**Resumen**

**Introducción**

Las Fake News:

El termino Fake News se ha vuelto muy popular en el año 2017 y cuyo uso ha ido aumentando con el pasar del tiempo. Pero ¿Qué son las Fake News? El periodista Iñaki Gabilondo (2021) presenta lo siguiente:

Las fake news son muchas cosas. Por un lado, las mentiras de toda la vida, amplificadas con la nueva realidad. Por otro lado, son una industria: hay gente que ha descubierto esa realidad y ha hecho maniobras intencionadamente dirigidas a buscar objetivos. Industrias operando para difundir mentira. Son una amenaza brutal, porque producen una gran desconfianza, no sabes ni puedes distinguir qué es lo que viene avalado por algo medianamente creíble. (conferencia "La falsa verdad: ¿Hay antídotos?")

Que inconvenientes tiene que se circulen noticias falsas? La difusión de noticias falsas puede causar efectos negativos masivos, a veces afectando o manipulando eventos públicos importantes. ([Detecting fake news by exploring the consistency of multimodal data - ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S0306457321001060" \l "preview-section-snippets))

¿Por qué si contamos con acceso a grandes cantidades de información, más propenso se vuelve la divulgación de las noticas falsas?

Introducción: (borrar)

En los últimos años es notable un avance en el uso de los Fake News, usadas para manipular al espectador, generando así una opinión colectiva sobre una persona (políticos, celebridades, figuras públicas) como también sobre temas especiales para divulgar información errónea. La importancia de identificar si una noticia es o no una Fake News es en torno a que si esta trata sobre un tema polémico, podría desatar un conflicto entre países, ya que hay diversos tipos de información falsa, tanto escrita, como en diarios, folletos, tweets, etc., como así también en contenidos audiovisuales, como videos o fotografías.

**Marco Teórico: (LISTO)**

La principal dificultad en la detección de noticias falsas consiste en encontrar las características que las hacen diferentes de las verdaderas (Xue et al., 2021). En la búsqueda de los atributos que diferencian las noticias verdaderas de las falsas, Xue et al. (2021) las encuentra observando que las noticias falsas hacen uso de imágenes y textos con una alta orientación sentimental para su rápida propagación. Por otro lado, Mottola (2020) analizó la relación de textos fraudulentos con estructuras fijas de expresión y un estilo de narración muy impactante (títulos en mayúscula, signos de puntuación exclamativos), como así también el uso de imágenes explicitas y llamativas.

Debido a la complejidad de la detección de las mismas, Zhou y Zafarani (2020) establecieron cuatro perspectivas para la detección de noticas fraudulentas: conocimiento, estilo, propagación y credibilidad. Los métodos basados en el primero, evalúan la autenticidad del escrito, mientras que, en el estilo, analizan la forma en la que está escrito. Por otro lado, en el estudio con perspectiva para la propagación, se analiza la información de la propagación y el usuario que la difunde. Por último, en el estudio de la credibilidad de la noticia se centra en el análisis en el perfil de quien crea y difunde estas publicaciones.

Zhou y Zafarani (2020) proponen realizar la tarea de detectar si una noticia es verdadera o no de dos formas: la primera y más antigua, verificar los datos de forma manual, esta forma es demasiado tardada, ya que se necesitaría que un grupo de personas se tomen el tiempo de realizar un análisis profundo de todos los artículos que se deseen clasificar o corroborar, teniendo en cuenta que los datos que se analizan se encuentran en grandes volúmenes de noticias que existen en las redes sociales o el internet, siendo esta además una forma no tan eficaz. Y la segunda es de forma automatizada con modelos de Aprendizaje Automático (ML, por sus siglas en ingles) y Procesamiento del Lenguaje Natural (NLP, por sus siglas en ingles). Esta segunda forma es mucho más veloz que la primera ya que se utiliza la ciencia de la computación, pudiendo así analizar los grandes volúmenes de datos de forma eficiente y pudiendo tener la certeza de la precisión obtenida en un modelo de ML o DL.

En nuestro trabajo analizaremos el contenido de los textos, y lo haremos de forma automática con procesos de NLP y modelos de ML. Con respecto a las herramientas de PLN que se basan en el contenido del texto, se han usado diferentes formas de analizar el contenido y clasificarlo. Las comúnmente utilizadas son TF (*Term Frecuency*), n-gramas, TF-IDF (*Term Frecuency-Inverse Document Frequency*), etiquetado POS (*Part of SpeechTagging*), incrustaciones de palabras (*embeddings*), Count Vectorizer y el uso de diccionarios lingüísticos. (Alvarez et al., 2020).

**Term Frecuency-Inverse Document Frequency**

**(Explicar TF-IDF)**

**Pasos para procesos de PLN**

El proceso de NLP tiene una secuencia de pasos:

1. **Tokenizacion:** explicar la importancia de tokenizar y su significado
2. **Eliminacion de StopWords**

El proceso de NLP tiene una secuencia de pasos, primero se tokeniza el texto a tratar (Dividir el texto en unidades funcionales para su análisis), eliminación de StopWords (palabras redundantes), Lematización (encuentra la base de las palabras) y por último uso de TF-IDF (número de veces que aparece un token en el texto comparado con todos los textos a analizar) (Zapata García, A., 2021). Esta estrategia puede variar dependiendo el trabajo a realizar, pero aun así siempre se seguirá la misma estructura hasta la acción de lematizar, de allí en adelante se puede proceder como sea necesario, o además se pueden agregar procesos entre medio, como puede ser la creación de n-gramas.

Por otro lado, los algoritmos de aprendizaje automático más utilizados son: Arboles de decisión, Regresión Logística, Random Forest y Suport Vector Machine (SVM). Los algoritmos que mejor resultados de accuracy tienen son SVM, llegando a alcanzar valores hasta de 99,90%, Regresión Logística con 91,60% y Gradient Descent con 91,05%, aunque este último es muy poco utilizado. (Álvarez et al., 2021; Zhou and Zafarani, 2020).

*Explicar los modelos que voy a utilizar en mi trabajo (Arbol de decisión, Random Forest, SVM, Regresión Logística*

**Marco metodológico:**

**Problema/Justificacion:**

Es de público conocimiento el problema que traen las Fake News, como explicamos anteriormente, la difusión de noticias falsas puede causar efectos negativos masivos, a veces afectando o manipulando eventos públicos importantes (Xue et al., 2021). Es por esto que Castillo et al. (2011) llega a la idea de que eliminar las noticias falsas es de gran necesidad para mejorar la calidad de las redes de información y mantener la estabilidad social, evitando su desinformación y manipulación. Por otro lado, es importante mejorar la detección de fake news para así disminuir las mismas, las cuales sufrieron un aumento en temas polémicos, como es el caso del COVID-19, gran cantidad de noticias falsas se pusieron en circulación, generando pánico entre los ciudadanos, como es el caso de la siguiente información “Todos deben asegurarse de que su boca y garganta estén húmedas y nunca secas. Tome sorbos de agua cada 15 minutos porque incluso si el virus llega a la boca al beber agua o líquidos, lo arrastrará a través del esófago y el estómago, donde el ácido del estómago matará al virus.” o proponiendo alimentos o medicamentos que curarían el COVID-19, que podría provocar que el infectado se auto medique, poniendo su vida en riesgo.

objetivos

Es por esto que se propone como objetivo general el de crear un algoritmo de Machine Learning que, con usos de Procesamiento del Lenguaje Natural, pueda clasificar si una noticia es o no una Fake News. Además, se tratará de alcanzar modelos con un 70% de exactitud (ver si llego) para así este pueda predecir, ya que si el modelo llegase a predecir en un 60% de los casos, no tendría la misma utilidad el trabajo realizado, ya que una persona, si bien se tardaría mas tiempo, es probable que clasifique con mayor exactitud.

**Herramientas para el desarrollo:**

El desarrollo de los modelos de predicción se realizó con el lenguaje de Python, teniendo este una gran cantidad de librerías que se utilizan para los procesos de PLN. En este trabajo utilizaremos específicamente las librerías de NLTK, Scikit-Learn, Spacy y Collections para el desarrollo de Procesamientos del Lenguaje Natural, además se utilizarán herramientas de análisis de datos como Pandas y Numpy. Por otro lado, la herramienta para la representación gráfica de los datos será Plotly, ya que nos aportará una gran cantidad de gráficos interactivos. Para la parte de modelos de aprendizaje automático se analizará como es el funcionamiento de 4 modelos, SVC, Arbol de decisión, Random Forest y Regresion Logística. Como puede ver, todos los modelos utilizados pertenecen a problemas de clasificación, puesto que lo que nosotros queremos determinar es saber si una noticia es verdadera o no.

**Mi set de datos:**

(Mostrar información y descripción acerca de mi set de datos (dividir mis datos en partes (testeos, entrenamientos (train, test))))

**Funcion de preprocesamiento**

Explicar lo que mi función de preprocesamiento realiza, poniendo detalle en la minusculizacion, eliminación de números, tokenizacion, filtrado de stop Words, lematizacion

**TF-IDF Vectorizer**

Explicar el funcionamiento de TF-IDF Vectorizer

**Explicacion de modelado de datos**

Explicar toda la búsqueda del mejor modelo, mostrando sus accuracy y por último la elección de hiperparametros con el mejor modelo.

**Analisis del resultado obtenido**

**Conclusiones:**

**Bibliografia**

fake news

[Iñaki Gabilondo: “Decir que nosotros hacíamos periodismo de verdad y el resto es basura es una visión cateta” - Información (informacion.es)](https://www.informacion.es/cultura/2021/11/12/decir-haciamos-periodismo-resto-basura-59454504.html)

[sklearn.feature\_extraction.text.TfidfVectorizer — scikit-learn 1.1.3 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html)

[10 'fake news' sobre el coronavirus – El Financiero](https://www.elfinanciero.com.mx/salud/10-fake-news-sobre-el-coronavirus/)

<https://normas-apa.org/formato/tamano-de-pagina/>

marco teorico

[Machine Learning: Modelo clasificador de textos en Python – Arnau Dunjó Workspace (arnaudunjo.com)](https://arnaudunjo.com/es/2021/04/25/machine-learning-modelo-clasificador-de-textos-en-python/)

[(3) ¡Nuevos mensajes! (upgrad.com)](https://www.upgrad.com/blog/fake-news-detection-project-in-python/)

[Fake News Detection Project in Python with Machine Learning - Project Gurukul](https://projectgurukul.org/fake-news-detection-project-python-machine-learning/#:~:text=Fake%20News%20Detection%20Project%20in%20Python%20with%20Machine,3%3A%20Train%20and%20Evaluate%20Fake%20News%20Detection%20Model)

[▷ ¿Qué es la tokenización en el procesamiento del lenguaje natural (NLP)? | Machine Learning Studio® 2022 (jaol.net)](https://mlstudio.jaol.net/que-es-la-tokenizacion-en-el-procesamiento-del-lenguaje-natural/)

informes pln

[Vista de Detección automática de noticias falsas usando representaciones textuales tradicionales y soluciones basadas en aprendizaje profundo (uaeh.edu.mx)](https://repository.uaeh.edu.mx/revistas/index.php/icbi/article/view/9008/9326)

paper xue e tal

[Detección de noticias falsas explorando la consistencia de los datos multimodales - ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S0306457321001060#preview-section-snippets)

atora

[Modelo de aprendizaje basado en características lingüísticas para la detección y clasificación de noticias falsas - ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S095741742030909X)

zhou

zhou-csur-survey.pdf

[Deteccion-de-Noticias-Falsas-en-Redes-Sociales-Basada-en-Aprendizaje-Automatico-y-Profundo-Una-Breve-Revision-Sistematica.pdf (researchgate.net)](https://www.researchgate.net/profile/Juan-Holgado-Terriza/publication/349120690_Deteccion_de_Noticias_Falsas_en_Redes_Sociales_Basada_en_Aprendizaje_Automatico_y_Profundo_Una_Breve_Revision_Sistematica/links/604e0e8d458515e529aa4949/Deteccion-de-Noticias-Falsas-en-Redes-Sociales-Basada-en-Aprendizaje-Automatico-y-Profundo-Una-Breve-Revision-Sistematica.pdf)

desarrollo de pln

[tfg-zap-ana.pdf (upct.es)](https://repositorio.upct.es/bitstream/handle/10317/9647/tfg-zap-ana.pdf?sequence=1&isAllowed=y)

castillo et al

[castillo\_mendoza\_poblete\_2010\_twitter\_credibility.pdf (chato.cl)](https://chato.cl/papers/castillo_mendoza_poblete_2010_twitter_credibility.pdf)

sdfsdfsdfssfsdfsd