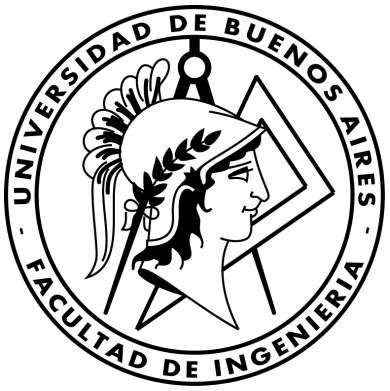

FACULTAD
DE INGENIERIA
Universidad de Buenos Aires

TRABAJO PROFESIONAL

INGENIERÍA EN INFORMÁTICA

2do Cuatrimestre 2023 - 1er Cuatrimestre 2024

Factify: Algoritmo de detección de noticias falsas en español con análisis de sentimiento

---

Integrantes

- Fuentes Azul Lucila
  <afuentes@fi.uba.ar>

- Iskandarani Roberto Ezequiel
  <riskandarani@fi.uba.ar>

- Pardo Lucía
  <lpardo@fi.uba.ar>

- Reinaudo Dante
  <dreinaudo@fi.uba.ar>

Tutor

- Martin Buchwald
  <mbuchwald@fi.uba.ar>

Resumen

El presente trabajo profesional busca generar un algoritmo capaz de abordar la problemática de la detección automática de noticias falsas en español, haciendo un enfoque mixto entre técnicas basadas en el contenido y técnicas basadas en la fuente. Esto se realizará utilizando distintos algoritmos del estado del arte del aprendizaje automático y/o aprendizaje profundo. Los resultados obtenidos serán evaluados a partir del uso de diversas métricas para comparar y analizar la calidad de cada modelo aplicado.

El abordaje del proyecto se llevará a cabo utilizando la metodología CRISP-DM, una de las metodologías más populares dentro del proceso de minería de datos. Esta consta de seis etapas: Comprensión del Negocio, Comprensión de los Datos, Procesamiento de los Datos, Modelado, Evaluación e Implementación.

Los datos utilizados serán recopilados a partir de distintas fuentes, como datasets preexistentes orientados a noticias falsas, datos recopilados de distintas páginas de *fact-checking* y recolección de datos de redes sociales, utilizando la técnica de *scraping*.

Una vez obtenido el algoritmo con mejor performance, se busca crear una página web que provea el servicio de clasificación y detección de noticias falsas en tiempo real. Esta herramienta permitirá a los usuarios corroborar la veracidad de la información que consumen.

Palabras clave

Noticias Falsas, Inteligencia Artificial, Aprendizaje Automático, Aprendizaje Profundo, Procesamiento de Lenguaje Natural, Redes Sociales

Abstract

The present project attempts to generate an algorithm capable of addressing the problem of automatic detection of fake news in Spanish language, using a mixed approach between content-based techniques and source-based techniques. This will be done using different algorithms from the state of the art of machine learning and/or deep learning. The results obtained will be evaluated through the use of several metrics to compare and analyze the quality of each applied model.

The project will be performed using the CRISP-DM methodology, one of the most popular methodologies within the data mining process. This methodology consists of six stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment.

The data used will be collected from different sources, such as pre-existing datasets, data collected from different fact-checking pages and data collection from social networks, using the scraping technique.
Once the algorithm with the best performance has been obtained, the aim is to create a web page that provides the service of classification and detection of fake news in real time. This tool will allow users to corroborate the veracity of the information they consume.

Keywords

Fake news, Artificial Intelligence, Machine Learning, Deep Learning, Natural Processing Language, Social Networks

1. Introducción

1.1 Contexto

Hoy en día las redes sociales son una de las principales fuentes de consumo de información de la sociedad, según un informe digital publicado en 2022[1] en la actualidad existen 4.620 millones de usuarios de redes en todo el mundo, lo que equivale a más del 58% de la población total mundial. La enorme cantidad de usuarios conectados entre sí ha logrado que la información se propague a altas velocidades. De acuerdo a Statista[2], actualmente en YouTube se suben más de 500 horas de video por minuto y se ven más de 5 billones de videos al día, mientras que en Twitter (X) se generan alrededor de 500 millones de tweets por día (350.000 por minuto). Esta abundante oferta de información junto con su rápida difusión, han generado un contexto en el cual resulta muy difícil distinguir la veracidad de los hechos difundidos. Como consecuencia, la proliferación de noticias falsas y campañas de desinformación han aumentado exponencialmente en los últimos años, ocasionando un impacto directo en nuestra sociedad, afectando la opinión pública, la toma de decisiones y agudizando la polarización política.

A su vez, la esencia misma de las redes sociales ha propiciado la emergencia de un fenómeno conocido como cámara de eco o efecto *echo chamber*. Este fenómeno se manifiesta cuando los usuarios tienden a reforzar y validar sus opiniones al cerrarse deliberadamente en un círculo informativo o burbuja de información, en donde el contenido al que están expuestos refleja, en gran medida, sus propias perspectivas, ideologías y ubicación geográfica, entre otros aspectos. Este proceso autoconfirmatorio conlleva a una reducción progresiva en la exposición a opiniones divergentes, contribuyendo así a la limitación del entendimiento de diversas perspectivas, profundizando así la polarización de opiniones. Por consiguiente, se genera un escenario en el que los usuarios se vuelven mucho más vulnerables a consumir y compartir información falsa la cual reafirme sus propias convicciones.

[1] We are Social & Hootsuite (2022). The global state of digital in October 2022. https://wearesocial.com/es/blog/2022/10/the-global-state-of-digital-in-october-2022/

[2] Statista (2022). Media usage in an online minute 2022. https://www.statista.com/statistics/195140/new-user-generated-content-uploaded-by-users-per-minute/

1.2 Definiciones

De acuerdo a (Lazer et al, 2023) una noticia falsa se puede definir como aquella información que ha sido publicada con la intención de engañar y confundir al lector. Los medios que fabrican estas noticias imitan la forma de publicación de los medios fiables, ignorando los procesos de verificación o directamente adulterando intencionadamente el contenido. De esta manera, pretenden hacer pasar una mentira, una exageración de los hechos, o en el mejor de los casos una opinión, por un hecho objetivo. Esta información puede presentarse en diversos formatos tales como texto, imágenes, audios y videos. Según (Claire Wardle, 2017), podemos clasificar a las noticias falsas en siete categorías:

- Sátira o parodia: tiene como objetivo principal causar gracia al lector, pero algunas personas pueden tomarlo en serio lo cual puede conducir a malentendidos y a la difusión involuntaria de información falsa.
- Conexión falsa: cuando el título y la imagen de la noticia no se corresponden con su contenido.
- Contenido engañoso: este tipo de desinformación puede no ser completamente falsa, pero se presenta de una manera engañosa para tergiversar la realidad. Puede incluir titulares sensacionalistas o información sacada de contexto.
- Contexto falso: el contenido de la noticia puede ser verdadero, pero el contexto en el que se ubica no lo es .
- Contenido impostor: el contenido se hace pasar por una fuente confiable o una entidad legítima. Esto puede incluir sitios web que imitan a medios de comunicación respetables para difundir información falsa.
- Contenido manipulado: se incluye contenido genuino que ha sido editado o manipulado para cambiar su significado con el fin de engañar al usuario.
- Contenido inventado: el contenido es completamente inventado pero se crea deliberadamente para parecer auténtico.

El mayor caudal de noticias falsas suele darse en contextos críticos, tales como la pandemia o periodos electorales, operando activamente sobre temas que generan incertidumbre, logrando fomentar sesgos y prejuicios.

1.2 Antecedentes

La propagación de noticias falsas ha tenido un gran impacto en nuestra sociedad a lo largo de los últimos años. En 2013, la cuenta oficial de twitter de una agencia de noticias estadounidense fue hackeada y usada para difundir información sobre supuestas explosiones en la Casa Blanca, que habían herido al por entonces presidente Barack Obama. Hasta ser desmentido, este tweet generó una gran repercusión en todos los medios y tuvo un impacto directo en el mercado de valores de Wall Street, ocasionando pérdidas de hasta 10 billones de dólares[3]. Durante la pandemia del COVID-19, se divulgaron gran variedad de *fake news* tales como:

- Recomendación de tratamientos caseros no probados contra el virus, como la ingestión de lejía o dióxido de cloro, cuyo consumo es perjudicial para la salud, pudiendo ocasionar daños severos o incluso ser letales.
- Desinformación acerca de las vacunas, circularon diversas afirmaciones acerca de que las vacunas contra el COVID-19 contienen microchips de seguimiento, lo cual genera miedo y desconfianza en el uso de las vacunas, dificultando el alcance de la inmunidad colectiva y el control de la pandemia.
- Negación de la existencia misma del virus, en algunos casos se ha llegado a negar la existencia del COVID-19, tratándolo como una conspiración. La desinformación en este caso, puede llevar a la falta de cumplimiento de los protocolos de salud pública y agudizar así la propagación del virus.

A su vez, se ha demostrado que los contenidos fraudulentos que circulan durante las elecciones ejercen efectos considerables sobre los resultados electorales, tal es el caso de las elecciones presidenciales de Estados Unidos de 2016 (Allcott & Gentzkow, 2017).

[3] CBC (2013). Fake White House bomb report causes brief stock market panic. https://www.cbc.ca/news/business/fake-white-house-bomb-report-causes-brief-stock-market-panic-1.1352024

1.3 Objetivo

En este contexto, para contrarrestar la oleada de desinformación a la que estamos expuestos cada día se vuelve fundamental poder verificar la veracidad de la información consumida en las redes, por lo que el presente proyecto propone como objetivo generar un algoritmo capaz de detectar noticias falsas, utilizando inteligencia artificial. Para lograrlo, el proyecto se llevará a cabo aplicando la metodología CRISP-DM. En primer lugar, se busca generar un set de datos de gran escala recopilando información de noticias falsas y verdaderas provenientes de distintas fuentes como redes sociales, páginas de *fact checking*, y data sets previamente clasificados. Una vez obtenido el conjunto de datos, se busca procesar la información, realizando un proceso de *feature engineering*, para obtener información útil que pueda agregar valor al modelado. Luego, a partir de los datos obtenidos, se procederá a entrenar distintos modelos de aprendizaje automático y/o aprendizaje profundo para la clasificación de noticias falsas, comparando y evaluando sus resultados. Por último, a partir del modelo que arroje mejores resultados, se desarrollará una página web que provea un servicio de detección de noticias falsas en tiempo real.

2. Estado del Arte

En esta sección, se detallarán los avances en el área de la detección de noticias falsas realizados en los últimos años. Se busca comparar los distintos enfoques utilizados, tanto en el set de datos como en los modelos de aprendizaje automático con el objetivo de advertir posibles desafíos y oportunidades existentes en la materia.

2.1 Enfoques para la detección de noticias falsas

Según (Zhou & Zafarani, 2018), en la actualidad existen principalmente cuatro enfoques para la detección de *fake news*, según: *Knowledge-based*, *Style-based*, *Propagation-based* y *Source-based*.

2.1.1 Basadas en conocimiento (*Knowledge-based*)

Cuando se detectan noticias falsas desde una perspectiva basada en el conocimiento, se suele utilizar un proceso conocido como verificación de datos. El *fact checking*, desarrollado inicialmente en el periodismo, tiene como objetivo evaluar la autenticidad de las noticias comparando el conocimiento extraído de la noticia por verificar, como sus afirmaciones o declaraciones, con hechos conocidos. Los hechos conocidos con los que se entrena el modelo son extraídos de forma manual o automática de fuentes fiables especializadas en la verificación de datos.

Esta técnica suele ser de las más utilizadas en la actualidad para la clasificación de noticias falsas, ya que presenta técnicas de extracción de información , procesamiento de lenguaje natural y de aprendizaje automático. No obstante presenta algunas desventajas a la hora de la detección temprana, debido a que las noticias falsas se expanden mucho más rápido que las verdaderas (Vosoughi et al., 2018), esperar a que una fuente desmienta una noticia puede llevar mucho tiempo.

2.1.2 Basadas en el Estilo (*Style-based*)

La detección de noticias falsas basada en el estilo también se centra en analizar el contenido de la noticia. Sin embargo, los métodos basados en el conocimiento evalúan principalmente la autenticidad de la noticia dada, mientras que los métodos basados en el estilo pueden evaluar la intención de las noticias, es decir, si existe la intención de engañar al público o no. La premisa detrás de los métodos basados en estilos es que las noticias falsas cuentan con un formato especial cuya intención es captar la atención del lector. Estudios realizados (Zhou et al, 2019) han demostrado que existen patrones en las noticias falsas que se diferencian de las noticias verdaderas: los textos de noticias falsas, en comparación con los de noticias verdaderas, tienen

mayor grado de informalidad (% de malas palabras), diversidad (% de verbos únicos), subjetividad (% de verbos de reporte) y mayor grado emocionales (% de palabras emocionales). A su vez, estudios realizados utilizando algoritmos de análisis de sentimiento muestran que las noticias falsas suelen inducir respuestas con mayor grado de sorpresa y disgusto que las verdaderas (Vosoughi et al., 2018). Por estas razones, este enfoque suele tener en cuenta características semánticas, sintácticas y léxicas del contenido de la noticia.

2.3 Basadas en la Propagación (*Propagation-based*)

Al detectar noticias falsas desde una perspectiva basada en la propagación, se utiliza información relacionada a cómo estas fueron difundidas. Los algoritmos utilizan técnicas de análisis de redes sociales para identificar patrones de difusión que puedan indicar la presencia de noticias falsas o información engañosa. En este contexto, el estudio de las relaciones entre los usuarios y su comportamiento en las redes toman un papel preponderante. Los modelos basados en este enfoque le dan especial importancia a datos como la información del autor, la hora de publicación, la cantidad de visualizaciones, respuestas y repercusiones.

2.4 Basadas en la Fuente (*Source-based*)

Este enfoque está orientado a evaluar las noticias falsas de acuerdo a la credibilidad de la fuente, está credibilidad se evalúa de acuerdo con el autor, editor y usuarios que podrían haber desarrollado o publicado la noticia. La aproximación se basa en la hipótesis de que si una fuente pública noticias falsas, tiene gran probabilidad de seguir publicando noticias de este estilo. Esto parece razonable, ya que existen gran cantidad de sitios web diseñados específicamente para la publicación de engaños o sitios hiper partidistas que se presentan como publicadores de noticias reales, así como usuarios que actúan como "trolls" y bots que contribuyen de manera sutil pero perjudicial a la difusión de noticias. Si bien los modelos basados en este enfoque puede padecer de errores como falsos positivos (una fuente confiable que publica una noticia falsa por error) y falsos negativos (una fuente dudosa que publica una noticia verdadera), ha demostrado ser una técnica bastante eficaz en la detección de noticias falsas.

Además del enfoque, otro aspecto de interés a la hora de analizar el estado del arte es si los modelos se centran en realizar una clasificación precisa o una detección temprana. La clasificación precisa busca la precisión en la determinación de la veracidad, mientras que la detección temprana se centra en la velocidad y la prevención. Muchos de los trabajos estudiados han desarrollado algoritmos que cuentan con una gran capacidad para la clasificación de noticias, pero dejan de lado la detección temprana . Dado que las noticias falsas se extienden mucho más rápido que la verdad, generando un impacto negativo en nuestra sociedad, es de suma importancia considerar ambos aspectos.

2.2 Trabajos Relacionados

A continuación, se analizan distintos *papers* y artículos académicos dedicados a los avances en el área de la detección de noticias falsas.

2.2.1 Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities

Este artículo, publicado en 2019 por la Delhi Technological University, realiza un estudio del estado del arte en la detección de noticias falsas. En él se lleva a cabo un análisis y comparación de gran variedad de trabajos relacionados, prestando atención a los distintos set de datos utilizados, sus *features*, las redes sociales de donde se obtiene la información y los modelos de inteligencia artificial aplicados.

En cuanto a los set de datos, podemos encontrar que los trabajos utilizan distintos *features* para la detección de *fake news*, algunos de estos son: información del usuario, tópico, tiempo, propagación, análisis de sentimiento, análisis de texto e información lingüística. Además, cabe destacar que todos los trabajos utilizan sus set de datos en inglés.

Por otro lado, podemos dividir a los trabajos según utilizan técnicas de *machine learning* o de *deep learning* para el modelado. Entre los algoritmos de *machine learning* usados para la clasificación encontramos Naive Bayes, Support Vector Machines, K-nearest neighbors, KStar, Decision Tree, Decision Rule, Random Forest, Logistic Regression, Linear Regression y Stochastic Gradient Descent, mientras que entre los de *deep learning* aparecen Convolutional Neural network, Recurrent Neural Network, Multilayer Perceptron, entre otros.

Por último, si bien se han realizado gran cantidad de avances en el área en los últimos años, el artículo destaca que aún existen muchas oportunidades de investigación por explotar. Algunas de estas oportunidades son las siguientes:

- Detección Multiplataforma: la gran mayoría de los algoritmos se centran en una única plataforma o red social, dado que una misma noticia falsa se puede expandir de una red a otra, esto dificulta la tarea de descubrir la fuente de origen. Por lo tanto, generar algoritmos que detectan noticias falsas entre distintas plataformas resulta muy importante .
- Aprendizaje en tiempo real: Implementar una aplicación web para verificación de hechos que puede aprender en tiempo real de nuevos artículos verificados manualmente y proporcione detección en tiempo real de información fraudulenta.

- Modelos no Supervisados: el trabajo actual se realiza principalmente mediante el uso de enfoques de aprendizaje supervisado, las noticias cuentan con una label que indica si es verdadera o falsa. Dado que el etiquetado manual de las noticias es una tarea ardua, el desarrollo de modelos no supervisados resulta crucial debido a la gran cantidad de información no etiquetada en la internet.
- Datasets: Los conjuntos de datos son fundamentales ya que en ellos reside la calidad de cualquier modelo. La mayor parte de las investigaciones se realizan con conjuntos de datos personalizados. Debido a la falta de conjuntos de datos a gran escala disponibles públicamente como punto de referencia resulta muy difícil poder realizar una comparación fiable entre los diferentes trabajos.
- Modelos de Detección para Distintos Idiomas: todos los trabajos estudiados en este artículo utilizan conjuntos de datos en inglés y sus características lingüísticas dejando de lado los distintos idiomas.
- Detección Temprana: poder detectar noticias falsas en una etapa temprana es una tarea fundamental, una vez estas se viralizan entre los usuarios, pueden generar un fuerte impacto trayendo consecuencias negativas en nuestra sociedad, incluso en algunos casos resulta difícil volver a cambiar la percepción de estas personas.
- Análisis entre dominios: la mayor parte de las investigaciones existentes se centra solo en una forma de detección de engaño: según el estilo, conocimiento, propagación o fuente. Poder utilizar los distintos enfoques unidos puede ayudar a mejorar la detección.

2.2.2 Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts

Este paper publicado en 2015 resulta muy interesante ya que ofrece un enfoque diferente de los estudiados anteriormente, este trabajo no se centra en la clasificación de noticias como falsas o verdadera sino que propone un algoritmo para la detección rápida de rumores sobre la red social Twitter, analizando el potencial que estos tienen de viralizarse. El supuesto principal sobre el que se fundamentan es que los *tweets* falsos o controversiales tienden a ser respondidos por preguntas, debido a que los usuarios buscan verificar dicha información. Bajo estas condiciones, utilizan la frecuencia de preguntas como indicador de credibilidad. Por lo tanto, usan palabras y frases claves (por ejemplo, *Is this true?, What?, Really?*) como sensores para detectar posibles rumores. De esta manera proponen una manera simple y eficiente para la detección temprana de posibles noticias falsas.

2.2.3 Early Detection of Fake News by Utilizing the Credibility of News, Publishers, and Users Based on Weakly Supervised Learning

Este trabajo trae un enfoque novedoso para detectar noticias falsas en las primeras etapas de su difusión utilizando el aprendizaje débilmente supervisado. El enfoque propuesto, llamado SMAN, combina la información de credibilidad de los usuarios y los editores de noticias con el contenido de las noticias para lograr un alto rendimiento en la detección de noticias falsas.

Por otro lado utiliza algunas ecuaciones del álgebra para formalizar el proceso de aprendizaje y predicción de la credibilidad de los usuarios y editores de noticias. Por ejemplo, se utilizan matrices y vectores para representar las características de los usuarios y editores, y se aplican operaciones de multiplicación de matrices y sumas ponderadas para obtener las puntuaciones de credibilidad. Además, se utilizan funciones de activación como ELU para transformar las características de entrada y mejorar el rendimiento del modelo.

2.2.4 Fake news detection with petrained transformers

Este trabajo de fin de máster analiza la eficacia de los modelos de lenguaje pre-entrenados basados en transformers en la detección automática de noticias falsas. Se presentan dos enfoques: un enfoque unimodal que trata la información adicional de la metadata como un elemento más del lenguaje y un enfoque híbrido multimodal en el que se hace uso de BERT para procesar la información textual mientras que para el resto de características adicionales se hace uso de técnicas de machine learning tradicionales.

Concluyendo que adaptar los enfoques de detección de noticias falsas a cada contexto comunicativo es importante, teniendo en cuenta las particularidades lingüísticas y estilísticas presentes en cada modalidad de comunicación. Además, se observa que el uso de las justificaciones de las noticias en el modelo es beneficioso para la clasificación únicamente cuando la longitud de estas justificaciones no es considerablemente mayor que el resto de cadenas con las que se procesa simultáneamente.

2.2.5 Modelo para la detección de noticias falsas en formato texto en la red social Twitter, aplicado al contexto político colombiano de las elecciones presidenciales de 2022

Este trabajo de grado de la Universidad Icesi de Cali, publicado en 2022 propone un modelo para la detección de noticias falsas en el contexto político colombiano, relacionado a las elecciones presidenciales de dicho año. Crea su propio set de datos en español y utilizan un enfoque orientado en el estilo para la detección. Además, comparan distintos algoritmos de machine learning tradicional y se quedan con el de mejor *accuracy.* Si bien el set de datos generado cuenta con muy pocos registros, tan solo 634, se destaca de este trabajo la búsqueda de generar un algoritmo para la detección de noticias falsas en el idioma español.

3. Problema detectado y/o faltante

Como se mencionó a lo largo de las secciones anteriores, las noticias falsas son nocivas para nuestra sociedad, afectan la opinión pública y polarizan las opiniones. En estas condiciones, resulta esencial contar con herramientas capaces de combatirlas. La importancia de este fenómeno es tal, que a lo largo de los últimos años se han realizado gran variedad de trabajos al respecto. No obstante, consideramos que aún existen grandes carencias en los modelos realizados hasta el momento, entre ellas destacamos:

- Ausencia de Conjuntos de Datos en Español: la mayoría de los estudios hasta la fecha utilizan sets de datos en el idioma inglés, mientras que los escasos que están en español cuentan con poca variedad de *features* y registros insuficientes.

- Ausencia de modelos capaces de detectar Noticias Falsas en Español: al no contar con conjuntos de datos en español para el entrenamiento, los modelos existentes se vuelven inútiles o incluso incapaces de detectar noticias falsas en este idioma.

- Ausencia de Modelos que utilicen distintos Enfoques de Detección: los modelos propuestos al día solo tienen en cuenta uno de los cuatro enfoques de detección estudiados en la sección anterior. Crear modelos capaces de combinar estos enfoques podría aportar mayor robustez a la tarea de detectar noticias falsas.

- Ausencia de servicios para la detección de Noticias Falsas: si bien existen diversas fuentes especializadas en el *fact-checking*, estas fuentes suelen verificar los datos manualmente lo cual puede demorar mucho tiempo. La existencia de una página web para la verificación de hechos en tiempo real y de manera automática puede ser de gran utilidad para los usuarios.

4. Solución propuesta

Dados los problemas detectados, este trabajo propone como solución realizar un proceso de minería de datos, aplicando la metodología CRISP-DM, para obtener un modelo capaz de detectar noticias falsas en español, haciendo hincapié en las noticias en formato de texto, dejando de lado las imágenes, audios y videos.

En primer lugar, se busca generar de un set de datos de gran escala en idioma español, el cual contenga una amplia variedad de noticias etiquetadas como falsas o verdaderas, junto con diversidad de *features* que agreguen valor relevante como la información de la fuente, la repercusión de la noticia, entre otras. Los datos utilizados serán recopilados a partir de distintas fuentes, como datasets preexistentes orientados a noticias falsas, datos recopilados de páginas de *fact-checking* y datos recolectados de redes sociales, utilizando la técnica de *scraping*. Luego de obtener los datos, se busca aplicar distintos algoritmos de análisis de sentimientos y procesamiento de lenguaje natural para obtener aún más atributos interesantes sobre los mismos Para el análisis de sentimiento se extraerá una clasificación acotada de los sentimientos a los cuales la noticia quiere apelar. Por otro lado, a partir del procesamiento del lenguaje natural también se analizaran características léxicas como la cantidad total de palabras, cantidad de palabras únicas, etc. junto con otras propiedades sintácticas como la puntuación, etiquetado POS (Part-of-Speech), técnicas de modelo de bolsa de palabras (Bag of Words) y n-gramas.

Una vez obtenido el *dataset*, se probarán diversos modelos en el estado del arte del aprendizaje automático y/o aprendizaje profundo, comparando los resultados obtenidos a través de métricas para evaluar el mejor modelo. Se busca generar modelos capaces de utilizar más de un enfoque para la detección de noticias falsas, en particular mixtos entre los basados en estilo y los basados en la fuente. El enfoque basado en el estilo, se centrará en el contenido de la noticia, tanto en el análisis de sentimiento como en sus características lingüísticas. Mientras que los basados en la fuente se orientan en la credibilidad de la misma a partir de la identificación de patrones. Se analizará tanto la fuente como el contenido de las noticias, ajustando las variables necesarias para que la clasificación sea la más precisa posible, realizando finalmente un estudio comparativo entre todos los modelos. No obstante, se elegirá el enfoque más adecuado una vez realizado un análisis preliminar de la información adquirida en las primeras etapas del proyecto, sin descartar la posibilidad de utilizar otros enfoques.

Por último, a partir de los mejores modelos obtenidos se desea implementar una aplicación web para la verificación automática de hechos en tiempo real. La idea es generar una página web que cuente con las siguientes funcionalidades:

- Análisis y verificación de noticias por parte del usuario: Esta herramienta permitirá a los usuarios interactuar con el modelo, insertar su propia noticia en formato de texto para poder así comprobar su veracidad. El resultado obtenido será una escala de fidelidad del contenido, junto con un análisis estadístico de la intencionalidad de la noticia, denotando al sentimiento que intenta apelar.

- Análisis de narrativas en tiempo real: Se consumirá información en tiempo real tanto redes sociales como de medios de comunicación, para poder estudiar cuales son los tópicos de mayor relevancia en la actualidad con el objetivo de identificar posibles narrativas desinformantes. Al mismo tiempo, dado que esta funcionalidad se ejecuta en vivo puede ser de gran ayuda para la detección temprana de noticias falsas.

- Panel de tendencias: Esta sección realizará un monitoreo en tiempo real de distintas fuentes tales como redes sociales (Twitter/X, Telegram, entre otras) así como también plataformas de noticias de la región. Esta sección proveerá estadísticas sobre las tendencias tanto del día como las semanales, haciendo un análisis de propagación de los contenidos que ofrecen estas plataformas.

Es importante destacar, que en todos los casos los resultados obtenidos de cada sección serán justificados, explicando el método o razonamiento por el cual fueron obtenidos.

5. Alcance

Esta sección tiene como objetivo establecer de manera específica los límites y metas del proyecto. Es importante destacar que el alcance del mismo puede estar sujeto a modificaciones a medida que avancemos en la primera etapa de la investigación. La recopilación y el análisis de datos iniciales desempeñarán un rol clave en la definición de futuras direcciones y ajustes en el proyecto. A medida que se adquiera un mayor entendimiento sobre el comportamiento de las noticias y la efectividad de los algoritmos desarrollados, se podrían considerar cambios, adiciones o ajustes en los objetivos del proyecto. A continuación, se detallan los elementos que se incluyen en el alcance del proyecto y aquellos que se excluyen.

5.1 Inclusiones:

- Generación de Set de Datos de Noticias en Español: Se llevará a cabo la creación de dataset de noticias en español mediante diversas técnicas de recopilación de datos. Los datos serán recopilados a partir de fuentes de noticias en línea, redes sociales y bases de datos de verificaciones de hechos.
- Desarrollo de Algoritmo de Análisis de Sentimiento: Se implementará un algoritmo de análisis de sentimiento con el propósito de generar una clasificación acotada de los sentimientos a los que una noticia apela.
- Implementación de Algoritmo de Detección de Noticias Falsas en Español: Se desarrollará un algoritmo con inteligencia artificial para la de detección de noticias falsas en español, el cual se basará en el análisis de estilo de las noticias, utilizando Procesamiento de Lenguaje Natural (NLP) y el algoritmo de análisis de sentimiento mencionado previamente. También se abordarán enfoques como el de análisis de la fuente para luego determinar a partir de los resultados obtenidos el mejor método a utilizar. Este algoritmo será capaz de analizar el contenido de las noticias, identificar patrones de desinformación y generar puntajes de veracidad.
- Diseño y Desarrollo de una Página Web: Como interfaz de usuario se desarrollará una página web que permita interactuar con el modelo obtenido para analizar la veracidad de las noticias a través de su contenido de texto. La página web también ofrecerá análisis de narrativas en tiempo real y contendrá un panel de tendencias de distintos medios. Esta interfaz permitirá a los usuarios cargar enlaces a noticias, visualizar los puntajes de veracidad generados por el algoritmo y acceder a estadísticas sobre la noticia.

5.2 Exclusiones:

- Análisis de Detección de Noticias Falsas en formato Imagen, Audio o Video: No se abordará el análisis de la detección de noticias falsas en formato de imágenes, audio o video en este proyecto.
- Detección de Noticias Falsas en Múltiples Idiomas: El enfoque se centrará en la detección de noticias falsas en el idioma español, y no se incluirá el análisis en otros idiomas.
- Análisis de Detección de Noticias Falsas basado en la Propagación: La detección de noticias falsas basada en la propagación de información no estará dentro del alcance de este proyecto.
- Análisis de Detección de Noticias Falsas basado en el Conocimiento: La detección de noticias falsas basada en el conocimiento no será parte de los objetivos de este proyecto.
- Validación del Algoritmo en Producción: La validación y despliegue del algoritmo a escala de producción, con altas cargas de peticiones en tiempo real, no son parte de los objetivos de este proyecto.

6. Herramientas y Tecnologías

A continuación, se detallan algunas de las herramientas clave que se utilizarán durante el desarrollo del proyecto como el lenguaje de programación principal, frameworks relevantes, sistema de control de versiones, entre otros.

- Python: Se selecciona Python como lenguaje de programación principal debido a su amplio conjunto de bibliotecas para el desarrollo de modelos de inteligencia artificial, además de ser uno de los lenguajes más utilizados en este área. Se utilizará tanto para la recopilación de datos junto con su procesamiento adecuado como para el modelado de los algoritmos de inteligencia artificial. A su vez, se usará para el desarrollo del servidor backend.

- React y JavaScript: Estas herramientas se usarán para la implementación de la interfaz de usuario y el frontend de la aplicación web. React es una biblioteca de JavaScript ampliamente adoptada que permite la creación de aplicaciones web interactivas.

- Figma: Figma es una herramienta de diseño web colaborativo que se utilizará para la creación de la interfaz de usuario.

- GitHub: Se utilizará como plataforma de control de versiones para el seguimiento y la colaboración en el desarrollo del proyecto.

- MongoDB: Como base de datos NoSQL, MongoDB proporciona una solución flexible para almacenar y gestionar datos, incluidos textos y metadatos recopilados. Su capacidad de escalabilidad horizontal será beneficiosa para el manejo de grandes conjuntos de datos.

- Trello: Sirve como plataforma de organización y gestión de proyectos. Ayuda en la asignación de tareas, el seguimiento del progreso y la priorización de actividades. Esto garantiza una gestión efectiva del proyecto y un cumplimiento oportuno de los objetivos.

- Grafana: Para la visualización de métricas y resultados del sistema, se utilizará Grafana. Esta plataforma de código abierto permite crear paneles de control personalizados y visualizar datos en tiempo real, lo que ayudará a evaluar el rendimiento del algoritmo de detección de noticias falsas.

7. Experimentación y/o validación

Para evaluar los distintos modelos de *machine learning* y/o *deep learning* propuestos en este proyecto se hará uso de distintas métricas para comparar la calidad de los resultados obtenidos. En primer lugar, dado que se trata de un problema de clasificación, se usará la accuracy para tener un tener una medida general del rendimiento del modelo. La *accuracy* o exactitud, es una métrica que mide la proporción de predicciones correctas realizadas en relación con el total de predicciones.

$$accuracy = \frac{\#\ predicciones\ correctas}{\#\ predicciones\ totales}$$

La exactitud es útil como métrica general de rendimiento cuando las clases en el set de datos están balanceadas. Sin embargo, puede no ser adecuada si las clases están desequilibradas, ya que un modelo que predice siempre la clase mayoritaria puede tener una alta exactitud, pero un rendimiento deficiente en la detección de la clase minoritaria. Por esta razón, también se utilizarán otras técnicas para el evaluado, como K-Fold Cross Validation. Esta técnica consiste en dividir el conjunto de datos en K particiones o "folds", y luego entrenar y evaluar el modelo K veces, utilizando una partición diferente como conjunto de prueba en cada iteración. Esto permite una evaluación más confiable del rendimiento del modelo al considerar múltiples divisiones de los datos. Además, es útil para estimar el comportamiento de un modelo con nuevos datos y ayuda a detectar problemas tanto de *overfitting* como de *underfitting*.

Por último, también se analizará la Curva ROC (Receiver Operating Characteristic Curve), esta es la representación gráfica de la efectividad del modelo binario de clasificación. Se realiza graficando la tasa de verdaderos positivos frente a la tasa de falsos positivos en diferentes umbrales de clasificación. Si se mide el área bajo la curva (AUC) se obtiene una medición del rendimiento general del modelo de clasificación binaria.

8. Plan de actividades

La metodología CRISP-DM (**CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining) es una modelo que sirve de base para los procesos de ciencia de datos. Esta consiste en seis fases, en donde la sucesión de las mismas no es rígida, se permite (y requiere) el movimiento entre ellas. El resultado de cada fase determina los pasos a seguir.

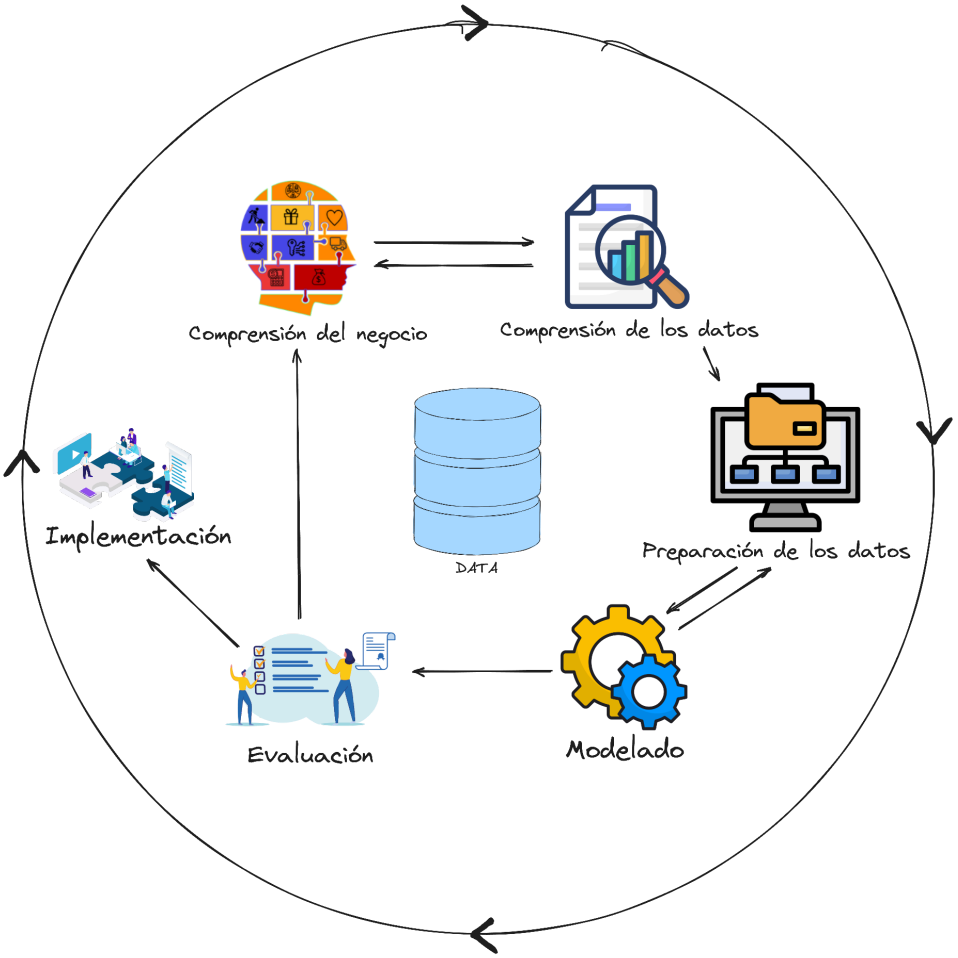

Figura 1: Diagrama de la metodología CRISP-DM

La misma consta de 6 fases:

- Comprensión del negocio: entendimiento exhaustivo de los objetivos y requerimientos del proyecto, determinar la definición técnica del problema y producir un plan de proyecto.
- Comprensión de los datos: esta etapa se enfoca en identificar, recolectar y analizar los datos obtenidos, también se verifica la calidad de los mismos.
- Preparación de los datos: se preparan los datos para el modelado, esto incluye selección, limpieza, estructuración, formateo e integración.

- Modelado: esta etapa consiste en la selección de técnicas de modelado, generación del diseño de las pruebas, armado y evaluación del modelo.
- Evaluación: se evalúan los resultados obtenidos, se realiza un proceso de revisión en el cual se corrigen los errores en caso de ser necesario, y por último se determinan los pasos a seguir.
- Implementación: en esa etapa se realiza el plan de implementación, el plan de monitoreo y mantenimiento, un reporte final y la evaluación del proyecto.

Como parte de este proyecto, se tiene la intención de incorporar una plataforma visual orientada al usuario, lo que conlleva una etapa de "desarrollo web" en la que se planifica y crea un sitio web.
Al implementar CRISP-DM de manera flexible, iterando de manera rápida y eficiente, en conjunto con otras metodologías ágiles, se obtiene un enfoque ágil de trabajo.
Es por esta razón que para complementar CRISP-DM se seguirá un marco de trabajo ágil, el cual consiste en una versión simplificada de Scrum[4], que posee ciclos iterativos conformados por los siguientes eventos/ceremonias de equipo:

- Organizar el *backlog*[5]: como equipo se deben elegir las tareas próximas a realizar, las cuales deben estar ordenadas por prioridad.
- Sprint planning: el equipo estima la dificultad de cada tarea y se evaluará en cuáles estará centrado el sprint.
- Sprint: se trabaja en las tareas pendientes establecidas durante la planificación.
- Sprint Retrospective: revisión del sprint, se discute lo que se podría mejorar para el siguiente. Se genera la documentación correspondiente a los procesos, diseño y especificaciones técnicas de lo realizado.

Los *sprint* tendrán una duración de 2 semanas cada uno, y la herramienta utilizada para organizar los ítems de cada uno de ellos será el software Trello, la cual permite administrar las tareas del *backlog* y del *sprint*, marcando el estado en que se encuentran (en desarrollo, en revisión, etc.) para que todo el equipo se encuentre al tanto de las mismas.

A su vez, cada 2 semanas se tendrá una reunión sincrónica con el tutor con el objetivo de tener un seguimiento continuo del progreso del proyecto y discutir cualquier pregunta o inquietud que pueda surgir durante ese período.

[4] Scrum: marco de gestión de proyectos de metodología ágil que ayuda a los equipos a estructurar y gestionar el trabajo mediante un conjunto de valores, principios y prácticas.
[5] Backlog: lista priorizada de todas las tareas y mejoras que deben realizarse en un proyecto de desarrollo de software.

Utilizaremos Git y Github como software de control de versiones para tener organizados los avances que se vayan realizando a lo largo del proyecto. Asimismo, se hará uso del método de flujo de trabajo *feature branch development,* en donde se crean ramas separadas para cada característica o tarea y una vez finalizadas se fusiona a la rama principal requiriendo que al menos un integrante más del equipo valide el trabajo a fusionarse con el método de Github llamado Pull Request.

Se realizó un estudio y una organización de las horas que serán destinadas para la realización del trabajo profesional, así como también un diagrama de Gantt para la división de tareas del equipo, detallando el tiempo destinado a cada etapa y fase del modelo CRISP que se llevará a cabo. En el Anexo 1, se encuentran los enlaces a los documentos de planificación de horas junto con un diagrama de Gantt más detallado.

Figura 2: Diagrama de Gantt de la planificación del proyecto.

Como se observa en la figura, la duración estimada para el desarrollo del proyecto es de veintisiete semanas, dividiéndolo en ocho fases acorde a la metodología utilizada. La fase cero se corresponde con la redacción del presente documento, mientras que las fases uno a seis están asociadas a las fases de la metodología CRISP-DM: Comprensión del Negocio, Comprensión de los Datos, Preparación de los Datos, Modelado, Evaluación e Implementación respectivamente. Por último, se agregó una fase llamada Finalización para corregir y modificar cualquier inconveniente que pueda presentarse en la etapa final del proyecto, junto con los requerimientos propios para la entrega del mismo.

A su vez, se hizo un estudio de los riesgos que se pueden presentar a lo largo del desarrollo, junto con los planes de contingencia y mitigaciones para contrarrestar dichas eventualidades. La tabla con las especificaciones se encuentra asimismo en el Anexo 1.

9. Referencias

1. We are Social & Hootsuite (2022). The global state of digital in October 2022. https://wearesocial.com/es/blog/2022/10/the-global-state-of-digital-in-october-2022/
2. Statista (2022). Media usage in an online minute 2022. https://www.statista.com/statistics/195140/new-user-generated-content-uploaded-by-users-per-minute/
3. David M. J. Lazer, Matthew Baum, Yochai Benkler, Adam J. Berinsky, Jonathan L. Zittrain (2018). The science of fake news. *SCIENCE.* (Vol 359, Issue 6380 , pp. 1094-1096)
4. Wardle, Claire (2020). Understanding Information Disorder. https://firstdraftnews.org/long-form-article/understanding-information-disorder/
5. CBC (2013). Fake White House bomb report causes brief stock market panic. https://www.cbc.ca/news/business/fake-white-house-bomb-report-causes-brief-stock-market-panic-1.1352024
6. Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* (Vol. 31, Issue 2, pp. 211–236).
7. Zhou, X., & Zafarani, R. (2018). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Surveys* (Vol. 53, Issue 5, pp 1–40)
8. Vosoughi, S., Roy, D., & Aral, S. (2018). False news is big news. the spread of true and false news online. *SCIENCE.* (Vol 359, Issue 6380, pp. 1146-1151)
9. Xinyi Zhou, Atishay Jain, Vir V Phoha, and Reza Zafarani. (2019). Fake News Early Detection: A Theory-driven Model. *Digital Threats: Research and Practice*. (Volume 1, Issue 2, pp 1–25)
10. Meel, P., Kumar Vishwakarma, D. (2019). Fake News, Rumor, Information Pollution in Social Media and Web: A Contemporary Survey of State-of-the-arts, Challenges and Opportunities. Expert Systems with Applications (Vol 153)
11. Zhao, Z., Resnick, P., Mei, Q. (2015) Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts.: https://doi.org/10.1145/2736277.2741637
12. Yuan, C., Ma, Q, Zhou, W. (2020) Early Detection of Fake News by Utilizing the Credibility of News, Publishers, and Users based on Weakly Supervised Learning. 10.18653/v1/2020.coling-main.475
13. Dominguez, V. (2023) *Fake news detection with pre trained transformers* (Tesis de maestría). Universidad Politécnica de Madrid. Madrid, España.

14. Flores Quinayás, J.E., Montaño Morcillo, J.G. (2022). Modelo para la detección de noticias falsas en formato texto en la red social Twitter, aplicado al contexto político colombiano de las elecciones presidenciales de 2022 (Tesis de maestría). Universidad ICESI. Santiago de Cali, Colombia.

10. Anexos

Anexo 1: Plan de Actividades

Para el acceso a los siguientes documentos es necesario contar con un correo electrónico de la facultad, dominio @fi.uba.ar.

- Análisis de Riesgo: Enlace Análisis de Riesgo
- Diagrama de Gantt: Enlace Diagrama de Gantt
- Planificación de horas: Enlace Planificación de Horas