**Diseño ético para la detección de noticias falsas en periodos electorales**

**Resumen**

En la era digital, la desinformación electoral representa una seria amenaza para la integridad de los procesos democráticos. Este artículo aborda la identificación y clasificación de desinformación en medios digitales, un problema crítico que afecta tanto a nivel local como internacional. La manipulación de la percepción pública durante las campañas electorales puede influir en las decisiones de voto y alterar los resultados electorales. Para abordar este desafío, se propone el desarrollo de un sistema de detección de noticias falsas que emplea procesamiento del lenguaje natural (PLN) y aprendizaje máquina (AM). La propuesta del proyecto implica la recopilación de conjuntos de datos desde Kaggle, abarcando noticias verdaderas y falsas. Se lleva a cabo un riguroso preprocesamiento de texto para eliminar ruido, símbolos y palabras vacias. La extracción de características se realiza mediante valores tf-idf de las palabras, y se entrena un modelo de aprendizaje máquina, específicamente un bosque aleatorio, para clasificar las noticias como verdaderas o falsas. El ajuste y evaluación del modelo son pasos cruciales para mejorar su rendimiento. El modelo será de uso libre y se recomienda que la implementación y visualización del modelo se lleve a cabo en una plataforma en línea, donde los usuarios pueden ingresar noticias y recibir clasificaciones de autenticidad. Este enfoque no solo destaca la importancia de abordar la desinformación electoral, sino que también presenta una metodología integral para desarrollar una solución efectiva y ética. El proyecto contribuye al avance tecnológico al tiempo que se alinea con la misión más amplia de preservar los principios democráticos y fortalecer la resistencia a intentos de manipulación.

**I Introducción**

La era digital ha amplificado la difusión de desinformación, especialmente durante las elecciones, donde la integridad de la información es crítica para el proceso democrático. La desinformación puede manipular la percepción pública, influir en las decisiones de voto y potencialmente alterar los resultados electorales [1]. Esto representa un desafío nacional e inclusive mundial ya que las fake news se han convertido en un fenómeno mundial en el consumo de información online [2], debido a esto la desinformación no conoce fronteras y puede tener un impacto significativo en la estabilidad política y social del país

El problema específico que aborda el proyecto es la identificación y clasificación de información electoral falsa en los medios digitales. Es importante abordar este problema porque afecta directamente la confianza pública en los procesos democráticos y puede llevar a decisiones mal informadas por parte de los votantes, otro problema importante es el de la alta propagación que tienen las noticias, las cuales se comparten tan rápido que superan por mucho la capacidad que una persona tiene para controlar o contener este fenómeno [3]. La justificación se refuerza por el hecho de que la desinformación no permite un proceso electoral justo, debido a la manipulación de los medios.

Actualmente, la detección de desinformación se aborda a través de varias metodologías, incluyendo la verificación manual, el uso de algoritmos basados en reglas, y técnicas más avanzadas de Inteligencia artificial que incluyen procesamiento del lenguaje natural y aprendizaje máquina. Estas soluciones varían en efectividad y a menudo enfrentan desafíos como la detección tardía de desinformación, la dificultad en discernir el contexto y el riesgo de sesgos en los datos históricos y algoritmos. Una manera buena de resolver este problema de las noticias falsas es con el uso de aprendizaje profundo, como en el trabajo de M. A. Espejel-Rivera, R. Calderón-Suárez, R. M. Ortega-Mendoza, C. J. Camacho-Bello y M. A. Márquez Vera quienes utilizaron tres tipos de redes de aprendizaje profundo: CNN, GRU y CRU con una capa de atención para intentar diferenciar entre noticias falsas y verdaderas [3]. Otro enfoque que se le ha dado es el uso de bots los cuales ayudan a mitigar en cierta medida estas noticias falsas haciendo que sean los últimos resultados en aparecer. [4]

El objetivo principal de este trabajo es diseñar un sistema ético y eficiente para la detección de información electoral falsa en las noticias utilizando NLP y ML, evaluando los riesgos e impactos asociados. Esto incluye garantizar la precisión, reducir el sesgo, proteger la privacidad de los datos y proporcionar explicabilidad en el proceso de clasificación. El trabajo busca no solo desarrollar una solución técnica sino también considerar su impacto social.

La propuesta es crear un sistema avanzado de detección de noticias falsas binario que integre técnicas de NLP y ML. Esta solución se destaca por su enfoque en el preprocesamiento de datos dónde se procesarán para eliminar el ruido, símbolos y palabras vacías que son aquellas que no aportan mucha información en el lenguaje, la extracción de características relevantes que se hará mediante la técnica de tf-idf, y el uso de un modelo de aprendizaje máquina sofisticado, el cuál será un algoritmo de bosque aleatorio. Este algoritmo pasará por una evaluación y se le harán los respectivos ajustes, unas de las métricas a utilizar son la precisión, la sensibilidad, cuantos clasifica correctamente la puntuación F1 y la matriz de confusión. Será una contribución significativa al proporcionar una herramienta más precisa y eficiente para combatir la desinformación electoral, mejorando así la integridad de los procesos democráticos.

Contenidos

II Estado del arte: Breve explicación de cómo lo abordan otros y posibles problemas éticos que se identifican en el estado del arte con respecto al problema abordado.

III Revisión de requisitos éticos: El diseño ético y principios a considerar, así como las políticas, normas y estándares del problema en cuestión.

IV Propuesta de diseño ético para el problema en cuestión: Cuál es la metodología con la cuál se va a trabajar.

V Conclusiones.

**II Estado del arte de noticias falsas en periodos electorales**

México, con su rica historia política y su diverso panorama mediático, enfrenta desafíos únicos en términos de desinformación electoral. La influencia de las redes sociales y los medios de comunicación digitales ha incrementado la velocidad y el alcance con que se difunde la información, incluida la desinformación [5]. Durante las elecciones, se observan campañas de desinformación que buscan influir en la opinión pública, a menudo con objetivos políticos específicos. Esto puede incluir la propagación de noticias falsas sobre candidatos, manipulación de hechos relacionados con políticas electorales, o incluso rumores malintencionados para socavar la credibilidad de las instituciones electorales.

La necesidad de abordar este problema es crucial en México debido a su impacto directo en la democracia y la confianza pública en los resultados electorales. La desinformación puede llevar a una toma de decisiones mal informada por parte de los votantes y puede eliminar la confianza en el sistema político el cual es vital para la estabilidad y la gobernabilidad del país.

En el contexto actual de las campañas electorales, la proliferación de noticias falsas representa un desafío significativo. Este trabajo se centra en la clasificación binaria de noticias, aplicando algoritmos de aprendizaje automático para discernir entre noticias reales y falsas, y ayudar a los usuarios a navegar en un entorno informativo cada vez más complejo.

Desarrollo de Técnicas de Clasificación de Noticias Falsas

Algoritmos de Aprendizaje Automático en Redes Sociales: Sparks y Frishberg describen cómo Facebook utiliza algoritmos de aprendizaje automático para identificar noticias potencialmente falsas y sensacionalistas​​ [6].

Clasificadores de Aprendizaje Automático para Detección de Noticias Falsas: Se ha demostrado el uso de clasificadores de aprendizaje automático en la detección de noticias falsas, especialmente en la clasificación de texto y redes sociales [7]​​.

Técnicas de Aprendizaje Automático con Selección Óptima de Características: Un estudio aplicó un clasificador de bosque aleatorio para predecir noticias falsas o reales, utilizando características textuales del conjunto de datos ISOT Fake News [8]​​.

Detección de Noticias Falsas Mediante Aprendizaje Profundo: Se investigó un enfoque híbrido CNN-RNN para la detección de noticias falsas, destacando la necesidad de experimentar con una amplia gama de conjuntos de datos​​ [9].

Aplicación en el Contexto de Campañas Electorales

La clasificación precisa de noticias es crucial en las campañas electorales para combatir la desinformación. La integración del aprendizaje automático y el procesamiento basado en texto permite la construcción de clasificadores efectivos​​​​ [10]. Investigaciones recientes han subrayado la importancia de diferenciar entre conceptos de desinformación y mala información​​. Para concluir la clasificación binaria de noticias reales y falsas es esencial para mantener una democracia informada. El avance continuo en algoritmos de aprendizaje automático y técnicas de aprendizaje profundo promete mejorar significativamente la precisión en la identificación de noticias falsas, especialmente en contextos sensibles como las campañas electorales [6].

A pesar de todas estas investigaciones recientes para combatir la desinformación mediante algoritmos de aprendizaje máquina no se ha hecho un análisis ético a profundidad sobre los impactos ya sean positivos o negativos que estas posibles soluciones pueden ofrecer, no se han planteado principios éticos relevantes o si los mencionan no son el eje principal, en este proyecto la Explicabilidad, el Control Humano , la Justicia y no discriminación son los pilares éticos sobre los cuales se va a desarrollar el proyecto [11].

**III Revisión de requisitos éticos**

Como se habló anteriormente en este proyecto no solo se busca encontrar un algoritmo para la identificación de noticias falsas y reales si no también darle un enfoque ético por lo que veremos cuales son los principios éticos más importantes que van a guiar este desarrollo y cuales también es buena idea tener en cuenta, la explicabilidad es el primero de ellos y aplicado al proyecto debe cumplir con lo siguiente: El modelo debe de mostrar cuales fueron las características y el porcentaje de impacto que tuvieron estas en la predicción final para generar confianza en los resultados. La no discriminación y justicia se ven reflejados en el proyecto de la siguiente manera: Asegurar que los datos de entrenamiento representen una variedad de fuentes y perspectivas para evitar sesgos en las predicciones igualmente que la herramienta sea accesible y justa para todos los usuarios y no favorezca a un grupo sobre otro. El modelo no discriminará o favorecerá a ninguna perspectiva política, cultural o social. El último principio ético fundamental es el Control Humano de la Tecnología que sería el asegurarse de que las decisiones y acciones del modelo de aprendizaje máquina puedan ser supervisadas, validadas y si es necesario ser corregidas por seres humanos para prevenir y mitigar consecuencias no deseadas o dañinas. También sería bueno considerar la Responsabilidad Profesional que es la mejora continua y un compromiso con la mejora constante del modelo y las prácticas del equipo, aprendiendo de los errores y retroalimentaciones, y ajustándose a nuevas evidencias y contextos.

En este proyecto también se examinan las políticas, normas y estándares que rigen la difusión de información durante las campañas electorales, con un enfoque específico en el marco legal mexicano.

Análisis de Políticas y Leyes:

Ley General de Instituciones y Procedimientos Electorales (LGIPE):

La LGIPE establece las reglas fundamentales para la organización y desarrollo de elecciones federales en México. Un análisis detallado de esta ley proporcionará información sobre las disposiciones relacionadas con la difusión de información durante las campañas y la prohibición de noticias falsas que puedan afectar el proceso electoral [12].

Ley General de Comunicación Social:

Esta ley regula la distribución de recursos públicos destinados a comunicación social en México, lo que podría influir en la veracidad de la información transmitida durante las campañas electorales [13].

Normas Éticas y Principios Profesionales:

Código de Ética del Periodista Mexicano:

El análisis de normas éticas, como el Código de Ética del Periodista Mexicano, proporciona una comprensión profunda de los estándares profesionales y la responsabilidad de los medios de comunicación en la difusión de información precisa y veraz (Federación de Asociaciones de Periodistas Mexicanos [14].

Protección de Datos Personales:

Ley Federal de Protección de Datos Personales en Posesión de Particulares (LFPDPPP):

En el contexto de la recopilación y tratamiento de datos para el desarrollo del clasificador, la LFPDPPP establece principios y obligaciones para la protección de datos personales [15].

Este análisis resalta la importancia de considerar las políticas, normas y estándares vigentes al desarrollar un clasificador binario para abordar la desinformación electoral. Un enfoque integral que cumple con estas regulaciones no solo garantiza la efectividad del modelo sino también su alineación con los principios éticos y legales que rigen el entorno electoral mexicano.

**IV Propuesta de diseño ético para el problema en cuestión**

La metodología empleada para desarrollar la propuesta es AMCR (Metodología de alcance, mapeo, artefactos y reflexión), una simplificación del método SMACTR. Esta versión condensa los elementos esenciales de SMACTR [16] para aumentar su agilidad, facilitar la documentación y promover un diseño ético, buscando así una mayor adopción. La metodología AMCR se estructura en cuatro etapas clave:

Fase de Alcance: Esta inicial etapa abarca la formulación de la propuesta del proyecto y la planificación inicial. Incluye también una revisión ética preliminar de alto nivel.

Fase de Mapeo: Se realiza un mapeo detallado de los involucrados, especificando sus roles y responsabilidades. Además, se lleva a cabo un análisis FMEA para definir métricas y establecer prioridades donde definimos cuales principales valores éticos a cuidar, cuáles son sus beneficios e impactos y las métricas para cuantificarlas.

Fase de Recolección de Artefactos: Esta fase se divide en dos partes principales. La primera es la Carta del Modelo, donde se documenta el modelo de aprendizaje automático empleado. La segunda parte consiste en la Hoja de Datos de la Base de Datos, que detalla la documentación de los datos utilizados en el modelo.

Fase de Reflexión: La etapa final incluye un análisis de las estrategias de mitigación de riesgos del proyecto. Se concluye con una evaluación crítica para determinar si el modelo está listo o no para ser implementado en producción, así como cuales son los requisitos técnicos, operacionales y organizacionales para desplegar y mantener el modelo.

Cada fase de AMCR está diseñada para ser iterativa y reflexiva, garantizando que el proyecto no solo cumpla con los requisitos técnicos, sino que también se adhiera a principios éticos fundamentales.

**Caso de uso**

El problema por abordar es la identificación y clasificación de desinformación electoral en medios digitales, ya que las campañas de desinformación buscan influir en la percepción y decisión del electorado, tanto a nivel local como internacional. Este fenómeno puede favorecer a un candidato o partido específico o socavar la confianza en el sistema electoral. La capacidad de detectar y contrarrestar esta desinformación es esencial para garantizar la integridad de las elecciones y para mantener una democracia informada y resistente a intentos de manipulación.

***Alcance***

La propuesta del proyecto final es desarrollar un sistema de detección de noticias falsas utilizando procesamiento del lenguaje natural y aprendizaje máquina. El proceso incluirá la recopilación de conjuntos de datos que contengan noticias verdaderas y falsas de diversas fuentes y categorías, seguida de un preprocesamiento de texto para eliminar ruido, símbolos y palabras vacias. Se extraerán características relevantes de los textos, como conteo de palabras, frecuencia de términos y características específicas de noticias, como la fuente, el tono y la estructura. Luego, se entrenará un modelo de aprendizaje máquina para clasificar las noticias como verdaderas o falsas, ajustando el modelo para obtener mejores resultados. La implementación se realizará en una plataforma en línea donde los usuarios podrán ingresar noticias y obtener una clasificación de autenticidad, con la posibilidad de visualizar estadísticas sobre noticias clasificadas.

***Mapeo***

Las personas involucradas en el proyecto son las siguientes:

Ingeniero de Machine Learning**:** Santiago Valera Barreiros

Responsabilidades: Desarrollar, entrenar, validar y mejorar los modelos de machine

learning que clasificarán las noticias.

Ingeniero de Datos**:** Santiago Valera Barreiros

Responsabilidades: Asegurarse de que los datos estén disponibles, limpios y en formato utilizable para el entrenamiento del modelo.

Analista de Datos: Santiago Valera Barreiros

Responsabilidades: Analizar y obtener insights de los datos, así como monitorear el desempeño del modelo en producción.

Especialista en Ética y Cumplimiento**:** Santiago Valera Barreiros

Responsabilidades: Garantizar que el desarrollo y la implementación del modelo cumplan con los principios éticos.

Especialista en Dominio / Experto en Contenidos**:** José Luis López Aguirre

Responsabilidades: Aportar conocimientos sobre periodismo y noticias para ayudar a afinar y validar el modelo y sus predicciones.

***Modo de Fallas y Análisis de efectos***

Explicabilidad

1. Que las personas puedan entender por qué el algoritmo clasifico una noticia como falsa o verdadera viendo cuales fueron las palabras que tomaron mayor relevancia para esta clasificación.
2. Que las personas aumenten su criterio de decisión usando el algoritmo de clasificación viendo que palabras tienden a dar una noticia falsa o verdadera.

Control Humano

1. Que las personas no le den la última palabra a la máquina para decidir si una noticia es falsa o verdadera.
2. Que las personas contemplen el modelo como una herramienta para tener otro punto de vista si una noticia es falsa o verdadera.

Justicia y no discriminación

Que una gran parte de la población puedan acceder a la herramienta

1. Que el algoritmo no le dé un valor muy importante al nombre de una persona para determinar si una noticia es falsa o verdadera.
2. Que el algoritmo no diga que una noticia sea 100% verdadera o falsa si se identifica a un grupo vulnerable, es decir que la presencia de estas palabras no sesgue la decisión

Métricas

|  |  |
| --- | --- |
| Valor cualitativo | Valor cuantitativo |
| Muy alto | 10 |
| Alto | 7 |
| Medio | 4 |
| Bajo | 1 |

Tabla 1

La tabla 1 muestra la escala de valores cualitativos para medir los impactos y beneficios con su correspondiente valor cuantitativo.

Criterios de evaluación de beneficios

|  |  |  |
| --- | --- | --- |
| Valor cualitativo | Valor cuantitativo | Descripción |
| Muy alto | 10 | El beneficio es muy bueno para la población |
| Alto | 7 | El beneficio es de mucha ayuda para la población |
| Medio | 4 | El beneficio contribuye en buena medida a la población |
| Bajo | 1 | El beneficio ayuda de manera limitada a la población |

Tabla 2

La tabla 2 muestra el valor cuantitativo y cualitativos en los que se pueden medir los beneficios.

Criterios de evaluación de probabilidad de beneficios

|  |  |  |
| --- | --- | --- |
| Valor cualitativo | Valor cuantitativo | Descripción |
| Muy alto | 10 | Es casi seguro que ocurra el beneficio |
| Alto | 7 | Hay una buena probabilidad de que ocurra el beneficio |
| Medio | 4 | Hay una probabilidad considerable de que ocurra el beneficio |
| Bajo | 1 | Las probabilidades de que ocurra el beneficio son bajas |

Tabla 3

La tabla 3 muestra el valor cuantitativo y cualitativos en los que se pueden medir las probabilidades de los beneficios.

Criterios de evaluación de riesgos

|  |  |  |
| --- | --- | --- |
| Valor cualitativo | Valor cuantitativo | Descripción |
| Muy alto | 10 | Este sería un impacto cuyas consecuencias perjudicaran mucho a la población |
| Alto | 7 | Este impacto perjudicara a la población |
| Medio | 4 | Este impacto perjudicara en cierta medida a la población |
| Bajo | 1 | El impacto no traerá consecuencias tan significativas en la población |

Tabla 4

La tabla 4 muestra el valor cuantitativo y cualitativos en los que se pueden medir los riesgos.

Criterios de evaluación de probabilidad de riesgos

|  |  |  |
| --- | --- | --- |
| Valor cualitativo | Valor cuantitativo | Descripción |
| Muy alto | 10 | Es casi seguro que ocurra el riesgo |
| Alto | 7 | Hay una buena probabilidad de que ocurra el riesgo |
| Medio | 4 | Hay una probabilidad considerable de que ocurra el riesgo |
| Bajo | 1 | Las probabilidades de que ocurra el riesgo son bajas |

Tabla 5

La tabla 5 muestra el valor cuantitativo y cualitativos en los que se pueden medir las probabilidades de los riesgos.

BENEFICIOS

1. Beneficio: las personas pueden entender por qué una noticia ha sido clasificada como falsa o verdadera y aprender a distinguir por sí mismas características que hacen verdadera o falsa a una noticia.
2. Beneficio: La herramienta permite a las personas pensar dos veces el tipo de noticia que tiene enfrente y le puede ayudar a prevenir la proliferación de noticias falsas.
3. Beneficio: El algoritmo de clasificación binaria refuerza el criterio de una persona para dictaminar la clasificación de la noticia, es decir que si una persona tiene fuertes sospechas de que una noticia es falsa, la persona podría ver con este algoritmo algunas características a considerar.
4. Beneficio: Las personas pueden entender que cosas en verdad son reales en las campañas electorales analizando las palabras que toman mayor relevancia para la clasificación de la noticia.
5. Beneficio: El algoritmo será imparcial y no beneficiara de manera injusta a ningún partido político, persona, empresa o grupo social, lo que permitirá que la población obtenga información verídica y justa.
6. Beneficio: Las elecciones serían más justas y transparentes.
7. Beneficio: Las personas son conscientes de los hechos que son verdaderos para juzgar a los candidatos/partidos políticos.

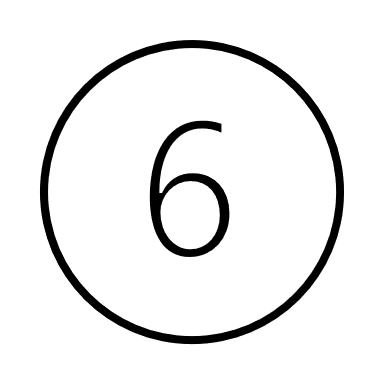
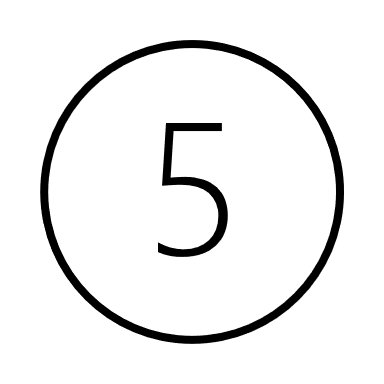
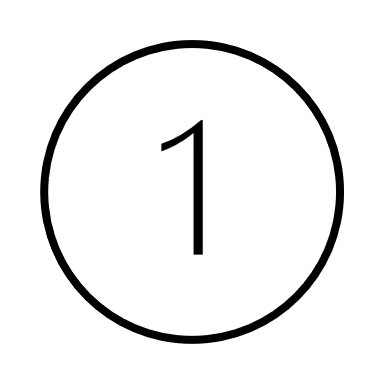
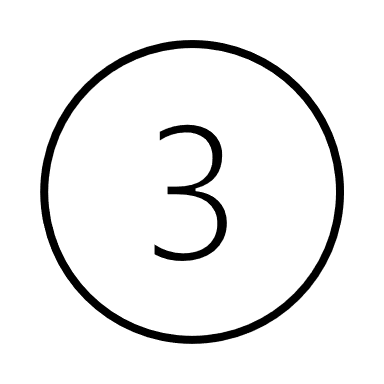
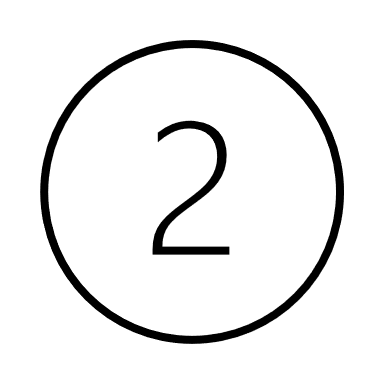
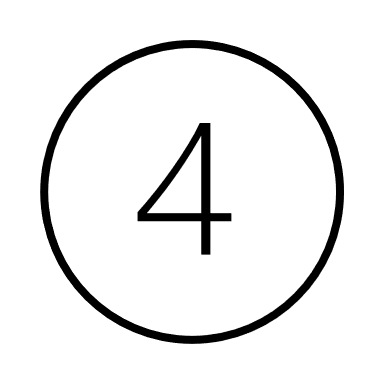
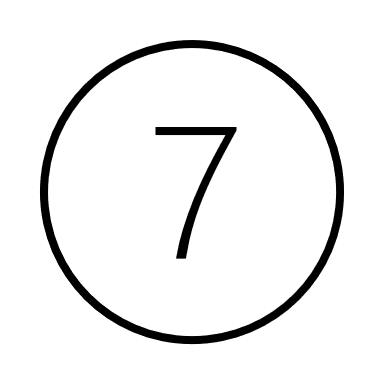
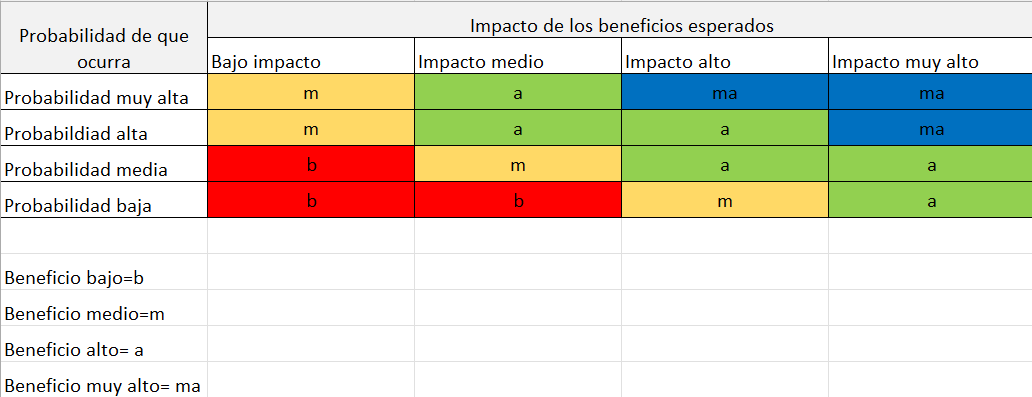
 

Tabla 6

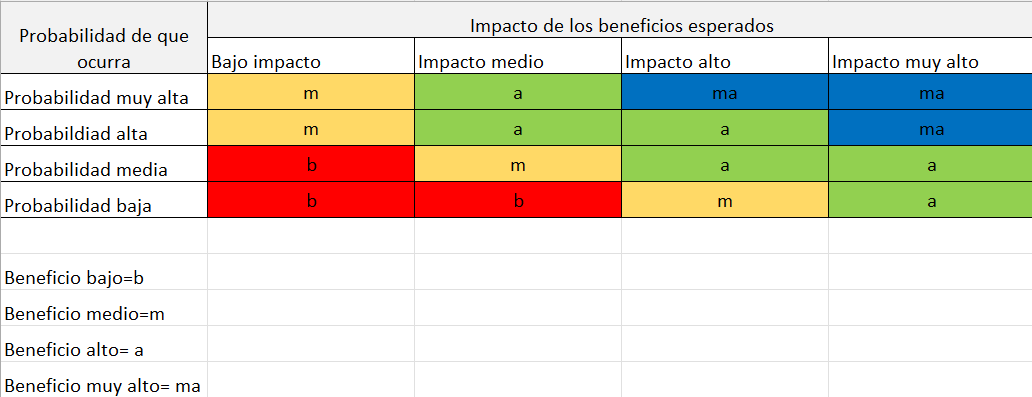


Tabla 7

En las tablas 6 y 7 se observa como se ha medido cada beneficio que está en el proyecto con su respectiva nomenclatura.

RIESGOS

1. Riesgo: Las personas podrían manipular la explicabilidad, es decir si las personas que se dedican a difamar y crear malas noticias pueden ver la explicación del algoritmo y modificar el cómo crean las noticias para que perezcan más verdaderas, lo que provocaría 3 cosas:

-Proliferación de noticias falsas

-Manipulación de la percepción del público

-Influir en la decisión de los votos

1. Riesgo: Las personas podrían tomar como verdadera el dictamen de la máquina sin dudar y en caso de una clasificación errónea puede ocasionar la expansión de una noticia falsa que se tome por verdadera, aumentando la desinformación electoral.
2. Riesgo: Las personas podrían comenzar a depender solo del criterio de la máquina y perder la capacidad para poner en sospecha una noticia
3. Riesgo: Que desacredite o quite voz a noticias donde una población vulnerable se esté viendo perjudicada y la máquina decida que no es así.
4. Riesgo: Que se descredite una noticia por el simple hecho de aparecer un nombre famoso, es decir que, si aparece en nombre de Barack Obama en automático se clasifica como falsa.
5. Riesgo: Que se acredite una noticia simplemente por el hecho de que aparezca un nombre famoso, es decir que, si en la noticia por aparecer Donald Trump en automático es verídica, lo que provocaría la proliferación de noticias falsas con el nombre de esta persona.

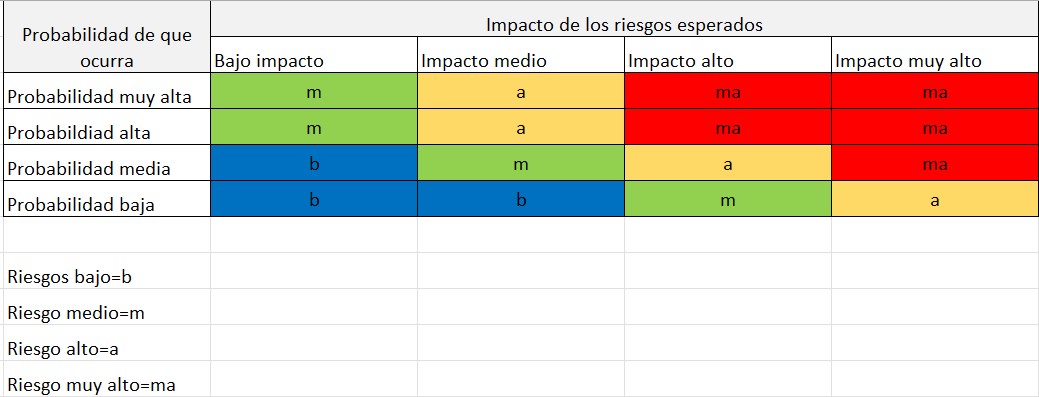
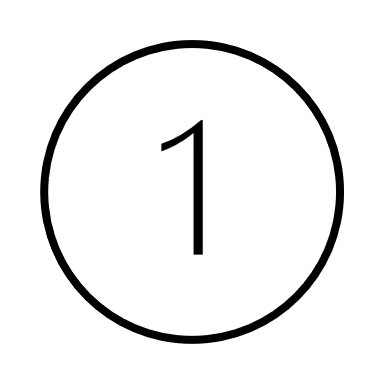
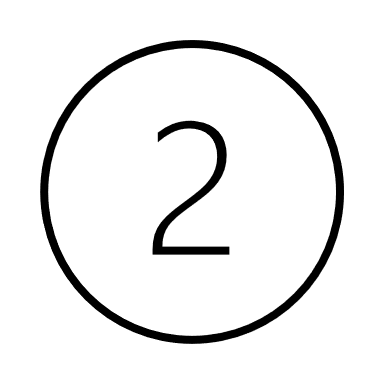
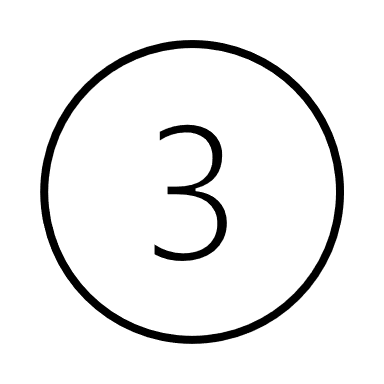
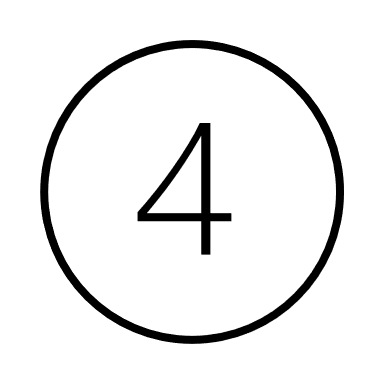
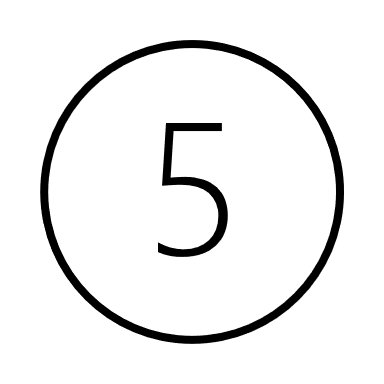
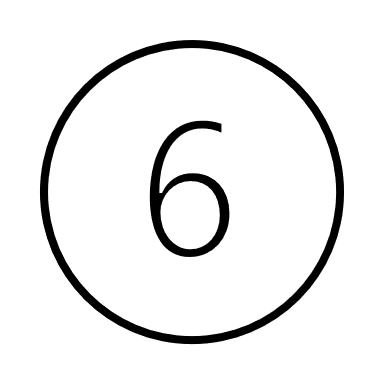
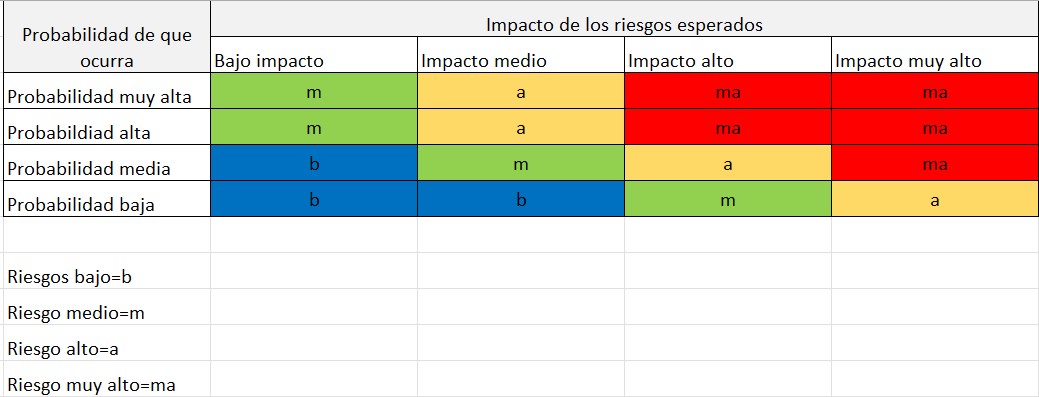


Tabla 8

  
Tabla 9

En las tablas 8 y 9 se observa cómo se ha medido cada riesgo que está en el proyecto con su respectiva nomenclatura.

***Artefactos***

**Carta del modelo**

Detalles del Modelo:

Información Básica sobre el Modelo

Organización o Persona Desarrolladora del Modelo:

Santiago Valera Barreiros

Fecha del Modelo:

16/11/2023

Versión del Modelo:

v1

Tipo de Modelo:

Clasificador binario basado en Bosque aleatorio

Información sobre Algoritmos de Entrenamiento

Algoritmo de Entrenamiento:

Bosque aleatorio

Parámetros Principales:

Profundidad= None es decir hasta que las hojas sean puras

Número de árboles:100

Características Utilizadas:

Se obtuvo el valor número de las palabras con tf-idf Vectorizer, se eliminaron las palabras cuya presencia en textos era menor al 5% y se eliminaron aquellas cuya presencia era mayor al 90%.

Dónde Enviar Preguntas o Comentarios sobre el Modelo:

Correo: santiagovalera@gmail.com

Uso Previsto

Usos Principales Previstos: Se espera que el modelo sea utilizado para identificar noticias falsas en plataformas de redes sociales o medios de comunicación en formato de texto.

Usuarios Previstos: Personas que se estén informando acerca de noticias políticas, principalmente usuarios que estén expuestos a redes sociales.

Casos de Uso Fuera del Alcance: Que las personas escuchen alguna noticia y de lo que recuerden le pregunten eso al modelo, debe ser la noticia tal cual redactada.

Factores

Factores Relevantes: Los hechos históricos influyen en el rendimiento del modelo y puede estar sesgado a ciertos acontecimientos.

Factores de Evaluación: Se busca que haya una buena relación entre las dos clases evaluadas y que no haya nombre de empresas, organizaciones o personas como característica principal para determinar la categoría de un modelo.

Métricas

Medidas de Rendimiento del Modelo: Precisión, Recall, F1-score y accuracy.

Umbrales de Decisión: Si la probabilidad es de 50% por ciento o menor se asigna a la categoría 0 (falsas)

Datos de Evaluación

Conjuntos de Datos: 20% de datos se utilizaron para la evaluación, equivale a 500 ejemplos

Motivación: Debido a que se contaba con una buena cantidad para ejemplos de entrenamiento se decidió dejar el 20% para evaluar.

Preprocesamiento: Primero se eliminaron las filas que las hacía falta ya sea el titular de la noticia o el texto de la noticia, después se lematizaron las palabras, después se pasaron a minúsculas, después se removieron las palabras vacías definición, después se calculó el valor td-idf para las palabras y eliminamos los valores más altos y pequeños, es decir las palabras cuya frecuencia era muy alta y las palabras cuya frecuencia era muy pequeña.

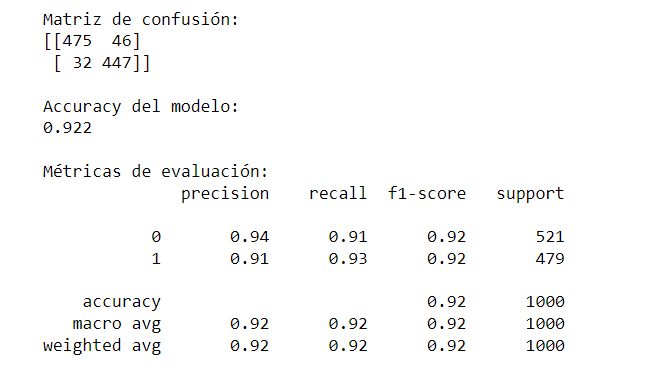
Datos de Entrenamiento

Detalles del Conjunto de Datos de Entrenamiento: Se utilizaron 4000 ejemplos.

Distribución sobre Factores Varios: Se utilizaron 2000 ejemplos falsos y 2000 reales

Análisis Cuantitativos

Resultados Unitarios



Consideraciones Éticas

Reflexiona sobre los posibles impactos éticos del modelo: El modelo podría traer algunos sesgos, en algunos análisis de explicabilidad se ha visto que el nombre de una empresa si está presente en la noticia suele tomar gran relevancia para la clasificación de la nota.

Advertencias y Recomendaciones

Advertencias: El modelo solo es capaz de realizar la clasificación de forma binaria y con textos en inglés.

Recomendaciones: Este modelo solo es una herramienta para quitar sospechas sobre la veracidad de una noticia no es una fuente de verdad absoluta.

**Hoja de datos de la base de datos**

Motivación

¿Con cuál propósito se creó la base de datos?

Se recolecto esta base de datos para hacer una competencia de clasificación binaria en el sitio web kaggle.

¿Quién creo la base de datos?

Saurabh Shahane, un usuario de la plataforma de kaggle

¿Quién financio la creación del conjunto de datos?

Saurabh Shahane

Composición

¿Qué representan las instancias que componen el conjunto de datos (por ejemplo, documentos, fotos, personas, países)? ¿Hay múltiples tipos de instancias (por ejemplo, películas, usuarios y calificaciones; personas e interacciones)?

Las instancias son noticias y sus títulos extraídos de otros conjuntos de datos de : Kaggle, Mcintire,BuzzFeed Political, Reuters, junto a un valor binario que nos dice si es falsa o verdadera la noticia, donde 0 es falso y 1 verdadero.

¿Cuántas instancias hay en total?

37,014 ejemplos que son noticias verdaderas.

34,965 ejemplos que son noticias falsas.

En total son 71989 instancias

¿El conjunto de datos contiene todas las posibles instancias o es una muestra de un conjunto de datos más grande?

Este conjunto de datos contiene todas las posibles instancias de los conjuntos de datos usados al momento de la creación de este.

¿En que consiste cada instancia?

Cada instancia tiene una columna que enumera a la instancia, una columna de texto que es el título de la noticia, otra columna de texto que es el cuerpo de la noticia y una última columna que es un número ya sea cero o uno y nos dice la clasificación de la instancia.

¿Hay alguna etiqueta asociada a cada instancia?

Si, cada instancia tiene una etiqueta que es 0 o 1 que representa falso o verdadero.

¿Hay valores faltantes en alguna instancia de manera individual?

Si, estos valores faltantes se encuentran en su mayoría en las instancias verdaderas, donde hace falta 10 cuerpos de texto, 562 títulos y hay 10 instancias que no tienen una etiqueta.

¿Hay relaciones entre las instancias hechas explícitamente?

No hay ninguna relación explicita.

¿Hay algún método recomendado para separar los datos?

Con una división de 80% para entrenamiento y 20% para la prueba y obtener métricas se obtuvo un muy buen rendimiento.

¿Hay errores, ruido en el origen, o redundancias en el conjunto de datos?

Si había unos símbolos que no pertenecían al alfabeto inglés, también había algunos datos que parecían estar corrompidos dentro del mismo data set o no estaban correctamente almacenados en sus celdas.

¿El conjunto de datos es completamente autónomo?

Si, el conjunto de datos es completamente autónomo.

¿El conjunto de datos contiene datos que podrían ser considerados confidenciales?

No, ya que son noticias públicas.

¿El conjunto de datos contiene datos que si son leídos directamente podrían ser ofensivos, amenazadores o causar ansiedad?

Si, debido a que son noticias, algunas son noticias falsas que pueden estar hechas precisamente para atacar a un sector o difamar a alguna persona.

¿El conjunto de datos contiene alguna subpoblación?

Si, dentro del conjunto de datos se pueden encontrar palabras que describen a estas subpoblaciones.

¿Es posible identificar individuos de manera directa o indirecta?

Si, si es posible identificar individuos de manera directa ya que sus nombres están ahí.

¿El conjunto de datos contiene información que podría considerarse sensible en cualquier sentido o ámbito?

Si, al ser noticias se pueden hablar sobre creencias religiosas, opiniones políticas, datos financieros, etc.

¿Otros comentarios?

No

Colección Procesos

¿Cómo se adquirieron los datos asociados con cada instancia?

¿Fueron los datos directamente observables, reportados por los sujetos, o inferidos/derivados indirectamente de otros datos? ¿se validaron los datos?

Los datos fueron extraídos de otros conjuntos de datos ya existentes, fueron directamente observados y no se validaron.

¿Qué mecanismo o procedimientos se usaron para recolectar los datos?

Desconocido para el autor del datasheet.

Si el conjunto de datos es una muestra de un conjunto más grande ¿cuál fue la técnica de muestreo?

No es un subconjunto de otros datos.

¿Quién estuvo envuelto en el proceso de la recolección de datos y como fueron compensados?

Quién recolecto los datos fue Saurabh Shahane y se desconoce si fue recompensado.

¿Durante que periodo de tiempo se recopilo el conjunto de datos?

¿Coincide este periodo de tiempo con el marco temporal de creación de los datos asociados a las instancias?

Esta recolección de dato se hizo en noviembre del 2021, desde ese momento se han ido agregando más noticias cada mes hasta la última actualización que fue en octubre del 2023.

¿Se realizaron procesos de revisión ética?

Desconocido para el autor del datasheet.

¿Recopilaste los datos directamente de los individuos en cuestión o los obtuviste a través de terceros u otras fuentes?

Los datos se obtuvieron del sitio web Kaggle.

¿Fuero los individuos en cuestión notificados de la recolección de datos?

No, se obtuvieron los datos de otros conjuntos de datos ya existentes.

¿Las personas en cuestión dieron su consentimiento para la recopilación y uso de sus datos?

No (ver la respuesta de la pregunta anterior)

Si se obtuvo el consentimiento, ¿se proporcionó a las personas que consintieron un mecanismo para revocar su consentimiento en el futuro o para ciertos usos?

No aplica.

¿Se ha realizado un análisis del impacto potencial del conjunto de datos y su uso en los sujetos de datos (por ejemplo, un análisis de impacto en la protección de datos)?

No aplica

¿Otros comentarios?

No

Preprocesamiento/ Limpieza/ Etiquetado

¿Se realizó algún preprocesamiento/limpieza/etiquetado de los datos (por ejemplo, discretización o agrupación, tokenización, etiquetado de partes del discurso, extracción de características SIFT, eliminación de instancias, procesamiento de valores faltantes)?

Si, primero se pasaron a minúsculas los textos, después se tokenizaron los textos utilizando la librería de spacy , con el modelo en\_core\_web\_sm una vez tokenizados se obtuvo el lemma de cada palabra y se reformo el texto original a la secuencia de lemmas del texto.

Después se filtraron stops words por definición, se removieron espacios múltiples, se removieron los signos de puntuación y las vocales que tenían acento se les quito.

¿Se guardaron los datos 'crudos' además de los datos preprocesados/limpiados/etiquetados (por ejemplo, para respaldar usos futuros no anticipados)?

Solo se guardaron los datos crudos, los procesados no.

¿Otros comentarios?

No

Uses

¿Se ha utilizado ya el conjunto de datos para alguna tarea?

Si, al ser un repositorio público en Kaggle muchas personas lo han descargado para hacer sus modelos.

¿Existe un repositorio que enlace a alguno o a todos los artículos o sistemas que utilizan el conjunto de datos?

Si, Shahane, S. (2023). Fake News Classification. Recuperado de https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification/versions/1/discussion

¿Para que otras tareas puede ser utilizado el conjunto de datos?

Este conjunto de datos puede ser utilizado para lo que sea relacionado con noticias o el modelado del lenguaje.

¿Hay algo en la composición del conjunto de datos o en la forma en que fue recopilado y preprocesado/limpiado/etiquetado que podría impactar usos futuros?

No se dio un tratamiento adecuado a nombres famosos, ya sea de empresas o personas, por lo que podría dañar la reputación de alguna persona.

¿Hay tareas para las cuales no se debería utilizar el conjunto de datos?

Este conjunto de datos se enfoca solamente en el análisis de noticias para ver si son verdaderas o falsas, no sirve para saber si lo que dice una persona es verdadero o falso dando alguna afirmación, por lo que este algoritmo solo se debería utilizar con textos catalogados como noticias.

¿Algún comentario adicional?

No

Distribución

¿Se distribuirá el conjunto de datos a terceros fuera de la entidad (por ejemplo, empresa, institución, organización) en nombre de la cual se creó el conjunto de datos?

Si, el conjunto de datos es publico y está disponible en internet.

¿Como se distribuirá el conjunto de datos?

El conjunto de datos será distrivuido en el sitio web de Kaggle en: Shahane, S. (2023). Fake News Classification. Recuperado de https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification/versions/1/discussion

¿Cuándo se distribuirá el conjunto de datos?

La primera distribución se hizo en el año 2021

¿Han impuesto terceras restricciones basadas en propiedad intelectual u otras restricciones sobre los datos asociados con las instancias?

No

¿Existen controles de exportación o restricciones regulatorias que apliquen al conjunto de datos o a instancias individuales?

Desconocido para el autor del datasheet

¿Otros comentarios?

No

Mantenimiento

¿Quién estará dando soporte/hosting/mantenimiento al conjunto de datos?

Saurabh Shahane es quién está dando soporte y mantenimiento al conjunto de datos mientras de kaggle los almacena.

¿Cómo se puede contactar al propietario/curador/administrador del conjunto de datos (por ejemplo, dirección de correo electrónico)?

En el siguiente link es donde se encuentra el conjunto de datos, ahí también hay una sección de discusión donde uno puede dejar sus comentarios y opiniones, el autor puede leer estas opiniones.

[Shahane, S. (2023). Fake News Classification. Recuperado de https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification/versions/1/discussion classification/versions/1/discussion](https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification/versions/1/discussion)

¿Hay algún erratum?

El link a la sitio web donde se almacenan los datos se van actualizando y corrigiendo, es decir ahí se encuentra la última versión

Shahane, S. (2023). Fake News Classification. Recuperado de https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification/versions/1/discussion¿El conjunto de datos será actualizado?

Si cada mes recibe una actualización.

Si el conjunto de datos se relaciona con personas, ¿existen límites aplicables a la retención de los datos asociados con las instancias (por ejemplo, se les informó a las personas en cuestión que sus datos serían retenidos por un período de tiempo fijo y luego eliminados)?

No aplica.

***Reflexión***

Riesgo1

Las personas podrían manipular la explicabilidad, es decir si las personas que se dedican a difamar y crear malas noticias pueden ver la explicación del algoritmo y modificar el cómo crean las noticias para que perezcan más verdaderas, lo que provocaría 3 cosas:

-Proliferación de noticias falsas

-Manipulación de la percepción del público

-Influir en la decisión de los votos

Mitigación 1

Dar una explicación limitada y local para que no entiendan el algoritmo de manera general, además de solo dar una lista corta, de las 9 principales características de esta manera no saben como afecta las palabras de manera en general.

Riesgo 2:

Las personas podrían tomar como verdadera el dictamen de la máquina sin dudar y en caso de una clasificación errónea puede ocasionar la expansión de una noticia falsa que se tome por verdadera, aumentando la desinformación electoral.

Mitigación 2

Poner un disclaimer después de cada output que diga que el modelo nos es 100% confiable y que siempre se debe verificarse con un experto o en fuentes oficiales.

Riesgo 3:

Las personas podrían comenzar a depender solo del criterio de la máquina y perder la capacidad para poner en sospecha una noticia

Mitigación 3

Después de cierto número de clasificaciones, presentar a los usuarios cuestionarios o encuestas breves que les hagan reflexionar sobre su proceso de pensamiento y cómo abordan la información que reciben.

Riesgo 4:

Que desacredite o quite voz a noticias donde una población vulnerable se esté viendo perjudicada y la máquina decida que no es así.

Mitigación 4

Establecer un mecanismo donde las noticias relacionadas con poblaciones vulnerables sean revisadas por un equipo humano antes de ser desacreditadas por el algoritmo. Esto puede ayudar a evitar que el algoritmo silencie injustamente asuntos importantes.

Riesgo 5:

Que se descredite una noticia por el simple hecho de aparecer un nombre famoso, es decir que, si aparece en nombre de Barack Obama en automático se clasifica como falsa.

Mitigación 5

Realiza auditorías y pruebas regulares para detectar cualquier sesgo en el algoritmo, particularmente en lo que respecta a la mención de nombres famosos.

Riesgo 6:

Que se acredite una noticia simplemente por el hecho de que aparezca un nombre famoso, es decir que, si en la noticia por aparecer Donald Trump en automático es verídica, lo que provocaría la proliferación de noticias falsas con el nombre de esta persona.

Mitigación 6

Realiza auditorías y pruebas regulares para detectar cualquier sesgo en el algoritmo, particularmente en lo que respecta a la mención de nombres famosos.

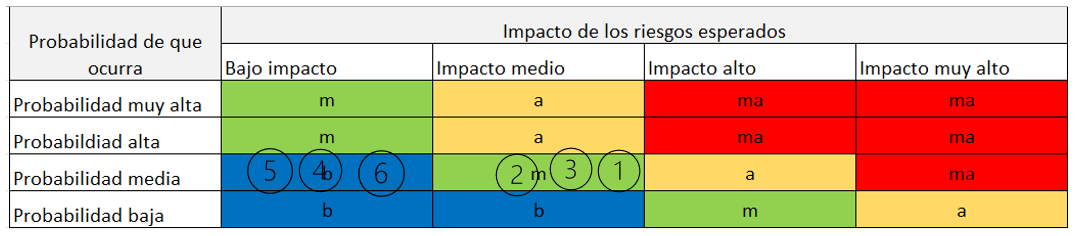


Tabla 10

En la tabla 10 se observa la nueva clasificación de los riesgos una vez analizados estos y dadas sus respectivas mitigaciones.

Requerimientos en producción

Requerimientos técnicos

Se requiere la versión 3.9 de Python con las siguientes librerías:

Numpy, Spacy,n úcleo de spacy: en\_core\_web\_sm, Nltk, Scikit-Sklearn, necesitando una infraestructura escalable y segura para el despliegue del modelo, asegurando la privacidad y seguridad de los datos.

Requerimientos operacionales

Se necesita establecer un monitoreo para detectar y corregir sesgo o errores en el modelo, así como regularmente actualizar el modelo con nuevos datos para mantener su relevancia.

Requerimientos organizacionales

Proporcionar al equipo de trabajo una formación continua en ética de la IA así como fomentar una cultura organizacional que priorice la ética y la responsabilidad.

Una vez completados el análisis ético y la evaluación de los beneficios y riesgos potenciales, junto con sus estrategias de mitigación correspondientes, se puede afirmar que el modelo está listo para ser implementado en producción, siempre y cuando se apliquen efectivamente las medidas de mitigación propuestas.

**V Conclusiones**

La implementación de una metodología que integre consideraciones éticas en el desarrollo despliegue y mantenimiento de sistemas inteligentes es una necesidad imperativa en esta era tecnológica. Esta investigación ha demostrado que, al adoptar un enfoque ético, podemos no solo mitigar posibles problemas que podrían surgir durante la fase de producción, sino también asegurar que estos sistemas actúen como facilitadores de bienestar y justicia social.

El enfoque en este artículo se centra en preservar la agencia humana, garantizando que las personas comprendan las razones detrás de las decisiones del sistema, como la clasificación de una noticia y el que puedan aprender a discernir entre una noticia real de una falsa. Esto es crucial para mantener la confianza en la tecnología y asegurar que los sistemas sean transparentes y responsables ante sus usuarios. Además, la metodología propuesta busca prevenir cualquier forma de injusticia o discriminación que pueda emanar del uso de algoritmos. Al priorizar estos aspectos éticos, los sistemas inteligentes se vuelven más confiables y robustos que se convierten en herramientas que pueden potenciar las capacidades humanas.

En el contexto específico de la clasificación de noticias falsas en campañas electorales, donde la información precisa y fiable es fundamental, nuestro enfoque ético adquiere aún mayor relevancia. Al integrar principios éticos en la clasificación binaria de noticias, podemos ayudar a salvaguardar el tejido democrático de nuestra sociedad, protegiendo a los ciudadanos de la desinformación y fortaleciendo su capacidad para tomar decisiones informadas.

En resumen, la metodología propuesta en este estudio no solo aborda los desafíos técnicos dentro del desarrollo de sistemas inteligentes, sino que también establece un nuevo estándar en la implementación de soluciones tecnológicas, uno que coloca a la ética y a la humanidad en el centro del progreso tecnológico.

**Referencias**

1-Magallón Rosa, R. (2019). Desinformación en campaña electoral.

2-Pérez, C. R. (2019). No diga fake news, di desinformación: una revisión sobre el fenómeno de las noticias falsas y sus implicaciones. *Comunicación*, (40), 65-74.

3- Espejel-Rivera, M. A., Calderón-Suárez, R., Ortega-Mendoza, R. M., Camacho-Bello, C. J., & Máquez-Vera, M. A. (2022). Detección automática de noticias falsas usando representaciones textuales tradicionales y soluciones basadas en aprendizaje profundo. *Pädi Boletín Científico de Ciencias Básicas e Ingenierías del ICBI*, *10*(Especial3), 120-127.

4-Flores Vivar, J. M. (2019). Inteligencia artificial y periodismo: diluyendo el impacto de la desinformación y las noticias falsas a través de los bots.

5- Ochando, M. B. (2018). El problema de las noticias falsas: detección y contramedidas. *XV Seminario Hispano–Mexicano de Investigación En Biblioteconomía y Documentación, Ciudad de México*, 1-18.

6- Sparks, A., & Frishberg, A. (2020). Approaches to Identify Fake News: A Systematic Literature Review. Recuperado de <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7250114/>

7 - Alim Al Ayub Ahmed,Ayman Aljarbouh,Praveen Kumar Donepudi & Myung Suh Choi. (2021). Detecting Fake News using Machine Learning: A Systematic Literature Review. Recuperado de https://arxiv.org/ftp/arxiv/papers/2102/2102.04458.pdf#:~:text=,Introduction%20World%20is%20changing%20rapidly

8- Muhammad Fayaz, Atif Khan, Muhammad Bilal & Sana Ullah Khan . (2022). Machine learning for fake news classification with optimal feature selection. Recuperado de https://link.springer.com/article/10.1007/s00500-022-06773-x

9-Jamal Abdul Nasir a b, Osama Subhani Khan b & Iraklis Varlamis. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. Recuperado de https://www.sciencedirect.com/science/article/pii/S2667096820300070

10- Alim Al Ayub Ahmed,Ayman Aljarbouh,Praveen Kumar Donepudi & Myung Suh Choi. (2021). Development of Fake News Model Using Machine Learning through Natural Language Processing: A Systematic Literature Review. Recuperado de <https://arxiv.org/ftp/arxiv/papers/2201/2201.07489.pdf#:~:text=,incorporating%20those%20features%20into%20classification>

11- Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI." Berkman Klein Center for Internet & Society, 2020.

12- Instituto Nacional Electoral (INE). (2014). Ley General de Instituciones y Procedimientos Electorales. https://www.ine.mx/wp-content/uploads/2020/07/Despen-LEGIPE-NormaINE.pdf.

13- Diario Oficial de la Federación (DOF). (2018). Ley General de Comunicación Social. https://www.dof.gob.mx/nota\_detalle.php?codigo=5522344&fecha=11/05/2018#gsc.tab=0.

14- Federación de Asociaciones de Periodistas Mexicanos (FAPERMEX). (2017). Código de Ética del Periodista Mexicano.https://www.fapermex.reportajesmetropolitanos.com.mx/.

15- Diario Oficial de la Federación (DOF). (2010). Ley Federal de Protección de Datos Personales en Posesión de Particulares. https://dof.gob.mx/nota\_detalle.php?codigo=5150631&fecha=05/07/2010#gsc.tab=0.

16- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020, January). Closing the AI accountability gap: defining an end-toend framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency