**Resumen**

**Introducción**

**Marco Teórico:**

**1- Formas de detectar las Fake News (pensar titulo)**

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

Debido a la complejidad de la detección de las mismas, Zhou y Zafarani establecieron cuatro perspectivas para la detección de las mismas: *conocimiento*, evalúan la autenticidad del texto, *estilo*, analizar la forma en el que el texto fue escrito, *propagación*, se analiza la información de la propagación y el usuario que la difunde, y por último *credibilidad*, se centra en el análisis en el perfil de quien crea y difunde estas publicaciones (2020). Una vez elegida la perspectiva a trabajar, Zhou y Zafarani proponen realizar la tarea de detectar noticias fraudulentas de dos formas: la primera y más antigua, verificar los datos de forma manual, esta forma es lenta, ya que es necesario que un grupo de personas realicen el análisis de la gran cantidad de artículos. La segunda forma de detectar una Fake News 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 inglés), siendo esta más veloz que la primera (2020).

**2- Procesamiento del Lenguaje Natural (PLN) y Herramientas**

Según Cortez Vásquez et al. PLN es la utilización de un lenguaje natural para comunicarnos con las computadoras, y esta entienda las oraciones, el uso de los lenguajes naturales y facilitar el desarrollo de programas que realicen tareas relacionadas con el lenguaje, como así también ayudar a comprenderlos. (2008). Lo mismo destacan Deng & Liu expresando que el PLN se encarga de investigar el uso de computadoras para poder procesar o entender el lenguaje natural con el propósito de realizar tareas útiles. Se define como un campo interdisciplinario que combina la lingüística computacional, ciencias de la computación, ciencia cognitiva e Inteligencia Artificial (2018). Ambos autores concuerdan que el uso de PLN se realiza para procesar o comprender el lenguaje natural y utilizar el mismo para realizar futuras tareas. Pero ¿De qué forma se realiza este análisis? Según Zapata el proceso de PLN debe realizarse de manera sistemática, dividiendo el elemento en partes, agregando elementos gramaticales e identificando elementos interesantes, como así eliminando elementos que no aportan información. Además, recalca que alguno de los elementos comunes del procesamiento en NLP son: Tokenización, Eliminación de palabras redundantes (Stop Words), Lematización y TF-IDF (2021).

* 1. **Tokenización**

Según Baciero Fernández, J la tokenización es la tarea de dividir en pedazos una secuencia de caracteres. Un token es un elemento de una secuencia de caracteres de un documento, y si estos se agrupan se puede encontrar relaciones para el procesamiento del documento (2020).

* 1. **Eliminación de palabras redundantes (StopWords)**

ZEOKAT afirma que el concepto de las stop words nace en el mundo de la lingüística computacional y se su uso en el PLN. El termino hace referencia a todas aquellas palabras que no tienen un significado por si solas, que suelen ser pronombres, preposiciones, artículos, y demás. (2017)

* 1. **Lematización**

Según un usuario de TechEdu llamado Katrina, la lematización es el proceso de reducir una palabra a su forma básica. Mayormente se suele realizar eliminando los sufijos de flexión (-ed, -ing o -es) (2022). Mientras tanto theastrologypage define a la lematizacion como un proceso de PLN, que implica trabajar con palabras de acuerdo con sus componentes léxicos raíz, permitiendo así agrupar palabras que tengan el mismo significado pero escritas de diferentes formas (2022)

* 1. **TF-IDF**

Según Kang et al. el algoritmo TF-IDF se usa para evaluar la importancia de las palabras en el corpus, y esta importancia es proporcional al número de veces que la palabra aparece en el documento, pero a su vez inverso a la frecuencia de palabras que aparecen en el corpus (2016). Chinty et al. expresa que el termino TF representa la frecuencia absoluta de una palabra, e IDF determina la importancia del termino en el conjunto, se obtiene calculando el logaritmo del número total de documentos contenidos en el corpus sobre el número de documentos donde aparece el término a analizar. Para concluir el cálculo, se multiplican los dos términos (TF e IDF) generando así nuestra variable TF-IDF, que determina si una palabra tiene mayor peso cuando esta sea más significativa en el texto (2022).

**3- Algoritmos de Machine Learning**

El autor Lugo sostiene que el termino *Machine Learning* hace referencia a la ciencia que permite que las computadoras aprendan y actúen como lo hacen los humanos, alimentándolas con datos e información que representan representaciones del mundo real (2020). Por otro lado, *Machine Learning* (ML) usa algoritmos de Inteligencia Artificial (IA), por lo cual, según Arbeláez-Campillo et al. si contamos con información cualitativa y cuantitativa, podría alcanzar varios objetivos propuestos (2021). Según Digital55 et al. utiliza algoritmos de ML con la finalidad de determinar los patrones de los datos para luego mediante modelos, realizar predicciones (2020). Cajahuanca et al. dan a conocer los modelos de clasificación de ML, estos son algoritmos que cuentan con IA, que ayudan a aprender a una máquina, produciendo modelos matemáticos capaces de generalizar el comportamiento de los datos

**3.1- Maquina de Vectores de Soporte (SVM, por siglas en inglés)**

Según indica Xia et al. SVM es un método de aprendizaje automático supervisado, que se basa en la teoría del aprendizaje estadístico (2022). Por otro lado, Fayed & Atiya ponen a este modelo como uno de los mejores métodos de clasificación. La desventaja de estos modelos es que son muy sensibles a los parámetros del mismo (2019). IBM explica que el modelo funciona correlacionando datos y representándolos en un plano de x dimensiones. Luego se detecta un separador entre las diferentes categorías. Realizado esto, las futuras observaciones se podrán predecir a que grupo pertenece dependiendo de donde se encuentre en este plano (2021).

**3.2- Regresión Logística**

Bravo et al. define al algoritmo de regresión logística (RL) como un modelo que genera una función exponencial, donde se calcula la probabilidad de que la variable a predecir adopte el valor, en función de las variables independientes (2022). Por otra parte (Granada et al., 2022) “Es un modelo estadístico que permite conocer la relación entre variables dependiente dicotómicas y un conjunto de variables independientes, su principal característica se da en las variables independientes, estas podrán ser cualitativas o cuantitativas dependiendo de la investigación”. Una de las facilidades que brinda el uso de RL es su fácil interpretación si se conoce el concepto de los *Odds ratio* (OR). Cerda et al. lo definen como la probabilidad de ocurrencia de un acontecimiento utilizando probabilidades calculadas de la siguiente manera: (2013)



**3.3- Random Forest**

Hosseinpour-Zarnaq et al. presentan que *Random Forest* (RF) es una evolución a partir de los árboles de decisión (DT), precisamente, RF consta de muchos DT. Este modelo es una herramienta de aprendizaje automático de conjunto, ya que, para el procesamiento de los datos, cada DT clasifica los datos de entrada y al finalizar, RF recopila las clasificaciones y elije la predicción más votada como su resultado. Este modelo tiene un buen rendimiento en los problemas donde el número de variables es mucho mayor que el de las muestras (2022).

**4- Herramientas y modulos para el desarrollo del trabajo**

**4.1 Python**

Según Pérez Álvarez Python es un lenguaje de programación orientado a objetos, y es el que más creció en uso en los últimos años, principalmente por su facilidad de uso y su gran cantidad de librerías de ML y PLN (2021). Además, Baciero Fernández indica que el modo de trabajo con Python es determinado por la combinación de módulos y paquetes que se pueden importar para añadir nuevas funciones al análisis o desarrollo de un proyecto, pudiendo estos módulos permitir la tarea de entrenamientos de modelos de ML, o importar herramientas para desarrollar PLN(2020)

**4.1.1 NLTK**

Baciero Fernández expresa que NLTK (Natural Language ToolKit por sus siglas en ingles) es una de los principales módulos para desarrollar PLN. Este proporciona algoritmos para el análisis de diferentes recursos léxicos, pertenecientes a un conjunto de bibliotecas para el procesamiento de texto tales para la casificacion de los mismos, tokenizacion, derivación, etiquetado, análisis sintáctico y razonamiento semántico. Está diseñada para trabajar con IA, y áreas de PLN (2020).

**4.1.2 Spacy**

Basándonos en lo que dice Baciero Fernández, spaCy es una librera usada para desarrollar PLN en Python. Contiene algoritmos diseñados para el desarrollo de soluciones en ámbito industrial y productivo, por lo que se utiliza para generar aplicaciones que procesen y comprendan grandes volúmenes de datos. Proporciona módulos como la Tokenizacion, Lematizacion, Reconocimiento de entidades nominales, entre otras. (2020)

**4.1.3 Pandas**

Pandas es un paquete de Python que permite crear e

**4.1.4 NumPy**

NumPy es un paquete de procesamiento de matrices de propósito general.

**4.1.5 Matplotlib**

**4.1.6 scikit-learn**

**Marco metodológico:**

**Problema/Justificacion:**

En los últimos años, es notable el incremento de la circulación de noticias, textos y demás artículos informativos gracias a la implementación de las redes sociales y del Internet, donde las noticias o rumores se difunden a altas velocidades, pero a su vez se vuelve más difícil identificar cuando una noticia es verdadera o no. “En tiempos de inundación, lo primero que escasea es el agua potable. En esa inundación de señales informativas, lo que escasea es esa agua potable informativa, que ahora es absolutamente vital” (Gabilondo, 2021). Esta frase de Gabilondo representa a la perfección lo anteriormente dicho, ya que en internet abundan estas noticias falsas, noticias erróneas, mala información, y se las denomina comúnmente como Fake News. El mismo término se ha vuelto muy popular en el año 2017. Pero ¿Qué es una Fake News? Gabilondo propone que son muchas cosas, por un lado, mentiras amplificadas, pero por otro lado es una industria para difundir mentiras, generando asi gran desconfianza de la sociedad hacia lo que leen en los medios (2021). Según Mishra et al. describe a las Fake News como cualquier material capaz de hacer creer a los lectores en información que no es real

Es de público conocimiento el problema que traen estas noticias erróneas, Xue et al. recalca que la difusión de noticias falsas puede causar efectos negativos masivos, a veces afectando o manipulando eventos públicos importantes (2021). Es por esto que Castillo et al. 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 (2011). 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 información que circulaba por las redes afirmando lo siguiente: “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. Es por esto que se desarrolla el trabajo a continuación explicado, pudiendo este en un posible futuro ser utilizado para disminuir estas noticias fraudulentas y así mejorar la calidad de información que consume la sociedad.

**Objetivo General:**

Generar una aplicación que detecte si una noticia es o no una Fake News que cuente con herramientas de Procesamiento del Lenguaje Natural. El mismo trabajo deberá finalizarse antes del 18/11/22

**Objetivos específicos:**

* Caracterizar y analizar los datos que se utilizarán en la aplicación.
* Desarrollar una función de pre-procesamiento aplicando herramientas de PLN
* Desarrollar un algoritmo de Aprendizaje Automático, generando modelos de predicción, cuyo nivel de precisión sea superior al 70%

**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, 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)

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

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

[TFG\_JULEN\_PEREZ\_ALVAREZ.pdf](file:///C:\Users\Usuario\Desktop\lautaro\ies%2021\IES21\4to%20cuatrimestre\PLN\informes%20academicos\TFG_JULEN_PEREZ_ALVAREZ.pdf)

[zhou-csur-survey.pdf](file:///C:\Users\Usuario\Downloads\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)

[El análisis del aprendizaje automático habilitó técnicas de detección de noticias falsas para conjuntos de datos diversificados (hindawi.com)](https://www.hindawi.com/journals/wcmc/2022/1575365/)

[14492 (upm.es)](https://oa.upm.es/62858/1/TFG_JOSE_IGNACIO_BACIERO_FERNANDEZ.pdf) (informe de 197 pag)

[Vista de Procesamiento de lenguaje natural (unmsm.edu.pe)](https://revistasinvestigacion.unmsm.edu.pe/index.php/sistem/article/view/5923/5121)

[El análisis del aprendizaje automático habilitó técnicas de detección de noticias falsas para conjuntos de datos diversificados (hindawi.com)](https://www.hindawi.com/journals/wcmc/2022/1575365/)

[GUÍA PARA LA (ug.edu.ec)](http://repositorio.ug.edu.ec/bitstream/redug/59656/1/B-CISC-PTG%232028-A%c3%b1o%202022%20Vera%20Guam%c3%a1n%20Kerlly%20del%20Pilar%20-%20Yela%20Garc%c3%ada%20Ricardo%20Alberto.pdf)

Stop words

[¿Qué son las stop words o palabras vacías? - Vozidea.com](https://www.vozidea.com/que-son-las-stop-words-o-palabras-vacias)

Lema

[Lematización - Definición y explicación (techlib.net)](https://techlib.net/techedu/lematizacion/)

[¿Qué es la lematización? - definición de techopedia - Audio 2022 (theastrologypage.com)](https://es.theastrologypage.com/lemmatization)

TF-IDF

[Netflix-Recommendation-System-based-on-TF-IDF-and-Cosine-Similarity-Algorithms.pdf (researchgate.net)](https://www.researchgate.net/profile/Mohamed-Chiny/publication/360856267_Netflix_Recommendation_System_based_on_TF-IDF_and_Cosine_Similarity_Algorithms/links/63035829aa4b1206facd2842/Netflix-Recommendation-System-based-on-TF-IDF-and-Cosine-Similarity-Algorithms.pdf)

Kang, G., Tang, M., Liu, J., Liu, X., and Cao, B., 2016. Diversifying web service recommendation results via exploring service usage history. IEEE Transactions on Services Computing, vol. 9.

Machine learning

Lugo, A. (2020). ¿Qué es el Machine Learning? INVID. https://invidgroup. com/es/machine-learning-metodos/

Arbeláez-Campillo, D. F., Villasmil, J. J., y Rojas-Bahamón, M. J. (2021). Inteligencia artificial y condición humana: ¿Entidades contrapuestas o fuerzas complementarias? Revista de Ciencias Sociales (Ve), XXVII(2), 502-513. [https://doi.org/10.31876/rcs. v27i2.35937](https://doi.org/10.31876/rcs.%20v27i2.35937)

[Dialnet-DesercionUniversitaria-8526463.pdf](file:///C:\Users\Usuario\Downloads\Dialnet-DesercionUniversitaria-8526463.pdf)

Digital55 (01 de Junio de 2020). Qué es Machine Learning: casos de éxito en empresas. Digital55. https:// www.digital55.com/innovacion/quees-machine-learning-casos-exitoempresas

SVM

[Performance optimization of support vector machine with oppositional grasshopper optimization for acute appendicitis diagnosis - ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S0010482521010003)

[Acelere la búsqueda en cuadrícula para la selección de parámetros de máquinas de vectores de soporte - ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S1568494619301632)

[Funcionamiento de SVM - Documentación de IBM](https://www.ibm.com/docs/es/spss-modeler/saas?topic=models-how-svm-works)

Regresion Logistica

[Vista de Degradación y deforestación en la cuenca del río Conchos (México): modelado predictivo mediante regresión logística (1985-2016) (ugr.es)](https://revistaseug.ugr.es/index.php/cuadgeo/article/view/21629/22756)

[FCE-CEST-CONDOR ELENA-CORREA SANTIAGO.pdf (uce.edu.ec)](http://www.dspace.uce.edu.ec/bitstream/25000/26706/3/FCE-CEST-CONDOR%20ELENA-CORREA%20SANTIAGO.pdf)

Cerda, J., Vera, C., & Rada, G.(2013). Scielo. Obtenido de Revista de medicina Interna Pontifica Universidad Católica de Chile: https://www.scielo.cl/scielo.php?script=sci\_arttext&pid=S0034- 98872013001000014

Random Forest

[Diagnóstico de fallos de la caja de cambios auxiliar del tractor mediante análisis de vibraciones y clasificador de bosques aleatorio - ScienceDirect](https://www.sciencedirect.com/science/article/pii/S2214317321000020)