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 alcoholes*".

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 viralizarce no solamente entre las personas, sino también también 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 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 intensión 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 canal tan efectivo para diseminar 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 faltas 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.
- 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 extractor 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, preprocesamiento, extracción de características, definición del modelo, entrenamiento del modelo, prueba del modelo, validación y evaluación del algoritmo entrenado [9], [11] y [10].

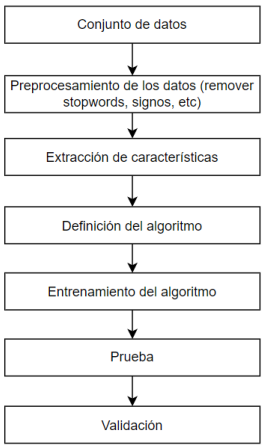


Figura 3: Proceso de análisis de noticias. Fuente: El autor

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 estás 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 "*desing thinking*" que incluye las fases de entendimiento, definición, ideación, prototipado y prueba [22] (Ver figura 4).

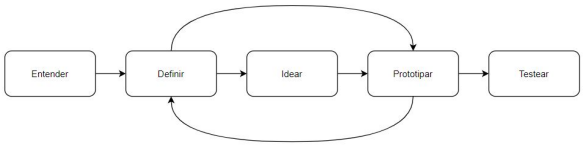


Figura 4: Metodología definida para el abordaje de la investigación. Fuente: el autor

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.

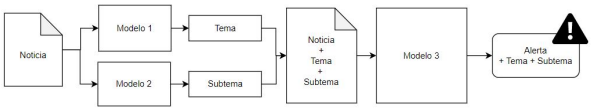


Figura 5: Componentes de la metodología. Fuente: el autor

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.

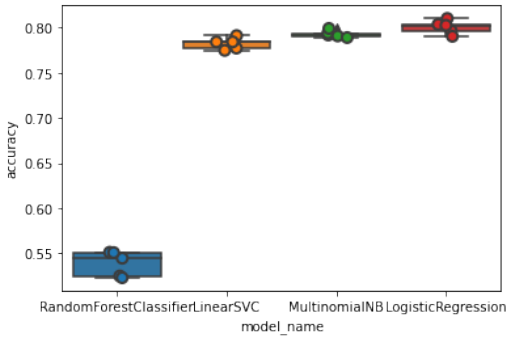
Estos conjuntos de datos se fusionaron en uno solo para contar con una mayor cantidad de ejemplos que facilitaran el entrenamiento de los modelos.

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.

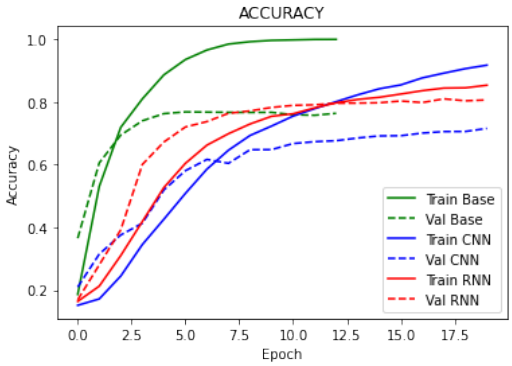
C. Desarrollo del modelo de predicción del tema

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

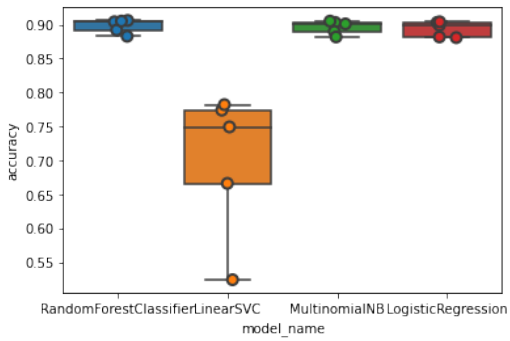
En la figura 6b se detalla la comparación de la precisión y el *loss* de los tres modelos generados. De acuerdo con este resultado es claro que el modelo más preciso fue el desarrollado con redes neuronales recurrentes LSTM, alcanzando un *accuracy* de alrededor de 0.8.

C.3. Comparación

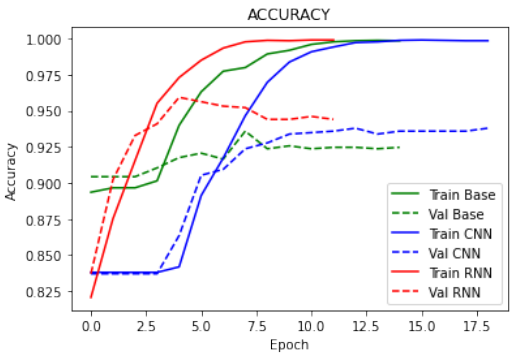
Se observa que tanto el clasificador *logistic* como el de redes neuronales recurrentes alcanzan una precisión de alrededor de 0.8. Considerando que el clasificador *logistic* es el más liviano computacionalmente, se seleccionará este método como parte de la metodología planteada.

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

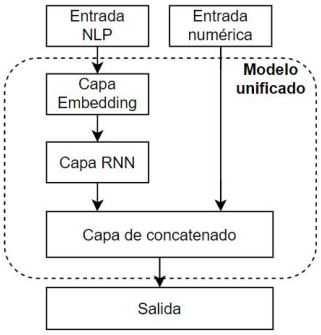


Figura 9: Modelo con múltiples entradas y una salida. Fuente: el autor

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 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.

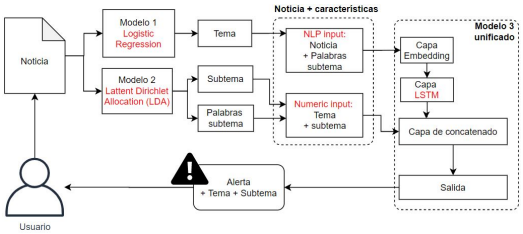


Figura 10: Planteamiento final de la metodología unificada. Fuente: el autor

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 relevantes[2].

Paso 1: Carga de conjunto de datos: Se carga del conjunto de datos y se realiza un preprocesamiento 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.

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

```
# Funcion para extraer subtematicas
def topics_lda(documento):
  unseen_document=documento
  bow_vector = dictionary.doc2bow(preprocess(unseen_document))
  prediction_lda=lda_model[bow_vector]
  probs=[]
  for i in range(0, len(prediction_lda)):
    probs.append(prediction_lda[i][1])
  max_probs=max(probs)
  for i in range(0,len(prediction_lda)):
    if max_probs==prediction_lda[i][1]:
      position=i
      break
  return position
# Aplico modelo LDA a conjunto de datos del dataframe
datos_revisar=datos_trabajo["Text"]
list_result_id=[]
list_result_words=[]
for i in range(0,len(datos_revisar)):
  id_predict=topics_lda(datos_revisar[i])
  list_result_id.append(id_predict)
  list_result_words.append(str(topics[id_predict]))
# Creo dataframe con columna adicional de prediccion y palabras
datos_trabajo_pred["pred_topics_id"]=list_result_id
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).

```
# Extraigo datos de dataframe
corpus_trabajo = datos_analizar["text_topics"] # Datos texto - Input 1
meta = datos_analizar["metadatos"] # Metadatos numericos - Input 2
results_trabajo = datos_analizar["label"].map(category_dict) # Prediccion
# Tokenizacion
corpus_trabajo = datos_analizar["text_topics"]
sequences = tokenizer.texts_to_sequences(corpus_trabajo.values)
X = pad_sequences(sequences, maxlen=max_len)
meta=meta.values.tolist()
meta_arr = np.array(meta)
# Train - Test split
x_train1,x_test1, y_train,y_test = train_test_split(X, results_trabajo,
      test_size=0.2, random_state=88 ) #  (Input 1)
x_train2,x_test2, y_train2,y_test2 = train_test_split(meta_arr,
      results_trabajo, test_size=0.2, random_state=88 ) # (Input 2)
# Definicion de la red
nlp_input = Input(shape=(seq_length,), name='nlp_input')
meta_input = Input(shape=(2,), name='meta_input')
emb = Embedding(output_dim=emb_dim, input_dim=embedding_size,
      input_length=seq_length)(nlp_input)
nlp_out = (LSTM(64, dropout=0.7, recurrent_dropout=0.7,
      kernel_regularizer=regularizers.l2(0.01)))(emb)
x = concatenate([nlp_out, meta_input])
x = Flatten()(x)
x = Dropout(0.3)(x)
x = Dense(32, activation='relu')(x)
x = Dense(1, activation='sigmoid')(x)
model = Model(inputs=[nlp_input , meta_input], outputs=[x])
# Compilacion
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 incluirlas. En el siguiente apartado de código se puede observar esta particularidad.

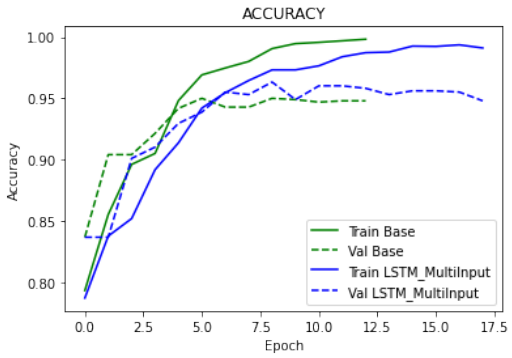
```
Model: "functional_35"
__________________________________________________________________________________________________
Layer (type)                    Output Shape         Param #     Connected to
==================================================================================================
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

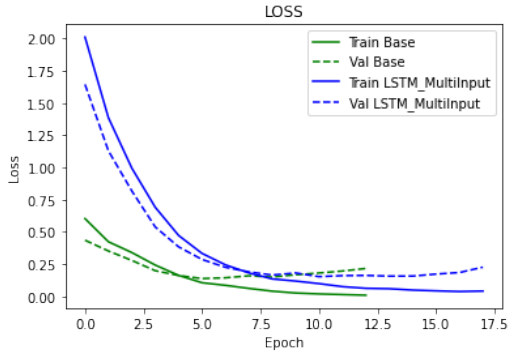
```
# Entrenamiento
historyFinal1=model.fit({'nlp_input': x_train1, 'meta_input': x_train2},
       y_train, epochs=epochs, batch_size=batch_size,validation_split=0.2,
       callbacks=[EarlyStopping(monitor='val_loss',patience=7, min_delta
       =0.0001)])
```

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`.

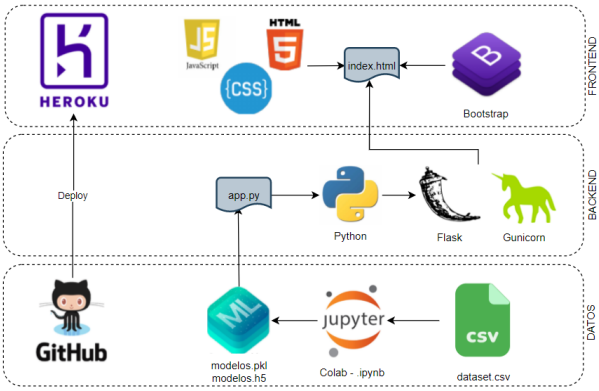
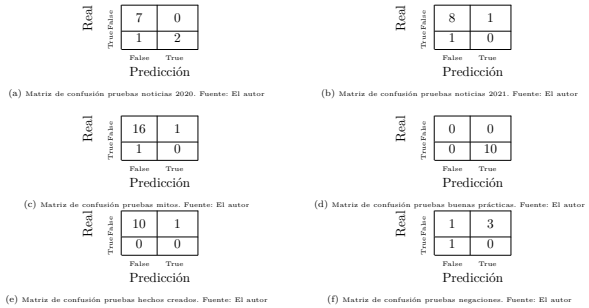


Figura 14: Arquitectura del modelo implementado. Fuente: el autor

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.



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 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.

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]

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

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', '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]

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 Explorer[3], 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 contraer el coronavirus, citando principalmente como fuente a la Organización Mundial de la Salud.

[3]https://toolbox.google.com/factcheck/explorer

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]

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

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-

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]

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	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 Bide 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 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.

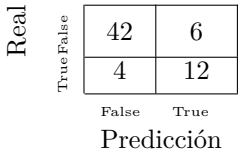
ID	Fuente	Texto	L	Tema	Subtema	P	D	Probabilidad
1	WHO	Washing hands often with antibacterial soap and water is imperative to protect yourself from covid.	1	Tech	['mask', 'kill', 'wear', 'chines', 'hand']	1	0	[0.2492171]
2	El autor	Washing hands often with antibacterial soap and water is not imperative 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', '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]

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

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.



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 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.

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 al 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 caracterízación 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 estar 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].