La inteligencia artificial en la mitigación de desinformación en la gobernanza del sistema democrático.

Víctor Hugo Montero López
Universidad Técnica de Manabí (UTM)
vmontero9383@utm.edu.ec
Portoviejo, Ecuador
Enrique Macias Arias
Universidad Técnica de Manabí (UTM)
enrique.macias@utm.edu.ec
Portoviejo, Ecuador

DOI: https://doi.org/10.56124/encriptar.v9i17.003

RESUMEN

Esta investigación presenta el desarrollo de un marco de evaluación y protección basado en los modelos algorítmicos de IA para mitigar la desinformación en la gobernanza democrática del Ecuador, para ello se utilizó la investigación aplicada con enfoque mixto, incorporado con una revisión de literatura, simulaciones de difusión en redes sociales y clasificación de noticias, contextualizado con encuestas a personas relevantes del ecosistema mediático. según los hallazgos obtenidos se evidenció en la simulación de redes sociales la formación de cámaras de eco y la amplificación de narrativas afines, fenómeno intensificado por los algoritmos de recomendación que priorizan el compromiso del usuario; para mitigar la difusión en nuestra simulación de modelos automatizados para la detección de noticias falsas se logró una precisión de 91,67%, sin embargo, un hallazgo clave fue referido a que las decisiones de los modelos, los que se apoyan en artefactos estadísticos y sesgos estilísticos del dataset más que en una comprensión semántica de la veracidad, revelando fragilidad y limitada generalización. En ese sentido se desarrolló un marco de evaluación y protección, orientado a garantizar transparencia, minimizar sesgos y fortalecer la confianza pública, para afianzar la gobernanza democrática, se propuso lineamientos concretos, como la exigencia de interpretabilidad obligatoria y un ciclo de vida de operaciones de aprendizaje automático, para combatir la obsolescencia del modelo, priorizando en un nivel ético la mitigación de la amplificación sobre la censura.

Palabras clave: inteligencia artificial; desinformación; aprendizaje automático; gobernanza democrática; ética algorítmica.

Artificial Intelligence in the Mitigation of Disinformation in the Governance of the Democratic System

ABSTRACT

This research presents the development of an evaluation and protection framework based on AI algorithmic models to mitigate disinformation in the democratic governance of Ecuador; to this end, an applied mixed-methods approach was used, incorporating a literature review, simulations of diffusion on social networks, and news classification, contextualized with surveys of relevant actors in the media ecosystem. According to the findings obtained, the social-network simulation showed the formation of echo chambers and the amplification of like-minded narratives, a phenomenon intensified by recommendation algorithms that prioritize user engagement; to mitigate diffusion in our simulation, automated models for detecting fake news achieved 91.67% accuracy; however, a key finding was that the models' decisions rely on statistical artifacts and stylistic biases of the dataset rather than on a semantic understanding of veracity, revealing fragility and limited generalization. In this sense, an evaluation and protection framework was developed, aimed at ensuring transparency, minimizing biases, and strengthening public trust to bolster democratic governance; concrete guidelines were proposed, such as the requirement of mandatory interpretability and a machine-learning operations lifecycle to combat model obsolescence, ethically prioritizing the mitigation of amplification over censorship.

Keywords: artificial intelligence; disinformation; machine learning; democratic governance; algorithmic ethics.

1. Introducción

La desinformación, nos enfrenta a potenciar problemas en nuestra sociedad, este fenómeno amenaza la gobernanza en los sistemas democráticos porque la desinformación distorsiona la percepción pública y genera que se socave los principios, valores y la institucionalidad democrática de un país, este fenómeno amenaza tres pilares centrales de la gobernanza democrática: representación, rendición de cuentas y, en última instancia, la moneda más importante en un sistema político: la confianza (Kreps & Kriner, 2023).

En ese sentido, para logar una comprensión amplia de la desinformación es preciso el estudio del ecosistema en el que se desarrolla (Duarte & Magallón-Rosa, 2023), por un lado, las empresas emplean algoritmos de recomendación para filtrar grandes volúmenes de contenido y presentarlo a los usuarios de manera que se maximice su atención (Sun, 2023), el cual fomenta la homofilia (Siderius, 2023) y crea cámaras de eco lo que, sumado a la lógica de la viralidad de las plataformas, acelera estructuralmente la propagación del contenido (García-Orosa, 2021).

Ante esta problemática, la IA emerge como herramienta para detectar y combatir la desinformación, puesto que los algoritmos avanzados y las técnicas de aprendizaje automático permiten analizar grandes cantidades de datos en tiempo real e identificar patrones (Sánchez Gonzales & Alonso-González, 2024), utilizando diversas técnicas métricas de precisión para detectarla como la exactitud, la precisión, el recall y la puntuación F1-score, se plantean modelos como el aprendizaje automático y el procesamiento de lenguaje natural, los cuales establecen bases confiables para detectar patrones dentro de contenidos textuales (López López et al., 2023), en ese sentido los hallazgos revelaron la efectividad de estos modelos para identificar patrones de desinformación y la propagación de narrativas falsas como menciona (Mahmood & Akar, 2024), a través de la incrustación de palabras preentrenadas (Kaliyar et al., 2020), se emplean algoritmos de aprendizaje para la clasificación, teniendo como resultado los modelos que demuestran resultados de vanguardia para predecir noticias falsas (Kaliyar et al., 2020). los algoritmos que alcanzaron mejor exactitud para detectar noticias falsas fueron, el SVM, la regresión logística y el algoritmo de potenciación de gradiente (Álvarez-Daza et al., 2021).

Por tanto, estos estudios y la mayoría de las investigaciones se enmarcan en la precisión de los modelos, con escasa consideración y análisis

a dos factores importantes: 1. Sesgo y mecanismos internos de los modelos de la IA, empleados a la detección de desinformación y 2. El planteamiento de que la desinformación implica un problema dual que corresponde a contenido y propagación, el estudio procura cubrir esta brecha a través de la valoración de eficiencia, interpretabilidad de sesgo en un entorno simulado de red social, de allí su pertinencia con las investigaciones previas.

Por esta razón, es fundamental desarrollar un marco de evaluación y protección, que permita brindar conocimientos sobre los algoritmos que actúan en los modelos de IA para combatir los efectos negativos de la desinformación, esto hace relevante el estudio de su aplicabilidad como instrumento en la mitigación de la desinformación; además de orientar y guiar su implementación, por otro lado contribuiría a comprender la magnitud del fenómeno considerando que la sociedad necesita educarse en alfabetización digital, específicamente en el reconocimiento de los patrones de la IA generativa, para ir enfrentando los desafíos que este fenómeno representa.

Por eso, el objetivo del estudio es desarrollar un marco de evaluación y protección, que permita evaluar su impacto dentro de la dimensión humana en su uso y aplicación, lo que obliga a consensuar un código de conducta ética (Martínez, 2024), es decir que debemos considerar los valores éticos haciendo énfasis en la transparencia y la veracidad, logrando de esta forma que estas tecnologías contribuyan a enfrentar los desafíos que este fenómeno representa en la gobernanza democrática del Ecuador. También desarrollar capacidades técnicas como modelos de IA basados en datos, los cuales serán objeto del presente estudio, es así como se pone de manifiesto el compromiso del análisis contextual del texto de los modelos de IA con el propósito de establecer modelos entrenados con un conjunto de datos que tome en cuenta las características lingüísticas locales y adaptarlos a su vez a las características culturales y políticas que presentan países como Ecuador, donde la desinformación tiene cualidades propias.

2. Métodos y herramientas

El estudio se desarrolla bajo un tipo de investigación aplicada, con un enfoque mixto cualitativo-cuantitativo para comprender y abordar la problemática del impacto de la inteligencia artificial en la propagación de la desinformación y sus implicaciones en la gobernanza democrática, dando lugar al análisis a través de una revisión sistemática de la literatura, compaginado con el método experimental de simulación, así como percepciones, por medio de encuestas aplicada a participantes cualificados en el contexto ecuatoriano.

2.1 Revisión de la Literatura

Se realizó una revisión rigurosa de la literatura sobre desinformación y su mitigación algorítmica, con referentes como Siderius, Narayanan, Mahmood y Akar, Kaliyar, Innerarity, entre otros, y documentos de la UNESCO. La búsqueda se ejecutó en Scopus, Google Scholar, SciELO y repositorios, aplicando curación de datos, resúmenes analíticos y codificación temática, durante la fase se mapeó, evaluó y sintetizó la evidencia para sustentar teóricamente el estudio y construir el marco de evaluación y protección, delimitando la influencia de los algoritmos y el rol de los modelos de detección dentro de marcos éticos.

2.2 Experimentos de Simulación Computacional

El diseño experimental de la investigación se constituyó en dos simulaciones computacionales complementarias, fundamentándose en los principios del diseño y análisis de experimentos de simulación, lo que busca asegurar el rigor, la validez y la replicabilidad de los resultados (Keijnen, 2015), siendo así, el proceso se estructuró en tres simulaciones complementarias:

2.2.1 Simulación 1: Dinámica de Propagación en Red Social

Se desarrolló una simulación basada en agentes, adaptada del modelo bayesiano de (Siderius, 2023), donde cada agente es un actor racional pero sesgado que decide compartir, desaprobar o ignorar en función de la

actualización bayesiana de creencias y la utilidad esperada; el propósito fundamental de la simulación fue entender la lógica de la viralidad propia de las plataformas (García-Orosa, 2021) dentro de un ecosistema digital segregado, donde el medio se convierte en un amplificador y altavoz de determinados contenidos el más relevante en este nuevo escenario (Duarte & Magallón-Rosa, 2023), los factores clave analizados fueron el nivel de homofilia y el sesgo ideológico, esperando medir su impacto en la tasa de propagación del contenido y la polarización estructural de la red social.

La implementación de la simulación se realizó en el lenguaje de programación Python utilizando la librería NetworkX para la construcción y visualización de la estructura de la red, nuestro modelo cuenta con 300 agentes, con una ejecución simultanea de 50 rondas, teniendo una homofilia fijada en un valor de 0,8 en una escala del 0 al 1 y en cada ronda de la simulación se sembró una noticia que se propagó por la red, se espera que cada agente, racional pero sesgado, actualiza su creencia bayesiana π combinando su sesgo previo con una credibilidad cuadrática $\phi(r) = r^2$, con $r \in [0,1]$, donde decide compartir, desaprobar o ignorar maximizando utilidad esperada.

Así mismo, la utilidad de compartir se modeló como $\pi * \mathrm{B}_{verdad} - (1 - \pi) * C_{mentira} + k$, con $k = 0{,}1 + Beta(\alpha, \beta)$, para reflejar un incentivo social variable por popularidad de contenido, siguiendo la intuición de incentivos y reputación del modelo (Siderius, 2023).

Al ejecutar el modelo con la configuración descrita fue posible generar una representación visual de la segregación de la red, mediante la tasa de propagación y la polarización estructural de la red, se permite evaluar el impacto estructural de la homofilia en una dinámica de desinformación donde los agentes actúan de manera racional, pero están influenciados por sesgos e incentivos sociales.

2.2.2 Simulación 2: Clasificación de Contenido

En el segundo experimento, se analizó la clasificación automática de noticias falsas y verdaderas mediante el entrenamiento y la evaluación de modelos de aprendizaje con el objetivo de determinar el modelo con mayor eficiencia en aspectos que van desde el preprocesamiento del texto hasta la optimización y la clasificación de las noticias.

Datos y preparación

Se utilizó el dataset spanishFakeNews.csv, que contiene 8,781 noticias rotuladas en un 57% verdaderas y un 43% falsas, se incorporaron así mismo pruebas de sensibilidad y de generalización para valorar la robustez de los modelos. El pipeline se implementó en Python y comprendió normalización a minúsculas, desacentuación y limpieza de caracteres no alfabéticos, tokenización para español y filtrado de stopwords ampliado con términos de alta frecuencia del propio corpus, la representación del texto se realizó con TF-IDF, limitando el vocabulario a 10,000 términos y usando n-gramas de 1–2 para capturar coocurrencias locales.

Modelado, optimización y evaluación

Se entrenaron tres familias de modelos: Random Forest (ensamble de árboles con votación), SVM (clasificador de margen máximo) y una LSTM bidireccional para capturar dependencias secuenciales del lenguaje. Los hiperparámetros de RF y SVM se ajustaron mediante búsqueda en malla con validación cruzada y, posteriormente, ambos se integraron en un VotingClassifier para promediar decisiones y mejorar estabilidad y precisión. El rendimiento se midió con métricas estándar (accuracy, precisión, recall y F1-score), utilizando particiones reproducibles y control de semillas para asegurar comparabilidad y trazabilidad de los resultados.

2.2.3 Simulación 3: Validación del marco

Para validar la efectividad de los lineamientos propuestos, se implementó un modelo que replica la primera simulación basada en agentes, en una red de 300 agentes interconectados según principios de homofilia, nuevamente con un coeficiente de 0.8, donde cada agente posee un sesgo ideológico específico y toma decisiones sobre compartir, desaprobar o ignorar contenido según su creencia bayesiana actualizada sobre la veracidad de la información.

En este caso, el modelo incorpora cuatro escenarios experimentales para evaluar el impacto diferencial de las intervenciones propuestas, en primer lugar tenemos un escenario control muy parecido a la primera simulación sin control, luego tenemos un escenario de mitigación algorítmica donde se implemente nuestro clasificador de IA previamente estudiado en la simulación 2, además de un escenario de verificador donde un 5% de los 300 agentes sirven como agente verificadores con mayor compresión critica y por lo tanto un menor incentivo social para difundir noticias falsas y por último un escenario que combina la mitigación algorítmica con el escenario de verificadores.

Cada escenario ejecutó 50 rondas de difusión de contenido con características aleatorias de veracidad y confiabilidad, registrando las acciones agregadas de compartir y desaprobar como indicadores de viralidad y resistencia crítica respectivamente.

2.3 Aplicación de Encuesta

La tercera técnica empleada fue la encuesta, aplicando un cuestionario como instrumento de apoyo, en el contexto ecuatoriano, que validan los hallazgos de la revisión y de la simulación, midiendo percepciones, conocimientos y actitudes de actores que interactúan con desinformación, el instrumento adopta una estructura con tres bloques: percepción del impacto de plataformas y algoritmos en la opinión pública; conocimiento y eficacia percibida de modelos de IA para detección (LSTM, SVM, Random Forest) y su

viabilidad de implementación; y necesidad/contenido de lineamientos ético-técnicos adaptados al país, se combinan preguntas cerradas tipo Likert de 5 puntos y reactivos de sondeo para recoger razonamientos y propuestas.

El muestreo es no probabilístico por juicio, orientado a informantes con experiencia relevante; se definen dos poblaciones objetivo de 20 casos cada una, es decir n=40: comunicación social/personalidades públicas y tecnología, los marcos muestrales se construyen a partir de personal de medios y estudiantes de comunicación, priorizando experiencia en redes.

La recolección se realiza en línea a través de Google Forms y de forma presencial, garantizando anonimato y consentimiento informado. El trabajo de campo se extendió durante cuatro semanas, el análisis de este es interpretativo con apoyo descriptivo, orientado a identificar tendencias, consensos y brechas que contextualicen y respalden los lineamientos del marco propuesto.

3. Resultados y discusión

3.1 Resultados

Durante el desarrollo del estudio se evidenció que la mitigación de la desinformación es un problema que debe ser abordado desde un enfoque sistémico, porque automatizando únicamente la detección del contenido no lograría resolver completamente el problema, se demostró de la misma forma que se pudo desarrollar un modelo con alta precisión numérica en la detección, la misma aun presenta deficiencia en la comprensión del contexto semántico. En este contexto tomando como referencia los objetivos planteados en el estudio, se exponen los siguientes resultados:

3.1.1 Influencia de los algoritmos en la amplificación de la desinformación en la opinión pública

3.1.1.1 Revisión de la Literatura

El estudio evidenció el impacto que tienen los algoritmos en redes sociales sobre la difusión de desinformación, mostrando cómo estos sistemas

complejos generan patrones de retroalimentación derivados de la interacción entre usuarios y plataformas (Narayanan, 2023), indudablemente los algoritmos de recomendación al filtrar grandes volúmenes de contenido para maximizar la atención del usuario amplifican la propagación de mensajes, incluida la información falsa o sesgada (Sun, 2023), de esta forma las plataformas actúan como actores políticos que influyen en la opinión pública mediante estrategias de polarización y *astroturfing (de posverdad)* (García-Orosa, 2021), por su parte (McLoughlin & Brady, 2024) explican que la interacción entre sesgos atencionales humanos y amplificación algorítmica crea un ecosistema donde la información moral y emocional es privilegiada, permitiendo que la desinformación que explota estas características se propague en línea, por lo que el análisis de redes permite identificar cómo la desinformación se propaga a través de cuentas centrales y redes inauténticas que amplifican estos mensajes (Vysotska & Nazarkevych, 2025).

3.1.1.2 Simulación de Dinámica de Red Social

Tomando en consideración lo anteriormente expuesto, se realizó una simulación de red social basado en el planteamiento de agentes de (Siderius, 2023), a fin de determinar la influencia de los algoritmos de las plataformas en la amplificación de la desinformación, se observa en un escenario propuesto con un alto grado de homofilia, la formación de clústeres ideológicos claramente segregados, como se visualiza en la ilustración 1

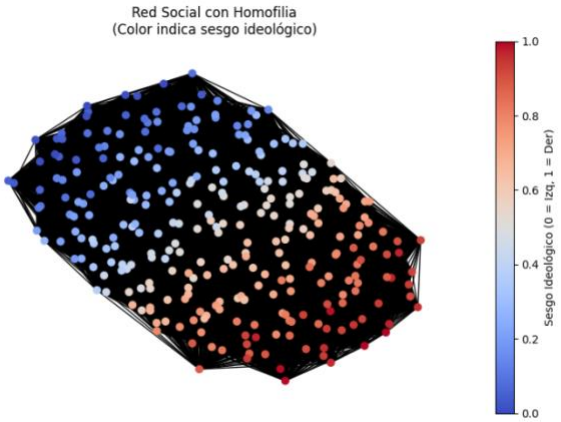


Ilustración 1 *Visualización de la Red Social Simulada con Homofilia.*

(Fuente: Elaboración propia, 2025)

Como se muestra en la ilustración 1, en nuestra red social el algoritmo de la plataforma amplifica la propagación de contenido en la red social con alta homofilia, donde los agentes maximizando su utilidad terminan generando una estructura notablemente segregada donde los nodos de similar color reflejan la formación de clústeres ideológicos densamente interconectados, este hallazgo comprueba la segmentación estructural de la población en cámaras de eco como afirma (Narayanan, 2023) , sumado a lo anterior la dinámica de compartir no solo depende de la veracidad percibida, sino también de incentivos sociales relevantes para capturar la búsqueda de validación grupal.

Los resultados de la simulación muestran que la curva de propagación de desinformación crece más rápido y alcanza mayor amplitud al existir algoritmos que potencian el contenido de alto impacto, con ratios superiores a 10:1 entre compartir y desaprobar, lo que con una ausencia de interacciones pasivas dicha tendencia se mantiene consistente tras múltiples réplicas del experimento, con medias estables en las acciones de compartir y desaprobar, y variabilidad nula en mensajes ignorados, finalmente, el análisis de redes confirma que la desinformación se disemina a través de cuentas centrales y estructuras inauténticas que maximizan la visibilidad de estos mensajes como afirma (Vysotska & Nazarkevych, 2025).

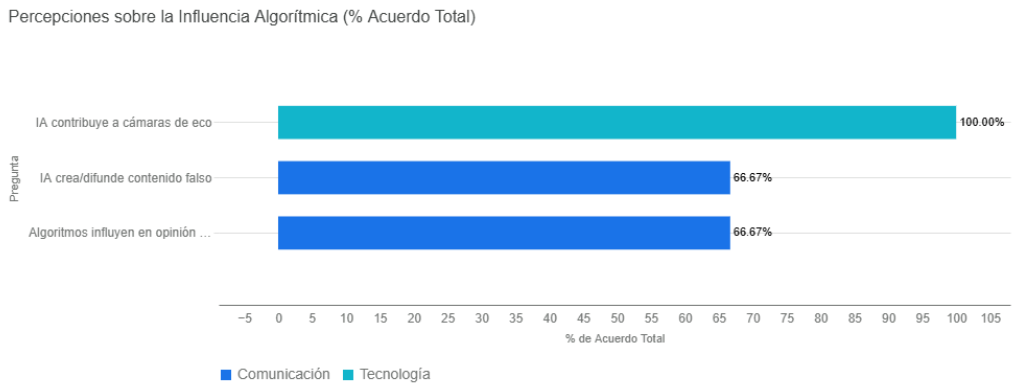
Los resultados de estas múltiples ejecuciones validan empíricamente la consistencia del modelo propuesto y demuestran que los patrones observados no constituyen artefactos estadísticos sino manifestaciones sistemáticas del fenómeno bajo estudio.

3.1.1.3 Encuesta sobre la Influencia Algorítmica

Los hallazgos descritos anteriormente se alinean perfectamente con las percepciones manifestadas por los participantes en la encuesta aplicada a profesionales de la comunicación y la tecnología, cuyos resultados reflejan un

alto nivel de conciencia entre profesionales de la comunicación y tecnología, sobre el impacto de la inteligencia artificial y los algoritmos en la propagación de desinformación y la polarización digital, encontramos que dos tercios de los encuestados del área de comunicación reconocen que la IA es actualmente utilizada para crear o difundir contenido engañoso, y atribuyen a los algoritmos un papel determinante en la formación de la opinión pública, adicional existe consenso entre los expertos en tecnología respecto a que los algoritmos de IA contribuyen activamente a la formación de cámaras de eco, dificultando la exposición a perspectivas diversas como podemos apreciar en la ilustración 2:

Ilustración 2 *Síntesis de percepciones sobre la Influencia Algorítmica*



Fuente: Elaboración propia, 2025

3.1.2 Eficacia y Características de los Modelos Algorítmicos de Detección

3.1.2.1 Revisión de la Literatura

En las investigaciones se evidencia el enfoque en la evaluación comparativa de diferentes modelos con el objetivo de identificar los más utilizados, se determinó que los enfoques más frecuentemente evaluados en la literatura incluyen modelos como Support Vector Machines y Random Forest (Mahmood & Akar, 2024), junto con modelos de Deep Learning como las Redes Neuronales Convolucionales (Kaliyar et al., 2020) y las redes LSTM, que capturan dependencias secuenciales y contextuales en el lenguaje (Alghamdi

et al., 2024), dentro de los estudios se evalua la eficacia cuantitativa mediante métricas estándar como Accuracy, Precision, Recall y F1-Score, alcanzando consistentemente valores superiores al 90% en datasets balanceados, lo que valida su aplicabilidad práctica en sistemas de detección automatizada (Vysotska & Nazarkevych, 2025).

3.1.2.2 Simulación de Clasificación de Contenido

Para evaluar comparativamente la eficacia de los modelos algorítmicos identificados en la literatura (SVM, Random Forest y LSTM), se implementó un experimento de clasificación de contenido utilizando el dataset público spanishFakeNews.csv de Kaggle (Zules, 2019), al respecto, el rendimiento de cada modelo se midió mediante métricas estándar consolidadas en la Tabla 1:

Tabla 1*Resultados Comparativos de los Modelos.*

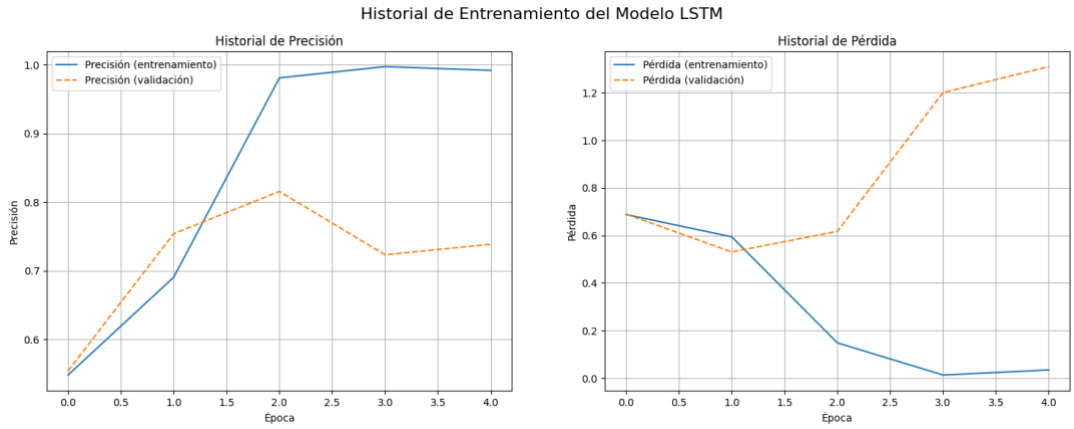
Modelo	Accuracy	Precision	Recall	F1-Score
Random Forest (Optimizado)	**0.8241**	**0.8387**	**0.8525**	**0.8455**
SVM (Optimizado)	**0.8981**	**0.8788**	**0.9508**	**0.9134**
Ensamblado (RF+SVM)	**0.9167**	**0.9062**	**0.9508**	**0.9280**
LSTM Bidireccional	**0.7870**	**0.9130**	**0.6885**	**0.7850**

fuente: elaboración propia,2025.

Los resultados de la simulación evidencian que el modelo ensamblado que combina el SVM optimizado y Random Forest, alcanzó el mejor desempeño general con un F1-Score de 0.9280 y una exactitud de 91.67%; por su parte el modelo SVM optimizado obtuvo un rendimiento cercano con F1-Score de 0.9134 y el mayor Recall con 0.9508, mostrando su efectividad en la identificación de noticias falsas, mientras que el Random Forest optimizado mostró métricas intermedias, con un F1-Score de 0.8455, y Accuracy de 82.41%, es destacable el modelo LSTM Bidireccional que exhibió un comportamiento diferenciado, aunque su exactitud fue de 78.70%, destacó en precisión con 91.30%, pero obtuvo un menor Recall con 0.6885, lo que indica mayor confiabilidad en casos positivos detectados a costa de omitir otros.

El análisis del proceso de entrenamiento del modelo LSTM Bidireccional reveló indicios de sobreajuste, como se observa en la ilustración 3, la precisión de entrenamiento superó el 99%, mientras que la de validación alcanzó un máximo cercano al 82%, de forma que la curva de pérdida mostró un incremento sostenido en la validación a partir de la tercera época, sugiriendo que el modelo memorizó patrones del conjunto de entrenamiento con limitada capacidad de generalización en datos no vistos.

Ilustración 3 Precisión y Pérdida de LSTM.



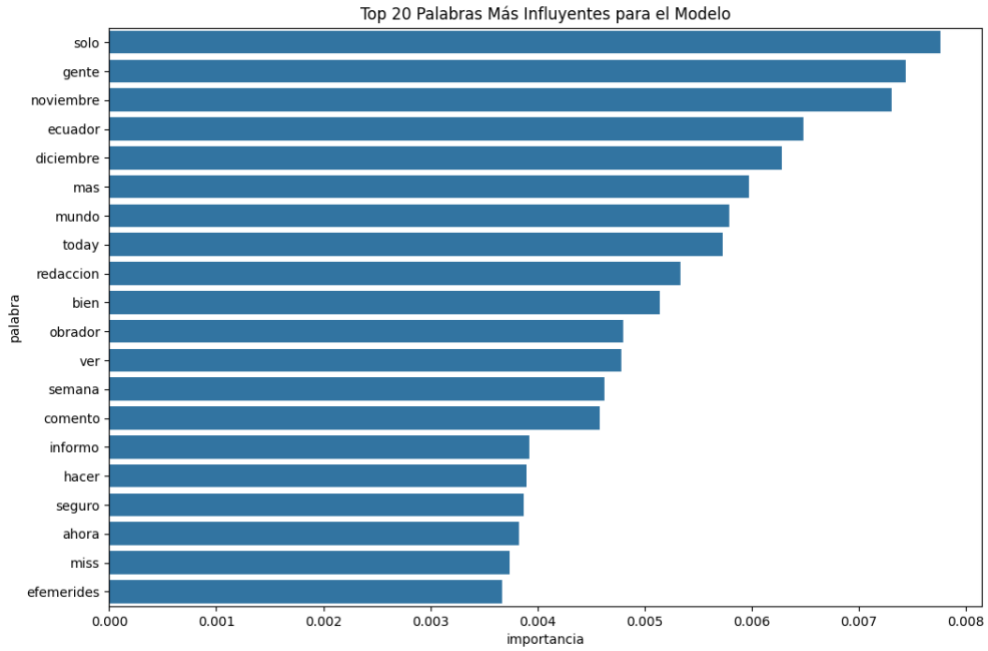
Fuente: Elaboración propia, 2025

Para una mayor comprensión del funcionamiento interno del modelo Random Forest, se implementó un análisis de interpretabilidad, cuyos resultados se presentan en la ilustración 4, el análisis reveló que entre las 20 características con mayor poder predictivo se identificaron términos específicos del dataset, como “today”, “obrador” y “Ecuador”, paralelamente se detectó que palabras estilísticas de uso frecuente, como solo, gente, comento e informo, también obtuvieron alta relevancia en las predicciones del modelo.

Para complementar este análisis, se realizó una evaluación cualitativa de errores sobre las predicciones del modelo LSTM utilizando el conjunto de prueba, con el propósito de examinar su comportamiento en casos específicos como ejemplo ilustrativo, se identificó que una noticia verificada como real,

proveniente de un conocido medio referente al aumento de delitos de odio, fue incorrectamente clasificada por el modelo como fake con una probabilidad del 95%, el resultado sugiere que los modelos pueden presentar limitaciones en la identificación de contenido legítimo cuando este aborda temas sensibles o controvertidos, lo que plantea interrogantes sobre la robustez de estos sistemas en escenarios reales de aplicación.

Ilustración 4 Top 20 de Palabras en RF.

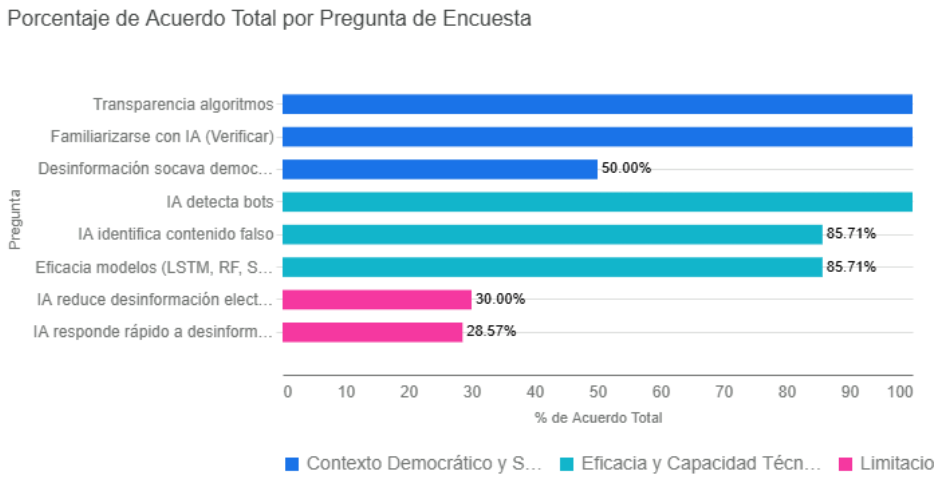


Fuente: Elaboración propia, 2025

3.1.2.2 Encuesta referente a modelos algorítmicos

Con el fin de situar la percepción de los encuestados con los hallazgos en la experimentación, la Ilustración 5 resume el nivel de acuerdo con las interrogantes planteadas.

Ilustración 5 *percepción referente a la IA y desinformación.*

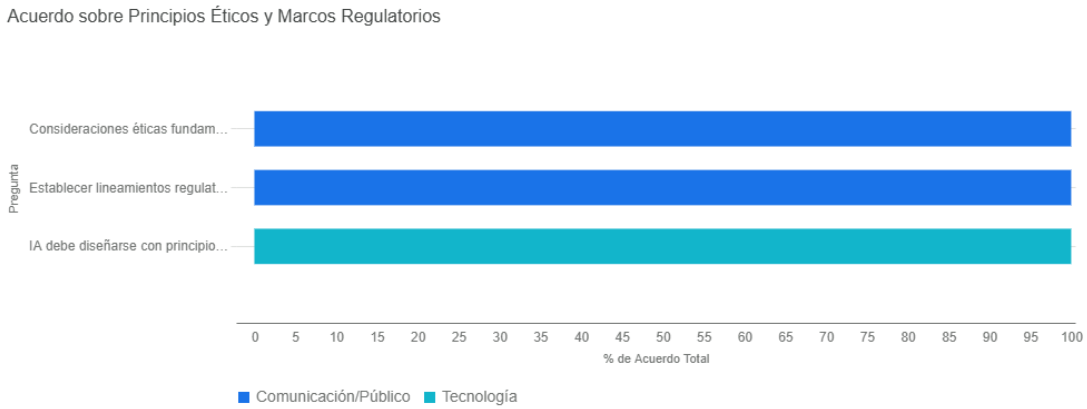


Fuente: Elaboración propia, 2025

Los hallazgos reflejan un consenso entre personas del área de tecnología sobre la capacidad de la IA para detectar la difusión inorgánica de contenido generada por bots y el 85.71% respalda la eficacia de modelos como LSTM, Random Forest y SVM para identificar contenido falso, no obstante, persiste escepticismo sobre su desempeño con solo el 28.57% que considera, que la IA responde con la celeridad requerida ante eventos de desinformación y apenas el 30% percibe reducción real de desinformación en época electoral, evidenciando una brecha entre capacidad potencial y efectividad observada en entornos reales. En consonancia, se constata un acuerdo pleno sobre requisitos ético-normativos: el 100% valora la transparencia algorítmica y la familiarización con herramientas de verificación, lo que refuerza la necesidad

de marcos que garanticen trazabilidad y control de sesgos en la aplicación de estos sistemas.

Ilustración 6 *consideración de implementación de principios éticos y marcos regulatorios*



Fuente: Elaboración propia, 2025

Sobre la base de los objetivos planteados en la investigación se propone desarrollar un marco de evaluación y protección basado en los modelos algorítmicos de inteligencia artificial para mitigar los efectos de la propagación de desinformación en la gobernanza democrática de Ecuador.

3.1.3 Propuesta de Marco de Evaluación y Protección Algorítmica para la Mitigación de la Desinformación en la Gobernanza Democrática del Ecuador

3.1.3.1 Introducción

La búsqueda de contención de la propagación de desinformación en plataformas digitales refleja el desafío que enfrentan nuestras sociedades frente al desarrollo de la tecnología y su adopción, por lo que se plantea un marco de evaluación y protección basado en IA que combine control de la amplificación con fortalecimiento de capacidades humanas, la evidencia del estudio muestra que aunque un modelo ensamblado alcanzó 91.67% de exactitud, los modelos dependen de sesgos del corpus y algunos patrones estilísticos; además en plataformas digitales existe una tendencia al

surgimiento de cámaras de eco donde la amplificación algorítmica pesa más que los atributos del contenido, lo que exige estrategias que atiendan simultáneamente mensaje y difusión, por esa razón, el marco propone tres dimensiones: ética, técnica y gobernanza social, con ello se busca mitigar la propagación y fortalecer la gobernanza democrática mediante sistemas transparentes y alineados con derechos.

3.1.3.2 Lineamientos del Marco de Evaluación y Protección

3.1.3.2.1 Lineamientos Éticos Fundamentales

Estos principios buscan un uso de la IA contra la desinformación que sea legítimo, transparente y compatible con los derechos democráticos: conforme a lo mencionado por (UNESCO, 2022), las mismas requieren transparencia y explicabilidad, documentación y evaluación de riesgos, trazabilidad y garantías sobre procedencia, calidad y veracidad de los datos, en este sentido el marco se concreta en tres lineamientos: (1) no dañar por "incompetencia semántica", limitando la IA al apoyo de moderadores cualificados y preservando discurso político legítimo, sátira y narrativas personales; (2) transparencia radical sobre capacidades y limitaciones mediante avisos visibles que identifiquen al sistema como clasificador, expliciten posibles sesgos temáticos o geográficos y aclaren que no emite decisiones finales; y (3) priorizar la mitigación de la amplificación sobre la censura, aplicando medidas graduales como etiquetas al estilo de "verificación en curso" o "necesita contexto adicional" y finalmente fricción a la compartición para desacelerar la viralidad mientras se verifica, preservando el acceso a la información.

3.1.3.2.2 Lineamientos Técnicos

El propósito de estos lineamientos es buscar interpretar los principios éticos en prácticas concretas de trabajo que aseguren el uso confiable de los modelos algorítmicos, por lo tanto, se propone la interpretabilidad obligatoria, en el sentido de que ningún modelo será desplegado sin un informe técnico

que explique qué patrones utiliza y por qué, revisado por un equipo multidisciplinario para verificar que las variables relevantes tengan sentido semántico y no sean simples artefactos del conjunto de entrenamiento; si el análisis evidencia atajos estadísticos sin fundamento, el sistema no será apto para producción.

A fin de prevenir la obsolescencia, se implementa un ciclo MLOps con monitorización continua, alertas y reentrenamiento controlado ante deriva de datos o de concepto, que será apoyado en baterías automáticas de pruebas distribucionales como podría ser la prueba Kolmogorov–Smirnov para continuas y Chi-cuadrado para categóricas, además de un registro y versionado de datos y validación fuera de distribución y umbrales de “roll-back” documentados, así se establecen estándares de calidad y diversidad del dataset, estableciendo conjuntamente la cobertura de fuentes diversas, balance geográfico y temporal, protocolos de anotación y control de clase, y una diversidad estilística con inclusión deliberada de contenidos verídicos pero emocionalmente intensos, sátira y discurso político, con el fin de permitir al modelo distinguir intensidad expresiva de falsedad, reduciendo sesgos sistemáticos y mejorando la generalización en contextos reales.

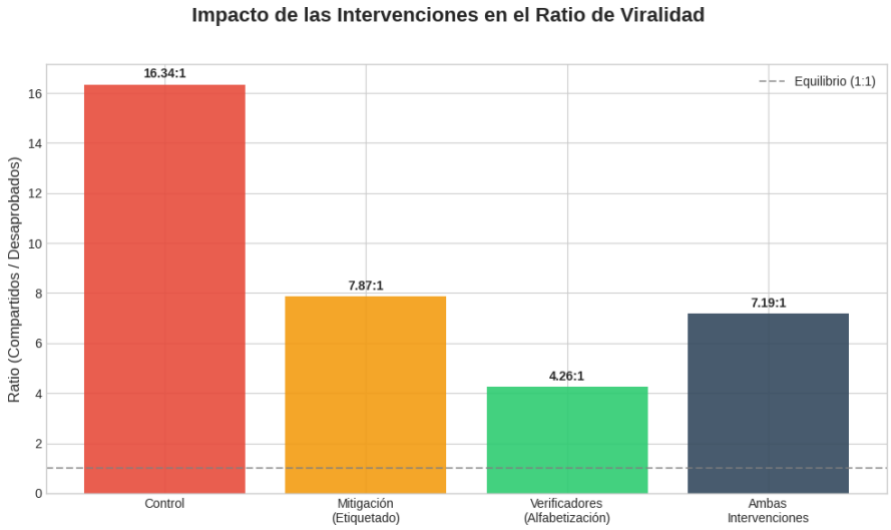
3.1.3.2.3 Dimensión de Gobernanza Social

La supervisión de sistemas de detección de desinformación con IA exige legitimidad social y rendición de cuentas mediante un consejo consultivo independiente y multiactor que integre a la sociedad civil, academia y sector privado para realizar auditorías periódicas de implementación, desempeño e impacto, garantizando transparencia y control público. Complementariamente, el marco debe incorporar políticas sostenidas de alfabetización digital y mediática que fortalezcan la capacidad crítica ciudadana, teniendo en cuenta que las soluciones tecnológicas son necesarias pero insuficientes, por lo que se prioriza comprender los mecanismos de propagación y el rol individual para prevenir la difusión viral de contenidos engañosos.

3.1.5 Validación Experimental del Marco Propuesto

Se validó experimentalmente el marco propuesto mediante una simulación comparativa que permitió conocer el efecto de distintas intervenciones sobre la difusión de contenidos en redes sociales, en ese sentido se modeló una red similar a la primera simulación, registrando nuevamente las decisiones de compartir, desaprobar o ignorar, pero con cuatro escenarios distintos: (1) escenario sin intervención; (2) un escenario con mitigación automatizada; (3) verificadores, asignando al 5% de la población con una alta competencia crítica y (4) aplicación de ambas intervenciones.

Ilustración 7 Impacto en el ratio de Viralidad

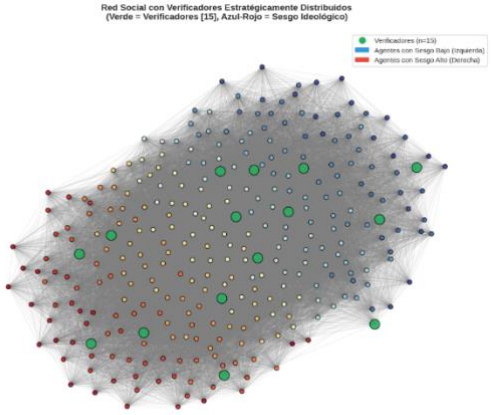


fuente: Elaboración propia, 2025

La ilustración 7 presenta la relación de viralidad que es el ratio de compartidos respecto a desaprobaciones para cada escenario, al respecto el escenario control estableció la relación más alta con 16.34:1, mientras que el escenario exclusivo de verificadores alcanzó la relación más baja con 4.26:1, entre tanto los escenarios de mitigación y ambas intervenciones mostraron relaciones intermedias de 7.87:1 y 7.19:1 respectivamente. La línea punteada horizontal

indica el punto de equilibrio teórico (1:1) donde compartidos y desaprobaciones serían equivalentes.

Ilustración 8 Red social con verificadores.



fuente: elaboración propia, 2025.

La ilustración 8 visualiza la topología de la red simulada utilizada en los escenarios de verificadores, los que están distribuidos estratégicamente para funcionar como puentes entre comunidades con diferentes sesgos ideológicos. Los nodos se colorean en un gradiente azul-rojo según el nivel de sesgo ideológico del agente, evidenciando la estructura de homofilia de la red y la posición estratégica de los verificadores en zonas de transición entre clusters polarizados.

3.2 Discusión

Durante el desarrollo de esta investigación se ha estudiado el fenómeno de la desinformación, donde se hace notable como la misma es un fenómeno multidimensional, en las simulaciones se observó que la arquitectura social funciona como superpropagador en ausencia de intervención, que orienta políticas hacia la reducción de la amplificación estructural y no únicamente al filtrado de piezas individuales. En consecuencia, se recomienda priorizar medidas que desaceleren la viralidad como el etiquetado, combinadas con alfabetización mediática y supervisión humana, además se sugiera en futuras investigaciones ampliar trabajos con datasets locales, diversidad estilística y

validación con usuarios reales para calibrar umbrales y formatos de advertencia.

4. Conclusiones

Durante el desarrollo de esta investigación se planteó la construcción de un marco de evaluación y protección basado en modelos algorítmicos de inteligencia artificial para mitigar la desinformación en la gobernanza democrática del Ecuador, por tal razón los hallazgos de esta, conducen a un enfoque de mitigación que integre el análisis de contenido y su dinámica de propagación, en ese sentido sobre la base de los resultados encontrados podemos establecer los siguientes puntos:

En un primer punto se demostró en una red social simulada con alta homofilia, la formación de cámaras de eco que actúan como aceleradores de narrativas los algoritmos de recomendación amplificaron la desinformación al elevar la relación comparticiones/desaprobaciones por encima de 5:1;por tanto al priorizar el compromiso sobre la veracidad, la arquitectura de la red actuó como superpropagador de narrativas, reforzó la polarización ciudadana y dificultó los acuerdos institucionales, con efectos adversos para la gobernanza democrática.

Se confirman que los modelos de aprendizaje automáticos en especial cuando se combinan en arquitecturas ensambladas logran altos niveles de precisión. No obstante, si bien los modelos alcanzan alta precisión numérica dentro de su funcionamiento se observa la detección de patrones léxicos y artefactos estadísticos del dataset, lo cual manifiesta fragilidad ante contenido fuera de la distribución de entrenamiento, si bien se tiene presente que parte de esta fragilidad es consecuencia misma de la elaboración del corpus del dataset, no invalida el comportamiento que presenta el modelo bajo contenido legítimo con alta carga emocional, es la razón por la que el presente estudio aboga por la asistencia en moderación bajo supervisión humana.

La validación del marco propuesto mostró que los sistemas automatizados pueden desacelerar la propagación, pero en última instancia solo la intervención humana cualificada es capaz de evaluar contexto, intencionalidad y matices semánticos, puede generar una resistencia activa y sostenida contra la desinformación.

Referencias bibliográficas

Alghamdi, J., Luo, S., & Lin, Y. (2024). A comprehensive survey on machine learning approaches for fake news detection. *Multimedia Tools and Applications*, *83*(17), 51009–51067. https://doi.org/10.1007/s11042-023-17470-8

Álvarez-Daza, N., Pico-Valencia, P., & Holgado-Terriza, J. A. (2021). Detección de Noticias Falsas en Redes Sociales Basada en Aprendizaje Automático y Profundo: Una Breve Revisión Sistemática. *Revista Ibérica de Sistemas e Tecnologias de Informação*, *1*(E41), 632–645. http://www.risti.xyz/issues/ristie41.pdf

Duarte, J. M. S., & Magallón-Rosa, R. (2023). Disinformation. *Eunomia. Revista en Cultura de la Legalidad*, *24*, 236–249. https://doi.org/10.20318/eunomia.2023.7663

García-Orosa, B. (2021). Disinformation, social media, bots, and astroturfing: the fourth wave of digital democracy. *Profesional de la Informacion*, *30*(6). https://doi.org/10.3145/epi.2021.nov.03

Kaliyar, R. K., Goswami, A., Narang, P., & Sinha, S. (2020). FNDNet – A deep convolutional neural network for fake news detection. *Cognitive Systems Research*, *61*, 32–44. https://doi.org/10.1016/J.COGSYS.2019.12.005

Keijnen, J. P. C. (2015). *Design and Analysis of Simulation Experiments* (2a ed.). Springer International Publishing Switzerland. https://doi.org/10.1007/978-3-319-18087-8

Kreps, S., & Kriner, D. (2023). How AI Threatens Democracy. *Journal of Democracy*, *34*(4), 122–131. https://doi.org/10.1353/jod.2023.a907693

López López, P. C., Andrea Mila Maldonado, A. M. M., & Ribeiro, V. (2023). La desinformación en las democracias de América Latina y de la península ibérica: De las redes sociales a la inteligencia artificial (2015-2022). *Uru: Revista de Comunicación y Cultura*, *8*, 69–89. https://doi.org/10.32719/26312514.2023.8.5

Mahmood, O. R. M., & Akar, F. (2024). Using and Comparison of Artificial Intelligence Techniques to Detect Misinformation and Disinformation on Twitter. *The European Journal of Research and Development*, *4*(2), 254–264. https://doi.org/10.56038/ejrnd.v4i2.467

Martínez, M. V. (2024). *De qué hablamos cuando hablamos de inteligencia artificial*. https://unesdoc.unesco.org/ark:/48223/pf0000391087.locale=es

McLoughlin, K. L., & Brady, W. J. (2024). Human-algorithm interactions help explain the spread of misinformation. En *Current Opinion in Psychology* (Vol. 56). Elsevier B.V. https://doi.org/10.1016/j.copsyc.2023.101770

Narayanan, A. (2023). *Understanding Social Media Recommendation Algorithms*. https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms

Sánchez Gonzales, H., & Alonso-González, M. (2024). Inteligencia artificial en la verificación de la información política. Herramientas y tipología. *Más Poder Local*, *56*, 27–45. https://doi.org/10.56151/maspoderlocal.215

Siderius, J. (2023). *Understanding Social Media: Misinformation, Attention, and Digital Advertising* [Massachusetts Institute of Technology]. https://hdl.handle.net/1721.1/150040

Sun, H. (2023). The Right to Know Social Media Algorithms. *Harvard Law & Policy Review*, *1*(18), 57. https://doi.org/10.2139/ssrn.4944976

UNESCO. (2022). *Recomendación sobre la ética de la inteligencia artificial Adoptada el 23 de noviembre de 2021*. https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa.locale=es

Vysotska, V., & Nazarkevych, M. (2025). Development of an information technology for detecting the sources and networks of disinformation dissemination in cyberspace based on machine learning methods. *Eastern-European Journal of Enterprise Technologies*, *4*(2), 35–51. https://doi.org/10.15587/1729-4061.2025.335501

Zules, F. A. (2019). *Spanish Fake and Real News*. Kaggle. https://www.kaggle.com/datasets/zulanac/fake-and-real-news/data