



Universidad Internacional de la Rioja (UNIR)

ESIT

Máster Universitario en Inteligencia Artificial

**Detección de desinformación
relacionada con la pandemia
de COVID-19 por medio
de técnicas de aprendizaje
automático, aprendizaje
profundo y procesamiento
de lenguaje natural.**

Trabajo Fin de Máster

Presentado por: Jorge Orlando Cifuentes Cifuentes

Dirigido por: Alberto Fernández Isabel

Ciudad: Bogotá

Fecha: 3 de febrero de 2021

Índice de Contenidos

Resumen	xii
Abstract	xiii
1. Introducción	1
1.1. Motivación	2
1.2. Planteamiento del trabajo	4
1.3. Estructura de la memoria	5
2. Contexto y Estado del Arte	7
2.1. Desinformación y noticias falsas	7
2.1.1. ¿Qué son las noticias falsas?	7
2.1.2. Historia y evolución de la desinformación en forma de noticias falsas	8
2.1.3. Tipos de desinformación en noticias falsas	13
2.1.4. Caracterización de una noticia falsa	15
2.1.5. El ciclo de vida de una noticia falsa	16
2.2. Técnicas de inteligencia artificial aplicadas a la detección de noticias falsas .	17
2.2.1. Detección de noticias falsas	17
2.2.2. Uso de técnicas de inteligencia artificial para la detección de noticias falsas	18
2.2.3. Pasos básicos de técnicas de inteligencia artificial aplicados a la clasificación de noticias	22
2.3. Fuentes de noticias falsas	22
2.4. Oportunidades de investigación identificadas	23
3. Objetivos y metodología de trabajo	25
3.1. Objetivo general	25

3.2. Objetivos específicos	25
3.3. Metodología de trabajo	26
4. Identificación de requisitos y planteamiento de la metodología	29
4.1. Características principales de la metodología	29
4.2. Propuesta conceptual de la metodología	30
4.2.1. Componentes de la metodología	31
4.2.2. Técnicas de inteligencia artificial a utilizar	33
4.3. Definición de datos necesarios	33
4.3.1. Buscadores de datos	33
4.3.2. Conjuntos de datos para el modelo de clasificación de noticias	34
4.3.3. Conjuntos de datos para exploración de subtemas y modelo de ge- neración de la alerta	35
4.3.4. Conjuntos de datos relacionados con la pandemia de COVID-19 . . .	36
5. Descripción y desarrollo de la metodología	37
5.1. Diseño detallado de la metodología	37
5.2. Recolección y preparación de los datos	39
5.2.1. Preparación de datos modelo de predicción del tema de la noticia .	40
5.2.2. Preparación de datos modelo de predicción de la alerta	40
5.3. Desarrollo de modelo de predicción de temática de la noticia	41
5.3.1. Exploración de los datos	41
5.3.2. Exploración de diferentes de técnicas de inteligencia artificial . . .	42
5.4. Desarrollo del modelo de predicción de la alerta	47
5.4.1. Exploración de los datos	47
5.4.2. Exploración de diferentes de técnicas de inteligencia artificial . . .	49
5.5. Desarrollo del modelo de extracción del subtema	53
5.6. Generación de la metodología unificada	57
5.6.1. Componentes principales	59
5.6.2. Planteamiento unificado de la metodología	59
5.6.3. Definición y desarrollo del modelo	59
6. Validación de la metodología	67
6.1. Comparación de resultados de la metodología con modelo base	67

6.2. Prueba del modelo y predicciones	69
6.3. Implementación del modelo	71
6.3.1. Interfaz	72
6.3.2. Despliegue	73
6.3.3. Repositorio de código	73
6.4. Pruebas de validación final	75
6.4.1. Pruebas con noticias del año 2020	75
6.4.2. Pruebas con noticias del año 2021	76
6.4.3. Pruebas con mitos sobre el COVID-19	77
6.4.4. Pruebas con buenas prácticas para mitigar el COVID-19	79
6.4.5. Pruebas con hechos creados artificialmente	80
6.4.6. Pruebas con negaciones de frases	81
6.4.7. Análisis de resultados	82
7. Conclusiones y trabajo futuro	85
7.1. Conclusiones	85
7.2. Líneas de trabajo futuro	86
A. Anexo 1. Artículo	95

Índice de Ilustraciones

1.1. Ejemplo de noticia falsa con tintes políticos en Ucrania. Fuente: stopfake.org	2
1.2. Tendencia de interés en el término ”Fake News” en las búsquedas de Google desde el año 2016 en todo el mundo. Fuente: trends.google.com	3
2.1. Denario de Marco Antonio y Cleopatra. Fuente: [Conversation(2017)]	8
2.2. Ilustración del artículo del 26 de agosto de 1835. Fuente: [Wikipedia(2021)]	9
2.3. Caricatura con tintes políticos. Fuente: [BoerWarArchive(2014)]	10
2.4. Orson Welles ensayando su representación radiofónica de ”La guerra de los mundos”de H.G. Wells. Fuente: [NewYorkTimes(2017)].	11
2.5. El desorden de la información. Fuente: [Ireton & Posetti(2018)]	13
2.6. Caracterización de una noticia falsa. Fuente: [Zhang & Ghorbani(2020)] .	15
2.7. Elementos de una noticia falsa. Fuente: [Zhang & Ghorbani(2020)]	16
2.8. Enfoques de detección de noticias falsas a lo largo de su ciclo de vida. Fuente: [Zhang & Ghorbani(2020)]	18
2.9. Proceso de detección de noticias falsas. Fuente: [Zhang & Ghorbani(2020)] .	19
2.10. Proceso de análisis de noticias. Fuente: El autor	23
3.1. Metodología definida para el abordaje de la investigación. Fuente: el autor .	27
4.1. Entradas y salidas de la metodología propuesta. Fuente: el autor	30
4.2. El ciclo de vida de la noticia. Fuente: el autor	31
4.3. Detalle del proceso de verificación. Fuente: el autor	32
4.4. Componentes de la metodología propuesta. Fuente: el autor	33
4.5. Noticias de reuters clasificadas en este caso en la categoría ”Health”. Fuente: reuters.com/news/health	35
5.1. Diseño detallado. Fuente: el autor	38

5.2.	Detalle de modelos a probar. Fuente: el autor	39
5.3.	Resultado del scraping de Reuters.com. Fuente: el autor	40
5.4.	Preparación de los datos con OpenRefine. Fuente: el autor	41
5.5.	Noticias y categorías. Fuente: el autor	42
5.6.	Clasificadores técnicas de aprendizaje automático	44
5.7.	Detalle modelos aprendizaje profundo modelo tema	45
5.8.	Comparación de modelos de aprendizaje automático. Fuente: el autor	46
5.9.	Comparación de modelos aprendizaje profundo	47
5.10.	Distribución de las variables	48
5.11.	Clasificadores técnicas de aprendizaje automático	51
5.12.	Detalle modelos aprendizaje profundo modelo alerta	52
5.13.	Comparación de modelos de aprendizaje automático. Fuente: el autor	53
5.14.	Comparación de modelos aprendizaje profundo	54
5.15.	Resultado del modelo LDA para 7 subtemas y 5 palabras. Fuente: el autor .	56
5.16.	Modelo con múltiples entradas y una salida. Fuente: el autor	58
5.17.	Planteamiento final de la metodología unificada. Fuente: el autor	60
5.18.	Planteamiento final de la metodología unificada. Fuente: el autor	61
5.19.	Resumen de la red LSTM con múltiples entradas planteada. Fuente: el autor	65
5.20.	Comparación de modelos aprendizaje profundo	66
6.1.	Comparación de modelo unificado LSTM de múltiples entradas con modelo base de única entrada	68
6.2.	Comparación de modelo unificado LSTM de múltiples entradas con modelo similar bidireccional	69
6.3.	Detalle de la interfaz desarrollada. Fuente: el autor	71
6.4.	Arquitectura del modelo implementado. Fuente: el autor	72
6.5.	Repositorio Interfaz. Fuente: el autor	74

Índice de Tablas

2.1. Tipología de las noticias falsas [Tandoc et al.(2018)Tandoc, Lim & Ling] . . .	14
2.2. Fact-checkers más comunes. Fuente: El autor	24
6.1. Pruebas con noticias de 2020, incluidas en el conjunto de datos (L: Label, P: Predicción, D: Diferencia, 0: FALSE, 1: TRUE). Fuente: El autor	75
6.2. Pruebas con noticias de 2021 (L: Label, P: Predicción, D: Diferencia, 0: FALSE, 1: TRUE). Fuente: El autor	77
6.3. Pruebas con mitos del Coronavirus encontrados en diversas fuentes (L: La- bel, P: Predicción, D: Diferencia, 0: FALSE, 1: TRUE). Fuente: El auto . . .	78
6.4. Pruebas con buenas prácticas recopiladas por la Organización Mundial de la Salud (L: Label, P: Predicción, D: Diferencia, 0: FALSE, 1: TRUE). Fuente: El autor	79
6.5. Pruebas con hechos artificialmente creados por el autor. (L: Label, P: Pre- dicción, D: Diferencia, 0: FALSE, 1: TRUE). Fuente: El auto	80
6.6. Pruebas con negaciones de noticias ya comprobadas. (L: Label, P: Predic- ción, D: Diferencia, 0: FALSE, 1: TRUE). Fuente: El auto	81
6.7. Matrices de confusión de pruebas finales a, b, c and d.	82
6.8. Matriz de confusión consolidada de todas las pruebas. Fuente: El autor . . .	83

Índice de Códigos

5.1.	Detalle del código del scraping y las categorías exploradas. Fuente: el autor	39
5.2.	Detalle generación de modelos de aprendizaje automático. Fuente: el autor .	42
5.3.	Detalle generación de modelos de aprendizaje automático. Fuente: el autor .	49
5.4.	Detalle de código de modelo LDA para la extracción del subtema de la noticia. Fuente: el autor	55
5.5.	Detalle de código de carga de datos y preprocesamiento. Fuente: el autor . .	60
5.6.	Detalle de carga de modelos previamente entrenados. Fuente: el autor . . .	61
5.7.	Detalle de aplicación de modelo de clasificación de temática. Fuente: el autor	61
5.8.	Detalle de aplicación del modelo de extracción de la subtemática. Fuente: el autor	62
5.9.	Detalle de planteamiento de la red LSTM con múltiples entradas. Fuente: el autor	64
5.10.	Entrenamiento de la red LSTM con múltiples entradas planteada. Fuente: el autor	65
6.1.	Definición de red LSTM bidireccional. Fuente: el autor	68
6.2.	Función para prueba de predicciones. Fuente: el autor	69

Resumen

Durante los años 2020 y 2021 hemos sido testigos de cómo la pandemia de COVID-19 se ha propagado de manera exponencial por todo el mundo al mismo tiempo en el que una infodemia de desinformación relacionada se expande a una velocidad casi tan rápida como la velocidad a la que se multiplica el propio virus. Es así como, en esta investigación, se realizó el planteamiento de una metodología para abordar el problema de la detección de desinformación por medio de la inteligencia artificial, explorando diferentes técnicas y enfocando su caso de uso específicamente hacia la detección de desinformación sobre la pandemia de COVID-19.

El resultado fue una metodología unificada que combina técnicas de aprendizaje automático y de aprendizaje profundo en un modelo de múltiples entradas que concatena internamente la información del texto (input NLP¹) con la información numérica en forma de metadatos (input numérico) del encabezado de la información, para obtener una alerta de predicción que incluya información adicional, como la temática y subtemática de la información analizada; entregándole al usuario un contexto ampliado a partir del cual puede tomar una decisión más informada sobre replicar o no la información consultada.

Con una precisión superior al 90 %, el modelo resultante demostró un buen comportamiento al probarlo con encabezados de noticias actuales e incluso al validarlos con mitos ampliamente conocidos del COVID-19, por ejemplo: “Puede protegerse del COVID-19 inyectando, tragando, frotándose o bañándose con lejía corporal, desinfectantes o alcoholes” en donde el modelo predice que efectivamente, en este caso se detecta una alta probabilidad de desinformación y se genera una alerta.

Palabras Clave: Inteligencia artificial, aprendizaje automático, aprendizaje profundo, noticias falsas, desinformación, COVID-19, infodemia

¹Natural language processing

Abstract

Throughout 2020 and 2021 we have witnessed how the COVID-19 pandemic has spread exponentially around the world at the same time as a related disinformation infodemic is spreading almost as fast as the speed of the virus. This research proposes a methodology to address the problem of detecting disinformation through artificial intelligence, exploring different techniques focused in COVID-19 related disinformation detection use case.

The result was an unified methodology that combines machine learning and deep learning techniques in a multi-input model that internally mix the text information (NLP input²) and the metadata information (numerical input) from the news headline, resulting in a prediction alert that includes additional information, such as the topic and sub-topic of the analyzed information, giving to the user a broader context to make a decision on whether or not to replicate the information.

With an accuracy greater than 90 %, the final model showed good behavior when tested with current news headlines and even when tested with widely known myths about COVID-19 such as: “You can protect yourself from COVID-19 by injecting, swallowing, bathing in or rubbing onto your body bleach, disinfectants or rubbing alcohols” where the model predicts that indeed, in this case a high probability of misinformation was detected and an alert is generated.

Keywords: Artificial intelligence, machine learning, deep learning, fake news, disinformation, COVID-19, infodemic.

²Natural language processing

Capítulo 1

Introducción

Tres años antes del referéndum del Brexit¹ y de las elecciones presidenciales de EE. UU. de 2016, sucesos que derivaron en la acuñación del término “Fake News” o “Noticias Falsas” en español, el Foro Económico Mundial en su Reporte de Riesgos Globales ya dedicaba un capítulo completo titulado “*Incendios forestales digitales en un mundo hiperconectado*” en donde advertía sobre el imminente peligro de la desinformación difundida de manera incontrolable en las redes sociales [Klaus & Børge(2018)].

Las facilidades para el acceso a la información, los avances en la electrónica, la computación y la hiperconectividad del mundo de hoy, han sido el caldo de cultivo perfecto para la proliferación de desinformación en forma de noticias falsas; noticias que unas veces inocentes y otras no tanto, han logrado influir en momentos críticos de nuestra historia. Tal es el caso de las elecciones de EE. UU. de 2016, en donde el mismo presidente Donald Trump se atribuyó la autoría del término “Fake News” [Andrew(2019)], y por supuesto, sin ir más lejos, durante el año 2020 las hemos visto resurgir a partir de la pandemia de COVID-19 que irónicamente ha llegado a viralizar no solamente a las personas, sino también también a las redes sociales a partir de noticias que sólo buscan generar desinformación en algo que bien se podría denominar paralelamente la “*Infodemia de COVID-19*”.

Es así como se hace relevante la búsqueda de herramientas que puedan aportar en la detección de desinformación, explorando en este caso diferentes técnicas de inteligencia artificial y planteando a partir de aquí la presente investigación, que busca plantear una

¹El referéndum de pertenencia a la Unión Europea del Reino Unido, comúnmente conocido como referéndum del Brexit, tuvo lugar el 23 de junio de 2016 con el fin de consultar al electorado sobre su preferencia de si su país debería seguir siendo miembro o abandonar la Unión Europea. Con un 51.89 % de los votos el resultado fue la decisión de abandonar la Unión Europea.

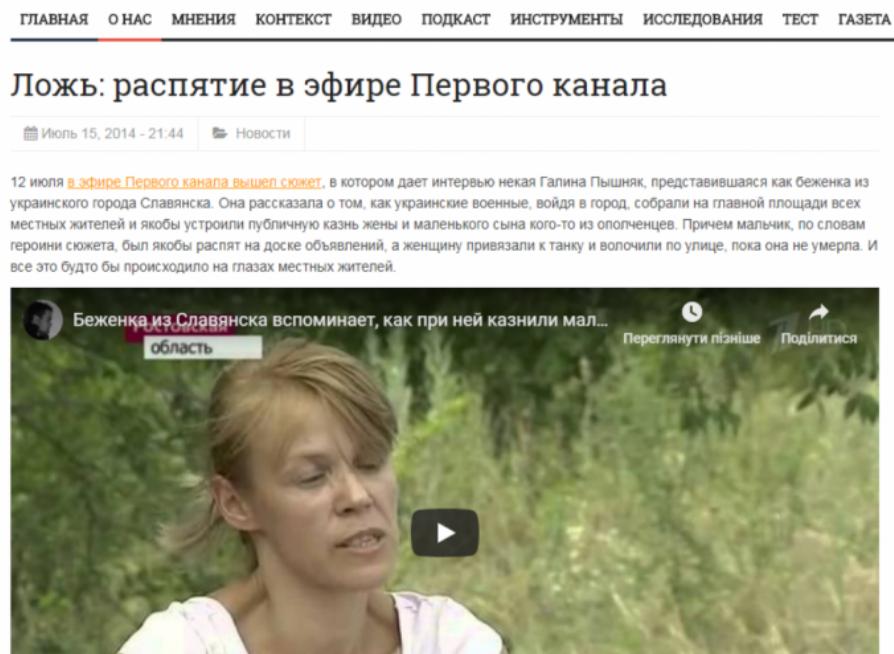


Figura 1.1: Ejemplo de noticia falsa con tintes políticos en Ucrania. Fuente: stopfake.org

metodología para la detección desinformación relacionada con la pandemia de COVID-19 por medio de técnicas de aprendizaje automático y procesamiento de lenguaje natural. A continuación se describe el desarrollo y resultado de la investigación.

1.1. Motivación

De acuerdo con el diccionario de Oxford, las noticias falsas se definen como: historias falsas que parecen ser noticias reales pero que tienen el objetivo de desinformar, difundidas a través de Internet u otros medios y que generalmente son creadas para influir en opiniones políticas o como bromas [Oxford(2020b)]. “*Si una historia es demasiado emocional o dramática, es probable que no sea real. La verdad suele ser aburrida*”, afirma Olga Yurkova, periodista cofundadora de la organización StopFake que combate las noticias falsas llenas de propaganda rusa que han inundado Ucrania desde la crisis de 2014 [Yurkova(2018)].

Un renombrado ejemplo de este tipo de noticias es una sobre la crucifixión de un niño en Ucrania. La noticia, distribuida en ese entonces por medios de comunicación rusos, presentaba a una mujer refugiada que aseguraba entre llantos que soldados ucranianos habían crucificado un niño de tres años en frente de su madre (Ver figura 1.1).

Resultó que la mujer era realmente la esposa de un militar y hasta el lugar de los hechos fue inventado, sin embargo, para cuando el engaño fue evidente ya era muy tarde;

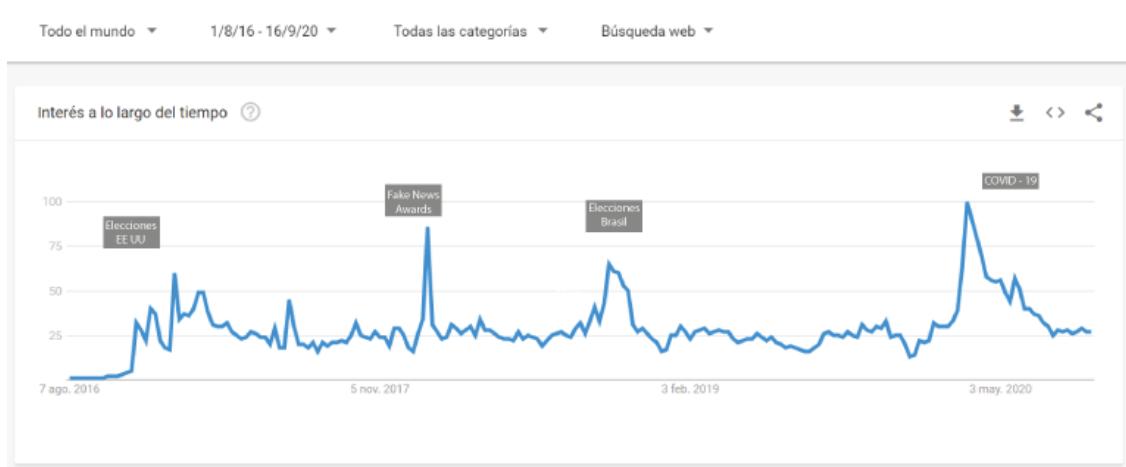


Figura 1.2: Tendencia de interés en el término "Fake News" en las búsquedas de Google desde el año 2016 en todo el mundo. Fuente: trends.google.com

"*la noticia del niño crucificado*" ya había motivado a ciudadanos ucranianos y rusos a que tomaran las armas. Es así como este tipo de noticias se vuelven peligrosas, convirtiéndose en verdaderas amenazas para la sociedad y la estabilidad de los países. Olga Yurkova lo expresa de forma tácita en su charla TED, cuando afirma: "*La gente ya no sabe lo que es real y lo que es falso, muchos han dejado de creer. Y esto es incluso más peligroso*" [Yurkova(2018)].

Noticias como estas han sido generadas cada vez con más frecuencia en los últimos años no solamente en Ucrania sino en todo el mundo, motivadas especialmente por acontecimientos importantes a nivel país, como elecciones presidenciales, o acontecimientos globales, como la reciente pandemia de COVID-19. Evidencia de esto se puede observar al realizar una consulta sobre las búsquedas del término "Fake News" en Google Trends² (Ver figura 1.2), en donde es evidente un aumento del interés durante el año 2016 (elecciones EE. UU. – Presidente Trump), durante el 2018 (elecciones Brasil – Presidente Bolsonaro) y un marcado aumento durante el año 2020 en línea con la infodemia mediática motivada por el nuevo coronavirus.

Es así como la proliferación de desinformación a través de noticias falsas se ha convertido recientemente en una problemática a resolver. Una de las formas más populares en la que se ha combatido esta problemática ha sido la verificación directa de las noticias, una estrategia de gran precisión pero que requiere un tiempo prolongado para realizar la respectiva investigación. Esto sin duda ha sido de gran ayuda a lo largo de mucho tiempo,

²Herramienta desarrollada por Google para buscar tendencias de búsqueda en una fecha específica.

pero hemos llegado a un punto donde la cantidad de noticias que se generan al día hace prácticamente imposible validarlas una a una, y la facilidad para su difusión y viralización hace que los tiempos disponibles para su validación sean cada vez más cortos. Por esta razón en años recientes, se ha iniciado la exploración de las técnicas de inteligencia artificial como herramientas para la detección de desinformación dentro de las noticias y es a partir de estas que se desarrollará esta investigación.

En este sentido el problema a resolver se concentrará en: ¿Cómo, por medio de técnicas de inteligencia artificial, podemos hacer una detección oportuna de desinformación relacionada con el COVID-19 dentro de una noticia, de manera que se evite su viralización y su potencial impacto negativo?

1.2. Planteamiento del trabajo

Una vez planteado el problema y evidenciado el gran interés en mitigarlo, el reto que se plantea para la investigación es definir una metodología que por medio de técnicas de inteligencia artificial logre disminuir el riesgo de la replicación de una noticia que potencialmente pueda contener desinformación. Gracias al desarrollo de la capacidad de procesamiento y a la disponibilidad de información, técnicas de inteligencia artificial ya se han venido aplicando en el problema de la detección de noticias falsas. Con el fin de establecer un punto de partida, se ha realizado una revisión inicial de trabajos relacionados, que se presentará en detalle en el capítulo 2.

Dentro de estos trabajos se reconoce la importancia de la problemática de la desinformación en forma de noticias noticias falsas y se plantean algunas oportunidades de investigación en temas relevantes que servirán de punto de partida para el desarrollo de la investigación, entre ellos:

- Incorporación de información adicional: Si un contenido o tema dado es digno de verificación se puede medir, por ejemplo, (i) su valor periodístico o su potencial para influir en la sociedad, por ejemplo, si está relacionado con los asuntos nacionales y puede conducir al pánico público, y (ii) su probabilidad histórica de ser una noticia falsa [Zhou & Zafarani(2018)].
- Uso de técnicas de aprendizaje profundo: Se ha demostrado su potencial en la investigación en detección de desinformación en forma de noticias falsas. Por ejemplo,

estudios recientes sobre noticias falsas han adoptado redes neuronales recurrentes (RNN) para representar publicaciones secuenciales y compromisos de usuarios o redes neuronales convolucionales (CNN) para capturar características locales de los textos e imágenes o ambos [Zhou & Zafarani(2018)].

- Clasificaciones multiclasa: investigaciones como la de Shoemaker proponen no solamente hacer una clasificación binaria (es o no es noticia falsa), sino incorporar una clasificación multiclasa que permita diferenciar por ejemplo las sátiras, otro tipo de noticia falsa que se crea con el ánimo de divertir [Shoemaker(2019)].
- Coincidencia entre contenido y titulares: Yang y Zhen incluyen dentro de sus trabajos futuros la exploración de la relación entre los títulos de las noticias y su contenido [Yang et al.(2018)Yang, Zheng, Zhang, Cui, Li & Yu].

Teniendo en cuenta lo anterior, el enfoque de esta investigación apuntará hacia las dos primeras oportunidades de investigación identificadas: la incorporación de información adicional sobre la noticia y el uso de técnicas de aprendizaje profundo.

El objetivo principal de la investigación será generar un modelo que permita identificar la existencia de desinformación relacionada con la pandemia de COVID-19 dentro de una noticia, marcándola como “*sospechosa*” para su doble chequeo. Se evaluará si la incorporación de las características anteriormente descritas puede ser de utilidad en la mejora de la efectividad del modelo.

1.3. Estructura de la memoria

La ruta de trabajo que se abordará en la investigación iniciará por una revisión del contexto y del estado del arte frente a la desinformación en forma de noticias falsas y el uso de inteligencia artificial para su detección, seguido por un planteamiento de los requerimientos, la definición de la metodología y finalizando con la prueba de esta para comprobar su precisión.

El desarrollo de la investigación se registrará a lo largo de los siguientes capítulos:

- En el *capítulo 2* se presenta el contexto y estado del arte como punto de partida para el planteamiento de la metodología.
- En el *capítulo 3* se definen formalmente los objetivos y la metodología de trabajo utilizada para abordar la investigación.

- En el *capítulo 4* se plantean los requisitos y componentes principales de la metodología.
- En el *capítulo 5* se detalla la descripción detallada de la metodología y su forma de aplicación.
- En el *capítulo 6* se describe la evaluación y validación de la metodología detallando los resultados obtenidos.
- Finalmente, en el *capítulo 7*, se presentan las conclusiones de la investigación y las recomendaciones para trabajos futuros que podrán surgir a partir de los resultados de obtenidos.

Capítulo 2

Contexto y Estado del Arte

Entendiendo la desinformación como información falsa o engañosa creada intencionalmente con el objetivo de engañar a la audiencia [Oxford(2020a)] y a las noticias falsas como historias falsas que parecen ser noticias creadas para engañar [Oxford(2020b)], es evidente que este tipo de noticias se ha convertido en nuestros tiempos en una de las formas más ampliamente usadas para la difusión de desinformación. Por esta razón y para fines de la presente investigación se acotará el estudio de la desinformación a su materialización en forma de “*noticias falsas*” difundidas a través de medios digitales o redes sociales.

Es así como para entender el contexto de la presente investigación, se propone dividirlo principalmente en dos partes. La primera parte está relacionada con la desinformación y las ampliamente conocidas noticias falsas, incluyendo su definición, principales características y tipologías. La segunda parte, está relacionada con las técnicas de inteligencia artificial utilizadas en los últimos años para la detección de desinformación dentro de una noticia. Al final del capítulo se identificarán las características comunes y oportunidades de investigación a partir de las cuales se plantea la investigación.

2.1. Desinformación y noticias falsas

2.1.1. ¿Qué son las noticias falsas?

Ha sido particularmente difícil encontrar una definición unificada de las noticias falsas, también conocidas popularmente en idioma inglés como “*Fake News*”. Un acercamiento interesante es presentado en el artículo “*Fake News: A definition*” en donde el autor, luego de revisar muchas de las definiciones existentes propone una primera definición



Figura 2.1: Denario de Marco Antonio y Cleopatra. Fuente: [Conversation(2017)]

formal que captura la mayoría de sus características distintivas: “*Las noticias falsas son una presentación deliberada de afirmaciones falsas o engañosas como noticias, donde las afirmaciones son engañosas por diseño*” [Gelfert(2018)].

Actualmente, en la definición también se han incorporado nuevos elementos como los canales de difusión, que en tiempos modernos han facilitado la explosión de este tipo de noticias, como por ejemplo el Internet. Vemos entonces que de acuerdo con el diccionario de Oxford, las noticias falsas se definen como: historias falsas que parecen ser noticias, difundidas a través de Internet u otros medios y que generalmente son creadas para influir en opiniones políticas o como bromas [Oxford(2020b)]. “*Si una historia es demasiado emocional o dramática, es probable que no sea real. La verdad suele ser aburrida*”, afirma Olga Yurkova, periodista cofundadora de la organización StopFake que se enfoca en combatir las noticias falsas de propaganda rusa que circulan en Ucrania [Yurkova(2018)].

2.1.2. Historia y evolución de la desinformación en forma de noticias falsas

La fabricación de información no es algo nuevo, a continuación se explorarán algunos hitos y hechos concretos a lo largo de la historia que pueden ser interesantes como referencia:

- 44 A.C - *La primera campaña de calumnias*: El primer emperador romano Augusto lanza una campaña de calumnias contra Marco Antonio con el fin de dañar su reputación que se distribuyó en frases muy cortas grabadas en monedas h. Las frases hacían ver a Marco Antonio como un títere de Cleopatra, un borracho y un mujeriego.



Figura 2.2: Ilustración del artículo del 26 de agosto de 1835. Fuente: [Wikipedia(2021)]

- **1493 - La invención de la imprenta:** La invención de la imprenta por parte de Johann Gutenberg sirvió como un nuevo canal de distribución y rápidamente se convirtió en un detonante que amplificó la difusión de la desinformación hacia un mayor volumen de público.
- **1835: La primera gran farsa periodística:** Conocida como “*El gran engaño de la Luna*” o “*Great Moon Hoax*”, en donde el periódico “*The Sun*” de Nueva York en una serie de 6 artículos afirmó que un prestigioso astrónomo británico de esa época, John Herschel (quien desconocía esta publicación), había descubierto a través de su telescopio vida en la Luna, describiendo desde ríos y bosques hasta humanoides con alas de murciélagos (Ver figura 2.2).
- **1899 – 1902: La guerra Bóer y el poder de la caricatura como sátira:** Durante esta guerra se difundieron estereotipos por medio de caricaturas satíricas con el objetivo de influir en la opinión pública y obtener mayor apoyo a la guerra. En la imagen vemos una caricatura inspirada en la novela *Los viajes de Gulliver*, en la que los pequeños Liliputienses, que representan a los Bóer, atan a un gigante que representa a los británicos mientras que la espada británica, etiquetada como “*British Prestige*” yace rota en el suelo. Este tipo de propaganda también fue común durante la Primera Guerra Mundial entre 1914 y 1918 (Ver figura 2.3).



Figura 2.3: Caricatura con tintes políticos. Fuente: [BoerWarArchive(2014)]

- **1938: La guerra de los mundos y la radio como canal de difusión:** Una radio novela de ciencia ficción basada en el libro *La Guerra de los Mundos* de H.G. Wells, hizo creer a casi un millón de personas en Estados Unidos que la tierra estaba bajo ataque de seres extraterrestres. Esto generó pánico en la población haciendo incluso que muchos oyentes huyeran de sus hogares con miedo (Ver figura 2.4).
- **1939 – 1945: La Segunda Guerra Mundial:** Sin duda la campaña de propaganda nazi en donde demonizaron y persiguieron a los judíos incitando al odio y a la cabeza de Joseph Goebbels fue esencial para motivar los horrores de esta guerra y legitimarlos públicamente hasta tal punto que aún hoy hay aquellos que niegan el Holocausto.
- **1947 – 1991: La Guerra Fría y otras guerras del siglo XX:** Desde la Guerra de Vietnam hasta la Guerra Fría, la propaganda fue usada como táctica de miedo para reprimir, para reclutar seguidores o para generar polarización entre las poblaciones.
- **1995 – 1998: La consolidación del Internet:** El nacimiento del Internet y su servicio estrella WWW (World Wide Web) permitió de forma sencilla la consulta de información de forma remota para todos. Este nuevo canal de difusión fue rápidamente aprovechado para la aparición de sitios webs dedicados a las noticias satíricas como por ejemplo “*The Onion*” que comenzaron la publicación en línea y que incluso se mantienen hasta hoy al aire ¹.

¹<https://www.theonion.com/>



Figura 2.4: Orson Welles ensayando su representación radiofónica de "La guerra de los mundos" de H.G. Wells.
Fuente: [NewYorkTimes(2017)].

- **2004 – 2006: El poder de las redes sociales - Surgimiento de Facebook y Twitter:** Ya en la década de los 2000, el surgimiento de las redes sociales potencializó la generación de noticias falsas y este fue el verdadero caldo de cultivo para lo que hoy en día conocemos con este término, habilitando su difusión a miles de millones de personas y minimizando su tiempo de replicación a segundos. Esto creó un nuevo concepto conocido como la *viralización de la información*.
- **2003 - 2011: La Guerra de Irak:** En 2003 el New York Times publicó una serie de artículos donde se afirmaba sobre la presencia de armas de destrucción masiva en Irak. Esta información nunca fue verificada y sin embargo Estados Unidos la citó como una de las razones para declarar la guerra contra Irak. Es así como estos artículos se han convertido en una de las noticias falsas que mayores consecuencias ha traído en términos de geopolítica global.
- **2014: Rusia y Ucrania:** En el marco del conflicto entre Rusia y Ucrania se conocen evidencias de un plan para inundar de propaganda rusa anti-occidental y pro-Kremlin diversos blogs, foros y redes sociales de medios electrónicos de comunicación ucranianos. Según las fuentes en un día promedio, los "*soldados de infantería digital*" debían publicar 50 veces artículos en mínimo de 6 cuentas de Facebook y replicarlos entre sí. Al final del mes se esperaba que se incrementaran los suscriptores de los blogs. En Twitter se esperaba que se manejaran diez cuentas con hasta 2000

seguidores cada una y se publicaran al menos 50 veces al día.

- *2016: Elecciones Presidenciales EE. UU. y acuñación del término “Fake News.”* La batalla por la presidencia de los Estados Unidos de 2016 motivó como nunca antes hasta ese momento, la propagación de noticias falsas especialmente en las redes sociales que llegaron a afirmar incluso el respaldo del Papa sobre la candidatura de Donald Trump. Esto motivó en ese momento una “fiebre de oro digital”, llegando a crear las llamadas “granjas de noticias falsas” que se lucraban de los ingresos publicitarios generados a través de la indignación provocada por noticias falsas que movilizaba a los usuarios a compartirlas. Irónicamente en este mismo año, el presidente electo de EE.UU. Donald Trump, utiliza la frase “fake news” al acusar a sus críticos y opositores (incluyendo a CNN) de publicar “fake news”. A partir de este momento el término “fake news” se popularizó a nivel mundial y fue usado para referirse a este tipo de noticias.
- *2017: Facebook inicia investigaciones sobre detección de noticias falsas:* Desde el 2017 y tras una acusación por facilitar la difusión de la desinformación, Facebook anuncia su interés en la investigación en el tema de las noticias falsas con el fin de marcarlas y reportarlas.
- *2018: Premios de Noticias Falsas:* El presidente de EE.UU. de ese momento, Donald Trump, a modo desafiante otorga a través de su cuenta de Twitter los “Fake News Awards” a medios de comunicación y críticos de su gobierno que considera que han tergiversado sus declaraciones.
- *2020: Las fake news cobran relevancia nuevamente:* Durante el año 2020 dos temáticas le vuelven a dar relevancia a las fake news, el primero relacionado nuevamente con las Elecciones Presidenciales de EE.UU. entre Joe Biden y Donald Trump, y el segundo relacionado con la pandemia de COVID-19.

Con este recuento histórico es evidente cómo los conflictos, cambios de régimen, crisis y catástrofes naturales, han sido los mayores caldos de cultivo para la generación de información falsa. Las noticias falsas han existido desde los principios de la historia y estas, además de su contenido construido deliberadamente con la intención de desinformar, están acompañadas a un canal de difusión que ha venido evolucionando; desde las monedas romanas usadas en contra de Marco Antonio, pasando por la imprenta y en nuestros tiempos



Figura 2.5: El desorden de la información. Fuente: [Ireton & Posetti(2018)]

el Internet y sus aplicaciones como las redes sociales que se fusionan en un mundo cada vez más hiperconectado. Nunca en la historia de la humanidad habíamos tenido un canal tan efectivo para diseminar la desinformación en forma de noticias falsas y esto, es lo que hoy ha hecho la diferencia.

2.1.3. Tipos de desinformación en noticias falsas

En su Manual de Educación y Capacitación en Periodismo [Ireton & Posetti(2018)], la UNESCO plantea que el término “*fake news*” o “*noticias falsas*” se queda corto para expresar toda la complejidad de la contaminación de información a la que estamos expuestos día a día y plantea algunos conceptos complementarios que permiten entender realmente la magnitud de las noticias falsas: información errónea, desinformación e información maliciosa (Ver figura 2.5).

- *Información errónea*: Es información falsa pero la persona que la está difundiendo cree que es verdad. No es intencional
- *Desinformación*: Es información falsa pero la persona que la difunde sabe que es falsa. Es deliberada e intencional
- *Información maliciosa*: Este tipo de información se basa en la realidad, pero se usa para hacer daño a una persona, organización o país.

Un estudio que compiló 34 artículos académicos que utilizaron el término *noticias falsas* entre 2003 y 2017 [Tandoc et al.(2018)Tandoc, Lim & Ling], dio como resultado una aproximación a la clasificación de los tipos de noticias falsas en: sátira, parodia, fabricación,

Intencionalidad de engañar del autor		
Nivel de veracidad	Alta	Baja
Alta	Publicidad Propaganda	Sátira
Baja	Manipulación Fabricación	Parodia

Tabla 2.1: Tipología de las noticias falsas [Tandoc et al.(2018)Tandoc, Lim & Ling]

manipulación, publicidad y propaganda. Para realizar estas definiciones el estudio se basó en dos dimensiones: nivel de veracidad e intencionalidad de engaño, muy en línea con lo definido también por la UNESCO (Ver figura 2.5).

- *Sátira*: Se refiere a noticias simuladas que suelen utilizar el humor o la exageración para presentar el contenido al público. Se considera la tipología más amplia.
- *Parodia*: Comparte características de la sátira ya que también se basa en el humor para atraer la audiencia, sin embargo en este caso la información no es real, mientras que en la sátira sí.
- *Manipulación*: Se refiere a la creación de narrativas falsas de manera deliberada. Es muy común en los contenidos visuales en donde se modifican imágenes o videos por medio de software de edición.
- *Publicidad*: Está relacionado con la generación de materiales publicitarios en forma de informes o noticias que parezcan genuinos que justifiquen o respalden productos o servicios.
- *Propaganda*: Se refiere a las noticias creadas por una entidad política para influir en las percepciones del público. Ha aumentado su interés debido a la relevancia de los acontecimientos políticos.

En la tabla (ver tabla 2.1) se presenta un resumen de las diferentes tipologías mencionadas, organizadas de acuerdo con las dimensiones de nivel de veracidad e intencionalidad de engaño. La primera dimensión, nivel de veracidad, está relaciona con qué tanto las noticias falsas se basan en hechos, mientras que la segunda dimensión tiene que ver con la intención directa del autor de engañar a la audiencia.

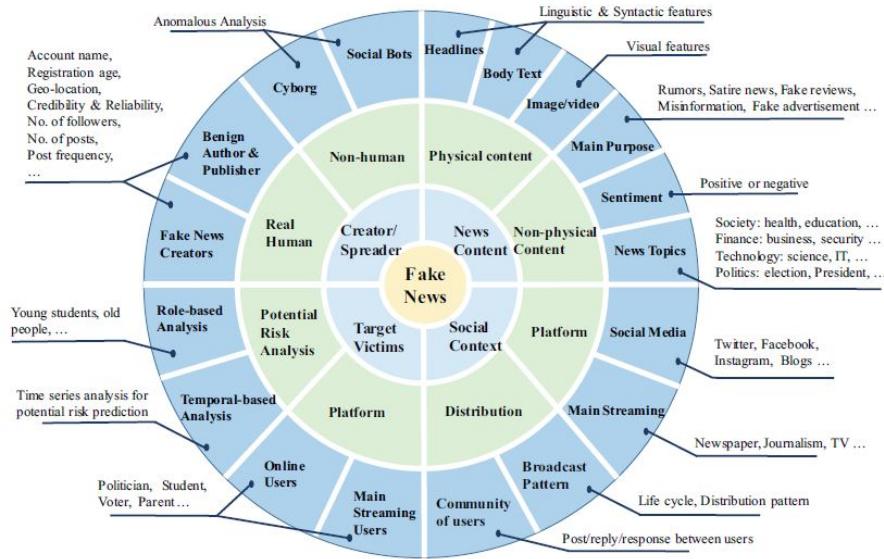


Figura 2.6: Caracterización de una noticia falsa. Fuente: [Zhang & Ghorbani(2020)]

2.1.4. Caracterización de una noticia falsa

Una completa caracterización de las noticias falsas se incluye en el trabajo de investigación [Zhang & Ghorbani(2020)], en donde se plantea un modelo basado en capas (Ver figura 2.6) con cuatro componentes principales:

- *Creador/difusor:* Puede ser humanos o no. Incluye aquellos que crean o publican la noticia con o sin intención
- *Víctima:* Son el principal objetivo de las noticias falsas. Pueden ser usuarios de las redes sociales en línea o de otras plataformas de noticias. Según el propósito de la noticias, las víctimas pueden ser estudiantes, votantes, padres, personas mayores, etc. También es quien toma una acción a partir del mensaje que puede ser ignorarlo, compartirlo en apoyo o compartirlo en oposición.
- *Contenido de noticias:* Se refiere al cuerpo de la noticia. Contiene tanto contenido físico (por ejemplo, título, cuerpo, multimedia) como contenido no físico (por ejemplo, tema, propósito, sentimiento)
- *Contexto social:* Indica cómo se distribuyen las noticias a través de Internet. El contexto incluye al usuario, sus redes y el patrón temporal de transmisión de la noticia a lo largo de esas redes.



Figura 2.7: Elementos de una noticia falsa. Fuente: [Zhang & Ghorbani(2020)]

Para verlo con un ejemplo de esta caracterización, en la figura 2.7 se presenta la caracterización de una noticia falsa muy famosa publicada en el 2017 en medio de la campaña por la presidencia de EE.UU. en donde se afirmaba que el Papa Francisco apoyaba la candidatura de Donald Trump. En este caso:

- **A** representa el Creador/Difusor: El usuario que lo comparte, en este caso Bob y Facebook.
- **B** representa el Contenido: Incluye el título de la noticia, el texto y el contenido multimedia (fotos y videos).
- **C** representa el contexto social: Incluye todas las interacciones entre otros usuarios y esta noticia (comentarios, me gusta/no me gusta, marca de tiempo)
- **D** representa las víctimas: Incluye a cualquier usuario que se involucre con la noticia por medio de las interacciones.

2.1.5. El ciclo de vida de una noticia falsa

En relación con la desinformación y en particular con las noticias falsas se plantean tres grandes momentos en su ciclo de vida: creación, publicación y distribución de la noticia [Ireton & Posetti(2018)]. Continuando con el ejemplo del caso de la noticia falsa del Papa

Francisco (Ver figura 2.7), a continuación, se identifica sobre la noticia cada uno de estos momentos:

- *Creación*: Artículo concebido por una persona no identificada.
- *Publicación*: Artículo publicado en el sitio de noticias WTOE5 (parte de una red de sitios de noticias fabricadas).
- *Distribución*: Artículo compartido en Facebook por integrantes de la red de sitios de noticias fabricadas y posteriormente replicado por partidarios de Donald Trump y partidarios de Hilary Clinton, difundiéndose en poco tiempo de manera viral.

Dentro del ciclo de vida de una noticia falsa y en especial en el momento de su distribución, es donde una noticia falsa se consolida como tal y se puede convertir en algo más destructivo que una bala de cañón. Se resalta la importancia de la persuasión, de quién comparte la noticia y en este caso voz a voz. Si bien los mensajes de los medios por sí solos generalmente no pueden convencer al público de algo contrario a sus actitudes o prejuicios existentes, si pueden reforzar poderosamente lo que la gente ya cree y, especialmente en la era de la información, las noticias falsas y la desinformación son más poderosas cuando se comparten, no sólo por las grandes fuentes noticiosas tradicionales, sino también por cualquier persona en las redes sociales, haciendo eco del llamado voz a voz.

2.2. Técnicas de inteligencia artificial aplicadas a la detección de noticias falsas

A continuación, se define de manera general el proceso de detección de noticias falsas a partir del cual en los últimos años se han utilizado las técnicas de inteligencia artificial. Se incluyen también algunas de las fuentes de noticias falsas verificadas más importantes, las cuales son conocidas como *Fack Checkers*, con el fin de tenerlos en cuenta para la recopilación de los conjuntos de datos necesarios para la investigación.

2.2.1. Detección de noticias falsas

La detección de noticias falsas puede estar relacionada directamente con la noticia y sus partes principales (título, cuerpo, autor, editor) o indirectamente con el contenido social

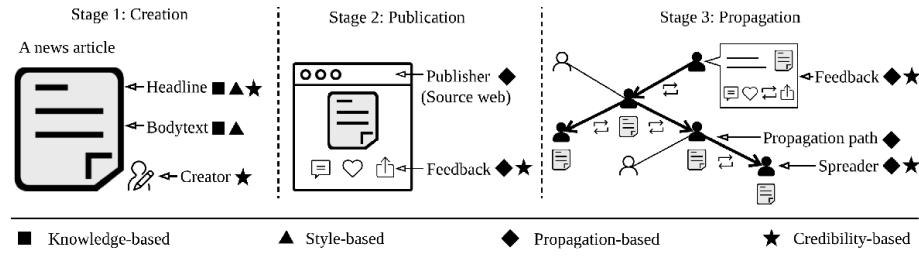


Figura 2.8: Enfoques de detección de noticias falsas a lo largo de su ciclo de vida. Fuente: [Zhang & Ghorbani(2020)]

(comentarios, red de propagación, difusores) cubriendo el ciclo de vida de la noticia (Ver figura 2.8, desde su creación, pasando por su publicación y finalizando con la distribución o difusión. En este sentido las noticias falsas se pueden estudiar desde cuatro enfoques [Zhou & Zafarani(2018)]:

- *Conocimiento*: ¿Qué tipo de contenido tienen las noticias falsas?
- *Estilo*: ¿Cómo se escriben usualmente las noticias falsas?
- *Propagación*: ¿Cómo se difunden las noticias falsas?
- *Credibilidad*: ¿Cuál es la reputación de sus creadores y difusores?

Independientemente del enfoque, el proceso más ampliamente difundido para la detección de una noticia falsa se basa en la verificación de hechos. Este proceso ha sido adoptado por organizaciones dedicadas a este fin, denominadas "*Fack-Checkers*" está dividida en 2 grandes fases: la extracción y la verificación. El proceso inicia con una entrada de un artículo sospechoso de noticia falsa, se procesa y tiene como salida un índice de confiabilidad.

La extracción de hechos tiene como fin construir una base de conocimiento, mientras que en la fase de verificación consiste en comparar los hechos descritos en la noticia con la base de conocimiento para calcular finalmente el índice de confiabilidad asociado al artículo sospechoso (Ver figura 2.9).

2.2.2. Uso de técnicas de inteligencia artificial para la detección de noticias falsas

Teniendo en cuenta el proceso de verificación de noticias descrito anteriormente, es fácil identificar que frente a un gran volumen de noticias de entrada, su verificación manual se convierte en un gran cuello de botella. El desarrollo reciente de las técnicas de inteligencia

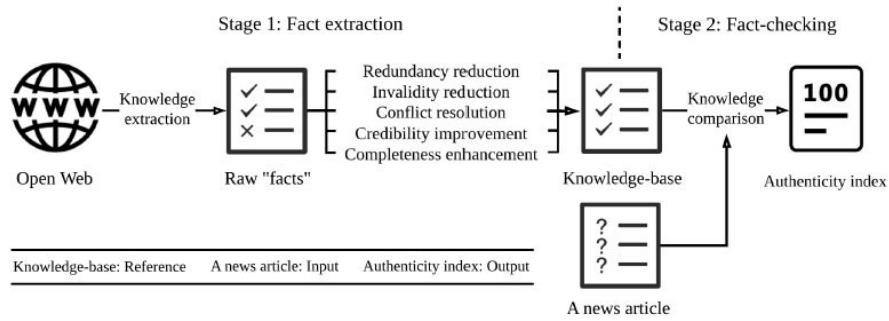


Figura 2.9: Proceso de detección de noticias falsas. Fuente: [Zhang & Ghorbani(2020)]

artificial ha motivado una gran cantidad de investigaciones orientadas a casos de uso relacionados con la detección de noticias falsas y se ha identificado que este puede ser uno de los caminos para automatizar el proceso de verificación de la información.

Dentro de los trabajos realizados se han definido claramente dos enfoques principales para abordar el problema de las noticias falsas y la detección de desinformación: la aplicación de técnicas de aprendizaje automático o *machine learning* y la aplicación de técnicas de aprendizaje profundo o *deep learning*.

Trabajos relacionados con técnicas de aprendizaje automático

- 2017: En [Shu et al.(2017)Shu, Sliva, Wang, Tang & Liu] los autores presentan una revisión integral de la detección de noticias falsas en las redes sociales, incluidas caracterizaciones de noticias falsas sobre psicología y teorías sociales desde una perspectiva de minería de datos, métricas de evaluación y conjuntos de datos representativos. También se discuten áreas de investigación relacionadas, problemas abiertos y futuras direcciones de investigación para la detección de noticias falsas en las redes sociales.
- 2018: En[Aphiwongsophon & Chongstitvatana(2018)] se propone el uso de técnicas de aprendizaje automático para detectar noticias falsas. En los experimentos se utilizan tres métodos populares: Naïve Bayes, Neural Network y Support Vector Machine (SVM), encontrando un mejor rendimiento con el método de Naïve Bayes.
- 2019: En[Agarwal et al.(2019)Agarwal, Sultana, Malhotra & Sarkar] se presenta una aproximación hacia la detección de noticias falsas basada en procesamiento de lenguaje natural y técnicas de aprendizaje automático. El modelo planeado realiza un preprocesamiento que incluye bag-of-words, n-grams y vectorización y posteriormente

te compara varios modelos de machine learning.

- 2020: En[Deb et al.(2020)Deb, Jha, Panjiyar & Gupta] se evalúan algunas de las técnicas de aprendizaje automático, principalmente la de Naïve Bayes, Random Forest, Decision Tree y SVM para problemas de clasificación automática de noticias de acuerdo a su temática o categoría.
- En[Kudarvalli & Fiaidhi(2020)] se realiza una aproximación a la detección de noticias falsas a través de métodos de machine learning incluyendo Radom Forest, Support Vector Machine y Naïve Bayes. Adicionalmente se incluyen algunas aproximaciones hacia un enfoque de aprendizaje profundo.

Trabajos relacionados con técnicas de aprendizaje profundo

- 2018: En [Ghosh & Shah(2018)], los autores proponen un método generalizado basado en redes neuronales profundas y procesamiento de lenguaje natural para identificar si una afirmación determinada es falsa o genuina. Se plantea un enfoque modular compuesto por dos partes principales, el primer submódulo a partir de la noticia a evaluar recupera artículos relevantes de una base de conocimiento que luego pueden ser de utilidad para verificar la veracidad de la afirmación. El segundo submódulo utiliza una red neuronal profunda para aprender el estilo de la falsificación del contenido.
- 2019: En [Bahad et al.(2019)Bahad, Saxena & Kamal] se aborda la problemática de la detección de noticias falsas a través de redes neuronales recurrentes del tipo LSTM bidireccional. Utiliza dos conjuntos de datos de artículos de noticias no estructurados disponibles públicamente para evaluar el rendimiento del modelo en comparación con otros métodos de redes neuronales profundas.
- 2020: En [Liu & fang Brook Wu(2020)], se enfocan en la detección temprana de noticias falsas utilizando datos observados en la etapa temprana de la propagación de noticias a partir de los cuales se genera un aprendizaje que permite identificar las noticias falsas. La investigación propone una red neuronal profunda con tres componentes: (1) un extracto de características de combinaciones de respuesta de texto de los usuarios y sus perfiles de usuario correspondientes, (2) un mecanismo de atención

que destaca respuestas importantes de los usuarios, y (3) un mecanismo de agrupación de medias de múltiples regiones para realizar la agregación de características.

- 2019: En [Dulhanty et al.(2019)] Dulhanty, Deglint, Ben Daya & Wong], se explora la noción de aprovechar modelos de lenguaje transformador bidireccional profundo a gran escala para codificar pares de reclamo-artículo en un esfuerzo por construir una detección de posturas de vanguardia orientada a identificar desinformación. Se enfoca en la detección de posturas, en la que, a partir de un reclamo y un artículo, se predice si el artículo está de acuerdo, en desacuerdo, no toma ninguna posición o no está relacionado con el reclamo.
- 2019: En [Drif et al.(2019)] Drif, Ferhat Hamida & Giordano], se propone un modelo de red neuronal convolucional (CNN) en conjunto con una arquitectura de red neuronal recurrente del tipo LSTM que aprovecha las características locales de grano grueso generadas por CNN y las dependencias de larga distancia aprendidas a través de la LSTM. Una evaluación empírica del modelo muestra una mejor precisión de predicción de la detección de noticias falsas, en comparación con Support Vector Machine y las líneas de base de CNN.
- 2020: En [Zhang et al.(2020)] Zhang, Dong & Yu] se presenta una red neuronal denominada FAKEDETECTOR. A partir de un conjunto de características explícitas y latentes extraídas de la información textual, FAKEDETECTOR construye un modelo de una red profunda del tipo DDNN (Deep Diffusive Neural Network) que permite conocer las representaciones de artículos periodísticos, creadores y sujetos de forma simultánea.
- 2018: En [Girgis et al.(2018)] Girgis, Amer & Gadallah], el objetivo fue la construcción de un clasificador que puede predecir si una noticia es falsa o no basándose únicamente en su contenido. El problema fue abordado desde una perspectiva puramente de aprendizaje profundo mediante modelos de técnica RNN (vainilla, GRU) y LSTM.
- 2018: En [Cai et al.(2018)] Cai, Li, Li & Wang] se plantea una red neuronal convolucional (CNN) para la clasificación de textos de noticias y se plantea de acuerdo a los resultados obtenidos que los clasificadores de texto tradicionales basados en métodos de aprendizaje automático tienen defectos relacionados con la escasez de

datos, explosión de dimensiones y poca capacidad de generalización, mientras que los clasificadores basados en redes de aprendizaje profundo mejoran en gran medida estos defectos, evita el engorroso proceso de extracción de características y tiene una gran capacidad de aprendizaje y una mayor precisión de predicción.

- 2020: En [Kula et al.(2020)Kula, Choraś, Kozik, Ksieniewicz & Woźniak], se presenta una solución innovadora para la detección de noticias falsas que utiliza métodos de aprendizaje profundo y adicionalmente combina el análisis de sentimientos.

Teniendo en cuenta la recopilación de investigaciones anteriormente listada, se observa que los métodos de aprendizaje automático más ampliamente utilizados son los de Random Forest, Naïve Bayes, Logistic y SVM, mientras que en el ámbito del aprendizaje profundo las técnicas más exploradas han sido las de redes neuronales convolucionales (CNN) y las redes neuronales recurrentes del tipo LSTM. Adicionalmente, es notable que en los últimos años la tendencia de las investigaciones se está orientando hacia las técnicas de aprendizaje profundo.

2.2.3. Pasos básicos de técnicas de inteligencia artificial aplicados a la clasificación de noticias

De acuerdo con los trabajos [Agarwal et al.(2019)Agarwal, Sultana, Malhotra & Sarkar], [Kudarvalli & Fiaidhi(2020)] y [Deb et al.(2020)Deb, Jha, Panjiyar & Gupta] los pasos básicos implementados en la clasificación de noticias falsas se resumen en la figura (Ver figura 4.1): captura de los datos, exploración, preprocesamiento, extracción de características, definición del modelo, entrenamiento del modelo, prueba del modelo, validación y evaluación del algoritmo entrenado.

En cuanto al preprocesamiento se puede subdividir en los siguientes pasos: eliminación de stopwords o palabras vacías, eliminación de espacios en blanco y signos de puntuación, tokenización, lematización y creación de conjuntos de características. Estos mismos pasos serán tenidos en cuenta en los próximos capítulos para la definición y desarrollo de la metodología.

2.3. Fuentes de noticias falsas

Una de las fuentes más confiables de noticias falsas son los conocidos *Fack Checkers*, organizaciones dedicadas exclusivamente a la verificación de hechos informativos. En la

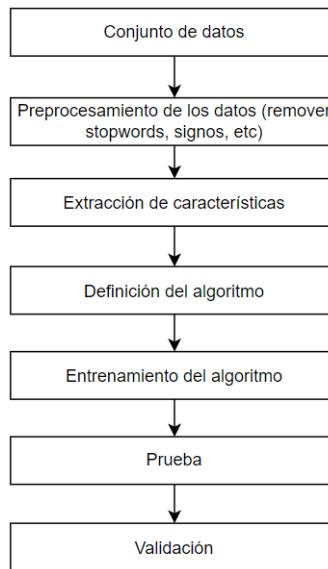


Figura 2.10: Proceso de análisis de noticias. Fuente: El autor

tabla 2.2 se listan algunos de los más conocidos. En Colombia, una de las organizaciones más conocidas que realiza la función de verificación de hechos es ColombiaCheck (<https://colombiacheck.com>).

2.4. Oportunidades de investigación identificadas

Trabajos como el de [Dulhanty et al.(2019) Dulhanty, Deglint, Ben Daya & Wong] y [Zhou & Zafarani(2018)], identifican claramente oportunidades y enfoques de investigación relacionados con la detección de noticias falsas las cuales. A continuación se incluyen algunos de ellos:

- *Incorporación de información adicional para la priorización:* Si un contenido o tema dado es digno de verificación, se puede medir (i) su valor periodístico, su potencial para influir en la sociedad (¿Está relacionado o no con asuntos nacionales?, ¿Puede generar pánico público?), (ii) su probabilidad histórica de ser una noticia falsa.
- *Uso de técnicas de aprendizaje profundo:* la capacidad de adaptación de estas técnicas ha demostrado que puede ser de utilidad, sin embargo es necesario tener en cuenta que se necesitará una gran cantidad de datos de entrenamiento.
- *Detección temprana:* Eficiencia en la verificación para identificar contenidos dignos de verificación.

Fact-Checker	Temática	Contenido analizado	Categorías de evaluación
Politifact	Política	Declaraciones	True, mostly true, half true, false, pants on fire
The Washington Post Fact Checker	Política	Declaraciones	One pinnochio, two pinnochio, three pinnochio, the geppetto checkmark, an upside-down pinnochio, veredit pending
FactCheck	Política	Declaraciones	true, no evidence, false
Snopes	Política Temas sociales	Artículos Videos	true, mostly true, mixture, mostly false, unproven, outdated, miscaptioned, correct attribution, misattributed, scam, legend
TruthOrFiction	Política Religión Naturaleza	Email rumors	truth, fiction
FullFact	Economía Salud Educación	Artículos	ambiguity
HoaxSlayer	Ambigüedades	Artículos Mensajes	hoaxes, scam, malware, bogus warning, fake news, misleading true, humour, spams
Poynter	Política Salud Economía	Artículos Noticias	True, false

Tabla 2.2: Fact-checkers más comunes. Fuente: El autor

- *Enfoque en el usuario y su intervención:* maximizar el compromiso del usuario en la verificación brindándole herramientas que le permitan mejorar su habilidad en la distinción de una noticia falsa de una verdadera.

Capítulo 3

Objetivos y metodología de trabajo

A partir de la problemática planteada en el *capítulo 1* y la línea base establecida en el *capítulo 2* con el contexto y estado del arte, a continuación se definen formalmente los objetivos que serán abordados a lo largo de la investigación.

3.1. Objetivo general

Investigar la viabilidad del uso de técnicas de procesamiento de lenguaje natural, aprendizaje automático y aprendizaje profundo para la clasificación y detección de desinformación dentro de titulares de noticias y a partir de un comparativo de estas técnicas, proponer una metodología que permita detectar efectivamente indicios de desinformación relacionada con la pandemia de COVID-19 permitiendo la verificación de la información sospechosa de manera oportuna por parte del usuario y evitando su replicación sin control.

3.2. Objetivos específicos

- Entender las características principales de la desinformación en forma de noticias falsas, su evolución a lo largo del tiempo, sus fuentes, canales de distribución y su potencial de impacto en la sociedad.
- Explorar las referencias relacionadas con la detección de desinformación en forma de noticias falsas por medio de técnicas de inteligencia artificial e identificar sus puntos comunes y oportunidades de investigación.

- Plantear las bases y requerimientos de la metodología a diseñar, detallando sus componentes, principales características, entradas y salidas, técnicas a utilizar y modo de funcionamiento.
- Diseñar y describir en detalle la metodología propuesta y el paso a paso para ser aplicada en la detección de desinformación.
- Realizar una prueba de la metodología planteada y analizar los resultados para detectar indicios de desinformación dentro de titulares de noticias.

3.3. Metodología de trabajo

Con el fin de alcanzar los objetivos planteados en la presente investigación, se plantea seguir una metodología de trabajo basada en el pensamiento de diseño o “*design thinking*” que incluye las fases de entendimiento, definición, ideación, prototipado y prueba [Brown(2008)]. Se ha identificado recientemente que este tipo de metodologías resultan de utilidad al momento de unirlas con el ámbito de la inteligencia artificial para realizar aproximaciones centradas en el usuario [Weller(2019)], lo cual se considera de gran relevancia teniendo en cuenta que siempre va a ser el usuario quien tomará la decisión de si replica o no una noticia. Para visualizarlo de una manera gráfica, en la figura 3.1 se presenta la metodología con cada uno de sus pasos y cómo se relacionan entre sí.

A continuación, se describen uno a uno los pasos, su alcance y el capítulo de esta memoria resultante de su ejecución:

- **Paso 1: Entender**

Contexto de las noticias falsas

Contexto de técnicas de inteligencia artificial

Este paso ya se surtió y tuvo como resultado el *Capítulo 2*

- **Paso 2: Definir: conectar, detectar, dimensionar**

Definición de requerimientos y especificaciones de la metodología

Definir los datos necesarios y sus características

Este paso tendrá como resultado el *Capítulo 4*

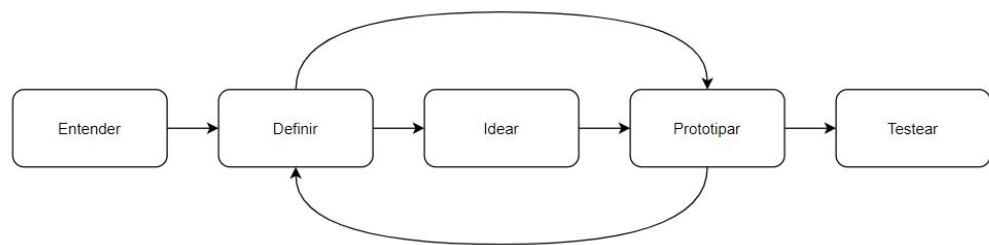


Figura 3.1: Metodología definida para el abordaje de la investigación. Fuente: el autor

■ **Paso 3: Idear – Prototipar**

- Recolección de los datos
 - Preparación de los datos
 - Análisis exploratorio
 - Comparativo de técnicas de inteligencia artificial
 - Definición del modelo
 - Entrenamiento del modelo
 - Construcción de modelo
- Este paso tendrá como resultado el *Capítulo 5*

■ **Paso 4: Probar**

- Prueba con datos reservados y nunca vistos por el modelo
- Este paso tendrá como resultado el *Capítulo 6*

Capítulo 4

Identificación de requisitos y planteamiento de la metodología

A partir de los objetivos planteados, en este capítulo se definieron las bases y requerimientos de la metodología, incluyendo: el detalle de sus componentes, características principales, entradas y salidas, técnicas a utilizar y modo de funcionamiento. También se definirán los datos necesarios para aplicar la metodología con el fin de que puedan ser recolectados y preparados para ser procesados.

4.1. Características principales de la metodología

Teniendo en cuenta las oportunidades de investigación identificadas en el capítulo anterior, se propone abordar la problemática de la detección de desinformación en forma de noticias falsas por medio de técnicas de inteligencia artificial a partir de una metodología enfocada en las siguientes características:

- *Que permita incorporar información adicional sobre la noticia a analizar:* La información adicional sobre la noticia podrá ser de utilidad para que el usuario pueda decidir si debe re-verificar la información dentro de la noticia. En este sentido se propone incorporar la temática de la noticia y posibles subtemas o temáticas dentro de la noticia.
- *Que permita una detección ágil:* Permitir que la metodología pueda ser implementada posteriormente para la detección de noticias de manera temprana (incluso antes de la replicación). Esto quiere decir que el tiempo de procesamiento será relevante a al



Figura 4.1: Entradas y salidas de la metodología propuesta. Fuente: el autor

hora de aplicar un modelo por lo que se propone inicialmente enfocar el análisis en el texto de la noticia y específicamente en su título o titular.

- *Que permita una priorización del contenido:* atacar el problema de las noticias falsas no en su detección directa sino en la priorización de su contenido para una posterior revisión a fondo y responder la pregunta de si el agregar información adicional al análisis como el tema de la noticia permite aportar en su detección.
- *Que funcione con noticias recientes:* y específicamente con la temática seleccionada relacionada de la pandemia de COVID-19, la cual le da una utilidad muy puntual y relevante a la metodología.
- *Que funcione inicialmente con noticias en idioma inglés:* teniendo en cuenta que la mayoría de las noticias inicialmente se generan en este idioma, además de que muchas de las técnicas y herramientas de procesamiento de lenguaje natural se encuentran también en inglés.

El enfoque planteado le permitirá al usuario detectar una noticia con posible contenido de desinformación de manera temprana y contando con información adicional para decidir si la replica o no (como por ejemplo la temática de la noticia). Esto permite que el usuario cuente con herramientas para que su proceso de decisión sea más objetivo y menos subjetivo (Ver figura 4.1).

4.2. Propuesta conceptual de la metodología

Conceptualmente la propuesta consiste en desarrollar una metodología que permita intervenir el ciclo de vida de una noticia con posible contenido de desinformación en el momento preciso en el que el usuario toma la decisión de replicarla o no a sus contactos.

En la parte superior de la figura 4.2 se observa el ciclo de vida actual de la noticia, donde el emisor o autor de la noticia la crea, la publica y la difunde a través de los canales

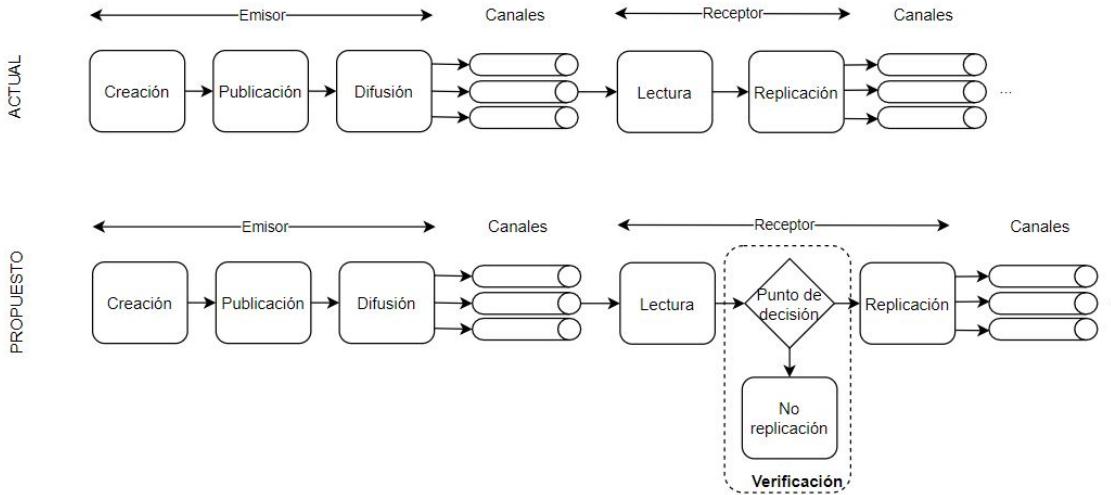


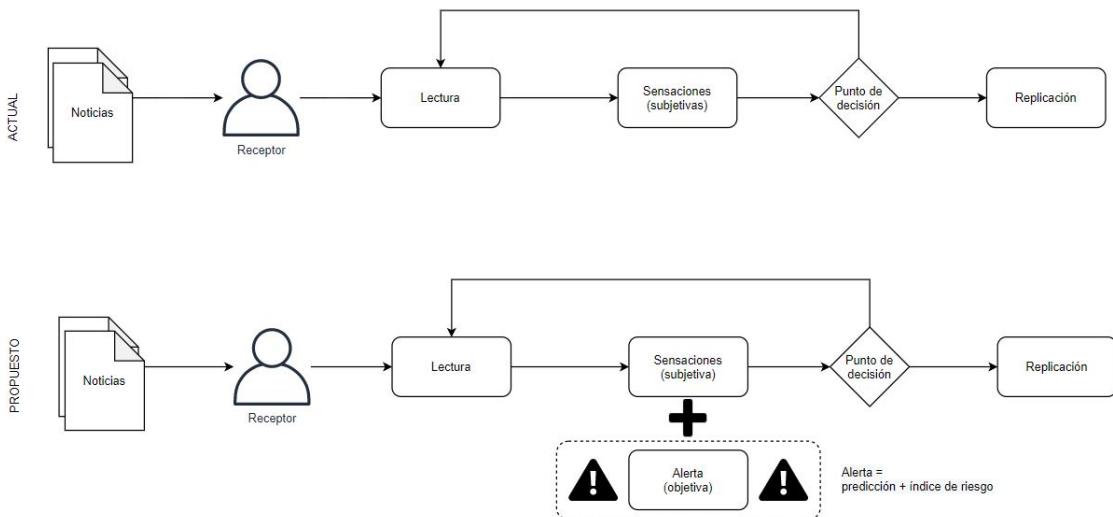
Figura 4.2: El ciclo de vida de la noticia. Fuente: el autor

digitales de su elección. La noticia llega al receptor, quien la lee y si es de su interés lo normal es que inmediatamente la replique a sus contactos casi sin pensarlo. En este punto el usuario no cuenta con información adicional sobre la noticia que le permita hacer un alto y evaluar si debe o no replicar la noticia. Para el ciclo de vida propuesto de la noticia que se detalla en la parte inferior de la figura 4.2 , la idea es crear un punto verificación en donde el usuario podrá hacer un alto antes de replicar una noticia y tomar una decisión basada en una mayor cantidad de información sobre la replicación de la noticia que ha llegado a sus manos.

La información adicional que se le entrega al usuario pretende complementar la información subjetiva, facilitando la decisión de replicar o no una noticia. Esta información será la predicción del modelo acompañado por información adicional de la noticia, como por ejemplo el tema de la noticia y conformarán una alerta que el usuario podrá combinar con las sensaciones que le genere la noticia para tomar una decisión más consciente de compartir la noticia o no (Ver figura 4.3).

4.2.1. Componentes de la metodología

Teniendo en cuenta que la metodología propuesta se basará en la extracción de características adicionales de la noticia como por ejemplo la temática principal y/o posibles subtemas dentro de esa temática que puedan ser de interés, se proponen tres componentes principales de la metodología:

**Figura 4.3:** Detalle del proceso de verificación. Fuente: el autor

- El primero estará relacionado con la *extracción del tema principal* de la noticia, es decir, deberá detectar si la noticia es de política, de medio ambiente, de deportes y por supuesto de salud, entre otros temas adicionales.
- Teniendo en cuenta que la mayoría de las noticias a analizar estarán dentro de la gran temática de “*salud*”, se propone que el segundo componente sea el *subtema de la noticia*, es decir, si estamos hablando de noticias de salud, de qué tema específicamente se trata la noticia, por ejemplo: salud pública, vacunas, epidemias, entre otros.
- El tercer gran componente de la metodología será la *predicción de la alerta* de posible contenido con desinformación. Esta parte considerará como entrada tanto el texto del titular de la noticia como el resultado de la clasificación de la temática y subtemática de la noticia.

En este sentido como se detalla en la figura 4.4), la metodología incluirá tres modelos de predicción. Los dos primeros estarán relacionados con la adquisición de la información adicional de la noticia, en este caso la temática y la subtemática, mientras que el tercer modelo corresponderá a la generación de la alerta de detección de posible contenido con desinformación.

Como resultado, a la salida de la aplicación de los tres modelos el usuario recibe un paquete de información resultante de analizar la noticia y puede contar con información no subjetiva para decidir si la replica o la reporta como posible noticia falsa para revisión

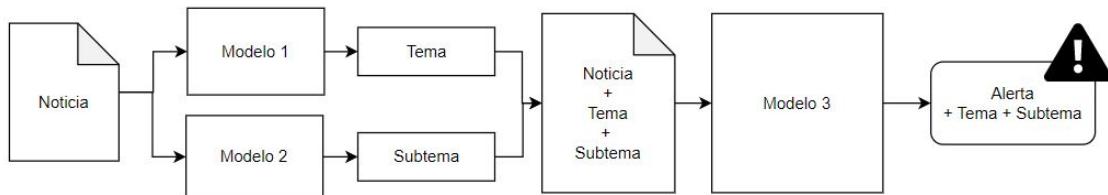


Figura 4.4: Componentes de la metodología propuesta. Fuente: el autor

a fondo de la noticia hacia alguna de las entidades de verificación de noticias conocidas (Ver tabla 2.2).

4.2.2. Técnicas de inteligencia artificial a utilizar

A pesar de que a partir de la revisión de la documentación del Estado del Arte (Ver Capítulo 3) es evidente que las investigaciones se están orientando hacia las técnicas de aprendizaje profundo, para definir objetivamente las técnicas a considerar como parte de la metodología, se propone realizar una comparación entre técnicas de aprendizaje automático y aprendizaje profundo. En ese sentido para la comparación se tendrán en cuenta las tres principales técnicas en cada caso, es decir para las técnicas de aprendizaje automático y de acuerdo a las mayormente usadas en la revisión bibliográfica se considerarán: Naïve Bayes, Random Forest, Logistic y Máquina de Vector de Soporte, mientras que para las técnicas de aprendizaje profundo se tendrán en cuenta: Redes Neuronales Fully Connected, Redes Neuronales Convolucionales (CNN) y Redes Neuronales Recurrentes (RNN) del tipo LSTM.

4.3. Definición de datos necesarios

Para el desarrollo de la metodología se ha identificado que se requieren al menos dos conjuntos de datos, el primero para el desarrollo del modelo de clasificación de la temática de la noticia y el segundo para la generación de la alerta de una posible noticia falsa.

4.3.1. Buscadores de datos

A continuación se listan algunas herramientas utilizadas durante la búsqueda de los conjuntos de datos requeridos.

- Google Dataset Search: <https://data.mendeley.com/>
- Kaggle Datasets: <https://www.kaggle.com/datasets>
- Mendeley Data: <https://data.mendeley.com/>

Adicionalmente, el artículo [Nakamura et al.(2020) Nakamura, Levy & Wang] incluye un comparativo de conjuntos de datos específicos para la detección de noticias falsas que también se consideró durante la búsqueda.

4.3.2. Conjuntos de datos para el modelo de clasificación de noticias

Inicialmente se exploraron los siguientes conjuntos de datos:

- BBC news classification: Este conjunto de datos contiene un total de 2225 artículos clasificados por la BBC en 5 categorías: sports, tech, business, entertainment, politics. Año: 2018, URL: <https://www.kaggle.com/c/learn-ai-bbc>
- News Category Dataset: Este conjunto de datos estructurado en .json contiene información de noticias desde 2012 a 2018 del periódico en línea HuffPost¹. Los artículos están clasificados en 41 categorías que van desde política, hasta educación. Año: 2018. URL: <https://www.kaggle.com/rmisra/news-category-dataset>
- AG's News Topic Classification Dataset: Este conjunto de datos estructurado en .json contiene información de noticias de diversas fuentes. El conjunto de entrenamiento contiene aproximadamente 120000 noticias y el de test contiene 8000. Los artículos están clasificados en 4 categorías: world, sports, business, science. Año: 2018. URL: <https://www.kaggle.com/amananandrai/ag-news-classification-dataset>

En el caso de el conjunto de datos para el desarrollo del modelo de clasificación de las temáticas de noticias, no se encontró un conjunto de datos actualizado con noticias recientes (2019-2020). En este sentido se exploraron directamente fuentes de noticias, entre ellas particularmente la de Reuters (<http://reuters.com>), en donde se publican las noticias agrupadas por categorías (Ver figura 4.5).

Es así como se decidió realizar un webscraping de las noticias publicadas en Reuters para el desarrollo del modelo. En el siguiente capítulo se describirá en detalle este proceso y los resultados obtenidos.

¹<https://www.huffpost.com/>

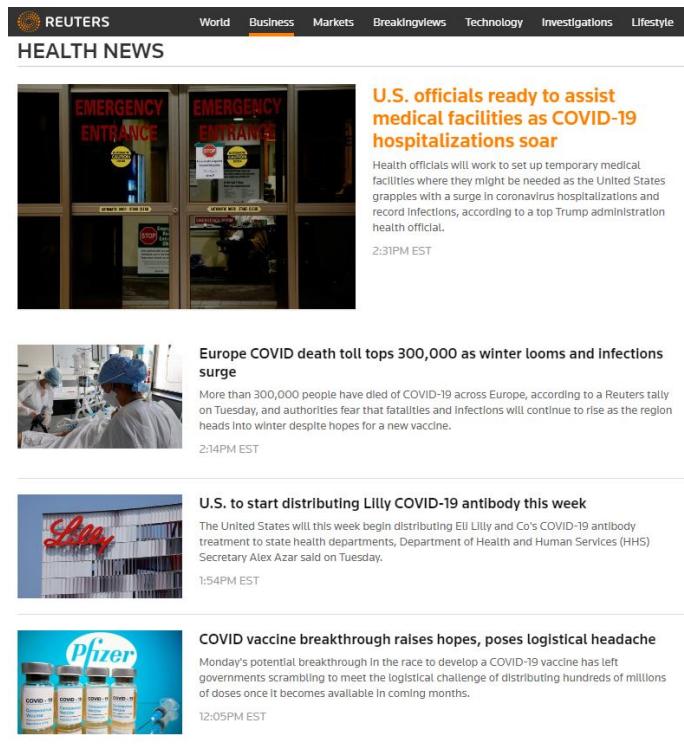


Figura 4.5: Noticias de reuters clasificadas en este caso en la categoría "Health". Fuente: reuters.com/news/health

4.3.3. Conjuntos de datos para exploración de subtemas y modelo de generación de la alerta

Inicialmente se exploraron los siguientes conjuntos de datos:

- Fake News: Este conjunto de datos contiene 5 columnas, un id del artículo, título, autor, cuerpo de la noticia y la etiqueta. Cuenta con 24000 noticias. Año: 2017, URL: <https://www.kaggle.com/c/fake-news/>
- Fake an real news dataset: Recopilación de alrededor de 40000 artículos de noticias agrupados en dos sets, uno de noticias reales y otro con noticias falsas que van desde el 2016 al 2018. Año: 2018, URL: <https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset>
- Fake News Challenge FNC-1: Este conjunto de datos se organizó para el concurso FakeNewsChallenge lanzado en 2016. Cuenta con 50000 artículos etiquetados. Año: 2017, URL: <https://github.com/FakeNewsChallenge/fnc-1>

Uno de los requerimientos de los conjuntos de datos para la investigación es que cuente con noticias recientes para que pueda ser probado adecuadamente, sin embargo se observa

que ninguno de los conjuntos de datos encontrado cuenta con noticias de actualidad. Se considera que este requisito es importante teniendo en cuenta que durante el presente año el volumen de noticias falsas se ha incrementado debido a la crisis generada por la pandemia de COVID-19 y seria ideal poder probar la metodología propuesta bajo el contexto actual. En ese sentido se realizó una nueva búsqueda de conjuntos de datos con noticias más recientes enfocadas en esta temática.

4.3.4. Conjuntos de datos relacionados con la pandemia de COVID-19

Se encontraron los siguientes conjuntos de datos de noticias falsas relacionados con la pandemia de COVID-19:

- FakeCovid [Shahi & Nandini(2020)]: Este conjunto de datos contiene artículos recolectados desde Poynter² y Snopes³. Incluye 5182 artículos en varios idiomas, incluyendo español que han circulado en 105 países y que han sido directamente verificados por fact checkers. El intervalo de fechas va desde enero hasta mayo de 2020. Año: 2020, URL: <https://gautamshahi.github.io/FakeCovid/>
- COVID19FN [AGARWAL(2020)]: Recopilación de alrededor de 2800 artículos de noticias etiquetadas recolectadas desde sitios webs de Fact Checkers desde enero hasta junio de 2020. Año: 2020, URL: <https://data.mendeley.com/datasets/b96v5hmfv6/3>

²<https://www.poynter.org/>

³<https://www.snopes.com/>

Capítulo 5

Descripción y desarrollo de la metodología

Este capítulo inicia con el diseño detallado de la metodología, a partir del cual se desarrollará paso a paso cada uno de los modelos que la conforman, desde la exploración y preparación de los datos hasta el diseño, compilación, entrenamiento, validación y prueba de los modelos. Se realizará una comparación de diferentes técnicas de aprendizaje automático y aprendizaje profundo con el fin de identificar las que obtengan un mayor rendimiento y de acuerdo a estas plantear un modelo unificado que mejore el rendimiento por separado de cada una de las técnicas.

5.1. Diseño detallado de la metodología

La metodología planteada estará compuesta principalmente por tres modelos que se combinan a la salida para presentar un único resultado al usuario que le sirva como orientación para tomar la decisión de si replica o no una determinada noticia que llega a sus manos. A partir de los conjuntos de datos definidos se realizará el desarrollo y entrenamiento de los modelos, para posteriormente ya con los modelos entrenados, aplicarlos a cualquier noticia *sospechosa* que reciba el usuario.

En la figura 5.1 se detalla el diseño de bloques que incluye el proceso desde los conjuntos de datos hasta los modelos resultantes entrenados para la predicción de la temática y subtemática, y cómo se deben combinar al final en un modelo unificado de alerta que genere un único resultado de cara al usuario.

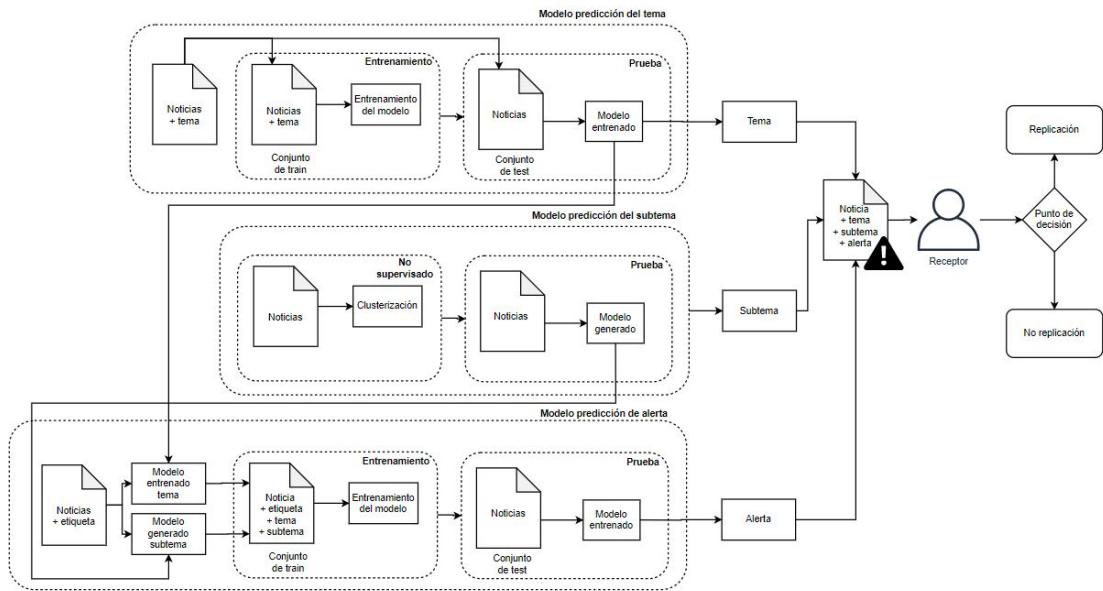
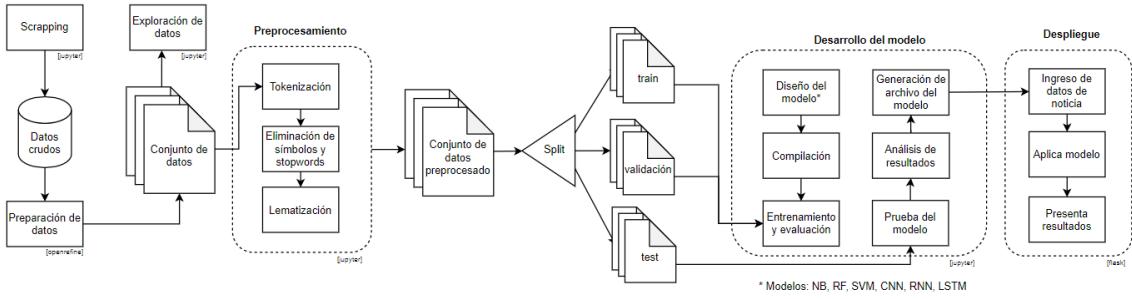


Figura 5.1: Diseño detallado. Fuente: el autor

El desarrollo de cada uno de los modelos incluye desde la recolección y preparación de los datos, su exploración, preprocesamiento, división en conjuntos de entrenamiento, validación y test, diseño del modelo, compilación, entrenamiento, validación y prueba. Como resultado se obtendrá un conjunto de modelos entrenados que podrán cargarse e implementarse para la predicción de desinformación dentro de noticias reales.

La figura 5.2 detalla paso a paso el proceso de transformación de los datos crudos, su preparación para conformar los conjuntos de datos, el preprocesamiento de los datos, la conformación de los conjuntos de entrenamiento, validación y prueba, el desarrollo del modelo (incluyendo su diseño, compilación, entrenamiento y evaluación), la prueba del modelo, el análisis de resultados, la generación del archivo del modelo y finalmente su despliegue de cara al usuario.

En cuanto a herramientas y plataformas para el desarrollo de los modelos en cada una de estas etapas se propone para las etapas de preprocesamiento y desarrollo del modelo utilizar Python con Jupyter Notebook, en este caso haciendo uso de la herramienta Google Colab para aprovechar las capacidades de procesamiento en la nube. En el caso del despliegue se propone utilizar Flask con el fin de facilitar el desarrollo de una sencilla interfaz de usuario para probar el modelo. Este proceso se detallará en el Capítulo 6.

**Figura 5.2:** Detalle de modelos a probar. Fuente: el autor

5.2. Recolección y preparación de los datos

En línea con lo definido en el Capítulo 4, para el modelo de clasificación de noticias y con el fin de garantizar la actualidad de las noticias, se realizó un webscraping por medio de Python y la librería Beautiful Soup¹, desplegando el proceso de captura desde un notebook de Google Colab. En cuanto al modelo de alerta se seleccionaron los conjuntos de datos relacionados con COVID-19 (FakeCovid [Shahi & Nandini(2020)] y COVID19FN [AGARWAL(2020)]) y se fusionaron en un único conjunto de datos. A continuación se detalla el proceso de preparación de cada uno de estos conjuntos de datos.

```

1 # Funcion lista paginas
2 def lista_de_paginas(numero , pagina):
3     feeds_list = []
4     for i in range(1,numero+1):
5         item=pagina+"?view=page&page="+str(i)
6         feeds_list.append(item)
7     return feeds_list
8 # Crear lista de paginas
9 N_pginas=100
10 Pag_ini="https://www.reuters.com/news/health" # Health
11 lista_pginas_cat1=lista_de_paginas(N_pginas , Pag_ini)
12 # Scraping de la lista de paginas
13 n_news=12 # Noticias por pagina
14 for feed in lista_pginas_cat1:
15     response = requests.get(feed)
16     xml_page = bs4.BeautifulSoup(response.text , "lxml")
17     val=0
18     for sub_heading in xml_page.find_all('p'):
19         val=val+1
20         if val <= n_news:
21             description_list.append(sub_heading.contents)

```

Código 5.1: Detalle del código del scraping y las categorías exploradas. Fuente: el autor

¹Beautiful Soup es una biblioteca de Python para capturar información desde documentos HTML

0	Diego Maradona has been admitted to hospital i...
1	The United States Golf Association added 10 pl...
2	Mainland China reported 49 new COVID-19 cases ...
3	The Danny O'Brien-trained King of Leogrance ha...
4	Chicago Bears wide receiver Javon Wims was sus...
...	...
1195	Following are facts and records ahead of the 1...
1196	Penpix of the top women's contenders at the 20...
1197	Novak Djokovic will release any pent up frustr...
1198	As a polarizing campaign that has shattered ea...
1199	Democratic presidential candidate Joe Biden sa...

1200 rows x 1 columns

Figura 5.3: Resultado del scraping de Reuters.com. Fuente: el autor

5.2.1. Preparación de datos modelo de predicción del tema de la noticia

El *scraping* se realizó el día 29 de octubre de 2020 para las categorías: *Politics*, *Health*, *Enviroment*, *Technology*, *Finance*, *Lifestyle*, *Science* y *Sports* (Ver código 5.1). Se capturaron aproximadamente 10000 titulares de noticias de cada una de las categorías. La idea principal del ejercicio fue contar con noticias actualizadas que permitirán generar un modelo que pueda ser probado con noticias actuales (Ver figura 5.3). Los datos capturados se incluyen en el repositorio de la investigación ubicado en la URL: <https://github.com/jorgecif/CovidMisinformationDetection/tree/main/data>.

5.2.2. Preparación de datos modelo de predicción de la alerta

Teniendo en cuenta que los conjuntos de datos de noticias falsas relacionadas con COVID-19 (FakeCovid [Shahi & Nandini(2020)] y COVID19FN [AGARWAL(2020)]) son relativamente pequeños, se decidió combinarlos en un nuevo conjunto de datos más grande. Además, también se realizó una depuración de los datos, eliminando registros duplicados, nulos y unificando categorías de las columnas. Para esta tarea se utilizó una herramienta desarrollada inicialmente por Google y liberada, denominada OpenRefine².

El resultado fue un conjunto de datos con alrededor de 20000 registros, a partir del cual se realizó el análisis y desarrollo de la metodología.

²OpenRefine (<https://openrefine.org/>)

11285 rows								Extensions: Zemanta ▾ Firebase ▾ RDF ▾ CK
Show as: rows records		Show: 5 10 25 50 rows						
All	Capital or Reven	Directorate	Transaction Num	Date	Service Area	Expenses Type	Amount	Supp
1. Revenue	Community Wellbeing & Social Care	5105695748	05.04.2013	Youth & Community	Operational Equipment	120	REDACTE PERSON	
2. Revenue	Community Wellbeing & Social Care	5105695748	05.04.2013	Youth & Community	Operational Equipment	80	REDACTE PERSON	
3. Revenue	Community Wellbeing & Social Care	5105698650	24.04.2013	Leaseholds by LA	Accommodation Costs - Leaseholder Payments	695.89	REDACTE PERSON	
4. Revenue	Community Wellbeing & Social Care	5105698650	24.04.2013	Leaseholds by LA	Accommodation Costs - Leaseholder Payments	695.89	REDACTE PERSON	
5. Revenue	Community Wellbeing & Social Care	5105698650	24.04.2013	Leaseholds by LA	Accommodation Costs - Leaseholder Payments	695.89	REDACTE PERSON	
6. Revenue	Community Wellbeing & Social Care	5105698650	24.04.2013	Leaseholds by LA	Accommodation Costs - Leaseholder Payments	695.89	REDACTE PERSON	
7. Revenue	Community Wellbeing & Social Care	5105698650	24.04.2013	Leaseholds by LA	Accommodation Costs - Leaseholder Payments	695.89	REDACTE PERSON	
8. Revenue	Chief Executive, Schools & Learning	5105698316	19.04.2013	L&A Commissioned Activity	Bought in Prof Services - Curriculum (Schools)	250	REDACTE PERSON	
9. Revenue	Chief Executive, Schools & Learning	5105698318	19.04.2013	L&A Commissioned Activity	Bought in Prof Services - Curriculum (Schools)	710	REDACTE PERSON	
10. Revenue	Economy & Environment	5105695879	05.04.2013	IV Biological Record	General Materials	220.2	REDACTE PERSON	
11. Revenue	Chief Executive, Schools & Learning	5105696514	12.04.2013	Adult Services Training	Training and Conferences	150	REDACTE PERSON	
12. Revenue	Community Wellbeing & Social Care	5105695832	10.04.2013	Short Breaks	Payments to Voluntary and Other Associations	1,280.00	REDACTE PERSON	
13. Capital Resources	Capital Resources	5105696504	12.04.2013	Capital Receipts	External Design and Supervision Fees	400	REDACTE PERSON	
14. Capital Resources	Capital Resources	5105696505	12.04.2013	Capital Receipts	External Design and Supervision Fees	1,350.00	REDACTE PERSON	
15. Revenue	Economy & Environment	5105698707	12.04.2013	School Reorganisation	Security of Buildings	300	REDACTE PERSON	
16. Revenue	Economy & Environment	5105698717	12.04.2013	School Reorganisation	Security of Buildings	300	REDACTE PERSON	

Figura 5.4: Preparación de los datos con OpenRefine. Fuente: el autor

5.3. Desarrollo de modelo de predicción de temática de la noticia

Como primer componente de la metodología se desarrollará el modelo de predicción de la categoría del titular de la noticia. En este caso el resultado de aplicar el modelo retornará una etiqueta definida para la temática de la noticia, como por ejemplo: *enviroment, sports, lifestyle, politics, technology, health, science, finance*.

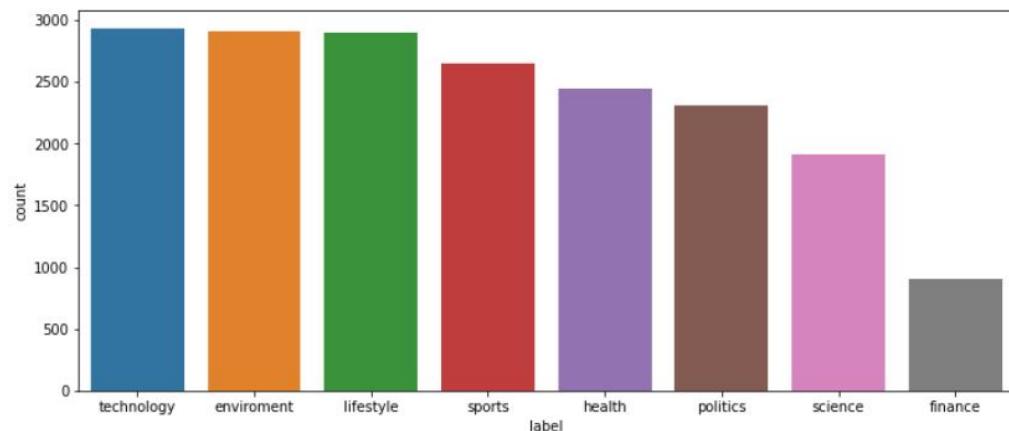
5.3.1. Exploración de los datos

Los datos se cargan directamente desde el repositorio de la investigación³. Contienen dos columnas, la primera es el “headline” es decir, el titular o título de la noticia, y la segunda columna corresponde al “label” o etiqueta que indica la temática general de la noticia.

En la figura 5.5 se presenta el resultado de la exploración inicial de los datos, en donde se observa que las categorías “technology, enviroment y lifestyle” son las más frecuentes, mientras que la categoría “finance” es la menos frecuente.

Posteriormente se hace una revisión de datos nulos y de titulares duplicados. No se encuentran datos nulos, sin embargo si se encuentran algunos datos duplicados los cuales se eliminan para evitar ambigüedades en la etiqueta.

³<https://github.com/jorgecif/CovidDisinformationDetection/blob/main/data/ReutersClasifiedNewsDataset.xlsx>

**Figura 5.5:** Noticias y categorías. Fuente: el autor

5.3.2. Exploración de diferentes de técnicas de inteligencia artificial

Antes de aplicar las diferentes técnicas de aprendizaje automático y aprendizaje profundo, es necesario preparar los datos. En este caso se realiza el mapeo de las categorías y la tokenización de los datos, convirtiendo las palabras a secuencias numéricas que facilitan su análisis.

Modelos de aprendizaje automático

Se realiza la comparación de los clasificadores Random Forest, SVC lineal, Naïve Bayes y Logistic, considerando que estos han sido los más ampliamente usados y con los que se han obtenido mejores resultados en las investigaciones revisadas en el estado del arte (Ver Capítulo 3).

```

1 # Listado de modelos
2 models = [RandomForestClassifier(n_estimators=200, max_depth=3, random_state=0),
            LinearSVC(), MultinomialNB(), LogisticRegression(random_state=0)]
3 # Generacion de los modelos con validacion cruzada
4 CV = 5
5 cv_df = pd.DataFrame(index=range(CV * len(models)))
6 entries = []
7 for model in models:
8     model_name = model.__class__.__name__
9     accuracies = cross_val_score(model, x_train, y_train, scoring='accuracy', cv=CV)
10    for fold_idx, accuracy in enumerate(accuracies):
11        entries.append((model_name, fold_idx, accuracy))
12 cv_df = pd.DataFrame(entries, columns=['model_name', 'fold_idx', 'accuracy'])
13 # Random Forest
14 modelRF = RandomForestClassifier(n_estimators=200, max_depth=3, random_state=0)
15 modelRF.fit(x_train, y_train)

```

```

16 y_pred = modelRF.predict(x_test)
17 # Logistic
18 modelLOG = LogisticRegression(random_state=0)
19 modelLOG.fit(x_train, y_train)
20 y_pred = modelLOG.predict(x_test)
21 # SVC
22 modelSVC = LinearSVC()
23 modelSVC.fit(x_train, y_train)
24 y_pred = modelSVC.predict(x_test)
25 #Naive Bayes
26 modelNB = MultinomialNB()
27 modelNB.fit(x_train, y_train)
28 y_pred = modelNB.predict(x_test)

```

Código 5.2: Detalle generación de modelos de aprendizaje automático. Fuente: el autor

Se plantea una comparación usando validación cruzada con 5 *folds*, para garantizar un entrenamiento con la gran mayoría de posibilidades de combinaciones de los datos (Ver código 5.2. El código completo se incluye en un cuaderno de Jupyter Notebook que se encuentra en el repositorio GitHub de la presente investigación⁴. Con los conjuntos de datos de entrenamiento y test debidamente creados, se realizó el entrenamiento y la ejecución de los diferentes modelos. En la figura 5.6, se presentan los resultados obtenidos para cada uno de los modelos en base a su matriz de confusión.

Teniendo en cuenta los resultados presentados sobre las diferentes matrices de confusión, es notable que el método de Random Forest no funcionó de forma precisa, ya que incluye una gran cantidad de datos clasificados por fuera de la diagonal. Mientras tanto los demás métodos se acercan a una clasificación más precisa de las categorías.

Modelos de aprendizaje profundo

A continuación se presenta el resultado de la comparación de tres clasificadores basados en conjuntos de palabras o “embedding”, el primero una red neuronal *fully connected* como modelo base, el segundo una red neuronal convolucional y el tercero una red neuronal recurrente de tipo *Long Short-term Memory (LSTM)*. Se describirá el detalle de las capas de cada una de las redes planteadas y los resultados obtenidos.

⁴<https://github.com/jorgecif/CovidDisinformationDetection/>

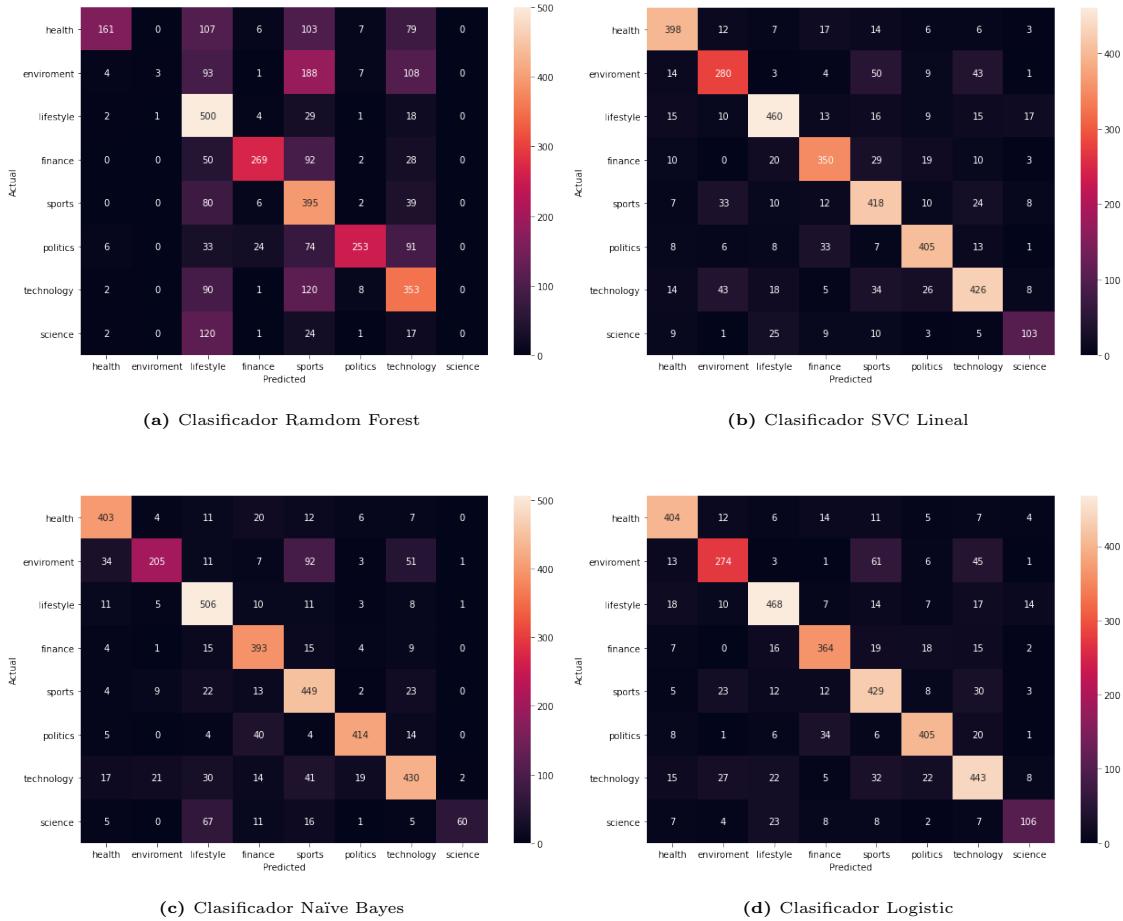


Figura 5.6: Clasificadores técnicas de aprendizaje automático

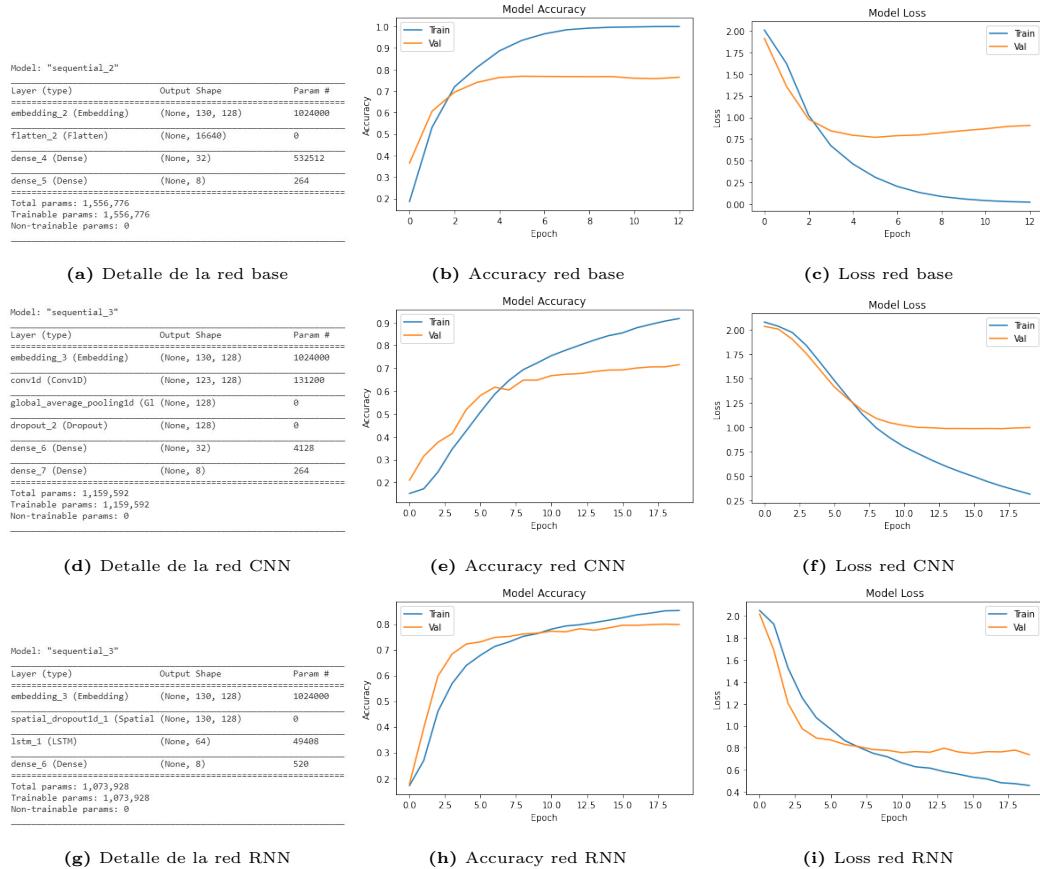
Modelo base fully connected

Con el fin de que el entrenamiento sea lo más rápido posible, la red neuronal base inicial se diseña de forma básica, incluyendo un par de capas densas, una capa *flatten* y la capa inicial de *embeddings*.

El resultado del modelo con datos de validación presenta una precisión cercana al 0.75, mientras que la pérdida alcanza valores cercanos a 0.02. El tiempo de entrenamiento fue bastante rápido, tardando aproximadamente 1 minuto (Ver detalle en las figuras 5.7a, 5.7b y 5.7c).

Modelo red neuronal convolucional (CNN)

Para este modelo, además de las capas densas del modelo base inicial, se agregaron un par de capas adicionales, la primera una ‘‘Conv1D’’ para darle el carácter convolucional y la segunda una capa ‘‘Global-average-pooling1D’’ para suavizar la salida de la capa convolucional.

**Figura 5.7:** Detalle modelos aprendizaje profundo modelo tema

El resultado del modelo con datos de validación presenta una precisión cercana al 0.71, mientras que la pérdida alcanza valores cercanos a 0.3. El tiempo de entrenamiento se incrementó a alrededor de 19 minutos (Ver detalle en las figuras 5.7d, 5.7e y 5.7f).

Modelo red neuronal recurrente (RNN)

Se plantea una red neuronal recurrente que incluye una capa del tipo Long Short-term Memory (LSTM), además de la capa *dense* de salida final y la capa *embedding* de entrada.

El resultado del modelo con datos de validación presenta una precisión cercana al 0.8, mientras que la pérdida alcanza valores cercanos a 0.4. El tiempo de entrenamiento se incrementó a 24 minutos (Ver detalle en las figuras 5.7g, 5.7h y 5.7i).

Comparación de los modelos

A continuación, se comparan los resultados obtenidos de los modelos en términos de precisión de la predicción y tiempos de ejecución.

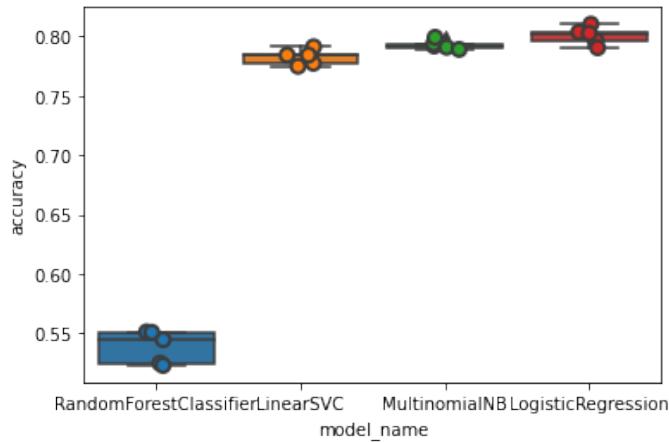


Figura 5.8: Comparación de modelos de aprendizaje automático. Fuente: el autor

Modelos de aprendizaje automático

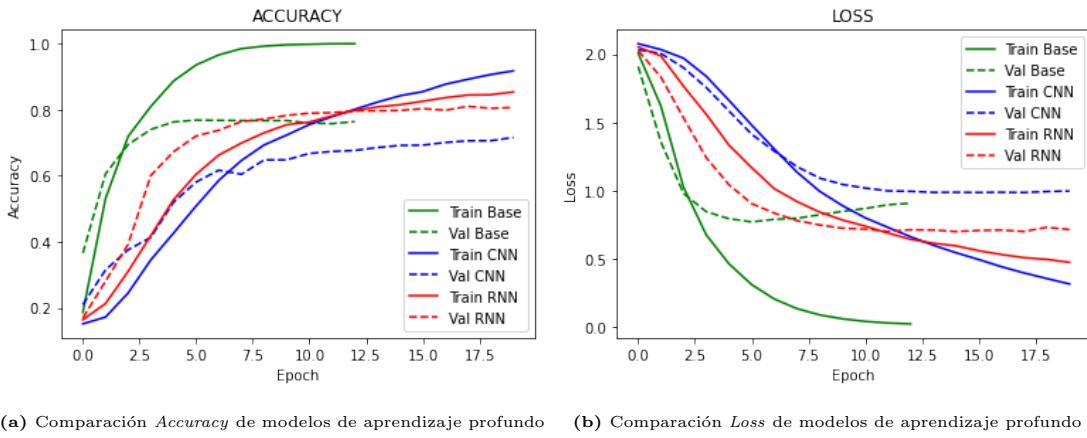
Con el fin de facilitar la comparación, se generó una gráfica en la que se puede evidenciar claramente la precisión alcanzada para cada uno de los modelos (ver figura 5.8). En este caso se confirma que el clasificador basado en *Random Forest* no alcanzó buenos resultados, mientras que el clasificador *Logistic* alcanzó el mejor desempeño, llegando hasta una precisión cercana a 0.8. En cuanto al tiempo de ejecución del entrenamiento de los modelos fue de alrededor de 8 minutos en correr cada uno de ellos y sin ninguna distinción significativa entre uno y otro.

Modelos de aprendizaje profundo

En la figura 5.9 se detalla la comparación de la precisión y el *loss* de los tres modelos generados. De acuerdo con este resultado es claro que el modelo más preciso fue el desarrollado con redes neuronales recurrentes, alcanzando un *accuracy* de alrededor de 0.8, sin embargo su tiempo de entrenamiento fue el más prolongado, llegando hasta los 24 minutos.

Teniendo en cuenta la comparación realizada frente a los métodos de aprendizaje automático y de aprendizaje profundo, se observa que tanto el clasificador *logistic* como el de redes neuronales recurrentes alcanzan una precisión de alrededor de 0.8. Ahora, considerando que el clasificador del tipo *logistic* es un método más liviano computacionalmente de acuerdo a su tiempo de ejecución, se seleccionará este método como parte de la metodología planteada en la presente investigación.

El código completo se incluye en un cuaderno de Jupyter Notebook que se encuentra

**Figura 5.9:** Comparación de modelos aprendizaje profundo

en el repositorio GitHub de la presente investigación⁵.

5.4. Desarrollo del modelo de predicción de la alerta

El modelo de predicción de la alerta es el segundo gran componente de la metodología propuesta. En este apartado se compararán diferentes abordajes para este modelo a través de técnicas de aprendizaje automático y aprendizaje profundo, con el fin seleccionar objetivamente la técnica más apropiada para la detección de desinformación relacionada con la pandemia de COVID-19 dentro del conjunto de datos disponible.

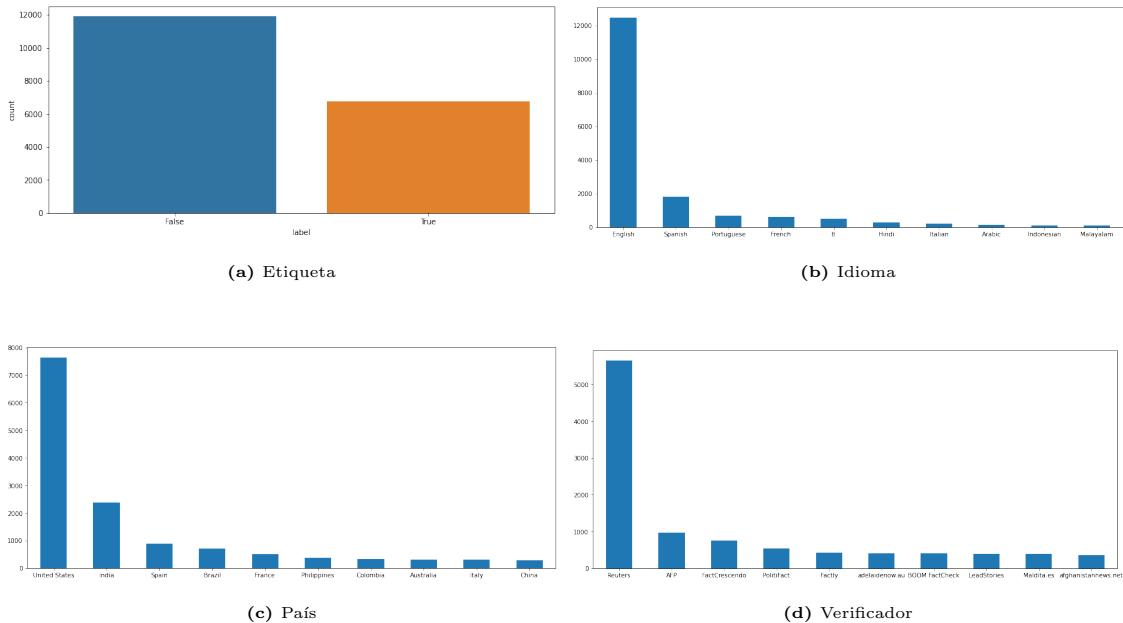
En este caso el resultado de aplicar el modelo retornará una etiqueta binaria sobre los datos que indica si detecta o no una alta probabilidad de desinformación dentro del titular de la noticia. Esta también será la salida final de la metodología que pretender mejorarse a través de la incorporación de características adicionales como la temática dentro del análisis, por lo tanto el resultado será tenido en cuenta para la comparación del modelo final propuesto.

5.4.1. Exploración de los datos

Los datos se cargan directamente desde el repositorio de la investigación y constan de siete columnas: “Text”: contiene el texto del titular de la noticia, “Country”: contiene el país de publicación de la noticia, “Language”: contiene el idioma en el que está escrita la noticia, “URL”: Contiene la URL donde se publicó la noticia, “Date”: Contiene la fecha de

⁵<https://github.com/jorgecif/CovidDisinformationDetection>

<https://github.com/jorgecif/CovidDisinformationDetection/blob/main/data/CovidHeadlinesDataset.xlsx>

**Figura 5.10:** Distribución de las variables

publicación de la noticia, “*Verificado por*”: Contiene la entidad o fact checker que realizó el etiquetado de la noticia y finalmente la columna “*Label*”: Contiene la etiqueta -True- para el caso en el que la información se considera mayoritariamente real, y -False- cuando se detecta riesgo de desinformación dentro del titular de la noticia. En la figura 5.10 se detalla las distribución de algunas de las variables más significativas.

Se observa que el conjunto datos es bastante variado, incluyendo noticias de diversas fuentes, países e idiomas (en su gran mayoría en inglés y en segundo lugar en español). La procedencia de las noticias es mayoritariamente de los EE.UU. pero también se encuentran noticias de la India, España, Brasil, Australia, e incluso algunas de Colombia. En cuanto a los entes verificadores, se observa una gran cantidad de noticias verificadas por Reuters y AFP, así como por otras entidades dedicadas exclusivamente a tal fin ya mencionadas en la tabla 2.2 como Politifact, entre otras.

Debido a que la mayoría de los titulares se encuentran en inglés y acorde a la disponibilidad mayoritaria de los datos dentro del conjunto definido, el enfoque de la metodología se realizará pensando en la detección de desinformación exclusivamente en este idioma. Al realizar este filtro y eliminando los datos nulos y ambigüedades, el conjunto de datos resultante contiene un aproximado de 12000 registros para trabajar.

5.4.2. Exploración de diferentes de técnicas de inteligencia artificial

Para poder aplicar las diferentes técnicas de aprendizaje automático y aprendizaje profundo, es necesario preparar los datos. En este caso se realiza el mapeo de las categorías y la tokenización de los datos, convirtiendo las palabras a secuencias numéricas que facilitan su análisis.

Modelos de aprendizaje automático

Se realiza la comparación de los clasificadores Random Forest, SVC lineal, Naïve Bayes y Logistic, considerando que estos han sido los más ampliamente usados y con los que se han obtenido mejores resultados en las investigaciones revisadas en el estado del arte (Ver Capítulo 3).

```

1 # Listado de modelos
2 models = [RandomForestClassifier(n_estimators=200, max_depth=3, random_state=0),
3            LinearSVC(), MultinomialNB(), LogisticRegression(random_state=0)]
4 CV = 5 # Generacion de los modelos con validacion cruzada
5 cv_df = pd.DataFrame(index=range(CV * len(models)))
6 entries = []
7 for model in models:
8     model_name = model.__class__.__name__
9     accuracies = cross_val_score(model, x_train, y_train, scoring='accuracy', cv=CV)
10    for fold_idx, accuracy in enumerate(accuracies):
11        entries.append((model_name, fold_idx, accuracy))
12 cv_df = pd.DataFrame(entries, columns=['model_name', 'fold_idx', 'accuracy'])
13 # Random Forest
14 modelRF = RandomForestClassifier(n_estimators=200, max_depth=3, random_state=0)
15 modelRF.fit(x_train, y_train)
16 y_pred = modelRF.predict(x_test)
17 # Logistic
18 modelLOG = LogisticRegression(random_state=0)
19 modelLOG.fit(x_train, y_train)
20 y_pred = modelLOG.predict(x_test)
21 # SVC
22 modelSVC = LinearSVC()
23 modelSVC.fit(x_train, y_train)
24 y_pred = modelSVC.predict(x_test)
25 #Naive Bayes
26 modelNB = MultinomialNB()
27 modelNB.fit(x_train, y_train)
28 y_pred = modelNB.predict(x_test)
```

Código 5.3: Detalle generación de modelos de aprendizaje automático. Fuente: el autor

Se plantea una comparación usando validación cruzada con 5 folds, para garantizar un entrenamiento con la gran mayoría de posibilidades de combinaciones de los datos (Ver código 5.3. El código completo se incluye en un cuaderno de Jupyter Notebook que se encuentra en el repositorio GitHub de la presente investigación⁶.

Con los conjuntos de datos de entrenamiento y test debidamente creados, se realizó el entrenamiento de los diferentes modelos. En la figura 5.11, se presentan los resultados obtenidos para cada uno de ellos en base a su matriz de confusión, en donde se encuentra que a diferencia del modelo de predicción del tema de la noticia, para este conjunto de datos el método *Random Forest* funcionó correctamente, alcanzando una precisión de alrededor del 0.8. En este caso el clasificador *Logistic* alcanzó una precisión cercana a 0.75, mientras que el *Naïve Bayes* y el *SVC* no superaron el 0.65 de precisión.

Modelos de aprendizaje profundo

A continuación, se presenta el resultado de la comparación de tres clasificadores basados en conjuntos de palabras o “embedding”, el primero una red neuronal *fully connected* como modelo base, el segundo una red neuronal convolucional y el tercero una red neuronal recurrente de tipo *Long Short-term Memory (LSTM)*. Se describirá el detalle de las capas de cada una de las redes planteadas y los resultados obtenidos.

Modelo base **fully connected**

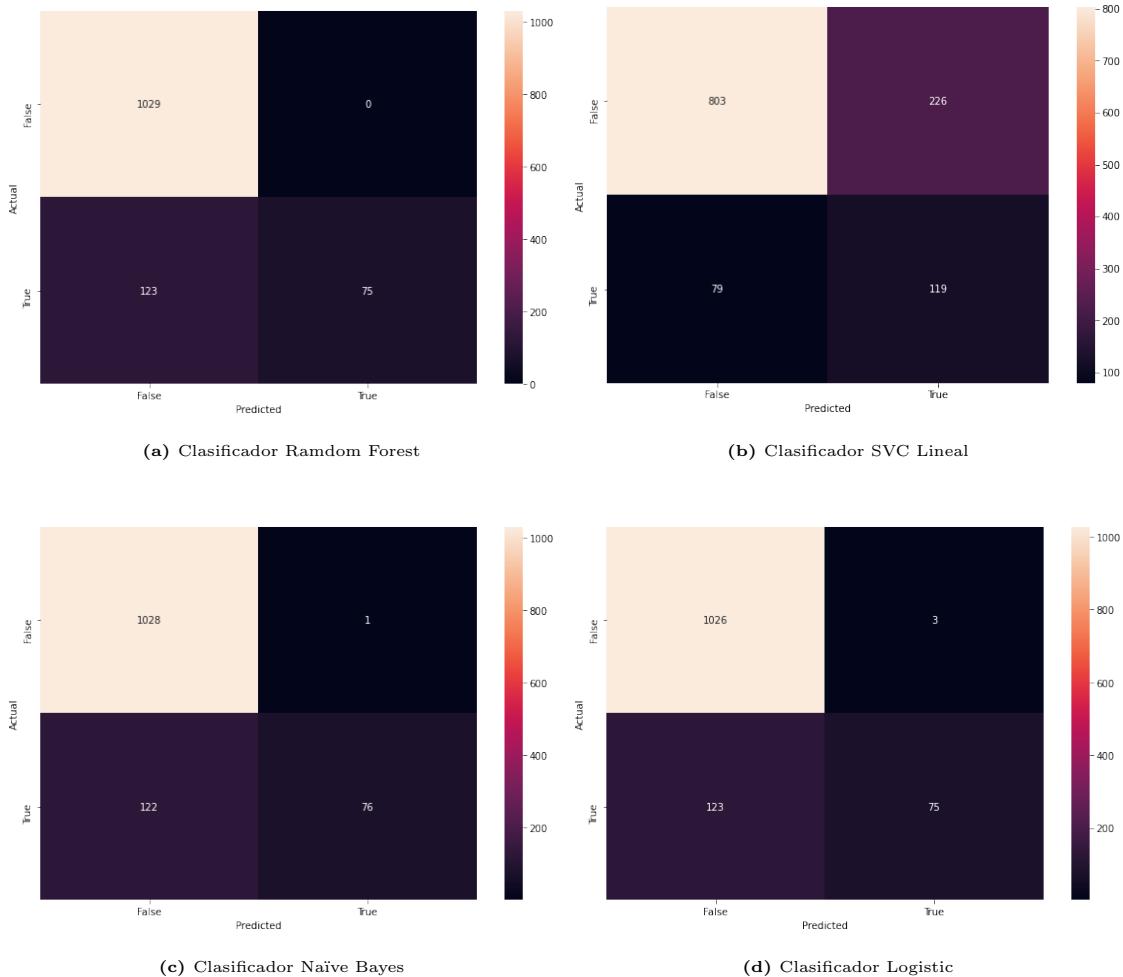
Con el fin de que el entrenamiento sea lo más rápido posible, la red neuronal base inicial se diseña de una forma básica, incluyendo un par de capas densas, una capa *flatten* y la capa inicial de *embeddings*.

El resultado del modelo con datos de validación presenta una precisión cercana al 0.92, mientras que la pérdida alcanza valores cercanos a 0.18. El tiempo de entrenamiento fue bastante rápido, tardando aproximadamente 2 minutos (Ver detalle en las figuras 5.12a, 5.12b y 5.12c).

Modelo red neuronal convolucional (CNN)

Para este modelo, además de las capas densas del modelo base inicial, se agregaron un par de capas adicionales, la primera una “*Conv1D*” para darle el carácter convolucional

⁶<https://github.com/jorgecif/CovidDisinformationDetection/>

**Figura 5.11:** Clasificadores técnicas de aprendizaje automático

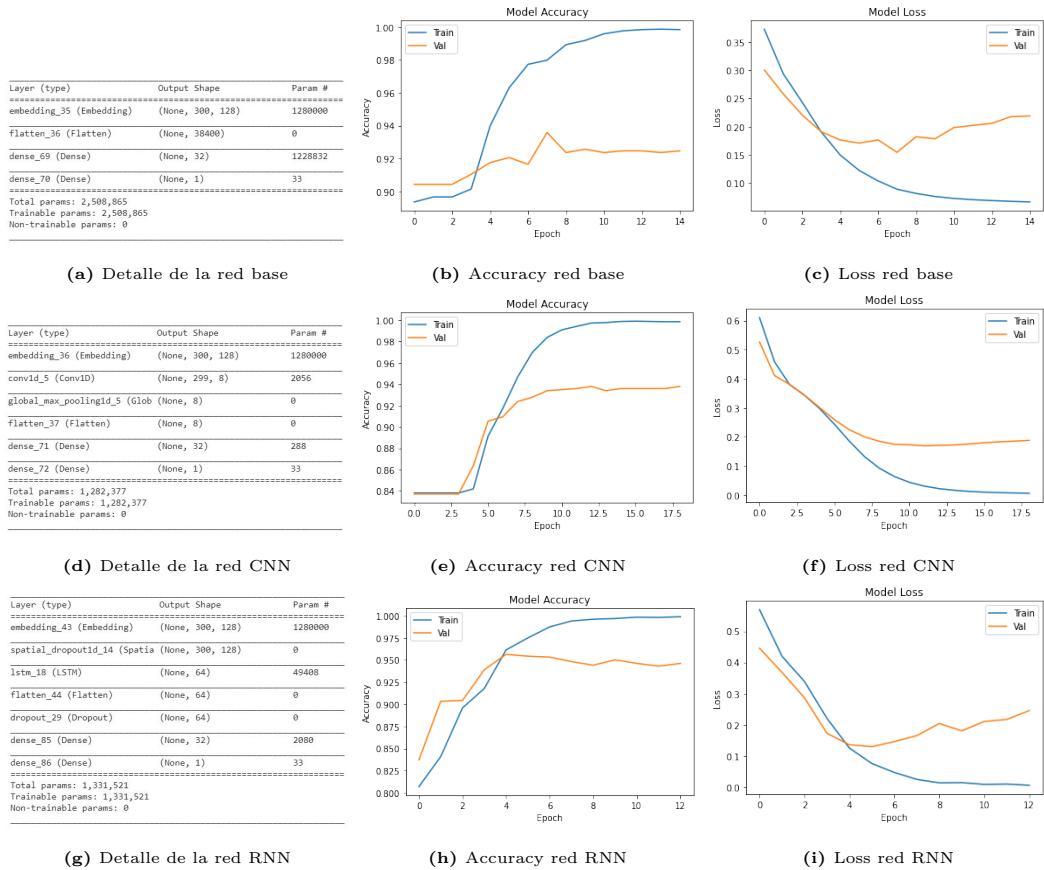
y la segunda, una capa “*Global-average-pooling1D*” para suavizar la salida de la capa convolucional.

El resultado del modelo con datos de validación presenta una precisión cercana al 0.93, mientras que la pérdida alcanza valores cercanos a 0.17. El tiempo de entrenamiento se mantuvo sobre los 2 minutos (Ver detalle en las figuras 5.12d, 5.12e y 5.12f).

Modelo red neuronal recurrente (RNN)

Se plantea una red neuronal recurrente que incluye una capa del tipo *Long Short-term Memory (LSTM)*, además de la capa *dense* de salida final y la capa *embedding* de entrada.

El resultado del modelo con datos de validación presenta una precisión cercana al 0.94, mientras que la pérdida alcanza valores cercanos a 0.17. El tiempo de entrenamiento se incrementó a 7 minutos (Ver detalle en las figuras 5.12g, 5.12h y 5.12i).

**Figura 5.12:** Detalle modelos aprendizaje profundo modelo alerta

Comparación de los modelos

A continuación, se comparan los resultados obtenidos de los modelos en términos de precisión de la predicción y tiempos de ejecución.

Modelos de aprendizaje automático

Con el fin de facilitar la comparación, se generó una gráfica en la que se puede evidenciar claramente la precisión alcanzada para cada uno de los modelos (Ver figura 5.13). En este caso se observa que los clasificadores Random Forest, Naïve Bayes y Logistic alcanzan buenos resultados, llegando hasta un 0.9, mientras que el clasificador SVC es el que obtiene el menor rendimiento con un máximo de precisión de 0.75. En cuanto al tiempo de ejecución de los modelos, se tardó alrededor de 2 minutos en correr cada uno de ellos y sin ninguna distinción significativa entre uno y otro.

Modelos de aprendizaje profundo

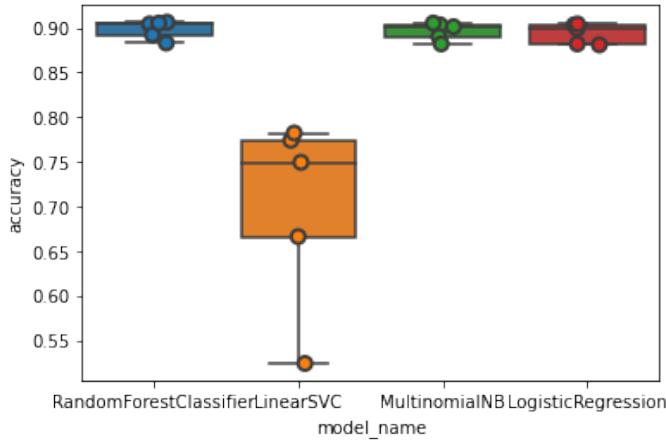


Figura 5.13: Comparación de modelos de aprendizaje automático. Fuente: el autor

En la figura 5.14 se detalla la comparación de la precisión y la pérdida de los tres modelos generados. De acuerdo con este resultado, es claro que el modelo más preciso fue el desarrollado con redes neuronales recurrentes del tipo LSTM, alcanzando un *accuracy* mayor al 0.9, sin embargo su tiempo de entrenamiento fue el más amplio, llegando hasta los 24 minutos.

Teniendo en cuenta la comparación realizada frente a los métodos de aprendizaje automático y de aprendizaje profundo, se observa que en este caso los modelos basados en redes neuronales profundas alcanzan un rendimiento más alto que los basados en técnicas de aprendizaje automático. En este caso el método que alcanza el mayor rendimiento es el de redes neuronales recurrentes basadas en *Long Short-term Memory (LSTM)* llegando a un 0.94 de precisión. Por esta razón, esta técnica será la seleccionada para la implementación de este componente dentro de la metodología propuesta.

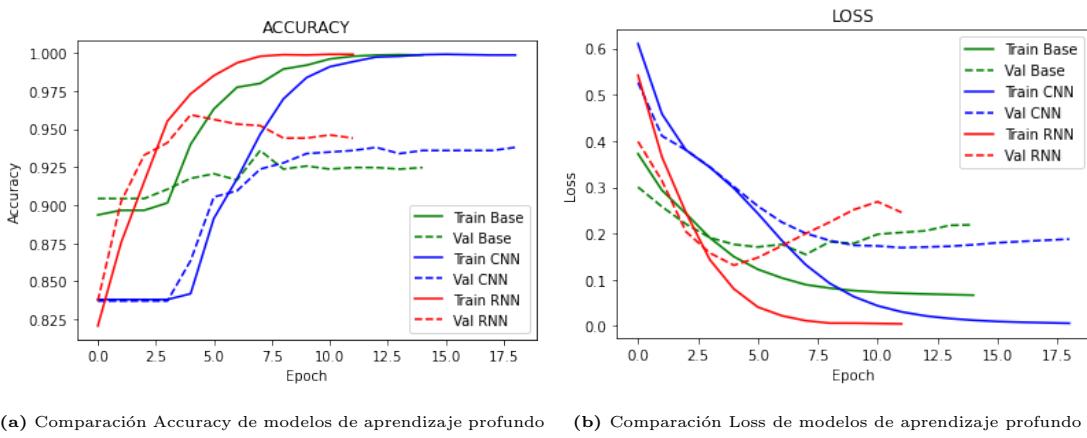
El código completo se incluye en un cuaderno de Jupyter Notebook que se encuentra en el repositorio GitHub de la presente investigación⁷.

5.5. Desarrollo del modelo de extracción del subtema

Si bien, para la clasificación de la noticia en temas se contaba con una etiqueta de entrenamiento en los datos relacionada con su temática, en este caso, no se cuenta con una etiqueta para el subtema. Por esta razón este tipo de problema se abordará a través de una técnica de aprendizaje no supervisado.

Una de las técnicas ampliamente usadas para la detección de temas de forma no

⁷<https://github.com/jorgecif/CovidDisinformationDetection>

**Figura 5.14:** Comparación de modelos aprendizaje profundo

supervisada es la *Asignación Latente de Dirichlet (ALD)* o *Latent Dirichlet Allocation (LDA)*, la cual permite agrupar elementos de un conjunto de datos a partir de parte de los datos que son semejantes. Ejemplos de este tipo de técnicas se pueden encontrar en investigaciones directamente relacionadas con la extracción de temáticas de noticias, como por ejemplo en los trabajos de [Xu et al.(2019)Xu, Meng, Chen, Qiu, Wang & Yao] y [Xu et al.(2020)Xu, Wang, Wang & Yang].

La técnica LDA parte de la suposición de que un documento está compuesto por una mezcla de varios temas y cada uno de estos temas se compone de una serie de elementos, en este caso palabras que son comunes y que tienen una mayor o menor probabilidad de aparecer en uno u otro tema. Un ejemplo de dos grupos en el contexto de la pandemia de COVID-19, podría ser: la *subtemática vacuna* cuyas palabras relacionadas podrían ser: vacuna, dosis, placebo, reacción. Mientras que si el otro grupo es la categoría relacionada con la *subtemática rebrote*, las palabras asociadas podrían ser: casos, cuarentena, medidas, pico.

En este sentido se plantea la generación de un modelo de extracción de la subtemática a partir del conjunto de datos relacionado con la pandemia de COVID-19, el cual permitirá extraer una mayor cantidad de información de los titulares de las noticias.

Para modelar esta técnica se utilizó *Gensim*⁸, una librería de código abierto para el modelado de temas no supervisados. A partir de la documentación de la librería, los principales parámetros a tener en cuenta para inicializar el modelo LDA son:

⁸<https://radimrehurek.com/gensim/>

- *Ntopics*: Número de temas
- *Eta*: Distribución del número de palabras por tema
- *Alpha*: Número de temas por documento revisado
- *Dictionary*: Diccionario de palabras

A continuación se incluye un extracto del código para implementar este modelo:

```

1 # Convierto plural a singular
2 stemmer = SnowballStemmer("english")
3 original_words = ['caresses', 'flies', 'dies', 'mules', 'denied', 'died', 'agreed',
4                   'owned',
5                   'humbled', 'sized', 'meeting', 'stating', 'siezing', 'itemization',
6                   'sensational',
7                   'traditional', 'reference', 'colonizer', 'plotted']
8 singles = [stemmer.stem(plural) for plural in original_words]
9 pd.DataFrame(data={'original word':original_words, 'stemmed':singles })
# Lematizacion
10 def lemmatize_stemming(text):
11     return stemmer.stem(WordNetLemmatizer().lemmatize(text, pos='v'))
# Tokenize and lemmatize
12 def preprocess(text):
13     result=[]
14     for token in gensim.utils.simple_preprocess(text) :
15         if token not in gensim.parsing.preprocessing.STOPWORDS and len(token) > 3:
16             result.append(lemmatize_stemming(token))
17     return result
18 # Creacion del modelo
19 num_topics=5
20 lda_model = gensim.models.LdaMulticore(bow_corpus,
21                                         num_topics = num_topics,
22                                         id2word = dictionary,
23                                         passes = 10,
24                                         workers = 2,
25                                         alpha=[0.01]*num_topics,
26                                         eta=[0.01]*len(dictionary.keys()))
27 # Imprimo los temas creados
28 for idx, topic in lda_model.print_topics(-1):
29     print("Topic: {} \nWords: {}".format(idx, topic ))
30     print("\n")

```

Código 5.4: Detalle de código de modelo LDA para la extracción del subtema de la noticia. Fuente: el autor

El código 5.4 inicia con el preprocesamiento de los datos convirtiendo las palabras en plural al singular, realizando la lematización, excluyendo las *stopwords* y posteriormente definiendo y ejecutando el modelo.

```
[[['test', 'health', 'hospit', 'clinic', 'australia'],
  ['mask', 'kill', 'wear', 'chines', 'hand'],
  ['infect', 'case', 'health', 'report', 'pictur'],
  ['case', 'australia', 'health', 'pictur', 'australian'],
  ['trump', 'presid', 'state', 'death', 'unit'],
  ['post', 'facebook', 'novel', 'show', 'share'],
  ['toilet', 'paper', 'australia', 'health', 'custom']]
```

Figura 5.15: Resultado del modelo LDA para 7 subtemas y 5 palabras. Fuente: el autor

El resultado del modelo permite crear los subtemas con sus respectivos conjuntos de palabras, cada una con su probabilidad de aparición. En la figura 5.15 se presenta el resultado final para un modelo con 7 subtemas resultando en la agrupación de palabras para cada uno de los 7 grupos.

Analizando los grupos resultantes:

- La primera temática podría estar relacionada con las pruebas y la hospitalización pues contiene las palabras “*test*”, “*health*”, “*hospit*”, “*clinic*”.
- La segunda categoría podría estar relacionada con los métodos preventivos pues contiene las palabras “*mask*”, “*wear*”, “*hand*”.
- El tercer grupo al contener las palabras “*infect*”, “*case*”, “*report*”, podría estar relacionado con el reporte de casos de infectados que día a día realizan todos los países.
- La cuarta categoría puede corresponder a casos específicamente relacionados con Australia, ya que incluye las palabras “*case*”, “*australia*”, “*australian*”.
- En cuanto al quinto grupo claramente se trataría de una categoría relacionada con los EE.UU. y su presidente, pues contiene palabras como “*unit*”, “*state*”, “*Trump*”, “*presi*”.
- La sexta categoría podría estar relacionada con publicaciones en redes sociales, pues tiene palabras como “*post*”, “*facebook*”, “*share*”.
- Finalmente, el último grupo por contener palabras como “*toilet*” y “*paper*”, podría estar relacionado con algunos casos sonados relacionados con el coronavirus, como el de la escasez de papel higiénico causado por el pánico generado muy al inicio de la pandemia.

De esta manera y luego del análisis planteado, a continuación se resumen los diferentes subtemas encontrados y sus palabras relacionadas:

- Pruebas de detección: test, health, hospit, clinic, australia
- Cuidados y recomendaciones: mask, kill, wear, chines, hand
- Reporte de casos: infect, case, health, report, pictur
- Australia: case, australia, health, pictur, australian
- EE.UU.: trump, presid, state, death, 'unit
- Redes sociales: post, facebook, novel, show, share
- Casos sonados: toilet, paper, australia, health, custom

Como se puede apreciar, la información de las subtemáticas podría ser muy relevante a la hora de identificar un titular de una noticia como potencialmente portadora de desinformación y es por esta razón que se incluirá dentro de la metodología propuesta.

El código completo se incluye en un cuaderno de Jupyter Notebook que se encuentra en el repositorio GitHub de la presente investigación⁹.

5.6. Generación de la metodología unificada

Una vez realizada la comparación de las diferentes técnicas de inteligencia artificial, ya se tiene una idea de cuáles lograron los mejores rendimientos al aplicarlas a los conjuntos de datos recolectados para la investigación. En este caso, para el componente de predicción de la temática de la noticia se ha seleccionado un clasificador tipo *Logistic*, mientras que para el modelo de predicción de la alerta de posible desinformación se seleccionó un modelo basado en redes neuronales recurrentes de tipo *Long Short-term Memory (LSTM)*. En cuanto a la predicción de la subtemática, de acuerdo a lo revisado en el numeral inmediatamente anterior, se realizará por medio del modelo Latent Dirchlet Allocation (LDA).

Ahora, el desafío tiene que ver con unificar estos componentes en una única metodología que le entregue al usuario una sola respuesta. Para este propósito se requiere una

⁹<https://github.com/jorgecif/CovidDisinformationDetection>

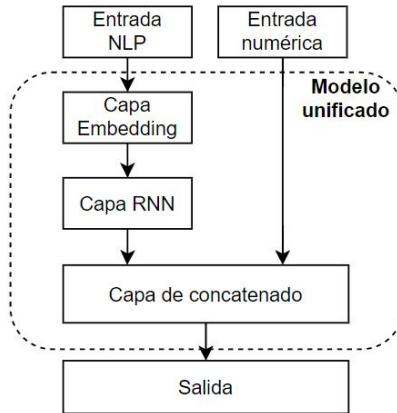


Figura 5.16: Modelo con múltiples entradas y una salida. Fuente: el autor

aproximación que permita combinar múltiples entradas y resultar en una única sola salida para técnicas de aprendizaje profundo.

En este sentido en trabajos como los de [Liu et al.(2019)Liu, Gherbi, Li & Cheriet] y [Turkoglu et al.(2019)Turkoglu, Hanbay & Sengur] se propone una opción novedosa de redes con múltiples entradas que se puede ajustar a los requerimientos de la metodología. El planteamiento corresponde a que las redes normalmente estudiadas incluyen solamente un vector de entrenamiento y una única salida con la predicción, sin embargo en escenarios más complejos, como el planteado en la presente investigación, los conjuntos de datos suelen tener diferentes tipos de información, combinando textos con datos numéricos y generando la necesidad de incluir estas características diferentes dentro de una misma red neuronal.

Para el escenario específico de la investigación, el principal problema sin duda está relacionado con clasificación de texto, sin embargo, teniendo en cuenta la necesidad de incorporar información adicional, se considerará una segunda entrada del modelo para la temática y subtemática de la noticia; información adicional que se puede considerar como un vector de metadatos asociados.

De acuerdo a lo propuesto en [Turkoglu et al.(2019)Turkoglu, Hanbay & Sengur] y [Liu et al.(2019)Liu, Gherbi, Li & Cheriet], existen dos acercamiento para abordar el problema de las múltiples entradas. El primero es sencillamente concatenar estos metadatos a los textos para que sean considerados en los *embeddings* y bolsas de palabras generadas en el preprocesamiento. Sin embargo este tipo de abordaje puede ser simplista y no aprovecha al máximo el potencial de la información adicional, ya que las nuevas palabras pueden diluirse frente al cuerpo del titular de la noticia. El segundo acercamiento, considera múltiples vectores de entrenamiento para el modelo, para así embeber por completo

los metadatos dentro del modelo generado y entrenado. Para este caso, la distribución de las entradas y salidas del modelo se resume en la figura 5.16).

5.6.1. Componentes principales

Con base en lo anterior, a continuación se definen los componentes principales de la metodología y su planteamiento unificado a partir del cual iniciará su construcción y desarrollo (Ver figura 4.4):

- El modelo de predicción de la *temática* de la noticia
- El modelo de extracción de *subtemas* de la noticia
- El modelo de predicción de la *alerta* de noticia potencialmente portadora de desinformación.

5.6.2. Planteamiento unificado de la metodología

En este apartado se detalla la arquitectura final de la metodología, el cual se define en la figura 5.17 e incluye los componentes principales (modelos a desarrollar), la forma como están interconectados, y las entradas y salidas del proceso. Se incluyen las técnicas específicas de inteligencia artificial a utilizar de acuerdo a los resultados obtenidos en la exploración realizada en los capítulos anteriores.

- *Componente de predicción de la temática:* Modelo basado en un clasificador del tipo *logistic*.
- *Componente de extracción de subtemas:* Modelo basado en un clasificador no supervisado del tipo *Latent Dirichlet Allocation (LDA)*.
- *Componente de predicción de la alerta:* Modelo basado en una red neuronal recurrente del tipo *LSTM* con múltiples entradas.

5.6.3. Definición y desarrollo del modelo

A partir de la figura 5.17 en donde se define la arquitectura final de la metodología, se realizó su implementación en un cuaderno de Jupyter Notebook con el fin de obtener un modelo unificado y posteriormente comparar sus resultados con los modelos base explorados en el Capítulo 4. En este numeral se detalla el paso a paso del desarrollo de

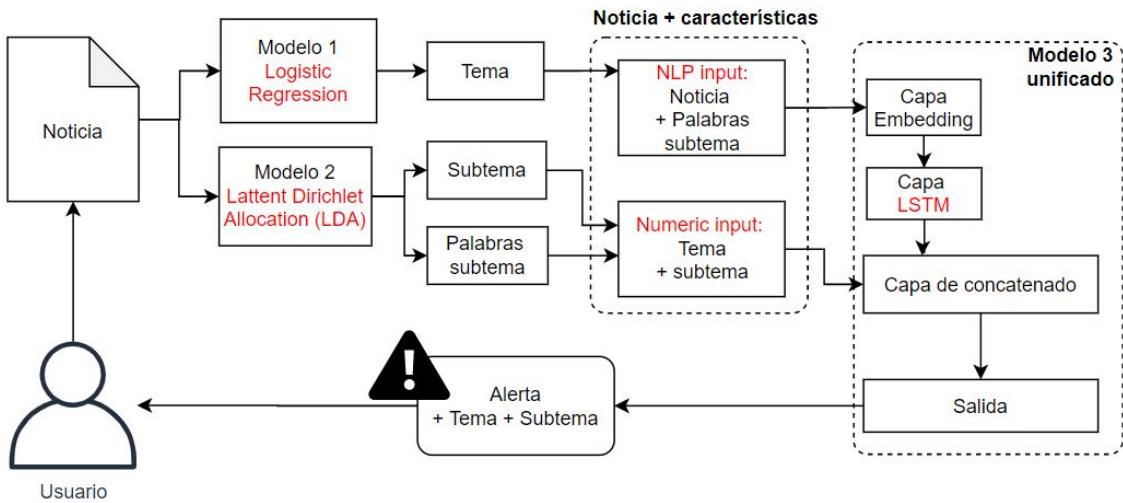


Figura 5.17: Planteamiento final de la metodología unificada. Fuente: el autor

la metodología referenciando los apartes de código correspondientes que se consideran de mayor relevancia, sin embargo el código completo podrá ser consultado en el repositorio de la presente investigación¹⁰.

Paso 1: Carga de conjunto de datos

Se carga el conjunto de datos directamente desde el repositorio GitHub y se realiza un preprocesamiento inicial para eliminar registros duplicados, datos nulos e inconsistencias en los datos. Este es el mismo proceso que se ha realizado en el desarrollo de los modelos de prueba explorados en los capítulos anteriores. A continuación se detalla un extracto del código implementado.

```

1 #Carga de datos
2 url_datos="https://github.com/jorgecif/CovidDisinformationDetection/blob/main/data/
  CovidHeadlinesDataset.xlsx?raw=true"
3 datos = pd.read_excel(url_datos)
4 #Elimino duplicados de todas las columnas
5 datos2 = datos.drop_duplicates()
6 # Elimino ambiguedades
7 datos3=datos2.drop_duplicates(['Text'], keep='first') # Elimino ambiguedades
  
```

Código 5.5: Detalle de código de carga de datos y preprocesamiento. Fuente: el autor

Paso 2: Carga de modelos

En este paso el objetivo es cargar los modelos probados que alcanzaron los mejores resultados en las comparaciones realizadas. Los modelos previamente han sido guardados en formato .pkl o .h5 (En la figura 5.18 se observa la respuesta de un *print* con el detalle de

¹⁰<https://github.com/jorgecif/CovidDisinformationDetection/>

```

CountVectorizer(analyzer='word', binary=False, decode_error='strict',
               dtype=<class 'numpy.int64'>, encoding='utf-8', input='content',
               lowercase=True, max_df=1.0, max_features=None, min_df=1,
               ngram_range=(1, 1), preprocessor=None, stop_words=None,
               strip_accents=None, token_pattern='(\\u)\\\\b\\\\w\\\\w+\\\\b',
               tokenizer=None, vocabulary=None)
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                    intercept_scaling=1, l1_ratio=None, max_iter=100,
                    multi_class='auto', n_jobs=None, penalty='l2',
                    random_state=0, solver='lbfgs', tol=0.0001, verbose=0,
                    warm_start=False)

```

Figura 5.18: Planteamiento final de la metodología unificada. Fuente: el autor

los modelos luego de que se han cargado. También se cargan los correspondientes *tokenizadores* o *vectorizadores* de los datos con el fin de poder aplicarlos en los nuevos textos en los próximos pasos. En el siguiente extracto de código se puede observar un mayor detalle sobre este paso.

```

1 # Cargo modelos desde archivos en repositorio github
2 mLink = 'https://github.com/jorgecif/CovidMisinformationDetection/blob/main/
          Modelos_clasificacion/vectorizer.pkl?raw=true'
3 mfile = BytesIO(requests.get(mLink).content)
4 vectorizer = joblib.load(mfile)
5 mLink = 'https://github.com/jorgecif/CovidMisinformationDetection/blob/main/
          Modelos_clasificacion/LOG_model.pkl?raw=true'
6 mfile = BytesIO(requests.get(mLink).content)
7 log_mod = joblib.load(mfile)

```

Código 5.6: Detalle de carga de modelos previamente entrenados. Fuente: el autor

Paso 3: Aplicación de modelo para clasificar la temática (Logistic Regression)

Una vez cargado el modelo, procedemos a aplicarlo y guardar el resultado para cada uno de los titulares en una nueva columna del *dataframe*. Para esto se implementó una función y un bucle que recorre y la aplica a lo largo de cada uno de los registros del *dataframe*. El detalle se incluye en el código a continuación.

```

1 # Funcion para aplicar modelo
2 def aplicar_modelo(datos_revisar, tokenizador, modelo):
3     tokenizador=tokenizador
4     modelo=modelo
5     datos_revisar=datos_revisar
6     len_datos_revisar=len(datos_revisar)
7     list_result=[]
8     for i in range(0,len_datos_revisar):
9         clear_output(wait=True)
10        linea_revisar=datos_revisar["Text"][i]
11        linea_revisar_token=tokenizador.transform([linea_revisar])
12        resultado=modelo.predict(linea_revisar_token)
13        list_result.append(resultado[0])

```

```

14     print("Progreso ", np.round(i/len_datos_revisar*100,2),"%")
15
16 # Calculo predicciones de modelos con funcion
17 start = time.clock()
18 pred_model_clasifica=aplicar_modelo(datos_trabajo, vectorizer, log_mod)
19 end = time.clock()
20 print("Tiempo de entrenamiento: ", (end-start)/60, " minutos")
21 # Creo dataframe con columna adicional de prediccion
22 datos_trabajo_pred=datos_trabajo
23 datos_trabajo_pred["pred_clasifica"]=pred_model_clasifica

```

Código 5.7: Detalle de aplicación de modelo de clasificación de temática. Fuente: el autor

Considerando que este tipo de bucles puede tardar un tiempo prolongado en recorrer uno a uno los elementos del *dataframe*, se implementó un código que permite verificar el progreso del bucle en forma de porcentaje y al final se calcula el tiempo total de ejecución de este paso. Para el caso particular, los tiempos estuvieron alrededor de los 3 minutos.

Paso 4: Aplicación de modelo para extraer la subtemática (LDA)

En este paso se aplica el modelo de extracción de la subtemática sobre la totalidad de los titulares incluidos en el *dataframe*. Los pasos iniciales son los mismos planteados en el código presentado en el Capítulo 4 (Ver código 5.8) complementado al final con la creación de dos columnas adicionales dentro del *dataframe*, la primera para guardar específicamente el subtema extraído, y la segunda columna para guardar las palabras claves asociadas a dicho subtema. En las líneas de código a continuación se detalla este proceso.

```

1 # Creo funcion para extraer subtematicas
2 def topics_lda(documento):
3     unseen_document=documento
4     bow_vector = dictionary.doc2bow(preprocess(unseen_document))      # Preprocesamiento
5     prediction_lda=lda_model[bow_vector]      # Aplico modelo
6     probs=[]
7     for i in range(0, len(prediction_lda)):
8         probs.append(prediction_lda[i][1])
9     max_probs=max(probs)
10    for i in range(0,len(prediction_lda)):
11        if max_probs==prediction_lda[i][1]:
12            position=i
13            break
14    return position
15 datos_revisar=datos_trabajo["Text"] # Aplico modelo LDA a dataframe
16 list_result_id=[]
17 list_result_words=[]

```

```

18 for i in range(0,len(datos_revisar)):
19     clear_output(wait=True)
20     id_predict=topics_lda(datos_revisar[i])
21     list_result_id.append(id_predict)
22     list_result_words.append(str(topics[id_predict]))
23 # Creo dataframe con columna adicional de predicción y palabras
24 datos_trabajo_pred["pred_topics_id"]=list_result_id
25 datos_trabajo_pred["pred_topics_words"]=list_result_words
26 # Concateno columnas de texto
27 datos_analizar=datos_trabajo_pred
28 datos_analizar["text_topics"] = datos_analizar["pred_topics_words"]+datos_analizar["Text"]
29 # Concateno columnas numéricas de metadatos
30 meta_list=[]
31 for i in range(0, len(datos_analizar)):
32     lista=np.asarray(datos_analizar["pred_clasifica"][i]).astype(np.int32),np.asarray
33         (datos_analizar["pred_topics_id"][i]).astype(np.int32)
34     lista=np.asarray(lista).astype(np.int32)
35     meta_list.append(lista)
36 # Creo en dataframe columna adicional
37 datos_analizar["metadatos"]=meta_list

```

Código 5.8: Detalle de aplicación del modelo de extracción de la subtemática. Fuente: el autor

En el código 5.8 se incluye la creación de una función para aplicar el modelo sobre todos los registros del conjunto de datos, la cual se implementa a través de un bucle con su respectivo indicador de progreso para verificar su avance. Para el caso, el tiempo promedio de aplicación del bucle fue de alrededor de 1 minuto. El resultado final de este paso será el *dataframe* con la columna “*Text-topics*” que contiene concatenado el texto del titular de la noticia con las palabras de la subtemática asociada, y la columna ”*metadatos*” que contendrá dos elementos, el primero correspondiente al tema predicho con anteriormente en el paso 3, y el segundo corresponderá a la subtemática seleccionada por el modelo LDA. Estas dos columnas serán las dos entradas del modelo *LSTM* del próximo paso.

Paso 5: Planteamiento de la red LSTM con múltiples entradas

Inicialmente se realiza la extracción de los datos del conjunto de datos que serán incluidos en el modelo, en este caso las columnas “*Text-topics*”(*input NLP*), “*metadatos*”(*input numérico*) y “*label*”(predicción). Con estas dos columnas extraídas el próximo paso será realizar el preprocesamiento con el fin de adaptar la información a una forma que pueda ser “*entendida*” por un modelo de redes neuronales profundas. Posteriormente se realiza el *split* o corte de los datos, en este caso sobre los dos conjuntos de entradas, resultando

los vectores de entrenamiento y prueba: *xtrain1*, *xtrain2*, *xtest1*, *xtest2* y sus respectivas salidas de etiquetas en las variables *ytrain* y *ytest*.

```

1 # Extraigo datos de dataframe
2 corpus_trabajo = datos_analizar["text_topics"] # Columna de text NLP - Input 1
3 meta = datos_analizar["metadatos"] # Columna de metadatos numerica - Input 2
4 results_trabajo = datos_analizar["label"].map(category_dict) # Prediccion
5 # Tokenizacion
6 corpus_trabajo = datos_analizar["text_topics"]
7 sequences = tokenizer.texts_to_sequences(corpus_trabajo.values)
8 X = pad_sequences(sequences, maxlen=max_len)
9 meta=meta.values.tolist()
10 meta_arr = np.array(meta)
11 # Train - Test split
12 x_train1,x_test1, y_train,y_test = train_test_split(X, results_trabajo, test_size
13 =0.2, random_state=88 ) # Datos NLP (Input 1)
13 x_train2,x_test2, y_train2,y_test2 = train_test_split(meta_arr, results_trabajo,
14 test_size=0.2, random_state=88 ) # Datos numericos - metadatos (Input 2)
14 # Definicion de la red
15 nlp_input = Input(shape=(seq_length,), name='nlp_input')
16 meta_input = Input(shape=(2,), name='meta_input')
17 emb = Embedding(output_dim=emb_dim, input_dim=embedding_size, input_length=
18 seq_length)(nlp_input)
18 nlp_out = (LSTM(64, dropout=0.7, recurrent_dropout=0.7, kernel_regularizer=
19 regularizers.l2(0.01)))(emb)
19 x = concatenate([nlp_out, meta_input])
20 x = Flatten()(x)
21 x = Dropout(0.3)(x)
22 x = Dense(32, activation='relu')(x)
23 x = Dense(1, activation='sigmoid')(x)
24 model = Model(inputs=[nlp_input , meta_input], outputs=[x])
25 # Compilacion
26 model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['acc'])

```

Código 5.9: Detalle de planteamiento de la red LSTM con múltiples entradas. Fuente: el autor

En cuanto a la definición de la red LSTM con múltiples entradas, se realiza incluyendo dos entradas de capa, una llamada *nlp-input* para *xtrain1* y la otra denominada *meta-input* para *xtrain2*. Las dos entradas se combinan dentro de la arquitectura de la red por medio de una capa de concatenado que incluye la librería Keras¹¹ para terminar con capas densas que generan la predicción unificada de detección de desinformación. El resultado del modelo implementado se detalla en la figura 5.19).

¹¹Keras es una biblioteca de Redes Neuronales de Código Abierto escrita en Python.

Model: "functional_35"			
Layer (type)	Output Shape	Param #	Connected to
nlp_input (InputLayer)	[(None, 300)]	0	
embedding_24 (Embedding)	(None, 300, 128)	1280000	nlp_input[0][0]
lstm_24 (LSTM)	(None, 64)	49408	embedding_24[0][0]
meta_input (InputLayer)	[(None, 2)]	0	
concatenate_19 (Concatenate)	(None, 66)	0	lstm_24[0][0] meta_input[0][0]
flatten_22 (Flatten)	(None, 66)	0	concatenate_19[0][0]
dropout_21 (Dropout)	(None, 66)	0	flatten_22[0][0]
dense_46 (Dense)	(None, 32)	2144	dropout_21[0][0]
dense_47 (Dense)	(None, 1)	33	dense_46[0][0]

Total params: 1,331,585
Trainable params: 1,331,585
Non-trainable params: 0

Figura 5.19: Resumen de la red LSTM con múltiples entradas planteada. Fuente: el autor

Paso 6: Entrenamiento de la red

Por tratarse de una arquitectura de múltiples entradas, el entrenamiento deberá tener esto en cuenta. En este caso se incorporan las dos entradas dentro del comando *model.fit*. En el siguiente apartado de código se puede observar esta particularidad.

```

1 # Entrenamiento
2 start=time.clock()
3 historyFinal1=model.fit({'nlp_input': x_train1, 'meta_input': x_train2}, y_train,
4                         epochs=epochs, batch_size=batch_size, validation_split=0.2, callbacks=[EarlyStopping(monitor='val_loss', patience=7, min_delta=0.0001)])
5 end = time.clock()
6 tiempo_base=(end-start)/60
7 print("Tiempo de entrenamiento: ", tiempo_base, " minutos")

```

Código 5.10: Entrenamiento de la red LSTM con múltiples entradas planteada. Fuente: el autor

Para el entrenamiento se ejecutan 30 *epochs*, sin embargo se incluyó un *callback* para ejecutar un *earlystopping* o parada anticipada en caso de que se detecte que la red no puede seguir mejorando. Para el caso de las pruebas realizadas, de los 30 *epochs* se ejecutaron la mayoría de veces alrededor entre 18 y 20 *epochs*. En este caso por la complejidad de la red, los tiempos de entrenamiento se incrementan a alrededor de los 15 minutos.

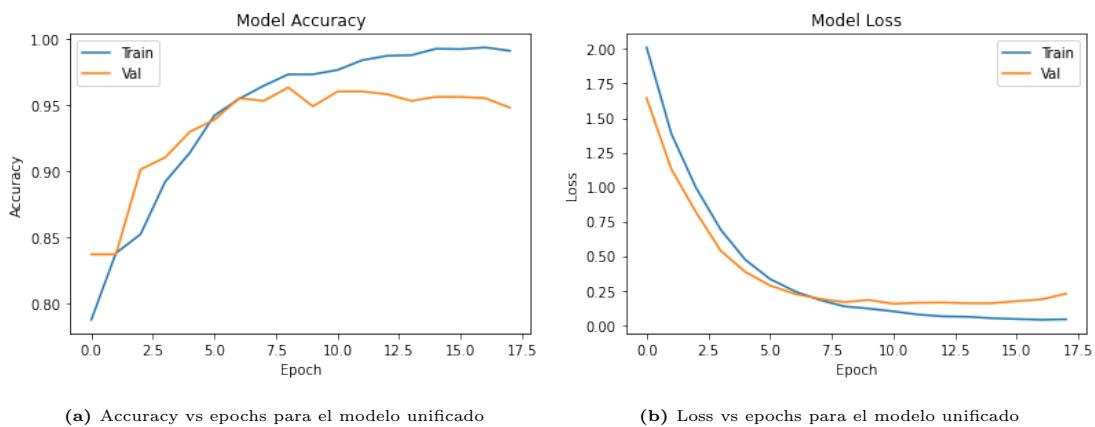


Figura 5.20: Detalle del rendimiento del modelo unificado en términos de los parámetros de *accuracy* y *loss* alcanzados

Paso 7: Resultados

Los resultados del rendimiento del modelo en cuanto a su *accuracy* y *loss*, alcanza valores cercanos a un 0.95 y 0.1, respectivamente. En la figura 5.20 se incluyen las gráficas de estas variables *accuracy* y *loss* a lo largo de los *epochs* de entrenamiento implementados.

Capítulo 6

Validación de la metodología

Este capítulo tiene como objetivo validar la metodología propuesta a partir de la comparación de los resultados y de la prueba del modelo con datos desconocidos o reservados para prueba. Se incluyen los siguientes apartados: comparación de resultados de la metodología con modelo base, prueba del modelo y predicciones, implementación del modelo y pruebas de validación final.

6.1. Comparación de resultados de la metodología con modelo base

Los resultados obtenidos a través de la metodología propuesta se contrastaron con los resultados del mejor de los modelos resultante de la exploración de técnicas para la generación de la alerta realizada en el numeral 5.4 (Ver figura 5.14). Esto quiere decir que en este caso el modelo base de la comparación será un modelo LSTM de única entrada que analiza solamente el texto del titular de la noticia; frente a un modelo LSTM de múltiples entradas que, en una de sus entradas incluye el texto del titular de la noticia junto a las palabras detectadas por el clasificador LDA, y en su segunda entrada incluye los metadatos de la clasificación de la temática y subtemática de la noticia.

En la figura 6.1, se presentan los resultados de este comparativo, alcanzando una mejora en el rendimiento del modelo de múltiples entrada propuesto en términos de *accuracy* y *loss*, obteniendo niveles superiores a 0.95 e inferiores a 0.25 respectivamente, sin embargo el tiempo de entrenamiento del modelo de múltiples entradas es superior al de una única entrada, por lo que habría que tener en cuenta esta característica para la generación de nuevos modelos entrenados.

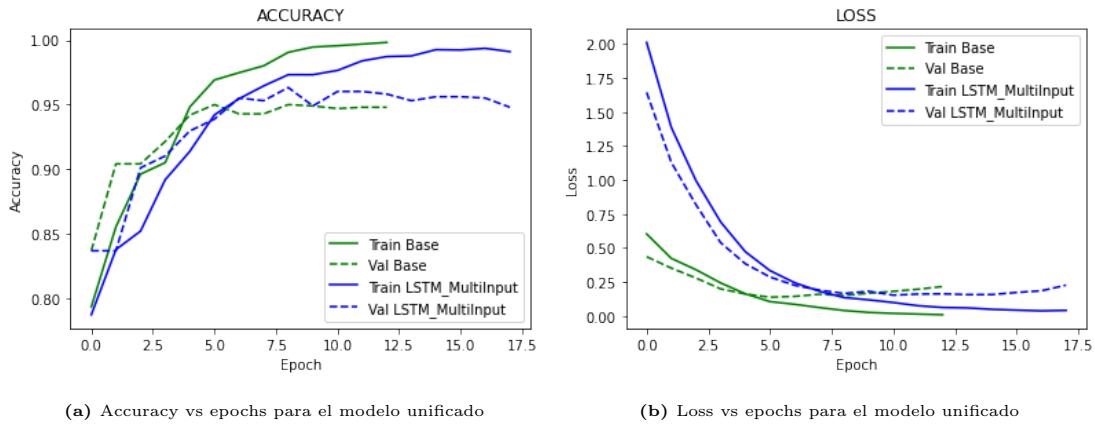


Figura 6.1: Comparación de modelo unificado LSTM de múltiples entradas con modelo base de única entrada

Adicionalmente, se realizó la prueba de ajustar la propuesta inicial del modelo LSTM de múltiples entradas, hacia un modelo similar pero de carácter bidireccional, es decir en este caso se entrenó un modelo LSTM bidireccional de múltiples entradas. En este sentido, a continuación se detalla el resumen del código planteado para definir esta red.

```

1 # Creacion de la red
2 nlp_input = Input(shape=(seq_length,), name='nlp_input')
3 meta_input = Input(shape=(2,), name='meta_input')
4 emb = Embedding(output_dim=emb_dim, input_dim=embedding_size, input_length=
      seq_length)(nlp_input)
5 nlp_out = Bidirectional(LSTM(64, dropout=0.7, recurrent_dropout=0.7,
      kernel_regularizer=regularizers.l2(0.01)))(emb)
6 x = concatenate([nlp_out, meta_input])
7 x = Flatten()(x)
8 x = Dense(classifier_neurons, activation='relu')(x)
9 x = Dense(1, activation='sigmoid')(x)
10 model_bLSTM = Model(inputs=[nlp_input, meta_input], outputs=[x])
11 # Compilacion
12 model_bLSTM.compile(optimizer='adam', loss='binary_crossentropy', metrics=['acc'])

```

Código 6.1: Definición de red LSTM bidireccional. Fuente: el autor

Los resultados obtenidos de la comparación de los modelos LSTM de múltiples entradas y el modelo LSTM bidireccional de múltiples entradas se presentan en la figura 6.2.

De acuerdo con los resultados, se observa que no hay una mejora sustancial en el rendimiento del modelo, mientras que el tiempo de entrenamiento se incrementó, alcanzando los 30 minutos.

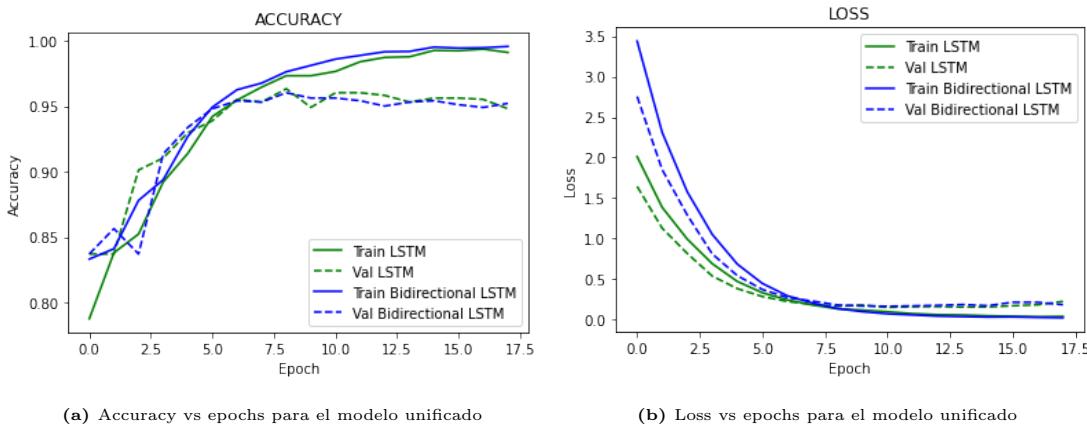


Figura 6.2: Comparación de modelo unificado LSTM de múltiples entradas con modelo similar bidireccional

6.2. Prueba del modelo y predicciones

Antes de entrenar los modelos, se realizó una reserva de datos especial para este momento de pruebas y predicciones, garantizando que estos datos nunca fueran “vistos” por los modelos desarrollados. En este caso los datos se guardaron en el dataframe “*datos-reserva*”, a partir del cual se extraerán registros para la ejecución de las pruebas.

A continuación se incluyen algunos de estos registros con los cuales se realizarán las primeras pruebas de predicciones:

- *Texto 1:* MTN Uganda is giving out 122GB of data to customers for free in response to COVID-19. Categoría real: Falso
- *Texto 2:* Turmeric And Lemon Help Fight Against coronavirus. Categoría real: Falso
- *Texto 3:* What we need to do to defeat the coronavirus is to consume more alkaline foods above the virus’ pH level. Categoría real: Falso

A continuación, se crea una función de prueba para aplicar la predicción al modelo.

```

1 # Funcion para prueba de predicciones
2 def news_alert(a, modelo_probar, tokenizer):
3     clf=modelo_probar
4     tok=tokenizer
5     # Tokenizacion
6     corpus_1=[]
7     corpus_1.append(a)
8     corpus_2=pd.Series(corpus_1)
9     sequences_reserva = tok.texts_to_sequences(corpus_2.values)

```

```

10 transform_vect_reserva= pad_sequences(sequences_reserva, maxlen=max_len)
11 prediccion=clf.predict(transform_vect_reserva)
12 prediccion_a = [np.array(prediccion)]
13 if prediccion > 0.5:
14     label= "NO"
15 else:
16     label = "SI"
17 alerta=[prediccion,label]
18 return alerta
19 # Aplico funcion
20 n=55
21 text1=datos_reserva["Text"][n]
22 clasificar_texto=text1
23 resultado_prediccion=news_alert(clasificar_texto, model_base, tokenizer_base)
24 print("Texto 1: ", text1)
25 print("Categoria real: ",datos_reserva["label"][n])
26 print(" ")
27 print("Prediccion: ")
28 print("Alerta de desinformacion: ", resultado_prediccion[1])
29 print("Probabilidad asociada: ", resultado_prediccion[0][0])

```

Código 6.2: Función para prueba de predicciones. Fuente: el autor

El resultado de la predicción para los titulares listados anteriormente fue el siguiente:

- *Texto 1:* MTN Uganda is giving out 122GB of data to customers for free in response to COVID-19. *Categoría real:* Falso, *Alerta de desinformación:* SI, *Probabilidad asociada:* [0.9999704]
- *Texto 2:* Turmeric And Lemon Help Fight Against coronavirus. *Categoría real:* Falso, *Alerta de desinformación:* SI, *Probabilidad asociada:* [0.99547905]
- *Texto 3:* What we need to do to defeat the coronavirus is to consume more alkaline foods above the virus' pH level. *Categoría real:* Falso, *Alerta de desinformación:* SI, *Probabilidad asociada:* [0.9980546]
- *Texto 4:* Bill Gates told us about the coronavirus in 2015. *Categoría real:* Verdadero, *Alerta de desinformación:* NO, *Probabilidad asociada:* [0.01032567]

Con el ejemplo del titular: “*Bill Gates told us about the coronavirus in 2015*” ocurre la particularidad de que a primera vista parecería falso, sin embargo revisando el hecho a profundidad dentro de los datos reservados, efectivamente este enunciado está etiquetado como verdadero. Esto se explica ya que por el hecho de que es factible que Bill Gates



Figura 6.3: Detalle de la interfaz desarrollada. Fuente: el autor

efectivamente hablaría del coronavirus en 2015 como enuncia el texto, haciendo referencia a otro coronavirus de ese momento y por supuesto no específicamente al actual coronavirus SARS-CoV-2 que produce la enfermedad de COVID-19.

6.3. Implementación del modelo

Con el fin de disponibilizar el modelo para pruebas, se realizó el diseño y desarrollo de una interfaz sencilla que le permite al usuario copiar y pegar un titular de una noticia y aplicar el modelo desarrollado. Una vez aplicado el modelo, la interfaz le devuelve al usuario la etiqueta del tema, el subtema en forma de palabras clave, la predicción de la alerta y su probabilidad asociada. En la figura 6.3 se presenta el diseño final de interfaz y un ejemplo de su respuesta al aplicar el modelo al hacer clic en el botón “*Predecir*”

Como una primera parte del desarrollo de la interfaz, se implementó toda la lógica de

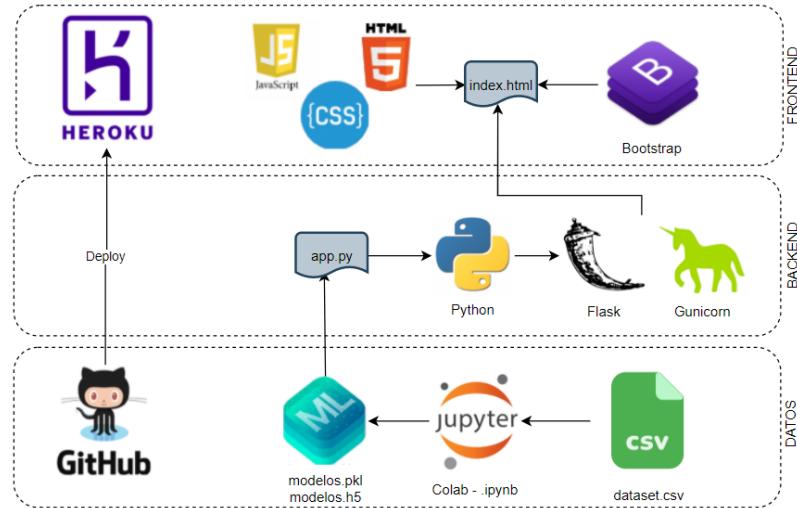


Figura 6.4: Arquitectura del modelo implementado. Fuente: el autor

la metodología en un cuaderno de Jupyter Notebook que incluyó la carga de los modelos desarrollados y almacenados en formatos *.pkl* y *.h5* y su aplicación a un titular de una noticia cualquiera. Este cuaderno se encuentra cargado en el repositorio de la investigación bajo el nombre: *ModeloFinalimplementacion.ipynb*¹ e incluye los siguientes pasos:

- Carga de modelos desde GitHub
- Carga del texto a revisar
- Aplicación de modelo para clasificar la temática
- Aplicación del modelo para extraer la subtemática y palabras clave
- Concatenación de resultados
- Aplicación de modelo unificado final

6.3.1. Interfaz

Posteriormente, se realizó el desarrollo de la interfaz de forma local, con una arquitectura basada en Python y Flask, agregando todo el código en los archivos *app.py* e *index.html* para desplegar la interfaz como una aplicación web en el navegador. En la figura 6.4 se define la arquitectura de la herramienta, incluyendo sus elementos en la *capa de datos*, en la *capa de backend* y en la *capa de frontend*.

¹<https://github.com/jorgecif/CovidDisinformationDetection>

6.3.2. Despliegue

Finalmente y luego de desarrollar la interfaz, se realizó su despliegue en un servicio de computación en la nube, en este caso se seleccionó Heroku² por permitir su integración directa con GitHub, en donde se dispuso todo el código³, y por las capacidades de su capa gratuita que en este caso fueron suficientes para soportar la herramienta completa. En este caso se realizaron algunos ajustes en las versiones de Python y de TensorFlow para el despliegue, con el fin de que no se superara el límite de los 500MB de la aplicación. La aplicación de prueba se encuentra desplegada en la siguiente url: <https://coviddisinformation.herokuapp.com/>.

6.3.3. Repositorio de código

Como ya se mencionó en el apartado anterior, el código de la investigación se dispuso en GitHub en el repositorio <https://github.com/jorgecif/CovidDisinformationDetection> y está organizado de la siguiente manera:

- Desarrollo (notebooks)

Scraping noticias Reuters:

ScrapingReuters.ipynb

- Comparación de métodos de clasificación de noticias:

ComparacionMetodosClasificacionNoticias.ipynb

- Comparación de métodos para la detección de desinformación:

ComparacionMetodosDeteccionDesinformacion.ipynb

- Modelo para extracción de subtema:

ExtraccionSubtematica.ipynb

- Modelo final - desarrollo:

ModeloFinalDeteccionDesinformacionCOVID.ipynb

²Heroku es una plataforma como servicio (PaaS) que permite a los desarrolladores crear, ejecutar y operar aplicaciones completamente en la nube: www.heroku.com.

³Repositorio de la investigación: <https://github.com/jorgecif/CovidDisinformationDetection>.

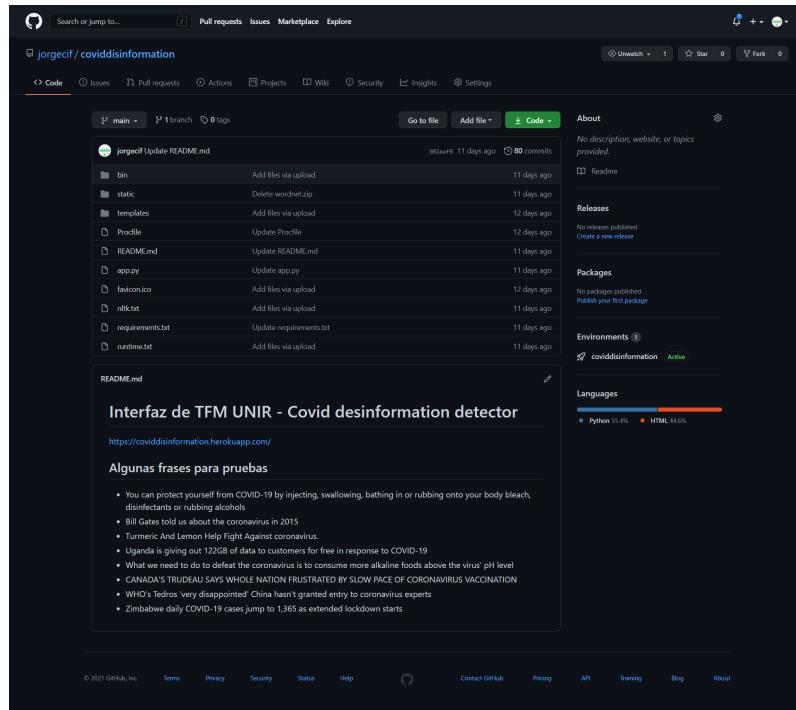


Figura 6.5: Repositorio Interfaz. Fuente: el autor

- Modelo final - implementación:

ModeloFinalimplementacion.ipynb

- Datos

data/CovidHeadlinesDataset2.xlsx

data/ReutersClasifiedNewsDataset.xlsx

- Modelos entrenados

Modelo predicción tema: TrainedModels/ModeloPrediccionTema

Modelo predicción alerta TrainedModels/ModeloPrediccionAlerta

Modelo subtema: TrainedModels/ModeloSubtema

Modelo final unificado: TrainedModels/ModeloFinalUnificado

- Interfaz de TFM UNIR - Covid desinformation detector

URL: <https://coviddisinformation.herokuapp.com/> (Con el fin de facilitar el despliegue directo en Heroku desde GitHub, la interfaz se creó en un repositorio independiente al del código principal (Ver figura 6.5)).

ID	Fuente	Texto	L	Tema	Subtema	P	D	Probabilidad
1	Dataset	Uganda is giving out 122GB of data to customers for free in response to COVID-19	0	Health	['trump', 'presid', 'state', 'death', 'unit']	0	0	[0.9999374]
2	Dataset	Bill Gates told us about the coronavirus in 2015	1	Health	['case', 'australia', 'health', 'pictur', 'australian']	1	0	[0.01032567]
3	Dataset	A tweet by Pakistani journalist Saadia Afzaal claiming that China has developed a COVID-19 vaccine	0	Health	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.9999794]
4	Dataset	Germany gave medical protection equipment like masks to China, now its missing in Germany	1	Health	['mask', 'kill', 'wear', 'chines', 'hand']	0	1	[0.9938691]
5	Dataset	Photos of Italian man committing suicide after he lost his entire family to COVID-19	0	Health	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.99996483]
6	Dataset	Harvard professor was arrested for creating and selling the coronavirus	0	Health	['toilet', 'paper', 'australia', 'health', 'custom']	0	0	[0.999812]
7	Dataset	Madagascar does not have any cases of the coronavirus	0	Health	['case', 'australia', 'health', 'pictur', 'australian']	0	0	[0.979228]
8	Dataset	COVID-19 is a bacterium that is easily treated with aspirin or a coagulant	0	Health	['test', 'health', 'hospit', 'clinic', 'australia']	0	0	[0.9554566]
9	Dataset	Trump Suspends Europe Travel, Announces New Economic Measures	1	Env	['trump', 'presid', 'state', 'death', 'unit']	1	0	[0.02297539]
10	Dataset	Social media users have shared a photo that claims to show a "Center for Global Human Population Reduction" affiliated with the Bill & Melinda Gates Foundation	0	Tech	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.9999998]

Tabla 6.1: Pruebas con noticias de 2020, incluidas en el conjunto de datos (L: Label, P: Predicción, D: Diferencia, 0: FALSE, 1: TRUE). Fuente: El autor

6.4. Pruebas de validación final

A partir del modelo desplegado se realizó un proceso de validación que consistió en una serie de pruebas con las que se buscó llevar al límite el modelo desarrollado para así identificar sus fortalezas y debilidades, revelando algunas pistas que pueden determinar posibles mejoras y trabajos futuros para continuar mejorando el modelo.

Las pruebas planteadas se dividieron en: pruebas con noticias de 2020, pruebas con noticias de 2021, pruebas con mitos sobre el COVID-19, pruebas con buenas prácticas para combatir el COVID-19, pruebas con hechos creados artificialmente, pruebas con negaciones o cambios de sentido. A continuación se detallan los resultados de cada una de ellas.

6.4.1. Pruebas con noticias del año 2020

Estas pruebas se realizaron considerando titulares de noticias incluidas dentro del rango de fechas del conjunto de datos, es decir aproximadamente hasta el mes de septiembre de 2020. Esto para asegurar que el modelo funciona correctamente con las generalidades, términos y hechos ocurridos hasta la fecha de entrenamiento del modelo.

En la tabla 6.1 se detalla el texto de los encabezados probado, su etiqueta real (L), y las predicciones realizadas por el modelo: tema, subtema , predicción (P), probabilidad

y un indicador de si existe alguna diferencia entre el la etiqueta real y la predicción (D). Para este caso los titulares de noticias probados se extrajeron del conjunto de datos reservado para pruebas, este conjunto de datos fue aislado desde el inicio y no se utilizó como entrenamiento, ni como validación de los modelos.

Luego de aplicar el modelo a cada uno de los encabezados se comprueba que el modelo predice correctamente la mayoría de este tipo de titulares de noticias. De los 10 encabezados revisados, solamente uno tuvo un resultado diferente a la etiqueta real, en este caso se trata del titular: “*Germany gave medical protection equipment like masks to China, now its missing in German*” el cual está etiquetado como verdadero y es predicho como falso. En la matriz de confusión de la tabla 6.7a se puede observar claramente el resultado de la prueba.

6.4.2. Pruebas con noticias del año 2021

Para esta prueba se seleccionaron titulares de noticias muy recientes, asegurando que estuvieran por fuera de las fechas incluidas dentro del conjunto de datos utilizado para desarrollar el modelo. El objetivo de esta prueba es comprobar que el modelo sigue siendo vigente a pesar de posibles cambios en el contexto de la temática, en este caso la pandemia de COVID-19.

En la tabla 6.2 se detalla el texto de los encabezados probado, su etiqueta real (L), y las predicciones realizadas por el modelo: tema, subtema , predicción (P), probabilidad y un indicador de si existe alguna diferencia entre el la etiqueta real y la predicción (D). Los titulares de las noticias se encontraron por medio de la herramienta Google Fact Check Explorer⁴, una nueva iniciativa de Google cuyo objetivo es recopilar hechos ya verificados y ofrecer una forma sencilla de para buscarlos por medio de palabras claves. El resultado de la búsqueda le muestra al usuario una lista de hechos relacionados con la búsqueda, su respectivo veredicto resultante del proceso de verificación, el organismo verificador y el enlace directo al artículo en el que se analiza el hecho.

Luego de aplicar el modelo a cada uno de los encabezados se comprueba que el modelo predice correctamente la mayoría de este tipo de titulares de noticias. De los 10 encabezados revisados, dos tuvieron un resultado diferente a la etiqueta real, el titular: “*COVID-19*

⁴<https://toolbox.google.com/factcheck/explorer>

ID	Fuente	Texto	L	Tema	Subtema	P	D	Probabilidad
1	Boomlive	COVID-19 vaccine do not eliminate the virus or stop the virus from transmitting	0	Health	['post', 'facebook', 'novel', 'show', 'share']	1	1	[0.08537072]
2	TLIndian	Pakistan Prime Minister Imran Khan has said, If Pakistan develops coronavirus vaccine then they will not give it to India or Israel.	0	Health	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.9995562]
3	Politifact	Says the new coronavirus vaccines contain toxic ingredients and are more dangerous than getting COVID-19.	0	Health	['trump', 'presid', 'state', 'death', 'unit']	0	0	[0.9998156]
4	TLIndian	Pope Francis said that Covid-19 vaccine will be required to enter heaven.	0	Health	['infect', 'case', 'health', 'report', 'pictur']	0	0	[0.9999604]
5	Forbes	China Deploys Anal Swab Tests To Detect High-Risk Covid-19 Cases	0	Health	['infect', 'case', 'health', 'report', 'pictur']	0	0	[0.18232581]
6	TLIndian	Countries in Europe have imposed lockdown against the second wave of the novel coronavirus.	1	Health	['infect', 'case', 'health', 'report', 'pictur']	0	1	[0.9998668]
7	Politifact	Says the new coronavirus vaccines contain toxic ingredients and are more dangerous than getting COVID-19.	0	Health	['trump', 'presid', 'state', 'death', 'unit']	0	0	[0.9998156]
8	Fullfact	The Covid vaccine will make you infertile.	0	Tech	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.9381714]
9	Fullfact	Covid vaccines contain aborted babies.	0	Health	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.9267428]
10	Boomlive	World Health Organization (WHO) ranked Sri Lanka fifth in a table of countries responses to the coronavirus pandemic.	0	Health	['case', 'australia', 'health', 'pictur', 'australian']	0	0	[0.9992931]

Tabla 6.2: Pruebas con noticias de 2021 (L: Label, P: Predicción, D: Diferencia, 0: FALSE, 1: TRUE). Fuente: El autor

vaccine do not eliminate the virus or stop the virus from transmitting" y el titular: "*Countries in Europe have imposed lockdown against the second wave of the novel coronavirus.*". La confusión del modelo se podría explicar debido a que en los dos casos se incluyen escenarios nuevos dentro del contexto actual que eran inexistentes en el momento en el que se entrenó el modelo, como lo son el hecho de que ya existe una vacuna desarrollada y que se esté presentando una nueva ola de una nueva variante del virus. En la matriz de confusión de la tabla 6.7b se puede observar claramente el resultado de la prueba.

6.4.3. Pruebas con mitos sobre el COVID-19

Una de los casos de uso de mayor utilidad del modelo desarrollado es la verificación de mitos relacionados con la pandemia de COVID-19, algunos de ellos se encuentran circulando en las redes sociales casi desde que comenzó la propia pandemia. En esta prueba se seleccionaron algunos famosos mitos recopilados por la Organización Mundial de la Salud [Organization(2020)] y la Facultad de Medicina de la Universidad Johns Hopkins [Medicine(2020)].

En la tabla 6.3 se detalla el texto de los mitos probados, su etiqueta real (L), y las predicciones realizadas por el modelo: tema, subtema , predicción (P), probabilidad y un

ID	Fuente	Texto	L	Tema	Subtema	P	D	Probabilidad
1	Hopkins	You can get a face mask exemption card so you don't need to wear a mask.	0	Tech	['mask', 'kill', 'wear', 'chines', 'hand']	0	0	[0.9065571]
2	Dataset	Turmeric And Lemon Help Fight Against coronavirus.	0	Health	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.99547905]
3	Hopkins	You can protect yourself from COVID-19 by injecting, swallowing, bathing in or rubbing onto your body bleach, disinfectants or rubbing alcohols.	0	Tech	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.97236454]
4	Hopkins	A vaccine to cure COVID-19 is available	1	Health	['post', 'facebook', 'novel', 'show', 'share']	0	1	[0.99695635]
5	Hopkins	The new coronavirus was deliberately created or released by people.	0	Health	['trump', 'presid', 'state', 'death', 'unit']	0	0	[0.99941945]
6	Hopkins	Ordering or buying products shipped from overseas will make a person sick.	0	Tech	['toilet', 'paper', 'australia', 'health', 'custom']	0	0	[0.9988524]
7	WHO	Can Covid-19 be transmitted through goods produced in countries where there is ongoing transmission?	0	Health	['toilet', 'paper', 'australia', 'health', 'custom']	0	0	[0.99999475]
8	WHO	Can Covid-19 be transmitted through mosquitoes?	0	Health	['infect', 'case', 'health', 'report', 'pictur']	0	0	[0.98874116]
9	WHO	How can we be sure that our clothes don't spread coronavirus?	0	Health	['mask', 'kill', 'wear', 'chines', 'hand']	1	1	[0.46550933]
10	WHO	Can drinking alcohol help prevent Covid-19?	0	Health	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.99845594]
11	WHO	Is it true that Covid-19 is transmitted in cold climate and not in hot and humid climate?	0	Sports	['infect', 'case', 'health', 'report', 'pictur']	0	0	[0.964204]
12	WHO	Can digital thermometers be 100 % effective in detecting Covid-19 patients?	0	Health	['test', 'health', 'hospit', 'clinic', 'australia']	0	0	[0.93550766]
13	WHO	Can UV bulbs used for disinfecting be used to kill Covid-19 on our body?	0	Health	['mask', 'kill', 'wear', 'chines', 'hand']	0	0	[0.9999908]
14	WHO	Can spraying alcohol or chlorine on your body kill the virus inside?	0	Env	['mask', 'kill', 'wear', 'chines', 'hand']	0	0	[0.99931556]
15	WHO	Can eating garlic prevent covid-19?	0	Health	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.9816327]
16	WHO	Can Pneumonia vaccine prevent Covid-19?	0	Health	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.7597852]
17	WHO	Can rinsing your nose regularly with saline solution prevent Covid-19?	0	Health	['infect', 'case', 'health', 'report', 'pictur']	0	0	[0.9985318]
18	WHO	Is there any drug that can prevent and treat Covid-19?	0	Health	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.9999687]

Tabla 6.3: Pruebas con mitos del Coronavirus encontrados en diversas fuentes (L: Label, P: Predicción, D: Diferencia, 0: FALSE, 1: TRUE). Fuente: El auto

indicador de si existe alguna diferencia entre el la etiqueta real y la predicción (D).

Se recolectaron en total 18 mitos a los cuales se les aplicó el modelo, resultando en solo un par de casos en los que la predicción fue diferente a la etiqueta real. Uno de los casos se trata de la afirmación: “*A vaccine to cure COVID-19 is available*”, la cual evidentemente es verdadera, sin embargo obtiene una predicción falsa.

Si tenemos en cuenta las fecha límites del conjunto de datos de entrenamiento del modelo, es evidente que hasta noviembre de 2020 efectivamente no existía ninguna vacuna disponible, por lo tanto la afirmación en ese contexto sería falsa, tal como lo predijo el modelo. En la matriz de confusión de la tabla 6.7c se puede observar claramente el resultado de la prueba.

ID	Fuente	Texto	L	Tema	Subtema	P	D	Probabilidad
1	WHO	The principal symptoms of COVID are fever, cough and tiredness	1	Health	['infect', 'case', 'health', 'report', 'pictur']	1	0	[0.04208112]
2	WHO	Keep the distance from others is a good to reduce the risk of coronavirus	1	Health	['case', 'australia', 'health', 'pictur', 'australian']	1	0	[0.01383418]
3	WHO	Washing hands often with antibacterial soap and water is imperative to protect yourself from covid.	1	Tech	['mask', 'kill', 'wear', 'chines', 'hand']	1	0	[0.2492171]
4	WHO	Animals and pets cannot catch coronavirus or transmit the virus to humans	1	Health	['infect', 'case', 'health', 'report', 'pictur']	1	0	[0.01292658]
5	WHO	Touching your face can lead to a fast transfer covid into the body	1	Politics	['infect', 'case', 'health', 'report', 'pictur']	1	0	[0.20195228]
6	WHO	Frequently disinfect surfaces such as door knobs, equipment handles, check-out counters is a best practice against covid	1	Politics	['infect', 'case', 'health', 'report', 'pictur']	1	0	[0.00035217]
7	WHO	Wear a mask is a good practice against covid	1	Politics	['mask', 'kill', 'wear', 'chines', 'hand']	1	0	[0.01882103]
8	WHO	Avoid poorly ventilated spaces and crowded spaces is a good practice to prevent coronavirus	1	Health	['case', 'australia', 'health', 'pictur', 'australian']	1	0	[0.01086292]
9	WHO	People with comorbidities have the highest risk of contracting covid virus	1	Health	['case', 'australia', 'health', 'pictur', 'australian']	1	0	[0.22701627]
10	WHO	Clean and disinfect frequently touched surfaces daily is a good practice against covid	1	Tech	['infect', 'case', 'health', 'report', 'pictur']	1	0	[0.00020418]

Tabla 6.4: Pruebas con buenas prácticas recopiladas por la Organización Mundial de la Salud (L: Label, P: Predicción, D: Diferencia, 0: FALSE, 1: TRUE). Fuente: El autor

6.4.4. Pruebas con buenas prácticas para mitigar el COVID-19

Teniendo en cuenta que hasta el momento la mayoría de las pruebas se han realizado con titulares falsos, se plantea esta prueba para verificar el comportamiento del modelo ante afirmaciones o enunciados con una etiqueta evidentemente verdadera. Para esto se realiza una búsqueda de buenas prácticas ya evidenciadas para, en este caso, mitigar el riesgo de contraer el coronavirus, citando principalmente a la Organización Mundial de la Salud.

En la tabla 6.4 se detalla el texto de las buenas prácticas probadas, su etiqueta real (L), y las predicciones realizadas por el modelo: tema, subtema , predicción (P), probabilidad y un indicador de si existe alguna diferencia entre el la etiqueta real y la predicción (D). En este caso se realiza una prueba con una selección de 10 buenas prácticas ya ampliamente probadas en la prevención del contagio del coronavirus y que han sido ampliamente divulgadas por la Organización Mundial de la Salud.

Luego de aplicar el modelo a cada uno de las buenas prácticas se comprueba que el modelo predice correctamente la totalidad de este tipo de enunciados. En la matriz de confusión de la tabla 6.7d se puede observar claramente el resultado de la prueba.

ID	Fuente	Texto	L	Tema	Subtema	P	D	Probabilidad
1	El autor	President Donald Trump dies by coronavirus	0	Politics	['trump', 'presid', 'state', 'death', 'unit']	0	0	[0.99752414]
2	El autor	Important research shows that covid is transmitted through water	0	Env	['infect', 'case', 'health', 'report', 'pictur']	0	0	[0.9999754]
3	El autor	In Colombia no covid contegy is reported	0	Politics	['infect', 'case', 'health', 'report', 'pictur']	0	0	[0.95423234]
4	El autor	A town in europe is detected where all its inhabitants are immune to the coronavirus	0	Health	['infect', 'case', 'health', 'report', 'pictur']	0	0	[0.96206295]
5	El autor	it is shown that the coronavirus was created in a laboratory as a biological weapon for the birth control of the population	0	Health	['trump', 'presid', 'state', 'death', 'unit']	0	0	[0.9999866]
6	El autor	the coronavirus vaccine is a business and the price at which it is sold is 100 times higher than its cost of production	0	Health	['toilet', 'paper', 'australia', 'health', 'custom']	0	0	[0.99994314]
7	El autor	First case of a dog with coronavirus detected in Australia	0	Health	['case', 'australia', 'health', 'pictur', 'australian']	1	1	[0.03450221]
8	El autor	in africa the coronavirus has not spread because its inhabitants have a very strong immune system	0	Health	['case', 'australia', 'health', 'pictur', 'australian']	0	0	[0.9943732]
9	El autor	Person who receives package from china by ebay gets coronavirus	0	Health	['mask', 'kill', 'wear', 'chines', 'hand']	0	0	[0.98594224]
10	El autor	International space station astronaut catches coronavirus	0	Env	['infect', 'case', 'health', 'report', 'pictur']	0	0	[0.99462175]
11	El autor	Joe Biden have coronavirus	0	Health	['trump', 'presid', 'state', 'death', 'unit']	0	0	[0.9040914]

Tabla 6.5: Pruebas con hechos artificialmente creados por el autor. (L: Label, P: Predicción, D: Diferencia, 0: FALSE, 1: TRUE). Fuente: El autor

6.4.5. Pruebas con hechos creados artificialmente

Esta prueba pretende ponerse en los zapatos de un generador de desinformación para, con un poco de creatividad, simular la creación de un conjunto de posibles noticias falsas relacionadas con la pandemia de COVID-19.

Se generaron un total de 11 titulares de noticias evidentemente falsas, las cuales se detallan en la tabla 6.5 incluyendo su etiqueta real (L), y las predicciones realizadas por el modelo: tema, subtema, predicción (P), probabilidad y un indicador de si existe alguna diferencia entre el la etiqueta real y la predicción (D).

Luego de aplicar el modelo a cada uno de los encabezados se comprueba que el modelo predice correctamente la mayoría de este tipo de titulares de noticias, sin embargo en el titular: “*First case of a dog with coronavirus detected in Australia*” no genera una alerta y es predicho como un hecho verdadero o sin contenido de desinformación. Este hecho es relevante pues ya se había probado anteriormente la afirmación “*Animals and pets cannot catch coronavirus or transmit the virus to humans*” y se había predicho como verdadera, sin embargo a pesar de que la frase “*First case of a dog with coronavirus detected in Australia*” se refiere a una mascota que contrae el coronavirus, es evidente una

ID	Fuente	Texto	L	Tema	Subtema	P	D	Probabilidad
1	WHO	Washing hands often with antibacterial soap and water is imperative to protect yourself from covid.	1	Tech	['mask', 'kill', 'wear', 'chiness', 'hand']	1	0	[0.2492171]
2	El autor	Washing hands often with antibacterial soap and water is not imperative to protect yourself from covid.	0	Tech	['mask', 'kill', 'wear', 'chiness', 'hand']	1	1	[[0.21670559]]
3	WHO	The principal symptoms of COVID are fever, cough and tiredness	1	Health	['infect', 'case', 'health', 'report', 'pictur']	1	0	[0.04208112]
4	El autor	These are not the main symptoms of COVID are fever, cough and tiredness	0	Health	['infect', 'case', 'health', 'report', 'pictur']	0	0	[0.5414618]
5	WHO	Animals and pets cannot catch coronavirus or transmit the virus to humans	1	Health	['infect', 'case', 'health', 'report', 'pictur']	1	0	[0.01292658]
6	El autor	Animals and pets can get coronavirus and can transmit the virus to humans	0	Env	['infect', 'case', 'health', 'report', 'pictur']	1	1	[0.04835919]
7	Hopkins	You can protect yourself from COVID-19 by injecting, swallowing, bathing in or rubbing onto your body bleach, disinfectants or rubbing alcohols.	0	Tech	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.97236454]
8	El autor	You cannot protect yourself from COVID-19 by injecting, swallowing, bathing in or rubbing onto your body bleach, disinfectants or rubbing alcohols.	1	Tech	['post', 'facebook', 'novel', 'show', 'share']	0	1	[0.9596821]
9	Dataset	Turmeric And Lemon Help Fight Against coronavirus.	0	Health	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.99547905]
10	El autor	Turmeric And Lemon not Help Fight Against coronavirus.	0	Health	['post', 'facebook', 'novel', 'show', 'share']	1	1	[0.9961331]

Tabla 6.6: Pruebas con negaciones de noticias ya comprobadas. (L: Label, P: Predicción, D: Diferencia, 0: FALSE, 1: TRUE). Fuente: El auto

contradicción en el modelo. En la matriz de confusión de la tabla 6.7e se puede observar claramente el resultado de la prueba.

6.4.6. Pruebas con negaciones de frases

Con el fin de explorar un poco más a fondo el tema de la contradicción detectada anteriormente, se plantea probar el comportamiento del modelo ante cambios de sentido de los enunciados. Para esto se redacta nuevamente la frase en forma de negación, tratando de agregarle la menor cantidad de palabras posible.

Un ejemplo de esto es la afirmación: “*Washing hands often with antibacterial soap and water is imperative to protect yourself from covid*”, la cual al ser redactada nuevamente como: “*Washing hands often with antibacterial soap and water is not imperative to protect yourself from covid*.” cambia completamente el sentido y pasaría de tener una etiqueta de verdadera a una etiqueta falsa.

En la tabla 6.6 se incluyen 5 afirmaciones, cada una con su redacción original y su redacción en negativo que le cambia el sentido. Para cada enunciado se indica en la tabla su etiqueta real (L), y las predicciones realizadas por el modelo: tema, subtema, predicción (P), probabilidad y un indicador de si existe alguna diferencia entre el la etiqueta real y la predicción (D).

	Real	
True	7	0
False	1	2
	False True	

Predicción

	Real	
True	8	1
False	1	0
	False True	

Predicción

(a) Matriz de confusión pruebas noticias 2020. Fuente: El autor

(b) Matriz de confusión pruebas noticias 2021. Fuente: El autor

	Real	
True	16	1
False	1	0
	False True	

Predicción

	Real	
True	0	0
False	0	10
	False True	

Predicción

(c) Matriz de confusión pruebas mitos. Fuente: El autor

(d) Matriz de confusión pruebas buenas prácticas. Fuente: El autor

	Real	
True	10	1
False	0	0
	False True	

Predicción

	Real	
True	1	3
False	1	0
	False True	

Predicción

(e) Matriz de confusión pruebas hechos creados. Fuente: El autor

(f) Matriz de confusión pruebas negaciones. Fuente: El autor

Tabla 6.7: Matrices de confusión de pruebas finales a, b, c and d.

Al aplicar el modelo, 4 de los 5 enunciados a los que se les cambió el sentido no son detectados correctamente por el modelo, evidenciando en este caso una limitación del mismo para este tipo de casos en particular en donde unas pocas palabras agregadas a un enunciado cambian completamente el sentido del mismo. En la matriz de confusión de la tabla 6.7f se puede observar en detalle el resultado de la prueba.

6.4.7. Análisis de resultados

Si bien en el cuadro 6.7 se han generado las matrices de confusión de cada una de las pruebas realizadas, en el cuadro 6.8 se incluye una matriz de confusión consolidada que agrupa la totalidad de las pruebas y que permite comparar objetivamente la totalidad de los resultados de las pruebas.

A partir de estos resultados se plantean las siguientes observaciones y puntos de discusión, los cuales se espera orienten nuevos trabajos para complementar y mejorar la metodología planteada.

		Real		Predicción
		True	False	
Real	True	42	6	
	False	4	12	
		False	True	

Tabla 6.8: Matriz de confusión consolidada de todas las pruebas. Fuente: El autor

- Teniendo en cuenta que el modelo desarrollado en la presente investigación está orientado hacia la detección de desinformación, la cantidad de elementos catalogados como verdaderos fue considerablemente menor que la de los elementos catalogados como falsos, esto hace que para el modelo sea más difícil distinguir los elementos catalogados como verdaderos. Una posible mejora podría ser recopilar una mayor cantidad de titulares verdaderos y re-entrenar el modelo.
- En una temática que actualmente es tan popular como la de la pandemia de COVID-19, se genera día a día una gran cantidad de información y se corre el riesgo de que el contexto bajo el cual se realizó el entrenamiento del modelo cambie tan rápidamente, y en tal proporción que logre desactualizar el modelo. La recolección de nuevos datos para con ellos actualizar periódicamente el modelo es clave para que no pierda vigencia con el paso del tiempo. Un proceso automatizado de entrenamiento continuo podría ser una gran mejora para la metodología planteada.
- Las pruebas en las que se realizó el cambio de sentido de los titulares al negarlos, dejó en evidencia la limitación del modelo para detectar estos cambios. Esto se debe a que a pesar de que el sentido de la frase se cambió totalmente, la mayoría de las palabras en la oración se conservan, por lo que no se genera un cambio significativo que pueda ser detectado por el modelo. La investigación en métodos que logren entender este tipo de cambios en el sentido de los titulares puede aportar una considerable mejora al modelo.
- Otro punto de mejora identificado es sin duda la necesidad de trabajar en el relacionamiento de las palabras para evitar así las contradicciones, como por ejemplo la detectada entre las frases: *Animals and pets cannot catch coronavirus or transmit the virus to humans* ” etiquetada como verdadera, y la frase “*First case of a dog with coronavirus detected in Australia*” etiquetada también como verdadera. El modelo efectivamente no reconoce la relación entre la palabra “*dog*” y la palabra “*pet*”,

generando una contradicción ya que si el modelo ha aprendido que las mascotas no pueden contagiarse de coronavirus y teniendo en cuenta que un perro es una mascota, la afirmación de que se detectó un primer caso de coronavirus en un perro debería ser etiquetada inmediatamente como falsa.

- Como fortalezas del modelo desarrollado se evidencia su aplicación para la validación de mitos y la verificación de buenas prácticas relacionadas con el coronavirus. Esta puede ser un caso de uso muy demandado que podría convertirse rápidamente en una aplicación de gran impacto para los usuarios y que a su vez permitiría continuar recolectando datos para alimentar el entrenamiento continuo del modelo.

Capítulo 7

Conclusiones y trabajo futuro

7.1. Conclusiones

- Es evidente que la desinformación en forma de noticias falsas cobra más sentido en tiempos de crisis o acontecimientos relevantes a nivel nacional o mundial. Tal es el caso de la pandemia de COVID-19, donde este tipo de noticias han cobrado relevancia y se han difundido casi tan rápido como el mismo virus. Es un momento clave para investigar este tipo de infodemias ya que contamos con la información y los datos disponibles, y así poder estar listos para las próximas olas de desinformación que seguramente van a seguir generándose en un mundo donde las personas estamos cada vez más interconectados.
- Se comprobó que las técnicas de inteligencia artificial independientemente de si se trata de técnicas de aprendizaje automático o de aprendizaje profundo pueden ser usadas para apoyar la detección de desinformación dentro de titulares de noticias y específicamente para apoyar la detección de desinformación relacionada con la pandemia de COVID-19. Estas técnicas podrán apoyar la tarea para entregarle más pistas al usuario sobre un posible contenido que no sea confiable pero siempre el usuario será quien tenga la última palabra para tomar la decisión de replicar o no replicar una información.
- Un desafío detectado durante la investigación fue el contexto actual frente a las noticias que se han generado en el marco de la pandemia de COVID-19. El contexto frente a la desinformación es dinámico y se adapta de acuerdo a los acontecimientos de la historia. En este caso los conjuntos de datos tradicionalmente usados para el

entrenamiento de clasificadores de noticias falsas no fueron de gran utilidad, pues no generaban un modelo coherente con el contexto de las noticias actuales. Esto quiere decir que es importante el desarrollo de modelos que puedan adaptarse fácilmente al entorno y puedan aprender rápidamente para poder seguir aportando buenos resultados.

- La posibilidad de implementar diferentes entradas dentro de un modelo de redes neuronales profunda permite experimentar con nuevas opciones para su entrenamiento como por ejemplo diferentes tipos de datos: secuenciales y no secuenciales, texto y numéricos, entre otros. Esta característica permite utilizar redes neuronales de extremo a extremo, combinando características diferentes dentro de la red en lugar de apilar modelos manualmente, expandiendo los casos de uso de este tipo de redes hacia nuevos horizontes.

7.2. Líneas de trabajo futuro

- Incorporación de atributos adicionales de caracterización de la información: Pensar en nuevas formas de caracterizar la información puede ser de gran utilidad para lograr mejores resultados en modelos más adaptables y moldeables al contexto. Los corpus de entrenamiento deberían poder incluir por ejemplo contextos relacionados con las noticias de actualidad que puedan ser utilizadas como fuente de información complementaria de referencia.
- Implementación de técnicas en idioma Español: La gran debilidad para desarrollar este tipo de técnicas y modelos está en la ausencia de los conjuntos de datos adecuados para realizar las investigaciones. El inicio de esta línea de trabajo futuro debería ser recolectar y preparar estos conjuntos de datos para luego iniciar con el desarrollo de estas técnicas. Otra posibilidad es pensar en extraer características que puedan ser independientes del idioma o que sean compartidas por varios idiomas.
- Incluir nuevos tipos de datos dentro del análisis: La posibilidad de incorporar varias entradas dentro de un modelo de redes neuronales profundas permite pensar en combinar diferentes tipos de elementos que también hacen parte de la noticia a verificar, como imágenes o metadatos relacionados. Esto puede derivar en modelos que alcancen un mayor rendimiento al considerar más información para la toma de

decisiones.

- El desafío del contexto cambiante: Este reto es uno de los más importantes en el ámbito de la desinformación en forma de noticias falsas, ya que a medida que pasa el tiempo los acontecimientos de la historia hacen que el contexto de las noticias cambie y la forma de generar desinformación también se adapta. Esto motiva a pensar en modelos resistentes al contexto o adaptables y que puedan aprender en el camino, como por ejemplo las técnicas de aprendizaje reforzado, las cuales podrían ser de utilidad en estos nuevos escenarios.

Bibliografía

- [AGARWAL(2020)] AGARWAL, D., Isha; Rana (2020). Covid19fn fake news dataset for covid-19. In *Sardar Vallabhbhai National Institute of Technology, Mendeley Data*. URL: <https://data.mendeley.com/datasets/b96v5hmfv6/3>.
- [Agarwal et al.(2019)Agarwal, Sultana, Malhotra & Sarkar] Agarwal, V., Sultana, H. P., Malhotra, S., & Sarkar, A. (2019). Analysis of classifiers for fake news detection. *Procedia Computer Science*, 165, 377 – 383. URL: <http://www.sciencedirect.com/science/article/pii/S1877050920300430>. doi:<https://doi.org/10.1016/j.procs.2020.01.035>. 2nd International Conference on Recent Trends in Advanced Computing ICRTAC -DISRUP - TIV INNOVATION , 2019 November 11-12, 2019.
- [Andrew(2019)] Andrew, B. (2019). Trump claims he invented the term “fake news”—here’s an interview with the guy who actually helped popularize it. *Washingtonian Magazine*, .
- [Aphiwongsophon & Chongstitvatana(2018)] Aphiwongsophon, S., & Chongstitvatana, P. (2018). Detecting fake news with machine learning method. (pp. 528–531). doi:10.1109/ECTICon.2018.8620051.
- [Bahad et al.(2019)Bahad, Saxena & Kamal] Bahad, P., Saxena, P., & Kamal, R. (2019). Fake news detection using bi-directional lstm-recurrent neural network. *Procedia Computer Science*, 165, 74–82.
- [BoerWarArchive(2014)] BoerWarArchive (2014). Political cartoons the boer war in drawings. URL: <http://www.boerwararchive.com/political-cartoons/index2.html> [Web; accedido el 02-11-2020].

[Brown(2008)] Brown, T. (2008). Design thinking. *Harvard business review*, 86, 84–92, 141.

[Cai et al.(2018)] Cai, J., Li, J., Li, W., & Wang, J. (2018). Deeplearning model used in text classification. In *2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*.

[Conversation(2017)] Conversation, T. (2017). The fake news that sealed the fate of antony and cleopatra. URL: URL{<https://theconversation.com/the-fake-news-that-sealed-the-fate-of-antony-and-cleopatra-7128>} [Web; accedido el 02-11-2020].

[Deb et al.(2020)] Deb, N., Jha, V., Panjiyar, A., & Gupta, R. (2020). A comparative analysis of news categorization using machine learning approaches. *International Journal of Scientific Technology Research*, 9, 2469–2472.

[Drif et al.(2019)] Drif, A., Ferhat Hamida, Z., & Giordano, S. (2019). Fake news detection method based on text-features.

[Dulhanty et al.(2019)] Dulhanty, C., Deglint, J., Ben Daya, I., & Wong, A. (2019). Taking a stance on fake news: Towards automatic disinformation assessment via deep bidirectional transformer language models for stance detection. [arXiv:1911.11951](https://arxiv.org/abs/1911.11951).

[Gelfert(2018)] Gelfert, A. (2018). Fake news: A definition. *Informal Logic*, 38, 84–117.

[Ghosh & Shah(2018)] Ghosh, S., & Shah, C. (2018). Toward automatic fake news classification. In *Proceedings of the Association for Information Science and Technology* (pp. 805–807). volume 55.

[Girgis et al.(2018)] Girgis, S., Amer, E., & Gadallah, M. (2018). Deep learning algorithms for detecting fake news in online text. In *2018 13th International Conference on Computer Engineering and Systems (ICCES)*.

[Ireton & Posetti(2018)] Ireton, C., & Posetti, J. (2018). Journalism, "fake news" disinformation. *UNESCO*, .

[Klaus & Børge(2018)] Klaus, S., & Børge, B. (2018). The global risks report 2018 13th edition. *World Economic Forum*, .

[Kudarvalli & Fiaidhi(2020)] Kudarvalli, H., & Fiaidhi, J. (2020). Detecting Fake News using Machine Learning Algorithms, . URL: https://www.techrxiv.org/articles/preprint/Detecting_Fake_News_using_Machine_Learning_Algorithms/12089133. doi:10.36227/techrxiv.12089133.v1.

[Kula et al.(2020)] Kula, Choraś, Kozik, Ksieniewicz & Woźniak] Kula, S., Choraś, M., Kozik, R., Ksieniewicz, P., & Woźniak, M. (2020). Sentiment analysis for fake news detection by means of neural networks. In V. V. Krzhizhanovskaya, G. Závodszky, M. H. Lees, J. J. Dongarra, P. M. A. Sloot, S. Brissos, & J. Teixeira (Eds.), *Computational Science – ICCS 2020* (pp. 653–666). Cham: Springer International Publishing.

[Liu et al.(2019)] Liu, Gherbi, Li & Cheriet] Liu, X., Gherbi, A., Li, W., & Cheriet, M. (2019). Multi features and multi-time steps lstm based methodology for bike sharing availability prediction. *Procedia Computer Science*, 155, 394 – 401. URL: <http://www.sciencedirect.com/science/article/pii/S187705091930969X>. doi:<https://doi.org/10.1016/j.procs.2019.08.055>. The 16th International Conference on Mobile Systems and Pervasive Computing (MobiSPC 2019),The 14th International Conference on Future Networks and Communications (FNC-2019),The 9th International Conference on Sustainable Energy Information Technology.

[Liu & fang Brook Wu(2020)] Liu, Y., & fang Brook Wu, Y. (2020). Fned: A deep network for fake news early detection on social media. *ACM Transactions on Information Systems*, 38, 1–33.

[Medicine(2020)] Medicine, J. H. (2020). Coronavirus disease 2019: Myth vs. fact. URL: URL{<https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/2019-novel-coronavirus-myth-versus-fact>} [Web; accedido el 28-01-2021].

[Nakamura et al.(2020)] Nakamura, Levy & Wang] Nakamura, K., Levy, S., & Wang, W. Y. (2020). r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection, . [arXiv:1911.03854](https://arxiv.org/abs/1911.03854).

[NewYorkTimes(2017)] NewYorkTimes (2017). Orson welles and the birth of fake news. URL: URL{<https://www.nytimes.com/2018/10/30/opinion/orson-welles-war-of-the-worlds-fake-news.html>} [Web; accedido el 02-11-2020].

[Organization(2020)] Organization, W. H. (2020). Coronavirus disease (covid-19) advice for the public: Mythbusters. URL: URL{<https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters>} [Web; accedido el 28-01-2021].

[Oxford(2020a)] Oxford (2020a). Disinformation meaning. *Oxford Dictionary*, . URL: <https://dictionary.cambridge.org/es/diccionario/ingles/disinformation>.

[Oxford(2020b)] Oxford (2020b). Fake news meaning. *Oxford Dictionary*, . URL: <https://dictionary.cambridge.org/es/diccionario/ingles/fake-news>.

[Shahi & Nandini(2020)] Shahi, G. K., & Nandini, D. (2020). FakeCovid – a multilingual cross-domain fact check news dataset for covid-19. In *Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media*. URL: http://workshop-proceedings.icwsm.org/pdf/2020_14.pdf.

[Shoemaker(2019)] Shoemaker, E. (2019). Using data science to detect fake news. *James Madison University, JMU Scholarly Commons*.

[Shu et al.(2017)Shu, Sliva, Wang, Tang & Liu] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *Sigkdd Explorations*, 19, 22–36.

[Tandoc et al.(2018)Tandoc, Lim & Ling] Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining “fake news”. *Digital journalism*, 6, 137–153.

[Turkoglu et al.(2019)Turkoglu, Hanbay & Sengur] Turkoglu, M., Hanbay, D., & Sengur, A. (2019). Multi-model lstm-based convolutional neural networks for detection of apple diseases and pests. *Ambient Intell Human Comput*, . doi:10.1007/s12652-019-01591-w.

[Weller(2019)] Weller, A. J. (2019). Design thinking for a user-centered approach to artificial intelligence. *She Ji: The Journal of Design, Economics, and Innovation*, 5, 394 – 396. URL: <http://www.sciencedirect.com/science/article/pii/S2405872619300887>. doi:<https://doi.org/10.1016/j.sheji.2019.11.015>.

[Wikipedia(2021)] Wikipedia (2021). Great moon hoax. URL: URL{https://es.wikipedia.org/wiki/Great_Moon_Hoax} [Web; accedido el 02-11-2020].

[Xu et al.(2019)] Xu, Meng, Chen, Qiu, Wang & Yao] Xu, G., Meng, Y., Chen, Z., Qiu, X., Wang, C., & Yao, H. (2019). Research on topic detection and tracking for online news texts. *IEEE Access*, 7, 58407–58418. doi:10.1109/ACCESS.2019.2914097.

[Xu et al.(2020)] Xu, Wang, Wang & Yang] Xu, K., Wang, F., Wang, H., & Yang, B. (2020). Detecting fake news over online social media via domain reputations and content understanding. *Tsinghua Science and Technology*, 25, 20–27. doi:10.26599/TST.2018.9010139.

[Yang et al.(2018)] Yang, Zheng, Zhang, Cui, Li & Yu] Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z., & Yu, P. S. (2018). Ti-cnn: Convolutional neural networks for fake news detection. [arXiv:1806.00749](https://arxiv.org/abs/1806.00749).

[Yurkova(2018)] Yurkova, O. (2018). Inside the fight against russia's fake news empi-
re. ted. *TED Conference*, . URL: https://www.ted.com/talks/olga_yurkova_inside_the_fight_against_russia_s_fake_news_empire.

[Zhang et al.(2020)] Zhang, Dong & Yu] Zhang, J., Dong, B., & Yu, P. S. (2020). Fake-detector: Effective fake news detection with deep diffusive neural network. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)* (pp. 1826–1829).

[Zhang & Ghorbani(2020)] Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing Management*, 57, 102025. URL: <http://www.sciencedirect.com/science/article/pii/S0306457318306794>. doi:<https://doi.org/10.1016/j.ipm.2019.03.004>.

[Zhou & Zafarani(2018)] Zhou, X., & Zafarani, R. (2018). Fake news: A survey of research, detection methods, and opportunities. *ArXiv*, *abs/1812.00315*.

Apéndice A

Anexo 1. Artículo

Detección de desinformación relacionada con la pandemia de COVID-19 por medio de técnicas de aprendizaje automático, aprendizaje profundo y procesamiento de lenguaje natural.

Jorge Orlando Cifuentes Cifuentes

Universidad Internacional de la Rioja, Logroño (España)

Febrero de 2021

RESUMEN

En 2020 y 2021 hemos sido testigos de cómo la *pandemia* de COVID-19 se ha propagado de manera exponencial por el mundo al mismo tiempo en el que una *infodemia* de desinformación se expande a una velocidad casi tan rápida como la del propio virus. Esta investigación plantea una metodología para abordar el problema de la detección de desinformación relacionada con el COVID-19 por medio de técnicas de inteligencia artificial buscando extraer la mayor cantidad de información posible para su clasificación. El resultado combina técnicas de aprendizaje automático y aprendizaje profundo en un modelo de múltiples entradas que le entrega al usuario, además de la clasificación final, información adicional como el tema y el subtema de la noticia. Con una precisión superior al 90 %, el modelo detectó correctamente desinformación en noticias actuales e incluso dentro de mitos ampliamente conocidos como por ejemplo: “*Puede protegerse del COVID-19 inyectando, tragando, frotándose o bañándose con desinfectantes o alcohol*”.



PALABRAS CLAVE

Artificial intelligence, machine learning, deep learning, fake news, disinformation, COVID-19, infodemia

I. INTRODUCCIÓN

Tres años antes del referéndum del Brexit y de las elecciones presidenciales de EE. UU. de 2016, sucesos que derivaron en la acuñación del término “Fake News” o “Noticias Falsas”, el Foro Económico Mundial en su Reporte de Riesgos Globales, ya dedicaba un capítulo completo titulado “*Incendios forestales digitales en un mundo hiperconectado*” en donde advertía sobre el inminente peligro de la desinformación difundida de manera incontrolable en las redes sociales [1].

Las facilidades para el acceso a la información, los avances en la electrónica y la hiperconectividad del mundo de hoy, han sido el caldo de cultivo perfecto para la proliferación de desinformación en forma de noticias falsas; noticias que unas veces inocentes y otras no tanto, han logrado influir en momentos críticos de nuestra historia. Desde el año el año 2020 hemos sido testigos de una nueva ola de desinformación a partir de la pandemia de COVID-19, una enfermedad que irónicamente ha llegado a viralizarse no solamente entre las personas, sino también tam-

bien en las redes sociales y medios digitales, a partir de noticias que sólo buscan generar desinformación en algo que bien se podría denominar paralelamente la “*Infodemia de COVID-19*”.

Es así como se hace relevante la búsqueda de herramientas que puedan aportar en la detección de desinformación desde el ámbito de la inteligencia artificial y se plantea esta investigación, la cual busca definir una metodología para la detección de desinformación relacionada con la pandemia de COVID-19 por medio de técnicas de aprendizaje automático, aprendizaje profundo y procesamiento de lenguaje natural. A continuación se describe el desarrollo y resultado de la investigación.

II. ESTADO DEL ARTE

Entendiendo la desinformación como información falsa o engañosa creada intencionalmente con el objetivo de engañar a la audiencia [2] y a las noticias falsas como historias falsas que parecen ser noticias creadas para engañar[3], es evidente que este tipo

de noticias se ha convertido en nuestros tiempos en una de las formas más ampliamente usadas para la difusión de desinformación. Por esta razón y para fines de la presente investigación se acotará el estudio de la desinformación a su materialización en forma de “*noticias falsas*” difundidas a través de medios digitales o redes sociales.

Para entender el contexto de la presente investigación, se ha dividido el estado del arte en dos grandes partes: La primera parte está relacionada con la desinformación y las noticias falsas, incluyendo su definición, principales características y tipologías. La segunda parte, está relacionada con las técnicas de inteligencia artificial utilizadas en los últimos años para la detección de desinformación dentro de una noticia.

A. Desinformación y noticias falsas

A.1. ¿Qué son las noticias falsas?

De acuerdo con el diccionario de Oxford, las noticias falsas se definen como: *historias falsas que parecen ser noticias, difundidas a través de Internet u otros medios y que generalmente son creadas para influir en opiniones políticas o como bromas* [3].

A.2. Historia y evolución

Es evidente cómo los conflictos, cambios de régimen, crisis y catástrofes naturales, han sido los mayores caldos de cultivo para la generación de información falsa.



Figura 1: Denario Marco Antonio-Cleopatra. Fuente: [4]

Las noticias falsas han existido desde los inicios de la historia y estas, además de su contenido construido deliberadamente con la intención de desinformar, están acompañadas a un canal de difusión que ha venido evolucionando; desde las monedas romanas usadas en contra de Marco Antonio (Ver figura 1), pasando por la Imprenta de Gutenberg y en nuestros tiempos, el Internet y sus aplicaciones como las redes sociales, que se fusionan en un mundo cada vez más hiperconectado. Nunca en la historia de la humanidad habíamos tenido un ca-

nal tan efectivo para disseminar la desinformación y esto, sin duda, es lo que hoy ha hecho la diferencia.

A.3. Caracterización

Una completa caracterización de las noticias falsas se incluye en el trabajo de investigación [5], en donde se plantea un modelo basado en capas con cuatro componentes principales:

- *Creador/difusor*: Puede ser humanos o no. Incluye aquellos que crean o publican la noticia con o sin intención
- *Víctima*: Son el principal objetivo de las noticias falsas. Pueden ser usuarios de las redes sociales en línea o de otras plataformas de noticias. Según el propósito de la noticias, las víctimas pueden ser estudiantes, votantes, padres, personas mayores, etc. También es quien toma una acción a partir del mensaje que puede ser ignorarlo, compartirlo en apoyo o compartirlo en oposición.
- *Contenido de noticias*: Se refiere al cuerpo de la noticia. Contiene tanto contenido físico (por ejemplo, título, cuerpo, multimedia) como contenido no físico (por ejemplo, tema, propósito, sentimiento)
- *Contexto social*: Indica cómo se distribuyen las noticias a través de Internet. El contexto incluye al usuario, sus redes y el patrón temporal de transmisión de la noticia a lo largo de esas redes.

Para verlo con un ejemplo, en la figura 2 se presenta la caracterización de una noticia falsa publicada en el 2017 en medio de la campaña por la presidencia de EE.UU. en donde se afirmaba que el Papa Francisco apoyaba la candidatura de Donald Trump. En este caso:

- **A** representa el Creador/Difusor: El usuario que lo comparte, en este caso Bob y Facebook.
- **B** representa el Contenido: Incluye el título de la noticia, el texto y el contenido multimedia (fotos y videos).
- **C** representa el contexto social: Incluye todas las interacciones entre otros usuarios y esta noticia (comentarios, me gusta/no me gusta, marca de tiempo)
- **D** representa las víctimas: Incluye a cualquier usuario que se involucre con la noticia por medio de las interacciones.



Figura 2: Elementos de una noticia falsa. Fuente: [5]

En relación con las noticias falsas se plantean tres grandes momentos en su ciclo de vida: creación, publicación y distribución de la noticia [6]. El momento de su distribución, es donde una noticia falsa se consolida como tal y genera mayor impacto y en este caso se resalta la importancia de la persuasión de quién comparte la noticia y en este caso del voz a voz. Las noticias falsas y la desinformación son más poderosas cuando se comparten, no sólo por las grandes fuentes noticiosas tradicionales, sino también por cualquier persona de confianza, este es el mencionado voz a voz.

A.4. Uso de técnicas de inteligencia artificial para la detección de noticias falsas

Dentro de los trabajos realizados se han definido claramente dos enfoques para abordar el problema de la detección de desinformación: la aplicación de técnicas de aprendizaje automático o *machine learning* y la aplicación de técnicas de aprendizaje profundo o *deep learning*. A continuación, se mencionan algunos de ellos.

Trabajos relacionados con técnicas de aprendizaje automático

- 2017: En [7] se presenta una revisión integral de la detección de noticias falsas en las redes sociales desde una perspectiva de minería de datos y se discuten futuras direcciones de investigación.
- 2018: En[8] se propone el uso de técnicas de aprendizaje automático como: Naïve Bayes, Neural Network, Logistic y Support Vector Machine (SVM), definiendo un mejor rendimiento a través del método de Naïve Bayes.
- 2019: En[9] se presenta una aproximación hacia la detección de noticias falsas basada en

procesamiento de lenguaje natural y técnicas de aprendizaje automático. El modelo planeando realiza un preprocesamiento que incluye bag-of-words, n-grams y vectorización.

- 2020: En[10] se evalúan algunas de las técnicas de aprendizaje automático, entre ellas las de Naïve Bayes, Random Forest, Decision Tree y SVM para problemas de clasificación de noticias de acuerdo a su temática o categoría.
- En[11] se realiza una aproximación a la detección de noticias falsas a través de métodos de machine learning incluyendo Radom Forest, SVM y Naïve Bayes. Adicionalmente se incluyen algunas aproximaciones hacia un enfoque de aprendizaje profundo.

Trabajos relacionados con técnicas de aprendizaje profundo

- 2018: En [12], se propone un método basado en redes neuronales profundas y PLN con un enfoque modular compuesto por dos partes principales; una base de conocimiento y una red neuronal profunda para aprender el estilo de la falsificación del contenido.
- 2019: En [13] se aborda la problemática de la detección de noticias falsas a través de redes neuronales recurrentes del tipo Long Short-term memory (LSTM) bidireccional.
- 2020: En [14], se enfocan en la detección temprana de noticias falsas utilizando datos observados en la etapa de propagación de la noticia y a partir de los cuales se genera el aprendizaje. Se compone de: (1) un extracto de características del perfil del usuario, (2) un mecanismo de atención que destaca respuestas importantes de los usuarios, y (3) un mecanismo de agrupación de características.
- 2019: En [15], se enfoca en la detección de posturas en la que, a partir de un reclamo y un artículo, se predice si el artículo está de acuerdo, en desacuerdo, no toma ninguna posición o no está relacionado con el reclamo.
- 2019: En [16], se propone un modelo de red neuronal convolucional (CNN) en conjunto con una arquitectura de red neuronal recurrente del tipo LSTM que aprovecha las características locales de grano grueso generadas por CNN y las dependencias de larga distancia aprendidas a través de la LSTM.

- 2020: En [17] se presenta una red neuronal denominada FAKEDETECTOR. A partir de un conjunto de características explícitas y latentes extraídas de la información textual, se construye un modelo de una red profunda del tipo DDNN (Deep Diffusive Neural Network) que permite aprender sobre artículos periodísticos, creadores y sujetos de forma simultánea.
- 2018: En [18] el objetivo fue la construcción de un clasificador que puede predecir si una noticia es falsa o no basándose únicamente en su contenido. El problema fue abordado desde una perspectiva puramente de aprendizaje profundo mediante modelos de técnica RNN (vainilla, GRU y LSTM).
- 2018: En [19] se plantea que los clasificadores basados en métodos de aprendizaje automático tienen limitaciones relacionadas con la escasez de datos, explosión de dimensiones y poca capacidad de generalización; mientras que los basados en métodos de aprendizaje profundo, facilitan la extracción de características, tienen gran capacidad de aprendizaje y una mayor precisión.
- 2020: En [20], se presenta una solución para la detección de noticias falsas que utiliza métodos de aprendizaje profundo y adicionalmente combina el análisis de sentimientos.

De acuerdo con la recopilación de investigaciones revisada, se observa que los métodos de aprendizaje automático más ampliamente utilizados son los de Random Forest, Naïve Bayes, Logistic y SVM, mientras que en el ámbito del aprendizaje profundo, las técnicas más exploradas han sido las de redes neuronales convolucionales (CNN) y las redes neuronales recurrentes del tipo LSTM. Adicionalmente, es notable que en los últimos años la tendencia de las investigaciones se está orientando hacia las técnicas de aprendizaje profundo.

A.5. Pasos generales de las técnicas aplicadas a la clasificación de noticias

Para aplicar las técnicas, es importante entender primero cómo se aplican de forma general. Los pasos generales para implementar un clasificador de noticias se resumen en la figura (Ver figura 3) y básicamente consisten en: captura de los datos, exploración, preprocessamiento, extracción de características, definición del modelo, entrenamiento del modelo, prueba del modelo, validación y evaluación del algoritmo entrenado [9], [11] y [10].

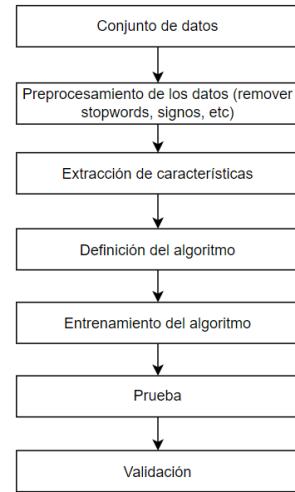


Figura 3: Proceso de análisis de noticias. Fuente: El autor

A.6. Oportunidades de investigación identificadas en el estado del arte

A continuación, se listan algunas oportunidades y enfoques de investigación identificadas en el estado del arte [15] y [21]:

- *Incorporación de información adicional para la priorización:* Medición de características adicionales como por ejemplo su valor periodístico, su potencial para influir en la sociedad, su probabilidad histórica de desinformar.
- *Uso de técnicas de aprendizaje profundo:* la capacidad de adaptación de estas técnicas ha demostrado que puede ser de utilidad.
- *Detección temprana:* La eficiencia en la verificación para identificar contenidos dignos de verificación es clave para ganar tiempo y minimizar el impacto.
- *Enfoque en el usuario y su intervención:* maximizar el compromiso del usuario en la verificación brindándole herramientas que le permitan mejorar su habilidad en la distinción de una noticia falsa de una verdadera.

III. OBJETIVOS Y METODOLOGÍA

A. Objetivo general

Investigar la viabilidad del uso de técnicas de procesamiento de lenguaje natural, aprendizaje automático y aprendizaje profundo para la clasificación y detección de desinformación dentro de titulares de

noticias y a partir de un comparativo de estas técnicas, proponer una metodología que permita detectar efectivamente indicios de desinformación relacionada con la pandemia de COVID-19 permitiendo la verificación de la información sospechosa de manera oportuna por parte del usuario y evitando su replicación sin control.

B. Objetivos específicos

- Entender las características principales de la desinformación en forma de noticias falsas, su evolución a lo largo del tiempo, sus fuentes, canales de distribución y su potencial de impacto en la sociedad.
- Explorar las referencias relacionadas con la detección de desinformación en forma de noticias falsas por medio de técnicas de inteligencia artificial e identificar sus puntos comunes y oportunidades de investigación.
- Plantear las bases y requerimientos de la metodología a diseñar, detallando sus componentes, principales características, entradas y salidas, técnicas a utilizar y modo de funcionamiento.
- Diseñar y describir en detalle la metodología propuesta y el paso a paso para ser aplicada en la detección de desinformación.
- Realizar una prueba de la metodología planteada y analizar los resultados para detectar indicios de desinformación dentro de titulares de noticias.

C. Metodología de trabajo aplicada

Se plantea seguir una metodología de trabajo basada en el pensamiento de diseño o “*design thinking*” que incluye las fases de entendimiento, definición, ideación, prototipado y prueba [22] (Ver figura 4).

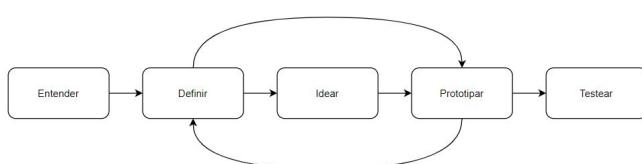


Figura 4: Metodología definida para el abordaje de la investigación. Fuente: el autor

Se ha identificado que este tipo de metodologías resultan de utilidad al momento de unirlas con el ámbito de la inteligencia artificial para realizar aproximaciones centradas en el usuario [23], lo cual

se considera de gran relevancia teniendo en cuenta que siempre va a ser el usuario quien tomará la decisión de si replica o no una noticia.

IV. CONTRIBUCIÓN

Teniendo en cuenta las oportunidades de investigación identificadas y los objetivos planteados, se propone abordar la problemática de la detección de desinformación en forma de noticias falsas a partir de una metodología enfocada en las siguientes características:

- *Que permita incorporar información adicional:* Incorporar la temática de la noticia y posibles subtemas para que el usuario cuente con más información al decidir.
- *Que permita una detección ágil:* Que la metodología pueda ser implementada para la detección de noticias de manera rápida y temprana.
- *Que permita la priorización:* Que permita priorizar su contenido para una posterior revisión a fondo.
- *Que funcione con noticias recientes:* Específicamente con la temática de la pandemia de COVID-19 con el fin de aportarle una utilidad muy puntual y relevante.
- *Que funcione con noticias en inglés:* la mayoría de las noticias inicialmente se generan en este idioma, y muchas de las técnicas y herramientas de procesamiento de lenguaje natural se encuentran también en inglés.

A. Componentes de la metodología

Se proponen tres componentes principales de la metodología (Ver figura 5):

- El primero estará relacionado con la *extracción del tema principal* de la noticia, es decir, deberá detectar si la noticia es de política, de medio ambiente, de deportes y por supuesto de salud, entre otros temas adicionales.
- Teniendo en cuenta que la mayoría de las noticias a analizar estarán dentro de la gran temática de “salud”, se propone que el segundo componente sea el *subtema de la noticia*, es decir, si estamos hablando de noticias de salud, de qué tema específicamente se trata la noticia, por ejemplo: salud pública, vacunas, epidemias, entre otros.

- El tercer gran componente de la metodología será la *predicción de la alerta* de posible contenido con desinformación. Esta parte considerará como entrada tanto el texto del titular de la noticia como el resultado de la clasificación de la temática y subtemática de la noticia.

En este sentido la metodología incluirá tres modelos de predicción. Los dos primeros modelos estarán relacionados con la adquisición de la información adicional de la noticia, en este caso la temática y la subtemática y el tercer modelo corresponderá a la generación de la alerta de detección de posible contenido con desinformación.

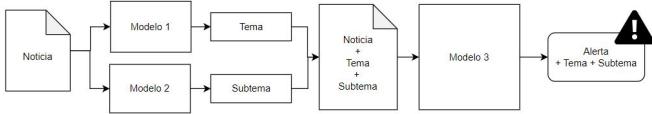


Figura 5: Componentes de la metodología. Fuente: el autor

B. Recolección y preparación de los datos

Para el desarrollo de la metodología se tomaron en cuenta los siguientes conjuntos de datos relacionados con la pandemia de COVID-19:

- FakeCovid [24]: Este conjunto de datos contiene artículos recolectados desde Poynter y Snopes. Incluye 5182 artículos en varios idiomas, que han circulado en 105 países y que han sido verificados por Fact Checkers. El intervalo de fechas va desde enero hasta mayo de 2020.
- COVID19FN [25]: Recopilación de alrededor de 2800 artículos en varios idiomas etiquetados directamente recolectados desde sitios webs de Fact Checkers desde enero hasta junio de 2020.

Estos conjuntos de datos se fusionaron en uno solo para contar con una mayor cantidad de ejemplos que facilitaran el entrenamiento de los modelos.

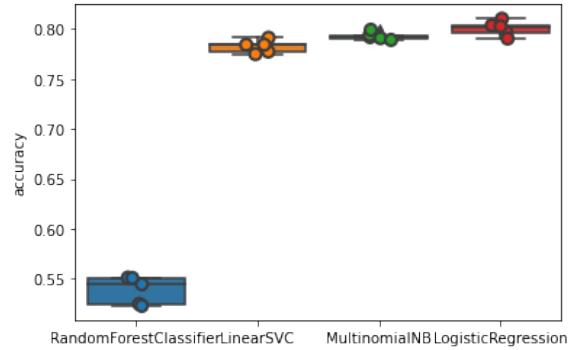
Adicionalmente se realizó un scraping de noticias en Reuters.com para las categorías: Politics, Health, Enviroment, Technology, Finance, Lifestyle, Science y Sports. Estos datos fueron clave para la generación del modelo de clasificación de la temática principal de la noticia.

C. Desarrollo del modelo de predicción del tema

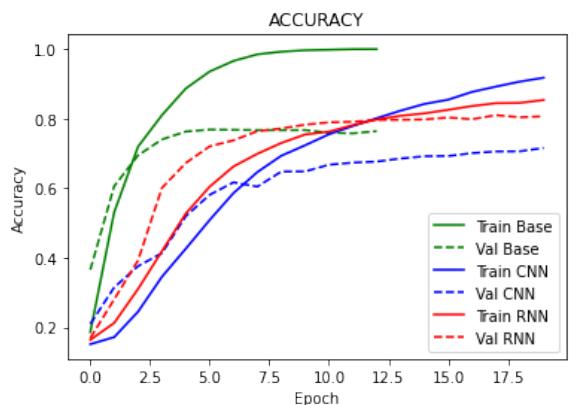
Como primer componente de la metodología se desarrolló un modelo de predicción de la categoría

del titular de la noticia. En este caso el resultado de aplicar el modelo retornará una etiqueta definida para la temática de la noticia que podrá estar entre las siguientes: “*enviroment, sports, lifestyle, politics, technology, health, science, finance*”.

En línea con lo revisado en el estado del arte, se realizó una comparación de las técnicas de aprendizaje automático (Random Forest, Naïve Bayes, Logistic y SVM) y de aprendizaje profundo (fully connected, redes neuronales convolucionales (CNN) y las redes neuronales recurrentes del tipo LSTM).



(a) Comparación de modelos de aprendizaje automático



(b) Comparación de modelos de aprendizaje profundo

Figura 6: Comparación de modelos tema

C.1. Modelos de aprendizaje automático

En la figura 6a se confirma que el clasificador *Random Forest* no alcanzó buenos resultados, mientras que el *Logistic* obtuvo el mejor desempeño, con una precisión superior a 0.8.

C.2. Modelos de aprendizaje profundo

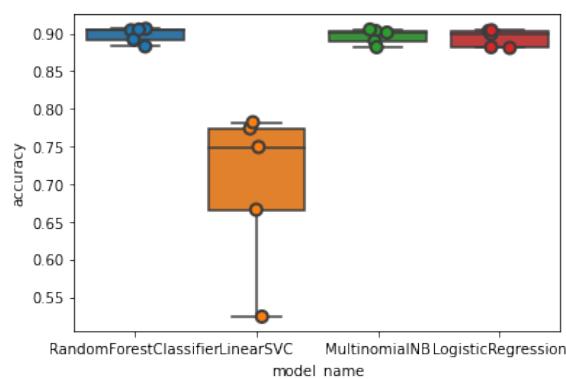
En la figura 6b se detalla la comparación de la precisión y el *loss* de los tres modelos generados. De acuerdo con este resultado es claro que el modelo más preciso fue el desarrollado con redes neuronales recurrentes LSTM, alcanzando un *accuracy* de alrededor de 0.8.

C.3. Comparación

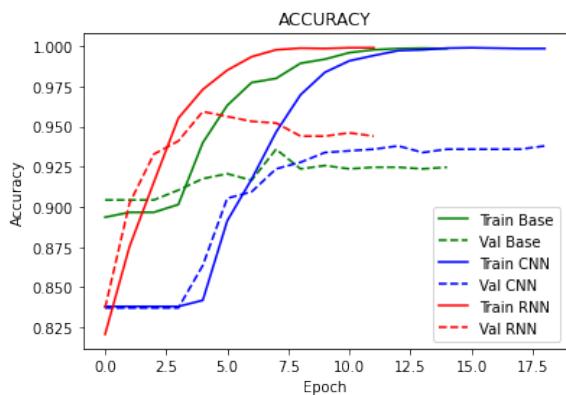
Se observa que tanto el clasificador *logistic* como el de redes neuronales recurrentes alcanzan una precisión de alrededor de 0.8. Considerando que el clasificador *logistic* es el más liviano computacionalmente, se seleccionará este método como parte de la metodología planteada.

D. Desarrollo del modelo de predicción de la alerta

El modelo de predicción de la alerta es el segundo gran componente de la metodología propuesta. En este caso el resultado de aplicar el modelo retornará una etiqueta binaria sobre los datos que indica si detecta o no una alta probabilidad de desinformación dentro del titular de la noticia. Se realizó una comparación de diferentes abordajes para este modelo a través de las mismas técnicas de aprendizaje automático y aprendizaje profundo que se tuvieron en cuenta para el modelo de predicción del tema.



(a) Comparación de modelos de aprendizaje automático



(b) Comparación de modelos de aprendizaje profundo

Figura 7: Comparación de modelos alerta

D.1. Modelos de aprendizaje automático

En la figura 7a se observa que los clasificadores Random Forest, Naïve Bayes y Logistic alcanzan un resultado cercano a un 0.9 de precisión, mientras

que el SVC obtiene el menor rendimiento con un máximo de 0.75.

D.2. Modelos de aprendizaje profundo

En la figura 7b es claro que el modelo más preciso fue el desarrollado con redes neuronales recurrentes, alcanzando una precisión superior a 0.9.

D.3. Comparación

En este caso los modelos basados en redes neuronales profundas alcanzan un rendimiento más alto que los basados en técnicas de aprendizaje automático. El método que alcanza un mayor rendimiento es el de redes neuronales recurrentes LSTM, llegando a un 0.94 de precisión. Esta técnica será la seleccionada para su implementación dentro de la metodología propuesta.

E. Desarrollo del modelo de extracción del subtema

Si bien para la clasificación de la noticia en temas se contaba con una etiqueta de entrenamiento en los datos relacionada con su temática, en este caso, no se cuenta con una etiqueta para el subtema. Por esta razón este tipo de problema se abordará a través de una técnica de aprendizaje automático no supervisada.

Una de las técnicas usadas para la detección de temas de forma no supervisada es la *Asignación Latente de Dirichlet o Latent Dirichlet Allocation (LDA)*, la cual permite agrupar elementos de un conjunto de datos a partir de parte de los datos que son semejantes. Ejemplos de este tipo de técnicas se pueden encontrar en trabajos directamente relacionados con la extracción de temáticas de noticias, como por ejemplo [26] y [27].

En este sentido se plantea un modelo de extracción de la subtemática que parte de el conjunto de datos relacionado directamente con la pandemia de COVID-19 y agrupa cada uno de los titulares en grupos relacionados. Un ejemplo de dos grupos resultantes podría ser: la *subtemática vacuna* cuyas palabras relacionadas podrían ser: vacuna, dosis, placebo, reacción. Mientras que si el otro grupo está relacionada con la *subtemática rebrote*, las palabras asociadas podrían ser: casos, cuarentena, medidas, pico.

Para implementar esta técnica se utilizó *Gensim*¹, una librería de código abierto para el modelado de temas no supervisados. En la figura 8 se presenta el resultado de generar el modelo a partir del conjunto de datos y generar una agrupación

¹<https://radimrehurek.com/gensim/>

```
[['test', 'health', 'hospit', 'clinic', 'australia'],
['mask', 'kill', 'wear', 'chines', 'hand'],
['infect', 'case', 'health', 'report', 'pictur'],
['case', 'australia', 'health', 'pictur', 'australian'],
['trump', 'presid', 'state', 'death', 'unit'],
['post', 'facebook', 'novel', 'show', 'share'],
['toilet', 'paper', 'australia', 'health', 'custom']]
```

Figura 8: Resultado del modelo LDA para 7 subtemas y 5 palabras. Fuente: el autor

de 7 subtemáticas. Analizando los grupos resultantes, la primera temática podría estar relacionada con las pruebas y la hospitalización pues contiene las palabras “test”, “health”, “hospit”, “clinic”. La segunda categoría podría estar relacionada con los métodos preventivos pues contiene las palabras “mask”, “wear”, “hand”. El tercer grupo al contener las palabras “infect”, “case”, “report”, podría estar relacionado con el reporte de casos de infectados que día a día realizan todos los países. La cuarta categoría puede corresponder a casos específicamente relacionados con Australia, ya que incluye las palabras “case”, “australia”, “australian”. En cuanto al quinto grupo claramente se trataría de una categoría relacionada con los EE.UU. y su presidente, pues contiene palabras como “unit”, “state”, “Trump”, “presi”. La sexta categoría podría estar relacionada con publicaciones en redes sociales, pues tiene palabras como “post”, “facebook”, “share”, mientras que el último grupo podría estar relacionado con algunos casos sonados relacionados con el coronavirus, como el de la escasez de papel higiénico por el pánico generado muy al inicio de la pandemia.

A continuación, se resumen los diferentes subtemas resultantes y sus palabras relacionadas:

- Pruebas: test, health, hospit, clinic, australia
- Cuidados: mask, kill, wear, chines, hand
- Reportes: infect, case, health, report, pictur
- Australia: case, australia, health, pictur, australian
- EE.UU.: trump, presid, state, death, 'unit'
- Redes sociales: post, facebook, novel, show, share
- Casos: toilet, paper, australia, health, custom

F. Generación de la metodología unificada

Una vez realizada la comparación de las diferentes técnicas de clasificadores, se tiene una idea de las técnicas que lograron los mejores rendimientos. En este caso para el componente de predicción de la temática de la noticia se ha seleccionado un clasificador tipo *Logistic*, mientras que para el modelo de

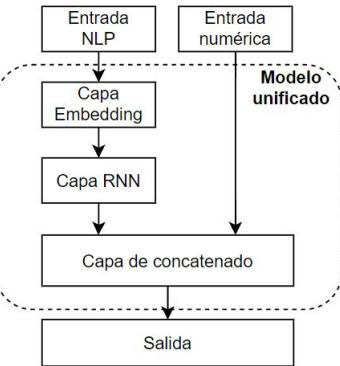


Figura 9: Modelo con múltiples entradas y una salida. Fuente: el autor

predicción de la alerta se selecciona un modelo basado en redes neuronales recurrentes de tipo *Long Short-term Memory (LSTM)*.

El siguiente paso es unificar estos componentes en una única metodología. Para este propósito se realizará una aproximación a un modelo que permita combinar múltiples entradas y resultar en una única salida como los propuestos en [28] y [29].

Para el escenario específico de la investigación, el principal problema sin duda está relacionado con clasificación de texto, sin embargo, teniendo en cuenta la necesidad de incorporar información adicional, se considerará una segunda entrada del modelo para la temática y subtemática de la noticia; información adicional que se puede considerar como un vector de metadatos asociados.

De acuerdo a lo propuesto en [28] y [29], existen dos acercamiento para abordar el problema de las múltiples entradas. El primero es sencillamente concatenar estos metadatos a los textos para que sean considerados en los *embeddings* y bolsas de palabras generadas en el preprocesamiento. Sin embargo este tipo de abordaje puede ser simplista y no aprovecha al máximo el potencial de la información adicional, ya que las nuevas palabras pueden diluirse frente al cuerpo del titular de la noticia. El segundo acercamiento, considera múltiples vectores de entrenamiento para el modelo, para así embeber por completo los metadatos dentro del modelo generado y entrenado. Para este caso, la distribución de las entradas y salidas del modelo se resume en la figura 9).

G. Arquitectura de la metodología

En la figura 10 se define la arquitectura de la metodología propuesta la cual incluye los siguientes componentes:

- Componente de predicción de la temática: Modelo basado en un clasificador del tipo logistic.

- Componente de extracción de subtemas: Modelo basado en un clasificador no supervisado del tipo Latent Dirichlet Allocation (LDA).

- Componente de predicción de la alerta: Modelo basado en una red neuronal recurrente del tipo LSTM con múltiples entradas.

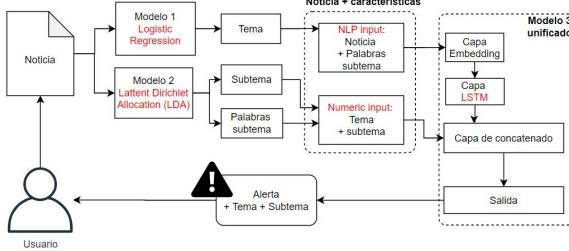


Figura 10: Planteamiento final de la metodología unificada.
Fuente: el autor

H. Desarrollo del modelo

Teniendo en cuenta la figura 10, a continuación se detalla el paso a paso de su desarrollo, referenciando algunos apartes de código relevantes².

Paso 1: Carga de conjunto de datos:

Se carga del conjunto de datos y se realiza un preprocessamiento inicial para eliminar registros duplicados, datos nulos e incostistencias.

Paso 2: Carga de modelos: Se cargan los modelos probados que alcanzaron los mejores resultados en las comparaciones realizadas. Los modelos previamente han sido guardados en formato .pkl o .h5. También se cargan los correspondientes tokenizadores o vectorizadores.

Paso 3: Aplicación de modelo para clasificar la temática (Logistic Regression): Se aplica el modelo de extracción de la temática y se guarda el resultado para cada uno de los titulares en una nueva columna del dataframe.

Paso 4: Aplicación de modelo para extraer la subtemática (LDA): Se aplica el modelo de extracción de la subtemática sobre la totalidad de los titulares incluidos en el dataframe almacenando los resultados en dos columnas adicionales, en la primera se guarda el ID del subtema extraído, y en la segunda se guardan las palabras claves asociadas. En las siguientes líneas de código se detalla un extracto del proceso.

²El código completo podrá ser consultado en: <https://github.com/jorgecif/CovidDisinformationDetection/>

```

1 # Funcion para extraer subtemáticas
2 def topics_lda(documento):
3     unseen_document=documento
4     bow_vector = dictionary.doc2bow(preprocess(unseen_document))
5     prediction_lda=lda_model[bow_vector]
6     probs=[]
7     for i in range(0, len(prediction_lda)):
8         probs.append(prediction_lda[i][1])
9     max_probs=max(probs)
10    for i in range(0,len(prediction_lda)):
11        if max_probs==prediction_lda[i][1]:
12            position=i
13            break
14    return position
15 # Aplico modelo LDA a conjunto de datos del dataframe
16 datos_revisar=datos_trabajo[["Text"]]
17 list_result_id=[]
18 list_result_words=[]
19 for i in range(0,len(datos_revisar)):
20     id_predict=topics_lda(datos_revisar[i])
21     list_result_id.append(id_predict)
22     list_result_words.append(str(topics[id_predict]))
23 # Creo dataframe con columna adicional de predicción y palabras
24 datos_trabajo_pred["pred_topics_id"] = list_result_id
25 datos_trabajo_pred["pred_topics_words"] = list_result_words
  
```

Código 1: Detalle de aplicación del modelo de extracción de la subtemática. Fuente: el autor

Paso 5: Planteamiento de la red LSTM con múltiples entradas: Inicialmente se realiza la extracción de las columnas que servirán de entrada para el modelo, en este caso *Input 1* que tiene asociada toda la información de texto, y la *Input 2*, que tiene asociada toda la información de metadatos.

La definición de la red LSTM planteada incluye dos entradas: *Input1: nlp-input* para *xtrain1*, e *Input 2: meta-input* para *xtrain2*. Las dos entradas se combinan dentro de la arquitectura de la red por medio de una capa de concatenado que incluye Keras, y continúa con capas densas que generan una predicción unificada de detección de desinformación considerando las dos entradas (Ver código 2 y figura 11).

```

1 # Extraigo datos de dataframe
2 corpus_trabajo = datos_analizar[["text_topics"]] # Datos texto - Input 1
3 meta = datos_analizar[["metadatos"]] # Metadatos numéricos - Input 2
4 results_trabajo = datos_analizar[["label"]].map(category_dict) # Predicción
# Tokenización
5 corpus_trabajo = datos_analizar[["text_topics"]]
6 sequences = tokenizer.texts_to_sequences(corpus_trabajo.values)
7 X = pad_sequences(sequences, maxlen=max_len)
8 meta=meta.values.tolist()
9 meta_arr = np.array(meta)
10
11 # Train - Test split
12 x_train1,x_test1,y_train,y_test = train_test_split(X, results_trabajo,
13 test_size=0.2, random_state=88 ) # (Input 1)
14 x_train2,x_test2,y_train2,y_test2 = train_test_split(meta_arr,
15 results_trabajo, test_size=0.2, random_state=88 ) # (Input 2)
# Definición de la red
16 nlp_input = Input(shape=(seq_length,), name='nlp_input')
17 meta_input = Input(shape=(2,), name='meta_input')
18 emb = Embedding(output_dim=emb_dim, input_dim=embedding_size,
19 input_length=seq_length)(nlp_input)
20 nlp_out = (LSTM(64, dropout=0.7, recurrent_dropout=0.7,
21 kernel_regularizer=regularizers.l2(0.01)))(emb)
22 x = concatenate([nlp_out, meta_input])
23 x = Flatten()(x)
24 x = Dropout(0.3)(x)
25 x = Dense(32, activation='relu')(x)
26 x = Dense(1, activation='sigmoid')(x)
model = Model(inputs=[nlp_input , meta_input], outputs=[x])
# Compilación
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['acc'])
  
```

Código 2: Detalle de planteamiento de la red LSTM con múltiples entradas. Fuente: el autor

Paso 6: Entrenamiento de la red: Para el entrenamiento de la red, por tratarse de una arquitectura de múltiples entradas, el comando *model.fit* deberá incluirlos. En el siguiente apartado de código se puede observar esta particularidad.

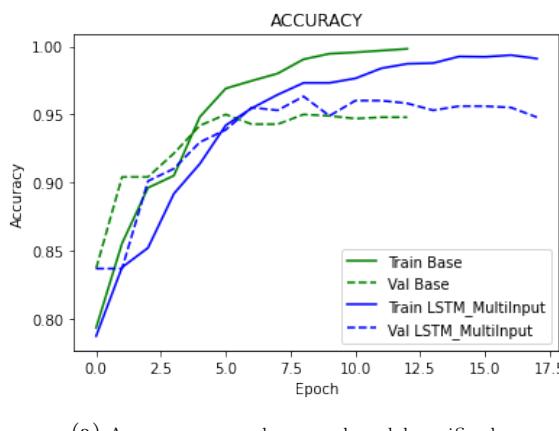
Model: "functional_35"				
Layer (type)	Output Shape	Param #	Connected to	
nlp_input (Inputlayer)	[None, 300]	0		
embedding_24 (Embedding)	(None, 300, 128)	1280000	nlp_input[0][0]	
lstm_24 (LSTM)	(None, 64)	49408	embedding_24[0][0]	
meta_input (InputLayer)	[None, 2]	0		
concatenate_19 (Concatenate)	(None, 66)	0	lstm_24[0][0] meta_input[0][0]	
flatten_22 (Flatten)	(None, 66)	0	concatenate_19[0][0]	
dropout_21 (Dropout)	(None, 66)	0	flatten_22[0][0]	
dense_46 (Dense)	(None, 32)	2144	dropout_21[0][0]	
dense_47 (Dense)	(None, 1)	33	dense_46[0][0]	
Total params:	1,331,585			
Trainable params:	1,331,585			
Non-trainable params:	0			

Figura 11: Resumen de la red LSTM con múltiples entradas planteada. Fuente: el autor

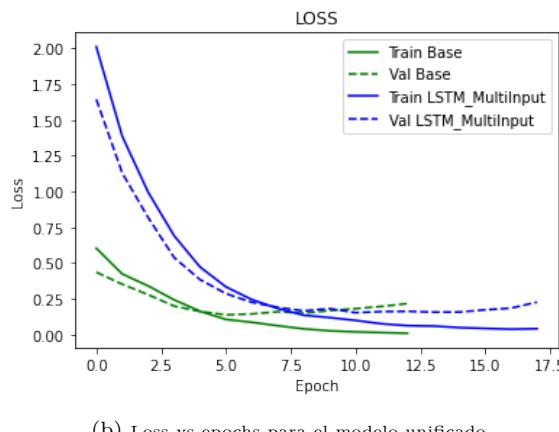
```
1 # Entrenamiento
2 historyFinal=model.fit({'nlp_input': x_train1, 'meta_input': x_train2},
   y_train, epochs=epochs, batch_size=batch_size, validation_split=0.2,
   callbacks=[EarlyStopping(monitor='val_loss', patience=7, min_delta
 =0.0001)])
```

Código 3: Entrenamiento de la red LSTM con múltiples entradas planteada. Fuente: el autor

El rendimiento del modelo entrenado en términos de su *accuracy* y *loss*, alcanza valores cercanos a un 0.95 y 0.1, respectivamente.



(a) Accuracy vs epochs para el modelo unificado



(b) Loss vs epochs para el modelo unificado

Figura 12: Comparación de modelo unificado LSTM de múltiples entradas con modelo base de única entrada

V. RESULTADOS

En la figura 12 se presenta una comparación del rendimiento del modelo final LSTM de múltiples entradas con un modelo base LSTM de entrada única. Los resultados evidencian una mejora en el rendimiento del modelo de múltiple entrada propuesto en términos de *accuracy* y *loss*, obteniendo niveles superiores a 0.95 e inferiores a 0.25, respectivamente.

A. Implementación del modelo

El modelo final desarrollado se implementó a través de una interfaz sencilla que le permite al usuario copiar y pegar un titular de una noticia y aplicar el modelo para obtener una predicción. La interfaz le devuelve al usuario la etiqueta del tema, el subtema en forma de palabras clave, la predicción de la alerta y su probabilidad. En la figura 13 se presenta el diseño final de interfaz y un ejemplo de su respuesta al aplicar el modelo al hacer clic en el botón “*Predecir*”.



Figura 13: Detalle de la interfaz desarrollada. Fuente: el autor

Se implementó una arquitectura basada en Python y Flask, agregando todo el código en los archivos *app.py* e *index.html* para desplegar la interfaz como una aplicación web en el navegador. En la figura 14 se define la arquitectura de la herramienta, incluyendo sus elementos en la capa de datos, en la capa de backend y en la capa de frontend.

Luego de desarrollar la interfaz se realizó su despliegue en Heroku directamente desde GitHub. La aplicación de prueba se encuentra desplegada en la url: <https://coviddisinformation.herokuapp.com/> y

el código de la aplicación se puede consultar en: <https://github.com/jorgecif/CovidDisinformationDetection>.

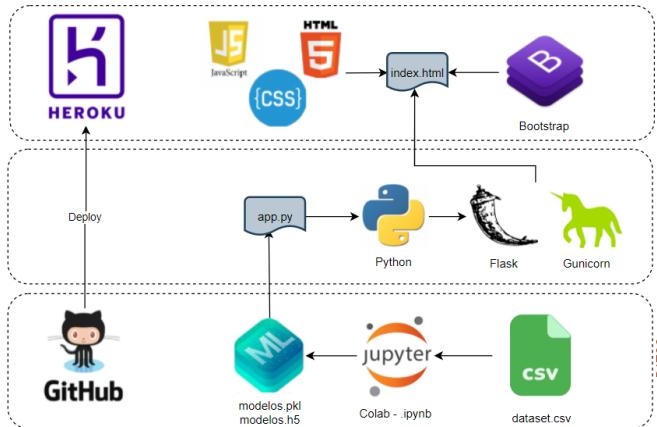


Figura 14: Arquitectura del modelo implementado. Fuente: el autor

B. Pruebas y validación

Con el fin de llevar al límite el modelo desarrollado para identificar sus fortalezas y debilidades, se plantearon las siguientes pruebas: pruebas con noticias de 2020, pruebas con noticias de 2021, pruebas con mitos sobre el COVID-19, pruebas con buenas prácticas para combatir el COVID-19, pruebas con hechos creados artificialmente, pruebas con negaciones o cambios de sentido.

(a) Matriz de confusión pruebas noticias 2020. Fuente: El autor	(b) Matriz de confusión pruebas noticias 2021. Fuente: El autor
(c) Matriz de confusión pruebas mitos. Fuente: El autor	(d) Matriz de confusión pruebas buenas prácticas. Fuente: El autor
(e) Matriz de confusión pruebas hechos creados. Fuente: El autor	(f) Matriz de confusión pruebas negaciones. Fuente: El autor

Cuadro 1: Matriz de confusión de pruebas finales

En el cuadro 1 se han generado las matrices de confusión de las pruebas realizadas. A continuación, se incluye el detalle de cada una de ellas.

B.1. Pruebas con noticias del año 2020

Se consideraron titulares de noticias incluidas dentro del rango de fechas del conjunto de datos, es decir aproximadamente hasta el mes de septiembre de 2020, con el fin de verificar el funcionamiento del modelo con las generalidades, términos y hechos ocurridos hasta la fecha de su entrenamiento.

En el cuadro 2 se detalla el texto de los titulares probados, luego de aplicar el modelo a cada

ID	Fuente	Texto	L	Tema	Subtema	P	D	Probabilidad
1	Dataset	Uganda is giving out 122GB of data to customers for free in response to COVID-19	0	Health	['trump', 'presid', 'state', 'death', 'unit']	0	0	[0.9999374]
2	Dataset	Bill Gates told us about the coronavirus in 2015	1	Health	['case', 'australia', 'health', 'pictur', 'australian']	1	0	[0.01032567]
3	Dataset	A tweet by Pakistani journalist Saadia Afzaal claiming that China has developed a COVID-19 vaccine	0	Health	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.9999794]
4	Dataset	Germany gave medical protection equipment like masks to China, now its missing in Germany	1	Health	['mask', 'kill', 'wear', 'chines', 'hand']	0	1	[0.9938691]
5	Dataset	Photos of Italian man committing suicide after he lost his entire family to COVID-19	0	Health	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.99996483]
6	Dataset	Harvard professor was arrested for creating and selling the coronavirus	0	Health	['toilet', 'paper', 'australia', 'health', 'custom']	0	0	[0.999812]
7	Dataset	Madagascar does not have any cases of the coronavirus	0	Health	['case', 'australia', 'health', 'pictur', 'australian']	0	0	[0.979228]
8	Dataset	COVID-19 is a bacterium that is easily treated with aspirin or a coagulant	0	Health	['test', 'health', 'hos-pit', 'clinic', 'australia']	0	0	[0.9554566]
9	Dataset	Trump Suspends Europe Travel, Announces New Economic Measures	1	Env	['trump', 'presid', 'state', 'death', 'unit']	1	0	[0.02297539]
10	Dataset	Social media users have shared a photo that claims to show a "Center for Global Human Population Reduction" affiliated with the Bill & Melinda Gates Foundation	0	Tech	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.999998]

Cuadro 2: Pruebas con noticias de 2020, incluidas en el conjunto de datos (L: Label, P: Predicción, D: Diferencia, 0: FALSE, 1: TRUE). Fuente: El autor

uno de los encabezados se comprueba que el modelo predice correctamente la mayoría de este tipo de titulares de noticias. De los 10 titulares revisados, solamente uno tuvo un resultado diferente a la etiqueta real, en este caso se trata del titular: “*Germany gave medical protection equipment like masks to China, now its missing in German*” el cual está etiquetado como verdadero y es predicho como falso. En la matriz de confusión del cuadro 1a se registra el resultado de la prueba.

B.2. Pruebas con noticias del año 2021

Se seleccionaron titulares de noticias muy recientes, asegurando que estuvieran por fuera de las fechas incluidas dentro del conjunto de datos del modelo (Ver cuadro 3). El objetivo de esta prueba es comprobar que el modelo sigue siendo vigente a pesar de posibles cambios en el contexto de la temática, en este caso la pandemia de COVID-19.

ID	Fuente	Texto	L	Tema	Subtema	P	D	Probabilidad
1	Boomlive	COVID-19 vaccine do not eliminate the virus or stop the virus from transmitting	0	Health	['post', 'facebook', 'novel', 'show', 'share']	1	1	[0.08537072]
2	TLIndian	Pakistan Prime Minister Imran Khan has said, If Pakistan develops coronavirus vaccine then they will not give it to India or Israel.	0	Health	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.9995562]
3	Politifact	Says the new coronavirus vaccines contain toxic ingredients and are more dangerous than getting COVID-19.	0	Health	['trump', 'presid', 'state', 'death', 'unit']	0	0	[0.9998156]
4	TLIndian	Pope Francis said that Covid-19 vaccine will be required to enter heaven.	0	Health	['infect', 'health', 'report', 'pictur']	0	0	[0.9999604]
5	Forbes	China Deploys Anal Swab Tests To Detect High-Risk Covid-19 Cases	0	Health	['infect', 'health', 'report', 'pictur']	0	0	[0.18232581]
6	TLIndian	Countries in Europe have imposed lockdown against the second wave of the novel coronavirus.	1	Health	['infect', 'health', 'report', 'pictur']	0	1	[0.9998668]
7	Politifact	Says the new coronavirus vaccines contain toxic ingredients and are more dangerous than getting COVID-19.	0	Health	['trump', 'presid', 'state', 'death', 'unit']	0	0	[0.9998156]
8	Fullfact	The Covid vaccine will make you infertile.	0	Tech	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.9381714]
9	Fullfact	Covid vaccines contain aborted babies.	0	Health	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.9267428]
10	Boomlive	World Health Organization (WHO) ranked Sri Lanka fifth in a table of countries responses to the coronavirus pandemic.	0	Health	['case', 'australia', 'health', 'pictur', 'australian']	0	0	[0.9992931]

Cuadro 3: Pruebas con noticias de 2021 (L: Label, P: Predicción, D: Diferencia, 0: FALSE, 1: TRUE). Fuente: El autor

Los titulares de las noticias se encontraron por medio de la herramienta Google Fact Check Explorer³, una nueva iniciativa de Google cuyo objetivo es recopilar hechos ya verificados y ofrecer una forma sencilla de para buscarlos por medio de palabras claves. De los 10 encabezados revisados, dos tuvieron un resultado diferente a la etiqueta real, el titular: “*COVID-19 vaccine do not eliminate the virus or stop the virus from transmitting*” y el titular: “*Countries in Europe have imposed lockdown against the second wave of the novel coronavirus.*”. La confusión del modelo se podría explicar debido a que en los dos casos se incluyen escenarios nuevos dentro del contexto actual que eran inexistentes en el momento en el que se entrenó el modelo, como lo son el hecho de que ya existe una vacuna desarrollada y que se esté presentando una nueva ola de una nueva variante del virus. En la matriz de confusión del cuadro 1b se registra el resultado de la prueba.

B.3. Pruebas con mitos sobre el COVID-19

Uno de los casos de uso de mayor utilidad del modelo desarrollado es la verificación de mitos relacionados con la pandemia de COVID-19. En esta prueba se seleccionaron algunos mitos recopilados por la Organización Mundial de la Salud [30] y la Facultad de Medicina de la Universidad Johns Hopkins [31].

Se recolectaron en total 18 mitos (Ver cuadro 4) a los cuales se les aplicó el modelo, resultando en solo un par de casos en los que la predicción fue diferente a la etiqueta real. Uno de los casos se trata de la afirmación: “*A vaccine to cure COVID-19 is available*”, la cual evidentemente es verdadera, sin embargo obtiene una predicción falsa.

Si tenemos en cuenta las fecha límites del conjunto de datos de entrenamiento del modelo, es evidente que hasta septiembre de 2020 efectivamente no existía ninguna vacuna disponible, por lo tanto la afirmación en ese contexto sería falsa, tal como lo predijo el modelo. En la matriz de confusión del cuadro 1c se registra el resultado de la prueba.

B.4. Pruebas con buenas prácticas para mitigar el COVID-19

Teniendo en cuenta que hasta el momento la mayoría de las pruebas se han realizado con titulares falsos, se plantea esta prueba para verificar el comportamiento del modelo ante afirmaciones o enunciados con una etiqueta evidentemente verdadera. Para esto se realiza una búsqueda de buenas prácticas ya evidenciadas para reducir el riesgo de con-

ID	Fuente	Texto	L	Tema	Subtema	P	D	Probabilidad
1	Hopkins	You can get a face mask exemption card so you don't need to wear a mask.	0	Tech	['mask', 'kill', 'wear', 'chines', 'hand']	0	0	[0.9065571]
2	Dataset	Turnmeric And Lemon Help Fight Against coronavirus.	0	Health	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.99547905]
3	Hopkins	You can protect yourself from COVID-19 by injecting, swallowing, bathing in or rubbing onto your body bleach, disinfectants or rubbing alcohols.	0	Tech	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.97236454]
4	Hopkins	A vaccine to cure COVID-19 is available	1	Health	['post', 'facebook', 'novel', 'show', 'share']	0	1	[0.99695635]
5	Hopkins	The new coronavirus was deliberately created or released by people.	0	Health	['trump', 'presid', 'state', 'death', 'unit']	0	0	[0.99941945]
6	Hopkins	Ordering or buying products shipped from overseas will make a person sick.	0	Tech	['toilet', 'paper', 'australia', 'health', 'custom']	0	0	[0.9988524]
7	WHO	Can Covid-19 be transmitted through goods produced in countries where there is ongoing transmission?	0	Health	['toilet', 'paper', 'australia', 'health', 'custom']	0	0	[0.99999475]
8	WHO	Can Covid-19 be transmitted through mosquitoes?	0	Health	['infect', 'case', 'health', 'report', 'picture']	0	0	[0.98874116]
9	WHO	How can we be sure that our clothes don't spread coronavirus?	0	Health	['mask', 'kill', 'wear', 'chines', 'hand']	1	1	[0.46550933]
10	WHO	Can drinking alcohol help prevent Covid-19?	0	Health	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.99845594]
11	WHO	Is it true that Covid-19 is transmitted in cold climate and not in hot and humid climate?	0	Sports	['infect', 'case', 'health', 'report', 'picture']	0	0	[0.964204]
12	WHO	Can digital thermometers be 100 % effective in detecting Covid-19 patients?	0	Health	['test', 'health', 'hospit', 'clinic', 'australia']	0	0	[0.93550766]
13	WHO	Can UV bulbs used for disinfecting be used to kill Covid-19 on our body?	0	Health	['mask', 'kill', 'wear', 'chines', 'hand']	0	0	[0.9999908]
14	WHO	Can spraying alcohol or chlorine on your body kill the virus inside?	0	Env	['mask', 'kill', 'wear', 'chines', 'hand']	0	0	[0.99931556]
15	WHO	Can eating garlic prevent covid-19?	0	Health	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.9816327]
16	WHO	Can Pneumonia vaccine prevent Covid-19?	0	Health	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.7597852]
17	WHO	Can rinsing your nose regularly with saline solution prevent Covid-19?	0	Health	['infect', 'case', 'health', 'report', 'picture']	0	0	[0.9985318]
18	WHO	Is there any drug that can prevent and treat Covid-19?	0	Health	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.9999687]

Cuadro 4: Pruebas con mitos del Coronavirus encontrados en diversas fuentes (L: Label, P: Predicción, D: Diferencia, 0: FALSE, 1: TRUE). Fuente: El auto

traer el coronavirus, citando principalmente como fuente a la Organización Mundial de la Salud.

En el cuadro 5 se incluye una selección de 10 buenas prácticas populares en la prevención del contagio del coronavirus y que han sido ampliamente divulgadas por la Organización Mundial de la Salud. Luego de aplicar el modelo a cada una de ellas se comprueba que el modelo predice correctamente la totalidad de este tipo de enunciados. En la matriz de confusión del cuadro 1d se registra el resultado de la prueba.

B.5. Pruebas con hechos creados artificialmente

Esta prueba pretende ponerse en los zapatos de un generador de desinformación para simular la creación de un conjunto de posibles noticias falsas relacionadas con la pandemia de COVID-19.

Se generaron un total de 11 titulares de noticias evidentemente falsas, las cuales se detallan en el cuadro 6. Luego de aplicar el modelo se comprueba que el modelo predice correctamente la mayoría de este tipo de titulares de noticias, sin embargo en el titular: “*First case of a dog with coronavirus detected in Australia*” no genera una alerta y es predicho como un hecho verdadero o sin contenido de desin-

³<https://toolbox.google.com/factcheck/explorer>

ID	Fuente	Texto	L	Tema	Subtema	P	D	Probabilidad
1	WHO	The principal symptoms of COVID are fever, cough and tiredness	1	Health	['infect', 'health', 'pictur']	1	0	[0.04208112]
2	WHO	Keep the distance from others is a good to reduce the risk of coronavirus	1	Health	['case', 'australia', 'health', 'australian']	1	0	[0.01383418]
3	WHO	Washing hands often with antibacterial soap and water is imperative to protect yourself from covid.	1	Tech	['mask', 'kill', 'wear', 'chines', 'hand']	1	0	[0.2492171]
4	WHO	Animals and pets cannot catch coronavirus or transmit the virus to humans	1	Health	['infect', 'case', 'health', 'pictur']	1	0	[0.01292658]
5	WHO	Touching your face can lead to a fast transfer covid into the body	1	Politics	['infect', 'case', 'health', 'pictur']	1	0	[0.20195228]
6	WHO	Frequently disinfect surfaces such as door knobs, equipment handles, check-out counters is a best practice against covid	1	Politics	['infect', 'case', 'health', 'pictur']	1	0	[0.00035217]
7	WHO	Wear a mask is a good practice against covid	1	Politics	['mask', 'kill', 'wear', 'chines', 'hand']	1	0	[0.01882103]
8	WHO	Avoid poorly ventilated spaces and crowded spaces is a good practice to prevent coronavirus	1	Health	['case', 'australia', 'health', 'australian']	1	0	[0.01086292]
9	WHO	People with comorbidities have the highest risk of contracting covid virus	1	Health	['case', 'australia', 'health', 'australian']	1	0	[0.22701627]
10	WHO	Clean and disinfect frequently touched surfaces daily is a good practice against covid	1	Tech	['infect', 'case', 'health', 'pictur']	1	0	[0.00020418]

Cuadro 5: Pruebas con buenas prácticas recopiladas por la Organización Mundial de la Salud (L: Label, P: Predicción, D: Diferencia, 0: FALSE, 1: TRUE). Fuente: El autor

ID	Fuente	Texto	L	Tema	Subtema	P	D	Probabilidad
1	WHO	Washing hands often with antibacterial soap and water is imperative to protect yourself from covid.	1	Tech	['mask', 'kill', 'wear', 'chines', 'hand']	1	0	[0.2492171]
2	El autor	Washing hands often with antibacterial soap and water is not imperative to protect yourself from covid.	0	Tech	['mask', 'kill', 'wear', 'chines', 'hand']	1	1	[0.21670559]
3	WHO	The principal symptoms of COVID are fever, cough and tiredness	1	Health	['infect', 'health', 'report']	1	0	[0.04208112]
4	El autor	These are not the main symptoms of COVID are fever, cough and tiredness	0	Health	['infect', 'case', 'report']	0	0	[0.5414618]
5	WHO	Animals and pets cannot catch coronavirus or transmit the virus to humans	1	Health	['infect', 'health', 'report']	1	0	[0.01292658]
6	El autor	Animals and pets can get coronavirus and can transmit the virus to humans	0	Env	['infect', 'health', 'report']	1	1	[0.04835919]
7	Hopkins	You can protect yourself from COVID-19 by injecting, swallowing, bathing in or rubbing onto your body bleach, disinfectants or rubbing alcohols.	0	Tech	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.97236454]
8	El autor	You cannot protect yourself from COVID-19 by injecting, swallowing, bathing in or rubbing onto your body bleach, disinfectants or rubbing alcohols.	1	Tech	['post', 'facebook', 'novel', 'show', 'share']	0	1	[0.9596821]
9	Dataset	Turneric And Lemon Help Fight Against coronavirus.	0	Health	['post', 'facebook', 'novel', 'show', 'share']	0	0	[0.99547905]
10	El autor	Turneric And Lemon not Help Fight Against coronavirus.	0	Health	['post', 'facebook', 'novel', 'show', 'share']	1	1	[0.9961331]

Cuadro 7: Pruebas con negaciones de noticias ya comprobadas. (L: Label, P: Predicción, D: Diferencia, 0: FALSE, 1: TRUE). Fuente: El auto

ID	Fuente	Texto	L	Tema	Subtema	P	D	Probabilidad
1	El autor	President Donald Trump dies by coronavirus	0	Politics	['trump', 'presid', 'state', 'death', 'unit']	0	0	[0.99752414]
2	El autor	Important research shows that covid is transmitted through water	0	Env	['infect', 'case', 'health', 'pictur']	0	0	[0.9999754]
3	El autor	In Colombia no covid contegy is reported	0	Politics	['infect', 'case', 'health', 'pictur']	0	0	[0.95423234]
4	El autor	A town in europe is detected where all its inhabitants are immune to the coronavirus	0	Health	['infect', 'case', 'health', 'pictur']	0	0	[0.96206295]
5	El autor	it is shown that the coronavirus was created in a laboratory as a biological weapon for the birth control of the population	0	Health	['trump', 'presid', 'state', 'death', 'unit']	0	0	[0.9999866]
6	El autor	the coronavirus vaccine is a business and the price at which it is sold is 100 times higher than its cost of production	0	Health	['toilet', 'paper', 'australia', 'health', 'custom']	0	0	[0.99994314]
7	El autor	First case of a dog with coronavirus detected in Australia	0	Health	['case', 'australia', 'health', 'pictur', 'australian']	1	1	[0.03450221]
8	El autor	in africa the coronavirus has not spread because its inhabitants have a very strong immune system	0	Health	['case', 'australia', 'health', 'pictur', 'australian']	0	0	[0.9943732]
9	El autor	Person who receives package from china by ebay gets coronavirus	0	Health	['mask', 'kill', 'wear', 'chines', 'hand']	0	0	[0.98594224]
10	El autor	International space station astronaut catches coronavirus	0	Env	['infect', 'case', 'health', 'pictur']	0	0	[0.99462175]
11	El autor	Joe Biden have coronavirus	0	Health	['trump', 'presid', 'state', 'death', 'unit']	0	0	[0.9040914]

Cuadro 6: Pruebas con hechos artificialmente creados por el autor. (L: Label, P: Predicción, D: Diferencia, 0: FALSE, 1: TRUE). Fuente: El auto

formación. Este hecho es relevante pues ya se había probado anteriormente la afirmación “*Animals and pets cannot catch coronavirus or transmit the virus to humans*” y se había predicho como verdadera, sin embargo a pesar de que la frase “*First case of a dog with coronavirus detected in Australia*” se refiere a una mascota que contrae el coronavirus, es evidente una contradicción en el modelo. En la matriz de confusión del cuadro 1e se registra el resultado de la prueba.

B.6. Pruebas con negaciones de frases

Con el fin de explorar con mayor profundidad la contradicción detectada anteriormente, se plantea probar el comportamiento del modelo ante cambios de sentido de los enunciados. Para esto se redacta nuevamente la frase en forma de negación, tratando de agregar la menor cantidad de palabras posible.

Un ejemplo de esto es la afirmación: “*Washing hands often with antibacterial soap and water is*

imperative to protect yourself from covid”, la cual al ser redactada nuevamente como: “*Washing hands often with antibacterial soap and water is not imperative to protect yourself from covid.*” cambia completamente el sentido y pasaría de tener una etiqueta de verdadera a una etiqueta falsa.

En el cuadro 7 se incluyen 5 afirmaciones, cada una con su redacción original y su redacción en negativo que le cambia el sentido. Al aplicar el modelo, 4 de los 5 enunciados a los que se les cambió el sentido no son detectados correctamente por el modelo, evidenciando en este caso una limitación para este tipo de casos en los que unas pocas palabras agregadas a un enunciado le cambian completamente el sentido. En la matriz de confusión del cuadro 1f se registra el resultado de la prueba.

VI. DISCUSIÓN

Si bien en el cuadro 1 se han generado las matrices de confusión de cada una de las pruebas realizadas, en el cuadro 8 se incluye una matriz de confusión consolidada que agrupa la totalidad de las pruebas y que permite comparar objetivamente la totalidad de los resultados de las pruebas.

		Real	
		True	False
		42	6
		4	12
True	False		
		4	True
		12	False
		Predicción	

Cuadro 8: Matriz de confusión consolidada de todas las pruebas. Fuente: El autor

A partir de estos resultados se plantean las siguientes observaciones y puntos de discusión, los

cuales se espera orienten nuevos trabajos para complementar y mejorar la metodología planteada.

- Teniendo en cuenta que el modelo desarrollado en la presente investigación está orientado hacia la detección de desinformación, la cantidad de elementos catalogados como verdaderos fue considerablemente menor que la de los elementos catalogados como falsos, esto hace que para el modelo sea más difícil distinguir los elementos catalogados como verdaderos. Una posible mejora podría ser recopilar una mayor cantidad de titulares verdaderos y re-entrenar el modelo.
- En una temática que actualmente es tan popular como la de la pandemia de COVID-19, se genera día a día una gran cantidad de información y se corre el riesgo de que el contexto bajo el cual se realizó el entrenamiento del modelo cambie tan rápidamente, y en tal proporción que logre desactualizar el modelo. La recolección de nuevos datos para con ellos actualizar periódicamente el modelo es clave para que no pierda vigencia con el paso del tiempo, Un proceso automatizado de entrenamiento continuo podría ser una gran mejora para la metodología planteada.
- Las pruebas en las que se realizó el cambio de sentido de los titulares al negarlos, dejó en evidencia la limitación del modelo para detectar estos cambios. Esto se debe a que a pesar de que el sentido de la frase se cambió totalmente, la mayoría de las palabras en la oración se conservan, por lo que no se genera un cambio significativo que pueda ser detectado por el modelo. La investigación en métodos que logren entender este tipo de cambios en el sentido de los titulares puede aportar una considerable mejora al modelo.
- Otro punto de mejora identificado es sin duda la necesidad de trabajar en el relacionamiento de las palabras para evitar así las contradicciones, como por ejemplo la detectada entre las frases: *“Animals and pets cannot catch coronavirus or transmit the virus to humans”* etiquetada como verdadera, y la frase “*“First case of a dog with coronavirus detected in Australia”* etiquetada también como verdadera. El modelo efectivamente no reconoce la relación entre la palabra “*dog*” y la palabra “*pet*”, generando una contradicción ya que si el modelo ha aprendido que las mascotas no pueden contagiarse de coronavirus y teniendo en cuenta que un perro es una mascota, la afirmación de

que se detectó un primer caso de coronavirus en un perro debería ser etiquetada inmediatamente como falsa.

- Como fortalezas del modelo desarrollado se evidecia su aplicación para la validación de mitos y la verificación de buenas prácticas relacionadas con el coronavirus. Esta puede ser un caso de uso muy demandado que podría convertirse rápidamente en una aplicación de gran impacto para los usuarios y que a su vez permitiría continuar recolectando datos para alimentar el entrenamiento continuo del modelo.

VII. CONCLUSIONES

- Es evidente que la desinformación en forma de noticias falsas cobra más sentido en tiempos de crisis o acontecimientos relevantes a nivel nacional o mundial. Tal es el caso de la pandemia de COVID-19, donde este tipo de noticias han cobrado relevancia y se han difundido casi tan rápido como el mismo virus. Es un momento clave para investigar este tipo de infodemias ya que contamos con la información y los datos disponibles, y así poder estar listos para las próximas olas de desinformación que seguramente van a seguir generándose en un mundo donde las personas estamos cada vez más interconectados.
- Se comprobó que las técnicas de inteligencia artificial independientemente de si se trata de técnicas de aprendizaje automático o de aprendizaje profundo pueden ser usadas para apoyar la detección de desinformación dentro de titulares de noticias y específicamente para apoyar la detección de desinformación relacionada con la pandemia de COVID-19. Estas técnicas podrán apoyar la tarea para entregarle más pistas al usuario sobre un posible contenido que no sea confiable pero siempre el usuario será quien tenga la última palabra para tomar la decisión de replicar o no replicar una información.
- Un desafío detectado durante la investigación fue el contexto actual frente a las noticias que se han generado en el marco de la pandemia de COVID-19. El contexto frente a la desinformación es dinámico y se adapta de acuerdo a los acontecimientos de la historia. En este caso los conjuntos de datos tradicionalmente usados para el entrenamiento de clasificadores de noticias falsas no fueron de gran utilidad,

pues no generaban un modelo coherente con el contexto de las noticias actuales. Esto quiere decir que es importante el desarrollo de modelos que puedan adaptarse fácilmente al entorno y puedan aprender rápidamente para poder seguir aportando buenos resultados.

- La posibilidad de implementar diferentes entradas dentro de un modelo de redes neuronales profunda permite experimentar con nuevas opciones para su entrenamiento como por ejemplo diferentes tipos de datos: secuenciales y no secuenciales, texto y numéricos, entre otros. Esta característica permite utilizar redes neuronales de extremo a extremo, combinando características diferentes dentro de la red, en lugar de apilar modelos manualmente, expandiendo los casos de uso de este tipo de redes hacia nuevos horizontes.

A. Trabajos futuros

- Incorporación de atributos adicionales de caracterización de la información: Pensar en nuevas formas de caracterizar la información puede ser de gran utilidad para lograr mejores resultados en modelos más adaptables y moldeables al contexto. Los corpus de entrenamiento deberían poder incluir por ejemplo contextos relacionados con las noticias de actualidad que puedan ser utilizadas como fuente de información complementaria de referencia.
- Implementación de técnicas en idioma Español: La gran debilidad para desarrollar este tipo de técnicas y modelos está en la ausencia de los conjuntos de datos adecuados para realizar las investigaciones. El inicio de esta línea de trabajo futuro debería ser recolectar y preparar estos conjuntos de datos para luego iniciar con el desarrollo de estas técnicas. Otra posibilidad es pensar en extraer características que puedan ser independientes del idioma o que sean compartidas por varios idiomas.
- Incluir nuevos tipos de datos dentro del análisis: La posibilidad de incorporar varias entradas dentro de un modelo de redes neuronales profundas permite pensar en combinar diferentes tipos de elementos que también hacen parte de la noticia a verificar, como imágenes o metadatos relacionados. Esto puede derivar en modelos que alcancen un mayor rendimiento al considerar más información para la toma de decisiones.
- El desafío del contexto cambiante: Este reto es uno de los más importantes en el ámbito de la

desinformación en forma de noticias falsas, ya que a medida que pasa el tiempo los acontecimientos de la historia hacen que el contexto de las noticias cambie y la forma de generar desinformación también se adapta. Esto motiva a pensar en modelos resistentes al contexto o adaptables y que puedan aprender en el camino, como por ejemplo las técnicas de aprendizaje reforzado, las cuales podrían ser de utilidad en estos nuevos escenarios.

Referencias

- [1] Schwab Klaus and Brende Børge. The global risks report 2018 13th edition. *World Economic Forum*, 2018.
- [2] Oxford. Disinformation meaning. *Oxford Dictionary*, 2020.
- [3] Oxford. Fake news meaning. *Oxford Dictionary*, 2020.
- [4] The Conversation. The fake news that sealed the fate of antony and cleopatra, 2017. [Web; accedido el 02-11-2020].
- [5] Xichen Zhang and Ali A. Ghorbani. An overview of online fake news: Characterization, detection, and discussion. *Information Processing Management*, 57(2):102025, 2020.
- [6] Ireton Cherilyn and Posetti Julie. Journalism, "fake news" disinformation. *UNESCO*, 2018.
- [7] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *Sigkdd Explorations*, 19(1):22–36, 2017.
- [8] Supanya Aphiwongsophon and Prabhas Chongstitvatana. Detecting fake news with machine learning method. pages 528–531, 07 2018.
- [9] Vasu Agarwal, H. Parveen Sultana, Srijan Malhotra, and Amitrajit Sarkar. Analysis of classifiers for fake news detection. *Procedia Computer Science*, 165:377 – 383, 2019. 2nd International Conference on Recent Trends in Advanced Computing ICRTAC -DISRUP - TIV INNOVATION , 2019 November 11-12, 2019.
- [10] Nabamita Deb, Vishesh Jha, Alok Panjiyar, and Roshan Gupta. A comparative analysis of news categorization using machine learning approaches. *International Journal of Scientific Technology Research*, 9:2469–2472, 01 2020.

- [11] Harika Kudarvalli and Jinan Fiaidhi. Detecting Fake News using Machine Learning Algorithms. 4 2020.
- [12] Souvick Ghosh and Chirag Shah. Toward automatic fake news classification. In *Proceedings of the Association for Information Science and Technology*, volume 55, pages 805–807, 2018.
- [13] Pritika Bahad, Preeti Saxena, and Raj Kamal. Fake news detection using bi-directional lstm-recurrent neural network. *Procedia Computer Science*, 165:74–82, 2019.
- [14] Yang Liu and Yi fang Brook Wu. Fned: A deep network for fake news early detection on social media. *ACM Transactions on Information Systems*, 38(3):1–33, 2020.
- [15] Chris Dulhanty, Jason L. Deglint, Ibrahim Ben Daya, and Alexander Wong. Taking a stance on fake news: Towards automatic disinformation assessment via deep bidirectional transformer language models for stance detection, 2019.
- [16] Ahlem Drif, Zineb Ferhat Hamida, and Silvia Giordano. Fake news detection method based on text-features. 08 2019.
- [17] Jiawei Zhang, Bowen Dong, and Philip S. Yu. Fakedetector: Effective fake news detection with deep diffusive neural network. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1826–1829, 2020.
- [18] Sherry Girgis, Eslam Amer, and Mahmoud Gadallah. Deep learning algorithms for detecting fake news in online text. In *2018 13th International Conference on Computer Engineering and Systems (ICCES)*, 2018.
- [19] Jingjing Cai, Jianping Li, Wei Li, and Ji Wang. Deeplearning model used in text classification. In *2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, 2018.
- [20] Sebastian Kula, Michał Choraś, Rafał Kozik, Paweł Ksieniewicz, and Michał Woźniak. Sentiment analysis for fake news detection by means of neural networks. In Valeria V. Krzhizhanovskaya, Gábor Závodszky, Michael H. Lees, Jack J. Dongarra, Peter M. A. Sloot, Sérgio Brissos, and João Teixeira, editors, *Computational Science – ICCS 2020*, pages 653–666, Cham, 2020. Springer International Publishing.
- [21] Xinyi Zhou and R. Zafarani. Fake news: A survey of research, detection methods, and opportunities. *ArXiv*, abs/1812.00315, 2018.
- [22] Tim Brown. Design thinking. *Harvard business review*, 86:84–92, 141, 07 2008.
- [23] Amanda J. Weller. Design thinking for a user-centered approach to artificial intelligence. *She Ji: The Journal of Design, Economics, and Innovation*, 5(4):394 – 396, 2019.
- [24] Gautam Kishore Shahi and Durgesh Nandini. FakeCovid – a multilingual cross-domain fact check news dataset for covid-19. In *Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media*, 2020.
- [25] Dipti AGARWAL, Isha; Rana. Covid19fn fake news dataset for covid-19. In *Sardar Vallabhbhai National Institute of Technology, Mendeley Data*, 2020.
- [26] G. Xu, Y. Meng, Z. Chen, X. Qiu, C. Wang, and H. Yao. Research on topic detection and tracking for online news texts. *IEEE Access*, 7:58407–58418, 2019.
- [27] K. Xu, F. Wang, H. Wang, and B. Yang. Detecting fake news over online social media via domain reputations and content understanding. *Tsinghua Science and Technology*, 25(1):20–27, 2020.
- [28] M. Turkoglu, D. Hanbay, and A. Sengur. Multi-model lstm-based convolutional neural networks for detection of apple diseases and pests. *Ambient Intell Human Comput*, 2019.
- [29] Xu Liu, Abdelouahed Gherbi, Wubin Li, and Mohamed Cheriet. Multi features and multi-time steps lstm based methodology for bike sharing availability prediction. *Procedia Computer Science*, 155:394 – 401, 2019. The 16th International Conference on Mobile Systems and Pervasive Computing (MoBiSPC 2019), The 14th International Conference on Future Networks and Communications (FNC-2019), The 9th International Conference on Sustainable Energy Information Technology.
- [30] World Health Organization. Coronavirus disease (covid-19) advice for the public: Mythbusters, 2020. [Web; accedido el 28-01-2021].
- [31] Johns Hopkins Medicine. Coronavirus disease 2019: Myth vs. fact, 2020. [Web; accedido el 28-01-2021].