponde a la Fase 1 del *pipeline*.

Características	P	R	F_1	Acc
LEAD	0,851	0,772	0,810	0,851
BODY	0,960	0,964	0,962	0,929
CONCLUSION	0,710	0,836	0,768	0,648
media macro	0,840	0,857	0,846	0,809

Tabla 5.2: Rendimiento de la segmentación de la estructura periodística.

En general, este módulo obtiene una puntuación macro F_1 de 0,809 en un conjunto de prueba independiente del 20 % de las noticias. La configuración de los parámetros se detalla en el Apéndice C.

Rendimiento de la Fase 2. Extracción del contenido esencial (5W1H)

La Tabla 5.3 presenta el rendimiento a nivel de token del módulo de segmentación de las 5W1H que corresponde a la Fase 2 en el *pipeline*. Como se explica en la Fase 2, se entrena un modelo jerárquico en diferentes subconjuntos de clases para abordar el desequilibrio de etiquetas en el corpus.

El modelo jerárquico se entrena primero sólo en NONE, WHAT y REST (que agrupa todas las etiquetas restantes), en un conjunto de prueba del 20 % de las noticias, y después se entrena un segundo modelo sólo en el subconjunto de tokens con las etiquetas HOW, WHY, WHEN, WHERE y WHO en el mismo conjunto de prueba, dando lugar a los resultados que se muestran en la Tabla 5.3. El primer nivel utiliza únicamente las características sintácticas y semánticas de Spacy, mientras que el segundo nivel incluye también las características de roles semánticos de Freeling. Esta configuración mostró mejores resultados, probablemente porque los roles semánticos no son útiles para el reconocimiento de la etiqueta WHAT, a diferencia del resto de elementos 5W1H. Como puede observarse, cada modelo es significativamente mejor (en términos de macro F_1) en el subproblema correspondiente.

Curiosamente, el primer nivel es capaz de reconocer exactamente la clase REST, lo que significa que podemos estimar el rendimiento global del modelo agregando los resultados de ambos modelos. La media macro estimada combinada F_1 para este modelo de dos niveles es de 0,661. Además, el peor rendimiento se obtiene con las etiquetas HOW y WHY, que tienen el menor número de instancias. Si descartamos estas etiquetas y sólo tenemos en cuenta las cinco etiquetas restantes (incluida NONE), la macro global F_1 sería de 0,774. Por último, se obtiene un buen rendimiento con la etiqueta WHAT (F_1=0,948), que es un elemento importante a la hora de determinar la veracidad de una noticia.

	Primer nivel		
	P	R	F_1
NONE	0,999901	0,983090	0,991425
WHAT	0,901966	0,999378	0,948177
REST	1,000000	1,000000	1,000000
media macro	0,967289	0,994156	0,979867

	Segundo nivel		
	P	R	F_1
HOW	0,262500	0,750000	0,388889
WHEN	0,788732	0,629213	0,700000
WHERE	0,489583	0,566265	0,525140
WHY	0,336957	0,462687	0,389937
WHO	0,844444	0,639731	0,727969
media macro	0,544443	0,609579	0,546387

Tabla 5.3: Resultados del primer nivel y del segundo nivel del modelo jerárquico entrenado para la extracción de las 5W1H.

Las etiquetas HOW y WHY son importantes durante el proceso de verificación de datos para determinar la veracidad de una noticia, ya que añaden detalles y, por tanto, podrían cambiar el significado de la noticia. Por este motivo, es probable que su detección fallida en esta fase provoque una disminución significativa del rendimiento general del *pipeline*. En cambio, una alta precisión en la extracción de la etiqueta WHAT podría compensar la pérdida de rendimiento.

Rendimiento de la Fase 4. Predicción de la veracidad del contenido esencial (5W1H)

Esta fase se evalúa con diferentes configuraciones. Utilizando los elementos 5W1H *gold standard* del corpus, se mide la validez de este módulo de forma aislada.

La Tabla 5.4 presenta el rendimiento del módulo de la predicción de la veracidad del contenido esencial (5W1H) con tres configuraciones:

Deep NN (Texto) utiliza únicamente características textuales de los tokens dentro de cada elemento 5W1H anotado en el corpus *gold standard*.

Deep NN (FC) utiliza únicamente características de *fact-checking* de los elementos 5W1H, obtenidas automáticamente en la Fase 3.

Deep NN (Combinado) utiliza tanto las características textuales como las de *fact-checking* de los elementos 5W1H.

Con fines comparativos, se implementan dos modelos de referencia, usando una estrategia que siempre predice la clase mayoritaria (*Dummy*) y usando la representación *Term Frequency-Inverse Document Frequency* (TF-IDF) del texto de cada elemento 5W1H para entrenar con LR. Los valores corresponden a la

media de *precision*, *recall*, F_1 y *accuracy* de cada modelo para cada etiqueta de veracidad (es decir, ***U**nknown*, ***T**rue* y ***F**alse)*, situados en una media de 10 ejecuciones independientes con un 80 % de entrenamiento y un 20 % de prueba.

Modelos	Modelo de referencia		Deep Learning		
	Dummy	TF-IDF	Texto	FC	Combinado
Precision (T)	0,000	0,601	0,592	0,370	0,592
Recall (T)	0,000	0,471	0,547	0,930	0,523
F_1 (T)	0,000	0,528	0,565	0,529	0,554
Precision (F)	0,000	0,476	0,512	0,000	0,507
Recall (F)	0,000	0,234	0,374	0,000	0,452
F_1 (F)	0,000	0,313	0,424	0,000	0,468
Precision (U)	0,513	0,630	0,733	0,512	0,753
Recall (U)	1,000	0,837	0,837	0,993	0,821
F_1 (U)	0,678	0,719	0,780	0,675	0,784
Accuracy	0,513	0,607	0,658	0,542	**0,660**
Macro_F_1	0,226	0,520	0,590	0,409	**0,602**

Tabla 5.4: Resultados del rendimiento de diferentes configuraciones de la predicción de veracidad de las 5W1H utilizando la segmentación 5W1H *gold standard*.

Como se deduce de los resultados obtenidos en la Tabla 5.4, determinar la veracidad de cada uno de los contenidos esenciales de una noticia no es una tarea trivial. Las cifras obtenidas reflejan que el uso de herramientas textuales para la predicción de veracidad es prometedor, pero aún necesita de cierto margen de mejora. Asimismo, el uso del *fact-checking* con las herramientas actuales disponibles, aunque proporciona una mejoría, esta es bastante limitada, por lo que sería interesante profundizar en las características textuales.

Hay que señalar que, aunque el corpus tiene un tamaño limitado (200 noticias), esta fase se entrena con frases individuales de las 5W1H; por lo tanto, hay un mayor número de ejemplos de entrenamiento. En total hay 2788 frases 5W1H diferentes, de las cuales 2230 se utilizan para el entrenamiento (80 %) y 558 para la validación (20 %).

Para comprender mejor el comportamiento del módulo de predicción de veracidad de las 5W1H en diferentes tipos de elementos 5W1H, la Tabla 5.5 muestra las métricas de evaluación agregadas por etiqueta 5W1H.

Los resultados obtenidos entre los distintos elementos 5W1H son bastante similares, excepto en el caso de la etiqueta WHO, que, como se indica en la Tabla 3.3, presenta un alto grado de incertidumbre (valor de veracidad desconocido), lo que se traduce en un *accuracy* limitado. Los resultados indican la necesidad de añadir información más compleja que implique conocimiento externo y contexto para mejorar la predicción de la veracidad de cada elemento.

Etiqueta 5W1H		HOW	WHEN	WHERE	WHY	WHAT	WHO
macro	precision	0,496	0,610	0,573	0,630	0,502	0,332
	recall	0,492	0,495	0,477	0,609	0,491	0,333
	F_1	0,442	0,485	0,457	0,573	0,485	0,332
	accuracy	0,495	0,558	0,535	0,571	0,505	0,197

Tabla 5.5: Métricas de evaluación del modelo de predicción de veracidad de las 5W1H utilizando características sintácticas y de *fact-checking* combinadas y agregadas por tipo de elemento 5W1H.

Rendimiento de la Fase 5. Predicción de la veracidad del artículo de la noticia

Los experimentos con el módulo de predicción de la veracidad del artículo de la noticia representan los resultados más interesantes porque demuestran que, al considerar la veracidad de las partes de la estructura de la noticia y las 5W1H, se proporciona una solución adecuada al problema de la detección automática de noticias falsas. Por lo tanto, para evitar los problemas derivados de las fases anteriores, este módulo se mide de forma aislada utilizando los elementos 5W1H *gold standard* y su valor de veracidad asignado manualmente. El experimento demuestra que esta información es valiosa a la hora de determinar la veracidad de toda la noticia. La Tabla 5.6 presenta el rendimiento de esta última fase del *pipeline*. Se muestran los resultados de los distintos enfoques de ML aplicados, así como dos modelos de referencia para determinar si existe una mejora al utilizar nuestra propuesta: i) un modelo de referencia aleatorio y ii) un modelo de referencia que utiliza TF-IDF de todo el documento anotado con un único valor de veracidad para el documento.

Modelo	True			False				
	P	R	F_1	P	R	F_1	Acc	Macro F_1
Referencia (aleatorio)	0,523	0,503	0,510	0,483	0,502	0,489	0,502	0,500
Referencia (TF-IDF)	0,609	0,868	0,715	0,726	0,381	0,494	0,637	0,605
DT	0,971	0,976	0,972	0,976	0,965	0,969	0,971	0,971
LR	0,964	0,997	**0,980**	0,996	0,958	**0,976**	**0,978**	**0,978**
MNB	0,920	0,995	0,956	0,995	0,902	0,945	0,951	0,950
SVM	0,934	0,994	0,962	0,993	0,919	0,953	0,958	0,958

Tabla 5.6: Resultados del rendimiento del modelo de predicción de la veracidad del artículo de la noticia utilizando la veracidad de los elementos 5W1H *gold standard*.

Como se puede concluir de los resultados de la tabla, todos los modelos propuestos superan significativamente a los dos modelos de referencia propuestos. Aún así, el modelo que obtiene mejores resultados es el de LR tanto para detectar noticias falsas como para determinar qué noticias son verdaderas, obteniendo

un 0,978 de macro F_1. Es especialmente destacable que utilizando todo el documento anotado con un único valor de veracidad (TF-IDF de referencia) la macro F_1 sea de 0,605. Estos resultados validan la hipótesis principal de esta investigación, es decir, que los elementos individuales 5W1H son un buen indicador de predicción de la veracidad global de las noticias.

Rendimiento de las Fases 3 + 4 + 5. Capa de veracidad

Para medir todo el rendimiento de la Capa de Veracidad, pero evitando los errores producidos por la Capa de Estructura, es decir, los módulos de segmentación (Fase 1 y Fase 2), se utilizan los elementos *gold standard* del corpus y se mide el rendimiento de la arquitectura de la Fase 3 a la Fase 5.

Para ello, el módulo de predicción de la veracidad de las 5W1H (Fase 4) se ejecutó 10 veces independientes en diferentes divisiones de *training* (80 %) y *test* (20 %) y se concatenaron los resultados de las etiquetas predichas en cada conjunto de *test* independiente. De este modo, se dispone de un "nuevo" conjunto de entrenamiento remuestreado para entrenar y evaluar en la Fase 5, lo que permite entrenar el módulo de esta fase directamente con las etiquetas de veracidad predichas, en lugar de con las etiquetas *gold standard*. Por lo tanto, si el módulo de predicción de la veracidad de las 5W1H (Fase 4) comete errores consistentes en diferentes etiquetas 5W1H, el módulo de la Fase 5 podría ser capaz de corregir estos errores en la predicción agregada mediante la asignación de menos pesos a esas etiquetas. Los resultados se presentan en la Tabla 5.7.

	True			False				
Modelo	P	R	F_1	P	R	F_1	Acc	Macro F_1
Referencia (aleatorio)	0,551	0,549	0,548	0,498	0,500	0,497	0,526	0,522
Referencia (TF-IDF)	0,609	0,868	0,715	0,726	0,381	0,494	0,637	0,605
DT	0,736	0,752	0,741	0,724	0,696	0,706	0,726	0,723
LR	0,842	0,783	**0,809**	0,780	0,835	**0,805**	**0.807**	**0.807**
MNB	0,794	0,827	0,808	0,804	0,760	0,778	0,795	0,793
SVM	0,802	0,768	0,781	0,761	0,786	0,770	0,777	0,775

Tabla 5.7: Resultados del rendimiento del módulo de predicción de la veracidad del artículo de la noticia entrenado y evaluado con las etiquetas predichas de la Fase 4.

Como puede observarse, aunque los resultados son peores que cuando se utilizan anotaciones *gold standard*, son mejores de lo que cabría esperar si todos los errores de la Fase 4 se llevaran a la Fase 5. Dado que la Fase 4 obtiene por el momento una precisión máxima de 0,660, el hecho de que pueda obtenerse una media de 0,805 agregando estimaciones de baja precisión para cada 5W1H sugiere algún tipo de efecto regularizador. Podemos argumentar que la Fase 5 efectivamente aprende a corregir algunos de los errores de la Fase 4. Esto no es sorprendente si tenemos en cuenta que, en la Fase 5, cada una de las etiquetas

de veracidad individuales para cada elemento 5W1H en un único artículo puede verse como la salida de un único clasificador, los cuales se agregan de forma conjunta. Por lo tanto, aunque los elementos individuales no sean muy fiables (es decir, cada elemento 5W1H acierta una media del 66% de las veces), el clasificador global es mucho más fiable. Se sabe que los modelos de conjunto pueden superar considerablemente a cada uno de sus elementos, especialmente cuando los elementos individuales cometen errores que son en su mayoría independientes unos de otros. Parece que en este caso se produce un efecto similar. La configuración de los parámetros se detalla en el Apéndice C.

Análisis entre dominios

Para explorar la aplicabilidad de nuestra propuesta en distintos ámbitos, se llevaron a cabo dos experimentos diferentes. En primer lugar, se creó un corpus pequeño en el ámbito político y se anotó de acuerdo con nuestro esquema de anotación. Contiene 17 noticias falsas y 14 verdaderas, y se utilizó únicamente para llevar a cabo la prueba. En segundo lugar, para poder probar el sistema en ámbitos distintos al político, el sistema de detección de noticias falsas se probó utilizando el corpus español de (Posadas-Durán y cols., 2019)[5], que es un corpus de sitios web de noticias que cubre diferentes ámbitos (ciencia, deporte, economía, educación, entretenimiento, política, salud, seguridad y sociedad). Dado que este corpus sólo está anotado con dos etiquetas (*real* y *fake*), no podemos utilizarlo aquí para entrenar nuestro sistema, pero sí para probarlo como experimento entre dominios. En la Tabla 5.8 se muestran los resultados obtenidos en estos dos escenarios de dominios cruzados.

Training	Test	*Pipeline* completo			
		Acc	F_1 (True)	F_1 (False)	Macro F_1
Corpus salud	Corpus salud	0,75	0,79	0,69	0,74
Corpus salud	Corpus política	0,62	0,17	0,75	0,46
Corpus salud	Corpus Posadas	0,52	0,31	0,59	0,45

Tabla 5.8: Análisis entre dominios de la propuesta.

No es sorprendente que haya una pérdida en el *accuracy* y F_1 en comparación con los resultados dentro del dominio que se muestran en la primera fila de la Tabla 5.8. Disminución de rendimientos similares se produjeron en la literatura cuando se lleva a cabo un análisis entre dominios (Pérez-Rosas y cols., 2017) (Huang y Chen, 2020) (Hanselowski y cols., 2018). En cuanto al corpus de (Posadas-Durán y cols., 2019), también hay una pérdida considerable de F_1 y de *accuracy* en comparación con los resultados obtenidos por los autores. Una de las principales causas es que el corpus de Posadas está compuesto por docu-

[5]Disponible en `https://github.com/jpposadas/FakeNewsCorpusSpanish`

mentos de nueve dominios diferentes, cuyo vocabulario es muy diverso y, por lo tanto, diferente al vocabulario de salud sobre el que se entrena nuestro sistema. Por ello, hay que tener en cuenta que (Posadas-Durán y cols., 2019) entrenó sobre su corpus, de ahí que sea de esperar que sus resultados sean superiores.

Los resultados del experimento entre dominios muestran que aún se puede mejorar el modelo para resolver el problema de la aplicabilidad entre dominios. Aunque el sistema obtiene resultados de *accuracy* razonables, la puntuación F_1 disminuye significativamente, sobre todo en la clase de Verdadero. Esto puede explicarse teniendo en cuenta el desequilibrio en términos de características en nuestro conjunto de entrenamiento, es decir, es más difícil que una noticia se clasifique como Verdadera, ya que casi cualquier evidencia de declaraciones falsas apunta a noticias Falsas. Esto se debe a que las noticias con declaraciones falsas y verdaderas se consideran Falsas, al igual que las noticias con declaraciones falsas únicamente. Por lo tanto, nuestro modelo aprende inherentemente un sesgo hacia la clasificación de noticias como Falsas, a menos que esté presente un número suficiente de elementos 5W1H Verdaderos. En el caso extremo de no tener ninguna prueba, nuestro modelo clasifica por defecto una noticia como Falsa. Obsérvese que se trata de un valor por defecto razonable, que no está codificado, sino que se aprende implícitamente del corpus anotado. Al aplicar nuestro modelo fuera del dominio, se extraen con éxito muchos menos elementos 5W1H, ya que las características léxicas de los otros dominios difieren de aquellas en las que se entrenaron los modelos CRF. Este fallo en las primeras partes del proceso explica el sesgo hacia la clase del valor Falso.

5.2.4 Comparación de la propuesta FNDeepML con el estado del arte

El objetivo de una comparación con el SOTA es realizar una comparación fiable. Debido a la novedad y particularidades de nuestro corpus, donde se detecta cada parte esencial de la noticia y se le asigna un valor de veracidad, y dado que esto no ocurre en ningún otro corpus SOTA, a nuestro parecer, no es posible una comparación directa de los resultados de los diferentes sistemas publicados en la literatura en esos corpus. Por ello, llevamos a cabo una comparación de nuestra propuesta con sistemas del SOTA, entrenándolos y probándolos en nuestro corpus.

Nuestra propuesta vs sistemas SOTA

Para realizar esta comparativa SOTA se han analizado dos investigaciones destacadas en la literatura: (Pérez-Rosas y cols., 2017) y (Rashkin y cols., 2017). Sin embargo, en ambos casos los sistemas no estaban disponibles y han sido replicados. Además, también se incluyó otra investigación destacada cuyo código estaba disponible (Potthast y cols., 2018). La configuración de los parámetros se detalla en el Apéndice C.

Los sistemas SOTA utilizados en esta comparación trabajan con corpus en

inglés en los que a las noticias se les asigna un valor de veracidad. Por lo tanto, para comparar el rendimiento de nuestra propuesta con otros sistemas SOTA de detección de noticias falsas, nuestro corpus fue traducido al inglés y los tres sistemas SOTA mencionados fueron entrenados y probados en el corpus traducido, con una configuración de entrenamiento y prueba de 80 %/20 % en 30 evaluaciones independientes. Los resultados obtenidos se muestran en la Tabla 5.9.

Sistema	Acc	F_1 (True)	F_1 (False)	Macro-F_1
Nuestro sistema	**0,75**	**0,79**	**0,69**	**0,74**
Potthast (2018)	0,66	0,63	0,69	0,66
Pérez-Rosas (2018)	0,56	0,63	0,46	0,52
Rashkin (2017)	0,53	0,46	0,55	0,51

Tabla 5.9: Comparación con sistemas SOTA: entrenamiento y prueba con nuestro corpus.

Como se presenta en la Tabla 5.9, nuestra propuesta supera a los demás sistemas. Respecto a (Potthast y cols., 2018), nuestro sistema obtiene una mejora de 13,6 % en *accuracy*, 25,4 % en F_1 en las noticias clasificadas como *True* y obtiene un resultado muy similar en el F_1 de noticias clasificadas como *False*. Respecto a (Pérez-Rosas y cols., 2017), nuestro sistema obtiene una mejora del 33,45 % de *accuracy*, 25,40 % en F_1 en noticias Verdaderas y 50,33 % en F_1 en noticias Falsas. En cuanto a (Rashkin y cols., 2017), nuestro sistema obtiene una mejora de 41,51 % en *accuracy*, 71,73 % en F_1 en noticias Verdaderas y 25,45 % en F_1 en noticias Falsas.

Esta comparación SOTA demuestra que nuestro enfoque mejora los resultados obtenidos en nuestro corpus y que es una solución robusta. Además, nuestro enfoque pretende ir un paso más allá abordando el problema a un nivel superior al de la clasificación de textos, pues estos sistemas actúan como cajas negras. De ahí que nuestro objetivo sea proporcionar al usuario los elementos concretos de la información que lleva al sistema a una conclusión final sobre la veracidad del artículo periodístico.

5.3 Evaluación del esquema RUN-AS

Este apartado se centra en la evaluación de la anotación RUN-AS, enfocada en el criterio de confiabilidad. Es necesario profundizar en la aportación de las características textuales al problema de las fake news y en concreto al problema de la confiabilidad de la información. Si bien es cierto que la verificación de datos es fundamental para la detección y clasificación de las fake news, esta tarea no es el objetivo de la presente tesis, por lo que a continuación se presenta la evaluación de la propuesta de la anotación de características para determinar la confiabilidad de un texto. De esta forma, se cambia el criterio de veracidad del

apartado anterior (Verdadero, Falso, Desconocido) por el de confiabilidad (Confiable, No confiable). Así, la guía ha sido orientada de forma que se lleve a cabo una anotación puramente lingüística y semántica (sin acudir a conocimiento externo), de forma que proporcione al lector indicios que le permitan evaluar la confiabilidad de una noticia antes de verificar su contenido en fuentes externas.

Para validar la guía RUN-AS y respaldar las hipótesis mencionadas en el apartado 1.4, especialmente la que defiende que una evaluación detallada de la confiabilidad de múltiples elementos semánticos puede influir en la confiabilidad global de una noticia, se han llevado a cabo varios experimentos. Para determinar si el esquema de anotación propuesto es factible para abordar la detección de la desinformación, se utilizaron los métodos de ML y DL más avanzados del SOTA, ampliamente aplicados en la tarea de clasificación de la desinformación. Esta propuesta de anotación de grano fino proporciona características lingüísticas y semánticas que enriquecen el proceso de entrenamiento de los modelos de clasificación. De los tres niveles de anotación (Estructura, Contenido y Elementos de Interés) se extrajeron dos tipos de características: numéricas y categóricas. En total, se extrajeron 42 características diferentes por noticia.

Del nivel de Estructura (Pirámide Invertida) se extrajeron un total de 7 características: 5 características categóricas que indican la presencia de estas partes de la estructura de la noticia (titular, subtítulo, entradilla, cuerpo y conclusión); y otras 2 características categóricas extraídas de los atributos del titular (*title_stance* y *style*). En cuanto a los niveles de Contenido (5W1H) y de Elementos de Interés, hay un total de 35 características numéricas que hacen referencia al número de etiquetas de cada uno. Respecto al nivel de contenido 5W1H, se extrajeron 6 características relacionadas con cada 5W1H (qué, quién, dónde, cuándo, por qué y cómo). Para cada etiqueta 5W1H, se contabilizó el número de atributos de tipo *Reliable/Unreliable* (12 características), así como el número de atributos de tipo *lack_of_ information* (6 características), el atributo de tipo *role* (3 características), el atributo de tipo *main_event* (1 característica). En cuanto al nivel de Elementos de Interés, se extrajeron un total de 4 rasgos numéricos (*figure, key_expression, orthotypography, quote*), así como el número de atributos de tipo *author_stance* (3 rasgos).

A continuación se presenta un ejemplo simplificado de las características numéricas y categóricas extraídas del titular y la entradilla de una noticia (solo se muestran algunas de las características para ejemplificar la generación de las mismas):

```
{
    TITLE_style: Objective,
    TITLE_title_stance: Agree,
    TITLE_WHAT_Reliable: 0,
    TITLE_WHAT_Unreliable: 1,
    TITLE_WHO_Reliable: 0,
    TITLE_WHO_Unreliable: 1,
    TITLE_WHEN_Reliable: 0,
    TITLE_WHEN_Unreliable: 1,
```

```
    # ...
    LEAD_WHAT_Reliable: 2,
    LEAD_WHAT_Unreliable: 2,
    LEAD_WHO_Reliable: 0,
    LEAD_WHO_Unreliable: 1,
    LEAD_WHEN_Reliable: 0,
    LEAD_WHEN_Unreliable: 3,
    # ...
}
```

El mismo tipo de características se generará a partir de las demás partes de la estructura del documento. Cada característica indica el número de elementos 5W1H con una etiqueta y atributo de confiabilidad específicos que aparecen en cada parte de la noticia. Por ejemplo, en el caso de **LEAD_WHAT_Reliable: 2**, el **LEAD** contiene dos elementos **WHAT** anotados con un valor **Reliable**. A partir de estas características numéricas y categóricas, el modelo se entrena para predecir la etiqueta de confiabilidad global del documento.

5.3.1 Experimentación con el corpus RUN

Para validar la propuesta RUN-AS, esta sección presenta la experimentación propuesta con el corpus RUN mediante métodos de ML y DL (modelos *Transformers* preentrenados, que son modelos de DL que permiten extraer características detalladas y semánticas a partir de fragmentos de texto o código (García Soto, 2022)) utilizados en las tareas de desinformación, con el objetivo de mejorar el rendimiento de la tarea al entrenar los modelos con el corpus anotado con RUN-AS. Con este fin, se llevaron a cabo los dos experimentos siguientes:

1. **Rendimiento de** ML: se utilizaron los siguientes algoritmos de clasificación de ML para crear sistemas de referencia y entrenar los siguientes clasificadores: SVM, *Random Forest* (RF), LR, DT, *Multilayer Perceptron* (MLP), *Adaptive Boosting* (AdaBoost) y *Gaussian Naive Bayes* (GaussianNB). De los algoritmos mencionados, se utilizaron dos configuraciones:

   - *Modelo de referencia*: codificación de textos de noticias usando vectores de tipo TF-IDF.
   - *Modelo con características RUN-AS*: concatenación de los vectores TF-IDF con las 42 características obtenidas de la anotación.

   Este experimento se llevó a cabo utilizando la biblioteca *scikit-learn*[6] y puede replicarse en [7].

2. **Rendimiento** DL (modelo *Transformer* preentrenado): para crear dos modelos clasificadores se utilizó el modelo lingüístico BETO[8] basado en la

[6] https://scikit-learn.org/stable/
[7] https://bit.ly/37KNHnM
[8] https://github.com/dccuchile/beto

arquitectura de *Transformer* (para más detalle consultar (Canete y cols., 2020)). Ambos modelos clasificadores consisten en afinar el modelo utilizando el corpus anotado y están formados por dos componentes principales: un modelo lingüístico (BETO) y una red neuronal de clasificación.

- *Modelo de referencia*: el primer modelo es un sistema de referencia que utiliza las noticias como entrada para el modelo lingüístico (BETO).
- *Modelo con características RUN-AS*: el segundo modelo utiliza la arquitectura propuesta por (Sepúlveda-Torres, Saquete Boró, y cols., 2021), que modificó la base de BETO para incluir características externas. Tanto el texto como las 42 características se utilizan como entrada. Las características se concatenan con la salida del modelo del lenguaje BETO para alimentar la entrada a la red neuronal de clasificación.

Para crear los clasificadores, se utilizó la biblioteca *Simple Transformers*[9]. Estos experimentos se pueden reproducir en el repositorio[10].

En todos los experimentos se utilizó la estrategia de validación cruzada, la cual es una técnica estadística que consiste en dividir los datos en subconjuntos, entrenar los datos en un subconjunto y utilizar el otro subconjunto para evaluar el rendimiento del modelo. La validación cruzada permite utilizar todos los datos disponibles para el entrenamiento y las pruebas (Bergmeir y Benítez, 2012).

5.3.2 Resultados y discusión de la evaluación del corpus RUN

Esta sección presenta los resultados obtenidos en cada uno de los experimentos descritos en la sección 5.3.1.

Resultados de los experimentos con ML y DL

La Tabla 5.10 presenta el rendimiento de los experimentos 1 y 2 explicados en la sección 5.3.1 y llevados a cabo en el subconjunto anotado manualmente del corpus RUN, que está compuesto por un total de 80 noticias Confiables y No confiables.

Como puede deducirse de los resultados de la tabla 5.10, todos los modelos que utilizaron las características obtenidas mediante la anotación superan significativamente los modelos de referencia propuestos. Los mejores resultados se obtuvieron con DT utilizando la anotación RUN-AS, con un 0,948 de F_1 macro ($F_1 m$), y BETO utilizando la anotación RUN-AS, obteniendo un 0,854 de $F_1 m$.

[9] https://simpletransformers.ai/
[10] https://github.com/rsepulveda911112/BETO_RUN_AS.git

Experimentos	Modelo de referencia (TF-IDF)		Modelo con características RUN-AS	
	Acc	$F_1\,m$	Acc	$F_1\,m$
SVM	0,662	0,395	0,937	0,925
RF	0,75	0,639	0,912	0,898
LR	0,65	0,392	0,912	0,875
DT	0,737	0,683	**0,95**	**0.948**
MLP	0,712	0,57	0,925	0,912
AdaBoost	**0,787**	**0,748**	**0,95**	0,945
GaussianNB	0,612	0,456	0,687	0,57
BETO	0,85	0,80	**0,887**	**0,854**

Tabla 5.10: Resultados de los experimentos utilizando métodos de ML y DL en un subconjunto del corpus RUN anotado de forma manual.

Cabe destacar que cuando se utiliza todo el documento anotado con un único valor de confiabilidad (modelos de referencia) el mejor valor de $F_1\,m$ lo obtiene AdaBoost con 0,748 de $F_1\,m$, seguido de DT y RF. Sin embargo, para el resto de enfoques, los resultados utilizando el documento con un único valor de confiabilidad son bastante bajos. Todos los enfoques mejoran significativamente al utilizar la información proporcionada por las etiquetas de anotación del esquema RUN-AS.

Posteriormente, tras completar la fase de anotación semiautomática, se realizó este mismo experimento con el corpus RUN completo, el cual presenta 170 noticias Confiables y No confiables. La Tabla 5.11 muestra los resultados obtenidos tras entrenar con los mismos modelos:

Experimentos	Modelo de referencia (TF-IDF)		Modelo con características RUN-AS	
	Acc	$F_1\,m$	Acc	$F_1\,m$
SVM	0,464	0,317	0,958	**0,958**
RF	0,723	0,722	0,958	**0,958**
LR	0,805	**0,805**	0,947	0,946
DT	0,635	0,619	0,923	**0,922**
MLP	0,788	0,786	0,929	0,928
AdaBoost	0,694	0,693	**0,941**	0,94
GaussianNB	0,729	0,727	0,741	0,74
BETO	0,81	0,80	0,858	0,855

Tabla 5.11: Resultados de los experimentos utilizando métodos de ML y DL en el corpus RUN completo anotado de forma semiautomática.

Para validar la eficacia de la metodología propuesta para la construcción de este corpus, se creó un modelo de referencia para predecir un conjunto de *test* aleatorio con 20 noticias[11]. Este conjunto de *test* se obtuvo antes del proceso de AL, utilizando el conjunto de noticias inicial (seleccionado aleatoriamente)

[11] https://github.com/livinglang/NewsReliabilityAnnotation

para garantizar que los datos del *test* y los datos del *training* inicial (Fase 1 de la metodología) presentasen aproximadamente la misma distribución. El modelo de referencia se entrenó con el corpus RUN para predecir el conjunto de *test* y utiliza como entrada el contenido de la noticia (TITLE y BODY) codificado en vectores de tipo TF-IDF.

Como puede apreciarse en la Tabla 5.11, en este caso todos los modelos que utilizaron las características del esquema RUN-AS también mejoran notablemente en comparación con los modelos de referencia. Los mejores resultados se obtuvieron con SVM y RF utilizando la anotación RUN-AS, con un 0,958 de $F_1 m$. El mejor valor de $F_1 m$ cuando se utiliza el documento anotado con un único valor de confiabilidad (modelos de referencia) lo obtiene LR con 0,805 de $F_1 m$. En resumen, al igual que con la Tabla 5.10, todos los enfoques mejoran significativamente al utilizar la información proporcionada por las etiquetas de anotación del esquema RUN-AS. Por lo tanto, estos resultados validan la hipótesis 2 presentada en el apartado 1.4, es decir, que la confiabilidad de los elementos individuales de las 5W1H son un buen indicador de la confiabilidad global de la noticia.

Se puede concluir que, a pesar de ser un esquema de anotación con cierto grado de complejidad y costoso de generar, es una solución muy adecuada para abordar la tarea para la que está diseñado.

5.3.3 Detección automática de la confiabilidad de los elementos 5W1H

La sección 5.3.1 presenta la experimentación que demuestra que nuestro esquema de anotación manual de la estructura periodística y del contenido semántico mejora la detección de la confiabilidad de las noticias. Para demostrar la viabilidad real de la propuesta, se realizó un experimento sobre el corpus anotado exclusivamente con la estructura y se entrenó un modelo para detectar la confiabilidad de los elementos 5W1H. En futuros trabajos se abordará en profundidad la detección automática de la confiabilidad de estos elementos, ya que ello implica el uso de recursos de PLN adicionales para anotar automáticamente estos elementos, lo que queda fuera del alcance de la presente investigación.

Para llevar a cabo la clasificación de confiabilidad de las 5W1H, se utilizó el modelo BETO con una configuración similar a la de los experimentos de la sección 5.3.1. En este caso, el texto de las etiquetas 5W1H se utilizó como entrada para el modelo BETO. Se llevó a cabo la misma estrategia de validación cruzada para entrenar y validar, obteniendo 0,9 de *accuracy* y 0,73 de $F_1 m$. Estos resultados en la detección automática de la confiabilidad de las etiquetas corroboran la viabilidad de la propuesta en un futuro *pipeline* completamente automático, desde el texto plano hasta la anotación de la confiabilidad de las partes y de la noticia global.

5.3.4 Experimentación con el corpus RUN-AS-SFN

Para evaluar la eficacia y eficiencia de la metodología propuesta en el Capítulo 4, así como de la calidad del corpus semiautomático generado, se proponen dos evaluaciones: (i) en primer lugar, se evaluará si las características anotadas aplicando la anotación RUN-AS permiten determinar la confiabilidad de una noticia y (ii) se evaluará si la anotación de la confiabilidad es útil para determinar la veracidad de una noticia cuando se aplica a la tarea de detección de noticias falsas. Al igual que se hizo en la sección 5.2.4 con la primera versión de la guía, los resultados de la guía RUN-AS se compararon con los del SOTA de la literatura del corpus *The Spanish Fake news dataset* (Posadas-Durán y cols., 2019), que se ha utilizado como base de esta investigación. Por último, se realiza un análisis de la relación entre la veracidad y la confiabilidad de las noticias.

En los experimentos llevados a cabo para evaluar la tarea de detección de confiabilidad se utilizaron algoritmos clásicos de ML aplicados a la tarea de clasificación de la desinformación, para así determinar si el corpus anotado semiautomáticamente es útil para abordar la tarea de detección de la confiabilidad de noticias. Los enfoques de ML que se utilizaron fueron: SVM, LR, DT, RF, y MLP. Además, se crearon los modelos de referencia para predecir las 100 noticias que componen el corpus de *test*.

Para evaluar los modelos de clasificación, se creó un conjunto de validación a partir del conjunto de entrenamiento (300 noticias) utilizando el 20 % de los ejemplos. Para predecir el conjunto de prueba, los modelos de referencia se entrenaron y validaron con el corpus anotado RUN-AS-SFN (Sección 4.7.3). El modelo de referencia utiliza como entrada el contenido de las noticias (texto del titular y del cuerpo) codificado en vectores de tipo TF-IDF. Para el modelo que utiliza las características de RUN-AS, los vectores TF-IDF se concatenan con 42 características numéricas y categóricas extraídas de la anotación RUN-AS. Al igual que en la sección 5.3, las características se extraen de los tres niveles de anotación (Estructura, Contenido y Elementos de Interés).

La Tabla 5.12 muestra los resultados del modelo de referencia en términos de F_1 y *accuracy* para predecir el conjunto de *test*. Se entrenó un modelo de referencia sin utilizar las características RUN-AS y utilizando únicamente los vectores TF-IDF del titular y del contenido de las noticias, y un modelo que utiliza características RUN-AS que concatena las características extraídas del corpus anotado RUN-AS-SFN. Los resultados presentados en la tabla 5.12 se pueden replicar en el enlace de Github [12].

Como puede observarse en los resultados presentados en la tabla 5.12, los mejores resultados se obtienen con el enfoque SVM (95 % de *accuracy* y $F1$). Esta evaluación muestra que en todos los enfoques de ML utilizados se consigue un aumento muy elevado en los resultados de detección de confiabilidad cuando el modelo utiliza las características RUN-AS, a pesar de que la anotación de estas características se está realizando únicamente teniendo en cuenta el con-

[12] https://github.com/rsepulveda911112/RUN-AS-SFN

Experimentos	Modelo de referencia SIN características		Modelo con características RUN-AS	
	Acc	$F_1\,m$	Acc	$F_1\,m$
MLP	49,00	48,58	93,00	92,44
SVM	62,00	38,27	**95,00**	**94,66**
LR	38,00	36,76	94,00	93,56
DT	41,00	40,27	87,00	85,97
RF	52,00	51,69	88,00	86,97

Tabla 5.12: Resultados de los experimentos utilizando enfoques clásicos de ML para la tarea de detección de confiabilidad.

tenido esencial de las noticias que proporcionan los resúmenes (Sección 4.7.1). Por tanto, puede concluirse que tanto la metodología que utiliza los resúmenes como el corpus semiautomático generado son muy eficaces para la tarea de detección de confiabilidad.

5.3.5 Resultados y discusión de la evaluación del corpus RUN-AS-SFN

Para determinar cómo la anotación de la confiabilidad apoya la tarea de detección de veracidad de una noticia, se aplicó el corpus anotado RUN-AS-SFN a la tarea de detección de noticias falsas. Se realizaron los mismos experimentos explicados en el apartado anterior pero en este caso para predecir la veracidad de las noticias. Los resultados se presentan en la Tabla 5.13.

Experimentos	Modelo de referencia SIN características		Modelo con RUN-AS características	
	Acc	$F_1\,m$	Acc	$F_1\,m$
MLP	56,00	44,19	74,00	73,62
SVM	52,00	34,21	75,00	74,57
LR	52,00	43,89	77,00	76,60
DT	54,00	52,45	72,00	71,98
RF	60,00	57,55	**78,00**	**77,96**

Tabla 5.13: Resultados de los experimentos utilizando enfoques clásicos de ML para la tarea de detección de noticias falsas.

Como se presenta en la Tabla 5.13, los mejores resultados de *accuracy* y $F1m$ se obtienen con RF (78 % y 77,96 % respectivamente). El uso de las características RUN-AS mejora la tarea, con un incremento de 20 puntos porcentuales en $F1$ para el enfoque de RF y de 33 puntos porcentuales en $F1$ para el enfoque de LR.

En cuanto a la comparación con el SOTA, (Posadas-Durán y cols., 2019) entrenaron un clasificador para generar un modelo que pueda diferenciar entre noticias verdaderas y falsas y experimentaron con cuatro clasificadores de ML: SVM, LR; RF; y *Boosting* (BO). Utilizaron dos representaciones de características: una de ellas es el modelo estándar de *Bag-of-Words* (BoW) y las otras dos

representaciones son la representación de n-gramas de caracteres y n-gramas de etiquetas *Part-Of-Speech* (POS).

(Posadas-Durán y cols., 2019) presentaron el *accuracy* obtenido en el conjunto de *test* cuando entrenaron los clasificadores en un conjunto de características individuales, como BoW y POS, y combinaron esos conjuntos de características. Su mejor resultado fue un 76,94 % de *accuracy* con RF y BoW+POS. Nuestra experimentación supera este resultado, ya que se obtiene un 78,00 % de *accuracy* con RF al aplicar las características anotadas de confiabilidad. Además, es importante destacar que solo se anota el contenido esencial y no todo el texto de la noticia. Estos resultados validan la hipótesis 3 de que la anotación de la confiabilidad de las noticias favorece la tarea de detección de noticias falsas y desinformación.

Dado que los resultados de (Posadas-Durán y cols., 2019) se obtuvieron a partir de la totalidad de su corpus (*Spanish Fake News Corpus*) y teniendo en cuenta que nosotros sólo utilizamos un subconjunto de dicho corpus (el 50 % de las noticias), para llevar a cabo una comparación adecuada replicamos el modelo de LR con BoW, de la misma forma que se describe en (Posadas-Durán y cols., 2019). En este caso, el conjunto de entrenamiento es del 80 % de 300 noticias anotadas (240 noticias), con 60 noticias a validar, obteniendo un *accuracy* del 68,00 %. En la experimentación utilizando características RUN-AS, el modelo de LR obtiene un *accuracy* del 77 %. Sin embargo, es importante destacar que el *accuracy* replicado es inferior al obtenido por (Posadas-Durán y cols., 2019) debido a que en nuestro caso se utilizó menos de la mitad de las noticias del conjunto de entrenamiento y no se conocía la configuración específica de hiperparámetros utilizada por (Posadas-Durán y cols., 2019).

Análisis de la relación Confiabilidad-Veracidad

Existe un consenso general en que las noticias falsas contienen tanto información confiable como no confiable y en que existen patrones lingüísticos que pueden añadir o restar confiabilidad a las noticias. Por ello, hemos aplicado nuestro esquema de anotación RUN-AS, que clasifica la confiabilidad de los elementos de una noticia, al corpus publicado por (Posadas-Durán y cols., 2019), el cual sólo proporciona una clasificación global de la veracidad de la noticia.

En este apartado se analiza la relación existente entre las dos anotaciones, así como la influencia que tiene la anotación de la confiabilidad en la clasificación de veracidad de una noticia. El análisis se llevó a cabo mediante coincidencias y divergencias en el corpus anotado RUN-AS-SFN, el cual combina nuestra anotación con la del corpus de (Posadas-Durán y cols., 2019). Consideramos que existe una coincidencia de anotación cuando una noticia es Confiable-Verdadera o, al contrario, No confiable-Falsa; mientras que consideramos que existe una divergencia de anotación cuando una noticia es Confiable-Falsa o No confiable-Verdadera.

Como se muestra en la tabla 5.14, en el conjunto de *training* anotado, 253 noticias coinciden, mientras que 47 noticias divergen. En la tabla 5.15, en el con-

junto de *test*, 80 noticias coinciden y sólo 20 divergen.

TRAINING	REAL	FAKE
Unreliable	15	125
Reliable	128	32

Tabla 5.14: Relación entre Confiabilidad-Veracidad en el conjunto de *training*.

TEST	REAL	FAKE
Unreliable	3	35
Reliable	45	17

Tabla 5.15: Relación entre Confiabilidad-Veracidad en el conjunto de *test*.

Las noticias que divergen entre la confiabilidad obtenida por la guía de anotación RUN-AS y la clasificación de veracidad del corpus original de (Posadas-Durán y cols., 2019), tanto las pertenecientes al conjunto de *training* como de *test*, se analizaron minuciosamente para justificar esa divergencia. A continuación se presentan algunos ejemplos.

En primer lugar, se estudió la divergencia **No confiable-Verdadero (NC-V)**, en cuyo caso se consideró una noticia como No confiable (*Unreliable*) cuando la anotación de (Posadas-Durán y cols., 2019) la clasifica como Verdadera (*Real*). De acuerdo con la guía RUN-AS, los siguientes ejemplos que fueron anotados originalmente como Verdaderos fueron anotados como No confiables según nuestros criterios de confiabilidad:

- Presencia de titulares poco objetivos, mal construidos, inacabados o de tipo *clickbait*, creados para atraer la atención del usuario:
  - `Ponen en duda investigación de Carlos Trejo !y cuentan la verdad!`
  - `Lo que NBC no mostró de la entrevista con Putin`
- Imprecisión o vaguedad de la información, como estructuras impersonales, falta de claridad del tema o sujeto, o momento impreciso en el tiempo:
  - `Varios años más tarde`
  - `Incluso se ha comentado que no será sancionada`
- Contenido que influye en la neutralidad del anotador, porque lingüísticamente la información anotada es objetiva y está bien presentada, pero el contenido en sí parece poco creíble o el anotador tiene un conocimiento del mundo que no le permite ser objetivo e influye en su anotación:

  - `Niño de 8 años se prepara para entrar a la universidad`

- Información exagerada, por ejemplo, con el uso de superlativos:

  - `Retrataba el caso más escalofriante ocurrido en una casa en la Ciudad de México`

- Observaciones personales mediante el uso de la primera persona o de experiencias personales:

  - `La secretaria me dijo que se trataba de una ocasión...`

- Presencia de expresiones clave que intentan influir en la opinión del lector o incitarle a difundir y creer la información:

  - `Comparte este contenido`

- Subjetividad reflejada en la información dirigida al lector, como preguntas, opiniones personales o consejos y recomendaciones sin base científica.

  - `Si estás embarazada, el mejor remedio para olvidarte de las náuseas es consumir una pequeña dosis`
  - `Si no lo crees, te decimos cuáles son estos beneficios`

  En muchos casos, esas frases poco fiables que aportan subjetividad o polarización al discurso se mezclan con información fiable y completa, como se ve a continuación:

  - `Entre el 3 de febrero de 2016 y el 30 de mayo de 2019, el Gobierno de España ha concedido 1884 subvenciones...El chollo de ser feminista en España.`

- Falta de evidencia científica o fuentes, lo que resta credibilidad a la información al no saber de dónde procede ni en qué se basa:

  - `Diversos estudios han demostrado que...`

- Errores tipográficos, uso indebido de mayúsculas, errores gramaticales o incluso el uso de puntos suspensivos para generar dudas en el lector.

  - `Algunos se pasan de frenada en eso de la reivindicación de derechos e interpretación de la ley...`
  - `AUSENCIA DE INDÍGENAS DEL CAUCA Y CAQUETÁ...`

En segundo lugar, se analizó la divergencia **Confiable-Falso (C-F)**, es decir, aquellos casos en los que se clasificó como Confiable (*Reliable*) una noticia anotada en el corpus de (Posadas-Durán y cols., 2019) como Falsa (*Fake*). Tras estudiar este tipo de divergencia, nos dimos cuenta de que las etiquetas Confiables superan con creces a las No Confiables, lo que refleja, según los criterios de confiabilidad presentados en RUN-AS, la objetividad y neutralidad de la información compartida. En estos casos, dado que lingüísticamente no se puede demostrar la presencia de contenido engañoso, sería necesario recurrir al conocimiento externo para contrastar la información y detectar estas noticias falsas.

- Los ejemplos más representativos y comunes que hacen que las noticias se clasifiquen como Confiables son fragmentos de noticias que presentan fechas, lugares o sujetos concretos, así como hechos relatados de forma imparcial. A continuación se citan algunos ejemplos:

  - `El pasado 25 de enero las autoridades chinas anunciaban la construcción de dos hospitales`
  - `A las 7 en punto de la mañana del 25 de enero más de 500 trabajadores de la construcción y más de 10 vehículos de maquinaria de construcción aparecieron...`

  Como puede verse en estos ejemplos, hay una gran cantidad de datos concretos que podrían verificarse con técnicas de *fact-checking*, lo que aporta un alto nivel de confiabilidad a nivel de contenido por el mero hecho de ser información precisa que puede verificarse fácilmente. Sin embargo, tras la verificación de dichas noticias se demostró que los datos aportados no eran ciertos. Las organizaciones de verificación desmienten ambos ejemplos[13] [14].

- En algunos casos también detectamos la presencia de citas en las que no se muestra la postura del autor, sino que se utilizan las citas para ampliar la información y las pruebas correspondientes, lo que es un signo de neutralidad y, por tanto, de confiabilidad.

  - `"Los virus se vuelven más violentos en temas de virulencia, no necesariamente más letales", dijo Quintero en BLU Radio.`
  - `Facebook en un comunicado emitido a Europa Press aseguró haber "formado equipos especializados y trabajado con expertos para frenar a actores maliciosos"`

De este estudio comparativo entre ambos criterios de clasificación de desinformación (confiabilidad y veracidad), se puede concluir que en el caso de las

[13] `http://bit.ly/3FlN9TJ`
[14] `https://bit.ly/3UkQ9E8`

noticias Verdaderas clasificadas por nuestra guía como No confiables (NC-V), a pesar de encontrar información inexacta y subjetiva, la noticia resultó ser cierta, ya que las técnicas periodísticas actuales tienden a intentar captar la atención del lector mediante expresiones llamativas con la esperanza de que lea todo el artículo. Por lo tanto, las noticias No confiables no siempre son Falsas, sino que se pueden encontrar indicadores de subjetividad o con carga emocional, pero luego hay otros criterios, como el estilo del periodista, que influyen en esta forma de redactar la noticia sin que esta sea necesariamente No confiable. En cuanto a las noticias Falsas clasificadas como Confiables (C-F), el uso de técnicas de *fact-checking* es necesario, pues las noticias falsas adoptan la apariencia de una noticia creíble, mezclando datos verdaderos y falsos con la intención de que al lector le resulte más difícil detectar los elementos falsos.

Por ello, aunque el análisis lingüístico puede proporcionar numerosas pistas sobre la confiabilidad de una noticia en un primer nivel textual, cuando a nivel lingüístico no se pueden detectar suficientes indicadores de desinformación, es necesario recurrir al conocimiento externo de fuentes fiables para contrastar este tipo de noticias. El análisis lingüístico y la verificación de datos deben ser un equipo que trabajen de la mano, cada uno en su dominio, siendo el lenguaje un apoyo previo para detectar la confiabilidad de las noticias de forma más rápida y precisa. Así pues, el presente trabajo apoya que, junto al análisis lingüístico, el cual se ha demostrado que puede ayudar en la detección de noticias falsas en términos de confiabilidad, se combine la verificación automática de datos mediante conocimiento del mundo, tal y como se incorporó en la experimentación del esquema FNDeepML.

La Figura 5.3 muestra un resumen de toda la experimentación llevada a cabo a lo largo de este capítulo. Para cada evaluación (la de veracidad y la de confiabilidad) se indican los algoritmos entrenados, el tipo de característica o anotación que se toma como entrada para el entrenamiento, el esquema de anotación utilizado y el corpus sobre el que se entrena en cada caso. En la evaluación de la veracidad (color verde) se entrena con el esquema de anotación FNDeepML y con el corpus FNDeep, mientras que en la evaluación de la confiabilidad (color rosa) se entrena con el esquema RUN-AS, pero unos experimentos se llevan a cabo con el corpus RUN y otros con el corpus RUN-AS-SFN.

5.4 Conclusiones

En este capítulo se ha presentado toda la experimentación llevada a cabo con respecto a las dos guías de anotación (FNDeepML y RUN-AS), evaluadas con varios recursos y aplicando diversos modelos de ML y DL. En primer lugar, se presenta la guía diseñada para la anotación de la veracidad (FNDeepML), para la cual se diseñó una arquitectura basada en cinco fases organizadas en dos capas: la de Estructura (encargada de estructurar la noticia y extraer el contenido esencial) y la de Veracidad (basada en enriquecer los elementos 5W1H con

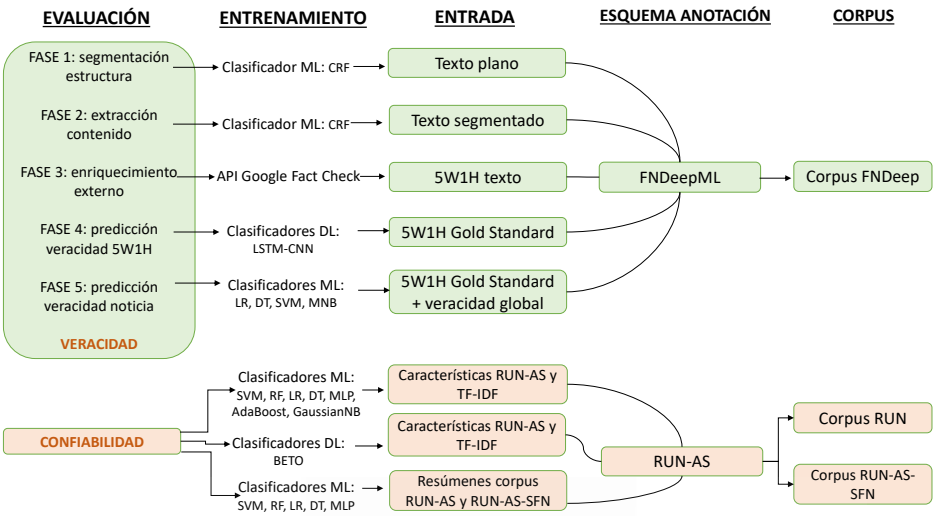


Figura 5.3: Marco de evaluación y experimentación de las guías FNDeepML y RUN-AS.

conocimiento externo y en clasificar tanto la veracidad del contenido como la de la noticia completa). La arquitectura planteada se basa en (i) segmentar la noticia según la estructura de la Pirámide Invertida, (ii) clasificar las etiquetas 5W1H mediante una clasificación jerárquica en dos niveles, (iii) enriquecer cada elemento 5W1H con conocimiento externo mediante el uso de *fact-checking*, (iv) predecir el valor de veracidad de cada elemento de las 5W1H, teniendo en cuenta tanto sus características textuales como las evidencias recogidas en la fase de *fact-checking* y (v) predecir la noticia global teniendo en cuenta el valor de veracidad de cada 5W1H.

Los experimentos con esta guía de anotación han demostrado que el uso del valor de veracidad de los distintos elementos estructurales y del contenido esencial 5W1H dentro de la noticia proporciona una solución adecuada al problema. El mejor rendimiento de la Capa de Veracidad utilizando la veracidad de los elementos 5W1H *gold standard* se obtuvo con un modelo de LR, que dio como resultado un $F_1 m$=0,978, en comparación con un modelo de referencia (TF-IDF), anotado con un valor de veracidad único para el documento, que dio como resultado un $F_1 m$=0,605. Estos resultados demuestran la validez de la guía FNDeepML para la tarea de detección de la veracidad de las noticias.

Tras evaluar el esquema FNDeepML y el recurso FNDeep, se analizó igualmente la versión de la guía enfocada a la anotación de la confiabilidad (RUN-AS) y el recurso generado de forma semiautomática (RUN). Para evaluar el esquema RUN-AS se llevaron a cabo dos experimentos con clasificadores de ML y DL Los resultados de la evaluación de esta guía muestran que los modelos mejoran significativamente al utilizar las características del esquema RUN-AS y también

que la confiabilidad individual de cada uno de los elementos anotados contribuye a predecir la confiabilidad global de una noticia, llegando a obtener un $F_1\,m$=0,958 con SVM y RF, frente a $F_1\,m$=0,317 y $F_1\,m$=0,722 respectivamente sin utilizar las características (modelos de referencia TF-IDF).

Finalmente, se llevó a cabo una segunda evaluación del esquema RUN-AS, pero en otro corpus diferente: el corpus RUN-AS-SFN. Se trata de un corpus publicado y ya anotado con un valor de veracidad global al que se le ha aplicado la guía de anotación RUN-AS y la clasificación de confiabilidad. Con este corpus se quiso evaluar si el criterio de confiabilidad era útil en la tarea de detección de la veracidad. Para ello, se comparó la clasificación de confiabilidad proporcionada por la guía RUN-AS con la clasificación de veracidad atribuida por el autor del corpus y se analizaron las divergencias y similitudes, así como los indicadores lingüísticos clave en cada caso, para extraer conclusiones. Además, a diferencia de los experimentos con el corpus RUN o FNDeep, la novedad del corpus RUN-AS-SFN reside en el uso de extracción de resúmenes durante el proceso de anotación, lo que agilizó la tarea de anotación. Así pues, de los resultados puede concluirse que tanto la guía de veracidad como la de confiabilidad son eficaces en la tarea de detección de desinformación, pues su anotación detallada contribuye a conocer su veracidad/confiabilidad global.

CAPÍTULO 6

Conclusiones y trabajo futuro

6.1 Conclusiones generales

El fenómeno de la desinformación y de las fake news se ha convertido en un problema social a nivel mundial: la desinformación preocupa y altera a la sociedad, juega con los sentimientos de las personas e influye en su ideología, por no hablar del riesgo que puede ocasionar cuando trata temas de salud. Así pues, además de por su rápida propagación, el gran peligro de la desinformación reside en su poder de calar y persuadir de tal forma que es muy difícil desmentirla. Este poder de persuasión se debe a su componente emocional intrínseco, el cual consigue embaucar a la razón mediante sentimientos como el odio, la emoción, el temor, la compasión, la euforia o simplemente la sensación de pertenencia a un grupo que comparte los mismos ideales.

La lucha contra la desinformación es un reto crucial para la sociedad actual y para los investigadores en particular. La pandemia de la desinformación debe combatirse en el mismo medio en el que se genera y viraliza: el medio digital. Por ello, de la misma forma que el uso de algoritmos favorece la propagación de noticias falsas o no confiables, estos pueden ser a su vez la solución para identificar y detectar automáticamente la desinformación. Por ello, las tecnologías de la información y, en concreto, la Inteligencia Artificial y el Procesamiento del Lenguaje Natural se han vuelto imprescindibles para combatir la propagación exponencial de la desinformación. No obstante, aunque el PLN es un gran protagonista, la intervención humana sigue siendo esencial para crear recursos equilibrados y bien construidos para entrenar a los modelos, así como para supervisar el trabajo automatizado.

La tarea de la detección de la desinformación ha sido objeto de estudio de muchas investigaciones, muchas de ellas relacionadas con la lingüística y el periodismo. Investigaciones basadas en enfoques de contenido, las cuales analizan las características textuales de las noticias, coinciden en que existe una relación

entre el lenguaje y la falsedad de las noticias. La forma en que se comunica o redacta una noticia, las palabras que se eligen, el estilo o tono empleado, la manera de estructurar la historia, las evidencias en las que se apoya, la neutralidad de la información o la carga emocional o subjetiva son algunos de los rasgos decisivos que pueden marcar la diferencia entre una noticia confiable y otra que no lo es.

Además del análisis de los rasgos lingüísticos, el enfoque periodístico es fundamental para estudiar este fenómeno. Los dos pilares sobre los que se sostiene una noticia bien construida son la neutralidad y la objetividad. Además de estos dos principios, existen dos técnicas periodísticas que se utilizan en las noticias bien construidas y que se han sido el foco de diversos estudios enfocados en tareas de detección de roles semánticos, entidades nombradas o eventos: la Pirámide Invertida (para estructurar la noticia de forma ordenada) y las 5W1H (para comunicar el contenido de la noticia de forma precisa y completa).

6.2 Principales aportaciones

- **Marco de evaluación de la desinformación en noticias**: se ha estudiado el modelado de la desinformación en noticias mediante un análisis profundo de su estructura y contenido, así como de su entorno periodístico, lingüístico y tecnológico. En este último se valoran diferentes enfoques y técnicas de PLN.

- **Diseño de dos guías de anotación para la detección de la desinformación en noticias**: partiendo de la hipótesis de que existen características lingüísticas que permiten diferenciar las noticias poco fiables de las más fiables, se han diseñado dos guías de anotación *ad hoc*. Por un lado, el esquema FNDeepML (*Fake News Deep Markup Language*) que se basa en anotar la veracidad de las noticias con base en la información externa de un desmentido y, por otro lado, el esquema RUN-AS (*Reliable and Unreliable News Annotation Scheme*), enfocado en anotar la confiabilidad de las noticias a partir de indicadores únicamente lingüísticos, sin conocimiento externo. Para ello, se han seguido unos criterios periodísticos y lingüísticos de confiabilidad, como la objetividad, la neutralidad, la carga emocional, el estilo o la evidencia. Ambos esquemas tienen como base las dos técnicas periodísticas de la Pirámide Invertida y las 5W1H.

- **Anotación lingüística y semántica multinivel de grano fino**: partiendo de la hipótesis de que las noticias falsas o no confiables suelen mezclar información contrastada con información sospechosa o falsa, el estudio de las partes y de los elementos de contenido por separado puede ayudar a predecir la confiabilidad global de una noticia. Según la literatura consultada, los corpus actuales suelen considerar la noticia como un todo y le asignan un único valor de veracidad. Aunque este valor de veracidad

puede presentar varios grados de certeza, estos corpus no determinan específicamente qué partes dentro de la noticia son verídicas y qué partes son falsas o no confiables. Por ello, se proponen dos esquemas que permiten una anotación exhaustiva de las noticias, ya que todas las partes son importantes para la clasificación y la información falsa puede encontrarse en cualquier parte o elemento de la noticia. En el caso del esquema FNDeepML, este presenta una anotación detallada de dos niveles (Estructura y Contenido) y una clasificación en Verdadero, Falso y Desconocido con base en el conocimiento externo proporcionado por desmentidos oficiales. Respecto al esquema RUN-AS, este permite una anotación de tres niveles (Estructura, Contenido y Elementos de Interés), basándose igualmente en las dos técnicas periodísticas, pero llevando a cabo además un análisis lingüístico y semántico que permite anotar las noticias sin acudir a conocimiento externo.

- **Construcción de dos corpus para la tarea de detección de desinformación**: en primer lugar, de la anotación con la guía de veracidad FNDeep se obtuvo un corpus compuesto por un total de 200 noticias en español (105 Verdaderas y 95 Falsas) pertenecientes al ámbito de la salud, que fue anotado manualmente en función de la información proporcionada por los desmentidos oficiales recopilados para la tarea. En segundo lugar, se generó el corpus RUN compuesto por 170 noticias (85 Confiables y 85 No Confiables), de las cuales 80 fueron anotadas mediante un proceso manual y las 130 restantes con ayuda de un proceso de anotación asistida. La anotación del corpus RUN se llevó a cabo sin conocimiento externo, únicamente a partir del análisis lingüístico y semántico del texto.

- **Diseño e implementación de una metodología semiautomática de anotación**: uno de los principales problemas en el PLN es la escasez de corpus debido a la costosa tarea de anotación, la cual requiere tiempo, esfuerzo y conocimiento. A esto hay que sumarle la escasez de los recursos en el idioma español generados para la tarea de detección de desinformación, siendo el inglés el idioma en el que más recursos se generan. Para paliar este problema se presenta una metodología de anotación semiautomática que combina la anotación manual con la automática, basada en diferentes estrategias de *Human-in-the-loop*, como *Active Learning*, y técnicas de *Machine Learning* y *Deep Learning*. Esta metodología permite asistir al anotador experto y facilitar la tarea mediante la reducción del tiempo y el aumento de los ejemplos anotados, con el fin de incrementar el corpus de forma eficaz para seguir entrenando la propuesta.

- **Marco de evaluación de la anotación, de los recursos y del rendimiento de diferentes modelos entrenados**: para cerrar la tesis se presenta una evaluación completa de todos los experimentos y resultados obtenidos con respecto a las dos guías de anotación evaluadas en varios recursos

y aplicando diversos modelos de ML y DL para detectar tanto la confiabilidad como la veracidad. Además, se presenta una arquitectura diseñada para la detección de las noticias falsas en dos capas que permite aplicar las características de nuestra anotación: segmentación estructural y extracción de contenido, por un lado, y predicción de los elementos individuales y de la noticia global, por otro lado.

6.3 Validación de las hipótesis

En el apartado 1.4 se presentaron cuatro hipótesis en torno a la presente investigación. A lo largo de la tesis se ha ido contestando a estas suposiciones, pero a continuación se resume la justificación de cada una de ellas:

- **¿Es posible detectar el engaño a partir del lenguaje?**: sí, el análisis profundo de las noticias y la experimentación llevada a cabo corroboran la relación entre el lenguaje y la falsedad, apreciada en indicadores o patrones lingüísticos que permiten alertar sobre la sospecha de la confiabilidad de una noticia. Entre esos patrones se puede destacar la subjetividad (marcas personales), la carga emocional (expresiones dirigidas al lector), títulos llamativos o de tipo *clickbait* o imprecisión (falta de evidencias o de datos).

- **¿La confiabilidad global de una noticia depende de la confiabilidad de cada uno de sus elementos?**: sí, en el Capítulo 5 se ha demostrado con la experimentación que el entrenamiento con las características de las guías FNDeepML y RUN-AS, las cuales permiten la anotación detallada de todos los elementos estructurales y de contenido de una noticia, mejoran notablemente la predicción de la noticia global. Esto demuestra que la confiabilidad individual de cada uno de los elementos anotados contribuye a conocer la confiabilidad global de una noticia. Además, considerando que uno de los problemas actuales de la IA es que los sistemas de aprendizaje automático actúan como cajas negras (se les delega decisiones, pero no se llega a extraer un razonamiento de dicha decisión o resultado), esta anotación detallada ayuda a su vez a conocer los elementos que influyen en la confiabilidad global de la noticia.

- **¿Es posible construir recursos de calidad a partir de una anotación lingüística basada en el criterio de confiabilidad?**: tal y como se ha demostrado en el marco de evaluación de los corpus FNDeep y RUN, estos recursos han sido validados con buenos resultados para la tarea para la que se han creado. A su vez, se ha demostrado que la anotación basada en el criterio de confiabilidad es útil y eficaz para un primer análisis textual de la noticia y, sin contrastar la información con técnicas de verificación de datos, permite hacerse una idea de su confiabilidad y en su mayoría acertar a la hora de deducir si existe información sospechosa o no.

- **¿Un sistema de anotación asistida podría detectar el lenguaje del engaño?**: la anotación semiautomática se ha aplicado de momento a la detección de los elementos 5W1H, los cuales se preanotan y son revisados y clasificados por el experto anotador. Al igual que se ha entrenado la detección automática de las 5W1H, la preanotación de la confiabilidad de estas etiquetas puede entrenarse a partir de los patrones de desinformación definidos en la presente investigación. Precisamente es la subjetividad del lenguaje y la complejidad de la anotación lo que permite clasificar la confiabilidad de las noticias y se ha demostrado en la experimentación que los indicadores analizados son eficaces en la tarea de detección de desinformación.

6.4 Trabajo futuro

- **Detección automática de la confiabilidad de las 5W1H**: como trabajo futuro, una de las tareas que requiere de más investigación es la de la detección automática de la confiabilidad de los elementos 5W1H. Hasta ahora, la preanotación de la metodología semiautomática propuesta permite anotar automáticamente las etiquetas del nivel de Contenido (5W1H), No obstante, la confiabilidad de las 5W1H, los atributos de las etiquetas y las etiquetas de los niveles de Estructura (Pirámide Invertida) y Elementos de Interés se completan de forma manual, lo que sigue requiriendo tiempo en el proceso de anotación. En un futuro, se pretende profundizar en la automatización completa de la anotación, así como en la anotación automática de la confiabilidad global y por partes del texto con el fin de reducir la intervención del anotador. Durante la presente investigación se llevó a cabo un pequeño experimento en el que se entrenó un modelo para detectar la confiabilidad de los elementos 5W1H que corroboró la viabilidad de esta propuesta, pero esta tarea quedaba fuera del alcance de la presente investigación y forma parte de los objetivos futuros.

- **Balanceo de corpus para evitar desequilibrios**: otra de las tareas propuestas es la mejora de los corpus, de forma que estén libres de sesgos y desequilibrios. Para ello, sería conveniente llevar a cabo un estudio de todas las características anotadas y las posibles necesidades de balanceo entre ellas para evitar cualquier tipo de sesgo. Esto requiere de un estudio en profundidad porque no todas las etiquetas necesitan estar balanceadas, sino ser representativas de la realidad. Si una determinada etiqueta sólo aparece un 5 % en el mundo real, entonces en nuestro corpus sólo debería aparecer ese 5 %, para así poder modelar fielmente la realidad.

- **Aplicabilidad entre dominios**: tal y como muestran los resultados del experimento entre dominios, aún se puede mejorar el modelo para resolver el problema de la aplicabilidad entre dominios. Para ello, se pretende trabajar en recopilar noticias que presenten más variedad de temas, para así

poder aprender del vocabulario y del estilo de cada uno de ellos. También sería interesante entrenar con otros idiomas. De esa forma, se podría validar que la propuesta de anotación no solo es eficaz en el idioma español o en un único tema, como el de salud, sino que se puede aplicar a cualquier idioma o ámbito.

- **Verificación de datos**: finalmente, una tarea importante a abordar para poder seguir avanzando en la presente investigación es la de combinar la anotación lingüística con la verificación de datos. Se ha concluido de los experimentos y del análisis lingüístico que el lenguaje sin contexto es muy útil como un primer filtro a la hora de detectar la desinformación. No obstante, este tiene sus límites y necesita del conocimiento externo para verificar ciertas afirmaciones que lingüísticamente pueden ser precisas u objetivas, pero que necesitan del contexto para poder contrastar los datos comunicados. El uso de un enfoque híbrido que combine tanto las características textuales del contenido como el conocimiento externo podría mejorar significativamente el rendimiento de los modelos para esta tarea.

6.5 Publicaciones

En el marco de la presente investigación, se ha contribuido al ámbito de la detección de desinformación mediante las siguientes publicaciones en revistas y congresos.

- **Bonet-Jover, A., Piad-Morffis, A., Saquete, E., Martínez-Barco, P., & García-Cumbreras, M. Á. (2021). *Exploiting discourse structure of traditional digital media to enhance automatic fake news detection.* Expert Systems with Applications, 169, 114340**: en este artículo se presenta una arquitectura para la detección automática de las noticias falsas que tiene en cuenta la estructura del discurso de las noticias en los medios digitales tradicionales, así como la influencia del contenido esencial de una noticia. La arquitectura se basa en dos capas: la de Estructura y la de Veracidad. Los resultados presentados validan la eficacia de nuestro enfoque (Bonet-Jover, Piad-Morffis, Saquete, Martínez-Barco, y García-Cumbreras, 2021).

- **Sepúlveda-Torres, R., Bonet-Jover, A., & Saquete, E. (2021). *Here Are the Rules: Ignore All Rules: Automatic Contradiction Detection in Spanish.* Applied Sciences, 11(7), 3060**: este trabajo aborda la detección automática de contradicciones en español en el ámbito de las noticias. Para esta tarea, se creó el corpus ES-Contradiction, el cual contiene información contradictoria, compatible y no relacionada. La novedad de la investigación es la anotación detallada de los distintos tipos de contradicciones del corpus (de negación, de antónimos, numérica y estructural). Los resultados obtenidos muestran que este corpus de contradicciones en español es

adecuado para generar un modelo lingüístico capaz de detectar contradicciones en el idioma español y distinguir el tipo específico de contradicción detectada (Sepulveda-Torres, Bonet-Jover, y Saquete, 2021).

- **Bonet-Jover, A. (2020). *The disinformation battle: Linguistics and Artificial Intelligence join to beat it* (pp. 31-37). http://ceur-ws.org/Vol-2802/**: este artículo presenta un modelo del lenguaje de la mentira utilizado en las noticias falsas mediante un análisis profundo a nivel léxico, estructural y de contenido que permite distinguir la información verídica de la falsa para su detección automática. Para ello se propone el uso de las Tecnologías del Lenguaje Humano (TLH) así como la sinergia entre la Lingüística y la IA, la relación entre humano-máquina, para poder avanzar en la tarea de detección de desinformación (Bonet-Jover, 2020).

- **Bonet-Jover, A. (2021). *Semi-automatic Annotation Proposal for Increasing a Fake News Dataset in Spanish* (pp. 14-22). http://ceur-ws.org/Vol-3030/**: este artículo presenta una anotación semiautomática que permite reducir el tiempo a la vez que incrementa la cantidad de ejemplos anotados. Esta propuesta, además de suponer un avance para la investigación, facilita la construcción y anotación del corpus con el que se entrena. Teniendo en cuenta que los corpus anotados son escasos en el campo de la detección de desinformación, especialmente en español, y que la anotación manual de un corpus es un proceso lento y difícil, se propone potenciar el corpus mediante la implementación de una anotación semiautomática que asista al anotador experto (Bonet-Jover, 2021).

- **Bonet-Jover, A. (2022). *Veracity vs. Reliability: Changing the Approach of Our Annotation Guideline* (pp. 9-14). https://ceur-ws.org/Vol-3270/**: este artículo presenta la evolución del esquema de anotación diseñado para la tarea de detección de desinformación. Inicialmente, el esquema consistía en anotar cada uno de los elementos de una noticia en Falso o Verdadero, para lo cual se necesita conocimiento del mundo. El cambio de enfoque consiste en anotar las noticias en Confiable o No confiable a partir de un análisis puramente lingüístico y semántico, sin recurrir al conocimiento externo. El artículo justifica el cambio de enfoque de la anotación, explica la diferencia entre veracidad y confiabilidad y muestra los cambios concretos adoptados en esta nueva propuesta de anotación. El enfoque de confiabilidad permite detectar indicadores de desinformación y captar la atención del lector con respecto a estas características, con el fin de que pueda valorar la noticia antes de tomar la decisión de compartirla o creérsela. Este cambio de enfoque justifica además que la guía de anotación presentada ha sido cuidadosamente diseñada, probada y continuamente actualizada y modificada hasta conseguir un recurso válido para la investigación (Bonet-Jover, 2022).

Además, otras cinco publicaciones presentadas que también refuerzan el presente trabajo se encuentran en proceso de revisión:

- Bonet-Jover, A., Sepúlveda-Torres, R., Saquete, E., Martínez-Barco, P., & Nieto-Pérez, M. (2022). *RUN-AS: A novel approach to annotate news reliability for disinformation detection* (Language Resources and Evaluation, LRE). Springer.
- Bonet-Jover, A., Sepúlveda-Torres, R., Saquete, E., Martínez-Barco, P., Piad-Morffis, A., & Estévez-Velarde, S. (2022). *Applying Human-in-the-Loop to construct a dataset for determining content reliability to combat disinformation* (Engineering Applications of Artificial Intelligence, EAAI). Elsevier.
- Bonet-Jover, A., Sepúlveda-Torres, R., Saquete, E., & Martínez-Barco, P. (2022). *Annotating reliability to enhance disinformation detection: annotation scheme, resource and evaluation* (Procesamiento del Lenguaje Natural, Revista Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)).
- Bonet-Jover, A., Sepúlveda-Torres, R., Saquete, E., & Martínez-Barco, P. (2022). *When machines assist humans to assist machines. Agile dataset building for automatic disinformation detection* (Computational Intelligence and Neuroscience). Hindawi LTD.
- Badenes-Olmedo, C., Saquete, E., Sepúlveda-Torres, R., & Bonet-Jover, A. (2023). *Elsevier. ELAINE: rELiability And evidence-aware News vErifier* (17th International Conference on Document Analysis and Recognition, ICDAR 2023).

APÉNDICE A

Guía de anotación FNDeepML

A.0.1 Introducción

En la creación de noticias falsas intervienen diversos factores y elementos. Las fake news son noticias disfrazadas de verdad que tienen el objetivo de engañar y de obtener cualquier beneficio, ya sea económico, político o ideológico. Además, tienden a mezclar elementos verdaderos y falsos o a manipular la información verdadera, lo que supone otro tipo de distorsión de la realidad. Debido a esta mezcla de información verdadera y falsa, la detección de noticias falsas se ha convertido en una tarea complicada. FNDeepML (*Fake News Deep Markup Language*) nace con el propósito de detectar de forma detallada e individual cada una de las partes y los elementos de una noticia y anotar la veracidad de cada uno de ellos en función de la información externa proporcionada por los desmentidos recopilados para la investigación.

A.0.2 Nivel de Estructura

La forma en la que se redacta una noticia varía en función de la formación del autor, el tema, el estilo, la fuente, etc. Sin embargo, hay dos características que comparten las noticias bien redactadas: la neutralidad y la estructura de la Pirámide Invertida, la cual defiende que cada una de las partes en las que se divide una noticia contiene información con distinto nivel de relevancia. Así, la información más relevante se sitúa al principio de la noticia y el resto de la información sigue en orden de relevancia, finalizando con la información menos importante. A continuación se describen las etiquetas de estructura por orden de relevancia.

TITLE

Esta etiqueta marca el titular de la noticia, el cual proporciona la idea principal de la historia. Normalmente en una frase resume la información básica y

esencial de la noticia. Su principal objetivo es captar la atención del lector.

SUBTITLE

Se trata de un segundo título que detalla un poco más el titular, pero de forma muy resumida. A veces completa la información dada en el titular y otras aporta datos no mencionados antes.

LEAD

Es el párrafo que desarrolla la información principal siguiendo la técnica de las 5W1H y presenta los elementos más relevantes de la noticia. Toda la información principal de la noticia debe presentarse de forma clara y concisa en la entradilla. La entradilla y el titular se consideran a veces como una unidad porque la entradilla suele repetir la idea dada por el titular, pero con más detalle y precisión.

BODY

Toda la información desarrollada se encuentra en el cuerpo de la noticia, el cual presenta de forma detallada todos los antecedentes, hechos y elementos importantes de la noticia. No debe desarrollar nuevos significados sino que más bien repite la información proporcionada en el titular o la entradilla. Las seis preguntas contestadas en la entradilla se desarrollarán en el cuerpo explicando todos los elementos que intervienen en la noticia. Cabe resaltar que en esta primera versión de la anotación la parte del cuerpo de la noticia no se anotó con las 5W1H por su extensión, que ralentizaba la anotación, por lo que se decidió dejarla sin anotar hasta que se comprobara que la anotación funcionaba correctamente con el resto de las partes.

CONCLUSION

La idea principal de la historia puede resumirse en una frase o en un párrafo, pero, aunque la conclusión forme parte de un artículo bien construido, no siempre aparece. La conclusión presenta la información menos importante, un resumen de la noticia, ya que toda la información importante ha sido desarrollada a lo largo del resto de las partes.

A continuación se muestra un ejemplo de la anotación del Nivel de Estructura:

<TITLE>Un vaso de agua caliente con limón puede salvarte la vida</TITLE>

<SUBTITLE>El limón, además de ser un componente ideal para nuestras comidas, puede salvarnos la vida, ya que previene y cura el cáncer.</SUBTITLE>

<LEAD>Son muchas las propiedades que tiene el limón, pero seguro que no sabías que desde hacía millones de años expertos médicos lo han utilizado para curar el cáncer, pues tomar un vaso de agua caliente con trozos de este cítrico

todos los días mata las células cancerígenas de nuestro cuerpo y crea un escudo protector que previene futuros tumores.</LEAD>

<BODY>Existen muchos estudios que a lo largo de los años han demostrado que el limón tiene propiedades milagrosas para nuestra salud. Se ha llegado a demostrar que hasta es 100 veces más efectivo que la quimioterapia. No obstante, hay que saber cómo prepararlo para que este cítrico tenga los efectos deseados en nuestro cuerpo. En primer lugar, se debe utilizar agua caliente con trocitos de limón y tomarlo en ayunas todos los días [...]</BODY>

<CONCLUSION>En resumen, el limón es un alimento anticáncer que puede salvar tu vida gracias a sus propiedades anticancerígenas. Tomar una infusión de agua caliente con una rodaja de limón te ayudará a prevenir y matar esta dura enfermedad, así que no dudes en difundir esta noticia a todo el mundo.</CONCLUSION>

<TITLE>*A glass of hot water with lemon can save your life*</TITLE>

<SUBTITLE>*The lemon, besides being an ideal component for our meals, can save our lives, since it prevents and cures cancer.*</SUBTITLE>

<LEAD>*The lemon has several properties, but surely you did not know that the medicine has used it for millions of years to cure the cancer, as drinking a glass of hot water with slices of this citrus fruit every day kills cancer cells of our body and creates a protective shield that prevents future tumours.*</LEAD>

<BODY>*There are many studies that have shown over the years that the lemon has miraculous properties for our organism. It has been shown that it is up to 100 times more effective than chemotherapy. However, we must know how to prepare it so that this citrus fruit has the desired effects on our body. First of all, you should use hot water with lemon slices and take it fasting every day [...]*</BODY>

<CONCLUSION>*To sum up, the lemon is an anti-cancer food that can save your life thanks to its anti-cancer properties. Its consumption with hot water will help you to prevent and kill this hard illness, so do not hesitate to spread this news to everybody.*</CONCLUSION>

QUOTE

Otra etiqueta que se incluye en el Nivel de Estructura, pero que no forma parte de la Pirámide Invertida, es la etiqueta QUOTE (cita). Esta etiqueta se diferencia de las etiquetas básicas de la estructura piramidal porque se utiliza para enmarcar un conjunto de información externa, como son las citas de terceros. Un ejemplo de esta etiqueta es:

Ha circulado una noticia falsa por WhatsApp que afirmaba que <QUOTE>un médico alemán había sido detenido por fabricar el coronavirus en un laboratorio de Berlín.</QUOTE>

A false story has circulated on WhatsApp claiming that <QUOTE>*a German doctor had been arrested for creating the coronavirus in a laboratory in Berlin.*</QUOTE>

A.0.3 Nivel de Contenido

El segundo nivel se centra en los elementos esenciales del contenido de las noticias que siguen la técnica periodística conocida como las 5W1H, las cuales permiten detectar los elementos clave necesarios para comunicar con precisión una noticia. Las preguntas periodísticas 5W1H permiten describir el acontecimiento principal de un artículo respondiendo a las 6 preguntas siguientes: quién, qué, cuándo, dónde, por qué y cómo.

WHO

Marca el sujeto del evento. Normalmente puede referirse a personas, a organizaciones o incluso a entidades personificadas, como un país (por ejemplo: Francia ha comprado vacunas).

WHAT

Hace referencia a las circunstancias, acontecimientos o hechos de la historia.

WHEN

Indica el momento en que ocurrieron los hechos. Esta etiqueta suele estar relacionada con expresiones temporales (por ejemplo: el miércoles, en 2010, el viernes pasado).

WHERE

Designa el lugar donde ocurrieron los hechos y se suele encontrar en expresiones relacionadas con la ubicación, ya sean físicas (por ejemplo: en un laboratorio) o no (por ejemplo: en Facebook).

WHY

Se refiere al motivo o la causa del acontecimiento. No siempre está presente en una noticia, ya que la causa puede estar implícita o no conocerse.

HOW

Indica el modo en que se han desarrollado los acontecimientos, la manera o el método en que se ha llevado a cabo una determinada acción.

A continuación se muestra un ejemplo de la anotación del Nivel de Contenido:

```
<WHO>Un científico italiano</WHO>
<WHAT>fue detenido</WHAT>
<HOW>mediante el uso de la fuerza</HOW>
```

<WHEN>ayer</WHEN>
<WHERE>en Milán</WHERE>
<WHY>por vender una vacuna no autorizada</WHY>
<WHO>*An Italian scientist*</WHO>
<WHAT>*was arrested*</WHAT>
<HOW>*by force*</HOW>
<WHEN>*yesterday*</WHEN>
<WHERE>*in Milan*</WHERE>
<WHY>*for selling an unauthorised vaccine*</WHY>

A.0.4 Atributos de FNDeepML

A continuación se presentan los atributos utilizados para algunas de las etiquetas de la presente guía de anotación.

type

Este atributo permite marcar el valor de veracidad, es decir, indica si una frase o un párrafo es verdadero o falso. Estos valores se indican de la siguiente manera: *True* (información verdadera), *False* (información falsa) o *Unknown* (información desconocida). De este modo, se pueden detectar partes falsas y partes verdaderas en la misma noticia. Un ejemplo del uso de este atributo es el siguiente:

<WHAT type ::= False>Un vaso de agua caliente con limón puede salvarte la vida</WHAT>

<WHAT type ::= False>*A glass of hot water with lemon can save your life*</WHAT>

author_stance

Este atributo se utiliza únicamente con la etiqueta QUOTE y sirve para anotar la postura del autor respecto al texto citado. Los valores de este atributo son: *Agree* (el autor está de acuerdo con la cita), *Disagree* (el autor no está de acuerdo con la cita) o *Unknown* (la postura del autor no está clara).

Un ejemplo del uso de este atributo es el siguiente:

Según una noticia falsa <QUOTE author_stance ::= Disagree>un médico alemán ha sido detenido por fabricar el coronavirus en un laboratorio de Berlín.</QUOTE>

According to a fake news item, <QUOTE author_stance ::= Disagree>*a German doctor had been arrested for creating the coronavirus in a laboratory in Berlin.*</QUOTE>

APÉNDICE B

Guía de anotación RUN-AS

B.0.1 Introducción

RUN-AS (*Reliable and Unreliable News Annotation Scheme*) es una propuesta de anotación de grano fino para determinar la confiabilidad de las noticias mediante el análisis lingüístico, sin depender del conocimiento externo. Se trata de una evolución de la guía FNDeepML que presenta un enfoque distinto. RUN-AS se apoya en la hipótesis de que existen características textuales y lingüísticas que permiten diferenciar las noticias confiables de las no confiables. Esta anotación se basa también en las dos técnicas periodísticas presentadas en la investigación y permite anotar la noticia en tres niveles: el Nivel de Estructura, el Nivel de Contenido y el nivel de Elementos de Interés. Cada nivel contiene varias etiquetas y atributos que se explican a continuación.

B.0.2 Nivel de Estructura

Este nivel divide una noticia en diferentes partes siguiendo la técnica periodística de la Pirámide Invertida. Según esta técnica, la información más importante se sitúa al principio de la noticia y la menos relevante al final. Las etiquetas de las estructuras se describen a continuación por orden de relevancia.

TITLE

El titular proporciona la idea principal de la noticia y resume, normalmente en una frase, la información básica y esencial. El principal objetivo del titular es captar la atención del lector.

SUBTITLE

El subtítulo explica el titular con más detalle, pero de forma breve. Completa la información del titular o aporta información adicional.

LEAD

La entradilla es el párrafo que desarrolla la información principal respondiendo a las seis preguntas clave que permiten comunicar la información de manera precisa y objetiva (las 5W1H). El objetivo de la entradilla es mantener la atención del lector y animarlo a leer la noticia completa.

BODY

El cuerpo de la noticia contiene toda la información desarrollada de la noticia y presenta todos los antecedentes, hechos y argumentos de la historia en detalle. Las preguntas clave respondidas en la entradilla se desarrollan en profundidad en el cuerpo de la noticia. En esta segunda versión de la anotación sí que se anota el cuerpo de la noticia al completo con las 5W1H.

CONCLUSION

La idea principal de la historia de la noticia se puede resumir en una oración o en un párrafo pero, aunque la conclusión se considere parte de un artículo bien estructurado, no siempre aparece. Presenta la información menos importante, ya que es solo un resumen de toda la información desarrollada en las partes anteriores de la noticia.

A continuación se muestra un ejemplo con todas las etiquetas del Nivel de Contenido (Pirámide Invertida):

<TITLE>Un vaso de agua caliente con limón puede salvarte la vida</TITLE>

<SUBTITLE>El limón, además de ser un componente ideal para nuestras comidas, puede salvarnos la vida, ya que previene y cura el cáncer.</SUBTITLE>

<LEAD>Son muchas las propiedades que tiene el limón, pero seguro que no sabías que desde hacía millones de años expertos médicos lo han utilizado para curar el cáncer, pues tomar un vaso de agua caliente con trozos de este cítrico todos los días mata las células cancerígenas de nuestro cuerpo y crea un escudo protector que previene futuros tumores.</LEAD>

<BODY>Existen muchos estudios que a lo largo de los años han demostrado que el limón tiene propiedades milagrosas para nuestra salud. Se ha llegado a demostrar que hasta es 100 veces más efectivo que la quimioterapia. No obstante, hay que saber cómo prepararlo para que este cítrico tenga los efectos deseados en nuestro cuerpo. En primer lugar, se debe utilizar agua caliente con trocitos de limón y tomarlo en ayunas todos los días [...]</BODY>

<CONCLUSION>En resumen, el limón es un alimento anticáncer que puede salvar tu vida gracias a sus propiedades anticancerígenas. Tomar una infusión

de agua caliente con una rodaja de limón te ayudará a prevenir y matar esta dura enfermedad, así que no dudes en difundir esta noticia a todo el mundo.</CONCLUSION>

<TITLE>*A glass of hot water with lemon can save your life*</TITLE>

<SUBTITLE>*The lemon, besides being an ideal component for our meals, can save our lives, since it prevents and cures cancer.*</SUBTITLE>

<LEAD>*The lemon has several properties, but surely you did not know that the medicine has used it for millions of years to cure the cancer, as drinking a glass of hot water with slices of this citrus fruit every day kills cancer cells of our body and creates a protective shield that prevents future tumours.*</LEAD>

<BODY>*There are many studies that have shown over the years that the lemon has miraculous properties for our organism. It has been shown that it is up to 100 times more effective than chemotherapy. However, we must know how to prepare it so that this citrus fruit has the desired effects on our body. First of all, you should use hot water with lemon slices and take it fasting every day [...]*</BODY>

<CONCLUSION>*To sum up, the lemon is an anti-cancer food that can save your life thanks to its anti-cancer properties. Its consumption with hot water will help you to prevent and kill this hard illness, so do not hesitate to spread this news to everybody.*</CONCLUSION>

B.0.3 Nivel de Contenido

Este nivel se centra en anotar los elementos esenciales del contenido de las noticias. El enfoque seguido en este nivel se basa en la técnica periodística de las 5W1H, la cual permite detectar los elementos clave necesarios para comunicar con precisión una historia. Las preguntas 5W1H utilizadas para este método periodístico son qué, quién, cuándo, dónde, por qué y cómo.

WHO

Anota el sujeto o entidad involucrada. Por lo general, puede referirse a personas, organizaciones o incluso entidades personificadas.

WHAT

Se refiere a las circunstancias, eventos o hechos de la noticia.

WHEN

Indica el momento en que ocurren los hechos. Se encuentra en expresiones temporales (por ejemplo: el miércoles, en 2010, el viernes pasado).

WHERE

Designa el lugar donde ocurren los hechos, la ubicación del evento, ya sea física o no.

WHY

Anota la causa del evento. No debe confundirse con el propósito (para qué).

HOW

Se refiere a la forma en la que se han desarrollado los acontecimientos.

A continuación se muestra un ejemplo de las etiquetas del Nivel de Contenido (5W1H):

<WHO>Un científico italiano</WHO>
<WHAT>fue detenido</WHAT>
<HOW>mediante el uso de la fuerza</HOW>
<WHEN>ayer</WHEN>
<WHERE>en Milán</WHERE>
<WHY>por vender una vacuna no autorizada</WHY>
<WHO>*An Italian scientist*</WHO>
<WHAT>*was arrested*</WHAT>
<HOW>*by force*</HOW>
<WHEN>*yesterday*</WHEN>
<WHERE>*in Milan*</WHERE>
<WHY>*for selling an unauthorised vaccine*</WHY>

B.0.4 Nivel de Elementos de Interés

Este nivel permite anotar información textual adicional que permite detectar indicadores interesantes que pueden marcar la diferencia entre noticias confiables y no confiables, como indicadores de subjetividad o de redacción de baja calidad.

KEY_EXPRESSION

Fraseología que insta a los lectores a compartir la información o que contiene una alta carga emocional que puede expresar miedo, desprecio, alarma, esperanza o fines económicos e ideológicos.

FIGURE

Etiqueta que permite marcar cifras en el texto, lo que puede ser un indicador de confiabilidad por el mero hecho de tratarse de una característica verificable mediante técnicas de *fact-checking*.

ORTHOTYPOGRAPHY

Esta etiqueta se utiliza para anotar errores gramaticales, ortográficos o de formato del texto. Algunos ejemplos de ortotipografía son frases enteras en mayúsculas, puntos suspensivos en medio del texto o incompletos, dobles espacios, muchos signos de exclamación, errores gramaticales, faltas de ortografía, falta de cohesión, etc.

QUOTE

Esta etiqueta permite la anotación de elementos u oraciones que citan textualmente o reproducen ideas de terceros. Esta etiqueta ya se utiliza en la guía FNDeepML, pero se ha movido ahora al nivel de Elementos de Interés, pues más que una etiqueta de estructura como las de la Pirámide Invertida, se trata de una etiqueta que aporta información textual adicional, entre ellas, marcas de objetividad o subjetividad.

A continuación se muestra un ejemplo de las etiquetas del Nivel de Elementos de Interés:

<KEY_EXPRESSION>Comparte esta información</KEY_EXPRESSION>
<KEY_EXPRESSION>*Share this information*</KEY_EXPRESSION>
<FIGURE>45</FIGURE> pacientes han dado positivo.
<FIGURE>*45*</FIGURE> *patients tested positive*.
<ORTHOTYPOGRAPHY>Camviará tu VIDA!!!</ORTHOTYPOGRAPHY>.
<ORTHOTYPOGRAPHY>*Dis will change your live!!!*</ORTHOTYPOGRAPHY>.
<QUOTE>"Es solo cuestión de tiempo"</QUOTE>, declaró el experto.
<QUOTE>*"It is only a matter of time"*</QUOTE>, *stated the expert.*

B.0.5 Atributos de RUN-AS

A continuación se presentan los atributos utilizados para algunas de las etiquetas de la presente guía de anotación.

reliability

Atributo que permite anotar una noticia como Confiable (*Reliable*) o No confiable (*Unreliable*) en función del nivel de precisión y objetividad. Este atributo se utiliza con todas las etiquetas 5W1H y sustituye al atributo *type* de la guía FNDeepML.

Un ejemplo del uso de este atributo es el siguiente:

<WHAT reliability ::= Unreliable>Un vaso de agua caliente con limón puede salvarte la vida</WHAT>

<WHAT reliability ::= Unreliable>*A glass of hot water with lemon can save your life*</WHAT>

author_stance

Este atributo se utiliza únicamente con la etiqueta QUOTE y sirve para anotar la postura del autor respecto al texto citado. Los valores de este atributo son: *Agree* (el autor está de acuerdo con la cita), *Disagree* (el autor no está de acuerdo con la cita) o *Unknown* (la postura del autor no está clara).

Un ejemplo del uso de este atributo es el siguiente:

La IARC lo califica como <QUOTE author_stance ::= Unknown>"probable carcinógeno humano".</QUOTE>

IARC classifies it as a <QUOTE author_stance ::= Unknown>*"probable human carcinogen".*</QUOTE>

style

Este atributo anota si el titular presenta información objetiva o subjetiva. Solo se utiliza con la etiqueta TITLE y tiene dos valores: *Objective* (Objetivo) y *Subjective* (Subjetivo).

Un ejemplo del uso de este atributo es el siguiente:

<TITLE style ::= Subjective>Un vaso de agua caliente con limón puede salvarte la vida</TITLE>

<TITLE style ::= Subjective>*A glass of hot water with lemon can save your life*</TITLE>

title_stance

Este atributo se utiliza únicamente con la etiqueta TITLE e indica si la información presentada en el cuerpo es coherente con la información del titular. Esta coherencia está representada por los siguientes valores: *Agree* (la información proporcionada es coherente en ambas partes), *Disagree* (la información proporcionada es incoherente en una de las partes) o *Unrelated* (la información proporcionada en el titular no tiene relación con el resto de la noticia).

Un ejemplo del uso de este atributo es el siguiente:

<TITLE title_stance ::= Agree>Un vaso de agua caliente con limón puede salvarte la vida</TITLE>

<TITLE title_stance ::= Agree>*A glass of hot water with lemon can save your life*</TITLE>

lack_of_information

Este atributo se usa con las etiquetas 5W1H y permite marcar si falta evidencia científica o datos importantes. Este atributo tiene un valor único (*Yes*), que indica cuándo falta dicha evidencia. Se utiliza con todas las etiquetas (si es necesario).

Un ejemplo del uso de este atributo es el siguiente:

<WHAT lack_of_information ::= Yes>Según algunos estudios</WHAT>.

<WHAT lack_of_information ::= Yes>*According to some studies*</WHAT>.

role

Este atributo solo se utiliza con la etiqueta WHO e indica el papel que desempeña el sujeto. Puede indicarse con uno de estos tres valores: *Subject* (el sujeto causa el evento), *Target* (el sujeto recibe los efectos del evento) o *Both* (el sujeto realiza ambas funciones).

Un ejemplo del uso de este atributo es el siguiente:

<WHO role ::= Target>Un científico italiano</WHO> fue detenido ayer en Milán.

<WHO role ::= Target>*An Italian scientist*</WHO> *was arrested yesterday in Milan*.

main_event

Este atributo solo se usa con la etiqueta WHAT y permite marcar el evento principal de la historia. Una noticia puede contener más de un *main_event*.

Un ejemplo del uso de este atributo es el siguiente:

<WHAT main_event>Un vaso de agua caliente con limón puede salvarte la vida</WHAT>.

<WHAT main_event>*A glass of hot water with lemon can save your life*</TITLE>

APÉNDICE C

Configuración de parámetros de la experimentación

La configuración de los parámetros de la experimentación aquí presentada se ha llevado a cabo dentro del marco del proyecto "LIVING-LANG: Living Digital Entities by Human Language Technologies" (RTI2018-094653-B-C21/C22).

C.0.1 Estrategias de Active Learning para la implementación de la metodología semiautomática

En la sección de la implementación de la metodología semiautomática del corpus RUN mediante estrategias de AL (apartado 4.4.1), se configuraron los siguientes parámetros.

Para calcular la informatividad de una noticia, el modelo entrenado se ejecuta en cada frase del documento completo y se almacenan las distribuciones de probabilidad de todas las etiquetas posibles en cada token. A partir de esta distribución, se calcula una medida de entropía a nivel de token, como se muestra en la ecuación donde $p^{(t)}_{label}$ es la probabilidad asociada a una etiqueta específica en el token t.

$$H(t) = -\sum_{label} p^{(t)}_{label} \log p^{(t)}_{label} \tag{C.1}$$

La entropía global de un documento D se define como la entropía media de todos sus tokens $t \in D$ (Ecuación C.2), que corresponde a una interpretación estándar del proceso de anotación como un proceso estocástico con decisiones independientes. Se trata de una simplificación, ya que las etiquetas de un token específico suelen estar correlacionadas con las etiquetas de los tokens cercanos. Sin embargo, esta simplificación hace que el problema se pueda abordar y no requiere suposiciones adicionales sobre la semántica del esquema de anotación, lo que lo hace extensible a otras anotaciones.

$$H(D) = \frac{1}{||D||} \sum_{t \in D} H(t) \tag{C.2}$$

Por último, se define un factor de similitud $\sim (D_i, \mathbf{D})$ entre cada nuevo documento D_i y el conjunto de documentos anotados **D**. Esta similitud se calcula como la similitud media del producto escalar entre el documento D_i y todos los documentos ya anotados en **D**, basándose en su representación doc2vec obtenida con la biblioteca de Python `gensim` (ver Ecuación C.3). Este factor de similitud se utiliza para disminuir la informatividad de posibles valores atípicos, por ejemplo, noticias en otros idiomas o documentos que no son noticias pero que, sin embargo, se incluyeron en el conjunto sin etiquetar durante la recopilación de datos.

$$sim(D_i, \mathbf{D}) = \frac{1}{||\mathbf{D}||} \sum_{D_j \in \mathbf{D}} \text{doc2vec}(D_i) \cdot \text{doc2vec}(D_j) \tag{C.3}$$

La puntuación informativa final de una noticia $I(D_i)$ se define como el producto de la entropía a nivel de documento y el factor de similitud, descontado por un factor β (en nuestros experimentos $\beta = 1$) que equilibra entre exploración y explotación (Ecuación C.4). Esta es la puntuación por la que se clasifican las noticias antes de presentarlas al anotador experto en la Fase 2.

$$I(D_i) = H(D_i) \times sim(D_i, \mathbf{D})^{\beta} \tag{C.4}$$

C.0.2 Ajuste del modelo de QA de la Fase 3 de la metodología semiautomática

Para ajustar el modelo de QA en la Fase 3 enfocada en la preanotación de las 5W1H (sección 4.4.2), la configuración inicial de hiperparámetros fue la siguiente: una longitud máxima de secuencia de 128, un tamaño de lote de 8, una tasa de entrenamiento de 4e-5 y un entrenamiento realizado en 3 iteraciones. Este modelo se puede replicar en el enlace de Github[1].

Por otro lado, la Figura C.1 muestra las curvas de pérdida en el entrenamiento y la evaluación donde se puede observar el comportamiento del modelo durante tres iteraciones de entrenamiento.

Según el gráfico de la Figura C.1 después de la primera iteración de entrenamiento, la pérdida en la curva del conjunto *training* disminuye de 1,41 a 0,96, y la pérdida en la curva del conjunto *development* aumenta de 6,19 a 6,49. Este comportamiento se mantiene en la tercera iteración, lo que indica que el modelo se está sobreajustando. Por lo tanto, se seleccionó la primera iteración para anotar las etiquetas 5W1H. En este punto, se inició la tercera fase de anotación con el modelo 5W1H ajustado.

[1] https://github.com/rsepulveda911112/BETO_QA_SPANISH_5W1H_fine_tuning

Figura C.1: Curva de pérdida utilizando los conjuntos de entrenamiento y desarrollo durante el entrenamiento.

C.0.3 Implementación y *fine-tuning* del modelo de QA

Respecto a la Figura 4.2 que muestra el proceso de definición de un umbral para la implementación del modelo QA (4.4.2), se utiliza un gráfico de dispersión para representar cada etiqueta 5W1H mediante un índice de etiqueta ficticio (eje x) y la puntuación asignada por el modelo de QA (eje y). Por último, se añade una línea de tendencia de regresión por mínimos cuadrados ordinarios para dividir las respuestas correctas y similares (puntos azules) y las incorrectas (puntos naranjas).

C.0.4 Parámetros de la Fase 4 de la arquitectura de detección

En la evaluación del esquema FNDeepML (sección 5.2), la configuración de los hiperparámetros para la Fase 4 de la arquitectura de detección (Predicción de la veracidad del contenido esencial 5W1H) se muestra a continuación.

Para predecir la veracidad de cada elemento, el módulo utiliza un modelo secuencial de tipo LSTM y otro convolucional con la siguiente arquitectura (Figura C.2).

1. Una capa de *embedding* entrenable con una dimensión de salida de 32, una longitud máxima de secuencia de 100 tokens (las secuencias más largas se truncan) y un número máximo de entradas de vocabulario de 1000 (construidas durante el entrenamiento a partir de los tokens de 1000 más frecuentes en el conjunto de entrenamiento).

2. Una capa *dropout* con una tasa *dropout* de 0,25.

3. Una capa convolucional 2D con 64 filtros y tamaño de núcleo de 5.

4. Una capa de agrupación máxima con un tamaño de agrupación de 4.
5. Una segunda capa *dropout* con una tasa *dropout* de 0,25.
6. Una capa LSTM con una dimensión de salida de 70.
7. Una tercera capa *dropout* con una tasa *dropout* de 0,25.
8. Una capa densa para la codificación de una sola vez de la etiqueta de los elementos 5W1H (es decir, WHAT, WHERE, WHY, etc.).
9. Una capa densa para la codificación de una sola vez de la etiqueta de la parte del artículo en la que aparece el elemento 5W1H en la noticia (es decir, LEAD, BODY, etc.).
10. Una concatenación de las tres capas anteriores.
11. Una capa densa final con 3 salidas (una para cada clase de *True*, *False*, *Unknown*) con una función de activación softmax.

Este modelo fue adaptado a partir de una arquitectura clásica para la clasificación de secuencias propuesta en la biblioteca de ML Keras[2] y modificado para ajustarse al número de características y ejemplos de entrenamiento disponibles en esta investigación.

Los parámetros exactos de cada capa (por ejemplo, el tamaño de las capas, tasa *dropout*, el número de filtros, etc.) se decidieron tras un breve ajuste manual entre una serie de parámetros previamente conocidos.

Cuando se dispone de la información de *fact-checking*, se añade una red densa *feed-forward fully connected*, con un total de 130 parámetros entrenables, cuya salida se concatena antes de la última capa densa con el modelo anterior.

El modelo global contiene 80 377 parámetros entrenables (80 507 cuando se añaden las características de *fact-checking*), y se entrena con el algoritmo de optimización Adam utilizando la entropía cruzada categórica como función de pérdida, con los hiperparámetros recomendados (Kingma y Ba, 2014). Para mejorar el rendimiento, este modelo se entrena haciendo uso de la técnica *early stopping*, basada en la pérdida medida en un 10 % separado del conjunto de entrenamiento, durante 3 iteraciones (Prechelt, 1998). El modelo se implementa en la biblioteca Python *keras*.

C.0.5 Parámetros de la Fase 5 de la arquitectura de detección

La configuración de los modelos utilizados en la Fase 5 de la evaluación del esquema FNDeepML (sección 5.2) es la siguiente:

- *Logistic Regression*, con un factor de regularización L_2 de 1,0 y un optimizador LBFGS.

[2] https://keras.io/

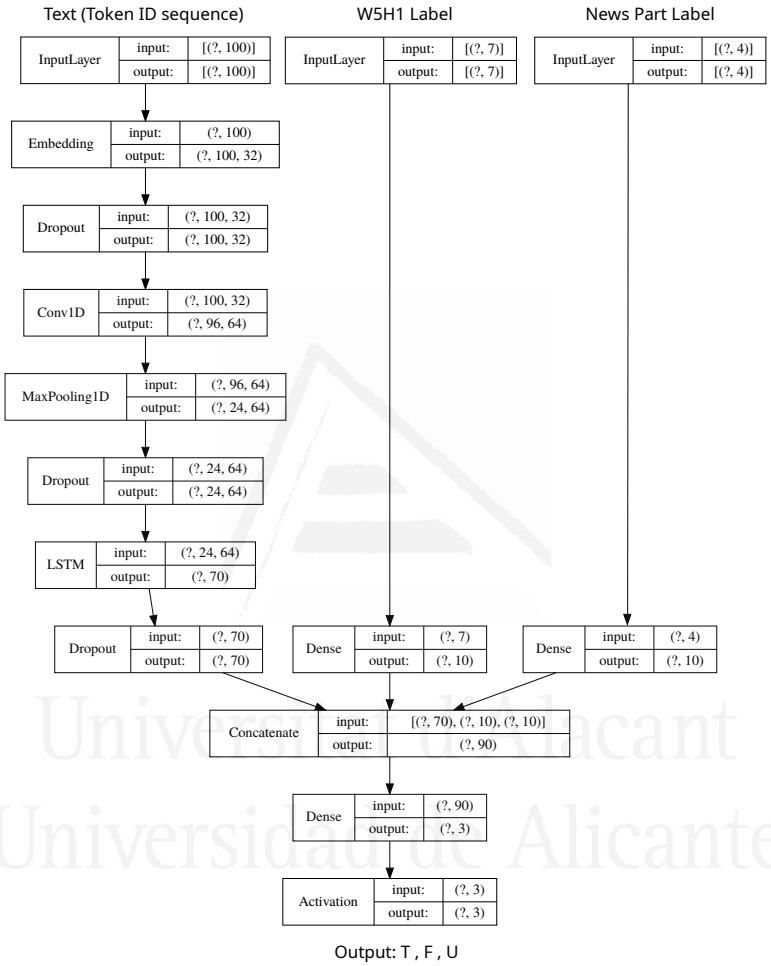


Figura C.2: Representación gráfica de la arquitectura de DL para la predicción de la veracidad de las 5W1H. Se informa del tipo de cada capa y de las formas del vector. Las formas con tamaño "?" indican la dimensión del lote, cuyo tamaño se determina en el momento del entrenamiento y no influye en el número total de parámetros.

- *Decision Trees*, utilizando GINI como criterio de selección de características.
- *Support Vector Machines*, con un núcleo de función de base radial y un factor de regularización de 1,0.
- *Multinomial Naive Bayes*, con un factor de suavizado de Laplace de 1,0.
- *Random Baseline* utilizando una estrategia aleatoria estratificada.

C.0.6 Matriz de confusión del rendimiento de la Fase 1 de la arquitectura

En la sección de resultados y discusión de la evaluación del corpus FNDeep (Sección 5.2.3), concretamente del rendimiento de la Fase 1 (Segmentación de la estructura periodística), se muestra la matriz de confusión del conjunto de prueba.

La Tabla C.1 muestra la matriz de confusión del conjunto de prueba. Como era de esperar de un modelo basado en CRF, no se produce confusión entre las clases que nunca se solapan, es decir, la entradilla y la conclusión. Dado que el cuerpo de la noticia es la clase mayoritaria (con un soporte de 23 708 tokens de un total de 28 154 en el conjunto de prueba, o 84,11 %), también es la clase con el mayor F_1. Sin embargo, a pesar de que el número de instancias de entrenamiento para el resto de clases es significativamente menor, sus puntuaciones F_1 son significativamente más altas de lo que cabría esperar de un modelo de referencia aleatorio. En comparación, si sólo se utiliza el índice relativo de tokens, se obtiene un F_1 global de 0,772, lo que indica que la mayoría de las noticias (del corpus) siguen una estructura relativamente similar en cuanto al tamaño relativo de cada segmento.

	LEAD	BODY	CONCLUSION
LEAD	2345	692	0
BODY	411	22 849	418
CONCLUSION	0	236	1203

Tabla C.1: Matriz de confusión del módulo de segmentación de la estructura periodística. Para cada uno de los 28 154 tokens de un conjunto de prueba de 20 %, las filas indican la etiqueta real y las columnas la etiqueta predicha.

C.0.7 Búsqueda de hiperparámetros para obtener el máximo rendimiento del *pipeline*

Para medir el rendimiento del *pipeline* completo evaluado en la Sección 5.2.3, se aplicó una búsqueda de hiperparámetros basada en la biblioteca de có-

digo abierto AutoGOAL (Estevez-Velarde y cols., 2020). La búsqueda de hiperparámetros permite probar un gran número de valores de parámetros para distintas partes del *pipeline* con el fin de encontrar la combinación que proporcione el mayor rendimiento. Se dedicó un total de 24 horas de recursos informáticos a la búsqueda de parámetros, lo que dio lugar a un total de 101 *pipelines* diferentes probados. El mejor *pipeline* alcanzó una precisión de 0,775 en una validación cruzada de 5 pasos con una división aleatoria de 80 % de los datos para el entrenamiento y 20 % para la prueba. El espacio de hiperparámetros contiene varios algoritmos de ML diferentes para cada fase, así como parámetros de configuración específicos como el tamaño de la ventana y la técnica de optimización para los etiquetadores CRF (Fases 2 y 3), el número de filtros y el tamaño de los vectores de *embedding* en la Fase 4, y si se cuentan los elementos 5W1H por parte del artículo (titular, cuerpo, conclusión) o agregados en la Fase 5. La mejor combinación de parámetros se resume en la Tabla C.2.

Fase	Parámetro	Valor
Fase 1	Optimizador	LBFGS
Fase 1	Tamaño ventana	3
Fase 2	Optimizador	Pasivo-Agresivo
Fase 2	Tamaño ventana	3
Fase 4	Tamaño vector *embedding*	32
Fase 4	Tamaño núcleo CNN	3
Fase 4	Filtros CNN	103
Fase 4	Tamaño de la agrupación CNN	4
Fase 4	Tamaño salida LSTM	75
Fase 4	*Dropout*	0.1
Fase 5	Algoritmo	Multinomial NB
Fase 5	Separar 5W1H en partes	Falso

Tabla C.2: Mejor combinación de parámetros encontrada para el *pipeline*

Tras la optimización, se realizó una prueba independiente en una selección aleatoria de conjuntos de prueba de 40 noticias, obteniéndose los resultados resumidos en la Tabla C.3. En general, el mejor *pipeline* encontrado obtiene una puntuación F_1 de 0,74 y una puntuación de precisión de 0,75. Obtiene una mayor precisión en la clase *True* y un mayor *recall* en la clase *False*, lo que indica un pequeño sesgo hacia la clasificación de noticias falsas.

C.0.8 Comparativa entre nuestra propuesta y los sistemas del estado del arte

Las configuraciones llevadas a cabo para comparar nuestra propuesta con los sistemas del estado del arte (Sección 5.2.4) se describen a continuación.

Respecto al enfoque de (Pérez-Rosas y cols., 2017), y teniendo en cuenta que su sistema no está disponible, lo hemos replicado considerando el mejor resultado obtenido en esta investigación. Como características se han utilizado: nú-

Modelo	True			False				
	P	*R*	F_1	*P*	*R*	F_1	*Acc*	Macro F_1
Modelo referencia (aleatorio)	0,551	0,549	0,548	0,498	0,500	0,497	0,526	0,522
Modelo referencia (TF-IDF)	0,609	0,868	0,715	0,726	0,381	0,494	0,637	0,605
***Pipeline* completo**	0,920	0,550	**0,790**	0,680	0,950	**0,690**	**0,750**	**0,740**

Tabla C.3: Rendimiento del *pipeline* completo.

mero de caracteres; palabras complejas; palabras largas; número de sílabas; tipos de palabras; número de párrafos; y métricas de legibilidad (*Flesch-Kincaid, Flesch Reading Ease, Gunning Fog y el Automatic Readability Index* (ARI)). De acuerdo con la forma en que describen la experimentación en su trabajo, hemos utilizado un clasificador lineal SVM y una validación cruzada de cinco iteraciones con nuestro corpus en inglés.

En cuanto al enfoque de (Rashkin y cols., 2017), replicamos su modelo de DL, que consiste en una capa *embedding* (utilizando *embeddings* preentrenados GLOVE 100-dim[3] que se ajustan durante el entrenamiento), una capa LSTM con 300 unidades ocultas, y una capa densa final. La única diferencia en nuestra réplica es que, dado que nuestro problema es binario, aplicamos una activación sigmoide y una pérdida de entropía cruzada binaria, en lugar de softmax y entropía cruzada categórica, como en su artículo original. Los parámetros de entrenamiento también se repiten, es decir, 10 iteraciones con un tamaño de lote de 64 elementos. Se realizaron 30 divisiones independientes de entrenamiento/prueba.

[3] `https://nlp.stanford.edu/projects/glove/`

Bibliografía

Abacha, A. B., Dinh, D., y Mrabet, Y. (2015). Semantic analysis and automatic corpus construction for entailment recognition in medical texts. En *Conference on artificial intelligence in medicine in europe* (pp. 238–242).

Afroz, S., Brennan, M., y Greenstadt, R. (2012). Detecting hoaxes, frauds, and deception in writing style online. En *2012 ieee symposium on security and privacy* (pp. 461–475).

Almela, A., Valencia-García, R., y Cantos, P. (2013). Seeing through deception: A computational approach to deceit detection in spanish written communication. *Linguistic Evidence in Security, Law and Intelligence*, *1*(1), 3–12.

Alzubi, J., Nayyar, A., y Kumar, A. (2018). Machine learning from theory to algorithms: an overview. En *Journal of physics: conference series* (Vol. 1142, p. 012012).

Appelman, A., y Sundar, S. S. (2016). Measuring message credibility: Construction and validation of an exclusive scale. *Journalism & Mass Communication Quarterly*, *93*(1), 59–79.

Assaf, R., y Saheb, M. (2021). Dataset for arabic fake news. En *2021 ieee 15th international conference on application of information and communication technologies (aict)* (pp. 1–4).

Bani-Hani, A., Adedugbe, O., Benkhelifa, E., Majdalawieh, M., y Al-Obeidat, F. (2020). A semantic model for context-based fake news detection on social media. En *2020 ieee/acs 17th international conference on computer systems and applications (aiccsa)* (pp. 1–7).

Barreiro, J. P. (2019). *Improving reading comprehension of narrative texts through summaries* (Tesis Doctoral no publicada). Universidad Casa Grande.

Bednarek, M., y Caple, H. (2012). *News discourse* (Vol. 46). A&C Black.

Benedikt, L., Joshi, C., Nolan, L., Henstra-Hill, R., Shaw, L., y Hook, S. (2020). Human-in-the-loop ai in government: a case study. En *Proceedings of the 25th international conference on intelligent user interfaces* (pp. 488–497).

Bergmeir, C., y Benítez, J. M. (2012). On the use of cross-validation for time

series predictor evaluation. *Information Sciences*, *191*, 192–213.

Bonet-Jover, A. (2020). The disinformation battle: Linguistics and artificial intelligence join to beat it.

Bonet-Jover, A. (2021). Semi-automatic annotation proposal for increasing a fake news dataset in spanish.

Bonet-Jover, A. (2022). Veracity vs. reliability: Changing the approach of our annotation guideline.

Bonet-Jover, A., Piad-Morffis, A., Saquete, E., Martínez-Barco, P., y García-Cumbreras, M. Á. (2021). Exploiting discourse structure of traditional digital media to enhance automatic fake news detection. *Expert Systems with Applications*, *169*, 114340.

Brown, S. A. (2018). *The effects of explicit main idea and summarization instruction on reading comprehension of expository text for alternative high school students* (Tesis Doctoral no publicada). Utah State University.

Budd, S., Robinson, E. C., y Kainz, B. (2021). A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, *71*, 102062.

Canete, J., Chaperon, G., Fuentes, R., y Pérez, J. (2020). Spanish pre-trained bert model and evaluation data. *PML4DC at ICLR*, *2020*.

Cañizares-Díaz, H., Piad-Morffis, A., Estevez-Velarde, S., Gutiérrez, Y., Cruz, Y. A., Montoyo, A., y Muñoz, R. (2021). Active learning for assisted corpus construction: A case study in knowledge discovery from biomedical text. En *Proceedings of the international conference on recent advances in natural language processing (ranlp 2021)* (pp. 216–225).

Chagas, L. J. (2019). The spiral model in the text of live radio journalism. *Journal of Radio & Audio Media*, *26*(2), 231–246.

Chakma, K., y Das, A. (2018). A 5w1h based annotation scheme for semantic role labeling of english tweets. *Computación y Sistemas*, *22*(3), 747–755.

Chakma, K., Swamy, S. D., Das, A., y Debbarma, S. (2020). 5w1h-based semantic segmentation of tweets for event detection using bert. En *International conference on machine learning, image processing, network security and data sciences* (pp. 57–72).

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37.

Conroy, N. K., Rubin, V. L., y Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the association for information science and technology*, *52*(1), 1–4.

Dale, R. (2017). Nlp in a post-truth world. *Natural Language Engineering*, *23*(2), 319–324. doi: doi: 10.1017/S1351324917000018

Daniel, A. M. (2021). Human-in-the-loop disinformation detection: Stance, sentiment, or something else? *ArXiv*, *abs/2111.05139*.

DeAngelo, T. I., y Yegiyan, N. S. (2019). Looking for efficiency: How online news structure and emotional tone influence processing time and memory. *Journalism & Mass Communication Quarterly*, *96*(2), 385–405.

Demartini, G., Mizzaro, S., y Spina, D. (2020). Human-in-the-loop artificial intelligence for fighting online misinformation: Challenges and opportunities. *IEEE Data Eng. Bull.*, *43*(3), 65–74.

Devlin, J., Chang, M.-W., Lee, K., y Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*.

Dhoju, S., Main Uddin Rony, M., Ashad Kabir, M., y Hassan, N. (2019). Differences in health news from reliable and unreliable media. En *Companion proceedings of the 2019 world wide web conference* (pp. 981–987).

Dijkman, R., y Wilbik, A. (2017). Linguistic summarization of event logs–a practical approach. *Information Systems*, *67*, 114–125.

El Naqa, I., y Murphy, M. J. (2015). What is machine learning? En *machine learning in radiation oncology* (pp. 3–11). Springer.

Engelen, J. A., Camp, G., van de Pol, J., y de Bruin, A. B. (2018). Teachers monitoring of students text comprehension: can students keywords and summaries improve teachers judgment accuracy? *Metacognition and learning*, *13*(3), 287–307.

Estevez-Velarde, S., Piad-Morffis, A., Gutiérrez, Y., Montoyo, A., Munoz, R., y Almeida-Cruz, Y. (2020). Solving Heterogeneous AutoML Problems with AutoGOAL. En *Proceedings of the 7th icml workshop on automated machine learning*.

Evrard, M., Uro, R., Hervé, N., y Mazoyer, B. (2020). French tweet corpus for automatic stance detection. En *Proceedings of the 12th language resources and evaluation conference* (pp. 6317–6322).

Fallis, D. (2014). The varieties of disinformation. *The philosophy of information quality*, 135–161.

Fanton, M., Bonaldi, H., Tekiroglu, S. S., y Guerini, M. (2021). Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. *arXiv preprint arXiv:2107.08720*.

Färber, M., Burkard, V., Jatowt, A., y Lim, S. (2020). A multidimensional dataset based on crowdsourcing for analyzing and detecting news bias. En *Procee-*

dings of the 29th acm international conference on information & knowledge management (pp. 3007–3014).

Feller, D. J., Zucker, J., Srikishan, B., Martinez, R., Evans, H., Yin, M. T., Gordon, P., Elhadad, N., y cols. (2018). Towards the inference of social and behavioral determinants of sexual health: development of a gold-standard corpus with semi-supervised learning. En *Amia annual symposium proceedings* (Vol. 2018, p. 422).

Ferreira, W., y Vlachos, A. (2016). Emergent: a novel data-set for stance classification. En *Proceedings of the conference of the north american chapter of the association for computational linguistics* (pp. 1163–1168). Association for Computational Linguistics. doi: doi: 10.18653/v1/N16-1138

García, M. A. (2018). *Fake news: La verdad de las noticias falsas.* Plataforma.

García Soto, E. (2022). *Diseño e implementación de una arquitectura de detección de aplicaciones android maliciosas mediante modelos transformer de código* (Tesis Doctoral no publicada). Telecomunicacion.

Grandini, M., Bagli, E., y Visani, G. (2020). Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*.

Gruppi, M., Horne, B. D., y Adali, S. (2018). An exploration of unreliable news classification in brazil and the us. *arXiv preprint arXiv:1806.02875*.

Habgood-Coote, J. (2019). Stop talking about fake news! *Inquiry*, *62*(9-10), 1033–1065.

Hamborg, F., Breitinger, C., Schubotz, M., Lachnit, S., y Gipp, B. (2018). Extraction of main event descriptors from news articles by answering the journalistic five w and one h questions. En *Proceedings of the 18th acm/ieee on joint conference on digital libraries* (pp. 339–340).

Hammad, M., y Hemayed, E. (2013). Automating credibility assessment of arabic news. En *International conference on social informatics* (pp. 139–152).

Hanselowski, A., PVS, A., Schiller, B., Caspelherr, F., Chaudhuri, D., Meyer, C. M., y Gurevych, I. (2018). A retrospective analysis of the fake news challenge stance-detection task. En *Proceedings of the 27th international conference on computational linguistics* (pp. 1859–1874). Association for Computational Linguistics. Descargado de https://www.aclweb.org/anthology/C18-1158

Hartling, L., Gates, A., Pillay, J., Nuspl, M., y Newton, A. (2018). *Development and usability testing of epc evidence review dissemination summaries for health systems decisionmakers* (Methods Research Report. Technical Report no EHC027-EF). Rockville (MD): Agency for Healthcare Research and Quality (US).

Hernández Rubio, J. E., y cols. (2022). Implementación de ids con machine learning en redes iot con dispositivo de edge computing.

Hernon, P. (1995). Disinformation and misinformation through the internet: Findings of an exploratory study. *Government information quarterly*, *12*(2), 133–139.

Hinsley, A., y Holton, A. (2021). Fake news cues: Examining the impact of content, source, and typology of news cues on peoples confidence in identifying mis-and disinformation. *International Journal of Communication*, *15*, 20.

Hordofa, B. A. (2020). Event extraction and representation model from news articles. *vol*, *16*, 1–8.

Horne, B. D., y Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. En *Eleventh international aaai conference on web and social media*.

Hou, L., Li, J., Wang, Z., Tang, J., Zhang, P., Yang, R., y Zheng, Q. (2015). Newsminer: Multifaceted news analysis for event search. *Knowledge-Based Systems*, *76*, 17–29.

Hsueh, P.-Y., Melville, P., y Sindhwani, V. (2009). Data quality from crowdsourcing: a study of annotation selection criteria. En *Proceedings of the naacl hlt 2009 workshop on active learning for natural language processing* (pp. 27–35).

Huang, Y.-F., y Chen, P.-H. (2020). Fake news detection using an ensemble learning model based on self-adaptive harmony search algorithms. *Expert Systems with Applications*, *159*, 113584. doi: doi: 10.1016/j.eswa.2020.113584

Ireton, C., y Posetti, J. (2018). *Journalism, fake news & disinformation: handbook for journalism education and training*. Unesco Publishing.

Jani, D., Patel, N., Yadav, H., Suthar, S., y Patel, S. (2022). A concise review on automatic text summarization. *Computational Intelligence in Data Mining*, 523–536.

Juez, L. A., y Mackenzie, J. L. (2019). Emotion, lies, and bullshit in journalistic discourse. *Ibérica*(38), 17–50.

Jung, W., y Jazizadeh, F. (2019). Human-in-the-loop hvac operations: A quantitative review on occupancy, comfort, and energy-efficiency dimensions. *Applied Energy*, *239*, 1471–1508.

Khalil, A., Jarrah, M., Aldwairi, M., y Jararweh, Y. (2021). Detecting arabic fake news using machine learning. En *2021 second international conference on intelligent data science technologies and applications (idsta)* (pp. 171–177).

Khan, S. U. R., Islam, M. A., Aleem, M., Iqbal, M. A., y Ahmed, U. (2018). Section-based focus time estimation of news articles. *IEEE Access*, *6*, 75452–75460.

Khodra, M. L. (2015). Event extraction on indonesian news article using multiclass categorization. En *2015 2nd international conference on advanced informatics: Concepts, theory and applications (icaicta)* (pp. 1–5).

Kholghi, M., Sitbon, L., Zuccon, G., y Nguyen, A. (2016). Active learning: a step towards automating medical concept extraction. *Journal of the American Medical Informatics Association*, *23*(2), 289–296.

Kim, J.-D., Son, J., y Baik, D.-K. (2012). Ca 5w1h onto: Ontological context-aware model based on 5w1h. *International Journal of Distributed Sensor Networks, 2012*. doi: doi: 10.1155/2012/247346

Kingma, D. P., y Ba, J. (2014). *Adam: A method for stochastic optimization*. Descargado de http://arxiv.org/abs/1412.6980 (cite arxiv:1412.6980 Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015)

Kirmani, M., Manzoor Hakak, N., Mohd, M., y Mohd, M. (2019). Hybrid text summarization: A survey. En K. Ray, T. K. Sharma, S. Rawat, R. K. Saini, y A. Bandyopadhyay (Eds.), *Soft computing: Theories and applications* (pp. 63–73). Singapore: Springer Singapore.

LeCun, Y., Bengio, Y., y Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Lee, S., Ha, T., Lee, D., y Kim, J. H. (2018). Understanding the majority opinion formation process in online environments: an exploratory approach to facebook. *Information Processing & Management*, *54*(6), 1115–1128.

Lin, X. G., Jhang, S.-E., y Dong, D. (2018). Investigating the effects of text summarization on linguistic quality of argumentative writing. *The New Korean Journal of English Language and Literature, 60*(4), 245–268.

Liu, Y., Song, X., y Chen, S.-F. (2019). Long story short: finding health advice with informative summaries on health social media. *Aslib Journal of Information Management*, *71*(6), 821–840.

Lloret, E., y Palomar, M. (2012). Text summarisation in progress: a literature review. *Artif. Intell. Rev.*, *37*(1), 1–41.

Mihalcea, R., y Strapparava, C. (2009). The lie detector: Explorations in the automatic recognition of deceptive language. En *Proceedings of the acl-ijcnlp 2009 conference short papers* (pp. 309–312).

Mihalcea, R., y Tarau, P. (2004). TextRank: Bringing order into text. En *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 404–411). Barcelona, Spain: Association for Computational

Linguistics. Descargado de https://aclanthology.org/W04-3252

Mitra, T., y Gilbert, E. (2015). Credbank: A large-scale social media corpus with associated credibility annotations. En *Proceedings of the international aaai conference on web and social media* (Vol. 9, pp. 258–267).

Monarch, R. M. (2021). *Human-in-the-loop machine learning: Active learning and annotation for human-centered ai*. Simon and Schuster.

Moreda, P., Llorens, H., Saquete, E., y Palomar, M. (2011). Combining semantic information in question answering systems. *Inf. Process. Manag.*, *47*(6), 870–885. doi: doi: 10.1016/j.ipm.2010.03.008

Mottola, S. (2020). Las fake news como fenómeno social. análisis lingüístico y poder persuasivo de bulos en italiano y español. *Discurso & Sociedad*(3), 683–706.

Narayanan, S., y Harabagiu, S. (2004). *Question answering based on semantic structures* (Inf. Téc.). INTERNATIONAL COMPUTER SCIENCE INST BERKELEY CA.

Névéol, A., Doğan, R. I., y Lu, Z. (2011). Semi-automatic semantic annotation of pubmed queries: a study on quality, efficiency, satisfaction. *Journal of biomedical informatics*, *44*(2), 310–318.

Newman, M. L., Pennebaker, J. W., Berry, D. S., y Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, *29*(5), 665 - 675.

Norambuena, B., Horning, M., y Mitra, T. (2020). Evaluating the inverted pyramid structure through automatic 5w1h extraction and summarization. En *Computational journalism symposium*.

Nycyk, M. (2015). The power gossip and rumour have in shaping online identity and reputation: A critical discourse analysis. *Qualitative Report*, *20*(2).

Okoro, E., Abara, B., Umagba, A., Ajonye, A., y Isa, Z. (2018). A hybrid approach to fake news detection on social media. *Nigerian Journal of Technology*, *37*(2), 454–462.

Orosa, B. G., Santorum, S. G., y García, X. L. (2017). El uso del clickbait en cibermedios de los 28 países de la unión europea. *Revista Latina de Comunicación Social*(72), 1261–1277.

Padró, L., y Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. En *Proceedings of the language resources and evaluation conference (lrec 2012)*. Istanbul, Turkey.

Paka, W. S., Bansal, R., Kaushik, A., Sengupta, S., y Chakraborty, T. (2021). Cross-sean: A cross-stitch semi-supervised neural attention model for covid-19

fake news detection. *Applied Soft Computing*, *107*, 107393.

Patwa, P., Sharma, S., Pykl, S., Guptha, V., Kumari, G., Akhtar, M. S., Ekbal, A., Das, A., y Chakraborty, T. (2021). Fighting an infodemic: Covid-19 fake news dataset. En *International workshop on combating online hostile posts in regional languages during emergency situation* (pp. 21–29).

Pennebaker, J. W., Boyd, R. L., Jordan, K., y Blackburn, K. (2015). *The development and psychometric properties of liwc2015* (Inf. Téc.).

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., y Mihalcea, R. (2017). Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*.

Pérez-Rosas, V., y Mihalcea, R. (2015). Experiments in open domain deception detection. En *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1120–1125).

Petkovic, J., Welch, V., Jacob, M., Yoganathan, M., Ayala, A. P., Cunningham, H., y Tugwell, P. (2016). The effectiveness of evidence summaries on health policymakers and health system managers use of evidence from systematic reviews: A systematic review. *Implementation Science*, *11*.

Posadas-Durán, J.-P., Gómez-Adorno, H., Sidorov, G., y Escobar, J. J. M. (2019). Detection of fake news in a new corpus for the spanish language. *Journal of Intelligent & Fuzzy Systems*, *36*(5), 4869–4876.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., y Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. En *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 231–240). Melbourne, Australia: Association for Computational Linguistics. Descargado de https://www.aclweb.org/anthology/P18-1022 doi: doi: 10.18653/v1/P18-1022

Prechelt, L. (1998). Early stopping-but when? En *Neural networks: Tricks of the trade* (pp. 55–69). Springer.

Rajpurkar, P., Zhang, J., Lopyrev, K., y Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Ramamoorthy, C. V., y Li, H. F. (1977). Pipeline architecture. *ACM Computing Surveys (CSUR)*, *9*(1), 61–102.

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., y Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. En *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2931–2937).

Rikkers, L. F. (2002). *The bandwagon effect.* Springer.

Rouhiainen, L. (2018). Inteligencia artificial. *Madrid: Alienta Editorial*.

Rubin, V. L. (2019). Disinformation and misinformation triangle: A conceptual model for fake news epidemic, causal factors and interventions. *Journal of documentation*.

Ruchansky, N., Seo, S., y Liu, Y. (2017). Csi: A hybrid deep model for fake news detection. En *Proceedings of the 2017 acm on conference on information and knowledge management* (pp. 797–806).

Sadiku, M., Eze, T., y Musa, S. (2018). Fake news and misinformation. *International Journal of Advances in Scientific Research and Engineering*, *4*(5), 187–190.

Salaverría, R., Buslón, N., López-Pan, F., León, B., López-Goñi, I., y Erviti, M.-C. (2020). Desinformación en tiempos de pandemia: tipología de los bulos sobre la covid-19. *El profesional de la información (EPI)*, *29*(3).

Salem, F. K. A., Al Feel, R., Elbassuoni, S., Jaber, M., y Farah, M. (2019). Fa-kes: A fake news dataset around the syrian war. En *Proceedings of the international aaai conference on web and social media* (Vol. 13, pp. 573–582).

Saquete, E., Tomás, D., Moreda, P., Martínez-Barco, P., y Palomar, M. (2020). Fighting post-truth using natural language processing: A review and open challenges. *Expert systems with applications*, *141*, 112943.

Schmitt-Beck, R. (2015). Bandwagon effect. *The international encyclopedia of political communication*, 1–5.

Seddari, N., Derhab, A., Belaoued, M., Halboob, W., Al-Muhtadi, J., y Bouras, A. (2022). A hybrid linguistic and knowledge-based analysis approach for fake news detection on social media. *IEEE Access*, *10*, 62097–62109.

Sepulveda-Torres, R., Bonet-Jover, A., y Saquete, E. (2021). Here are the rules: Ignore all rules: Automatic contradiction detection in spanish. *Applied Sciences*, *11*(7), 3060.

Sepúlveda-Torres, R., Saquete Boró, E., y cols. (2021). Gplsi team at checkthat! 2021: Fine-tuning beto and roberta. *CEUR*.

Settles, B. (2009). Active learning literature survey.

Shahi, G. K., y Nandini, D. (2020). Fakecovid–a multilingual cross-domain fact check news dataset for covid-19. *arXiv preprint arXiv:2006.11343*.

Shahi, G. K., Struß, J. M., y Mandl, T. (2021). Overview of the clef-2021 checkthat! lab task 3 on fake news detection. *Working Notes of CLEF*.

Shao, C., Ciampaglia, G. L., Varol, O., Flammini, A., y Menczer, F. (2017). The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592*, *96*, 104.

Shu, K., Bhattacharjee, A., Alatawi, F., Nazer, T. H., Ding, K., Karami, M., y Liu, H.

(2020). Combating disinformation in a social media age. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(6), e1385.

Shu, K., Sliva, A., Wang, S., Tang, J., y Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, *19*(1), 22–36.

Shu, K., Wang, S., Lee, D., y Liu, H. (2020). Mining disinformation and fake news: Concepts, methods, and recent advancements. En *Disinformation, misinformation, and fake news in social media* (pp. 1–19). Springer.

Shu, K., Wang, S., y Liu, H. (2019). Beyond news contents: The role of social context for fake news detection. En *Proceedings of the twelfth acm international conference on web search and data mining* (pp. 312–320).

Silva, R. M., Santos, R. L., Almeida, T. A., y Pardo, T. A. (2020). Towards automatically filtering fake news in portuguese. *Expert Systems with Applications*, *146*, 113199.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., y Tsujii, J. (2012). Brat: a web-based tool for nlp-assisted text annotation. En *Proceedings of the demonstrations at the 13th conference of the european chapter of the association for computational linguistics* (pp. 102–107).

Stone, P. J., Dunphy, D. C., y Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.

Sutton, C., McCallum, A., y cols. (2012). An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, *4*(4), 267–373.

Tandoc Jr, E. C., Lim, Z. W., y Ling, R. (2018). Defining fake news a typology of scholarly definitions. *Digital journalism*, *6*(2), 137–153.

Tarrés, J. (2000). El rumor como sustituto de la noticia. *monografías. com.[Publicación en línea]. Disponible en Internet< https://www.monografias.com/trabajos11/rumonot/rumonot2>[fecha de acceso: 9 de septiembre de 2010]*.

Thomson, E. A., White, P. R., y Kitley, P. (2008). objectivity and hard news reporting across cultures: Comparing the news report in english, french, japanese and indonesian journalism. *Journalism studies*, *9*(2), 212–228.

Tomanek, K., Wermter, J., y Hahn, U. (2007). An approach to text corpus construction which cuts annotation costs and maintains reusability of annotated data. En *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (emnlp-conll)* (pp. 486–495).

Vásquez, A. C., Quispe, J. P., Huayna, A. M., y cols. (2009). Procesamiento de lenguaje natural. *Revista de investigación de Sistemas e Informática*, *6*(2),

45–54.

Visvikis, D., Cheze Le Rest, C., Jaouen, V., y Hatt, M. (2019). Artificial intelligence, machine (deep) learning and radio (geno) mics: definitions and nuclear medicine imaging applications. *European journal of nuclear medicine and molecular imaging*, *46*(13), 2630–2637.

Viviani, M., y Pasi, G. (2017). Credibility in social media: opinions, news, and health informationa survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, *7*(5), e1209.

Vlachos, A., y Riedel, S. (2014). Fact checking: Task definition and dataset construction. En *Proceedings of the acl 2014 workshop on language technologies and computational social science* (pp. 18–22).

Volkova, S., Shaffer, K., Jang, J. Y., y Hodas, N. (2017). Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. En *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 647–653). Vancouver, Canada: Association for Computational Linguistics. Descargado de https://www.aclweb.org/anthology/P17-2102 doi: doi: 10.18653/v1/P17-2102

Vosoughi, S., Roy, D., y Aral, S. (2018). The spread of true and false news online. *science*, *359*(6380), 1146–1151.

Wang, W., Zhao, D., Zou, L., Wang, D., y Zheng, W. (2010). Extracting 5w1h event semantic elements from chinese online news. En *International conference on web-age information management* (pp. 644–655).

Wang, W. Y. (2017). Liar, liar pants on fire: A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Wardle, C., y Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking*. Council of Europe Strasbourg.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., y He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*.

Zabala, M. V. C. (2021). Desinformación, propaganda y guerra. fundamentos y orígenes. *Comunicación política en el mundo digital: tendencias actuales en propaganda, ideología y sociedad.*, *15*.

Zhang, A. X., Ranganathan, A., Metz, S. E., Appling, S., Sehat, C. M., Gilmore, N., Adams, N. B., Vincent, E., Lee, J., Robbins, M., y cols. (2018). A structured response to misinformation: Defining and annotating credibility indicators in news articles. En *Companion proceedings of the the web conference 2018* (pp. 603–612).

Zhang, H., Chen, X., y Ma, S. (2019). Dynamic news recommendation with hierarchical attention network. En *2019 ieee international conference on data mining (icdm)* (pp. 1456–1461).

Zhang, H., y Liu, H. (2016). Visualizing structural inverted pyramids in english news discourse across levels. *Text & Talk*, *36*(1), 89–110.

Zhou, L., y Zhang, D. (2008). Following linguistic footprints: Automatic deception detection in online communication. *Communications of the ACM*, *51*(9), 119–122. doi: doi: 10.1145/1378727.1389972

Zhou, X., y Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, *53*(5), 1–40.