Inteligencia artificial contra la desinformación: fundamentos, avances y retos

Fighting disinformation with artificial intelligence: fundamentals, advances and challenges

Andrés Montoro-Montarroso; Javier Cantón-Correa; Paolo Rosso; Berta Chulvi; Ángel Panizo-Lledot; Javier Huertas-Tato; Blanca Calvo-Figueras; M. José Rementeria; Juan Gómez-Romero

Note: This article can be read in its English original version on:
https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/87328

Cómo citar este artículo.
Este artículo es una traducción. Por favor cite el original inglés:

Montoro-Montarroso, Andrés; **Cantón-Correa, Javier**; **Rosso, Paolo**; **Chulvi, Berta**; **Panizo-Lledot, Ángel**; **Huertas-Tato, Javier**; **Calvo-Figueras, Blanca**; **Rementeria, M. José**; **Gómez-Romero, Juan** (2023). "Fighting disinformation with artificial intelligence: fundamentals, advances and challenges". *Profesional de la información*, v. 32, n. 3, e320322.
https://doi.org/10.3145/epi.2023.may.22

Artículo recibido el 27-03-2023
Aceptación definitiva: 17-05-2023

Andrés Montoro-Montarroso ✉
https://orcid.org/0000-0003-1893-3346
Universidad de Granada
Decsai
Citic-UGR
Periodista Rafael Gómez Montero, 2
18014 Granada, España
andres.montoro@ugr.es

Javier Cantón-Correa
https://orcid.org/0000-0002-8466-1679
Universidad Internacional de La Rioja
Fac. de Ciencias Sociales y Humanidades
Universidad de Granada
Decsai
Citic-UGR, España
javicanton@ugr.es

Paolo Rosso
https://orcid.org/0000-0002-8922-1242
Universitat Politècnica de València
Pattern Recognition and Human Language Technologies (*PRHLT*) *Research Center*
Camí de Vera, s/n
46022 Valencia, España
prosso@dsic.upv.es

Berta Chulvi
https://orcid.org/0000-0003-1169-0978
Universitat Politècnica de València
Pattern Recognition and Human Language Technologies (*PRHLT*) *Research Center*
Camí de Vera, s/n
46022 Valencia, España
berta.chulvi@upv.es

Ángel Panizo-Lledot
https://orcid.org/0000-0002-2195-3527
Universidad Politécnica de Madrid
Escuela Técnica Superior de Ingeniería de Sistemas Informáticos
Alan Turing, s/n
28031 Madrid, España
angel.panizo@upm.es

Javier Huertas-Tato
https://orcid.org/0000-0003-4127-5505
Universidad Politécnica de Madrid
Escuela Técnica Superior de Ingeniería de Sistemas Informáticos
Alan Turing, s/n
28031 Madrid, España
javier.huertas.tato@upm.es

Blanca Calvo-Figueras
https://orcid.org/0000-0001-6939-3576
Barcelona Supercomputing Center (*BSC*)
Language Technologies Unit
Plaça Eusebi Güell, 1-3
08034 Barcelona, España
blanca.calvo@bsc.es

M. José Rementeria
https://orcid.org/0000-0002-3140-1160
Barcelona Supercomputing Center (*BSC*)
Social and Media Impact Evaluation
Plaça Eusebi Güell, 1-3
08034 Barcelona, España
maria.rementeria@bsc.es

Juan Gómez-Romero
https://orcid.org/0000-0003-0439-3692

Universidad de Granada
Decsai
Citic-UGR
Periodista Rafael Gómez Montero, 2
18014 Granada, España
jgomez@decsai.ugr.es

Resumen

Internet y las redes sociales han revolucionado la forma en la que se distribuye y consume la información. Sin embargo, la enorme cantidad de contenidos disponibles en estas plataformas dificulta la tarea distinguir entre lo verdadero y lo falso, más aún con la proliferación de actores malintencionados que difunden bulos. Desmentir la desinformación es un proceso muy costoso, por lo que en los últimos años se han desarrollado múltiples investigaciones sobre el potencial de la inteligencia artificial (IA) –y, más concretamente, del aprendizaje automático (AA)– como una solución al problema. Este trabajo revisa la bibliografía reciente sobre las técnicas de IA y AA que han sido propuestas para combatir la desinformación, que van desde la clasificación automática de texto hasta la extracción de características, así como el papel relevante que pueden jugar en la creación de contenido artificial. La principal conclusión del estudio es que los avances en IA se han centrado principalmente en la clasificación automática y que su utilización fuera de los laboratorios de investigación ha sido escasa. Esto se debe principalmente a que los modelos de AA dependen mucho de los conjuntos de datos con los que son entrenados, lo cual limita su aplicación y su efectividad en diferentes ámbitos. En consecuencia, se propone que los esfuerzos de investigación ha de dirigirse hacia el desarrollo de sistemas de IA que sean explicables, confiables y que apoyen a las personas, en lugar de sustituirlas, en la detección temprana de desinformación.

Palabras clave

Periodismo; Desinformación; Computación; Inteligencia artificial; IA; Aprendizaje automático; Verificación periodística; Fact-checking; Conjuntos de datos; Datasets; Procesamiento del lenguaje natural; PLN; Análisis de redes sociales; Ultrafalsificaciones; Modelos de lenguaje de gran tamaño.

Abstract

Internet and social media have revolutionised the way news is distributed and consumed. However, the constant flow of massive amounts of content has made it difficult to discern between truth and falsehood, especially in online platforms plagued with malicious actors who create and spread harmful stories. Debunking disinformation is costly, which has put artificial intelligence (AI) and, more specifically, machine learning (ML) in the spotlight as a solution to this problem. This work revises recent literature on AI and ML techniques to combat disinformation, ranging from automatic classification to feature extraction, as well as their role in creating realistic synthetic content. We conclude that ML advances have been mainly focused on automatic classification and scarcely adopted outside research labs due to their dependence on limited-scope datasets. Therefore, research efforts should be redirected towards developing AI-based systems that are reliable and trustworthy in supporting humans in early disinformation detection instead of fully automated solutions.

Keywords

Journalism; Disinformation; Computing; Artificial intelligence; AI; Machine learning; Fact-checking; Datasets; Natural language processing; NLP; Social network analysis; Deepfakes; Large language models.

Financiación

Este trabajo ha sido financiado por la *Comisión Europea*, proyecto *Iberifier* (*Iberian digital media research and fact-checking hub*), convocatoria CEF-TC-2020-2 (*European Digital Media Observatory*), número 2020-EU-IA-0252.

1. Introducción

En medio de la imperante Era de la Posverdad, las personas se ven abrumadas por un flujo enorme e ininterrumpido de información, lo que dificulta discernir entre el contenido veraz y el que nos intenta engañar, ya sea de manera intencionada (en inglés, *disinformation*) o no intencionada (*misinformation*) (**Wardle**; **Derakhshan**, 2017). Como resultado, la desinformación supone una amenaza significativa, con potencial para influir negativamente en el tejido político, económico y cultural de la sociedad y erosionar los principios democráticos.

Aunque a los expertos les pueda resultar relativamente fácil desmentir algunos bulos, se necesitan recursos para impulsar y agilizar su trabajo, así como para capacitar a los ciudadanos y organizaciones no especializados. No sorprende, por tanto, que haya crecido el interés por elaborar tecnologías para la verificación automática de información, sobre

todo en el ámbito de las redes sociales. En particular, el aprendizaje automático (AA), un subcampo de la inteligencia artificial (IA), ha contribuido significativamente a combatir la desinformación en los últimos años. Los algoritmos de AA permiten identificar de manera automática patrones propios de la desinformación en conjuntos de datos y aplicar estos patrones aprendidos para determinar la veracidad o falsedad de otros contenidos. Por su parte, el aprendizaje profundo (AP, en inglés *deep learning*) es un subconjunto del AA que ha demostrado ser muy útil en múltiples dominios (**LeCun**; **Bengio**; **Hinton**, 2015) y, actualmente, es el enfoque predominante en la lucha contra la desinformación mediante IA (**Xu**; **Sheng**; **Wang**, 2023). Sin embargo, el aprendizaje profundo también se puede utilizar para generar contenidos sintéticos que aumentan el impacto y el alcance de la desinformación (**Masood** *et al.*, 2022).

> El aprendizaje automático (AA), un subcampo de la inteligencia artificial (IA), ha contribuido significativamente a combatir la desinformación en los últimos años

El AA es un área de investigación muy técnica y compleja, por lo que resulta difícil para los profanos aprovechar las soluciones que puedan surgir. Al mismo tiempo, los investigadores de AA deben conocer las múltiples facetas de un problema social como es la desinformación y no abordarlo desde una perspectiva únicamente tecnológica. Por todo ello, el objetivo de este artículo es ofrecer una breve guía multidisciplinar que permita navegar por la bibliografía reciente sobre la lucha contra la desinformación mediante IA, centrándose en AA y AP. En este documento se analiza la eficacia de estas técnicas para detectar y combatir la desinformación y se identifican los retos y las limitaciones de los enfoques actuales. También se proponen líneas futuras de investigación para elaborar sistemas de IA que permitan ayudar a los verificadores (en inglés *fact-checkers*) y a los investigadores en sus tareas de detección temprana de la desinformación.

La desinformación en los medios digitales se difunde principalmente a través de texto. Por ello, a la hora de desarrollar algoritmos de AA, las principales características identificativas de la desinformación que se suelen tener en cuenta están relacionadas con el léxico, la sintaxis, el estilo y la semántica de los textos, cuyo análisis cae dentro del ámbito del procesamiento del lenguaje natural (PLN). Por otra parte, el análisis de redes sociales (ARS) investiga la topología de las redes de desinformación, analizando su estructura e identificando comunidades de usuarios sospechosos de generar y difundir contenidos nocivos, ya sea de forma coordinada o descoordinada. En consecuencia, centramos este trabajo en el PLN y el ARS, por ser las áreas de IA más frecuentemente relacionadas con el tratamiento de la desinformación. También consideraremos los avances recientes en generación de contenidos multimedia artificiales, como imágenes y vídeos, que pueden contribuir poderosamente a amplificar las narrativas desinformativas.

Nuestro análisis de las tecnologías de IA para la desinformación se organiza en tres áreas no disjuntas:

- identificación de la desinformación mediante clasificación automatizada;
- extracción de características de la desinformación;
- apoyo a las tareas de verificación de información.

Esta organización es coherente con los planteamientos de los trabajos de investigación revisados y refleja la evolución histórica del área:

- Clasificación de la desinformación. La clasificación automatizada es la forma más directa de analizar la desinformación, ya que usa un conjunto de datos ya etiquetados para entrenar un modelo de AA que se encarga de diferenciar los contenidos veraces de los falsos. Esta metodología tiene el inconveniente de que los modelos entrenados en un dominio concreto son difícilmente extensibles a otros dominios.
- Identificación de desinformación basada en características. La extracción de características se centra en encontrar distintos rasgos y particularidades de la desinformación que se puedan utilizar para detectar, de forma manual o automática, los contenidos falsos y las comunidades implicadas en su difusión.
  Verificación híbrida o semiautomatizada. La verificación de bulos por periodistas especializados ha demostrado ser muy eficaz para atajar la desinformación, pero la cantidad de contenidos para analizar es muy superior a la capacidad de los verificadores. Esta limitación ha llevado a la aparición de un tercer enfoque denominado verificación semiautomatizada.

El resto de este artículo se divide en tres partes. La primera describe las técnicas y métodos de IA utilizados para la detección de contenidos desinformativos. La segunda se centra en los métodos de IA propuestos en la bibliografía para combatir la desinformación, incluidas las características utilizadas para entrenar estos modelos y cómo los verificadores pueden aprovechar estos avances tecnológicos. La tercera describe el desafío que supone el creciente uso de IA para generar contenidos desinformativos de manera automatizada. Para concluir, realizamos un resumen de las principales conclusiones alcanzadas y de las líneas de investigación más prometedoras para el futuro.

2. Antecedentes

Las técnicas de aprendizaje automático son una herramienta dentro de la IA enormemente útil a la hora de abordar el problema de la desinformación, ya que permiten la detección y el análisis rápido de contenidos falsos. Esta sección ofrece una visión general de los fundamentos del AA y, en su contexto, del procesamiento del lenguaje natural y el análisis de redes sociales. Aquellos lectores más familiarizados con estas tecnologías pueden pasar a la sección siguiente; en caso contrario, pueden ampliar esta información en los trabajos clásicos de **Russell** y **Norvig** (2020) y **Bishop** (2006).

2.1. Aprendizaje automático (AA)

El aprendizaje automático es un campo de la IA que comprende una serie de métodos, técnicas y herramientas para construir sistemas inteligentes a partir de grandes volúmenes de datos. En concreto, el AA se inscribe en el paradigma del reconocimiento de patrones, es decir, trata de identificar características repetidas en una muestra de datos mediante procesos estadísticos y computacionales. Estos patrones cumplen dos funciones principales:

- hacer predicciones sobre acontecimientos futuros (análisis predictivo);
- descubrir relaciones entre los datos (análisis descriptivo).

En función del proceso de obtención de patrones y del modo de aprendizaje, existen tres grandes familias dentro del AA:

- Aprendizaje supervisado (s*upervised learning*);
- No supervisado (*unsupervised learning*);
- Aprendizaje por refuerzo (*reinforcement learning*).

El aprendizaje profundo, basado en redes neuronales, se encuadra habitualmente en el aprendizaje supervisado, pero también puede aplicarse a problemas de aprendizaje no supervisado y por refuerzo. Esta subsección se centra en las técnicas supervisadas y no supervisadas (incluido el aprendizaje profundo), por ser las técnicas de AA más utilizadas en la lucha contra la desinformación.

El aprendizaje supervisado construye modelos predictivos a partir de datos de entrenamiento previamente etiquetados para estimar el valor de esas etiquetas en otros conjuntos de datos desconocidos. El aprendizaje supervisado se puede dividir en dos categorías básicas, dependiendo de la naturaleza de la variable objetivo de la predicción: clasificación y regresión:

- En la clasificación, la variable objetivo tiene un número limitado de valores discretos o categorías. Métodos de este tipo son los árboles de decisión, la regresión logística, las máquinas de vector soporte (*support vector machine*) o el algoritmo de los vecinos más cercanos (*K-nearest neighbors*).
- En la regresión, la variable objetivo es numérica. Ejemplos de algoritmos de regresión son la regresión lineal, la regresión polinómica, los *splines* y los árboles de regresión.

Los métodos de aprendizaje supervisado suelen combinarse entre sí para aumentar su precisión, dando lugar a modelos como los de

- *Bagging* (con agregaciones sencillas);
- *Boosting* (múltiples modelos entrenados progresivamente);
- *Random forest* (combinan la salida de múltiples árboles de decisión).

El aprendizaje no supervisado se usa para modelado de datos que no han sido previamente etiquetados, a diferencia del supervisado. La técnica más extendida es la agrupación o *clustering*, que se utiliza para identificar grupos dentro de un conjunto de datos con fines descriptivos. Distinguimos entre el

- *clustering* particional, aquel en el que los clusters son disjuntos (sin elementos en común), como por ejemplo los algoritmos *Dbscan* y *k-means*;
- *clustering* jerárquico, en el que los grupos se organizan en categorías relacionadas.

Otra técnica destacable dentro del aprendizaje no supervisado son las reglas de asociación, cuyo objetivo es descubrir dependencias entre los ítems de una base de datos.

Actualmente, la tendencia dominante en AA es el aprendizaje profundo (**Goodfellow**; **Bengio**; **Courville**, 2016). El AP se enmarca, en principio, dentro del aprendizaje supervisado, aunque en la actualidad se ha extendido su aplicación a otros paradigmas. El AP está basado en las redes neuronales, un modelo computacional inspirado en las sinapsis neuronales, e incorpora múltiples capas de procesamiento para captar relaciones complejas en grandes conjuntos de datos. Dentro del AP se incluyen distintos tipos de algoritmos como, por ejemplo:

- las redes neuronales convolucionales, que son redes neuronales especializadas en el procesamiento de datos con una estructura regular (como las imágenes);
- las redes neuronales recurrentes, que permiten bucles de retroalimentación en su cálculo y se aplican a datos secuenciales como las series temporales o el texto;
- los *transformers*, que aprenden a identificar secciones relevantes de secuencias aplicando modelos de atención y, en consecuencia, resultan útiles con datos textuales.

2.2. Procesamiento del lenguaje natural (PLN)

El procesamiento del lenguaje natural (PLN) investiga técnicas de lingüística computacional para analizar textos en un idioma concreto (**Manning**; **Schütze**, 1999). Antes de desarrollar un modelo de AA para el PLN (por ejemplo, un modelo lingüístico), es fundamental abordar tres retos: el preprocesamiento del texto, la extracción de características y la representación numérica.

1) El preprocesamiento del texto consiste en limpiar el texto y eliminar elementos superfluos para que sólo quede la información útil. Las etapas fundamentales del preprocesamiento de textos son

- tokenización: división del texto bruto en fragmentos, normalmente palabras;
- eliminación de *stopwords*: palabras comunes no significativas para el análisis;
- *stemming*: reglas de tipo heurístico para cortar los extremos de las palabras o eliminar afijos;
- lematización: transformación de las palabras en su forma base o lema.

2) La extracción de características busca identificar y seleccionar los rasgos básicos de los datos textuales que resulten adecuados para la tarea. Algunas de las técnicas más utilizadas para la extracción de características son:

- etiquetado de partes del discurso (*part-of-speech tagging*) para identificar categorías léxicas;
- reconocimiento de entidades con nombre (*named-entity recognition*) para identificar entidades dentro del texto;
- bolsas de palabras (*bag-of-words*) para representar unidades lingüísticas en función de su frecuencia de aparición.

Otra técnica de extracción de características es el análisis de sentimientos, también llamado minería de opiniones, cuyo objetivo es captar automáticamente los sentimientos, opiniones, emociones o actitudes que subyacen a un texto (**Serrano-Guerrero** *et al.*, 2015). Esta tarea también puede incluir la obtención de los rasgos psicológicos del autor a través de léxicos anotados con fines específicos (**John**; **Srivastava**, 1999; **Pennebaker** *et al.*, 2015).

3) La representación de textos conlleva crear una codificación numérica del texto para que otros algoritmos de AA puedan realizar cálculos. Existen muchas técnicas para obtener esta representación, siendo una de las más utilizadas actualmente las incrustaciones de palabras (*word embeddings*), ya que permiten capturar parcialmente la semántica del texto. *Word-2Vec* (**Mikolov** *et al.*, 2013) y *GloVe* (**Pennington**; **Socher**; **Manning**, 2014) son las técnicas más utilizadas para obtener estos *embeddings* a partir de un conjunto de textos cualquiera. También existen *embeddings* públicamente disponibles para términos comunes que han sido precalculados a partir de grandes recursos de texto, como *Wikipedia*, y que pueden ser reutilizados en otras aplicaciones. Una vez que un documento está representado de forma numérica, se pueden aplicar técnicas de AA (y en particular métodos de AP) para resolver una tarea posterior (por ejemplo, clasificar o predecir texto).

Actualmente, las redes de tipo *transformer* con mecanismos de atención son las más empleadas, ya que superan las limitaciones de otros métodos previos (por ejemplo, las redes neuronales recurrentes) aprendiendo a identificar las partes esenciales del texto de entrada (**Vaswani** *et al.*, 2017). Este tipo de redes son usadas por los modelos masivos del lenguaje (*large language models*, LLMs), que son sistemas especializados en la predicción de la siguiente palabra de una secuencia y pueden utilizarse para la generación automática de textos de alta calidad. Un LLM especialmente destacable es el *generative pre-trained transformer* (GPT) (**Brown** *et al.*, 2020). Sus versiones actuales *GPT-3* y *GPT-4* son capaces de realizar una amplia gama de tareas de PLN, como la generación de texto, la traducción automática y la respuesta a preguntas (**Zhu**; **Luo**, 2022). Los LLMs de la familia GPT han sido integrados en el software *ChatGPT*, un agente conversacional entrenado mediante interacciones con personas para entablar conversaciones realistas (**Megahed** *et al.*, 2023).

2.3. Análisis de redes sociales (ARS)

Se ocupa de estudiar las relaciones entre las entidades de un sistema para comprender su funcionamiento global, el papel de los diferentes actores y los vínculos entre ellos (**Barabási**, 2016). En particular, el ARS utiliza métodos matemáticos y computacionales para analizar los datos de los medios sociales a través de dos enfoques (**Aggarwal**, 2011; **Camacho** *et al.*, 2020):

- el análisis estructural (topología de la red, comunidades y nodos importantes);
- el análisis basado en el contenido (información sobre los usuarios de los medios sociales y el contenido compartido).

El análisis estructural se centra en estudiar la topología de una red aplicando la teoría de grafos. Entre las métricas estructurales más utilizadas se encuentran medidas locales como la centralidad, el grado, la cercanía o la intermediación, utilizadas para identificar la importancia de determinados nodos (usuarios) dentro de la red, así como medidas globales como la densidad, el diámetro, el radio o la transitividad (utilizadas para estudiar la estructura global de la red). Un problema esencial en ARS es la detección de comunidades, cuyo objetivo es identificar conjuntos de nodos más estrechamente conectados (**Bedi**; **Sharma**, 2016). La tarea de detección de comunidades está estrechamente relacionada con el problema de la agrupación (*clustering*), por lo que la mayoría de las técnicas pertenecen a esta amplia familia de algoritmos (**Fortunato**, 2010). Otros enfoques se basan en la maximización de la modularidad, una medida que equilibra el número de conexiones internas y externas de una comunidad. Algunos algoritmos de este tipo son el método de **Newman** (2004) y el método de **Blondel** *et al.* (2008).

El análisis basado en el contenido examina tanto el contenido como las conexiones entre nodos, incorporando por ejemplo el texto de los mensajes intercambiados para ofrecer un contexto adicional a la red (**Cambria**; **Wang**; **White**, 2014). El análisis de contenido suele aplicarse de diversas formas:

- al perfil de usuario, recopilando información adicional sobre los actores humanos (por ejemplo, comportamiento o características físicas) en una red (**Harrigan** *et al.*, 2021);
- a la extracción de temas, identificando los principales temas de discusión entre un grupo de nodos (**Yin** *et al.*, 2012), o los intereses de los usuarios a través de sus conexiones sociales (**Wang** *et al.*, 2013);

- al análisis de sentimientos, que examina el tono de los mensajes intercambiados entre los nodos (**Camacho** *et al.*, 2020).

3. Clasificación automática de desinformación mediante aprendizaje automático

El aprendizaje supervisado es el método más utilizado para la identificación automática de la desinformación, problema que se modela como una clasificación binaria (es o no es desinformación). Formalmente, dado un conjunto de características representativas de un elemento de información I, la tarea consiste en predecir si I es veraz o no, es decir:

$$f(I) = \begin{cases} 1, \text{si } I \text{ es un elmento desinformativo} \\ 0, \text{si } I \text{ no es desinformativo} \end{cases}$$

donde f es la función que queremos aprender a partir de los datos disponibles. La selección de características y la manera de combinarlas para dar forma a f puede hacerse de manera manual o automática. En el primer caso, se puede aplicar la toma de decisiones multicriterio (*multi-criteria decision-making*) para definir criterios y pesos de probabilidad con los que poder calcular una puntuación de credibilidad de la información y clasificar así las soluciones candidatas (**Pasi**; **De-Grandis**; **Viviani**, 2020). En el segundo caso, se aplica AP para aprender de forma automática las características y los pesos con los que poder clasificar la veracidad de la información (**Amador**; **Molina-Solana**; **Gómez-Romero**, 2019; **Molina-Solana**; **Amador**; **Gómez-Romero**, 2018).

El enfoque anterior tiene como principal limitación que la desinformación no se presenta en términos absolutos de blanco o negro, sino que se encuentra en un amplio espectro de grises. En la bibliografía, encontramos definiciones más detalladas de etiquetas para captar estos matices sutiles entre diferentes grados de desinformación. Por ejemplo, **Wang** (2017) elaboró un conjunto de datos anotado manualmente con seis etiquetas, donde se evaluaba el grado de veracidad de miles de afirmaciones con las categorías mentira, falsedad, algo de cierto, media verdad, mayormente cierto y verdad. **Nakamura**, **Levy** y **Wang** (2020) utilizaron una jerarquía de etiquetado de dos, tres y seis categorías para cada muestra de su conjunto de datos multimodal, lo cual permitía la implementación de modelos de clasificación a diferentes niveles de granularidad.

> El rendimiento de los métodos de aprendizaje automático depende directamente de la calidad de los datos

El rendimiento de los métodos de AA (en concreto, del aprendizaje supervisado) depende directamente de la calidad de los datos etiquetados. Puesto que estos datos suelen capturar situaciones y eventos muy particulares, la aplicación de estos modelos a otros dominios no suele ser efectiva. Esta limitación se vuelve aún más evidente en la detección automática de desinformación, ya que resulta todo un desafío construir conjuntos de datos con la calidad suficiente para abarcar los matices de un fenómeno tan heterogéneo (**Shu** *et al.*, 2017). La construcción de conjuntos de datos es un proceso costoso, ya que implica:

- la extracción de datos ya sea a través de las APIs (*application programming interfaces*) proporcionadas por los propietarios de las plataformas o de métodos de *web scraping*;
- la anotación, que puede ser manual (lo cual requiere mucho tiempo), o semiautomática (lo que incrementa la probabilidad de errores en el etiquetado) (**Simko** *et al.*, 2021).

En el Anexo incluimos una tabla con conjuntos de datos publicados y usados para entrenar modelos de AA para la clasificación de desinformación.

Las investigaciones que emplean aprendizaje no supervisado para la detección de desinformación son mucho más escasas (**Guo** *et al.*, 2020; **Meel**; **Vishwakarma**, 2020; **Zhang**; **Ghorbani**, 2020). Algunos trabajos formulan la identificación automática de desinformación como un problema de detección de anomalías en redes sociales, empleando un *autoencoder* como método de aprendizaje no supervisado (**Li** *et al.*, 2021), mientras que otros enfoques emplean estadística bayesiana para calcular la veracidad de las noticias y evaluar la credibilidad de sus autores (**Yang** *et al.*, 2019). No obstante, la mayoría de los estudios utilizan el aprendizaje no supervisado de forma complementaria al aprendizaje supervisado; es decir, utilizan un enfoque denominado semi-supervisado (**De-Souza** *et al.*, 2022; **Dong**; **Victor**; **Qian**, 2020; **Li**; **Lu** *et al.*, 2022; **Meel**; **Vishwakarma**, 2021; **Paka** *et al.*, 2021).

4. Características para la detección de desinformación con aprendizaje automático

Como se ha mencionado, los métodos de detección de desinformación requieren de características representativas y relevantes los elementos que se van a analizar para ser efectivos. Tradicionalmente, las características empleadas para la clasificación de desinformación se han dividido en dos categorías: basadas en el contenido y basadas en el contexto (**Bondielli**; **Marcelloni**, 2019):

- las características basadas en el contenido son atributos relevantes extraídos directamente del ítem desinformativo, normalmente un texto que afirma o apoya el posible engaño, y que a menudo suele estar asociado con imágenes o vídeos que lo refuerzan;
- las características basadas en el contexto se refieren a los datos o metadatos que rodean el ítem desinformativo.

Esta sección se centrará en las distintas características que pueden extraerse y emplearse para detectar información falsa.

4.1. Caracterización estilística de mensajes con procesamiento del lenguaje natural

Los métodos basados en el contenido hacen uso de las características lingüísticas de la información falsa, incluyendo su sintáctica y semántica (**Zhou** *et al.*, 2020), que pueden analizarse con técnicas de PLN (**Ruffo** *et al.*, 2023). Entre las características sintácticas podemos encontrar el etiquetado gramatical y la búsqueda de grupos de palabras relevantes (bigramas, trigramas o n-gramas). Las características semánticas pueden obtenerse mediante el análisis de sentimientos, la detección de temas o las codificaciones con *word embeddings*.

Un tipo especial de características lingüísticas son aquellas basadas en el estilo. Los métodos de AA que emplean estas características pueden captar el estilo distintivo que usan los actores maliciosos para incrementar la difusión de sus contenidos (**Zhou**; **Zafarani**, 2020). El estilo de los textos desinformativos se ha medido en términos de la frecuencia de ciertos patrones morfológicos (**Castelo** *et al.*, 2019; **Vogel**; **Meghana**, 2020), la presencia de elementos estructurales (**Bonet-Jover** *et al.*, 2021), la variedad léxica y el uso de símbolos de puntuación (**Azevedo** *et al.*, 2021), la complejidad y el nivel de legibilidad del texto (**Castelo** *et al.*, 2019) y el tono emocional (**Giachanou**; **Rosso**; **Crestani**, 2019).

En cuanto a los patrones morfológicos, **Afroz**, **Brennan** y **Greenstadt** (2012) identificaron información falsa analizando el número de sílabas y palabras, el vocabulario y la complejidad gramatical. También se observó que los difusores de contenidos engañosos utilizaban un lenguaje más informal (**Giachanou** *et al.*, 2022), por ejemplo, por el uso de pronombres personales y palabras mal sonantes (**Rashkin** *et al.*, 2017). En cuanto al tono emocional del discurso, **Del-Vicario** *et al.* (2016) demostraron que el estado emocional de los usuarios de las redes sociales está relacionado con su nivel de interacción: más actividad conduce a emociones más negativas, y viceversa. En consecuencia, el uso de patrones lingüísticos polarizados se percibe como una estrategia para aumentar el impacto provocando emociones negativas en el receptor como la ira, el asco o el miedo (**Giachanou**; **Rosso**; **Crestani**, 2021), siendo por tanto un indicador de baja credibilidad (**Ghanem** *et al.*, 2021; **Stella**; **Ferrara**; **De-Domenico**, 2018).

Al mismo tiempo, los desinformadores pueden aprender características basadas en el estilo para replicar los estilos de escritura de fuentes de información fiables y pasar desapercibidos. Por ejemplo, **Schuster** *et al.* (2020) demostraron que los modelos de PLN para la identificación de desinformación basados en rasgos estilísticos funcionan bien con la escritura humana, pero fallan cuando se enfrentan a textos sintéticos creados por modelos lingüísticos entrenados para replicar la apariencia de medios de comunicación fiables.

4.2. Aspectos contextuales de la desinformación en las redes sociales

Las características contextuales se extraen teniendo en cuenta los datos relevantes que rodean un ítem desinformativo, incluidos los metadatos u otros elementos externos. Esta información está disponible principalmente en las redes sociales, donde el contexto puede estar relacionado con los usuarios, sus mensajes o la estructura de la red (**Guo** *et al.*, 2020).

4.2.1. Características basadas en el contexto de los usuarios

Las características basadas en el usuario incluyen el número de publicaciones, el número de seguidores, los datos demográficos, si la cuenta está verificada o la antigüedad de la cuenta en la plataforma. Una métrica habitual construida a partir de estos datos de perfil es la credibilidad del usuario, que puede indicar la probabilidad de compartir información falsa (**Shu**; **Wang**; **Liu**, 2019). La credibilidad puede obtenerse a partir de metadatos de la red social para analizar si existe una correlación entre un perfil de usuario y la publicación de información falsa (**Shu** *et al.*, 2019). Además, la interacción de los usuarios (*likes*, *retweets* y respuestas) con los tweets escritos por otros usuarios verificados también se puede utilizar con este fin (**Yang** *et al*., 2019).

Un tipo muy interesante de usuario de redes sociales son los *bots*. Son programas informáticos que llevan a cabo acciones autónomas, incluyendo la generación automática de información falsa y la amplificación de la desinformación durante las etapas iniciales de difusión (**Shao**; **Ciampaglia** *et al.*, 2018). Los *bots* suelen tener características particulares en las redes sociales; por ejemplo, suelen ser cuentas recientes (**Davis** *et al.*, 2016) con nombres de usuario largos y con caracteres raros (**Oehmichen** *et al.*, 2019). Su comportamiento también es diferente al de los humanos (**Ruffo** *et al.*, 2023); por ejemplo, retweetean más, reciben menos retweets, reciben menos respuestas y menciones, y publican menos tweets originales (**Ferrara** *et al.*, 2016). Todas estas características pueden extraerse de perfiles públicos y el gráfico de retweets, lo que permite la identificación automatizada de *bots*, ya sea utilizando estos datos por separado (**Des-Mesnards** *et al.*, 2022) o combinados con los datos de las publicaciones (**Kudugunta**; **Ferrara**, 2018).

La desinformación está estrechamente relacionada con la personalidad y los procesos mentales del usuario. Dado que las características psicológicas regulan el comportamiento y la interacción en el mundo físico, es lógico suponer que también repercuten en las comunidades virtuales. Los rasgos psicológicos pueden influir en la forma en que los individuos interpretan e interactúan con la información, aumentando la probabilidad de difundir información falsa y narrativas tóxicas. Por ejemplo, los sesgos cognitivos inherentes al ser humano, como la percepción limitada de la realidad y el sesgo de confirmación, pueden aumentar la probabilidad de percibir las noticias falsas como reales y fomentar así su propagación (**Shu** *et al.*, 2017). A diferencia de los difusores de información veraz, se ha observado que

> Las características empleadas para la clasificación de desinformación se dividen en dos categorías: basadas en el contenido y basadas en el contexto

los desinformadores son extrovertidos, menos neuróticos y presentan más estrés en sus tweets (**Shrestha**; **Spezzano**, 2022). Por el contrario, **Srinivas**, **Das** y **Pulabaigari** (2022) sugieren que los usuarios que difunden información política falsa son neuróticos, conservadores y presentan rasgos psicopáticos. La diferencia en las conclusiones de estos trabajos se debe principalmente a la forma de detectar y medir estos rasgos psicológicos.

4.2.2. Características basadas en el contexto de los mensajes

Las características contextuales del usuario y las basadas en el mensaje a menudo son a veces tratadas en conjunto (**Guo** *et al.*, 2020) e incluso en ocasiones son indistinguibles (**Yang** *et al.*, 2019). Aun así, para mayor claridad, consideramos por separado el contexto de los mensajes publicados, ya que estos mensajes presentan más diversidad que las características de los usuarios (**Tacchini** *et al.*, 2017). Así, los metadatos sobre las publicaciones en las redes sociales se han utilizado principalmente para aumentar la eficacia del análisis basado en otra característica principal (**Della-Vedova** *et al.*, 2018). Del mismo modo, los recursos multimedia asociados a los mensajes se han empleado para complementar los modelos de AA, dando lugar a análisis multimodales de desinformación (**Hangloo**; **Arora**, 2022).

El análisis multimodal se ha centrado hasta la fecha en las imágenes y se ha abordado de tres formas principales:

- forense: evalúa si una imagen ha sido objeto de modificación o manipulación (**Qi** *et al.*, 2019);
- contextual: valora si hay o no coherencia entre imagen y texto (**Kang**; **Hwang**; **Yu**, 2020; **Xiong** *et al.*, 2023);
- híbrida: la imagen se procesa para extraer información adicional que se utilizará en la clasificación (**Giachanou**; **Zhang**; **Rosso**, 2020; **Jing** *et al.*, 2023; **Khattar** *et al.*, 2019; **Li**; **Yao** *et al.*, 2022; **Singh** *et al.*, 2023; **Wang** *et al.*, 2018). Por ejemplo, **Zhang**, **Giachanou** y **Rosso** (2022) combinaron información textual, visual y contextual para construir la "escena" representada en la publicación, obteniendo diferencias estadísticamente significativas en la aparición de lugares específicos, el clima y las estaciones entre contenidos falsos y veraces.

4.2.3. Características basadas en la estructura de la red

Las características basadas en la red se refieren tanto a la estructura estática de la red social, como los nodos centrales, las comunidades basadas en las conexiones entre los usuarios, como a la propagación más dinámica de la (des)información, incluidos los actores críticos, las vías de difusión y el "contagio" de una comunidad a otra (**Bondielli**; **Marcelloni**, 2019; **Zhou**; **Zafarani**, 2020).

La mayoría de los trabajos científicos se centran en la detección de información falsa mediante el modelado de la red de difusión de información, asumiendo que la información verdadera y falsa tienen diferentes patrones de propagación (**De-Souza** *et al.*, 2022; **Liu**; **Wu**, 2018; **Liu**; **Xu**, 2016; **Song** *et al.*, 2022). Otras investigaciones han combinado el análisis de las rutas de propagación con las características de los propagadores para la clasificación de la desinformación (**Grinberg** *et al.*, 2019; **Shao**; **Ciampaglia** *et al.*, 2018; **Shao**; **Hui**; *et al.*, 2018). Este enfoque resulta altamente efectivo para frenar la propagación de la desinformación, ya que prioriza la identificación y eliminación de contenido desinformativo sobre el análisis de las publicaciones individuales, proceso este último que resulta más costoso. En concreto, se han investigado las características de red de los usuarios implicados en la difusión de información falsa a través de iniciativas como los retos PAN (**Buda**; **Bolonyai**, 2020; **Vogel**; **Meghana**, 2020). Además, recientemente se han aplicado técnicas modernas de AA a este problema; por ejemplo, **Rath**, **Salecha** y **Srivastava** (2022) propusieron un modelo de red neuronal de grafos para identificar nodos propensos a difundir información falsa utilizando la topología de la red y datos históricos de actividad de los usuarios.

5. Verificación de hechos asistida por IA

La verificación de hechos (más conocida por su término en inglés *fact-checking*) es un tipo de periodismo centrado en la comprobación de afirmaciones públicas (**Graves**; **Nyhan**; **Reifler**, 2016). Aunque la verificación de información es de por sí una parte fundamental del periodismo, el fact-checking hace hincapié en la relevancia del proceso de comprobación y en el desarrollo de métodos y programas informáticos para hacerlo de forma eficaz y transparente. Las primeras propuestas para automatizar la comprobación de hechos en línea aparecieron hace más de 15 años (**Graves**, 2018), destacando que la automatización completa es prácticamente imposible debido al juicio crítico, la sensibilidad y la experiencia necesarias para tomar una decisión que no sea binaria (**Arnold**, 2020). La comunidad de verificadores reconoce que la rápida difusión de información falsa presenta problemas de escalabilidad, es decir, difundir una mentira es mucho más rápido que desacreditarla (**Vosoughi**; **Roy**; **Aral**, 2018). No obstante, constatar este hecho no debería suponer sacrificar el rigor del proceso de comprobación de hechos.

En consecuencia, los distintos enfoques analizados tienden a un fact-checking asistido por herramientas de IA en lugar de la comprobación automatizada de hechos, por lo que se denominan sistemas "*human-in-the-loop*" (personas-en-el-proceso) (**La-Barbera**; **Roitero**; **Mizzaro**, 2022; **Shabani** *et al.*, 2021; **Yang** *et al.*, 2021). Las técnicas de IA pueden apoyar la verificación de información en distintas fases del flujo de trabajo de verificación (**Guo**; **Schlichtkrull**; **Vlachos** 2022; **Nakov**; **Corney** *et al.*, 2021):

> La tendencia actual se inclina hacia la verificación asistida por herramientas de inteligencia artificial en lugar de una comprobación de hechos completamente automatizada

(1) supervisión, reconocimiento y priorización de contenidos susceptibles de verificación;

(2) evaluación de la verificabilidad de las afirmaciones y priorización de temas;

(3) búsqueda de verificaciones anteriores que se apliquen al mismo caso;

(4) recuperación de pruebas para una investigación más profunda;

(5) clasificación semiautomatizada en categorías (bulo, contenido engañoso, contexto falso, etc.);

(6) difusión de las verificaciones;

(7) agilización de la redacción y documentación de las comprobaciones de hechos.

Las propuestas en la bibliografía se han centrado principalmente en las etapas 1 a 4. Para la etapa 5, podrían aplicarse las contribuciones descritas en la sección 3, aunque muestren limitaciones en su aplicabilidad a múltiples dominios como se ha descrito.

Se han propuesto varios métodos para comprobar la idoneidad de las afirmaciones (etapas 1 y 2). Por ejemplo, algunos se basan en la clasificación de las afirmaciones falsas mediante predicciones de puntuación, (**Kartal**; **Kutlu**, 2023; **Nakov**; **Da-San-Martino** *et al.*, 2021) mientras que otros utilizan anotaciones específicas para la clasificación de afirmaciones falsas (**Konstantinovskiy** *et al.*, 2021). Dado que los sistemas automatizados pueden introducir sesgos en la selección de afirmaciones a verificar, la investigación ha pivotado hacia aplicaciones como las alertas de noticias, el reconocimiento de voz y los modelos de traducción para mejorar la eficacia del filtrado de afirmaciones (**Rashkin** *et al.*, 2017).

La detección de afirmaciones previamente verificadas, incluidas las verificadas en otros idiomas o países, se ha abordado con técnicas de PLN y recuperación de información (etapas 3 y 4). En el primer caso, se ha aplicado la similitud semántica textual para emparejar las nuevas afirmaciones con las ya verificadas en inglés (**Thorne**; **Vlachos**, 2018) y español (**Martín** *et al.*, 2022). En el segundo caso, se han elaborado softwares con diferentes niveles de inteligencia para la recuperación de pruebas, incluyendo la extracción de datos estructurados, el reconocimiento de voz, la búsqueda inversa de imágenes, el análisis forense de vídeo o la búsqueda en lenguaje natural (**Das** *et al.*, 2023).

Una herramienta notable que cubre diferentes etapas es *InVid*, una plataforma gratuita que reúne varias aplicaciones para comprobar la fiabilidad y autenticidad de imágenes y vídeos:
https://www.invid-project.eu

Se espera que el proyecto *vera.ai* continúe y amplíe en Europa la investigación en programas informáticos y servicios de verificación apoyados en IA.
https://www.veraai.eu

6. Desafío de la generación automática de desinformación

Los modelos masivos del lenguaje presentados en la Sección 2.2, suponen uno de los principales desafíos actuales por su capacidad para generar desinformación textual a gran escala. Por ejemplo, *GPT-3* y *ChatGPT* pueden usarse de muy diversas formas para amplificar la difusión de la desinformación (**Solaiman** *et al.*, 2019):

- para camuflar los contenidos falsos bajo una apariencia de información real;
- para crear *bots* y webs que reproduzcan una narrativa desinformativa;
- para esquivar los detectores basados en características estilísticas, etc.

Además, como no se realiza un control demasiado exhaustivo sobre las fuentes que se usan para entrenar estos grandes modelos del lenguaje, gran parte de los contenidos que generan pueden ser falsos o estar sesgados (**Marcus**, 2022). En consecuencia, es de vital importancia desarrollar métodos para detectar la desinformación generada por los LLMs y mitigar su impacto. Desafortunadamente, los avances realizados hasta la fecha han sido poco efectivos (**Mitchell** *et al.*, 2023).

La desinformación generada automáticamente no se limita al texto; de hecho, existen técnicas de IA para crear imágenes, vídeos y audio que podrían ser aún más dañinas. El término ultrafalsificación (*deepfake*) denota los contenidos multimedia muy realistas generados con técnicas de AP, como las redes generativas adversarias (**Goodfellow** *et al.*, 2014). Estas ultrafalsificaciones pueden servir para crear avatares falsos, alterar el rostro o el discurso de una persona en un audio o un vídeo y sustituir personas y entornos en fotos y vídeos, entre otras muchas otras manipulaciones. Las ultrafalsificaciones se han utilizado con diversos fines desinformativos en los últimos tiempos, desde socavar la reputación de una persona hasta manipular procesos electorales (**Greengard**, 2019; **Masood** *et al.*, 2022). Para combatir las ultrafalsificaciones hay que comprender su proceso de generación para poder crear sistemas capaces de detectarlas (**Dagar**; **Vishwakarma**, 2022; **Mirsky**; **Lee**, 2022; **Saif**; **Tehseen**, 2022), aunque en la actualidad las tecnologías de generación van muy por delante de las de detección.

> Los modelos masivos del lenguaje y las ultrafalsificaciones representan uno de los principales desafíos actuales en la lucha contra la desinformación debido a su capacidad para generar desinformación de forma automática y a gran escala

La manipulación de caras en imágenes y vídeos es una de las áreas de investigación en generación de contenidos artificiales más activas y con mayor relación con la desinformación. En la bibliografía podemos encontrar técnicas tanto para generación de caras completas (**Serengil**; **Ozpinar**, 2021) utilizando arquitecturas de redes neuronales como *ProGAN* (**Karras** *et al.*, 2018) o *StyleGAN* (**Karras**; **Laine**; **Aila**, 2019), como para la manipulación parcial de las caras, como el intercambio de una cara por otra, la modificación de los atributos faciales (pelo, tono de piel, ojos, etc.), la animación de caras o la re-sincronización de los labios con un discurso diferente al original (**Tolosana** *et al.*, 2020). En sentido contrario, existe también un gran número de técnicas de AA para la detección de ultrafalsificaciones de caras. Por ejemplo, se han usado con cierto éxito las redes neuronales convolucionales con mecanismos de atención (**Dagar**; **Vishwakarma**, 2022; **Rana** *et al.*, 2022; **Tolosana** *et al.*, 2020), aunque su efectividad es muy limitada en comparación con los avances en generación de ultrafalsificaciones y la posibilidad de refinarlos manualmente en posproducción.

7. Conclusiones y trabajo futuro

> La detección temprana de la desinformación y la promoción de la alfabetización mediática son cruciales para mitigar su impacto

La cantidad de datos disponibles en la actualidad y la velocidad con la que se propagan hacen que sea difícil distinguir la información de la desinformación, pues a menudo esta última se disfraza de legítima y apela a las emociones y creencias más profundas de las personas. Las tecnologías computacionales son instrumentos adecuados para abordar la desinformación, pero también pueden exacerbar este problema a través de la invención y la falsificación de contenidos. En este artículo hemos descrito las tendencias actuales en IA y AA aplicadas a la detección y caracterización de desinformación, así como los desafíos que plantean por su capacidad para la generación sintética de textos elaborados e imágenes realistas.

La mayoría de las propuestas de la bibliografía se centran en el análisis a posteriori de la desinformación, identificando las características que se pueden emplear para su identificación automática. No obstante, si bien los enfoques existentes asumen que las soluciones para problemas y dominios específicos pueden extenderse a otros, en realidad son muy dependientes de los conjuntos de datos utilizados para su entrenamiento. Por este motivo, destacamos la necesidad de crear conjuntos de datos nuevos, de alta calidad y libres de sesgo, y particularmente en idiomas distintos al inglés. Asimismo, consideramos conveniente aumentar los esfuerzos de investigación en la transferencia de los modelos entrenados de un dominio a otro y la evaluación de su efectividad. Finalmente, cabe destacar que una gran cantidad de los sistemas de análisis de desinformación basados en IA aún no se encuentran ampliamente disponibles y/o carecen de la madurez necesaria para ser utilizadas por usuarios no técnicos.

Asumiendo que la erradicación total de la desinformación es imposible, tanto la detección temprana como la actitud de las personas hacia ella son cruciales para limitar su impacto. Por lo tanto, remarcamos como líneas de investigación prioritarias para el futuro las dos siguientes:

- el estudio de los patrones de creación y propagación de la desinformación para comprender mejor y anticipar la difusión de propaganda dañina y teorías de la conspiración;
- la aplicación de tecnologías inteligentes para amplificar el alcance de las verificaciones de hechos y la alfabetización mediática, de manera similar a como los desinformadores diseñan sus mensajes para llegar a audiencias más amplias.

Estas iniciativas requerirán la creación de métodos de IA explicables, capaces de justificar los resultados obtenidos y de facilitar la interacción de los profesionales de la información con las tecnologías; en particular, facilitando las tareas de los verificadores de hechos, de los expertos en seguridad y de los responsables de la toma de decisiones. Al abordar estos desafíos, avanzaremos hacia sistemas basados en IA que puedan detectar y combatir la desinformación de manera más efectiva, contribuyendo en última instancia a una sociedad mejor informada.

8. Referencias

Afroz, Sadia; **Brennan, Michael**; **Greenstadt, Rachel** (2012). "Detecting hoaxes, frauds, and deception in writing style online". In: *IEEE symposium on security and privacy*, pp. 461-475.
https://doi.org/10.1109/SP.2012.34

Aggarwal, Charu C. (2011). "An introduction to social network data analytics". In: Aggarwal, Charu C. (ed.). *Social network data analytics.* Springer.
https://doi.org/10.1007/978-1-4419-8462-3

Amador, Julio; **Molina-Solana, Miguel**; **Gómez-Romero, Juan** (2019). "Towards easy-to-implement misinformation automatic detection for online social media". In: *Proceedings of the conference for truth and trust online 2019*.
https://doi.org/10.36370/tto.2019.4

Arnold, Phoebe (2020). "The challenges of online fact checking". *Full fact*, 17 December.
https://fullfact.org/blog/2020/dec/the-challenges-of-online-fact-checking-how-technology-can-and-cant-help

Azevedo, Lucas; **D'Aquin, Mathieu**; **Davis, Brian**; **Zarrouk, Manel** (2021). "LUX (linguistic aspects under examination): discourse analysis for automatic fake news classification". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 41-56.
https://doi.org/10.18653/v1/2021.findings-acl.4

Barabási, Albert-László (2016). *Network science*. Cambridge University Press. ISBN: 978 1 107 07626 6
http://networksciencebook.com

Bedi, Punam; **Sharma, Chhavi** (2016). "Community detection in social networks". *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, v. 6, n. 3, pp. 115-135.
https://doi.org/10.1002/widm.1178

Bishop, Christopher M. (2006). *Pattern recognition and machine learning*. Springer. ISBN: 978 0 387 31073 2
https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

Blondel, Vincent D.; **Guillaume, Jean-Loup**; **Lambiotte, Renaud**; **Lefebvre, Etienne** (2008). "Fast unfolding of communities in large networks". *Journal of statistical mechanics: theory and experiment*, n. 10, pp. P10008.
https://doi.org/10.1088/1742-5468/2008/10/P10008

Bondielli, Alessandro; **Marcelloni, Francesco** (2019). "A survey on fake news and rumour detection techniques". *Information sciences*, v. 497, pp. 38-55.
https://doi.org/10.1016/j.ins.2019.05.035

Bonet-Jover, Alba; **Piad-Morffis, Alejandro**; **Saquete, Estela**; **Martínez-Barco, Patricio**; **García-Cumbreras, Miguel-Ángel** (2021). "Exploiting discourse structure of traditional digital media to enhance automatic fake news detection". *Expert systems with applications*, v. 169, 114340.
https://doi.org/10.1016/j.eswa.2020.114340

Brown, Tom B.; **Mann, Benjamin**; **Ryder, Nick**; **Subbiah, Melanie**; **Kaplan, Jared**; **Dhariwal, Prafulla**; **Neelakantan, Arvind**; **Shyam, Pranav**; **Sastry, Girish**; **Askell, Amanda**; **Agarwal, Sandhini**; **Herbert-Voss, Ariel**; **Krueger, Gretchen**; **Henighan, Tom**; **Child, Rewon**; **Ramesh, Aditya**; **Ziegler, Daniel M.**; **Wu, Jeffrey**; **Winter, Clemens**; **Hesse, Christopher**; **Chen, Mark**; **Sigler, Eric**; **Litwin, Mateusz**; **Gray, Scott**; **Chess, Benjamin**; **Clark, Jack**; **Berner, Christopher**; **McCandlish, Sam**; **Radford, Alec**; **Sutskever, Ilya**; **Amodei, Dario** (2020). "Language models are few-shot learners". *Advances in neural information processing systems*, v. 33, pp. 1877-1901.
https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Buda, Jakab; **Bolonyai, Flora** (2020). "An ensemble model using n-grams and statistical features to identify fake news spreaders on *Twitter*". In: *Working notes of CLEF 2020 - Conference and labs of the evaluation forum*, v. 2696.
https://ceur-ws.org/Vol-2696/paper_189.pdf

Camacho, David; **Panizo-Lledot, Ángel**; **Bello-Orgaz, Gema**; **González-Pardo, Antonio**; **Cambria, Erik** (2020). "The four dimensions of social network analysis: an overview of research methods, applications, and software tools". *Information fusion*, v. 63, pp. 88-120.
https://doi.org/10.1016/j.inffus.2020.05.009

Cambria, Erik; **Wang, Haixun**; **White, Bebo** (2014). "Guest editorial: big social data analysis". *Knowledge-based systems*, v. 69.
https://doi.org/10.1016/j.knosys.2014.07.002

Castelo, Sonia; **Almeida, Thais**; **Elghafari, Anas**; **Santos, Aécio**; **Pham, Kien**; **Nakamura, Eduardo**; **Freire, Juliana** (2019). "A topic-agnostic approach for identifying fake news pages". In: *Companion proceedings of the 2019 World Wide Web conference*, pp. 975-980.
https://doi.org/10.1145/3308560.3316739

Dagar, Deepak; **Vishwakarma, Dinesh K.** (2022). "A literature review and perspectives in deepfakes: generation, detection, and applications". *International journal of multimedia information retrieval*, v. 11, n. 3, pp. 219-289.
https://doi.org/10.1007/s13735-022-00241-w

Das, Anubrata; **Liu, Houjiang**; **Kovatchev, Venelin**; **Lease, Matthew** (2023). "The state of human-centered NLP technology for fact-checking". *Information processing & management*, v. 60, n. 2, 103219.
https://doi.org/10.1016/j.ipm.2022.103219

Davis, Clayton-Allen; **Varol, Onur**; **Ferrara, Emilio**; **Flammini, Alessandro**; **Menczer, Filippo** (2016). "BotOrNot: a system to evaluate social bots". In: *Proceedings of the 25th International conference companion on World Wide Web*, pp. 273-274.
https://doi.org/10.1145/2872518.2889302

Della-Vedova, Marco L.; **Tacchini, Eugenio**; **Moret, Stefano**; **Ballarin, Gabriele**; **DiPierro, Massimo**; **De-Alfaro, Luca** (2018). "Automatic online fake news detection combining content and social signals". In: *22nd Conference of open innovations association (Fruct)*, pp. 272-279.
https://doi.org/10.23919/FRUCT.2018.8468301

De-Souza, Mariana C.; **Nogueira, Bruno-Magalhães**; **Rossi, Rafael-Geraldeli**; **Marcacini, Ricardo-Marcondes**; **Dos-Santos, Brucce-Neves**; **Rezende, Solange-Oliveira** (2022). "A network-based positive and unlabeled learning approach for fake news detection". *Machine learning*, v. 111, n. 10, pp. 3549-3592.
https://doi.org/10.1007/s10994-021-06111-6

Del-Vicario, Michela; **Vivaldo, Gianna**; **Bessi, Alessandro**; **Zollo, Fabiana**; **Scala, Antonio**; **Caldarelli, Guido**; **Quattrociocchi, Walter** (2016). "Echo chambers: emotional contagion and group polarization on facebook". *Scientific reports*, v. 6, 37825.
https://doi.org/10.1038/srep37825

Des-Mesnards, Nicolas-Guenon; **Hunter, David-Scott**; **El-Hjouji, Zakaria**; **Zaman, Tauhid** (2022). "Detecting bots and assessing their impact in social networks". *Operations research*, v. 70, n. 1.
https://doi.org/10.1287/opre.2021.2118

Dong, Xishuang; **Victor, Uboho**; **Qian, Lijun** (2020). "Two-path deep semisupervised learning for timely fake news detection". *IEEE transactions on computational social systems*, v. 7, n. 6, pp. 1386-1398.
https://doi.org/10.1109/TCSS.2020.3027639

Ferrara, Emilio; **Varol, Onur**; **Davis, Clayton**; **Menczer, Filippo**; **Flammini, Alessandro** (2016). "The rise of social bots". *Communications of the ACM*, v. 59, n. 7, pp. 96-104.
https://doi.org/10.1145/2818717

Fortunato, Santo (2010). "Community detection in graphs". *Physics reports*, v. 486, n. 3-5, pp. 75-174.
https://doi.org/10.1016/j.physrep.2009.11.002

Ghanem, Bilal; **Ponzetto, Simone P.**; **Rosso, Paolo**; **Rangel, Francisco** (2021). "FakeFlow: fake news detection by modeling the flow of affective information". In: *Proceedings of the 16th Conference of the European chapter of the Association for Computational Linguistics*, pp. 679-689.
https://doi.org/10.18653/v1/2021.eacl-main.56

Giachanou, Anastasia; **Ghanem, Bilal**; **Ríssola, Esteban A.**; **Rosso, Paolo**; **Crestani, Fabio**; **Oberski, Daniel** (2022). "The impact of psycholinguistic patterns in discriminating between fake news spreaders and fact checkers". *Data & knowledge engineering*, v. 138, 101960.
https://doi.org/10.1016/j.datak.2021.101960

Giachanou, Anastasia; **Rosso, Paolo**; **Crestani, Fabio** (2019). "Leveraging emotional signals for credibility detection". In: *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pp. 877-880.
https://doi.org/10.1145/3331184.3331285

Giachanou, Anastasia; **Rosso, Paolo**; **Crestani, Fabio** (2021). "The impact of emotional signals on credibility assessment". *Journal of the Association for Information Science and Technology*, v. 72, n. 9, pp. 1117-1132.
https://doi.org/10.1002/asi.24480

Giachanou, Anastasia; **Zhang, Guobiao**; **Rosso, Paolo** (2020). "Multimodal multi-image fake news detection". In: *IEEE 7th International conference on data science and advanced analytics (DSAA)*, pp. 647-654.
https://doi.org/10.1109/DSAA49011.2020.00091

Goodfellow, Ian; **Bengio, Yoshua**; **Courville, Aaron** (2016). *Deep learning*. MIT Press. ISBN: 978 0 262 035613

Goodfellow, Ian J.; **Pouget-Abadie, Jean**; **Mirza, Mehdi**; **Xu, Bing**; **Warde-Farley, David**; **Ozair, Sherjil**; **Courville, Aaron**; **Bengio, Yoshua** (2014). "Generative adversarial nets". *Advances in neural information processing systems*, v. 27.
https://papers.nips.cc/paper/5423-generative-adversarial-nets

Graves, Lucas (2018). *Understanding the promise and limits of automated fact-checking*. Reuters Institute, University of Oxford.
https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-02/graves_factsheet_180226%20FINAL.pdf

Graves, Lucas; **Nyhan, Brendan**; **Reifler, Jason** (2016). "Understanding innovations in journalistic practice: a field experiment examining motivations for fact-checking". *Journal of communication*, v. 66, n. 1, pp. 102-138.
https://doi.org/10.1111/jcom.12198

Greengard, Samuel (2019). "Will deepfakes do deep damage?". *Communications of the ACM*, v. 63, n. 1, pp. 17-19.
https://doi.org/10.1145/3371409

Grinberg, Nir; **Joseph, Kenneth**; **Friedland, Lisa**; **Swire-Thompson, Briony**; **Lazer, David** (2019). "Fake news on *Twitter* during the 2016 U.S. presidential election". *Science*, v. 363, n. 6425, pp. 374-378.
https://doi.org/10.1126/science.aau2706

Guo, Bin; **Ding, Yasan**; **Yao, Lina**; **Liang, Yunji**; **Yu, Zhiwen** (2020). "The future of false information detection on social media: new perspectives and trends". *ACM computing surveys*, v. 53, n. 4.
https://doi.org/10.1145/3393880

Guo, Zhijiang; **Schlichtkrull, Michael**; **Vlachos, Andreas** (2022). "A survey on automated fact-checking". *Transactions of the Association for Computational Linguistics*, v. 10, pp. 178-206.
https://doi.org/10.1162/tacl_a_00454

Hangloo, Sakshini; **Arora, Bhavna** (2022). "Combating multimodal fake news on social media: methods, datasets, and future perspective". *Multimedia systems*, v. 28, n. 6, pp. 2391-2422.
https://doi.org/10.1007/s00530-022-00966-y

Harrigan, Paul; **Daly, Timothy M.**; **Coussement, Kristof**; **Lee, Julie A.**; **Soutar, Geoffrey N.**; **Evers, Uwana** (2021). "Identifying influencers on social media". *International journal of information management*, v. 56, 102246.
https://doi.org/10.1016/j.ijinfomgt.2020.102246

Jing, Jing; **Wu, Hongchen**; **Sun, Jie**; **Fang, Xiaochang**; **Zhang, Huaxiang** (2023). "Multimodal fake news detection via progressive fusion networks". *Information processing & management*, v. 60, n. 1, 103120.
https://doi.org/10.1016/j.ipm.2022.103120

John, Oliver P.; **Srivastava, Sanjay** (1999). "The big five trait taxonomy: history, measurement, and theoretical perspectives". In: Pervin, Lawrence A.; John, Oliver P. (eds.). *Handbook of personality: Theory and research*, pp. 102-138.
https://pages.uoregon.edu/sanjay/pubs/bigfive.pdf

Kang, SeongKu; **Hwang, Junyoung**; **Yu, Hwanjo** (2020). "Multi-modal component embedding for fake news detection". In: *14th international conference on ubiquitous information management and communication (Imcom)*.
https://doi.org/10.1109/IMCOM48794.2020.9001800

Karras, Tero; **Aila, Timo**; **Laine, Samuli**; **Lehtinen, Jaakko** (2018). "Progressive growing of GANs for improved quality, stability, and variation". In: *6th International conference on learning representations*.
https://research.nvidia.com/sites/default/files/pubs/2017-10_Progressive-Growing-of/karras2018iclr-paper.pdf

Karras, Tero; **Laine, Samuli**; **Aila, Timo** (2019). "A style-based generator architecture for generative adversarial networks". In: *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*, pp. 4396-4405.
https://doi.org/10.1109/CVPR.2019.00453

Kartal, Yavuz-Selim; **Kutlu, Mucahid** (2023). "Re-think before you share: a comprehensive study on prioritizing check-worthy claims". *IEEE transactions on computational social systems*, v. 10, n. 1, pp. 362-375.
https://doi.org/10.1109/TCSS.2021.3138642

Khattar, Dhruv; **Goud, Jaipal-Singh**; **Gupta, Manish**; **Varma, Vasudeva** (2019). "MVAE: multimodal variational autoencoder for fake news detection". In: *The World Wide Web conference*, pp. 2915-2921.
https://doi.org/10.1145/3308558.3313552

Konstantinovskiy, Lev; **Price, Oliver**; **Babakar, Mevan**; **Zubiaga, Arkaitz** (2021). "Toward automated factchecking: developing an annotation schema and benchmark for consistent automated claim detection". *Digital threats: research and practice*, v. 2, n. 2.
https://doi.org/10.1145/3412869

Kudugunta, Sneha; **Ferrara, Emilio** (2018). "Deep neural networks for bot detection". *Information sciences*, v. 467, pp. 312-322.
https://doi.org/10.1016/j.ins.2018.08.019

La-Barbera, David; **Roitero, Kevin**; **Mizzaro, Stefano** (2022). "A hybrid human-in-the-loop framework for fact checking". In: *Proceedings of the 6th Workshop on natural language for artificial intelligence (NL4AI 2022)*, v. 3287.
https://ceur-ws.org/Vol-3287/paper4.pdf

LeCun, Yann; **Bengio, Yoshua**; **Hinton, Geoffrey** (2015). "Deep learning". *Nature*, v. 521, n. 7553, pp. 436-444.
https://doi.org/10.1038/nature14539

Li, Dun; **Guo, Haimei**; **Wang, Zhenfei**; **Zheng, Zhiyun** (2021). "Unsupervised fake news detection based on autoencoder". *IEEE access*, v. 9, pp. 29356-29365.
https://doi.org/10.1109/ACCESS.2021.3058809

Li, Shuo; **Yao, Tao**; **Li, Saifei**; **Yan, Lianshan** (2022). "Semantic-enhanced multimodal fusion network for fake news detection". *International journal of intelligent systems*, v. 37, n. 12, pp. 12235-12251.
https://doi.org/10.1002/int.23084

Li, Xin; **Lu, Peixin**; **Hu, Lianting**; **Wang, Xiao-Guang**; **Lu, Long** (2022). "A novel self-learning semi-supervised deep learning network to detect fake news on social media". *Multimedia tools and applications*, v. 81, n. 14, pp. 19341-19349.
https://doi.org/10.1007/s11042-021-11065-x

Liu, Yang; **Wu, Yi-Fang** (2018). "Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks". *Proceedings of the AAAI conference on artificial intelligence*, v. 32, n. 1, pp. 354-361.
https://doi.org/10.1609/aaai.v32i1.11268

Liu, Yang; **Xu, Songhua** (2016). "Detecting rumors through modeling information propagation networks in a social media environment". *IEEE transactions on computational social systems*, v. 3, n. 2, pp. 46-62.
https://doi.org/10.1109/TCSS.2016.2612980

Manning, Christopher D.; **Schütze, Hinrich** (1999). *Foundations of statistical natural language processing*. MIT Press. ISBN: 978 0 262 133609

Marcus, Gary (2022). "AI platforms like *chatGPT* are easy to use but also potentially dangerous". *Scientific American*, 19 December.
https://www.scientificamerican.com/article/ai-platforms-like-chatgpt-are-easy-to-use-but-also-potentially-dangerous

Martín, Alejandro; **Huertas-Tato, Javier**; **Huertas-García, Álvaro**; **Villar-Rodríguez, Guillermo**; **Camacho, David** (2022). "FacTeR-check: semi-automated fact-checking through semantic similarity and natural language inference". *Knowledge-based systems*, v. 251, 109265.
https://doi.org/10.1016/j.knosys.2022.109265

Masood, Momina; **Nawaz, Mariam**; **Malik, Khalid M.**; **Javed, Ali**; **Irtaza, Aun**; **Malik, Hafiz** (2022). "Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward". *Applied intelligence*, v. 54, pp. 3974-4026.
https://doi.org/10.1007/s10489-022-03766-z

Meel, Priyanka; **Vishwakarma, Dinesh K.** (2020). "Fake news, rumor, information pollution in social media and web: a contemporary survey of state-of-the-arts, challenges and opportunities". *Expert systems with applications*, v. 153, 112986.
https://doi.org/10.1016/j.eswa.2019.112986

Meel, Priyanka; **Vishwakarma, Dinesh K.** (2021). "A temporal ensembling based semi-supervised convnet for the detection of fake news articles". *Expert systems with applications*, v. 177, 115002.
https://doi.org/10.1016/j.eswa.2021.115002

Megahed, Fadel M.; **Chen, Ying-Ju**; **Ferris, Joshua A.**; **Knoth, Sven**; **Jones-Farmer, L. Allison** (2023). "How generative AI models such as *chatGPT* can be (mis)used in SPC practice, education, and research? An exploratory study". *ArXiv*.
https://doi.org/10.48550/arXiv.2302.10916

Mikolov, Tomas; **Chen, Kai**; **Corrado, Greg**; **Dean, Jeffrey** (2013). "Efficient estimation of word representations in vector space". In: *1st International conference on learning representations (ICLR)*.
https://arxiv.org/abs/1301.3781

Mirsky, Yisroel; **Lee, Wenke** (2022). "The creation and detection of deepfakes". *ACM computing surveys*, v. 54, n. 1.
https://doi.org/10.1145/3425780

Mitchell, Eric; **Lee, Yoonho**; **Khazatsky, Alexander**; **Manning, Christopher D.**; **Finn, Chelsea** (2023). "DetectGPT: zero-shot machine-generated text detection using probability curvature". *ArXiv*.
https://doi.org/10.48550/arXiv.2301.11305

Molina-Solana, Miguel; **Amador, Julio**; **Gómez-Romero, Juan** (2018). "Deep learning for fake news classification". In: *I Workshop on deep learning*, pp. 1197-1201.
https://sci2s.ugr.es/caepia18/proceedings/docs/CAEPIA2018_paper_207.pdf

Nakamura, Kai; **Levy, Sharon**; **Wang, William Y.** (2020). "Fakeddit: a new multimodal benchmark dataset for fine-grained fake news detection". In: *Proceedings of the 12th International conference on language resources and evaluation*, pp. 6149-6157.
https://aclanthology.org/2020.lrec-1.755.pdf

Nakov, Preslav; **Corney, David**; **Hasanain, Maram**; **Alam, Firoj**; **Elsayed, Tamer**; **Barrón-Cedeño, Alberto**; **Papotti, Paolo**; **Shaar, Shaden**; **Da-San-Martino, Giovanni** (2021). "Automated fact-checking for assisting human fact-checkers". In: *Proceedings of the Thirtieth international joint conference on artificial intelligence (IJCAI)*, pp. 4551-4558.
https://doi.org/10.24963/ijcai.2021/619

Nakov, Preslav; **Da-San-Martino, Giovanni**; **Elsayed, Tamer**; **Barrón-Cedeño, Alberto**; **Míguez, Rubén**; **Shaar, Shaden**; **Alam, Firoj**; **Haouari, Fatima**; **Hasanain, Maram**; **Mansour, Watheq**; **Hamdan, Bayan**; **Ali, Zien-Sheikh**; **Babulkov, Nikolay**; **Nikolov, Alex**; **Shahi, Gautam-Kishore**; **Struß, Julia-Maria**; **Mandl, Thomas**; **Kutlu, Mucahid**; **Kartal, Yavuz-Selim** (2021). "Overview of the clef-2021 checkthat! Lab on detecting check-worthy claims, previously fact-checked claims, and fake news". In: *International conference of the cross-language evaluation forum for European languages*. *Experimental IR meets multilinguality, multimodality, and interaction*, pp. 264-291.
https://doi.org/10.1007/978-3-030-85251-1_19

Newman, Mark E. J. (2004). "Fast algorithm for detecting community structure in networks". *Physical review E*, v. 69, n. 6, 066133.
https://doi.org/10.1103/PhysRevE.69.066133

Oehmichen, Axel; **Hua, Kevin**; **Amador, Julio**; **Molina-Solana, Miguel**; **Gómez-Romero, Juan**; **Guo, Yi-ke** (2019). "Not all lies are equal. A study into the engineering of political misinformation in the 2016 US presidential election". *IEEE access*, v. 7, pp. 126305-126314.
https://doi.org/10.1109/ACCESS.2019.2938389

Paka, William-Scott; **Bansal, Rachit**; **Kaushik, Abhay**; **Sengupta, Shubhashis**; **Chakraborty, Tanmoy** (2021). "Cross-sean: a cross-stitch semi-supervised neural attention model for Covid-19 fake news detection". *Applied soft computing*, v. 107.
https://doi.org/10.1016/j.asoc.2021.107393

Pasi, Gabriella; **De-Grandis, Marco**; **Viviani, Marco** (2020). "Decision making over multiple criteria to assess news credibility in microblogging sites". In: *IEEE International conference on fuzzy systems (FUZZ-IEEE)*.
https://doi.org/10.1109/FUZZ48607.2020.9177751

Pennebaker, James W.; **Boyd, Ryan L.**; **Jordan, Kayla**; **Blackburn, Kate** (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin.
https://repositories.lib.utexas.edu/handle/2152/31333

Pennington, Jeffrey; **Socher, Richard**; **Manning, Christopher** (2014). "GloVe: global vectors for word representation". In: *Proceedings of the 2014 Conference on empirical methods in natural language processing (Emnlp)*, pp. 1532-1543.
https://doi.org/10.3115/v1/D14-1162

Qi, Peng; **Cao, Juan**; **Yang, Tianyun**; **Guo, Junbo**; **Li, Jintao** (2019). "Exploiting multi-domain visual information for fake news detection". In: *IEEE International conference on data mining (ICDM)*, pp. 518-527.
https://doi.org/10.1109/ICDM.2019.00062

Rana, Md-Shohel; **Nobi, Mohammad-Nur**; **Murali, Beddhu**; **Sung, Andrew H.** (2022). "Deepfake detection: a systematic literature review". *IEEE access*, v. 10, pp. 25494-25513.
https://doi.org/10.1109/ACCESS.2022.3154404

Rashkin, Hannah; **Choi, Eunsol**; **Jang, Jin Y.**; **Volkova, Svitlana**; **Choi, Yejin** (2017). "Truth of varying shades: analyzing language in fake news and political fact-checking". In: *Proceedings of the 2017 Conference on empirical methods in natural language processing*, pp. 2931-2937.
https://doi.org/10.18653/v1/D17-1317

Rath, Bhavtosh; **Salecha, Aadesh**; **Srivastava, Jaideep** (2022). "Fake news spreader detection using trust-based strategies in social networks with bot filtration". *Social network analysis and mining*, v. 12, n. 66.
https://doi.org/10.1007/s13278-022-00890-z

Ruffo, Giancarlo; **Semeraro, Alfonso**; **Giachanou, Anastasia**; **Rosso, Paolo** (2023). "Studying fake news spreading, polarisation dynamics, and manipulation by bots: a tale of networks and language". *Computer science review*, v. 47, 100531.
https://doi.org/10.1016/j.cosrev.2022.100531

Russell, Stuart; **Norvig, Peter** (2020). *Artificial intelligence: a modern approach*. Pearson Series. ISBN: 978 0 134 610993

Saif, Shahela; **Tehseen, Samabia** (2022). "Deepfake videos: synthesis and detection techniques - a survey". *Journal of intelligent and fuzzy systems*, v. 42, n. 4, pp. 2989-3009.
https://doi.org/10.3233/JIFS-210625

Schuster, Tal; **Schuster, Roei**; **Shah, Darsh J.**; **Barzilay, Regina** (2020). "The limitations of stylometry for detecting machine-generated fake news". *Computational linguistics*, v. 46, n. 2, pp. 499-510.
https://doi.org/10.1162/coli_a_00380

Serengil, Sefik I.; **Ozpinar, Alper** (2021). "HyperExtended lightface: a facial attribute analysis framework". In: *International conference on engineering and emerging technologies (Iceet)*.
https://doi.org/10.1109/ICEET53442.2021.9659697

Serrano-Guerrero, Jesús; **Olivas, José A.**; **Romero, Francisco P.**; **Herrera-Viedma, Enrique** (2015). "Sentiment analysis: a review and comparative analysis of web services". *Information sciences*, v. 311, pp. 18-38.
https://doi.org/10.1016/j.ins.2015.03.040

Shabani, Shaban; **Charlesworth, Zarina**; **Sokhn, Maria**; **Schuldt, Heiko** (2021). "SAMS: human-in-the-loop approach to combat the sharing of digital misinformation". *CEUR workshop proceedings*, v. 2846.
https://ceur-ws.org/Vol-2846/paper27.pdf

Shao, Chengcheng; **Ciampaglia, Giovanni-Luca**; **Varol, Onur**; **Yang, Kai-Cheng**; **Flammini, Alessandro**; **Menczer, Filippo** (2018). "The spread of low-credibility content by social bots". *Nature communications*, v. 9, n. 1, pp. 4787.
https://doi.org/10.1038/s41467-018-06930-7

Shao, Chengcheng; **Hui, Pik-Mai**; **Wang, Lei**; **Jiang, Xinwen**; **Flammini, Alessandro**; **Menczer, Filippo**; **Ciampaglia, Giovanni-Luca** (2018). "Anatomy of an online misinformation network". *Plos one*, v. 13, n. 4, e0196087.
https://doi.org/10.1371/journal.pone.0196087

Shrestha, Anu; **Spezzano, Francesca** (2022). "Characterizing and predicting fake news spreaders in social networks". *International journal of data science and analytics*, v. 13, n. 4, pp. 385-398.
https://doi.org/10.1007/s41060-021-00291-z

Shu, Kai; **Sliva, Amy**; **Wang, Suhang**; **Tang, Jiliang**; **Liu, Huan** (2017). "Fake news detection on social media: a data mining perspective". *ACM SIGKDD explorations newsletter*, v. 19, n. 1, pp. 22-36.
https://doi.org/10.1145/3137597.3137600

Shu, Kai; **Wang, Suhang**; **Liu, Huan** (2019). "Beyond news contents: the role of social context for fake news detection". In: *Proceedings of the 12th ACM International conference on web search and data mining*, pp. 312-320.
https://doi.org/10.1145/3289600.3290994

Shu, Kai; **Zhou, Xinyi**; **Wang, Suhang**; **Zafarani, Reza**; **Liu, Huan** (2019). "The role of user profiles for fake news detection". In: *Proceedings of the 2019 IEEE/ACM International conference on advances in social networks analysis and mining*, pp. 436-439.
https://doi.org/10.1145/3341161.3342927

Simko, Jakub; **Racsko, Patrik**; **Tomlein, Matus**; **Hanakova, Martina**; **Moro, Robert**; **Bielikova, Maria** (2021). "A study of fake news reading and annotating in social media context". *New review of hypermedia and multimedia*, v. 27, n. 1-2, pp. 97-127.
https://doi.org/10.1080/13614568.2021.1889691

Singh, Prabhav; **Srivastava, Ridam**; **Rana, K. P. S.**; **Kumar, Vineet** (2023). "SEMI-fnd: stacked ensemble based multimodal inferencing framework for faster fake news detection". *Expert systems with applications*, v. 215, 119302.
https://doi.org/10.1016/j.eswa.2022.119302

Solaiman, Irene; **Brundage, Miles**; **Clark, Jack**; **Askell, Amanda**; **Herbert-Voss, Ariel**; **Wu, Jeff**; **Radford, Alec**; **Krueger, Gretchen**; **Kim, Jong-Wook**; **Kreps, Sarah**; **McCain, Miles**; **Newhouse, Alex**; **Blazakis, Jason**; **McGuffie, Kris**; **Wang, Jasmine** (2019). "Release strategies and the social impacts of language models". *ArXiv*.
https://doi.org/10.48550/arXiv.1908.09203

Song, Chenguang; **Teng, Yiyang**; **Zhu, Yangfu**; **Wei, Siqi**; **Wu, Bin** (2022). "Dynamic graph neural network for fake news detection". *Neurocomputing*, v. 505, pp. 362-374.
https://doi.org/10.1016/j.neucom.2022.07.057

Srinivas, P. Y. K. L.; **Das, Amitava**; **Pulabaigari, Viswanath** (2022). "Fake spreader is narcissist; real spreader is Machiavellian prediction of fake news diffusion using psycho-sociological facets". *Expert systems with applications*, v. 207, 117952.
https://doi.org/10.1016/j.eswa.2022.117952

Stella, Massimo; **Ferrara, Emilio**; **De-Domenico, Manlio** (2018). "Bots increase exposure to negative and inflammatory content in online social systems". *Proceedings of the National Scademy of Sciences*, v. 115, n. 49, pp. 12435-12440.
https://doi.org/10.1073/pnas.1803470115

Tacchini, Eugenio; **Ballarin, Gabriele**; **Della-Vedova, Marco L.**; **Moret, Stefano**; **De-Alfaro, Luca** (2017). "Some like it hoax: automated fake news detection in social networks". In: *CEUR Workshop proceedings*, v. 1960.
https://arxiv.org/abs/1704.07506

Thorne, James; **Vlachos, Andreas** (2018). "Automated fact checking: task formulations, methods and future directions". In: *Proceedings of the 27th International conference on computational linguistics*, pp. 3346-3359.
https://aclanthology.org/C18-1283

Tolosana, Rubén; **Vera-Rodríguez, Rubén**; **Fierrez, Julián**; **Morales, Aythami**; **Ortega-García, Javier** (2020). "Deepfakes and beyond: a survey of face manipulation and fake detection". *Information fusion*, v. 64, pp. 131-148.
https://doi.org/10.1016/j.inffus.2020.06.014

Vaswani, Ashish; **Shazeer, Noam**; **Parmar, Niki**; **Uszkoreit, Jakob**; **Jones, Llion**; **Gomez, Aidan N.**; **Kaiser, Łukasz**; **Polosukhin, Illia** (2017). "Attention is all you need". In: *31st Conference on neural information processing systems*.
https://papers.neurips.cc/paper/7181-attention-is-all-you-need.pdf

Vogel, Inna; **Meghana, Meghana** (2020). "Fake news spreader detection on *Twitter* using character n-grams". In: *CEUR Workshop proceedings*, v. 2696.
https://ceur-ws.org/Vol-2696/paper_59.pdf

Vosoughi, Soroush; **Roy, Deb**; **Aral, Sinan** (2018). "The spread of true and false news online". *Science*, v. 359, n. 6380, pp. 1146-1151.
https://doi.org/10.1126/science.aap9559

Wang, Tingting; **Liu, Hongyan**; **He, Jun**; **Du, Xiaoyong** (2013). "Mining user interests from information sharing behaviors in social media". In: *Pacific-Asia conference on knowledge discovery and data mining*, pp. 85-98.
https://doi.org/10.1007/978-3-642-37456-2_8

Wang, William Y. (2017). "'Liar, liar pants on fire': a new benchmark dataset for fake news detection". In: *55th Annual meeting of the Association for Computational Linguistics*, v. 2, pp. 422-426.
https://doi.org/10.18653/v1/P17-2067

Wang, Yaqing; **Ma, Fenglong**; **Jin, Zhiwei**; **Yuan, Ye**; **Xun, Guangxu**; **Jha, Kishlay**; **Su, Lu**; **Gao, Jing** (2018). "EANN: event adversarial neural networks for multi-modal fake news detection". In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 849-857.
https://doi.org/10.1145/3219819.3219903

Wardle, Claire; **Derakhshan, Hossein** (2017). *Information disorder: toward an interdisciplinary framework for research and policy making*. Council of Europe report.
https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

Xiong, Shufeng; **Zhang, Guipei**; **Batra, Vishwash**; **Xi, Lei**; **Shi, Lei**; **Liu, Liangliang** (2023). "Trimoon: two-round inconsistency-based multi-modal fusion network for fake news detection". *Information fusion*, v. 93, pp. 150-158.
https://doi.org/10.1016/j.inffus.2022.12.016

Xu, Fan; **Sheng, Victor S.**; **Wang, Mingwen** (2023). "A unified perspective for disinformation detection and truth discovery in social sensing: a survey". *ACM computing surveys*, v. 55, n. 1.
https://doi.org/10.1145/3477138

Yang, Jing; **Vega-Oliveros, Didier**; **Seibt, Tais**; **Rocha, Anderson** (2021). "Scalable fact-checking with human-in-the-loop". In: *IEEE International workshop on information forensics and security (WIFS)*.
https://doi.org/10.1109/WIFS53200.2021.9648388

Yang, Shuo; **Shu, Kai**; **Wang, Suhang**; **Gu, Renjie**; **Wu, Fan**; **Liu, Huan** (2019). "Unsupervised fake news detection on social media: a generative approach". *Proceedings of the AAAI Conference on artificial intelligence*, v. 33, n. 1, pp. 5644-5651.
https://doi.org/10.1609/aaai.v33i01.33015644

Yin, Zhijun; **Cao, Liangliang**; **Gu, Quanquan**; **Han, Jiawei** (2012). "Latent community topic analysis". *ACM transactions on intelligent systems and technology*, v. 3, n. 4.
https://doi.org/10.1145/2337542.2337548

Zhang, Guobiao; **Giachanou, Anastasia**; **Rosso, Paolo** (2022). "SceneFND: multimodal fake news detection by modelling scene context information". *Journal of information science*, Online first.
https://doi.org/10.1177/01655515221087683

Zhang, Xichen; **Ghorbani, Ali A.** (2020). "An overview of online fake news: characterization, detection, and discussion". *Information processing and management*, v. 57, n. 2.
https://doi.org/10.1016/j.ipm.2019.03.004

Zhou, Xinyi; **Jain, Atishay**; **Phoha, Vir V.**; **Zafarani, Reza** (2020). "Fake news early detection". *Digital threats: research and practice*, v. 1, n. 2.
https://doi.org/10.1145/3377478

Zhou, Xinyi; **Zafarani, Reza** (2020). "A survey of fake news: fundamental theories, detection methods, and opportunities". *ACM computing surveys*, v. 53, n. 5.
https://doi.org/10.1145/3395046

Zhu, Q.; **Luo, J.** (2022). "Generative pre-trained transformer for design concept generation: an exploration". *Proceedings of the design society*, v. 2, pp. 1825-1834.
https://doi.org/10.1017/pds.2022.185

9. Anexo. Conjuntos de datos

Tabla 1. Conjuntos de datos utilizados en la bibliografía para entrenar modelos de clasificación de desinformación

Nombre	Aplicación	Fuente	Tamaño	Fuente de información	Etiquetas	Anotación	Características	Idioma	Disponibilidad pública	URL
CREDBANK	Evaluación de la credibilidad	*Twitter*	> 60 millones	Publicaciones en redes sociales sobre 1.049 eventos	Tupla <grado (seguro, probable, incierto), polaridad (exacto, inexacto, incierto)>	*Mechanical Turk*	Características basadas en el contenido y en el contexto	Inglés	Sí	*https://compsocial.github.io/CREDBANK-data/*
PHEME	Detección de rumores	*Twitter*	5.802	Publicaciones en redes sociales sobre 5 eventos	Rumor (1.972), no-rumor (3.830)	Anotación por expertos	Características basadas en el contenido	Inglés	Sí	*https://figshare.com/articles/dataset/PHEME_dataset_of_rumours_and_non-rumours/4010619*
LIAR	Detección de información falsa	*PolitiFact.com*	12.800	Declaraciones políticas	Mentira (*pants on fire*) (1.047), falso (2.507), apenas-verdadero (2.103), medio-verdadero (2.627), casi-verdadero (2.454) y verdadero (2.053)	Anotación por expertos	Características basadas en el contenido y en el contexto	Inglés	Sí	*https://www.cs.ucsb.edu/~william/data/liar_dataset.zip*
FakeNewsNet	Estudiar la información falsa en las redes sociales	*BuzzFeed.com* y *PolitiFact.com*	422	Noticias	Falso (211), real (211)	Anotación por expertos	Características basadas en el contenido y en el contexto	Inglés	Sí	*https://github.com/KaiDMML/FakeNewsNet*
MuMiN	Detección de información errónea (*misinforma-tion*)	*Twitter* y 115 organizaciones de verificación	12.914 afirmaciones verificadas y 21.565.018 tweets	Publicaciones en redes sociales y afirmaciones contrastadas	Información errónea, hecho	Semiautomática	Características basadas en el contenido y en el contexto	Multidioma	Sí	*https://mumin-dataset.github.io/gettingstarted/*
MediaEval	Detección de información errónea y conspiraciones	*Twitter*	3.389	Publicaciones en redes sociales	Promueve/apoya la conspiración, debate sobre conspiración y no conspiración	Anotación por expertos	Características basadas en el contenido y en el contexto	Inglés	Bajo petición	*https://multimediaeval.github.io*
BuzzFeedNews dataset	Detección de contenido falso	*Facebook*	2.282	Publicaciones en redes sociales procedentes de 9 fuentes (3 de tendencia derechista, 3 de tendencia izquierdista y 3 fidedignas)	Muy verdadero (1.669), sin contenido fáctico (264), mezcla de verdadero y falso (245), muy falso (104)	Anotación por expertos	Características basadas en el contenido y en el contexto	Inglés	Sí	*https://webis.de/data/buzzfeed-webis-fake-news-16.html*
BuzzFace dataset	Detección de contenido falso y bots	*Facebook*	> 1,6 millones	Publicaciones en redes sociales verificadas por *BuzzFeed* más comentarios y reacciones sobre estas publicaciones	Sólo se etiquetan los datos de origen (conjunto de datos de *BuzzFeedNews*)	Anotación por expertos	Características basadas en el contenido y en el contexto	Inglés	Sí	*https://github.com/gsantia/BuzzFace*
FacebookHoax	Detección de bulos	*Facebook*	15.500	Publicaciones en redes sociales de 32 páginas (14 conspiratorias y 18 científicas)	Bulo (8.923), no bulo (6.577)	En base a la temática de la página	Características basadas en el contenido y en el contexto	Inglés	Sí	*https://github.com/gabll/some-like-it-hoax*
FACTOID	Detección de difusores de contenido falso	*Reddit*	4.150	3.354.450 publicaciones en redes sociales de 4.150 usuarios	Difusor de noticias (3.071), difusor de contenido falso (1.079)	Anotación automática guiada por expertos	Características basadas en el contenido y en el contexto	Inglés	Sí	*https://github.com/caisa-lab/FACTOID-dataset*
Spanish Fake News Corpus	Detección de contenido falso	Webs de noticias	971	Noticias sobre 9 temas diferentes	Falso (480), real (491)	Anotación por expertos	Características basadas en el contenido	Español	Sí	*https://github.com/jpposadas/FakeNewsCorpusSpanish*

Nombre	Aplicación	Fuente	Tamaño	Fuente de información	Etiquetas	Anotación	Características	Idioma	Disponibilidad pública	URL
Spanish Fake News Corpus 2.0	Detección de contenido falso	Webs de noticias y redes sociales	1.543	Noticias y publicaciones en redes sociales sobre 12 temas diferentes	Falso (766), real (777)	Anotación por expertos	Características basadas en el contenido	Español	Sí	*https://github.com/jpposadas/FakeNewsCorpusSpanish*
NLI19-SP	Detección de información errónea	*Twitter*	46.919	Publicaciones en redes sociales relacionadas con un conjunto de 61 bulos identificados por verificadores	Contradicción (406), vínculo (2.521), neutral (43.992)	Anotación automática	Características basadas en el contenido y en el contexto	Español e inglés	Bajo petición	*https://aida.etsisi.upm.es/download/nli19-sp-dataset-facter-check*
PAN-AP-2020 corpus	Detección de difusores de contenido falso	*Twitter*	500	Usuarios de medios sociales a partir de noticias publicadas en Twitter	Difusor de noticias (250), difusor de contenido falso (250)	Anotación por expertos	Características basadas en el contenido y en el contexto	Español e inglés	Bajo petición	*https://zenodo.org/record/4039435#.Y2z2fi8ryRs*